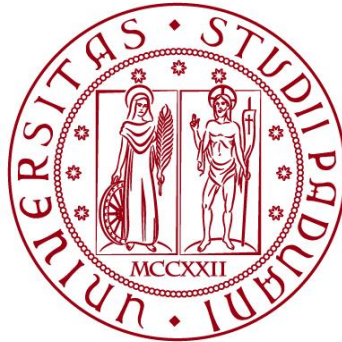


UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea in Biologia Molecolare



ELABORATO DI LAUREA

ANALISI AD ALTA PROCESSIVITÀ DI GENOMI MICROBICI A SINGOLA CELLULA CON RISOLUZIONE A LIVELLO DI CEPPO

Tutor: Prof. Stefano Campanaro
Dipartimento di Biologia

Co-tutor: Dott. Flavio Agostini
Dipartimento di Biologia

Laureando: Alessandro Vettore

ANNO ACCADEMICO 2025/2026

Abstract

Lo studio delle comunità microbiche complesse rappresenta, da sempre, una grande sfida per la biologia. Questa tesi presenta un metodo di analisi genomica single-cell ad alta processività, che è in grado di identificare le specie con risoluzione a livello di ceppo, senza la necessità di coltivare i microrganismi. Anziché applicare la più diffusa tecnica di metagenomica shotgun, viene introdotto un sistema di microfluidica in grado di incapsulare singole cellule in droplet, all'interno delle quali il DNA viene liberato e successivamente amplificato, marcato con barcode droplet-specifici e sequenziato. I single-amplified genomes (SAGs) così ottenuti vengono co-assemblati in genomi per consentire l'identificazione delle specie e dei ceppi. Analizzando il microbiota intestinale umano, Zheng et al., hanno ottenuto 20,000 SAGs partendo da un singolo donatore, i quali sono stati co-assemblati a livello di specie generando circa 100 genomi, alcuni dei quali hanno evidenziato la presenza di diversi ceppi.

Lo studio è stato approfondito durante il periodo di tirocinio curriculare, svolto presso il laboratorio di Genomica e Bioinformatica, con la realizzazione di una pipeline in Python in grado di raggruppare i SAGs su base Average Nucleotide Identity e Jaccard Index al fine di identificare le varie specie presenti in un bioreattore oggetto di indagine.

Indice

1 Stato dell'arte	1
1.1 Comunità microbiche e sequenziamento single-cell	1
1.2 Analisi computazionali per l'identificazione delle specie	2
2 Metodi.....	3
2.1 Ottenimento dei campioni.....	3
2.2 Sistema di microfluidica	3
2.3 Sequenziamento	3
2.4 Processamento dei dati grezzi e co-assembly.....	4
2.5 Identificazione dei ceppi.....	5
2.6 FastANI e Sourmash.....	6
2.7 Community detection.....	6
3 Risultati	10
3.1 FastANI.....	10
3.2 ANI stimato.....	12
3.3 Jaccard Index	14
3.4 SBM	16
4 Discussione.....	17
5 Bibliografia	19

1 Stato dell'arte

1.1 Comunità microbiche e sequenziamento single-cell

Le comunità microbiche si ritrovano in tutti gli ecosistemi naturali e rappresentano un importante caso di studio, in particolare per la caratterizzazione delle specie che vi risiedono. Lo studio di Zheng et al. (2022), si concentra sul microbiota intestinale umano, popolato da trilioni di microbi. La caratterizzazione del microbiota intestinale risulta di particolare rilevanza in quanto è associato alla salute umana e a diverse malattie che comprendono disturbi metabolici, cognitivi e malattie autoimmuni. Si rende quindi evidente come la distinzione delle diverse specie presenti nel microbiota e la caratterizzazione delle loro funzioni e interazioni costituisca un importante ambito di ricerca per la biomedicina.

La composizione del microbiota è specifica per ogni individuo e, anche se le specie presenti sono molto spesso simili tra persona e persona, ogni individuo presenta ceppi microbici caratteristici, contraddistinti da peculiari differenze genomiche [1], che possono portare in sé alcuni tratti distintivi come resistenza agli antibiotici ed interazioni con il sistema immunitario.

La conoscenza di come i diversi microbi che popolano l'intestino guidino i comportamenti complessivi della flora batterica parte, innanzitutto, dalla conoscenza del genoma di questi microrganismi. Questo è reso possibile grazie alla ricostruzione di genomi a singola cellula ad alta qualità che permettono la risoluzione a livello di ceppo.

Uno degli approcci maggiormente utilizzati per esplorare il microbioma è la metagenomica shotgun, che prevede la lisi e il sequenziamento di intere comunità microbiche. Questa tecnica, però, non è spesso in grado di permettere una risoluzione a livello di ceppo, in quanto non è efficace nell'assegnare con sicurezza sequenze di DNA comuni a più taxa, soprattutto quando si tratta di specie con molti ceppi o quando più sequenze omologhe sono presenti in taxa diversi [2].

In alternativa, genomi ad alta qualità con risoluzione a livello di ceppo sono ottenibili dalle colonie di batteri coltivati *in vitro* [1]; questa tecnica, tuttavia, può risultare molto laboriosa e non funzionale, in quanto la maggior parte delle specie batteriche non sono coltivabili.

Un'altra tecnica consiste nell'isolare singole cellule o una dozzina di esse da un pozzetto o una piastra, amplificarne il contenuto genomico e infine sequenziarlo. Per tutte le tecniche fin qui descritte, la quantità di materiale biologico è un fattore fortemente limitante nella risoluzione del genoma a livello di ceppo e rende queste tecnologie poco processive.

Un modo per superare queste limitazioni è l'utilizzo di un sistema di microfluidica in grado di incapsulare singole cellule in piccoli droplet, permettendo l'amplificazione e il sequenziamento del DNA [3]. Questa tecnologia nasce per il sequenziamento single-cell di cellule di mammifero, in cui ogni cellula è incapsulata e lisata per liberarne il materiale genetico, che viene poi amplificato e marcato. Tuttavia, questa tecnica applicata ai batteri richiede diversi step di microfluidica, i quali sono stati eseguiti singolarmente e mai combinati in un workflow unificato. La tecnologia prevede l'uso di sistemi di microfluidica per

incapsulare singoli batteri in droplet, all'interno dei quali le cellule sono lisate e il DNA è amplificato e marcato per poter essere sequenziato. Grazie a questa tecnica, Zheng et al. hanno ottenuto 21,914 SAGs partendo da sette campioni longitudinali di feci acquisiti da un singolo soggetto sano. I SAGs sono stati raggruppati e co-assemblati in 76 specie diverse, dieci delle quali risultano essere ulteriormente diversificate a livello di ceppo.

1.2 Analisi computazionali per l'identificazione delle specie

Durante il periodo di tirocinio curricolare svolto presso il laboratorio di Genomica e Bioinformatica, lo studio di Zheng et al. è stato approfondito attraverso la realizzazione di una pipeline in Python volta a identificare le specie presenti all'interno di un bioreattore oggetto di studio. Partendo dai SAGs ottenuti attraverso il sistema di microfluidica sopra citato e già assegnati ad una tassonomia, il lavoro che è stato svolto ha previsto la comparazione dei SAGs tra loro al fine di raggrupparli per poterne co-assemblare il genoma a livello di specie.

Il fulcro di questo lavoro è la Community Detection, processo matematico in grado di identificare comunità all'interno di sistemi complessi, molto spesso rappresentati da grafici a rete con nodi e archi [4].

La comparazione dei SAGs è stata svolta attraverso l'implementazione di due diverse metriche: l'Average Nucleotide Identity (ANI), percentuale di identità media dei nucleotidi condivisi dalle sequenze di DNA, e il Jaccard Index, un indice statistico utilizzato per il confronto tra campioni diversi, calcolati rispettivamente tramite i software FastANI [9] e Sourmash [8]. Grazie a questi indici sono state costruite delle reti di similarità genomica, in cui ogni nodo rappresenta un SAG e gli archi connettono coppie di nodi solo nel caso in cui la similarità tra i SAGs superi una determinata soglia. Per l'ANI la soglia è stata fissata al 95% di similarità, classica soglia oltre la quale due genomi sono considerati appartenenti alla stessa specie, mentre per il Jaccard Index, la soglia è stata fissata a 0.09, valore valutato empiricamente per massimizzare le metriche descrittive delle varie reti. Le reti sono state analizzate tramite algoritmi di Community Detection, quali Louvain e Leiden [5], Markov Clustering (MCL) [4] e Stochastic Block Model (SBM) [6], i quali sono stati valutati in base alla loro capacità di creare cluster completi e coerenti che a loro volta rappresentano le differenti specie che popolano la comunità.

I risultati ottenuti dal clustering hanno evidenziato la presenza di una quarantina di cluster distinti, con i SAGs appartenenti alla specie *Coprothermobacter proteolyticus*, la più abbondante all'interno del campione in analisi, divisi a metà all'interno di due cluster, suggerendo la probabile presenza di due ceppi diversi di questa specie.

2 Metodi

2.1 Ottenimento dei campioni

Nello studio di Zheng et al., i campioni di feci sono stati ottenuti da OpenBiome, una banca dati senza scopo di lucro. I campioni sono stati raccolti da un uomo sano di 28 anni, sottoposto a screening per escludere la presenza di patogeni contaminanti.

È stata creata una mock community coltivando quattro ceppi batterici noti, *Bacillus subtilis* ATCC 6051-U, *Escherichia coli* ATCC 25922, *Klebsiella pneumoniae* ATCC 35657 e *Staphylococcus aureus* ATCC 6538, in 1 mL di terreno LB liquido, al fine di testare le performance della pipeline implementata.

2.2 Sistema di microfluidica

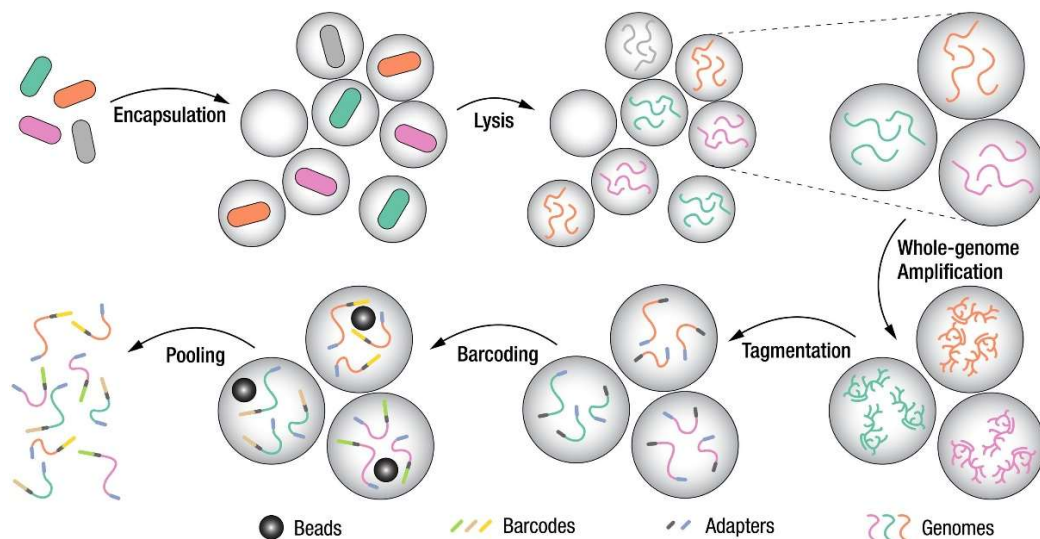


Fig. 1 Rappresentazione schematica del sistema di microfluidica (Zheng et al, 2022)

Lo studio ha previsto l'uso di quattro dispositivi di microfluidica distinti generati su polidimetilossano (PDMS) a partire da degli stampi in silicio [7]. I dispositivi sono stati utilizzati inizialmente per creare dei droplet contenenti al loro interno i campioni batterici, virtualmente un singolo batterio per goccia. Successivamente, l'utilizzo di tali dispositivi in sequenza, ha permesso di fondere i droplet contenenti i batteri con altri droplet portanti i reagenti necessari per la lisi cellulare, l'amplificazione del DNA single-cell, la tagmentazione del DNA amplificato, il barcoding dei frammenti per il riconoscimento dei vari campioni in seguito al sequenziamento e infine una ulteriore amplificazione di frammenti tramite PCR (Fig. 1). Alla fine, è stata ottenuta una emulsione al cui interno si trovano tutti i droplet originati dalla metodica.

2.3 Sequenziamento

L'emulsione contenente i droplet è stata rotta e il DNA è stato purificato tramite l'uso di beads magnetiche. Attraverso una ulteriore PCR sono stati aggiunti gli

adattatori per il sequenziamento Illumina e gli indici Nextera per distinguere campioni diversi sequenziati in contemporanea.

Il sequenziamento è stato svolto su piattaforma Illumina con una profondità di campionamento tra le 10,000 e le 200,000 reads per ogni batterio. Il sequenziamento è stato ripetuto due volte: una prima run da 100 paia di basi (pb) e una seconda run da 300 pb.

2.4 Processamento dei dati grezzi e co-assembly

Le sequenze di ogni campione sono state raggruppate in 36,864 cluster sulla base dei barcode associati ad ogni droplet, andando ad escludere quelli con troppe o con troppe poche reads. Le sequenze associate ad un singolo barcode formano i SAGs. Attraverso l'uso del software Trimmomatic [11] (versione 0.36), da ogni gruppo sono state rimosse le reads a bassa qualità e le sequenze degli adattatori, in modo da avere SAGs costituiti esclusivamente da sequenze microbiche.

Per valutare la qualità del metodo adoperato, cioè che ogni SAG contenga le informazioni relative ad un solo batterio, i SAGs ottenuti dalla mock community sono stati allineati ai quattro genomi di riferimento (RefSeq: GCF_002055965.1, GCF_004151095.1, GCF_001936035.1, GCF_002025145.1) tramite il software Bowtie2 [12] (versione 2.2.6, parametri di default); è stata quindi calcolata la purezza di ogni SAG sulla base della percentuale di reads di quel SAG allineate al genoma di riferimento. È emerso che l'84% dei SAGs aveva una purezza superiore al 95%. È stata inoltre valutata anche la copertura genomica di ogni singolo SAG, rivelando un coverage compreso tra il 17% e il 25% per i batteri Gram negativi e tra l'8% e il 9% per i batteri Gram positivi.

Per tale motivo, sono state combinate assieme le informazioni contenute in SAGs diversi appartenenti alla stessa specie (Fig. 2) per ottenere una copertura genomica più completa.

Passando all'analisi dei campioni di feci, sono stati raccolti in totale 21,914 SAGs ognuno contenente una media di 70,000 reads.

Tramite l'utilizzo del software Sourmash [8] (versione 2.0, $k = 51$, parametri di default) i SAG sono stati convertiti in signatures e confrontati tramite la funzione

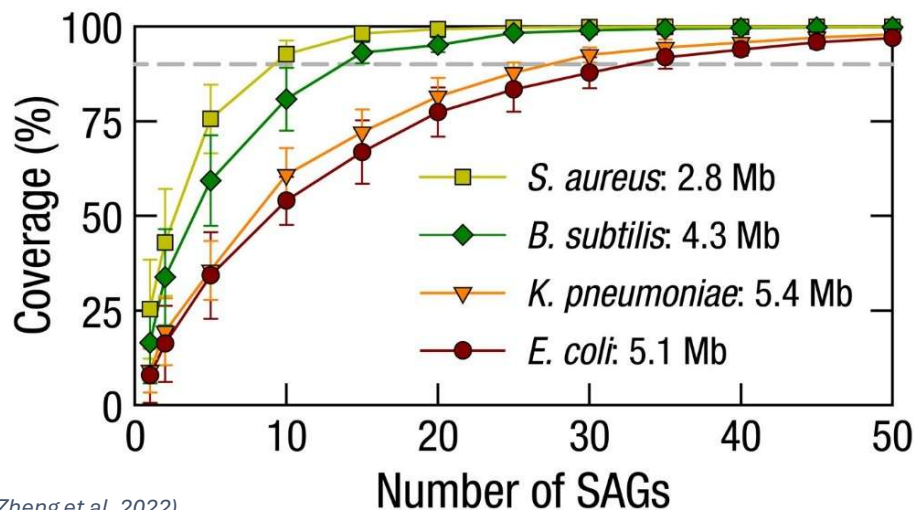


Fig. 2 (Zheng et al, 2022)

compare ottenendo una matrice di similarità tra SAGs. Attraverso un metodo di clustering gerarchico (SciPy, versione 1.1.0), i SAGs sono stati raggruppati in bin (cluster) utilizzando come soglia di raggruppamento 0.95, precedentemente verificata sul campione di controllo. Tutte le reads presenti in un bin sono state co-assemblate tramite SPAdes [13] (versione 3.13) per ottenere il genoma provvisorio di una certa specie. Poiché il genoma provvisorio è più completo di ogni singolo SAG, esso può essere utilizzato per creare dei signatures più precisi per confrontare i diversi gruppi tra loro e unirli in nuovi bin. Il processo è stato iterato finché più del 10% dei genomi assemblati presentava più del 10% di contaminazione, al fine di limitare la contaminazione di specie diverse all'interno di un bin. Il processo è stato ripetuto quattro volte ottenendo 364 bin.

Al fine di separare gruppi comprendenti più specie al loro interno, le reads di ciascun SAG sono state riallineate ai genomi provvisori tramite Bowtie2 e sono stati generati dei vettori per ogni contig superiore a 1000 pb contenenti il numero di reads allineate al contig. Attraverso un metodo di clusterizzazione gerarchica, i vettori sono stati suddivisi in due gruppi e i SAGs sono stati redistribuiti all'interno di essi sulla base del gruppo a cui si allineavano prevalentemente le reads, scartando i SAGs ambigui con meno del 95% delle reads allineate a uno dei due gruppi. Il processo è stato iterato finché non sono stati generati 400 bin, ognuno contenente i SAGs di una singola specie.

Infine, tramite l'utilizzo del software FastANI [9] (versione 1.2, parametri di default), sono stati confrontati a coppie tutti i SAGs appartenenti ai 400 bin attraverso il valore di ANI, al fine di identificare bin della stessa specie che erano rimasti separati. Utilizzando la soglia ANI > 95%, oltre la quale i due genomi vengono considerati appartenenti alla stessa specie, sono stati creati 234 bin. A questo punto, le reads appartenenti a ciascun bin sono state nuovamente co-assemblate, eliminando poi i contigs più corti di 500 pb. Tramite CheckM [14] è stata valutata la qualità dei bin, 76 dei quali sono stati definiti ad alta qualità (più del 90% di completeness e meno del 5% di contaminazione) o media qualità (più del 50% di completeness e meno del 10% di contaminazione). Tramite FastANI i 76 genomi ottenuti sono stati confrontati con tutti i genomi microbici presenti in RefSeq a settembre 2019 e con una collezione di più di mille batteri coltivati, al fine di identificare le specie batteriche presenti nel campione di feci.

2.5 Identificazione dei ceppi

Partendo dai genomi ad alta e media qualità ottenuti, ogni SAG è stato allineato al genoma co-assemblato e tramite l'utilizzo di *bfctools* sono stati identificati polimorfismi di singolo nucleotide (SNP) ad alta qualità. Si procede con un processo di filtraggio per garantire che le informazioni portate da SAG e SNP abbiano valenza statistica: vengono esclusi i SAGs con un'unica read allineata a un SNP, quelli con meno del 99% delle reads concordanti su un determinato SNP e quelli con copertura insufficiente delle posizioni SNP (meno dell'1% o meno di 10 posizioni totali). Dall'altro lato, vengono eliminati gli SNP poco rappresentati, presenti in meno del 5% dei SAGs, e quelli rari, con meno dell'1% dei SAGs con l'allele di riferimento e meno dell'1% con l'allele mutato, mentre vengono validati

solo gli SNP supportati da almeno due SAGs per ciascun allele (di riferimento e mutato).

Per ogni SAG superstite è stato costruito un vettore di identità delle basi alle posizioni SNP, usato per generare un dendrogramma tramite clustering gerarchico e successivamente una heatmap di SAGs nella quale sono emersi cluster corrispondenti a ceppi distinti. Per ciascun cluster è stato definito un genotipo di consenso, attribuendo a ogni posizione SNP la base condivisa da più del 90% dei SAGs. I SAGs sono stati poi assegnati con certezza a un ceppo nel caso in cui condividevano più del 95% delle posizioni SNP con il relativo consenso. Infine, le reads di tutti i SAG appartenenti a ciascun ceppo sono state co-assemblate con SPAdes per ricostruirne il genoma completo.

2.6 FastANI e Sourmash

Durante il periodo di tirocinio curricolare, applicando la metodica di microfluidica, sono stati ottenuti 1,512 SAGs, che sono stati confrontati attraverso vari metodi.

