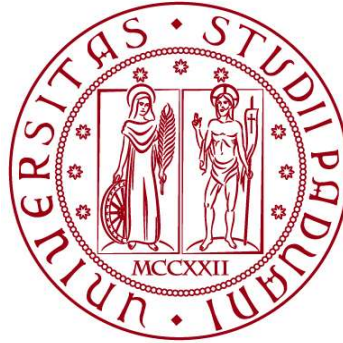


**UNIVERSITÀ DEGLI STUDI DI PADOVA**

**DIPARTIMENTO DI BIOLOGIA**

**Corso di Laurea in Biologia**



**ELABORATO DI LAUREA**

**NUOVI APPROCCI GENOMICI NELL'ANALISI DELLE  
INTERAZIONI VIRUS-PROCARIOTI NEI BIOFILM DELLE  
SORGENTI TERMALI**

**Tutor: Prof. Stefano Campanaro  
Dipartimento di Biologia**

**Co-tutor: Dott. Alessandro Rossi  
Dipartimento di Biologia**

**Laureando: Simone Carraro**

**ANNO ACCADEMICO 2021/2022**

## ABSTRACT

I procarioti sono i modelli di vita più antichi presenti sulla terra. La grande capacità di adattamento che possiedono permette loro di colonizzare ogni tipo di ambiente, anche quelli più estremi, come quelli caratterizzati dalle elevate temperature delle sorgenti termali analizzati in questo studio. In questi ambienti questi microorganismi secernono materiale organico, creando così un ambiente per molti aspetti isolato dall'esterno. I biofilm tendono ad essere multi-stratificati e accolgono al loro interno un complesso ecosistema di diverse specie interdipendenti. Date le condizioni presenti all'interno di questi biofilm si crea un ambiente ideale ad ospitare un gran numero di virus che infettano archaea e batteri. Lo studio qui proposto, con tecniche innovative come il sequenziamento a singola cellula, si pone come obiettivo una migliore comprensione delle interazioni che coinvolgono questi virus e i loro ospiti. I risultati ottenuti hanno evidenziato una chiara prevalenza del ciclo lisogenico. Inoltre, è emersa una compartimentalizzazione delle varie specie di virali delimitate dagli strati dei biofilm. Le tecniche di *single cell genomics* hanno permesso infine di individuare nuove specie virali e procariotiche desumendo anche i rapporti che li uniscono.

## INDICE

<b>1. INTRODUZIONE</b> .....	1
<b>1.1 I Biofilm</b> .....	1
<b>1.2 <i>Next Generation Sequencing</i></b> .....	2
<b>2. METODOLOGIE</b> .....	4
<b>2.1 Prelievo e Preparazione Campione</b> .....	4
<b>2.2 <i>Fluorescence Activated Cell Sorting</i></b> .....	5
<b>2.3 Lisi Cellulare</b> .....	6
<b>2.4 <i>Multiple Displacement Amplification</i></b> .....	7
<b>2.5 Illumina Nextseq</b> .....	8
<b>3. RISULTATI</b> .....	10
<b>3.1 <i>Piggyback the winner</i></b> .....	10
<b>3.2 Relazioni virus-procariote</b> .....	12
<b>4. CONCLUSIONE</b> .....	13
<b>5. BIBLIOGRAFIA</b> .....	15

## 1. INTRODUZIONE

### 1.1 I Biofilm

La produzione di sostanze adesive per aderire al suolo e per protezione da ambienti ostili o da predatori è un adattamento che si presume sia evoluto diversi milioni di anni fa. Questi tappeti microbici sono strutture bentoniche multi-stratificate ed ospitano al loro interno centinaia di specie di procarioti. Questa stratificazione riflette le diverse specializzazioni e adattamenti delle singole specie: in particolare, negli strati superiori, dove l'esposizione alla luce solare e all'atmosfera è maggiore, sono localizzati procarioti fotosintetici come i cianobatteri, mentre negli strati più interni sono presenti specie anaerobiche. L'abbondanza e la varietà di individui crea un ambiente dove le relazioni interspecie e intraspecie sono estremamente complesse e dove eventi di competizione per le risorse, commensalismo e simbiosi si mescolano.

In particolare, i biofilm di sorgenti termali presentano alcune caratteristiche tipiche che li differenziano da altri ecosistemi. L'elevata presenza di solfuri e temperature generalmente tra i 50°C e 90°C rendono l'ambiente particolarmente ostile alla vita, perciò la diversità di specie è notevolmente ridotta rispetto agli ambienti mesofili. Tra le specie di fondamentale importanza, oltre ai già citati cianobatteri termofili, ci sono i batteri anossici *Chloroflexi* e *Chlorobium*. Questo restringimento delle varietà di specie rende i biofilm delle sorgenti termali un ottimo modello di studio perché facilita notevolmente la comprensione della presenza e delle interazioni dei vari organismi presenti al loro interno. Un'altra caratteristica che distingue questi biofilm è la relativa stabilità dovuta all'assenza di variazioni dovute al cambiare delle stagioni.

La presenza di virus all'interno di biofilm è ben nota, tuttavia, questo è ancora un argomento che presenta diverse lacune nella ricerca scientifica. È certo che questi microrganismi hanno un ruolo fondamentale nel mantenere stabile la popolazione di procarioti. Si ritiene anche che i batteriofagi possano avere un ruolo nel ciclo del carbonio e di altri nutrienti. L'elevata densità di questi microrganismi aumenta gli eventi di trasduzione batterica, favorendo così lo scambio genetico orizzontale tra i procarioti. L'evoluzione di virus e procarioti è strettamente correlata in un rapporto di coevoluzione. Studi precedenti hanno avuto scarsi risultati nell'identificare i taxa di appartenenza dei virus che colonizzano i biofilm. Ancora più complesso è stato dimostrare i rapporti di dipendenza con le precise specie di batteri o archei. Questo ha reso difficile la comprensione degli adattamenti adottati dai virus nei tappeti microbici. Per esempio, non è conosciuto il *range* di infezione dei virus, cioè se in genere sono altamente specializzati infettando poche specie di procarioti o se invece infettano taxa molto diversi tra loro. Queste limitazioni nelle conoscenze delle relazioni virus-ospite sono dovute alla complessità dei sistemi ecologici dei biofilm. Per questo motivo la ricerca storicamente si è limitata su colture in laboratorio, che tuttavia necessitano di una semplificazione dei modelli di studio, ignorando così parte dei rapporti e delle dipendenze realmente presenti in natura. Ciò è dovuto al fatto che molti microrganismi difficilmente possono essere coltivati in laboratorio.