In primis, come presentato anche da Zheng et al., i SAGs sono stati confrontati a coppie tutti contro tutti grazie al software FastANI [9] (versione 1.34, parametri di default), il quale ha restituito come output l'ANI e l'alignment fraction (AF) per ogni coppia di SAGs.

In secondo luogo, i SAGs sono stati convertiti in signatures tramite l'uso di Sourmash [8] (versione 4.9.4) e confrontati poi grazie a delle funzioni proprie del software. I signatures creati da Sourmash si basano sulla divisione dei SAG in k-meri e sulla conversione di questi in un valore numerico, o "hash", permettendo un più rapido confronto tra sequenze rispetto al confronto basato sulla sequenza di basi. Per ogni SAG sono state generate due signatures diverse, una con $k = 31$ (parametro di default) e una con $k = 51$ (valore utilizzato nello studio di Zheng et al.). L'utilizzo di k-meri corti consente di avere una maggiore sensibilità nel confronto tra sequenze, ma possono identificare falsi match di sequenza. L'utilizzo di k-meri più lunghi, per contro, diminuisce la sensibilità di confronto, rendendo più difficile l'identificazione di regioni omologhe brevi o altamente divergenti.

I signatures sono stati confrontati all-vs-all attraverso la funzione *compare* al fine di ottenere una stima dell'ANI (*compare --ani --max-containment*) e l'indice di similarità Jaccard (*compare*) tra coppie di SAGs. L'indice Jaccard [10] è un indice di similarità definito dal rapporto tra la dimensione dell'intersezione e la dimensione dell'unione di due insiemi campionari:

$$J = \frac{|A \cap B|}{|A \cup B|} (1)$$

La stima dell'ANI e il calcolo dell'indice Jaccard sono stati eseguiti per entrambi gli insiemi di k-meri ($k = 31$ e $k = 51$). In entrambi i casi, l'output fornito da Sourmash è una matrice quadrata sfruttabile per le successive analisi.

2.7 Community detection

Il confronto a coppie di tutti i SAGs ha reso possibile la realizzazione di grafi grazie alle funzioni della libreria NetworkX di Python. I grafi sono stati originati facendo in modo che ad ogni SAG corrisponda un nodo e aggiungendo archi di

collegamento solo tra nodi il cui indice di confronto superava una certa soglia prestabilita. Con l'ANI calcolato da FastANI, l'arco è stato aggiunto tra SAGs con $ANI \geq 95\%$, tipica soglia oltre la quale due genomi sono considerati appartenenti alla stessa specie, e $AF \geq 0.1$, valore che tiene conto della poca sovrapposizione tra SAGs della stessa specie dovuta alla scarsa copertura genomica. La stessa soglia è stata utilizzata altresì per l'ANI stimato da Sourmash, anche se in questo caso l'AF non viene restituito dal software e non incide sul calcolo dell'ANI. Per quanto riguarda il Jaccard Index, invece, non essendoci una soglia classicamente utilizzata per la definizione della specie, è stata valutata una soglia in grado di massimizzare le metriche di valutazione degli algoritmi che verranno descritte in seguito. La soglia è stata fissata a 0.09.

Le reti così ottenute sono state analizzate tramite gli algoritmi di community detection Louvain, Leiden, MCL e SBM, tutti implementati nella libreria NetworkX, al fine di unire i SAGs di una stessa specie all'interno di uno stesso cluster. Gli algoritmi Louvain e Leiden si basano sul concetto di **modularità**, uno dei principali metodi di community detection che tenta di massimizzare la differenza tra il numero di archi all'interno di una comunità e il numero atteso di archi all'interno di questa [6]. La modularità viene definita come

$$H = \frac{1}{2m} \sum_c \left(e_c - \gamma \frac{K_c^2}{2m} \right) \quad (2)$$

dove e_c è il numero di archi presenti nella comunità c , K_c rappresenta il grado di ogni nodo, m il numero totale di archi nella rete e $\frac{K_c^2}{2m}$ il numero atteso di archi all'interno della comunità; γ viene detto parametro di risoluzione, è sempre > 0 e definisce il numero di comunità all'interno della rete. Il principale problema della community detection è massimizzare la modularità. L'algoritmo Louvain tende a spostare i nodi della rete nella comunità che comporta il maggior aumento di modularità, per poi aggregare le comunità così ottenute in un unico nodo. Il processo viene quindi ripetuto finché la modularità non può essere ulteriormente aumentata [5]. Un problema di Louvain è che può portare a comunità disconnesse sulla base di come un certo nodo viene spostato. L'algoritmo Leiden nasce per superare questa problematica, rendendo più efficiente e meno ridondante lo spostamento dei nodi tra le comunità e unendo i nodi in una comunità solo se questi sono significativamente connessi alla comunità stessa. L'algoritmo MCL, invece, si basa sul concetto di **camminatore casuale**, secondo il quale se una rete ha una forte struttura in comunità, un camminatore casuale che si muove lungo gli archi che uniscono i nodi di una comunità rimarrà all'interno della comunità stessa per lungo tempo. MCL parte dalla **matrice di trasferimento**, la matrice trasposta della matrice di adiacenza divisa per il grado del relativo vertice, in cui ogni elemento T_{ij} rappresenta la probabilità che un camminatore casuale che si trova in j si sposti in i . L'algoritmo eleva la matrice di trasferimento ad una potenza p generando una nuova matrice in cui ogni nuovo elemento M_{ij} rappresenta la probabilità del camminatore che si trova in j di spostarsi in i con p passi. Ogni elemento viene elevato ad una potenza α che rappresenta il peso di ogni cammino. A questo punto, gli elementi di ogni colonna vengono divisi per la loro somma generando una nuova matrice di trasferimento. Il processo viene ripetuto fino a quando la matrice che si

ottiene non varia più, cioè quando i suoi valori sono 0 e 1. La rete generata dalla matrice presenta delle componenti tra loro internamente connesse che sono le comunità [4]. Infine, SBM si basa sul concetto di **multigrafo stocastico**: dato un insieme di nodi e di relazioni tra nodi, viene creata una matrice di adiacenza che contiene vettori di adiacenza, ognuno dei quali comprende lo schema di legami tra il nodo i e il nodo j , statisticamente indipendenti l'uno dall'altro. Se due nodi appartengono allo stesso blocco, i loro vettori devono avere la stessa distribuzione di probabilità, cioè che la probabilità degli eventi riguardanti la rete non cambi scambiando tali nodi tra loro. Il modello crea quindi dei blocchi, le comunità, in cui i nodi sono stocasticamente equivalenti tra loro [6].

Una volta analizzati i grafi e ottenuta la suddivisione in cluster, si è resa necessaria l'implementazione di una metodica per valutare la bontà delle suddivisioni effettuate; conoscendo già la tassonomia di appartenenza di molti dei SAGs, si è deciso di utilizzare una valutazione basata sul loro ground truth. Tuttavia, molti SAGs risultavano essere unclassified, cioè non assegnati ad una tassonomia, e questi non sono stati utilizzati nella valutazione dei grafi. In particolare, sono state utilizzate le metriche presenti nella libreria scikit-learn di Python *completeness_score* ed *homogeneity_score*: *completeness_score* valuta se tutti i SAGs di una stessa tassonomia sono stati assegnati allo stesso cluster, mentre *homogeneity_score* valuta se i SAGs presenti in un cluster appartengono ad una sola tassonomia. Entrambe le metriche restituiscono un valore compreso tra 0 e 1. È stata poi implementata una terza metrica in grado di riassumere tutta l'informazione portata da ogni grafo a rete: questa metrica, definita *custom_score*, è una media pesata tra completeness e homogeneity, penalizzata dal numero di singleton presenti nella rete, che è stata così definita

$$custom_score = (0.67 \cdot h_score + 0.33 \cdot c_score)(1 - 0.1 \cdot p_singleton)$$

(3) dove h_score è homogeneity, c_score è completeness e $p_singleton$ rappresenta la proporzione di singleton sul numero totale di cluster generati dall'algoritmo di community detection. Si è deciso di dare ad homogeneity un peso maggiore rispetto a completeness in quanto, al fine di co-assemblare i SAGs appartenenti allo stesso cluster, è più importante avere cluster omogenei piuttosto che cluster completi. Il peso 0.1 moltiplicato per la proporzione di singleton rappresenta la penalità associata alla generazione dei singleton stessi.

Dalla media degli score ottenuti dai quattro algoritmi applicati alla rete costruita sulla base del Jaccard Index, è stato costruito un grafico della distribuzione delle metriche al variare della soglia, per valutare la soglia dell'indice Jaccard in grado di massimizzare il valore delle metriche (Fig. 3). Come citato in precedenza, in seguito a questa analisi, è stato deciso di utilizzare come soglia 0.09.

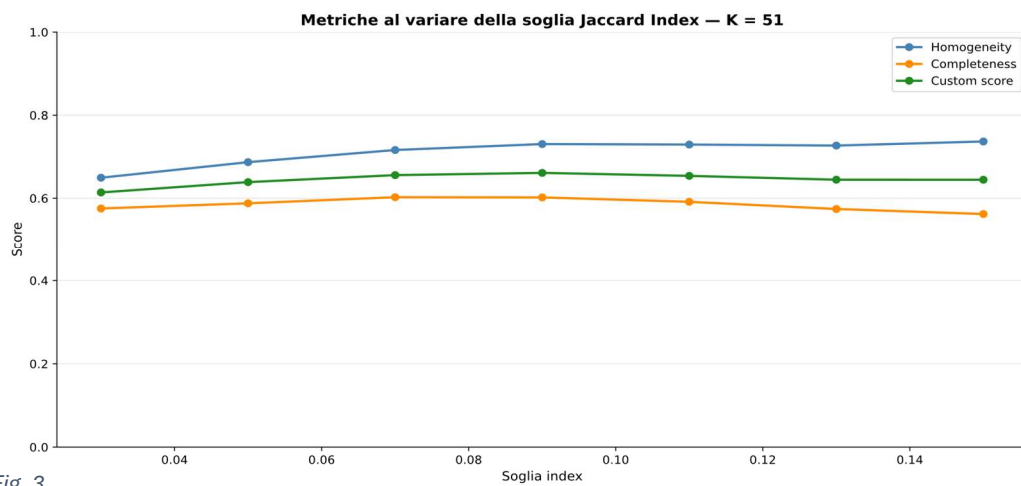
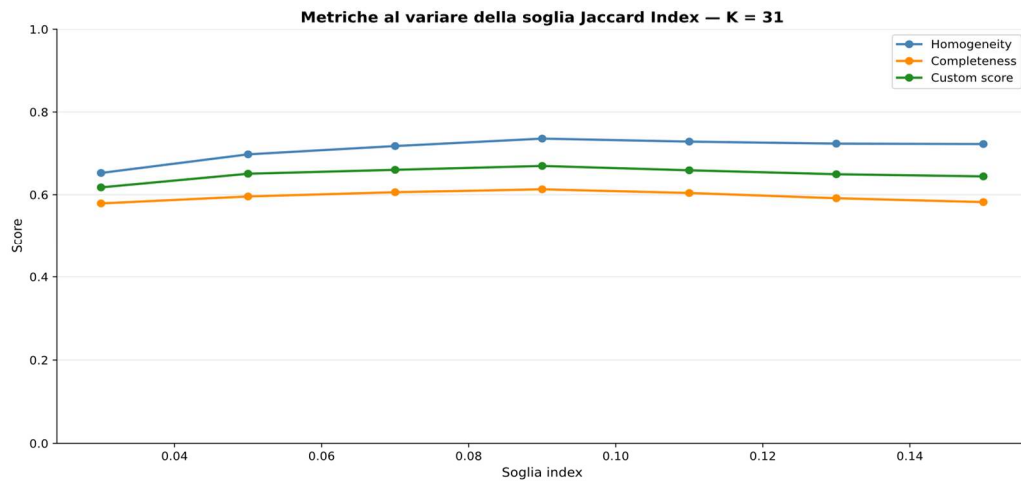


Fig. 3

Dataset	Rete	Homogeneity	Completeness	Custom_score	N_singleton
FastANI	Louvain	0,7451	0,4328	0,6029	925
	Markov	0,8092	0,4281	0,6397	972
	SBM	0,3701	0,4456	0,3709	925
	Leiden	0,7467	0,4337	0,6043	925
Sourmash ANI	Louvain k = 31	0,8953	0,7926	0,8459	274
	Markov k = 31	0,7937	0,9956	0,8448	274
	SBM k = 31	0,1083	0,1552	0,1216	274
	Leiden k = 31	0,1075	0,1526	0,1202	274
	Louvain k = 51	0,7625	0,9708	0,8178	245
	Markov k = 51	0,7788	0,9699	0,8277	255
	SBM k = 51	0,0655	0,1528	0,0928	245
	Leiden k = 51	0,0655	0,1528	0,0928	245
Sourmash Jaccard Index	Louvain k = 31	0,9316	0,7858	0,8564	466
	Markov k = 31	0,9210	0,6835	0,7996	776
	SBM k = 31	0,1597	0,1987	0,1673	466
	Leiden k = 31	0,9316	0,7858	0,8564	466
	Louvain k = 51	0,9257	0,7804	0,8487	502
	Markov k = 51	0,9133	0,6511	0,7825	814
	SBM k = 51	0,1580	0,1958	0,1649	502
	Leiden k = 51	0,9257	0,7804	0,8487	502

Tab. 1 Metriche di valutazione

3 Risultati

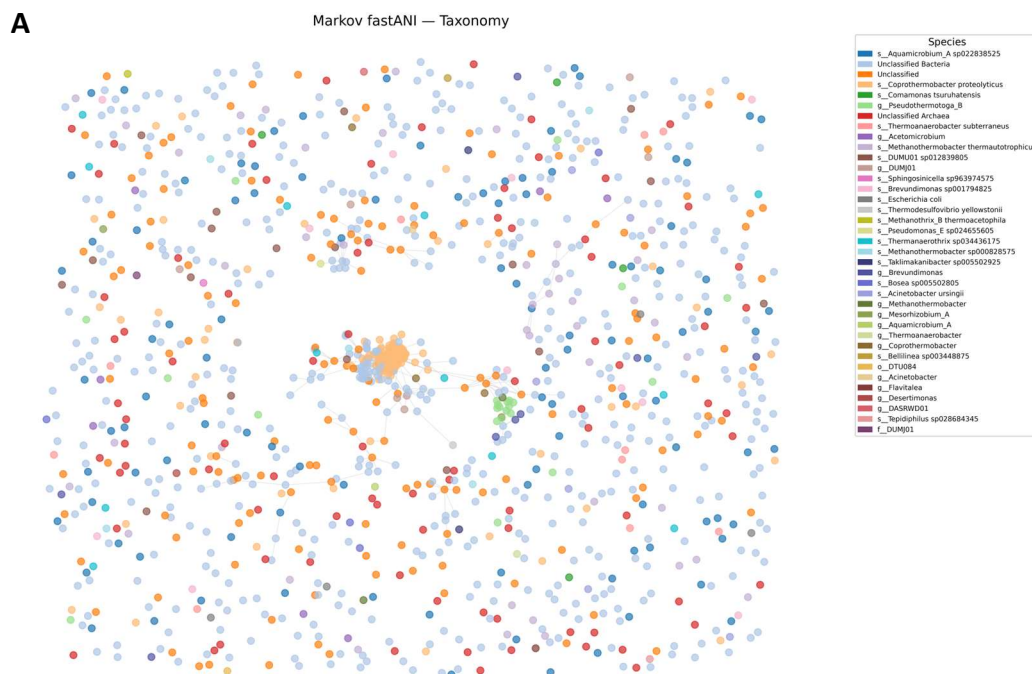
Nella tabella (Tab. 1) vengono presentati i valori delle metriche di valutazione ottenute da ogni algoritmo di community detection applicato alle reti create come descritto nella sezione Metodi.

Si presentano qui le tre reti migliori, cioè con *custom_score* più alto, ottenute per ogni blocco di analisi. Per ogni rete viene riportato il grafico della stessa con i nodi, cioè i SAG/droplet, colorati sia in base alla tassonomia che al cluster di appartenenza, il grafico della distribuzione di SAGs clusterizzati per ogni tassonomia e il grafico che mostra la composizione in specie di ogni cluster che non sia un singlet.

3.1 FastANI

Analizzando il *custom_score*, si nota come gli algoritmi di community detection applicati alle reti ottenute tramite l'ANI di FastANI diano dei risultati meno soddisfacenti rispetto agli stessi applicati alle reti ottenute con gli altri due metodi; questo è riconducibile all'alto numero di singleton generate dal raggruppamento in comunità, che supera la metà dei droplet. Ciò sottolinea come, in generale, l'utilizzo dell'ANI come tecnica per la classificazione delle SAG in specie non sia molto efficace.

Si riporta di seguito il risultato di MCL, quello a punteggio più alto sulle reti di FastANI (Fig. 4).



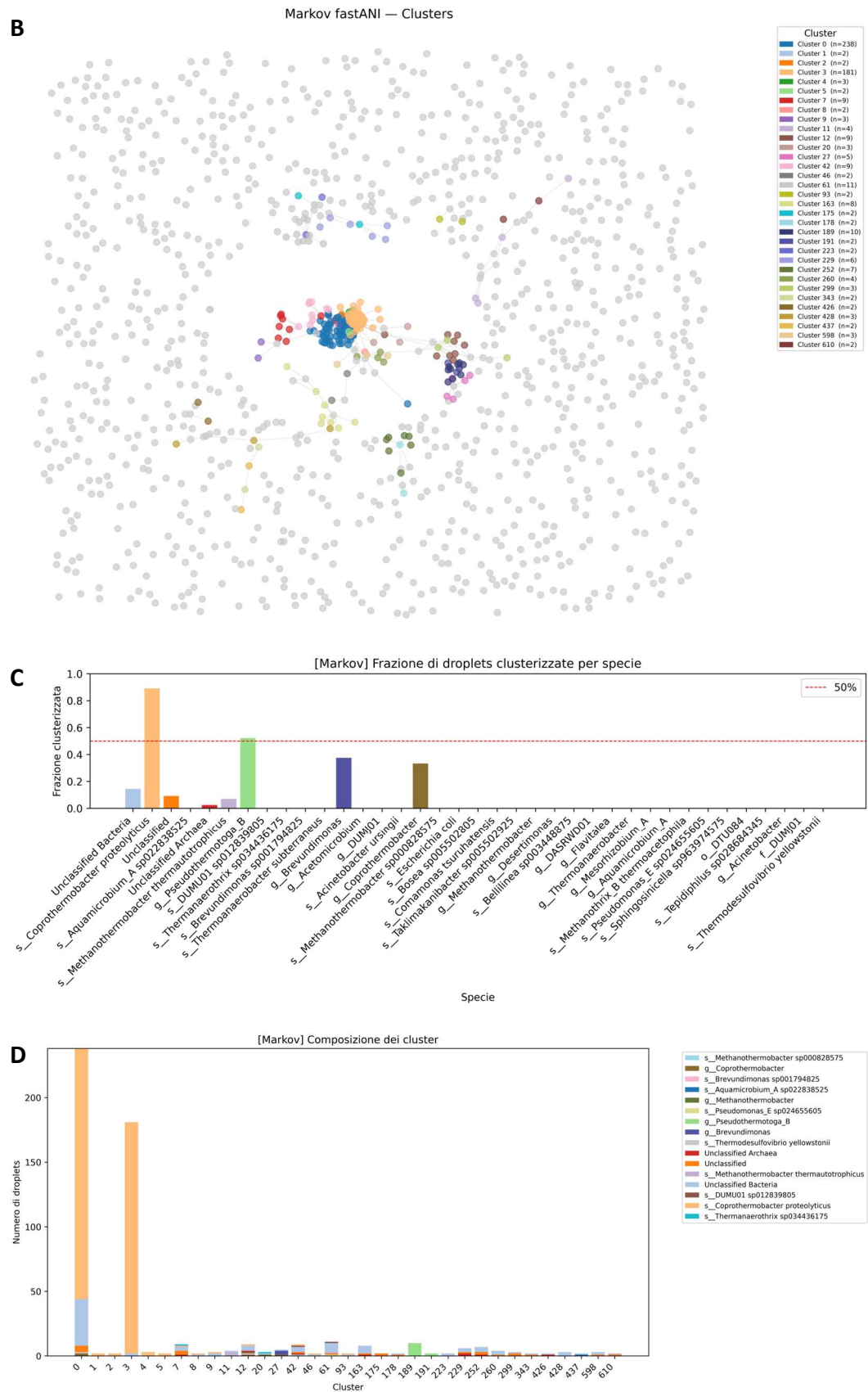


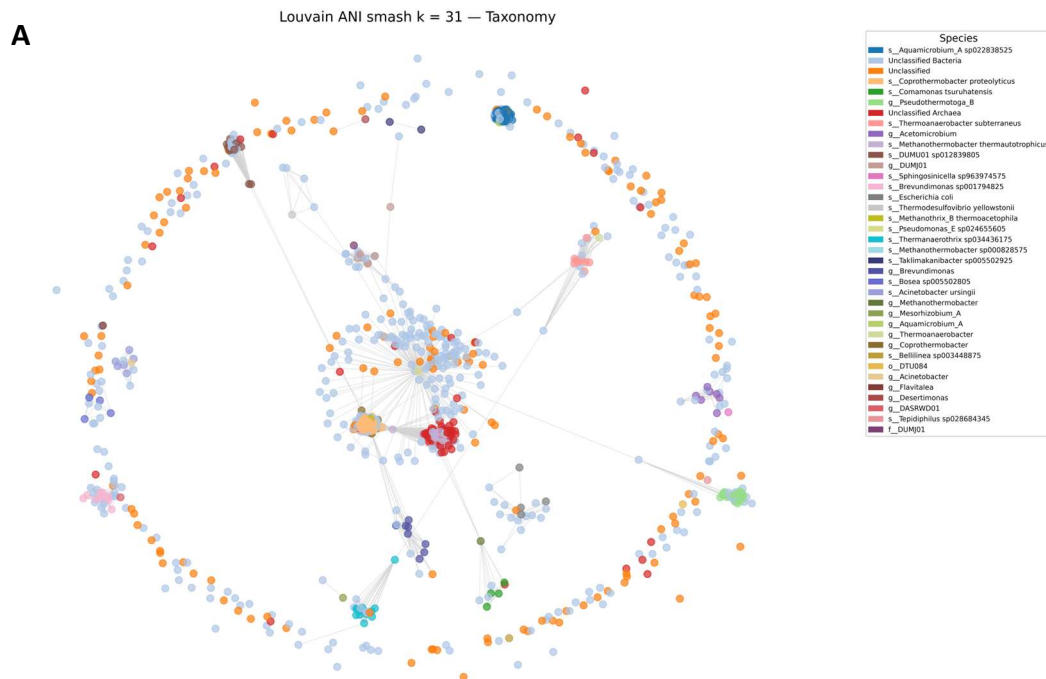
Fig. 4 (A) droplet colorate per tassonomia, (B) droplet colorate per appartenenza ad un cluster, (C) frazione di droplet clusterizzate per ogni tassonomia, (D) composizione in specie di ogni cluster.

Analizzando la struttura del grafico della rete è possibile notare come la maggiorparte dei droplet non sia raggruppato (Fig. 4B). Si nota, inoltre, che non è presente alcuna tassonomia per la quale i SAGs siano stati completamente clusterizzati (Fig. 4C); per quelle specie in cui nessun droplet risulta clusterizzato, tutte i droplet formano delle singlet. Tutto ciò sottolinea come questa analisi di community detection non sia ad alta qualità, rendendo la suddivisione in specie su base ANI molto ostica. Tuttavia, risulta interessante l'analisi della composizione dei cluster (Fig. 4B e Fig. 4D) dalla quale spicca la suddivisione all'incirca a metà dei SAGs di *C. proteolyticus* in due cluster distinti, caratteristica evidente in tutti le reti qui presentate.

3.2 ANI stimato

Per quanto riguarda le reti costruite valutando il valore di ANI stimato da Sourmash, si nota come i valori delle metriche siano in media decisamente migliori. I motivi verranno discussi nella prossima sezione.

Come esempio vengono riportati i risultati ottenuti tramite algoritmo di Louvain applicato alla rete costruita dall'analisi dei signatures con k-meri lunghi 31 nucleotidi (Fig.5)



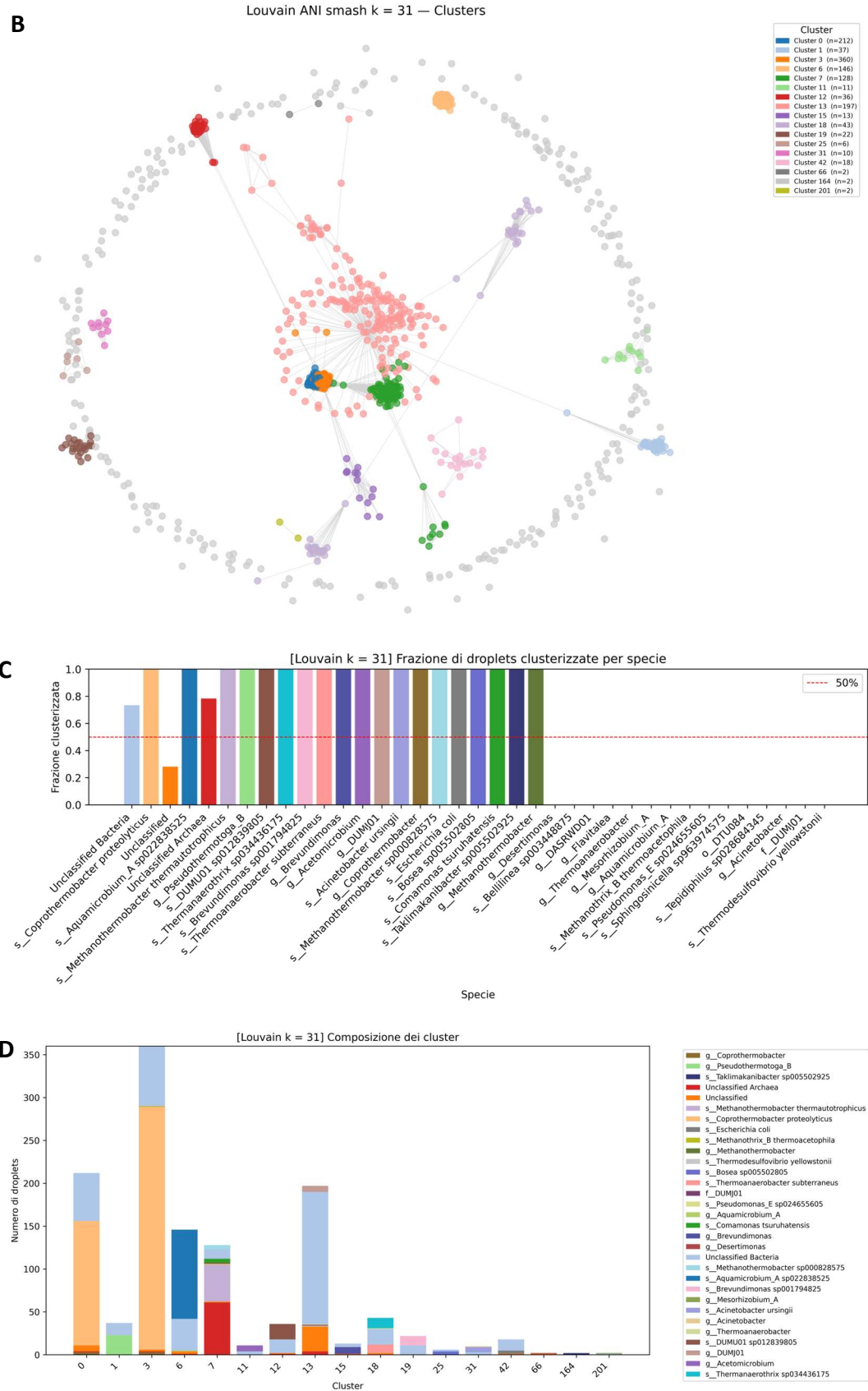
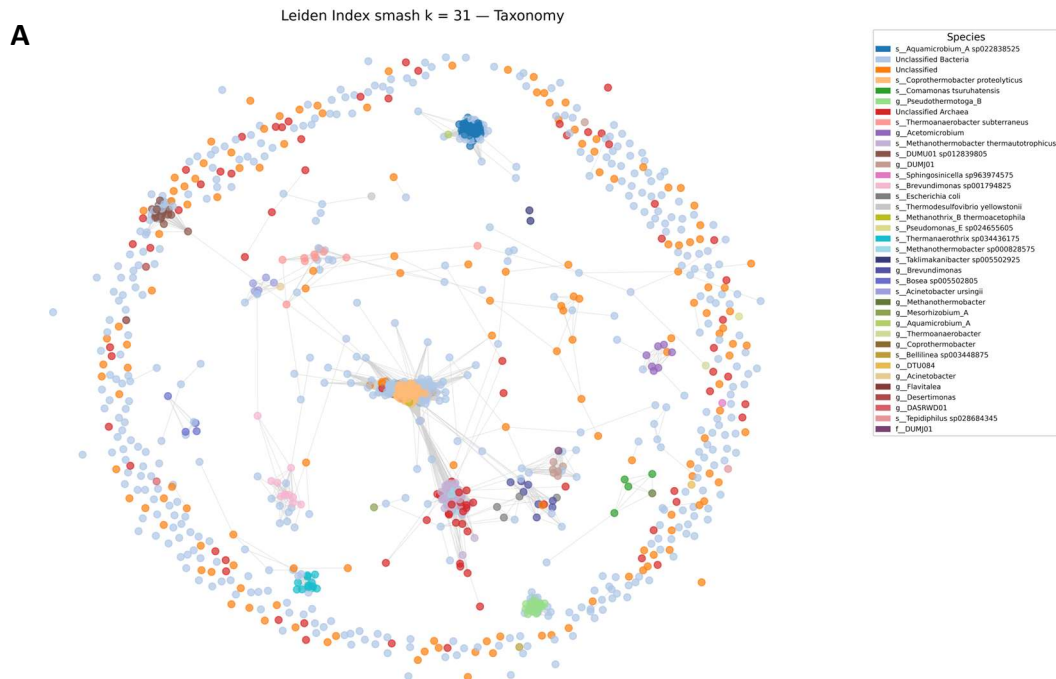


Fig. 5 (A) droplet colorate per tassonomia, (B) droplet colorate per appartenenza ad un cluster, (C) frazione di droplet clusterizzate per ogni tassonomia, (D) composizione in specie di ogni cluster.

Dal grafico della rete (Fig. 5A e Fig. 5B) risulta evidente come il numero di singleton sia molto inferiore rispetto alla rete precedente, così come il numero di cluster generati. Si nota, inoltre, come il numero di droplet clusterizzati per tassonomia sia molto più elevato (Fig. 5C), con la maggior parte delle specie che presentano i loro droplet completamente clusterizzati. Come nel caso precedente, i droplet di *C. proteolyticus* sono stati suddivisi in due cluster tra loro adiacenti (Fig. 5B e Fig. 5D) nei quali sono presenti anche dei droplet di specie *Unclassified*, i quali con tutta probabilità, appartengono a SAGs di *proteolyticus* non assegnati alla tassonomia. Si notano, inoltre, altri cluster che presentano al loro interno una tassonomia certa e alcuni unclassified, come il cluster 6 che mostra principalmente *s_Acquamicrobium_A* insieme a vari unclassified o il cluster 7 che presenta principalmente Archea (Fig. 5D). È interessante notare come alcuni cluster siano connessi tra loro da alcuni droplet che fungono da ponte tra di essi (Fig. 5A e 5B); questo lascia intuire che i droplet ponte potrebbero contenere materiale genetico di specie batteriche diverse, dovuto all'inserimento di due batteri diversi all'interno delle gocce create dal sistema di microfluidica.

3.3 Jaccard Index

Gli algoritmi di clustering applicati alle reti costruite su base Jaccard Index restituiscono i risultati migliori dell'intera analisi. In particolare, pur generando più singleton rispetto alle reti su base ANI stimato, l'omogeneità dei cluster è estremamente alta (Tab. 1), rendendo questi raggruppamenti i migliori ai fini dello studio. Dalle metriche in Tab. 1, si nota che l'algoritmo Louvain e l'algoritmo Leiden applicati al confronto dei signatures con $k = 31$, restituiscono gli stessi identici risultati. Questo conferma la bontà dell'uso del Jaccard Index per il confronto tra sequenze genomiche, in quanto due algoritmi diversi, seppur simili, restituiscono gli stessi raggruppamenti. Viene qui presentata la rete Leiden (Fig. 6).



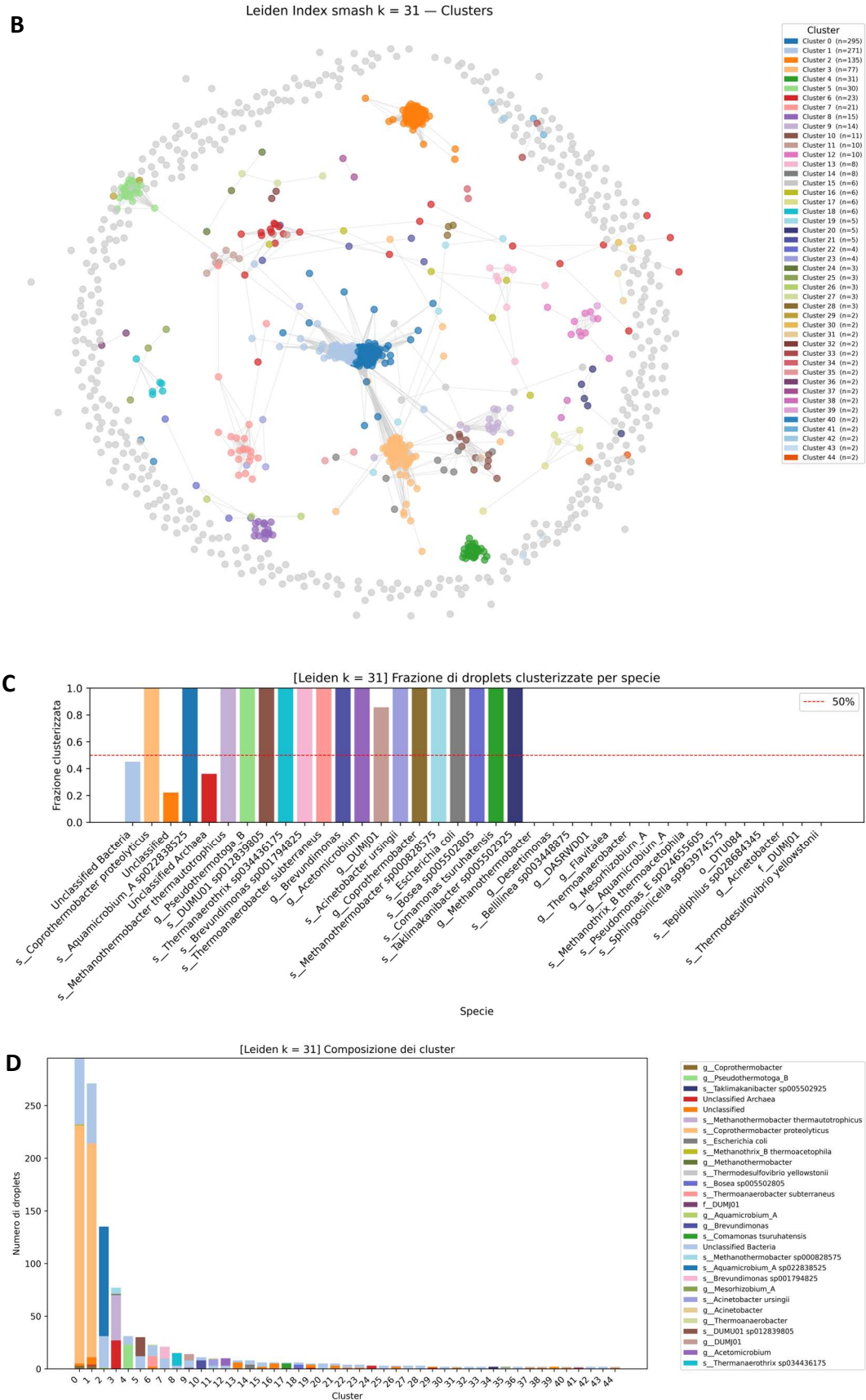
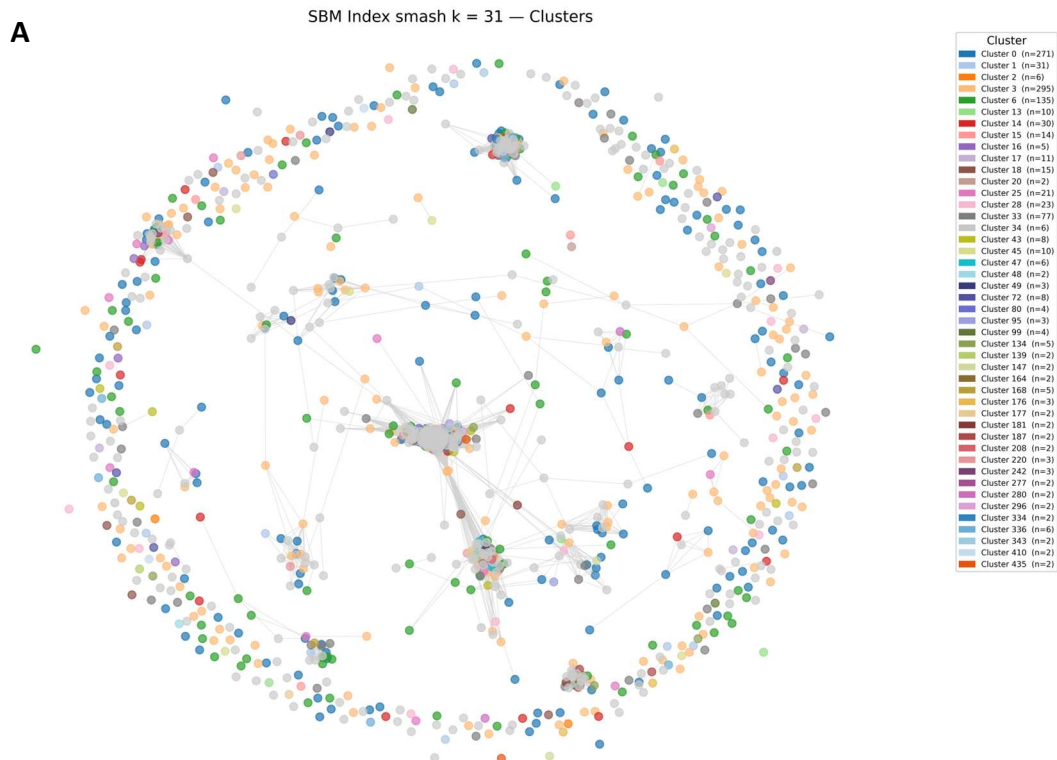


Fig. 6 (A) droplet colorate per tassonomia, (B) droplet colorate per appartenenza ad un cluster, (C) frazione di droplet clusterizzate per ogni tassonomia, (D) composizione in specie di ogni cluster.

Come nel caso precedente, sono presenti alcuni cluster ben distinti e altri uniti tra loro da droplet ponte; ciò sottolinea che l'ipotesi che tali droplet contengano due genomi differenti è fondata, in quanto metodi di analisi basati su indici diversi restituiscono lo stesso risultato. Inoltre, anche in questo caso, i due cluster più grandi sono costituiti dalla divisione dei droplet di *C. proteolyticus* (Fig. 4A, 4B e 4D) e, per la maggior parte delle tassonomie, quasi tutte i droplet vengono raggruppati (Fig. 4C).

3.4 SBM

I risultati mostrano come, in generale, il clustering effettuato tramite algoritmo SBM sia peggiore rispetto agli altri a causa della mancata clusterizzazione di circa metà dei SAGs delle tassonomie con alta completezza riscontrabili dagli altri algoritmi e a causa dell'elevata eterogeneità presente in ciascun cluster (Fig. 7). Questo è dovuto al particolare metodo di community detection implementato dall'SBM che, come mostrano i risultati, non si addice all'identificazione di specie.