In genere, alcune specie virali possono essere litiche o temperate a seconda dell'ambiente in cui si trovano. Nel primo caso un'infezione ha come conseguenza l'immediata replicazione del genoma virale e la sintesi dei suoi geni formando nuovi virioni che sono rilasciati nell'ambiente tramite lisi cellulare. Un virus temperato invece intraprende un ciclo lisogenico dove il genoma viene integrato nel DNA della cellula ospite, prendendo il nome di provirus; quindi, il provirus silenzia gran parte dei suoi geni per riattivarli in seguito a stimoli esterni. Dalla comparsa dell'ecologia virale nel 1980 fino agli ultimi anni la teoria principale riguardante il ciclo di infezione predominante era quella che favoriva il ciclo litico rispetto a quello lisogenico. In particolare, si riteneva che il ciclo litico fosse diffuso specialmente in ambienti ricchi di possibili ospiti in cui la replicazione avveniva immediatamente, mentre l'integrazione del genoma virale nei procarioti era ritenuto un meccanismo di difesa in ambienti ostili e poveri di organismi da infettare. La teoria *Kill-the-Winner* (KtW) inoltre stabilisce che, con un modello simile a quello osservato in alcune specie predatorie e parassitiche, il virus predilige l'infezione e la riproduzione diretta di ceppi di ospiti con un elevato fitness, comportandone così il declino. La variante dell'ospite che quindi prende il posto della prima in termini di tasso di crescita diventerà a sua volta l'ospite prediletto; in questo modo i vari ceppi della specie procariotiche si alternano in un andamento fluttuante nel tempo. Recentemente si è sviluppata una teoria detta *Piggyback-the-Winner* (PtW) che sostiene, tramite analisi metagenomiche (Figura 1) integrate con il conteggio delle popolazioni sia virali che degli ospiti e con studi sul tasso di crescita microbica, la predominanza del ciclo lisogenico, specialmente in ambienti ricchi di nutrienti e di microrganismi come lo sono alcuni biofilm. Viceversa, il ciclo litico, secondo PtW, è favorito in ambienti poveri di nutrienti e con microbi interspersi.

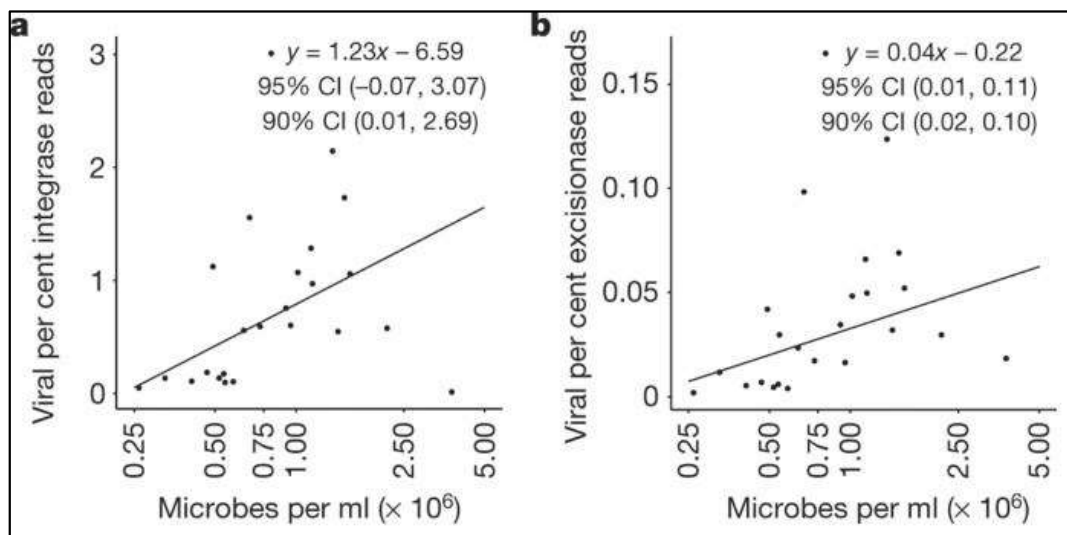


Figura 1: Correlazione positiva tra densità microbica e reads di integrasi virale (a) e reads di escissionasi virale (b). (Knowles et al., 2016, modificato)

## 1.2 Next Generation Sequencing

Nello studio della microbiologia ambientale le tecniche di *next generation sequencing* (NGS) usate nell'analisi e sequenziamento sono principalmente tre:

1. Quella storicamente più usata data la sua semplicità e lo scarso costo è il sequenziamento di ampliconi. Questo approccio prende come unico elemento di analisi un gene target, in genere l'rRNA ribosomiale 16S, che codifica una regione della subunità 30S dei ribosomi procariotici, permettendo così di distinguere l'organismo da una popolazione eterogenea. L'amplificazione del gene avviene con PCR. I prodotti sono sequenziati con tecniche di NGS e, con l'uso di database come SILVA, RDP o Greengenes si possono confrontare le sequenze ottenute con quelle già conosciute. Data l'estrema selettività di questa metodologia, essa non offre dati sufficienti per risalire né al ceppo né alla specie di appartenenza. Inoltre, non è possibile ottenere informazioni sulle vie metaboliche e i ruoli svolti dai microorganismi. Infine, non è possibile ottenere informazioni sulle popolazioni di virus. Queste limitazioni hanno reso necessario lo sviluppo di tecniche più complesse che permettono l'analisi dell'intero genoma microbico.
2. Nel sequenziamento shotgun metagenomico il DNA viene estratto e frammentato casualmente. Questi frammenti poi vengono sequenziati indipendentemente in *reads* e riassembleti nella fase di assembly. I programmi di assembly non generano un filamento completo del genoma microbico, ma piuttosto organizzano le sequenze in contigui di sequenza e in scaffolds separati da gaps. Il *binning* quindi organizza i dati dell'assembly in clusters che identificano il singolo genoma, separandolo così dalla comunità microbica. A seconda di parametri prestabiliti, tra cui completezza e contaminazione, se i *bins* sono di elevata qualità vengono selezionati come *metagenome assembled genomes* (MAG) utili al passo successivo, l'annotazione genica. Sono diversi gli approcci dell'annotazione, tra cui tools che ricercano similarità delle sequenze con programmi basati su BLAST. BLAST allinea le sequenze metagenomiche con sequenze simili contenute in database come KEGG. Questo tool, tuttavia, può risultare inadatto ad elaborare le elevate quantità di geni presenti nei MAGs. L'insieme dei geni sequenziati non è solo utile a identificare la specie di appartenenza, ma anche a identificare i pathway metabolici. In questo modo la metagenomica permette di ottenere informazioni riguardanti i vari taxa presenti nella comunità ed anche le funzioni biologiche e i ruoli svolti dagli individui nell'ecosistema. Tuttavia, pur risolvendo parte delle mancanze del sequenziamento di ampliconi, offrendo un'analisi più raffinata e precisa, la metagenomica shotgun presenta essa stessa delle limitazioni. Sequenze altamente ripetute rendono complesso il processo di *assembly* poiché i reads che li rappresentano possono provenire da regioni differenti del genoma. Questo ha come conseguenza la formazione di gaps, ambiguità ed errori nei risultati prodotti. In comunità complesse, inoltre, le grandi quantità di dati prodotti generano *bottlenecks*, perciò sono richiesti computer con un'elevata potenza di elaborazione. Anche il coverage non uniforme che

generalmente si sviluppa con la metagenomica shotgun influisce negativamente nel processo di *assembly*. Alcune soluzioni proposte a correggere errori e migliorare la qualità dei risultati, come RiboFR-Seq (collega profili 16S rRNA a *contig* metagenomici) e epicPCR (collega geni funzionali a sequenze con valore filogenetico), risolvono solo in parte la difficoltà di associazione: non tutti i geni funzionali di un procariote, infatti, possono essere ricondotti al phylum di origine.