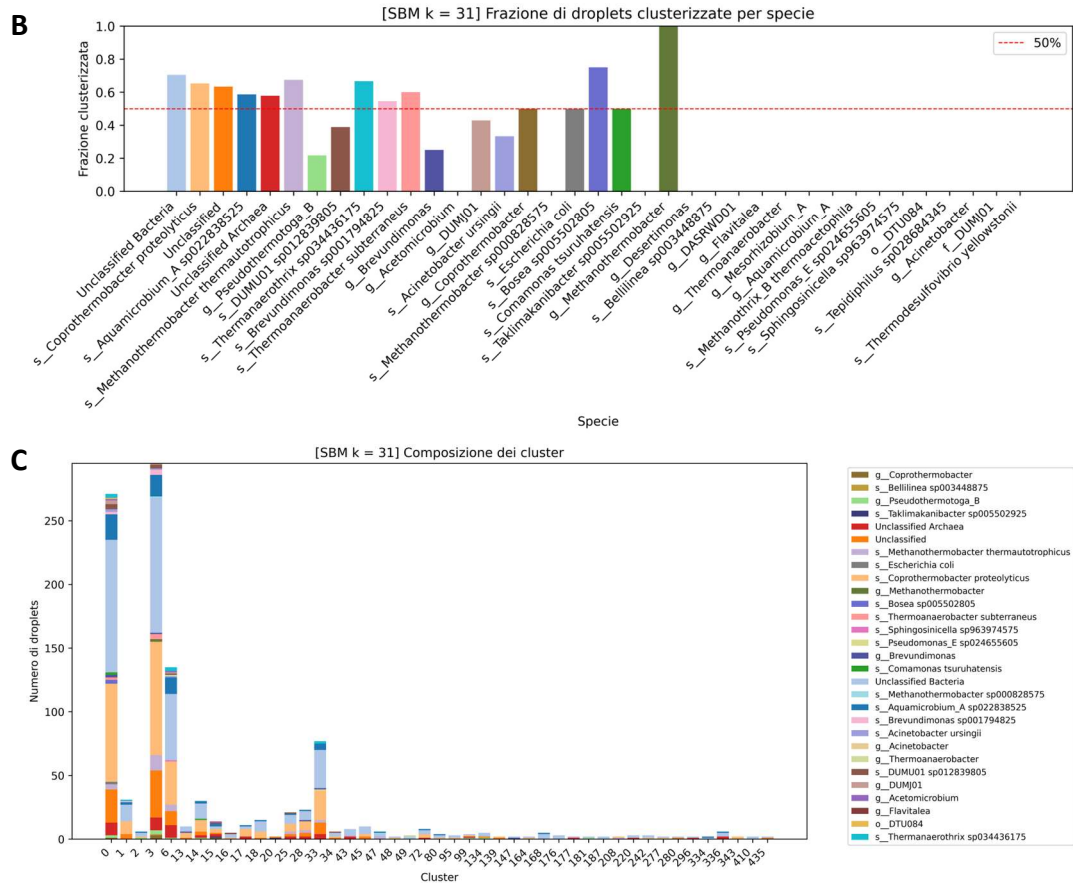


Fig. 7 Esempio di raggruppamento SBM su rete creata con Jaccard Index $k = 31$. (A) Rete con droplet colorate per cluster, (B) frazione di droplet raggruppate per ogni tassonomia, (C) composizione in specie di ogni cluster.

4 Discussione

L'utilizzo degli algoritmi di community detection applicati alle reti costruite sul confronto di SAGs attraverso indici diversi, ha dato la possibilità di ottenere varie informazioni sull'identificazione delle specie. Innanzitutto, i risultati ottenuti hanno mostrato come il confronto di SAGs incompleti attraverso l'ANI calcolato da FastANI, basato sulle sequenze nucleotidiche dei SAGs stessi, non costituisce un ottimo metodo di confronto al fine di valutare l'appartenenza dei SAGs ad un certo gruppo. Questo è dovuto al fatto che FastANI richiede un certo grado di overlap continuo tra le sequenze confrontate, indentificato dall'AF, che difficilmente si riscontra in SAGs incompleti, portando ad assegnare bassi ANI anche a coppie di SAGs della stessa specie e generando quindi molti singleton. Si ottengono ottimi risultati solo per specie ampiamente rappresentate nel campione in analisi, come *C. proteolyticus*, per le quali il maggior numero di SAGs aumenta la probabilità di ritrovare sequenze con overlap sufficiente.

Come mostrato dall'analisi di Zheng et al., FastANI risulta essere un buon algoritmo nel confronto di sequenze già parzialmente co-assemblate e quindi più complete, tuttavia il suo utilizzo per il confronto diretto tra SAGs, come qui dimostrato, non emerge come metodo ottimale.

Il confronto migliora utilizzando l'ANI stimato da Sourmash tramite il confronto dei signatures generati dalle sequenze; in questo caso il numero di singleton generati è ampiamente ridotto e vengono originati dei cluster ben distinti dagli altri. Questo miglioramento è da ritrovarsi nel fatto che Sourmash, creando dei signatures, valuta quanti sono i k-meri in comune tra due sequenze, indipendentemente dal fatto che queste abbiano delle sovrapposizioni

Sorprendentemente, i risultati migliori si ottengono attraverso il confronto dei SAGs tramite il Jaccard Index calcolato da Sourmash, che consente di creare cluster ben definiti e molto omogenei. Il motivo è da ritrovarsi nel fatto che, su SAGs appartenenti ad una stessa specie, la probabilità che la dimensione dell'intersezione dei k-meri in comune ai due SAGs sia simile alla dimensione dell'unione dei k-meri presenti tra i due SAGs (1) è molto alta. Questo è in grado di spiegare l'alta omogeneità dei cluster e il più alto numero di singleton rispetto all'ANI stimato, in quanto SAGs appartenenti alla stessa specie, ma con pochi k-meri in comune, ricevono un Jaccard Index più basso rispetto al relativo ANI.

Per quanto riguarda gli algoritmi di clustering, in generale, si nota come i raggruppamenti effettuati siano tra loro comparabili, ad eccezione dell'algoritmo SBM che non riesce a catturare la presenza di specie nella rete. Questo potrebbe essere dovuto alla generazione stocastica di blocchi di nodi che non consentirebbe di dare il giusto peso alle relazioni tra SAGs.

Tutte le reti qui mostrate concordano nel suddividere la popolazione di *C. proteolyticus* in due cluster distinti. Il fatto che utilizzando metodi di confronto diversi si osservi sempre questa caratteristica, lascia intuire che tale suddivisione potrebbe essere significativa. Nello specifico, si può ipotizzare che *C. proteolyticus* presenti due ceppi particolarmente divergenti all'interno del campione analizzato, i quali non differiscono per semplici mutazioni puntiformi, che sarebbero difficili da risolvere grazie ai SAGs incompleti, bensì dalla presenza di differenze su regioni genomiche più ampie.

Per concludere, questo studio ha mostrato come l'approccio di unire sistemi di microfluidica con analisi computazionali permette il sequenziamento single-cell ad alta processività ottenendo dei genomi ad alta qualità risolti a livello di ceppo, senza la necessità di coltivare i microrganismi. Tale approccio risulta molto versatile in quanto consente l'analisi di comunità microbiche con caratteristiche diverse, come possono essere il microbiota umano analizzato da Zheng et al. e il bioreattore analizzato durante il periodo di tirocinio, senza la necessità di conoscere a priori i microbi da sequenziare, così da rendere la metodica estremamente promettente nello studio dell'ambiente e della salute umana.

5 Bibliografia

- [1] M. Poyet, M. Groussin, S. M. Gibbons, J. Avila-Pacheco, X. Jiang, S. M. Kearney, A. R. Perrotta, B. Berdy, S. Zhao, T. D. Lieberman, P. K. Swanson, M. Smith, S. Roesemann, J. E. Alexander, S. A. Rich, J. Livny, H. Vlamakis, C. Clish, K. Bullock, A. Deik, J. Scott, K. A. Pierce, R. J. Xavier, E. J. Alm, A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nature medicine*, 2019
- [2] I. L. Brito, E. J. Alm, Tracking Strains in the Microbiome: Insights from Metagenomics and Models. *Frontiers Microbiology*, 2016
- [3] S.-Y. Teh, R. Lin, L.-H. Hung, A. P. Lee: Droplet microfluidics. *Lab on a Chip*, 2008
- [4] S. Fortunato, Community detection in graphs. *Physics Reports*, 2010
- [5] Traag VA, Waltman L and van Eck NJ, From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 2019
- [6] P. Holland, K.B. Laskey, S. Leinhardt, Stochastic blockmodels: Some first steps. *Social Networks* 5, 1983
- [7] J. C. McDonald, G. M. Whitesides, Poly(dimethylsiloxane) as a material for fabricating microfluidic devices. *Accounts of Chemical Research*, 2002
- [8] N. T. Pierce, L. Irber, T. Reiter, P. Brooks, C. T. Brown, Large-scale sequence comparisons with sourmash. *F1000 Research*, 2019
- [9] C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, S. Aluru, High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, 2018
- [10] D. Prokopenko et al, Utilizing the Jaccard index to reveal population stratification in sequencing data: a simulation study and an application to the 1000 Genomes Project. *Bioinformatics*, 2015
- [11] A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 2014
- [12] B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 2012
- [13] A. Bankevich et al., SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 2012
- [14] D. H. Parks et al., Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 2015

RESEARCH ARTICLE

MICROBIOLOGY

High-throughput, single-microbe genomics with strain resolution, applied to a human gut microbiome

Wenshan Zheng^{1,2,†}, Shijie Zhao^{2,3,†}, Yehang Yin^{2,4}, Huidan Zhang⁵, David M. Needham^{2,6}, Ethan D. Evans^{2,3}, Chengzhen L. Dai^{2,3,7}, Peter J. Lu^{5,8*}, Eric J. Alm^{2,3*}, David A. Weitz^{5,8,9*}

Characterizing complex microbial communities with single-cell resolution has been a long-standing goal of microbiology. We present Microbe-seq, a high-throughput method that yields the genomes of individual microbes from complex microbial communities. We encapsulate individual microbes in droplets with microfluidics and liberate their DNA, which we then amplify, tag with droplet-specific barcodes, and sequence. We explore the human gut microbiome, sequencing more than 20,000 microbial single-amplified genomes (SAGs) from a single human donor and coassembling genomes of almost 100 bacterial species, including several with multiple subspecies strains. We use these genomes to probe microbial interactions, reconstructing the horizontal gene transfer (HGT) network and observing HGT between 92 species pairs; we also identify a significant *in vivo* host-phage association between crAssphage and one strain of *Bacteroides vulgatus*. Microbe-seq contributes high-throughput culture-free capabilities to investigate genomic blueprints of complex microbial communities with single-microbe resolution.

Microbial communities inhabit many natural ecosystems, including the ocean, soil, and the digestive tracts of animals (1–4). One such community is the human gut microbiome. Comprising trillions of microbes in the gastrointestinal tract (5), this microbiome has substantial associations with human health and disease, including metabolic syndromes, cognitive disorders, and autoimmune diseases (6, 7). The behavior and biological effects of a microbial community depend not only on its composition (8, 9) but also on the biochemical processes that occur within each microbe and the interplays between them (10, 11); these processes are strongly affected by the genomes of each individual microbe living in that community.

The composition of the gut microbiome is specific to each individual person; although people often carry similar sets of microbial species, different individuals have distinct subspecies strains (hereafter referred to simply as

“strains”), which exhibit substantial genomic differences, including point mutations and structural variations (2, 12–14). These genomic variations between strains can lead to differences in important traits such as antibiotic resistance, metabolic capabilities, and interactions with the host immune system (15, 16), which can have serious consequences to human health. For example, *Escherichia coli* are common in healthy human gut microbiomes but certain *E. coli* strains have been responsible for several lethal foodborne outbreaks (17). Microbial behavior in the gut microbiome is influenced not only by the presence of particular strains but also by the interactions among them, such as cooperation and competition for food sources (11), phage modulation of bacterial composition (18, 19), and transfer of genomic materials between individual microbial cells (20, 21). Improving our fundamental understanding of these behaviors depends on detailed knowledge of the genes and pathways specific to particular microbes (22); however, elucidating this information can present considerable challenges where taxa are only known at the species level, obscuring strain-level differences. Individual microbes from the same strain from a single microbiome largely share the same genome (12, 23); therefore, a substantial improvement in understanding would be provided by high-quality genomes resolved to the strain level from a broad range of microbial taxa within a given community.

Several approaches are used to explore the genomics of the human gut microbiome. One widely used general technique is shotgun metagenomics, in which a large number of microbes are lysed and their DNA sequenced to yield a broad survey of genomic content

from the microbial community (22, 24, 25). Metagenomics-derived sequences have been assigned to individual species and have been used to construct genomes; however, metagenomics is generally not effective in assigning DNA sequences that are common to multiple taxa in a single sample, such as when one species has multiple strains or when homologous sequences occur in the genomes of multiple taxa (26, 27). Consequently, shotgun metagenomics generally cannot resolve genomes with strain resolution, though recent technological advances such as long-read sequencing (28, 29), read-cloud sequencing (30), and Hi-C (31, 32) are beginning to contribute strain-level information for some species. By contrast, high-quality strain-resolved genomes of taxa from the human gut microbiome have been assembled from colonies cultured from individual microbes (12, 14, 33, 34); however, culturing colonies can be labor-intensive and biased toward microbes that are easy to culture. Alternatively, single-cell genomics or mini-metagenomics rely upon isolation and lysing of individual or around a dozen microbes in wells on a titer plate, and subsequently amplifying their whole genomes for sequencing (35–40). Such approaches might yield strain-resolved genomes and have been used to probe the association between phages and bacteria (41, 42). For all of these metagenomic, culture, and well-plate approaches, however, available resources severely limit the number of strain-resolved genomes that originate from the same community (12, 33), thereby constraining our knowledge of the genomic structure and dynamics of the human gut microbiome of a given person.

One practical way to overcome this throughput limitation is droplet microfluidics (43), in which individual cells are encapsulated in nanoliter to picoliter droplets. These techniques have been used to analyze the transcriptomics of thousands of individual mammalian cells; more specifically, each cell is encapsulated in a single microfluidic step, and its genetic material liberated and labeled (44, 45). By contrast, lysing, whole-genome amplification, and labeling of bacterial DNA require multiple microfluidic steps; consequently, although each of these steps has been performed individually in droplets they have not thus far been combined into a unified droplet-based workflow that takes in bacteria and outputs whole genomes in which each DNA sequence can be traced back to its single host microbe (35, 46, 47). Thus, substantial improvement in our understanding of the human gut microbiome requires a new, practical, high-throughput method to obtain single-microbe genomic information at the level of detail given by culture-based or single-cell genomics, while simultaneously sampling the broad spectrum of microbes typically accessed by shotgun metagenomics.

¹Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA. ²Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. ³Center for Microbiome Informatics and Therapeutics, Massachusetts Institute of Technology, Cambridge, MA, USA.

⁴College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang, China. ⁵School of Engineering and Applied Sciences (SEAS), Harvard University, Cambridge, MA, USA. ⁶Ocean Ecosystems Biology, GEOMAR, Helmholtz Centre for Ocean Research, Kiel, Germany.

⁷Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA.

⁸Department of Physics, Harvard University, Cambridge, MA, USA. ⁹Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, USA.

*Corresponding author. Email: ejalm@mit.edu (E.J.A.); plu@post.harvard.edu (P.J.L.); weitz@seas.harvard.edu (D.A.W.)

†These authors contributed equally to this work.

‡Present address: Mzbio, Inc., Cambridge, MA, USA.

We introduce Microbe-seq, a high-throughput method for obtaining the genomes of large numbers of individual microbes. We use microfluidic devices to encapsulate individual microbes into droplets, and within these droplets we lyse, amplify whole genomes, and barcode the DNA. Consequently, we achieve substantially higher throughput than what is practically accessible with titer plates. We investigate the human gut microbiome, analyzing seven longitudinal stool samples collected from one healthy human subject, and acquire 21,914 single-amplified genomes (SAGs). Comparing with metagenomes from the same samples, we find that these SAGs capture a similar level of diversity. We group SAGs from the same species and coassemble them to obtain the genomes of 76 species; 52 of these genomes are high quality with more than 90% completeness and less than 5% contamination. We achieve single-strain resolution and observe that ten of these species have multiple strains, the genomes of which we then coassemble. With Microbe-seq, we can probe the genomic signatures of microbial interactions within the community. For instance, we construct the network of the horizontal gene transfer (HGT) of the bacterial strains in a single person's gut microbiome and find substantially greater transfer between strains within the same bacterial phylum, relative to those in different phyla. Unexpectedly, through use of Microbe-seq we detect association between phages and bacteria; we find that the most common bacteriophage in the human gut microbiome, crAssphage, has significant *in vivo* association with only a single strain of *B. vulgatus*.

Results

High-throughput sample preparation using droplet-based microfluidic devices

We use a microfluidic device to encapsulate individual microbes into droplets (fig. S1 and movie S1) containing lysis reagents, as shown in the schematic in Fig. 1A. We collect the droplets in a tube and incubate to lyse the microbes; the DNA from each individual microbe remains within its own single droplet. We reinject each droplet into a second microfluidic device (48) that uses an electric field to merge it with a second droplet containing amplification reagents (49, 50); we collect the resulting larger droplets and incubate them to amplify the DNA. We then use similar procedures with a third microfluidic device to merge each droplet with another droplet containing reagents to fragment and add adapters (Nextera) to the DNA (51). We subsequently employ a fourth microfluidic device to merge each droplet with an additional droplet containing a barcoding bead, a hydrogel microsphere with DNA barcode primers attached; these primers are generated through combinatorial barcode extension. Each primer con-

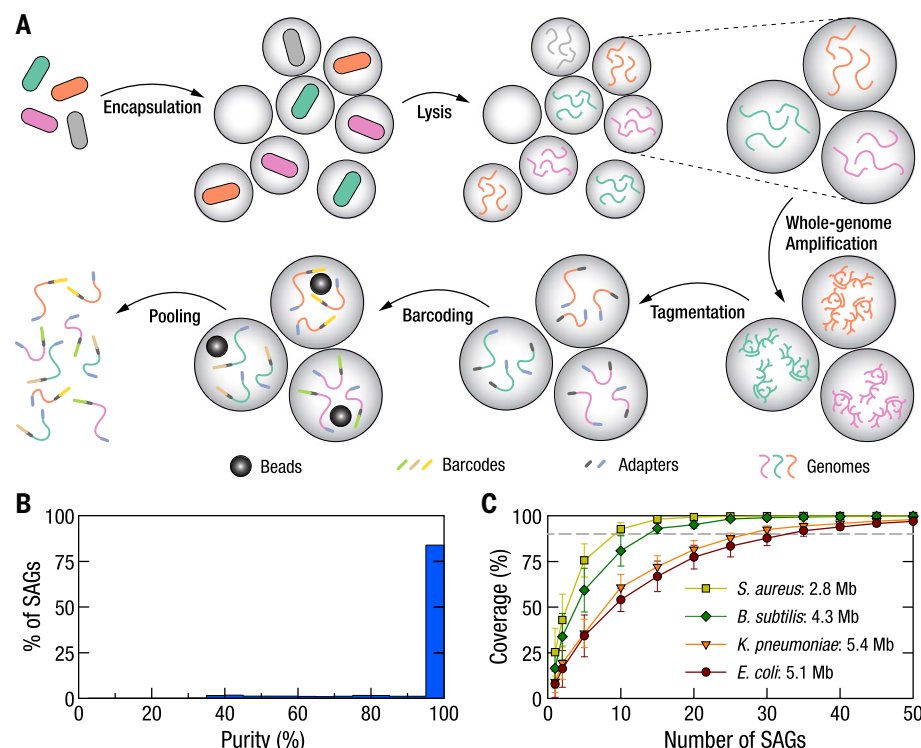


Fig. 1. Schematic of the Microbe-seq workflow and application in a community of known bacterial strains. (A) Schematic of the Microbe-seq workflow. Microbes are isolated by encapsulation with lysis reagents into droplets. Each microbe is lysed to liberate its DNA; after lysis, amplification reagents are added to each droplet to amplify the single-microbe genome within each. Tagmentation reagents are added into each droplet to fragment amplified DNA and tag them with adapters. PCR reagents and a bead with DNA barcodes are added to each droplet. PCR is performed to label the genomic materials with these primers, and droplets are broken to pool barcoded single-microbe DNA together. (B) Purity distribution of all SAGs from the mock community sample, which for a large majority of SAGs exceeds 95%, demonstrating single-microbe origin for the DNA in each of these SAGs. (C) Combined genome coverage of reads as a function of the number of SAGs from which these reads originate; error bars denote standard deviation. The dashed horizontal line indicates a coverage of 90%. In all cases, a few dozen SAGs contain essentially all the information of the microbial genome.

tains two parts: one barcode sequence that is specific to each droplet and another sequence that anneals to the previously added adapters. We attach these barcode primers to the fragmented DNA molecules within each droplet using polymerase chain reaction (PCR). We then break the droplets, add sequencing adapters, and sequence (Illumina). We illustrate all of these steps in the schematic in Fig. 1A and include schematics for all microfluidic devices in fig. S1.

The raw data constitutes sequencing reads, each containing two parts: a barcode sequence shared among all reads from the same droplet, and a sequence from the genome of the microbe originally encapsulated in that droplet. The collection of microbial sequences associated with a single barcode represents a SAG (38).

Single-microbe genomics in a community of known bacterial strains

To characterize the nature of the information contained within each SAG, we determine

whether each SAG contains genomes from one or multiple microbes and how much of a microbe's genome is contained in each SAG. Consequently, we apply our methods to a mock community sample that we construct from strains with genomes that are already known completely, providing an established reference to check the quality of each SAG. The mock sample contains four bacterial strains in similar concentrations, each with a complete, publicly available reference genome: Gram-negative *E. coli* and *Klebsiella pneumoniae*, and Gram-positive *Bacillus subtilis* and *Staphylococcus aureus*. From the mock sample, we recover 5497 SAGs, each containing an average of 20,000 reads (table S1).

To assess the extent to which each SAG contains genomic information from only a single microbe, we align each read against each genome and identify the genome containing the sequence that most closely matches each read as the closest-aligned genome (52). If a SAG includes reads from multiple microbes,

its constituent reads likely connect with a mix of different closest-aligned genomes; by contrast, if the reads from a SAG originate from only one microbe, then those reads will connect to the same closest-aligned genome. To test this, for each SAG we examine all reads that align successfully to at least one of the four genomes and determine the percentage of those reads that share the same closest-aligned genome; we define the highest of these four values as the purity of that SAG (47). Within the mock sample, we find that 84% (4612) of the SAGs have a purity exceeding 95%, which we designate as high purity; these data demonstrate that a large majority of SAGs represent single-microbe genomes, as shown in the distribution in Fig. 1B.

For each of these high-purity SAGs, we identify each base in the corresponding reference genome that has at least one read from that SAG that aligns successfully to it; we use this information to calculate genome coverage, defined as the ratio of these aligned bases to the total number of bases in the reference genome for each SAG. We find that genome coverage is broadly distributed around the average values of 17 and 25% for *B. subtilis* and *S. aureus*, respectively (fig. S2). The coverage for these Gram-positive strains is roughly double that of the coverage for the Gram-negative strains, which peaks more narrowly around the average values of 8 and 9% for *E. coli* and *K. pneumoniae*, respectively (fig. S2 and table S1); the comparatively smaller genome sizes of the Gram-positive strains likely contribute to this observed coverage difference.

The genome coverage of each individual SAG is incomplete, and one way to overcome this limitation is to combine the genomic information from multiple microbes belonging to the same strain, which are known to share nearly identical genomes. To explore how the genomic information contained within a group of SAGs depends on the number of SAGs in the group, we randomly select a subpopulation of SAGs from the group that matches each of the four reference genomes and determine the total combined coverage of all of the reads within that group of SAGs. We calculate the combined coverage as a function of the number of SAGs in that group and find that it increases with SAG group size. Although the specific number of SAGs needed to reach any given combined coverage varies between strains, in all cases the information that would be needed to reconstruct essentially complete genomes is, in principle, present within any randomly selected group of several dozen SAGs, as shown in Fig. 1C.

Human gut microbiome samples

To explore the utility of single-microbe sequencing, we apply the droplet-based approach to a

complex microbial community. We explore the human gut microbiome, which is expected to contain on the order of 100 species (22). We examine seven stool samples collected from one healthy human donor over a year and a half, for which both shotgun metagenomic datasets and cultured isolate genomes have been reported separately (12). We recover 1000 to 7000 SAGs per sample, for a total of 21,914 SAGs (table S2). Each SAG contains an average of about 70,000 reads so that each sample contains several hundred million reads.

Genomes of microbial species in the human gut microbiome

To explore the data acquired through the droplet-based methods the contents of each SAG must be identified, which is best done by comparison with known genomes. In the case of the mock sample, we identify each SAG by comparing its reads to preexisting reference genomes. By contrast, in the case of the human gut microbiome samples no complete set of genomes from all major strains exists, and certain species may not even appear in public reference databases; more generally, it is not possible to identify SAGs from complex microbial communities using comparison with preexisting reference genomes. Based on the data from the mock sample, we expect the coverage of the SAGs to be far from complete, thereby precluding an individual SAG from being used as a reference genome. Consequently, we develop an approach that does not consult external genomes but instead combines the genomic information from multiple SAGs to coassemble genomes and thus enable identification of individual SAGs.

In this approach, the first task is to identify SAGs that correspond to the same species. Within each SAG, we assemble the reads de novo with overlapping regions into contigs (53)—longer contiguous sequences of bases—and the resulting set of contigs forms that SAG's partial genome, which we expect from the mock sample to cover only a few percent of the total genome, somewhat less than the coverage of the reads themselves. The overlap between two genomes from a given species is expected to be roughly the square of this coverage, generally <1%; consequently, any two genomes from SAGs of the same species will likely share only a few or even no direct overlaps. This low overlap prevents direct sequence alignment from being a robust method for determining the similarity of two partial genomes; instead, for each SAG's genome, we use a hash function to extract a signature indicative of the complete genome (54). We compare the signatures of all pairs of genomes, using hierarchical clustering to group SAGs with similar partial genomes into preliminary data bins. For all SAGs within each of these bins, we treat all of the reads equally and coassemble

them into that bin's tentative genome. We then calculate new signatures for the tentative genomes and recompare their similarity, iterating this process to consolidate bins that should contain sequences from the same species.

This initial grouping process may generate bins containing reads from multiple taxa. In response, we examine how the reads within each bin align to the contigs in its tentative coassembled genome. For each contig, we examine the reads that align to that contig successfully; if two different contigs have non-overlapping subgroups of SAGs with reads that align successfully, then each of these subgroups likely correspond to different taxa (40). In these cases we create new bins from these subgroups and coassemble their tentative genomes; these genomes should, in principle, represent only a single taxon.

After this bin splitting process, multiple bins may contain genomes that correspond to the same species, which we may identify by comparing their genomes. However, in contrast to the earlier steps each bin at this stage contains a genome coassembled from many SAGs, which is large enough to share overlapping sequences with genomes from other bins that represent the same species; consequently, we can compare the sequences of tentative genomes directly without needing to rely on comparatively less precise hashes. For all pairs of these tentative coassembled genomes, we calculate their average nucleotide identity (ANI), a metric that estimates the similarity of two genomes by comparing their homologous sequences; we use an ANI value exceeding 95% to indicate that both genomes belong to the same species (55). Using this criterion, we merge all bins corresponding to the same species and coassemble their constituent reads to yield refined genomes of individual species.

To evaluate the quality of each of these refined coassembled genomes we count single-copy marker genes to estimate two metrics: completeness (the fraction of a taxon's genome that we recover) and contamination (the fraction of the genome from other taxa) (56). We find that 52 of the coassembled genomes have completeness >0.9 and contamination <0.05; we thus designate them high quality (33, 57, 58). We also find that 24 of the other coassembled genomes have completeness >0.5 and contamination <0.1; we thus designate them medium quality. More than three-quarters (16723) of the SAGs belong to one of these 76 species, demonstrating successful reconstruction of reference genomes for a large majority of SAGs; out of these 76 species, six have fewer than 24 SAGs.

To determine whether each genome corresponds to a single species known to occur in the human gut microbiome, we compare

each coassembled genome against a public database (GTDB-Tk) (59), using the ANI >95% criterion to identify matches of the same species. We obtain a broad mix of species from diverse phyla including Firmicutes, Bacteroidetes, Actinobacteria, Proteobacteria, and Fusobacteria (reported with assembly quality information in table S3). Several species well known in the human gut microbiome are abundant, including *Faecalibacterium prausnitzii*, *Bacteroides uniformis*, and *B. vulgatus*. For each of these 76 genomes, we list the name (colored according to corresponding phylum), illustrate its phylogenetic relationships with other species with a dendrogram, and indicate the number of SAGs used in its coassembly with the length of the outer bars, shaded for those of high quality, in Fig. 2.

Because there exists for these samples a large number of isolates cultured from the same human donor (12), we compare the coassembled genomes with the “gold standard” genomes derived from isolates. We find 19 species for which the coassembled genomes have corresponding isolate genomes, which we mark with an asterisk following each species name in Fig. 2. The ANI exceeds 99.5% in 17 species; these data provide strong evidence for the faithful reconstruction of genomes that closely match those of the cultured isolates, with low contamination.

With only a small set of culture-free experiments, we recover a broad set of accurate reference genomes from more species than those recovered from any other single gut microbiome. These genomes enable us to assign a large majority of single-microbe SAGs in the sample to one of these 76 species.

Microbial diversity in the human gut microbiome

Although species-level genomes provide one approach to assess microbiome diversity, the diversity of the human gut microbiome is typically assessed with metagenomics. We follow the spirit of this metagenomic approach and repurpose the droplet-based dataset to mimic that produced in metagenomics, by considering all reads from all SAGs in each sample. We classify each read in each sample by comparing it with the public database of microbial genomes (60); we also perform this comparison on each read from the corresponding metagenomic datasets (12). Each stool sample contains thousands of cells, in contrast to metagenomics which typically accumulates genomic data from millions of cells. Nevertheless, we recover 96.9 to 99.8% of the genera found by metagenomic analysis of the seven stool samples (figs. S3 and S4 and table S2).

The large collection of coassembled species-level genomes, however, provide an additional way to assess diversity with even greater precision at the species level. We align all meta-

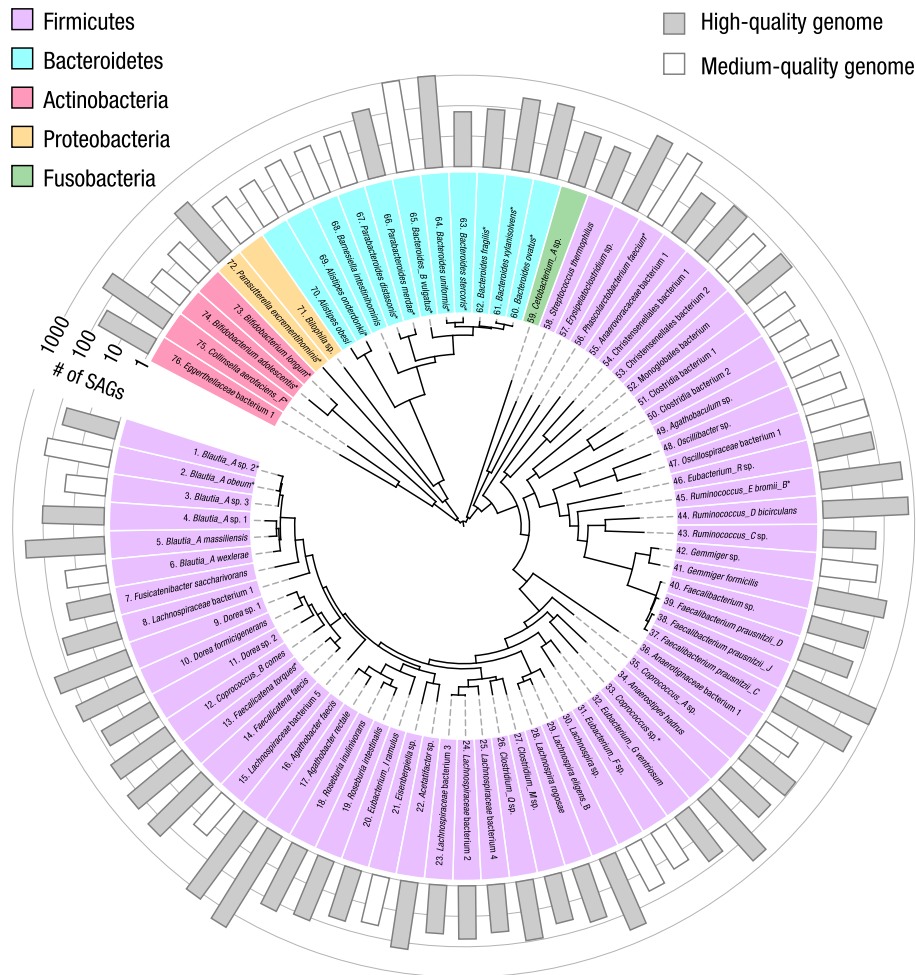


Fig. 2. Coassembled genomes of 76 bacterial species in the human gut microbiome of a single human donor. These 76 bacterial species have high- or medium-quality coassembled genomes. A phylogeny constructed from ribosomal protein sequences is represented by the dendrogram in the center of the circle. The phylum of each species is indicated by the background color behind each listed species name (GTDB-Tk database); the 19 species with genomes from isolates cultured from the same human donor are marked with an asterisk. The number of SAGs used for coassembly (abundance) is indicated by the bars in the outermost ring, shaded in gray for the 52 high-quality genomes and unshaded for the 24 medium-quality genomes.

genomic reads to the combined genome of all coassembled species irrespective of quality and find that 96 to 98% of these reads align, thereby providing further evidence that the droplet-based method does not miss any noticeable number of abundant taxa. For the 76 species with high- or medium-quality genome coassemblies, we estimate the relative abundance of each species in both metagenomics and the droplet-based approach. In metagenomics, the number of cells from a given species is proportional to the average read coverage over its genome; by contrast, in the droplet-based method we infer relative cell number by counting SAGs corresponding to the given species. We find that both abundance estimates are well correlated for the 76 species (fig. S5), though with one notable trend: In general, Gram-negative species—particularly

those from Bacteroidetes and Proteobacteria—are underrepresented in the droplet-based method; by contrast, Gram-positive species, including Firmicutes and Actinobacteria, are overrepresented—albeit with a few exceptions (fig. S6). These trends may result from differences in lysis methods: for the metagenomics samples, we follow standard lysing protocols that use mechanical bead beating; because such mechanical methods have not been demonstrated in droplets, we use purely enzymatic methods known to favor Gram-positive species.

Strain-resolved genomes in the human gut microbiome

Many species in the human gut microbiome are represented by multiple strains (61); different strains may play distinct roles within

complex microbial communities and express different sets of genes to carry out these roles (62). Linking specific genes and consequently their functionality to the strains which contain them requires knowledge of the genomes from those individual strains. Moreover, because each microbe inherently represents only a single strain, definitive identification of each SAG requires strain-resolved reference genomes.

To explore the possibility that the coassembled genomes contain contributions from more than a single strain, we further examine the comparison between the 19 coassembled genomes and cultured isolates of the same species; each of these isolates represents only a single strain. In general, the coassembled genome of a species with multiple strains contains some contigs specific to each strain; not all of these contigs appear in the single-strain genomes of the corresponding isolates. Consequently, we determine the shared genome fraction—the percentage of bases in each coassembled genome that are shared with isolate genomes from the same species. We find that for the comparison in 16 species, the shared genome fraction is above 96% and the ANI value exceeds 99.9%; these data suggest that each of these 16 coassembled genomes represents a single strain. By contrast, for the remaining three species, *Blautia obeum*, *B. vulgatus*, and *Parasutterella excrementihominis*, the shared genome fraction is far lower (between 70 and 90%) and ANI are all <99.6% (fig. S7). These lower values suggest that the genomes of these three species may include multiple strains or strains that do not appear among the cultured isolates. In principle, directly comparing all pairs of SAGs to estimate the fraction of their shared genomes could distinguish strains. However, the coverage of each SAG is expected to be <25% on average, for example 7% of the genome for *B. vulgatus*. This coverage suggests that such pairwise comparisons will not be reliable and instead motivates a different approach.

To distinguish strains, we develop a method that leverages the differences among homologous sequences between SAGs, specifically the single-nucleotide polymorphisms (SNPs). To illustrate this method we examine ~900 SAGs of *B. vulgatus*—the most abundant of the three species—and align reads from each SAG against the coassembled *B. vulgatus* genome, then identify ~12000 total SNP locations. For each SAG, we determine the SNP coverage, the fraction of all SNP locations in the genome that occur among the reads of that SAG; this SNP coverage is 8% on average, comparable to the average genome coverage. For each pair of SAGs, we measure the fraction of total SNP locations that occur in both and find this fraction to be ~0.7%, corresponding to ~80 SNPs, which is consistent with roughly the square

of the SNP coverage. Microbes of the same strain have nearly identical genomes (12, 14) such that two SAGs representing the same strain almost always have the same base at each SNP location shared by both SAGs; conversely, SAGs representing different strains show considerably lower similarity (61). Inferring the similarity of the bases at shared SNP locations in each pair of SAGs is governed by a binomial process; therefore, the average of 80 SNPs in each SAG pair should be sufficient for a robust inference, with an uncertainty of 6% or less. Consequently, the comparison of SNPs provides a promising approach to determine strains.

To test this possibility, in all pairs of SAGs, we examine the bases at all shared SNP locations and determine the fraction of locations where both SAGs have the same base. To probe whether these SAGs fall into any distinct groups, we visualize the SNP similarity between all pairs of SAGs with dimensional reduction (63). Notably, we find that the SAGs fall into four clearly distinct clusters as shown

in Fig. 3A. We independently validate the presence of these SAG groups with hierarchical clustering, which yields the same groupings with 99.8% overlap (fig. S8).

To test whether these clusters correlate with different strains, we examine the bases at SNP locations within each SAG cluster. We determine which base occurs most frequently at each SNP location; the set of these bases at each SNP location forms the consensus genotype of each SAG cluster. Then, for each SAG, we calculate the fraction of its SNPs that have the same base at the corresponding location in the consensus genotype of each of the four SAG clusters. Within each SAG cluster, we find that constituent SAGs share extremely high SNP similarity with the corresponding consensus genotype. For example, in the two clusters with the highest number of SAGs, almost all have the same base in >99% of the SNP locations as shown in the scatterplot and histograms in Fig. 3B. By contrast, SAG clusters show much lower overlap with the consensus genotypes of other clusters; for the two clusters with the

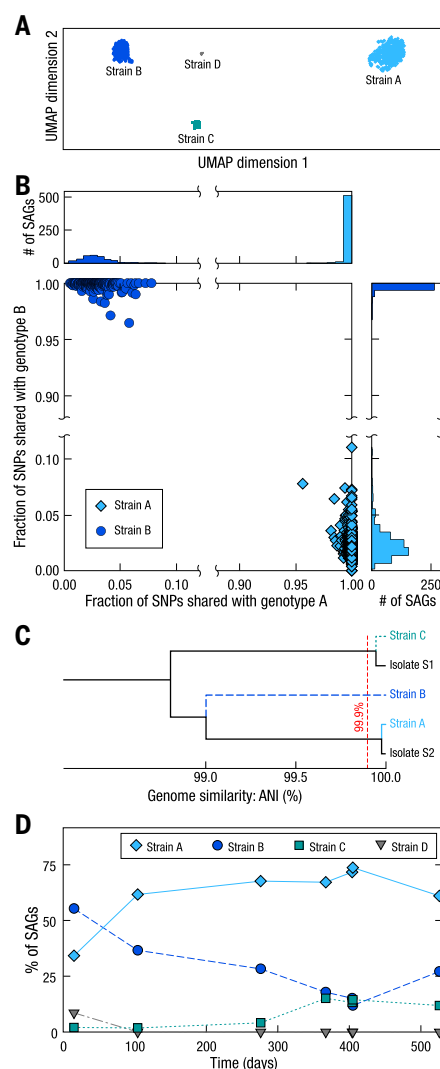


Fig. 3. Strain-resolved genomes of *B. vulgatus* in the human gut microbiome. (A) Dimensional reduction (UMAP) visualization of *B. vulgatus* SAGs, based on comparison of their sequences at SNP locations. SAGs fall into four distinct, widely separated clusters; the symbol for each SAG is colored according to the cluster in which it is grouped. (B) Scatterplot and histograms illustrating the fraction of SNPs from each SAG that match consensus genotypes for SAGs in the two most abundant clusters, A and B. In almost all cases, each SAG shares the same base in more than 99% of the SNP locations in its corresponding consensus genotype; by contrast, the SNP overlap with the consensus genotype of the other cluster is much lower, typically 5% or less. The symbols in each cluster are colored as in (A). (C) Phylogeny of the coassembled high- and medium-quality genomes of *B. vulgatus* strains and comparison with the corresponding genomes of strains of isolates cultured from the same human donor. The horizontal axis of the dendrogram represents the ANI values between these strain-resolved genomes, demonstrating that coassembled strain C and isolate S1 are the same strain; similarly, coassembled strain A and isolate S2 are the same strain. By contrast, the second most-abundant strain, B, does not appear among the isolates cultured from the same human donor. (D) Relative abundance of the four *B. vulgatus* strains in the seven longitudinal samples.

highest number of SAGs, all SAGs in each cluster share fewer than 10% of the bases at SNP locations with the consensus genotype of the other cluster, as shown in the figure. These trends persist among the other clusters (fig. S8). Together, these results provide strong evidence that SAGs within these clusters represent the same strain.

To further examine whether these four clusters correspond to actual *B. vulgatus* strains, we coassemble the reads within each SAG cluster. We obtain high-quality genomes for the two groups with the most SAGs, which we label candidate strains A and B; one medium-quality genome, C; and one additional genome of lower quality, D (table S4). We compare these coassembled genomes with the genomes of two distinct *B. vulgatus* isolate strains cultured from the same human donor (12). We find that both isolate genomes have closely matching coassembled counterparts (A and C) with ANI values and shared genome fractions exceeding 99.9 and 97%, respectively, as shown in Fig. 3C. These high values are consistent with those that occur between genomes of the same strain, thereby providing strong evidence that these coassembled genomes each represent a single, genuine strain of *B. vulgatus*. Notably, the second-most populous cluster—candidate strain B, with several hundred SAGs—does not appear among the nearly one hundred isolates of *B. vulgatus* cultured from the same human donor (12). Together these results demonstrate the capabilities of this SNP-based approach to correctly identify both the major known strains of *B. vulgatus* and potential new strains that have not been cultured, while at the same time enabling the accurate coassembly of their genomes.