3. L'ultima tecnica di sequenziamento utilizzata in microbiologia ambientale, presa in studio in questo elaborato, è il sequenziamento a singola cellula. Questo approccio allo studio di campioni microbiologici naturali permette di ottenere una migliore analisi delle attività metaboliche rispetto alla metagenomica shotgun ma produce anche dati utili alla comprensione delle interazioni virus-procarioti e interazioni cellula-cellula. È inoltre particolarmente efficiente nell'ottenere sequenze genomiche da specie presenti in scarse quantità nel campione di origine. Infine, l'affidabilità nella distribuzione filogenetica dedotta con la *single cell genomics* è decisamente migliore rispetto alle tecniche precedenti. Tuttavia, possono emergere degli errori come conseguenza di formazione di chimere nella fase di amplificazione, cioè segmenti di DNA che originano da due geni differenti. Questo riarrangiamento genetico si sviluppa con un tasso di una chimera su 10 kb rendendo sostanzialmente più difficile il successivo assemblaggio delle sequenze. Inoltre, il coverage, cioè il numero di *reads* che contengono un particolare nucleotide, può non essere distribuito uniformemente generando così *bias* nei risultati. Ciò è dovuto al priming casuale durante la reazione di *multiple displacement amplification* (MDA). L'utilizzo di primers casuali aumenta anche il rischio di contaminazione amplificando materiale genetico esogeno. Per questo motivo l'attrezzatura usata per l'isolamento cellulare è trattata con candeggina e pulita con acqua UV-c. Quindi, dati i vantaggi e svantaggi presenti nelle tecniche descritte sopra, queste spesso vengono applicate in concomitanza per ottenere risultati più comprensivi.

## 2. METODOLOGIE

### 2.1 Prelievo e Preparazione Campione

Il campione di studio è stato prelevato dalla sorgente termale *Cone Pool* (*Long Valley Caldera*, California, USA) attraverso una cannucchia sterile che ha permesso di raccogliere una sezione verticale completa del tappeto microbico. I 9 strati sono stati suddivisi in base a consistenza e colore. Tra questi sono stati scelti per studi successivi quelli con una più abbondante popolazione microbica non conosciuta. Per far ciò, campioni degli strati sono stati lisati e sottoposti ad analisi sequenziamento 16S rRNA. Quest'approccio permette una semplice ma efficace identificazione del phylum di appartenenza dato che questo particolare frammento di DNA è organizzato in regioni conservate, necessarie per il corretto *folding* della molecola, alternate a regioni ipervariabili. Sono queste ultime le sequenze che permettono di classificare gli organismi. Anche la scarsa probabilità

di scambio genico orizzontale caratteristica di questo gene rende il 16S rRNA un ottimo *marker*.

## **2.2 Fluorescence Activated Cell Sorting**

Il campione dello strato C è stato sottoposto a isolamento e sequenziamento a singola cellula. Il metodo più comune è quello usato in questo studio di isolamento è il *fluorescence-activated cell sorting* (FACS). Questa tecnica è un'espansione della citometria di flusso e consiste nell'isolare le cellule tramite *marker* fluorescenti. Il campione di studio è inserito in una soluzione salina contenente al suo interno specifici anticorpi affini a proteine di membrana. Ogni anticorpo, a sua volta, presenta un fluoroforo legato alle sue catene pesanti nelle regioni costanti e opposte al sito di legame dell'antigene. Per ogni tipologia di anticorpo presente nella soluzione è legato uno specifico fluoroforo in modo tale che il segnale di fluorescenza emesso dallo stesso è un indicatore preciso dell'espressione della proteina sulla cellula. Un citometro di flusso è costituito da un ugello, un generatore di gocce, da piastre deflettori, da una sorgente laser localizzata nel punto di interrogazione e dei rilevatori di dispersione frontale (FSC: *Forward scatter*) e di dispersione laterale (SSC: *Side scatter*), a cui si aggiungono, nel metodo FACS, una serie di rilevatori di fluorescenza (Figura 2). Il processo comincia all'interno dell'ugello dove sono presenti due fluidi. Il primo, detto fluido guaina, permette lo scorrimento della soluzione da analizzare. In questo modo le cellule vengono canalizzate in un sottile condotto e disposte in fila. Passando attraverso il raggio laser le cellule vengono analizzate per diversi caratteri. Il segnale FSC prosegue nella stessa direzione del laser e tramite un canale è raccolto da un rilevatore. I parametri FSC permettono di interpretare le dimensioni della cellula: un corpo di grosse dimensioni disperde più luce di un corpo di piccole dimensioni e quindi avrà un segnale FSC più forte. L'SSC, invece, è la luce rifratta dalla cellula a 90° rispetto la direzione del laser. Il *side scatter* è influenzato dalla granulosità e dalla complessità della cellula: una cellula povera di organelli produrrà un SSC debole rispetto a cellule più complesse, sia a livello membranoso che intracellulare. Infine, attraversando il punto di interrogazione, il fluoroforo viene eccitato dal laser emettendo luce con una particolare lunghezza d'onda unica per quell'anticorpo. Questa luce, attraversando filtri dicroici, viene direzionata verso rilevatori di fluorescenza. I dati raccolti dai vari rilevatori sono elaborati indipendentemente e possono essere plottati in istogrammi unidimensionali. Combinando più parametri insieme si possono ricavare istogrammi bidimensionali oppure grafici di densità. Dopo l'esposizione al laser, il particellato raggiunge il generatore di gocce dove la miscela viene separata e ogni goccia formata grazie a vibrazioni contiene un'unica cellula. Un anello di ricarica elettrica carica le cellule a seconda delle impostazioni dell'operatore e dal segnale di fluorescenza rilevato in precedenza. Quindi gli elettrodi posti a valle separano le varie gocce, a seconda della carica acquisita, verso delle piastre contenenti dai 94 a 384 pozzetti, ognuno dei quali conterrà con elevata precisione una singola cellula.

Il vantaggio di questo approccio rispetto ad altri sistemi, per esempio quelli optofluidici o microfluidici, è la facilità di operazione delle interfacce, lo scarso costo economico e infine la disponibilità di efficienti strumenti di visualizzazione



dei dati ottenuti. Tuttavia, il sistema FACS ha alcune limitazioni. Principalmente la necessità di avere anticorpi affini a specifiche proteine cellulari può portare alla mancata individuazione di specie non conosciute. Un'altra limitazione è dovuta all'alto tasso di falsi positivi che possono generarsi nel caso di ricerche su specie cellulari presenti in basse quantità nel campione. Infine, FACS necessita di grandi quantità di materiale prelevato. Tuttavia, quest'ultima limitazione nell'articolo preso in studio non è rilevante data l'abbondante disponibilità di campione reperibile. Possibili errori possono riscontrarsi nei risultati anche per la capacità intrinseca delle cellule, data la presenza di molecole come il NADPH, di eccitarsi ad esposizioni laser e quindi emettere segnali di fluorescenza. Misurazioni di campioni controllo sono sufficienti per eliminare *bias* in questo tipo dai risultati.