We further apply this SNP-based analysis to the remaining species with high- or medium-quality species-level genomes. We find nine additional species with multiple strains and coassemble their genomes (fig. S9 and table S4). We compare the genotype of each SAG to its corresponding strain-resolved consensus genotype and observe that <1% of the SAGs have <95% similarity with the consensus genotype (fig. S10); these results are similar to those from *B. vulgatus* and provide strong confirmation that the separation of SAGs from different strains are robust. In total, we obtain 86 high- and medium-quality strain-resolved genomes from 76 species—from just one set of experiments—and compare to corresponding isolate genomes cultured from the same human donor. We find excellent agreement for *B. obeum*, with an ANI of 99.9% and shared genome fraction of 95%; this again confirms—just as in the case for *B. vulgatus*—that the coassembled genome represents a single, genuine strain (for the remaining multi-strain species, we have no isolate genomes of the same strains with which to compare).

Notably, we are able to achieve this accurate identification of strains and the coassembly of their genomes even with a level of coverage that yields an average of <100 shared SNP locations between all pairs of SAGs.

The capability to identify the strain of each individual SAG also enables us to follow the relative abundances of these strains over time in the human donor, giving insight on bacterial population dynamics. The abundances of these strains appear to shift only gradually throughout the year and a half over which samples were collected; for instance, we observe quite similar abundances in *B. vulgatus* in the two samples collected on successive days around day 400, as shown in Fig. 3D. These observations are consistent with previous studies showing that different *Bacteroidetes* species can colonize the human gut for decades stably, and that different strains of the same *Bacteroidetes* species can coexist with stable relative abundance (64).

The results demonstrate the capability of this approach to resolve subspecies strains and reconstruct their strain-resolved genomes, even when the SAGs have coverage of only ~10% of the genome. Furthermore, the droplet-based approach can obtain strain-resolved genomes from strains which have not been cultured; this is of particular importance in the human gut microbiome, where many strains are difficult to culture. Consequently, this method contributes a new way to examine the strain-resolved structure and dynamics of the genomic information within the human gut microbiome independent of the bias imposed by what has been cultured. These high-quality, strain-resolved genomes from a broad range of strains from the gut microbiome of a single human donor not only allow greater precision in the identification of a large majority of SAGs, but further enable the probing of broader genomic aspects of the microbial community, particularly those involving microbes of different strains.

HGT within the human gut microbiome

One particularly notable genomic aspect of microbial communities is how microbes exchange genetic information; one of the most well-known mechanisms is HGT, which is frequently observed within the human gut microbiome (20, 21, 65, 66). In general, the genomes of different bacterial species will differ considerably; however, one of the major indicators of HGT is a nearly identical sequence shared between genomes from different species (21, 67). The large number of strain-resolved genomes originating from the gut microbiome of a single human donor offers the potential to detect HGT by identifying the common sequences shared between specific microbial taxa.

To explore this sequence matching approach, we designate an HGT event between genomes

from two species as the presence of a common sequence of at least 5 kb with 99.98% similarity. We apply these criteria to all 57 high-quality strain-resolved genomes, filter out potential contamination due to SAG merging (fig. S11), and observe 265 HGT sequences between 90 pairs of strains from different species, which are all HGT events within the same phylum: 65 strain pairs are within Firmicutes and 25 are within Bacteroidetes.

To evaluate whether these events might be false positives caused by contamination, we align the reads from all SAGs of each species pair against each HGT sequence, and determine the fraction of all SAGs that have adequate coverage; under a null hypothesis that if an observed HGT event were in fact a result of contamination and the sequence was absent from one of the species, then only a small fraction of its corresponding SAGs would align to the HGT sequence with sufficient coverage. Instead, we find that all of the observed HGT sequences align to a number of SAGs considerably greater than that expected under the null hypothesis in both species of each pair, thereby confirming that there are no false positives (fig. S12). Furthermore, we examine the HGT sequences from the pairs of species with corresponding cultured isolates and find that 100% of the HGT sequences determined from the coassembled genomes occur in the isolate genomes of both species.

The HGT sequences we observe encode genes involved in a variety of metabolic, cellular, and informational functions (table S5); genes indicative of phage, plasmid, and other forms of mobile genetic elements exist in ~80% of the observed HGT sequences. Among the 49 species with a single high-quality strain, we observe 66 HGT events, as shown in Fig. 4A. Notably, among the species with multiple high-quality strains we observe that individual strains of *Agathobacter faecis*, *Faecalicatena faecis*, and *Anaerostipes hadrus* exchange genes with different Firmicutes species whereas both strains of *B. vulgatus* exchange genes only with the same six other Bacteroides species, as shown in Fig. 4B. Together, these data demonstrate the ability to resolve HGT to the level of individual strains.

To determine whether any of these HGT events involve more than two strains, we identify all of the genes that occur within HGT regions and count the number of strains whose HGT sequences contain each gene. We observe that approximately half of the genes are shared among three or more species, providing strong evidence that these HGT events emerged within this single human donor. Within Bacteroidetes, genes detected from HGT sequences are shared by an average of 3.2 strain-resolved genomes versus 2.6 strains within Firmicutes, as shown in Fig. 4C (table S6).

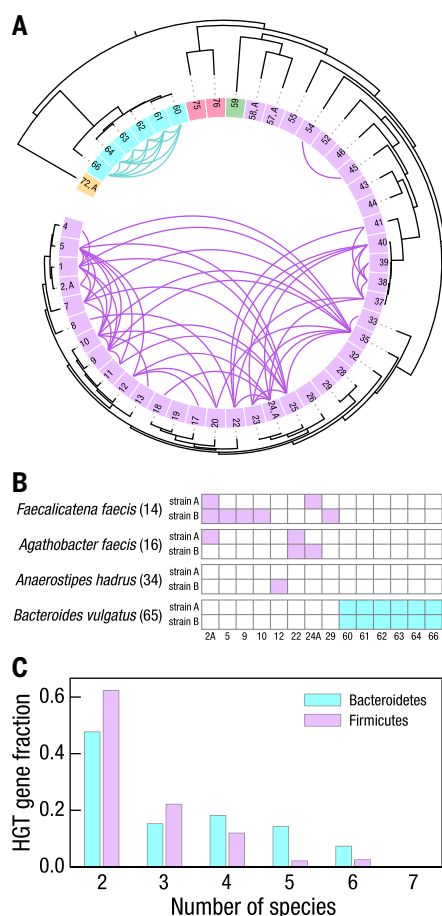


Fig. 4. HGT among bacterial strains within the human gut microbiome of a single donor.

(A) HGT among the 49 species with a single high-quality strain-resolved genome, following the order, numbers, and colors of Fig. 2. Detected HGT between two genomes indicated with a curve whose color matches that of the phylum of each species pair. (B) HGT between species with multiple high-quality strain-resolved genomes and species with single high-quality strain-resolved genomes, following the numbering in (A). For the bacteria in phylum Firmicutes (*Agathobacter faecis*, *Faecalicatena faecis*, and *Anaerostipes hadrus*), each strain has HGT with different sets of species. For the phylum Bacteroidetes, the only multistrain species is *B. vulgatus*, which has HGT between both of its strains and all other species in this phylum. (C) Distribution of the number of species in which HGT genes are shared. Approximately half of the genes in these HGT sequences are shared among more than two species; several genes occur in six or seven bacterial strains.

Notably, we find several genes that occur in the HGT sequences of six or seven Bacteroidetes strains. We examine the HGT sequences containing these particular genes and find that these sequences are connected with an integrative conjugative element containing a

type VI secretion system (T6SS), consistent with previous analysis using cultured isolates of *Bacteroides* from the same human donor (14); T6SS is one of the most-studied systems in *Bacteroides* that mediates interstrain competition between *Bacteroides* strains and has been shown to transfer between members of the same microbiome. In Firmicutes, we also observe genes shared among HGT sequences of six different strains; these HGT sequences contain genes annotated as recombinase, suggestive of an integrative mobile element or prophage.

Together, these data provide strong evidence that our methodology detects HGT widely and robustly, among strains of many species from multiple phyla within the gut microbiome of a single human donor. The detection of HGT among six or more species within this single microbiome suggests that HGT may have important functional consequences to the recipient strains. These methods provide new tools to investigate the interactions of multiple microbes within the human gut microbiome.

Host-phage association in the human gut microbiome

The ability to investigate microbial interactions within the human gut microbiome is not limited to only bacteria, but also includes other types of microbes. Indeed, the diversity analysis reveals the presence of viruses—specifically crAssphage, the most abundant bacteriophage recognized at present from the human gut microbiome (68, 69). The general regulatory role of bacteriophages, thought to modulate the abundance and behavior of bacteria, is only beginning to be understood within complex microbial communities (70, 71). The droplet-based method encapsulates not only an individual bacterium but also any bacteriophages physically colocalized with it, providing a direct means to probe host-phage association. To explore this association, we compare the reads in each SAG to the crAssphage genome; we find that a few dozen SAGs contain a substantial fraction of crAssphage-aligned reads. Moreover, many of these SAGs also contain a significant fraction of reads which do not align to the crAssphage genome but instead to bacterial taxa; we align these reads against the coassembled genomes of 76 species to identify which, if any, bacterial species might associate with crAssphage strain in this particular human donor.

Significantly, we find that 14 SAGs are associated with only one species, *B. vulgatus* (P value = 4×10^{-9} , Fisher's exact test) (table S7) and that no other species associates significantly with crAssphage, as shown in Fig. 5A. These data strongly suggest *B. vulgatus* as the in vivo host species for crAssphage in this human donor, consistent with previous

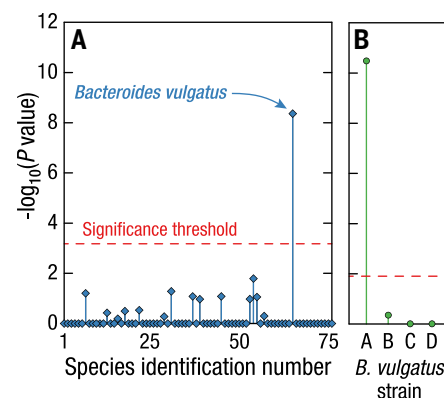


Fig. 5. Host-phage association with strain specificity in the human gut microbiome.

(A) Association between the bacteriophage crAssphage and bacterial species with high- or medium-quality genomes, with species numbers as in Fig. 2. All P values are calculated with one-sided Fisher's exact test. The only bacterial species that is significantly associated with crAssphage is *B. vulgatus*. (B) Association between the four strains of *B. vulgatus* and crAssphage. Only one specific strain of *B. vulgatus*—the most abundant strain, A—is significantly associated with crAssphage.

evidence that crAssphage is likely to be associated with *Bacteroides* species (68, 72). The statistical significance of the association indicates that this is not a result of simple random coencapsulation.

Furthermore, the unambiguous assignment of each SAG to one of the multiple strains of *B. vulgatus* enables even more precise characterization of in vivo host-phage association to the level of specific bacterial strains. We find that 13 SAGs represent the single *B. vulgatus* strain A, the most abundant (P value = 3×10^{-11}), as shown in Fig. 5B.

These data demonstrate the unique advantages of the droplet-based approach to establish accurate in vivo host-phage association not only for an individual species but even more precisely to a specific strain. We identify which bacterial strains interact with bacteriophages and which strains do not; the genomic differences between these strains provide preliminary data that may contribute to understanding of the molecular mechanisms underlying these host-phage interactions and their longitudinal dynamics in the human gut microbiome.

Discussion

Using Microbe-seq, a high-throughput method combining experiment and computation for single-microbe genomics, we obtain—without culturing—the genomic information of tens of thousands of individual microbes and de novo coassemble the strain-resolved genomes from 76 species, a large fraction of which

have not been cultured. This high-throughput microfluidics-based approach allows for more practical individual examination of a sufficient number of microbes to achieve these results, even with an average coverage of less than a quarter of the genome. The close agreement with strains for which we have corresponding cultured isolates confirms the accuracy of this approach. These strain-resolved genomes enable the reconstruction of an HGT network within a single human; when sampled over time, these data may allow the monitoring of microbe response, at the level of specific genes in specific strains, to selective pressures unique to that person, such as disease, diet, or antibiotic treatment. In addition, the *in vivo* association between specific strains of bacteriophages and bacteria could provide specific starting points to investigate how phages modulate microbial composition and possibly guide subsequent development of phage-based therapeutics.

Scaling up the analysis to examine an order of magnitude (or more) microbes from complex microbial communities would shed light on important questions without requiring any other qualitative changes to the existing procedures. In the human gut microbiome, sequencing hundreds of thousands of cells would likely allow for identification of nearly all of the present species and strains, thereby enabling far more accurate surveys of diversity and abundance. Moreover, expanding the present investigation to a larger population of humans could allow direct exploration of the effects on human health of key microbial pathways and genes, opening up potential directions for future therapeutic developments.

We envision several routes for further technical improvement. Integrating long-read sequencing technologies are likely to lengthen the coassembled contigs considerably, improving the quality and completeness of resulting genome assemblies (28). Exploring additional lysis conditions would improve the evenness and efficiency of lysis, potentially allowing investigation of microbes in other phyla or even other kingdoms such as fungi. Combining these methods with functional sorting, such as IgA bind-and-sort, would correlate functional outcomes with strain-level genomic information and single-cell resolution.

Microbe-seq provides a particularly effective and practical approach in a single laboratory-scale experiment to identify and sequence fully all of the major strains in microbial communities beyond the human gut microbiome, without any *a priori* knowledge of constituent microbes. The practical improvements provided by our methodology may make feasible the investigation of microbial communities that affect the environments, lives and health of human communities that otherwise lack

access to the resources to even begin to investigate these effects.

Materials and methods

Experimental model and subject details

We obtain stool samples from OpenBiome, a nonprofit stool bank, under a protocol approved by the institutional review boards at MIT and the Broad Institute (IRB protocol ID # 1603506899). The subject is a healthy male, 28 years old at initial sampling, screened by OpenBiome to minimize the potential of carrying pathogens and de-identified before receipt of samples. We homogenize stool samples from this donor, mix with 25% glycerol, and freeze at -80°C . For each experiment, we wash 1–3 μL of stool sample in 1 mL 1X PBS three to five times and resuspend it in 1X PBS with 15% (v/v) Optiprep density gradient medium (Sigma-Aldrich D1556) as the microbial suspension.

Mock community

We culture four bacteria strains, *Bacillus subtilis* ATCC 6051-U, *Escherichia coli* ATCC 25922, *Klebsiella pneumoniae* ATCC 35657, and *Staphylococcus aureus* ATCC 6538 in 1 mL LB liquid medium (L3522 Sigma Aldrich) overnight. We wash each bacterial culture with 1 mL 1X PBS three to five times and resuspend bacteria in 1X PBS with 15% (v/v) Optiprep density gradient medium (Sigma-Aldrich D1556). We combine approximately the same volume of these four bacterial strains and dilute to a final concentration of 5–50 million microbes/mL.

Microfluidic device fabrication

We print the device designs (fig. S1) as photomasks (CAD/Art Services, Inc.), and fabricate devices according to well-established soft-lithography procedures (73). We use photolithography and the photomasks to transfer each device design to a silicon wafer with SU8 photoresist. We cast polydimethylsiloxane (PDMS) (Sylgard 184) on the SU8 structure, where the SU8 structure on silicon wafer serves as a master for replica molding. We bake at 65°C for at least 2 hours to cure the PDMS and delaminate the resulting PDMS replicas off the master. We seal with glass slides (Corning, 2947) to create the microfluidic devices and make their surfaces hydrophobic by flowing Aquapel (PGW Auto Glass, LLC) through the channels. We remove excess residual Aquapel by flowing compressed air in the channels of microfluidic devices and bake the devices at 65°C overnight.

Isolation and lysis

We isolate microbes by encapsulating them into droplets with lysis reagents using a microfluidic device (fig. S1A and movie S1). We put the microbial suspension in a 1 mL syringe (BD Luer-Lok 1-mL syringe, 309628) and connect

the syringe to the microbial suspension device inlet via a needle (BD Precisionglide syringe needles, Z192384-100EA, Sigma Aldrich) and polyethylene tubing (BB31695-PE/2, Scientific Commodities, Inc.). We connect similarly the lysis reagents and oil, 2% (w/v) surfactant (RAN biotechnologies, 008-FluoroSurfactant) in HFE 7500 (3M), to the device. We use flow rates of 30 $\mu\text{L}/\text{h}$ for the microbial suspension, 120 $\mu\text{L}/\text{h}$ for lysis reagents, and 300 $\mu\text{L}/\text{h}$ for the oil. We collect droplets from the device outlet into a PCR tube and replace the oil from the bottom with 100 μL of 5% (w/v) oil. We add 100 μL mineral oil (MI499, Spectrum Chemical MFG Corp.) on top of the emulsion to avoid the evaporation of the aqueous phase in the droplets. We remove most of the oil from the bottom of the tube and incubate to lyse the microbes inside droplets.

We prepare an 80 μL lysis reagent mix for each experiment: 10 μL green buffer (prepGEM Bacteria, PBA 0100), 1 μL lysozyme (prepGEM Bacteria, PBA 0100), 1 μL prepGEM (prepGEM Bacteria, PBA 0100), 1 μL lysostaphin (1 mg/mL in 20 mM sodium acetate, pH 4.5, Sigma, L7386), 2 μL 20 mg/mL bovine serum albumin (BSA, B14, Thermofisher), 2 μL 10% tween-20 (diluted from Tween-20, Sigma-Aldrich, P9416-50mL), 1 μL 100 uM random hexamer with the last two 3' end bases phosphorothioated (IDT), and 62 μL water.

The incubation program for lysis is: 37°C for 30 min, 75°C for 15 min, 95°C for 5 min and sample storage at 4°C .

Whole-genome amplification

We transfer the droplet emulsion to a syringe and reinject droplets into a microfluidic merger device (48) (fig. S1B and movies S2 and S3). In the same device, we use a separate droplet maker to form droplets that encapsulate multiple displacement amplification (MDA) reagents. We synchronize the frequency of sample droplet re-injection and reagent droplet-making to form droplet pairs. Applying electric fields of 50–200 V at a frequency of 25 KHz through a pair of electrodes, we merge each droplet pair to add MDA reagents. We use flow rates of 60 $\mu\text{L}/\text{h}$ for sample droplets, 100 $\mu\text{L}/\text{h}$ for 2% (w/v) oil (fig. S1B, label 2), 75 $\mu\text{L}/\text{h}$ for MDA reagents, and 250 $\mu\text{L}/\text{h}$ for 2% (w/v) oil (fig. S1B, label 4). We incubate to amplify microbial genomes.

We prepare a 100 μL MDA mix for each experiment: 16 μL 10X phi29 DNA Polymerase Buffer (Lucigen, 30221-1), 0.5–2 μL 100 uM random hexamer with last two 3' end bases phosphorothioated (IDT), 0.8–3.2 μL 25 mM dNTPs (Thermo Fisher, R1121), 8 μL phi29 DNA Polymerase (Lucigen, 30221-1), 2 μL 20 mg/mL bovine serum albumin (BSA, B14, Thermofisher), and we add water to make the total volume to 100 μL .

The incubation program for MDA is: 30°C for 6–8 hours, 65°C for 10 min and sample storage at 4°C.

Tagmentation

We merge sample droplets with droplets containing commercially available tagmentation reagents (Nextera), utilizing a different droplet merger device (fig. S1C and movies S4 and S5). We use flow rates of 25 $\mu\text{L}/\text{h}$ for sample droplets, 100 $\mu\text{L}/\text{h}$ for 2% (w/v) oil (fig. S1C, label 2), 75 $\mu\text{L}/\text{h}$ for tagmentation reagents, and 300 $\mu\text{L}/\text{h}$ for 2% (w/v) oil (fig. S1C, label 4). We incubate to tagment these DNA products.

We prepare a 90 μL Nextera mix for each experiment: 60 μL TD Tagment DNA Buffer (Illumina, 15027866), 12 μL TDE1 Tagment DNA Enzyme (Illumina, 15027865), 1.8 μL 20 mg/mL bovine serum albumin (BSA, B14, Thermofisher), 1.8 μL 10% tween-20 (diluted in water from Tween-20, Sigma-Aldrich, P9416-50mL), and 14.4 μL water.

The incubation program for tagmentation is: 55°C for 10 min, and sample storage at 10°C.

Bead synthesis

We synthesize beads used for combinatorial barcoding by adopting a previously reported method (44, 74). In brief, we make droplets containing acrydite-modified DNA oligos using a photo-cleavable linker (table S8, Hydrogel DNA primer, IDT) and acrylamide:bisacrylamide solution. We keep these droplets at 65°C overnight to polymerize them into uniform soft gel beads covalently bonded to the DNA oligos by photo-cleavable linkers. We extend DNA oligos on beads enzymatically with a two-step split-and-pool synthesis protocol to prepare beads with a diverse barcode sequence library. At the first split-and-pool synthesis step, we evenly split beads into a 96-well plate where each well contains a unique barcode-1 oligo (table S8, IDT). We anneal these oligos with hydrogel oligos and extend them with Bst 2.0 DNA polymerase (M0537L, NEB). After the first split-and-pool synthesis step, we pool beads, wash them and evenly split them into a 384-well plate where each well contains a unique barcode-2 oligo (table S8, IDT). We perform the second barcode strand synthesis in the same way as we extend the first barcode strand. We avoid exposing beads to strong light.