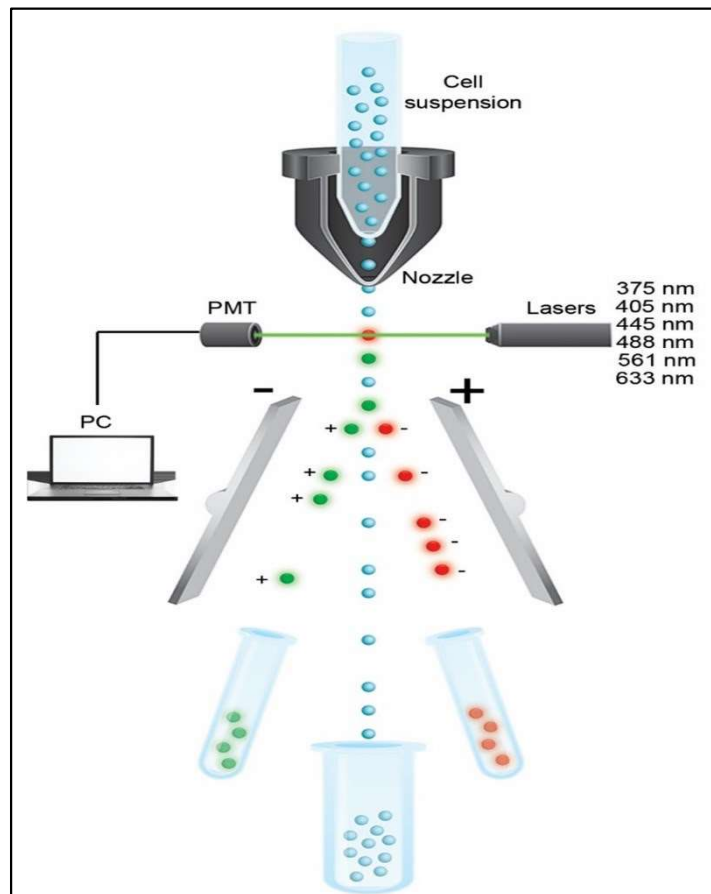


Figura 2: Schema della struttura di un citometro a fluorescenza

### 2.3 Lisi Cellulare

Per l'estrazione del DNA le cellule sono state lisate tramite il metodo congelamento-scongelo e la lisi chimica alcalina. La prima consiste in più cicli di congelamento e scongelamento del campione in sospensioni di etanolo. L'espansione della cellula durante il congelamento e la formazione di cristalli di ghiaccio al suo interno permettono la disintegrazione della membrana durante il restringimento della cellula quando questa è nuovamente esposta alla temperatura di 37°C. La lisi alcalina richiede invece ioni  $\text{OH}^-$  che reagiscono con i legami esteri tra glicerolo e acidi grassi rompendoli. Per questo motivo vengono usate delle forti basi come KOH o NaOH. È anche possibile l'utilizzo di enzimi litici

che, non necessitando di forze meccaniche eccessive o di pH elevati, più difficilmente degradano il DNA. Data la presenza di una moltitudine di specie nelle comunità microbiche naturali, sia gram negativi che positivi, può essere necessaria l'applicazione di una serie di agenti litici contemporaneamente.

#### 2.4 Multiple Displacement Amplification

Il materiale genetico raccolto è stato amplificato tramite la tecnica WGA-X, cioè un approccio di *Whole Genome Amplification* che non si basa su PCR ma su *Multiple Displacement Amplification* (MDA) (Figura 3). L'intero processo di amplificazione avviene alla temperatura costante di 30°C. L'enzima utilizzato è una polimerasi Phi29 che presenta una capacità di replicazione ad alta fedeltà data l'ottima capacità di *proofreading* e di attività esonucleasica 5'-3', quest'ultima capace di rimuovere primer o filamenti posizionati all'estremità 5'. La capacità di questa polimerasi di sintetizzare filamenti lunghi fino a 100kb è fondamentale per l'amplificazione di un intero genoma. I primer esamerici hanno una sequenza randomica, quindi il legame ai filamenti di DNA non avviene su sequenze precise ma piuttosto in multiple regioni del genoma differenti tra loro. Il filamento neosintetizzato non si appaia al DNA stampo ma invece è dislocato dall'enzima Phi29.

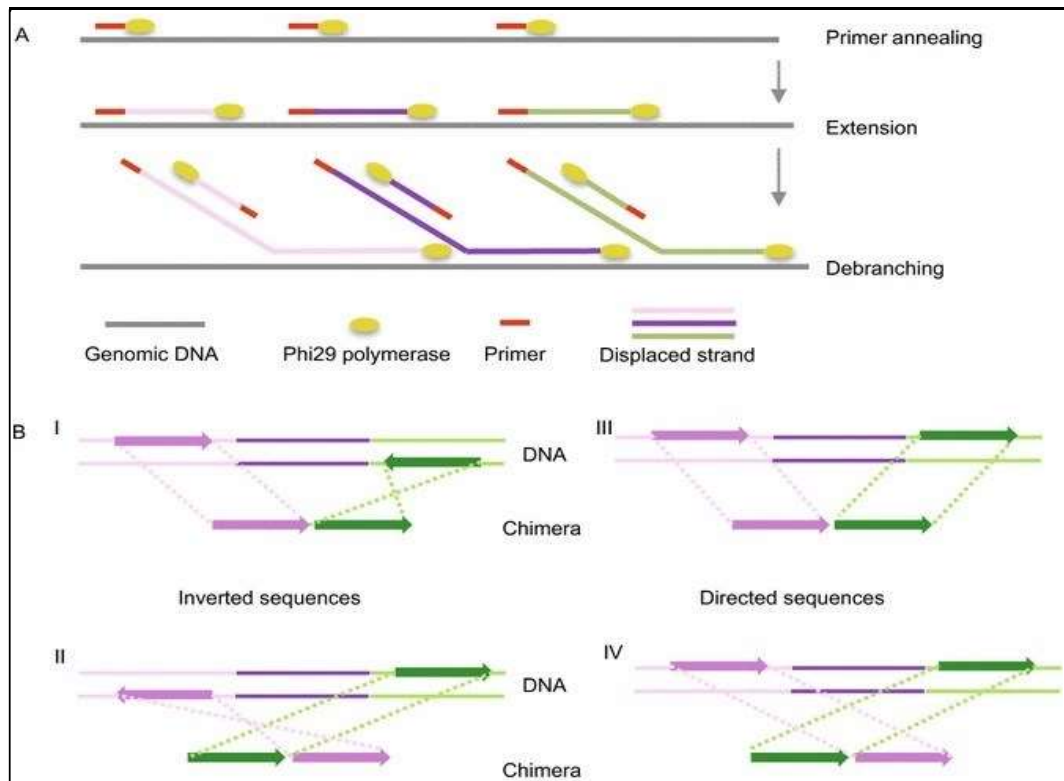


Figura 3: Illustrazione delle fasi di MDA (a) ed esempi di formazioni di chimere (b). (Xu and Zhao, 2018)

A sua volta il filamento template diventa valido target per un nuovo primer esamerico, permettendo così a una polimerasi di legarsi. In questo modo l'amplificazione del DNA è esponenziale e si sviluppa una struttura multi-ramificata. Un vantaggio di questa tecnica è la capacità della polimerasi Phi29 di

risolvere strutture secondarie nel DNA stampo e quindi di produrre un filamento templato continuo. Al contrario PCR, se effettuata con DNA organizzato in strutture secondarie, produce filamenti brevi o incompleti.