Each soft gel bead has millions of primers with the same sequence. Each full sequence contains two barcode regions: the first region has a diversity of 96; the second region, 384. Overall, the barcoding bead library has 36864 (96×384) possible sequences.

Bead preparation for barcoding

We wash 200 μL of beads with 1 mL bead wash buffer (10 mM pH 8.0 Tris-HCl, 0.1 mM EDTA and 0.1% (v/v) Tween-20), three times

in a tube. We withdraw supernatant from the top, leaving 500 μL in the tube. We add 300 μL water and 200 μL 5X Phusion HF detergent-free buffer (F520L, Thermo Fisher) to the tube. We vortex the beads and keep them at room temperature for 1 min. We centrifuge beads, remove supernatants, and use these beads for barcoding.

Barcoding

We merge sample droplets with droplets containing PCR reagents and a barcoding bead, using a droplet-merger microfluidic device (fig. S1D and movies S6 to S8). We use flow rates of 50 $\mu\text{L}/\text{h}$ for sample droplets, 100 $\mu\text{L}/\text{h}$ for 2% (w/v) oil (fig. S1D, label 2), 15–25 $\mu\text{L}/\text{h}$ for beads, 140 $\mu\text{L}/\text{h}$ for PCR reagents, and 400 $\mu\text{L}/\text{h}$ for 2% (w/v) oil (fig. S1D, label 5). We release barcode oligos from beads by exposing droplets to UV light (365 nm at $\sim 10 \text{ mW}/\text{cm}^2$, BlackRay Xenon Lamp) for 10 min. We perform PCR to barcode the DNA in the droplets.

We prepare a 240 μL PCR mix for each experiment: 136 μL water, 68 μL 5X Phusion HF detergent-free Buffer (F520L, Thermo Fisher), 8 μL 10 mM dNTPs (diluted from 25 mM dNTP mix, Thermo Fisher, R1121), 16 μL 10 μM RNS primer (table S8, IDT), 4 μL Phusion high-fidelity DNA polymerase (F530L, Thermo Fisher), 4 μL 20 mg/mL bovine serum albumin (BSA, B14, Thermofisher), 4 μL 10% tween-20 (diluted from Tween-20, Sigma-Aldrich, P9416-50mL).

The incubation program for barcoding is: 72°C for 4 min, 98°C for 30 s; 10 cycles of 98°C for 7 s, 60°C for 30 s and 72°C for 40 s; 72°C for 5 min, and sample storage at 4°C. We use slow ramping of 2°C/s at this step.

We observe the merger of some droplets after PCR, possibly during the high-temperature stage of PCR; such larger droplets may contain DNA from multiple microbes. We remove most of these droplets with droplet-size filter microfluidic device (75) (fig. S1E, movies S9 and S10) with flow rates of 120 $\mu\text{L}/\text{h}$ for sample droplets and 2 mL/h for 2% (w/v) oil.

Droplet pooling and sequencing library preparation

We break the emulsion of droplets by adding 200 μL 20% (v/v) PFO (1H,1H,2H,2H-Perfluoro-1-octanol, 370533 Sigma Aldrich) in HFE 7500 (3M) into each sample after PCR. We purify the aqueous phase with 1.1X volume AMPure beads (A63881, Beckman Coulter) and resuspend into 32 μL DNA suspension buffer (10 mM pH 8.0 Tris-HCl and 0.1 mM EDTA). We use PCR to add sequencing adapters for sequencing (Illumina) and a sample index (Nextera index) to each purified DNA sample so we can sequence multiple samples in one sequencing run.

We prepare a 50 μL PCR mix for each experiment: 2.5 μL water, 10 μL 5X Phusion HF detergent-free Buffer (F520L, Thermo Fisher),

1 μL 10 mM dNTPs (diluted from 25 mM dNTP mix, Thermo Fisher, R1121), 2 μL 10 μM P5PE1 primer (table S8, IDT), 2 μL Nextera i7 primer (Illumina), 0.5 μL Phusion high-fidelity DNA polymerase (F530L, Thermo Fisher), and 32 μL DNA sample in DNA suspension buffer.

The incubation program for PCR is: 98°C for 30 s; 5–10 cycles of 98°C for 7 s, 60°C for 30 s, and 72°C for 40 s; in the end, 72°C for 5 min and sample storage at 4°C.

We purify samples with 0.8X volume AMPure beads (A63881, Beckman Coulter) and resuspend DNA products into 20 μL DNA suspension buffer (10 mM pH 8.0 Tris-HCl and 0.1 mM EDTA). We store these products at -20°C before sequencing.

Illumina sequencing

We sequence at depths ranging between ten thousand and two hundred thousand reads for each microbe. A custom read-1 primer (table S8, IDT) is required for the sample to be sequenced. For a 100 base-pair (bp) sequencing run, we use the following sequencing length configurations: read-1 sequence: 45 bp, which contains the barcode sequence; index-1 sequence: 8 bp; read-2 sequence: remainder, which contains the microbial sequence. For a 300 bp sequencing run, we use the following sequencing length configurations: read-1 sequence: 150 bp, the first 45 bp are barcode sequences, the last 75 bp are microbial sequences, and those in the middle are adapter sequences; index-1 sequence: 8 bp; read-2 sequence: remainder, which contains the microbial sequence.

Preprocessing of raw sequencing data

We group raw sequencing reads based on the 36864 barcodes, excluding barcodes associated with too few reads ($\sim 15\%$ of total reads) and those with significantly more reads than other barcodes likely due to droplet merging ($\sim 5\%$ of total reads). For the remaining barcodes, we designate the collection of microbial sequences associated with a single barcode as a single amplified genome (SAG). We use Trimmomatic (76) (version 0.36, LEADING:25 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:30) to remove low-quality reads and adapter sequences from each SAG for following analysis.

Mock sample alignment, quality assessment, and coverage

We use Bowtie2 (52) (version 2.2.6, default parameters) to align reads from each SAG to the combined genome of the four reference genomes (RefSeq: GCF_002055965.1, GCF_004151095.1, GCF_001936035.1, GCF_002025145.1), which reports the best hit of each read. We use SAMtools (77) (version 1.9) to check the number of reads that align to each of the four genomes and to calculate the purity of each SAG. For each SAG with high purity (>0.95), we align

its reads to the most-aligned reference genome to determine its genome coverage.

Genome coassembly of microbial species in the human gut microbiome

We use SPAdes (53) (version 3.13.0, -sc-careful) to de novo assemble genomes from the reads of each of the 21914 SAGs. We compute and compare signatures of these assembled genomes using sourmash (78) (version 2.0, k-mer 51, default setting), which produces a matrix of estimated similarities between genomes. We use a hierarchical clustering method (SciPy version 1.1.0, method: complete, metric: Euclidean, criterion: “inconsistent”, and threshold: 0.95) to group SAGs into bins. We verify 0.95 as a threshold using mock samples. This set of parameters groups bins conservatively, minimizing the improper grouping of SAGs from different species. We use all the reads within each bin to coassemble a tentative genome, compare tentative genome similarities, and cluster the bins. We iterate this process until more than 10% of the assembled genomes have more than 10% contamination (estimated by CheckM version 1.0.13, default parameters) (56), which implies false clustering of SAGs; through four rounds, we group the 21914 SAGs into 364 bins.

To split bins that might contain SAGs from multiple species, we examine contig alignment patterns. Within each of the 364 bins, we align reads from each SAG to the de novo coassembled genome from that bin using bowtie2 (52) (default parameters). For each contig in the tentative genome with more than 1000 bp, we construct a vector for each contig with the number of reads aligned to the contigs from each SAG. We use a hierarchical clustering method (method: ward, default parameters) to group vectors of contigs into two groups. For each SAG, if >95% aligned reads are aligned to one of the two groups of contigs, it is designated as a SAG associated with that group of contigs. We assume that the remaining SAGs are a mixture of multiple species and exclude them from further analysis. We iterate this binary splitting process until we exclude more than 60% of the SAGs in the current bin, or both resulting new bins have fewer than 10 SAGs, or the change between the resulting new bin and the current bin is fewer than three SAGs. Using this process, we obtain 400 bins whose constituent SAGs we expect to represent a single species, with minimal contamination.

To combine bins of the same species for genome assembly, we use fastANI (55) (version 1.2, default parameters) to calculate average nucleotide identity (ANI) between all pairs of these 400 bins. Applying the commonly used ANI > 95% threshold, above which two genomes are considered to represent the same species, we generate 234 new species-level bins.

We de novo assemble reads from all SAGs within each of these 234 bins and remove contigs shorter than 500 bp. To further eliminate contigs that may originate from other species within each genome, e.g., as a result of random contamination in individual SAGs, we fit a normal distribution with the coverage of contigs on a log scale and remove those contigs with coverages that are more than two standard deviations away from the mean of the distribution.

Among these 234 genomes, 76 genomes are of high-quality (>90% completeness and <5% contamination) or medium-quality (>50% completeness and <10% contamination), as assessed by CheckM (56) (default parameters). We use fastANI (55) (default parameters) to compare the genomes of these 76 bins to all microbial genomes (RefSeq as of September 2019), and to the published collection of more than a thousand cultured-isolate whole genomes (12). We identify the closest corresponding species-level genomes with ANI > 95% in both databases. The closest genomes in RefSeq to species *Alistipes onderdonkii*, *Bacteroides fragilis*, and *Bacteroides ovatus* are cultured isolate whole genomes from the same donor, reported previously (79); we exclude these three genome pairs from the ANI and shared genome fraction analysis (fig. S7). We use BLASTn (BLAST+, version 2.10.0) (80) (default parameters) to compare overlapping sequences between genome pairs.

The names of the species-level genomes in RefSeq are not always labeled consistently; for example, we have four species that are named as *Blautia obeum* in RefSeq, though their ANI values are less than 95%. We use both GTDB-Tk (59) (version 1.0.2, reference data version r89) and comparison to RefSeq genomes (as of September 2019) to assign taxonomies to all species. In the main text, we use taxonomies classified with GTDB-Tk and remove sub-genus names, such as “A”.

Phylogeny analysis of genomes

To construct the phylogeny of the 76 species with high-quality or medium-quality genomes, we extract amino acid sequences of six ribosomal proteins (Ribosomal_L1, Ribosomal_L2, Ribosomal_L3, Ribosomal_L4, Ribosomal_L5, and Ribosomal_L6), concatenate and align them with Anvi'o (version 6.1) (81). We construct a maximum likelihood tree with RaxML (82) (version 8.2.12, standard LG model, 100 rapid bootstrapping). We use iTOL (version 5.5) (83) to visualize and annotate the resulting dendrograms.

Diversity of the human gut microbiome samples

For each of the seven samples, we temporarily ignore the barcode information and combine all reads from all SAGs from the sample. We use Kraken2 (84) (version 2.0.8, default parameters)

to classify reads from each Microbe-seq dataset and corresponding metagenomic dataset (12) (standard Kraken database as of April 2019). For the analysis shown in fig. S4, we keep only the reads classified to a specific genus and use only this genus-level information for the comparison; similar analyses using all operational taxonomic units (OTUs) show similar results (table S2). For each metagenomic dataset, we align reads to the combined genome co-assemblies from the 364 bins, irrespective of whether the bin is species level. Metagenomic reads are first quality filtered with fastp (version 0.12.4, parameters: -f 15 -t 15 -q 36 -u 10) and then aligned to the combined genomes using bowtie2 (parameter: -very-sensitive-local). We obtained overall alignment rates of 98.26%, 98.74%, 98.63%, 96.65%, 96.63%, 96.11%, and 98.64% for each of the seven metagenomic samples.

Abundance bias between Microbe-seq and metagenomics

We compare relative abundance from the 76 species with high- or medium-quality genome coassemblies. We estimate the cell number for each species in the metagenomic dataset by aligning metagenomic reads to each species-level reference genome and computing the average sequencing depth between the 20th and 80th percentiles in genome-wide sequencing depth. We infer cell number in the Microbe-seq dataset by counting the number of SAGs that we assign to each species; we normalize this cell-number inference across all these species and average across the seven longitudinal samples to obtain a single relative abundance inference for all species.

Differentiating strains of the same species

We use *B. vulgatus* as an example in the main text to illustrate the strain differentiation workflow; we use the same computational pipeline for all other species, without changing parameters, to resolve their constituent strains. The uncertainty in similarity of the bases at shared SNP locations in each pair of SAGs is the standard deviation of the normal approximation of the binomial distribution: $\text{uncertainty} = \sqrt{p(1-p)/n}$, where p is the probability of the event and n is the number of events. In the case of *B. vulgatus*, $n=80$ and the uncertainty is <6%.

Within each of the species with high- or medium-quality species-level genomes, we align (52) each SAG to the assembled genome. We use bcftools (77) (mpileup, filters: snps and %QUAL>30) to identify high-quality single-nucleotide polymorphism (SNP) mutations. We designate a SAG with fewer than 2 reads aligned to a SNP, as well as fewer than 99% of its reads being the same at a SNP as unknown/unaligned at this location. We remove SNPs with fewer than 5% of SAGs aligned to the

location, and SNPs with fewer than two SAGs being the reference allele or fewer than two SAGs being the mutation allele. We also remove any SNP with fewer than 1% SAGs being the reference allele or fewer than 1% SAGs being the mutation allele. We remove any SAG that covers less than 1% or fewer than 10 of the kept SNP locations.

We identify thousands of SNP locations and remove up to 6% of SAGs. We construct a SNP vector to represent the base identity sequence of each SAG at each SNP location. To identify the number of strains of the species in the samples, we build a dendrogram of SAGs with hierarchical clustering (method: “ward”) using the SNP vectors of all SAGs. Although the number of clusters is not obvious from the dendrogram, we obtain a sequence of SAGs; in this sequence, SAGs with similar SNP sequences are closer. We compare similarities of SNP vectors between SAGs at their shared SNP locations and construct a similarity heatmap with SAGs ordered in the same sequence as the corresponding dendrogram. We observe block-diagonal squares in the heatmap, which indicates that SAGs within each square are closer to each other than to SAGs in other squares. Using the block-diagonal squares in the heatmap, we determine the number of strains, though this number is challenging to determine accurately for species with relatively few SAGs (<200) and for species with potentially more than two strains. For *Blautia obeum*, it is unclear whether there are 3 or 4 strains in the sample; for *Parasutterella excrementihominis*, it is unclear whether there are 2 or 3 strains. We apply UMAP (63) (default parameters) to the SNP data to create dimensional-reduction plots (fig. S9).

To remove SAGs that have reads from microbes of multiple strains, we construct the consensus genotype of each strain by comparing the SNP vectors of SAGs of the same strain. If more than 90% of the values at a SNP location from all SAGs within the strain are the same, we use the value for this SNP in the consensus genotype for the strain; otherwise we drop this SNP location for this strain. We compare the SNP vector of each SAG to the consensus genotype of each strain and assign strains to those SAGs that match more than 95% locations at the consensus genotype of only one strain, which excludes fewer than about 1% of the SAGs from each species. We coassemble strain-resolved genomes with reads from all SAGs in each of these assigned strains with SPAdes (53) using default parameters.

Horizontal gene transfer analysis

We detect HGT events by searching for blocks of DNA sequences shared by a pair of strain-resolved genomes that are longer than 5000 bp and more than 99.98% identical (14, 67). Assuming that species from the gut microbiome evolve with a molecular clock of 1 SNP/genome/year

and that typical genome size is 5,000,000 bp, this set of criterion detects sequences that diverged within the past 1000 years and the HGT events likely emerged within the human host, based on known mutation rates (14). To filter out HGT sequences resulting from contaminated SAGs, we select all SAGs from each strain-resolved genome, and align reads from each SAG to the corresponding strain-resolved genome. We remove SAGs with an overall sequence alignment ratio below 90%, which eliminates three HGT sequences from two genome pairs, as no SAGs from one of the strain-resolved genomes have reads that cover the HGT sequences.

To further validate the remaining detected HGT sequences, we align reads from all the filtered SAGs from both HGT-associated species. We calculate the number of SAGs belonging to each strain-resolved genome with more than 500 bp coverage over the HGT sequence. We explore the statistical likelihood of the observed fraction of SAGs containing reads covering the HGT sequence. We build a null model that if we detect an artifactual HGT event between species A and B, that sequence actually only exists in the genome of species B, but appears in the SAGs of species A as a result of contamination. We assume a worse-than-real scenario that if a SAG from species A is contaminated by species B, this SAG will contain reads covering the false HGT sequence. We also assume a worse-than-real contamination rate of 20% SAGs for any strain and species. Under these assumptions, the upper limit for the probability that any SAG from species A is contaminated by B is: $20\% \times (\text{relative abundance of B}) = 0.2 \times N_b / N$, where N_b is the number of SAGs from species B, and N is the total number of SAGs. If the observed SAG number for species A is N_a , and the observed number of SAGs contaminated by B is up to x , then the probability that equal or more than x of the SAGs from species A are contaminated by species B is $1 - \text{binom.cdf}(x, N_a, 0.2 \times N_b / N)$; this calculated quantity represents the upper limit of the P value for the observed fraction of SAGs containing reads from the HGT sequences.

To explore whether these HGT events either emerged within this human subject or before both strains colonized the host, we compare our results to the baseline detectable HGT from strains that are not from the same human host. For 39 species that we find a corresponding high-quality genome assembly from the NCBI database, we select the single genome that most closely matches the strain-resolved genome from Microbe-seq using ANI. We apply our exact HGT criteria to this collection of 39 genomes from the NCBI database, and compare with the corresponding 39 strain-resolved genomes from Microbe-seq in fig. S13.

We predict genes (open reading frames, ORFs) from the HGT sequences using prokka

(85), (version 1.12, default parameters). We annotate ORFs using egg-nog-mapper (86) (version 3.0, parameter: -m diamond-tax_scope auto-go_evidence nonelectronic-target_orthologs all-seed_ortholog_evalue 0.001-seed_ortholog_score 60-query-cover 20-subject-cover 0). For each HGT sequence, we assign the sequence to a certain type of mobile element if ORF annotations contain signatures of mobile elements (detailed information in table S5). To examine how many species share the same HGT sequences, we cluster all the ORFs from all HGT sequences using CD-HIT (87) (version 4.7, 100% similarity). For each gene cluster, we count the number of species whose HGT sequences contain genes within the gene cluster (Fig. 4C and table S6). We cluster genes from only the HGT regions and the HGT sequences detected via our method, which are likely incomplete fragments of the original HGT events; therefore, the number of species we report for each gene is likely an underestimation.

Host-phase association analysis

To identify SAGs that are associated with both crAssphage and a bacterial cell, we use bowtie2 (52) (default parameters) to align reads in each SAG to the crAssphage genome (Refseq: GCF_000922395.1). We designate SAGs with more than 5% reads aligned to the crAssphage genome as containing significant crAssphage reads (raising this threshold to 10% of reads yields the same result); we align the non-crAssphage reads of these SAGs to each of the 76 high- or medium-quality genomes, as well as the combined genome of these 76 genomes. We define purity of these SAGs as the maximum number of reads aligned to individual genomes divided by the number of reads aligned to the combined genome. We identify SAGs with more than 50% of reads aligned to one of these 76 genomes, and with purity of more than 95%. We designate the species of the SAG as the species of the most aligned genome. We count the number of SAGs assigned to each species and perform the “one species versus remaining species” one-sided Fisher’s exact test.