## 2.5 Illumina Nextseq

Il passo successivo consiste nel *deep sequencing* con la piattaforma Illumina Nextseq. È una tipologia di NGS che sequenzia regioni di DNA ripetutamente, dato l'alto livello di coverage, in modo tale da rilevare specie cellulari presenti anche con abbondanze relativamente ridotte. Il workflow è distribuito in quattro tappe:

1. Preparazione libreria: Il DNA viene frammentato con enzimi nucleasici in regioni casuali. I frammenti a loro volta sono legati a adattatori alle estremità 3' e 5'. Gli adattatori sono organizzati in tre regioni con funzioni differenti: la regione esterna è affine agli oligonucleotidi del *flow cell*, la regione centrale è un indicatore univoco dell'appartenenza del frammento al campione d'origine e infine la regione interna costituisce il sito di legame dei primers di sequenziamento. I prodotti ottenuti sono quindi amplificati con PCR.
2. Formazione dei *cluster*: I frammenti di DNA sono distribuiti in un *flow cell* e qui si legano a oligonucleotidi affini all'adattatore. Sono presenti due tipi di oligonucleotidi, ciascuno affine all'estremità 5' o 3'. Inizialmente l'ibridazione avviene solamente all'estremità 3'. I frammenti vengono amplificati in ambiente isotermico. Una polimerasi sintetizza il filamento complementare; il doppio filamento viene quindi denaturato e, dei due, quello originale viene eliminato. L'estremità 5' del filamento neosintetizzato si ibridizza all'oligonucleotide finora libero formando così una struttura a ponte. La polimerasi, quindi, sintetizza un nuovo filamento complementare. Una nuova denaturazione libera due filamenti singoli di DNA, ognuno con l'estremità opposta legata all'oligonucleotide. Questo processo avviene ripetutamente generando dei *cluster*. Ad amplificazione completata i filamenti inversi sono tagliati ed eliminati. Il DNA rimanente, dopo aver bloccato l'estremità 3' con opportuni oligonucleotidi per evitare priming indesiderati, rappresenta il template per il sequenziamento.
3. Sequenziamento: questo processo avviene per "sequenziamento di sintesi" perciò nel *flow cell* vengono inserite polimerasi e nucleotidi fluorescenti. A ogni ciclo è aggiunta una base che venendo eccitata emette una luce. La lunghezza d'onda e l'intensità dell'emissione identificano univocamente la base. Il ciclo viene ripetuto quanto è necessario per produrre *reads* della lunghezza desiderata, fino a un massimo di 300 nucleotidi. A processo terminato, il filamento neosintetizzato è eliminato. In caso di sequenziamento *pair-end* si procede con un nuovo evento di *bridge amplification* per ottenere un filamento complementare. Il DNA stampo è eliminato e si ripete il sequenziamento di sintesi.

4. Allineamento e analisi dati: Le sequenze ottenute sono prima di tutto filtrate con programmi come BBTools che eliminano o riducono (*trimming*) *reads* da regioni di bassa qualità. Infatti, nel sequenziamento per sintesi gli errori tendono ad accumularsi verso l'estremità 3'. Quindi vengono separate a seconda dell'indice assegnato durante la preparazione della libreria. Con l'aiuto di software, SPAdes nell'articolo preso in studio, le sequenze ottenute vengono allineate, appaiando sequenze che condividono porzioni di sequenze simili o identiche. SPAdes è particolarmente adatto alla manipolazione di dati ottenuti con *single cell sequencing* perché permette di risolvere due dei problemi che si sviluppano durante l'amplificazione (MDA), cioè la formazione di chimere e la formazione di coverage delle *reads* non uniforme. Questo programma non studia le sequenze nella loro interezza come una singola unità ma invece li suddivide in *k-mers*, che assumono il ruolo di *edges*, e in prefissi e suffissi *k-1*, dove sono eliminate rispettivamente l'ultima e la prima base. Questi prendono il nome di *nodes*. La lunghezza dei frammenti, cioè il *k*, può essere determinata manualmente fino a un massimo di 128pb oppure può essere determinata automaticamente dal programma in proporzione alla lunghezza dei *reads*. L'assemblaggio di questi *k-mers* si basa su grafici di De Bruijn (Figura 4) che, al contrario dell'*hamiltonian cycle* (più difficile da risolvere), collega gli *edges* ai *nodes* in modo univoco. Quindi, i *k-mers* vengono connessi quando il suffisso di uno combacia con il prefisso di un altro. La lunghezza dei *k-mers* influenza la qualità del grafico. Brevi *k-mers* generano grafici complessi mentre *k-mers* di dimensioni maggiori

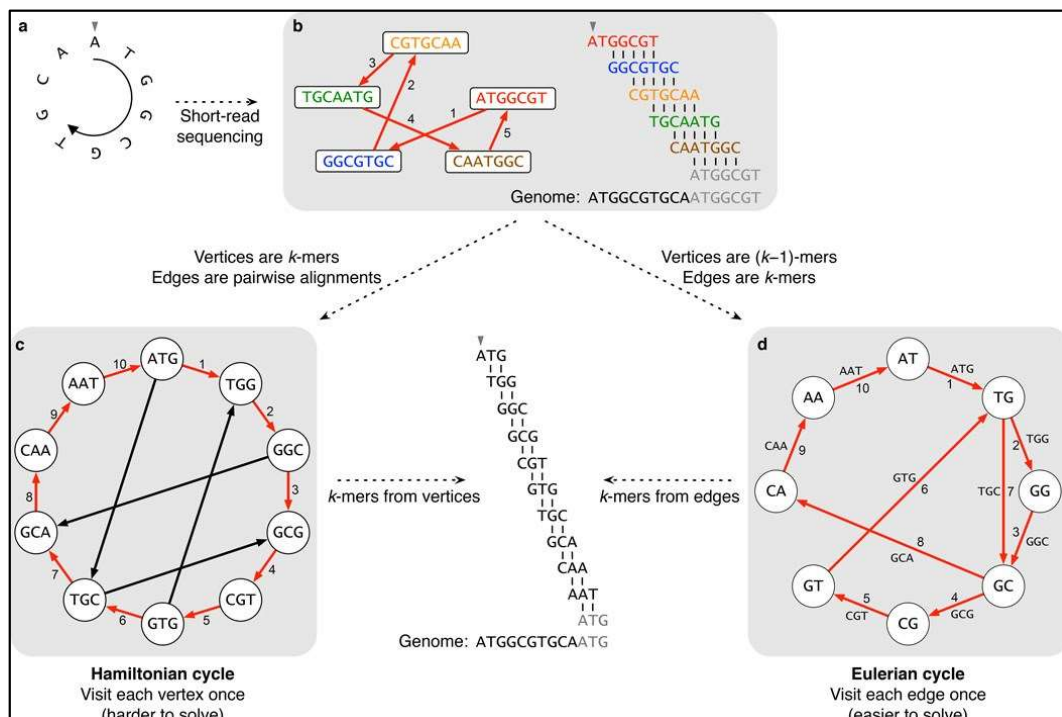


Figura 4: Esempio di un genoma circolare (a) ricostruito con il metodo Sanger (b), il metodo Hamiltoniano (c) e quello Euleriano (d) usato nel grafico de Bruijn. (Compeau et al., 2011)

generano grafici frammentati. Dato che MDA produce coverages non omogenei, SPAdes può usare grafici *multi-sized* che permettono l'uso di *k-mers* di varie dimensioni. Il grafico che si viene a generare sarà quindi il più adatto a trovare gli allineamenti corretti. Quindi l'*output* di SPAdes, derivato dall'assemblaggio dei *reads*, consiste in *contig* che a loro volta servono a costruire sequenze più comprensive del genoma (*scaffolds*).