REFERENCES AND NOTES

1. C. Huttenhower, D. Gevers, Human Microbiome Project Consortium, Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012). doi: 10.1038/nature11234; pmid: 22699609
2. J. Lloyd-Price et al., Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61–66 (2017). doi: 10.1038/nature23889; pmid: 28953883
3. S. Sunagawa et al., Tara Oceans coordinators, Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015). doi: 10.1126/science.1261359; pmid: 25999513
4. L. R. Thompson et al., A communal catalogue reveals Earth’s multiscale microbial diversity. *Nature* **551**, 457–463 (2017). doi: 10.1038/nature24621; pmid: 29088705
5. R. Sender, S. Fuchs, R. Milo, Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biol.* **14**, e1002533 (2016). doi: 10.1371/journal.pbio.1002533; pmid: 27541692

6. I. Cho, M. J. Blaser, The human microbiome: At the interface of health and disease. *Nat. Rev. Genet.* **13**, 260–270 (2012). doi: [10.1038/nrg3182](#); pmid: [22411464](#)
7. A. B. Shreiner, J. Y. Kao, V. B. Young, The gut microbiome in health and in disease. *Curr. Opin. Gastroenterol.* **31**, 69–75 (2015). doi: [10.1097/MOG.0000000000000139](#); pmid: [25394236](#)
8. V. K. Ridaura *et al.*, Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* **341**, 1241214 (2013). doi: [10.1126/science.1241214](#); pmid: [24009397](#)
9. T. Yatsunenko *et al.*, Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012). doi: [10.1038/nature11053](#); pmid: [22699611](#)
10. K. Z. Coyte, J. Schluter, K. R. Foster, The ecology of the microbiome: Networks, competition, and stability. *Science* **350**, 663–666 (2015). doi: [10.1126/science.1242602](#); pmid: [26542567](#)
11. S. Rakoff-Nahoum, K. R. Foster, L. E. Comstock, The evolution of cooperation within the gut microbiota. *Nature* **533**, 255–259 (2016). doi: [10.1038/nature17626](#); pmid: [27111508](#)
12. M. Poyet *et al.*, A library of human gut bacterial isolates paired with longitudinal multomics data enables mechanistic microbiome research. *Nat. Med.* **25**, 1442–1452 (2019). doi: [10.1038/s41591-019-0559-3](#); pmid: [31477907](#)
13. D. T. Truong, A. Tett, E. Pasoli, C. Huttenhower, N. Segata, Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017). doi: [10.1101/gr.216242.116](#); pmid: [28167665](#)
14. S. Zhao *et al.*, Adaptive Evolution within Gut Microbiomes of Healthy People. *Cell Host Microbe* **25**, 656–667.e8 (2019). doi: [10.1016/j.chom.2019.03.007](#); pmid: [31028005](#)
15. M. Scholz *et al.*, Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods* **13**, 435–438 (2016). doi: [10.1038/nmeth.3802](#); pmid: [26999001](#)
16. C. Yang *et al.*, Fecal IgA Levels Are Determined by Strain-Level Differences in *Bacteroides ovatus* and Are Modifiable by Gut Microbiota Manipulation. *Cell Host Microbe* **27**, 467–475.e6 (2020). doi: [10.1016/j.chom.2020.01.016](#); pmid: [32075742](#)
17. J. P. Nataro, J. B. Kaper, Diarrheagenic *Escherichia coli*. *Clin. Microbiol. Rev.* **11**, 142–201 (1998). doi: [10.1128/CMR.11.1.142](#); pmid: [9457432](#)
18. P. Manrique, M. Dills, M. J. Young, The Human Gut Phage Community and Its Implications for Health and Disease. *Viruses* **9**, 141 (2017). doi: [10.3390/v9060141](#); pmid: [28594392](#)
19. S. Minot *et al.*, Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 12450–12455 (2013). doi: [10.1073/pnas.1300833110](#); pmid: [23836644](#)
20. J. R. Huddleston, Horizontal gene transfer in the human gastrointestinal tract: Potential spread of antibiotic resistance genes. *Infect. Drug Resist.* **7**, 167–176 (2014). doi: [10.2147/IDR.S48820](#); pmid: [25018641](#)
21. C. S. Smillie *et al.*, Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241–244 (2011). doi: [10.1038/nature10571](#); pmid: [22037308](#)
22. Human Microbiome Project Consortium, A framework for human microbiome research. *Nature* **486**, 215–221 (2012). doi: [10.1038/nature11209](#); pmid: [22699610](#)
23. S. Zhao, C. L. Dai, E. D. Evans, Z. Lu, E. J. Alm, Tracking strains predicts personal microbiomes and reveals recent adaptive evolution. *bioRxiv* 2020.2009.2014.296970 [Preprint] (2020); doi: [10.1101/2020.09.14.296970](#)
24. J. Jovel *et al.*, Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. *Front. Microbiol.* **7**, 459 (2016). doi: [10.3389/fmicb.2016.00459](#); pmid: [2748170](#)
25. H. Xie *et al.*, Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. *Cell Syst.* **3**, 572–584.e3 (2016). doi: [10.1016/j.cels.2016.10.004](#); pmid: [27818083](#)
26. I. L. Brito, E. J. Alm, Tracking Strains in the Microbiome: Insights from Metagenomics and Models. *Front. Microbiol.* **7**, 712 (2016). doi: [10.3389/fmicb.2016.00712](#); pmid: [27427733](#)
27. T. Van Rossum, P. Ferretti, O. M. Maistrenko, P. Bork, Diversity within species: Interpreting strains in microbiomes. *Nat. Rev. Microbiol.* **18**, 491–506 (2020). doi: [10.1038/s41579-020-0368-1](#); pmid: [32499497](#)
28. E. L. Moss, D. G. Maghini, A. S. Bhatt, Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat. Biotechnol.* **38**, 701–707 (2020). doi: [10.1038/s41587-020-0422-6](#); pmid: [32042169](#)
29. C. M. Singleton *et al.*, Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nat. Commun.* **12**, 2009 (2021). doi: [10.1038/s41467-021-22203-2](#); pmid: [33790294](#)
30. A. Bishara *et al.*, High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nat. Biotechnol.* **36**, 1067–1075 (2018). doi: [10.1038/nbt.4266](#); pmid: [30320765](#)
31. E. Yaffe, D. A. Relman, Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation. *Nat. Microbiol.* **5**, 343–353 (2020). doi: [10.1038/s41564-019-0625-0](#); pmid: [31873203](#)
32. T. Stalder, M. O. Press, S. Sullivan, I. Liachko, E. M. Top, Linking the resistome and plasmidome to the microbiome. *ISME J.* **13**, 2437–2446 (2019). doi: [10.1038/s41396-019-0446-4](#); pmid: [31147603](#)
33. A. Almeida *et al.*, A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019). doi: [10.1038/s41586-019-0965-1](#); pmid: [30745586](#)
34. H. P. Browne *et al.*, Culturing of ‘unculturable’ human microbiota reveals novel taxa and extensive sporulation. *Nature* **533**, 543–546 (2016). doi: [10.1038/nature17645](#); pmid: [27144353](#)
35. R. Chijiwa *et al.*, Single-cell genomics of uncultured bacteria reveals dietary fiber responders in the mouse gut microbiota. *Microbiome* **8**, 5 (2020). doi: [10.1186/s40168-019-0779-2](#); pmid: [31969191](#)
36. R. S. Lasken, Single-cell genomic sequencing using Multiple Displacement Amplification. *Curr. Opin. Microbiol.* **10**, 510–516 (2007). doi: [10.1016/j.mbs.2007.08.005](#); pmid: [17923430](#)
37. M. G. Pachiadaki *et al.*, Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell* **179**, 1623–1635.e1611 (2019). doi: [10.1016/j.cell.2019.11.017](#)
38. C. Rinke *et al.*, Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nat. Protoc.* **9**, 1038–1048 (2014). doi: [10.1038/nprot.2014.067](#); pmid: [24722403](#)
39. L. Xu, I. L. Brito, E. J. Alm, P. C. Blainey, Virtual microfluidics for digital quantification and single-cell sequencing. *Nat. Methods* **13**, 759–762 (2016). doi: [10.1038/nmeth.3955](#); pmid: [27479330](#)
40. F. B. Yu *et al.*, Microfluidic-based mini-metagenomics enables discovery of novel microbial lineages from complex environmental samples. *eLife* **6**, e26580 (2017). doi: [10.7554/eLife.26580](#); pmid: [28678007](#)
41. M. Džunková *et al.*, Defining the human gut host-phage network through single-cell viral tagging. *Nat. Microbiol.* **4**, 2192–2203 (2019). doi: [10.1038/s41564-019-0526-2](#); pmid: [31384000](#)
42. B. A. Berghuis *et al.*, Hydrogenotrophic methanogenesis in archaeal phylum *Verstraetearchaeota* reveals the shared ancestry of all methanogens. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 5037–5044 (2019). doi: [10.1073/pnas.1815631116](#); pmid: [30814220](#)
43. S.-Y. Teh, R. Lin, L.-H. Hung, A. P. Lee, Droplet microfluidics. *Lab Chip* **8**, 198–220 (2008). doi: [10.1039/b715524g](#); pmid: [18231657](#)
44. A. M. Klein *et al.*, Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015). doi: [10.1016/j.cell.2015.04.044](#); pmid: [26000487](#)
45. E. Z. Macosko *et al.*, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015). doi: [10.1016/j.cell.2015.05.002](#); pmid: [26000488](#)
46. M. Hosokawa, Y. Nishikawa, M. Kogawa, H. Takeyama, Massively parallel whole genome amplification for single-cell sequencing using droplet microfluidics. *Sci. Rep.* **7**, 5199 (2017). doi: [10.1038/s41598-017-05436-4](#); pmid: [28701744](#)
47. F. Lan, B. Demaree, N. Ahmed, A. R. Abate, Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nat. Biotechnol.* **35**, 640–646 (2017). doi: [10.1038/nbt.3880](#); pmid: [28553940](#)
48. K. Ahn, J. Agresti, H. Chong, M. Marquez, D. A. Weitz, Electrocoalescence of drops synchronized by size-dependent flow in microfluidic channels. *Appl. Phys. Lett.* **88**, 264105 (2006). doi: [10.1063/1.2218058](#)
49. L. Blanco *et al.*, Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J. Biol. Chem.* **264**, 8935–8940 (1989). doi: [10.1016/S0021-9258\(18\)81883-X](#); pmid: [2498321](#)
50. F. B. Dean, J. R. Nelson, T. L. Giesler, R. S. Lasken, Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* **11**, 1095–1099 (2001). doi: [10.1101/gr.180501](#); pmid: [11381035](#)
51. A. Adey *et al.*, Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* **11**, R119 (2010). doi: [10.1186/gb-2010-11-12-r119](#); pmid: [21143862](#)
52. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012). doi: [10.1038/nmeth.1923](#); pmid: [22388286](#)
53. A. Bankevich *et al.*, SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012). doi: [10.1089/cmb.2012.0021](#); pmid: [22506599](#)
54. B. D. Ondov *et al.*, Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016). doi: [10.1186/s13059-016-0997-x](#); pmid: [27323842](#)
55. C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, S. Aluru, High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018). doi: [10.1038/s41467-018-07641-9](#); pmid: [30504855](#)
56. D. H. Parks *et al.*, Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015). doi: [10.1101/gr.186072.114](#); pmid: [25977477](#)
57. A. Almeida *et al.*, A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021). doi: [10.1038/s41587-020-0603-3](#); pmid: [32690973](#)
58. E. Pasoli *et al.*, Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649–662.e20 (2019). doi: [10.1016/j.cell.2019.01.001](#); pmid: [30661755](#)
59. P.-A. Chaumelle, A. J. Mussig, P. Hugenoltz, D. H. Parks, GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019). doi: [10.1093/bioinformatics/btz2848](#); pmid: [31730192](#)
60. D. E. Wood, S. L. Salzberg, Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014). doi: [10.1186/gb-2014-15-3-r46](#); pmid: [24580807](#)
61. N. R. Garud, B. H. Good, O. Hallatschek, K. S. Pollard, Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS Biol.* **17**, e3000102 (2019). doi: [10.1371/journal.pbio.3000102](#); pmid: [30673701](#)
62. D. Albanese, C. Donati, Strain profiling and epidemiology of bacterial species from metagenomic sequencing. *Nat. Commun.* **8**, 2260 (2017). doi: [10.1038/s41467-017-02209-5](#); pmid: [29273717](#)
63. E. Becht *et al.*, Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2018). doi: [10.1038/nbt.4314](#); pmid: [30531897](#)
64. J. J. Faith *et al.*, The long-term stability of the human gut microbiota. *Science* **341**, 1237439 (2013). doi: [10.1126/science.1237439](#); pmid: [23828941](#)
65. J.-H. Hehmann *et al.*, Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**, 908–912 (2010). doi: [10.1038/nature08937](#); pmid: [20376150](#)
66. P. J. Keeling, J. D. Palmer, Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.* **9**, 605–618 (2008). doi: [10.1038/nrg2386](#); pmid: [18591983](#)
67. M. Groussin *et al.*, Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell* **184**, 2053–2067. e2018 (2021). doi: [10.1016/j.cell.2021.02.052](#)
68. E. Guerin *et al.*, Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host Microbe* **24**, 653–664.e6 (2018). doi: [10.1016/j.chom.2018.10.002](#); pmid: [30449316](#)
69. B. A. Siranosian, F. B. Tamburini, G. Sherlock, A. S. Bhatt, Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages. *Nat. Commun.* **11**, 280 (2020). doi: [10.1038/s41467-019-1403-3](#); pmid: [31941900](#)
70. S. D. Ganeshan, Z. Hosseindoust, Phage Therapy with a Focus on the Human Microbiota. *Antibiotics (Basel)* **8**, 131 (2019). doi: [10.3390/antibiotics8030131](#); pmid: [31461990](#)
71. T. D. S. Sutton, C. Hill, Gut Bacteriophage: Current Understanding and Challenges. *Front. Endocrinol.* **10**, 784 (2019). doi: [10.3389/fendo.2019.00784](#); pmid: [31849833](#)
72. A. N. Shkoporov *et al.*, ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects *Bacteroides intestinalis*. *Nat. Commun.* **9**, 4781 (2018). doi: [10.1038/s41467-018-07225-7](#); pmid: [30429469](#)
73. J. C. McDonald, G. M. Whitesides, Poly(dimethylsiloxane) as a material for fabricating microfluidic devices. *Acc. Chem. Res.* **35**, 491–499 (2002). doi: [10.1021/ar010101q](#); pmid: [12118988](#)
74. R. Zilionis *et al.*, Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.* **12**, 44–73 (2017). doi: [10.1038/nprot.2016.154](#); pmid: [27929523](#)

75. R. Ding, W. L. Ung, J. A. Heyman, D. A. Weitz, Sensitive and predictable separation of microfluidic droplets by size using in-line passive filter. *Biomicrofluidics* **11**, 014114 (2017). doi: [10.1063/1.4976723](https://doi.org/10.1063/1.4976723); pmid: [28344725](https://pubmed.ncbi.nlm.nih.gov/28344725/)
76. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014). doi: [10.1093/bioinformatics/btu170](https://doi.org/10.1093/bioinformatics/btu170); pmid: [24695404](https://pubmed.ncbi.nlm.nih.gov/24695404/)
77. H. Li et al., 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). doi: [10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352); pmid: [19505943](https://pubmed.ncbi.nlm.nih.gov/19505943/)
78. N. T. Pierce, L. Irber, T. Reiter, P. Brooks, C. T. Brown, Large-scale sequence comparisons with sourmash. *F1000 Res.* **8**, 1006 (2019). doi: [10.12688/f1000research.19675.1](https://doi.org/10.12688/f1000research.19675.1); pmid: [31508216](https://pubmed.ncbi.nlm.nih.gov/31508216/)
79. X. Jiang, A. B. Hall, R. J. Xavier, E. J. Alm, Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. *PLOS ONE* **14**, e0223680 (2019). doi: [10.1371/journal.pone.0223680](https://doi.org/10.1371/journal.pone.0223680); pmid: [31830054](https://pubmed.ncbi.nlm.nih.gov/31830054/)
80. C. Camacho et al., BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). doi: [10.1186/1471-2105-10-421](https://doi.org/10.1186/1471-2105-10-421); pmid: [20003500](https://pubmed.ncbi.nlm.nih.gov/20003500/)
81. A. M. Eren et al., Anvi'o: An advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015). doi: [10.7717/peerj.1319](https://doi.org/10.7717/peerj.1319); pmid: [26500826](https://pubmed.ncbi.nlm.nih.gov/26500826/)
82. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014). doi: [10.1093/bioinformatics/btu033](https://doi.org/10.1093/bioinformatics/btu033); pmid: [24451623](https://pubmed.ncbi.nlm.nih.gov/24451623/)
83. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021). doi: [10.1093/nar/gkab301](https://doi.org/10.1093/nar/gkab301); pmid: [33885785](https://pubmed.ncbi.nlm.nih.gov/33885785/)
84. D. E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019). doi: [10.1186/s13059-019-1891-0](https://doi.org/10.1186/s13059-019-1891-0); pmid: [31779668](https://pubmed.ncbi.nlm.nih.gov/31779668/)
85. T. Seemann, Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014). doi: [10.1093/bioinformatics/btu153](https://doi.org/10.1093/bioinformatics/btu153); pmid: [24642063](https://pubmed.ncbi.nlm.nih.gov/24642063/)
86. J. Huerta-Cepas et al., eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019). doi: [10.1093/nar/gky1085](https://doi.org/10.1093/nar/gky1085); pmid: [30418610](https://pubmed.ncbi.nlm.nih.gov/30418610/)
87. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012). doi: [10.1093/bioinformatics/bts565](https://doi.org/10.1093/bioinformatics/bts565); pmid: [23060610](https://pubmed.ncbi.nlm.nih.gov/23060610/)
88. S. J. Zhao, shijiezhao/Microbe-seq: Scripts for Microbe-seq, version 2.0, Zenodo (2022); https://zenodo.org/record/6467400#_YoLEeJMI2w.

ACKNOWLEDGMENTS

We thank members of the Weitz and Alm laboratories for helpful discussions and Y. Cai, W. Chen, Z. Cheng, N. Cui, L. Dai, R. Ding, P. Ellis, Z. Ge, J. Gong, H. Li, F. Ling, B. Liu, H. Liu, H. Pei, R. Rosenthal, J. Tang, Y. Wang, J. Xia, Y. Yao, X. Yu, Z. Zhang, Z. Zhang, and Z. Zhao for general discussions and comments on the manuscript. P.J.L. acknowledges support from the Massachusetts DTA through the SNAP and HIP programs. We thank OpenBiome for providing stool samples. We thank the MIT Center for Microbiome Informatics and Therapeutics and The Bauer Core Facility at Harvard University for providing sequencing services. **Funding:** This work was supported by the following: US Department of Energy, Office of Science, Office of Biological & Environmental Research, grant DE-AC02-05CH11231 at Lawrence Berkeley National Laboratory by ENIGMA-Ecosystems and Networks Integrated with Genes and Molecular Assemblies; The National Science Foundation grant DMR-1708729; The National Science Foundation grant, through the Harvard University Materials Research Science and Engineering Center, DMR-2011754; National Institutes of Health grant P01HL120839; National Institutes of Health grant R21AI125990; National Institutes of Health grant R21AI128623; National Institutes of Health grant R01AI153156; National Aeronautics and Space

Administration grant NNX13A4Q8G; National Aeronautics and Space Administration grant 80NSSC19K0598. **Author contributions:** W.Z., S.Z., H.Z., P.J.L., E.J.A., and D.A.W. conceived and designed the methodology; W.Z. developed and performed the experiments with assistance from S.Z.; S.Z., W.Z., Y.Y., D.M.N., and C.L.D. performed the data analysis with input from all authors; P.J.L., W.Z., and S.Z. wrote the initial manuscript; D.A.W., P.J.L., W.Z., S.Z., and E.J.A. revised the manuscript; all authors read and commented on the manuscript; E.J.A. and D.A.W. supervised the study. **Competing interests:** E.J.A. is affiliated with Finch Therapeutics and Biobot Analytics. All other authors declare no other competing interests. **Data and materials availability:** Combined fastq files for each stool sample, with read header containing the unique SAG ID and adaptor removed and filtered for quality, are available from NCBI Sequence Read Archive (Bioproject: PRJNA803937). Metagenomic fastq files are available from the previous publication (Bioproject: PRJNA544527). Commented scripts, intermediary data, and genome coassemblies are available at (88). **License information:** Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. <https://www.sciencemag.org/about/science-licenses-journal-article-reuse>

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abm1483

Figs. S1 to S13
Tables S1 to S8
MDAR Reproducibility Checklist
Movies S1 to S10
Data S1

[View/request a protocol for this paper from Bio-protocol.](#)

Submitted 28 August 2021; accepted 29 April 2022
10.1126/science.abm1483



High-throughput, single-microbe genomics with strain resolution, applied to a human gut microbiome

Wenshan Zheng, Shijie Zhao, Yehang Yin, Huidan Zhang, David M. Needham, Ethan D. Evans, Chengzhen L. Dai, Peter J. Lu, Eric J. Alm, and David A. Weitz

Science **376** (6597), eabm1483. DOI: 10.1126/science.abm1483

Strain specific single-cell sequencing

Single-cell methods are the state of the art in biological research. Zheng *et al.* developed a high-throughput technique called Microbe-seq designed to analyze single bacterial cells from a microbiota. Microbe-seq uses microfluidics to separate individual bacterial cells within droplets and then extract, amplify, and barcode their DNA, which is then subject to pooled Illumina sequencing. The technique was tested by sequencing multiple human fecal samples to generate barcoded reads for thousands of single amplified genomes (SAGs) per sample. Pooling the SAGs corresponding to the same bacterial species allowed consensus assemblies of these genomes to provide insights into strain-level diversity and revealed a phage association and the limits on horizontal gene-transfer events between strains. —CA

View the article online

<https://www.science.org/doi/10.1126/science.abm1483>

Permissions

<https://www.science.org/help/reprints-and-permissions>

Use of this article is subject to the [Terms of service](#)

Science (ISSN 1095-9203) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.

Copyright © 2022 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works