### 3. RISULTATI

#### 3.1 *Piggyback the winner*

I nuovi metodi usati dai ricercatori per approcciarsi ai problemi insiti nello studio delle comunità microbiche, principalmente la *single cell sequencing*, hanno permesso di approfondire le attuali conoscenze sui biofilm. Di fondamentale importanza è il riscontro nel campione dei principi appartenenti alla teoria *Piggyback the Winner* proposta da Knowles et al. (2016). PtW sovverte le concezioni comuni sul ciclo vitale predominante dei virus batteriofagi, cioè che in ambienti abbondanti di ospiti ci sia una forte replicazione attiva virale. Infatti, questa teoria stabilisce che in ambienti ricchi di ospiti, come lo sono i biofilm, i virus sono temperati e si integrano nel DNA dell'ospite come provirus. Inversamente, il ciclo litico è maggioritario in ecosistemi poveri di batteri. Inoltre, questa teoria è correlata al *Bacteriophage Adherence to Mucus* (BAM) cioè la capacità di alcuni virus di legarsi alla matrice dei biofilm come adattamento per aumentare la possibilità di incontro virus-ospite, ed ha come conseguenza la riduzione della motilità virale e l'adozione di un'infezione litica. Nel suo insieme, quindi, PTW è dipendente dallo spazio e dalla distribuzione dei vari organismi nel biofilm. Generalmente gli strati superficiali sono più abbondanti di risorse nutritive necessarie allo sviluppo delle colonie batteriche e l'esposizione all'ossigeno favorisce lo sviluppo di popolazioni aerobiche. In contrasto, le regioni inferiori dei biofilm tendono ad avere un accrescimento microbico limitato dovuto alla scarsità di risorse. Quindi nelle regioni superficiali i virus sono integrati nel genoma dell'ospite data la maggiore motilità (la matrice del biofilm è meno densa) e la maggiore disponibilità di ospiti. Nelle regioni più profonde del biofilm invece i virus tendono a adottare preferibilmente il ciclo litico. I ricercatori dell'articolo hanno stabilito la generale prevalenza dei virus temperati tramite l'applicazione di diversi metodi, tra i quali lo studio della correlazione tra il coverage genomico virale e quello del suo ospite. Si suppone che se è favorito il ciclo litico nel campione allora il numero di materiale genetico virale sia nettamente superiore rispetto a quello batterico, sia per l'attiva replicazione virale all'interno dell'ospite, sia per l'alto numero di progenie che si sviluppa. Viceversa, proporzioni simili sono indicative di un'integrazione del DNA virale nel procariote. Se invece è il coverage dell'ospite a prevalere è possibile che l'infezione virale è limitata a particolari ceppi della specie microbica, oppure l'infezione rilevata può essere anomala. Anche cellule poliploidi, cioè con più copie di cromosomi omologhi, possono portare ad un coverage dell'ospite più abbondante. Tuttavia, questa caratteristica è tipica di cellule di grandi dimensioni che non sono state analizzate durante FACS, perciò si può assumere che la poliploidia non ha influito sul rapporto tra coverage dell'ospite e del virus.

Dei 35 accoppiamenti ospite-virus usati come riferimento, distribuiti in più strati, è stata rilevata una leggera abbondanza del coverage procariotico rispetto a quello virale (Figura 5).

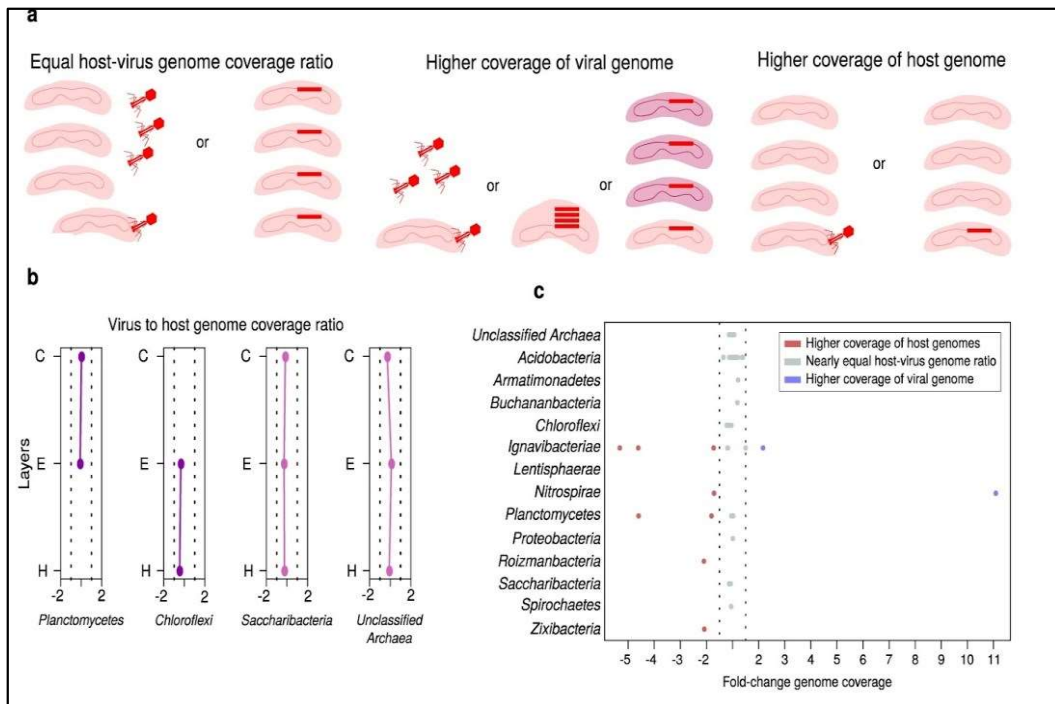


Figura 5: Illustrazioni dei possibili rapporti tra coverage virale e quello dell'ospite (a). Grafici del rapporto tra genoma virale e dell'ospite (b) (c). (Fonte: Jarett et al., 2020, modificato)

In generale, il 76% degli accoppiamenti dimostra un rapporto tra genoma virale e batterico molto simile. È possibile che all'interno di queste popolazioni avvengano eventi litici, tuttavia, dati i risultati riscontrati, questi sono casi isolati. Quindi, la formazione di profagi, in certe condizioni, offre aspetti vantaggiosi per entrambe le parti. Per i microbi l'infezione virale offre un incremento della fitness dell'individuo rispetto ad altri individui della stessa specie privi di profagi. La protezione da superinfezioni ne è un esempio: la presenza del profago, infatti, rende il microbo resistente a infezioni successive da parte di virus della stessa famiglia. Lo svantaggio che gli ospiti incorrono avendo un dispendio energetico maggiore data l'integrazione del genoma virale, che quindi porta alla trascrizione e mantenimento di nuovi geni, è trascurabile in un ambiente dove i nutrienti non sono un fattore limitante. Per i virus, invece, la difficoltà di diffusione e la bassa probabilità di trovare nuovi ospiti rende vantaggioso il ciclo lisogenico. Inoltre, al contrario di KtW dove le varianti con più successo vengono eliminate, i virus sfruttano la diffusione e l'alto tasso di crescita per aumentare la loro stessa densità di popolazione.

Riguardo la distribuzione spaziale il campione di biofilm dimostra un'elevata separazione degli organismi ospitati al suo interno. Infatti, dei 35 accoppiamenti virus-ospite, solo 4 erano presenti in più strati. Nei 323 contigs virali di cui non si è riusciti a rilevare l'ospite è stato riscontrato un simile risultato, con il 75% distribuito in un solo strato. Come ci si aspettava, dei tre strati presi in analisi,

quello più superficiale ha mostrato una diversità maggiore della popolazione virale, e ciò può essere spiegato dalla maggiore esposizione all'esterno dello stesso, e quindi maggiormente soggetto a colonizzazione da parte di nuovi virus. L'elevata divisione delle popolazioni virali, suddivise nei vari strati, esclude la diffusione passiva dovuta alla gravità come fattore di migrazione. Inoltre, l'assenza di una distribuzione virale uniforme su più strati può essere interpretata come un'indicazione del limitato range di infezione.

### 3.2 Relazioni virus-procariote

Lo studio si è posto principalmente come obiettivo il miglioramento delle conoscenze delle varie specie che abitano i biofilm e le strette relazioni che le connettono. A questo scopo l'analisi completa di tutti gli strati del campione tramite il sequenziamento del gene 16S rRNA svolta sia in una ricerca precedente (per lo strato più superficiale) che in quella qui descritta ha evidenziato la stratificazione delle comunità microbiche. Nello strato superiore sono molto diffusi, data l'esposizione ad un ambiente estremo, organismi termofili. In particolare, sono stati rilevati in abbondanza (1-10%) *Operational Taxonomic Units* (OTU) vicini alla specie di cianobatteri *Gloeomargarita lithophora*. Inoltre, sono presenti in minore abbondanza planctomiceti del genere *Isosphaera* e *Synechococcus*. Lo strato immediatamente inferiore mostra caratteristiche simili. Gli strati successivi, invece, ospitano popolazioni di phyla candidati, cioè organismi mai coltivati in vitro e perciò poco conosciuti. Questi strati sono stati sottoposti ad ulteriore analisi. In particolare, lo strato C, E, H sono stati selezionati per analisi di metagenomica shotgun e ancora lo strato C per il sequenziamento a livello di singola cellula già descritto sopra. I 130 *Single cell Amplified Genomes* (SAGs) e gli 88 *Metagenome Assembled Genomes* (MAGs) che hanno superato i requisiti minimi di qualità sono stati sottoposti ad *Average Nucleotide Identity* (ANI) *Based Clustering*, cioè il confronto delle similarità tra le sequenze ottenute. Con un range di similarità maggiore del 95% sono state identificate un totale di 159 specie (di cui 15 archei) appartenenti a 36 diversi phyla. Tra i phyla più diffusi si trovano *Proteobacteria*, *Ignavibacteriae*, *Planctomycetes* e il già citato *Chloroflexi*.

Sono stati individuati 385 *contigs* virali applicando una ricerca di sequenze di provirus con VIRsorter (individua sequenze di provirus in maniera *reference*-indipendente che dipendente) e IMG/VR (il più grande database di UViGs, cioè *uncultivated viral genomes*) l'insieme dei *contigs* metagenomici e dei SAGs. Delle sequenze conservate nel database IMG/VR, solo 88 sono state le similarità riscontrate con i *contig* del campione. Inoltre, questi virus presenti già in IMG/VR provenivano tutti da biofilm termali e nessuno di essi aveva un ospite conosciuto. Tutto ciò evidenzia quanti siano ancora poco conosciute queste aggregazioni di microrganismi e virus. Solo 14 dei 36 phylum riscontrati nel biofilm erano ospiti dei 59 virus (56 sequenze singole e 3 piccoli cluster di sole due sequenze) contenuti nei MAGs e SAGs. Il phylum più affetto da infezioni virali è *Acidobacteria* mentre tra i phyla in cui si è rilevata un'assenza di profagi, pur essendo molto diffusi nel campione, si trovano *Aminicenantes*, *WOR-3* e *Nanoarchaeota* (Figura 6).

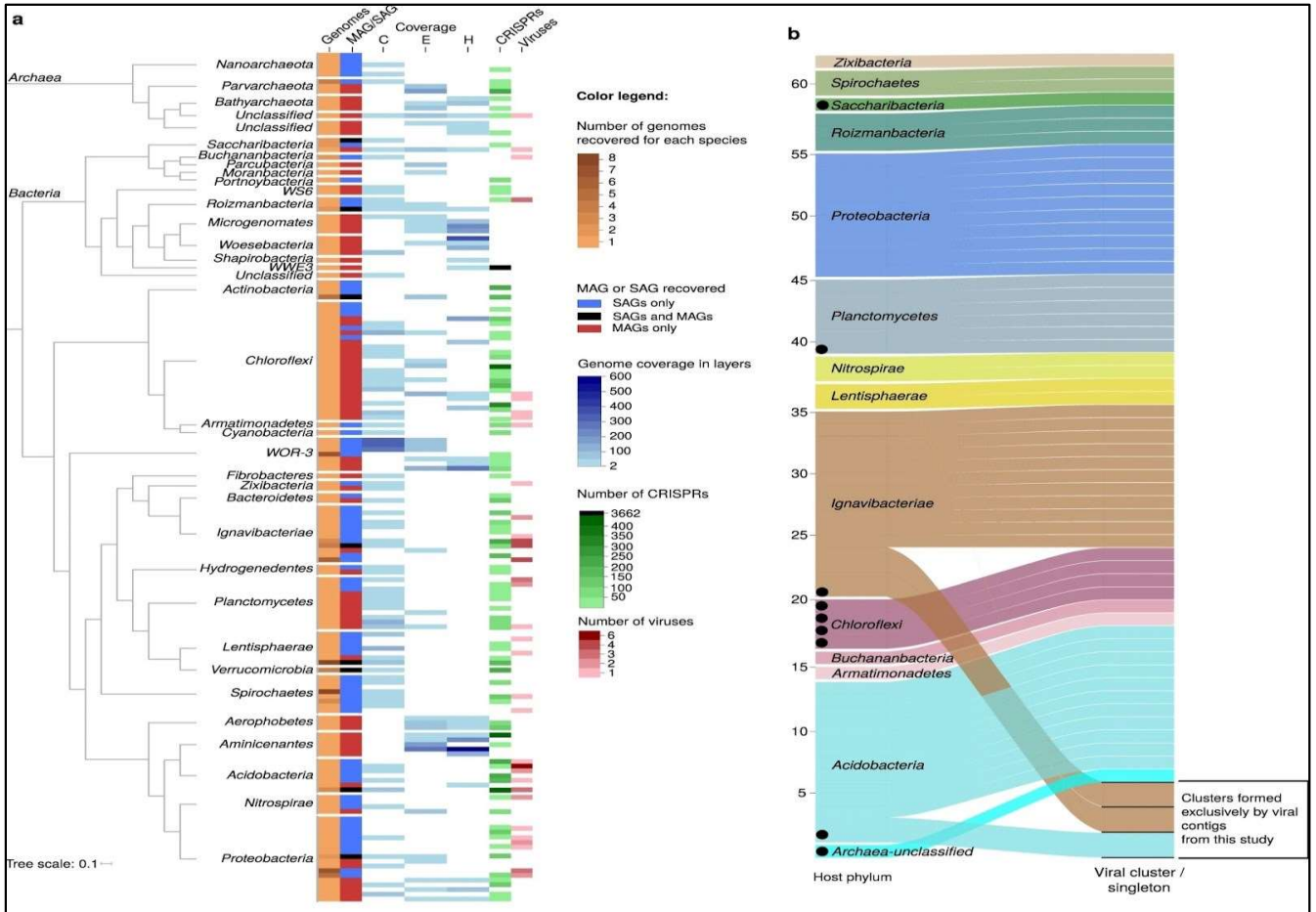


Figura 6: Specie procariotiche individuate tramite analisi di singola cellula e metagenomica (a). Relazioni tra virus e i loro ospiti (b). (Fonte: Jarett et al., 2020)

Per fare delle ipotesi sul range di infezione sono stati analizzati solo i tre cluster dato che gli altri virus, individuati con solo una singola sequenza, inevitabilmente conducono ad un unico ospite. Anche questi cluster, ognuno contenente due sequenze, comunque sono stati collegati ciascuno alla stessa specie batterica ospite. Questo, quindi, è indicativo di un'alta specializzazione delle specie virali per singoli ospiti. Tuttavia, su questo aspetto rimangono ancora delle incertezze sia perché ipotesi prevedevano che in un ambiente ricco di specie a stretto contatto e con scarsa mobilità fosse favorito lo sviluppo di un range infettivo elevato, sia perché i cluster analizzati sono composti ciascuno da due sequenze, un campione troppo piccolo per poter rappresentare l'intera popolazione.

#### 4. CONCLUSIONE

Questo elaborato si è proposto come obiettivo lo studio delle metodologie che hanno prodotto più efficacemente nuovi risultati e conoscenze sui biofilm. L'applicazione della metodologia *single cell genomics*, precedentemente all'articolo preso in studio, non era mai stata applicata per l'analisi di biofilm con l'obiettivo di studiare le interazioni virus-ospite. Le tecniche usate storicamente, come il 16S rRNA *sequencing*, ignorano la presenza di batteriofagi e di virus degli archei poiché viene sequenziato solo una regione precisa del genoma batterico. Nella metagenomica, invece, pur sequenziando il DNA virale oltre a quello



microbico, difficilmente si riesce a stabilire le relazioni tra virus e ospiti, un aspetto riconfermato in questo studio. *Single cell genomics* quindi si propone come una valida alternativa poiché, dato che l'analisi è al livello della singola cellula e non dell'intera comunità, permette di individuare efficacemente gli ospiti dei virus. Questo, assieme alla conferma della presenza del modello PtW permette di comprendere meglio come le popolazioni virali e microbiche sono strettamente legate tra loro in un rapporto che spesso risulta essere mutualmente vantaggioso. Questo non è di utilità solamente per un aspetto ecologico, infatti, essendo i biofilm ubiquitari, questi hanno un impatto su molti ambiti. L'industria alimentare, gli impianti di biogas, il controllo della corrosione in condutture idrauliche, l'industria farmaceutica, la depurazione delle acque e l'ambito sanitario, in particolare la sterilizzazione e disinfezione, beneficiano tutti di una migliore comprensione della struttura e delle popolazioni che abitano il biofilm.

Tuttavia, i risultati ottenuti dai ricercatori, data la complessità del tema affrontato, non offrono un'immagine completa dell'ecosistema del campione in studio, ma piuttosto un punto di partenza per future indagini. Principalmente, e come evidenziato dagli stessi autori dell'articolo, l'analisi delle specie è avvenuta completamente tramite sequenziamento del DNA procariotico e virale; quindi, in questo studio è ignorata la presenza di virus ssRNA e dsRNA. Anche se i batteriofagi, nella classificazione di Baltimore, sono raggruppati maggiormente nella prima e seconda categoria, cioè dsDNA e ssDNA, non si può ignorare l'importanza delle altre categorie. Dato l'elevato tasso di replicazione, di mutazione e il pattern differente di integrazione del genoma virale nei retrovirus, questi possono mostrare caratteristiche e adattamenti differenti rispetto a ciò che è stato rilevato per i virus a DNA. Futuri studi, quindi, potrebbero concentrarsi su virus a RNA applicando tecniche simili a quelle già citate in questo elaborato come l'*RNA single cell sequencing*.

Un altro aspetto per un possibile approfondimento riguarda l'analisi di *single cell sequencing* in altri strati oltre a quello già analizzato. Lo strato C preso in analisi, localizzato medialmente nella colonna del biofilm, è stato scelto principalmente perché ricco di specie poco caratterizzate. Tuttavia, dato che la teoria PtW è dipendente anche dalla profondità in cui le specie si trovano potrebbe essere utile effettuare analisi successive per verificare se sono presenti variazioni nell'adattamento virale. Il contributo dei dati ottenuti dalla metagenomica hanno permesso di individuare che in più strati, a diverse profondità, è prevalente il ciclo di infezione virale temperata. Ciononostante, data la maggiore efficacia riscontrata nella tecnica di *single cell sequencing*, se applicata in regioni differenti, potrebbero emergere risultati più dettagliati. In ogni caso, questo sicuramente permetterebbe di identificare altri accoppiamenti virus-ospite.

## 5. BIBLIOGRAFIA

Jarett, J.K., Džunková, M., Schulz, F., Roux, S., Paez-Espino, D., Eloë-Fadrosh, E., Jungbluth, S.P., Ivanova, N., Spear, J.R., Carr, S.A., Trivedi, C.B., Corsetti, F.A., Johnson, H.A., Becraft, E., Kyrpides, N., Stepanauskas, R., Woyke, T., 2020. Insights into the dynamics between viruses and their hosts in a hot spring microbial mat. *ISME J.* 14, 2527–2541. <https://doi.org/10.1038/s41396-020-0705-4>

Molecular ecology of microbial mats, 2014. . *FEMS Microbiol. Ecol.* <https://doi.org/10.1111/1574-6941.12408>

Prieto-Barajas, C.M., Valencia-Cantero, E., Santoyo, G., 2018. Microbial mat ecosystems: Structure types, functional diversity, and biotechnological application. *Electron. J. Biotechnol.* 31, 48–56. <https://doi.org/10.1016/j.ejbt.2017.11.001>

Silveira, C.B., Rohwer, F.L., 2016. Piggyback-the-Winner in host-associated microbial communities. *Npj Biofilms Microbiomes* 2, 1–5. <https://doi.org/10.1038/npjbiofilms.2016.10>

Saliba, A.-E., Westermann, A.J., Gorski, S.A., Vogel, J., 2014. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.* 42, 8845–8860. <https://doi.org/10.1093/nar/gku555>

D'Argenio, V., Tomaiuolo, R., Cariati, F., 2016. Whole genome amplification on single cell 40, 293–301. [https://doi.org/10.19186/BC\\_2016.034](https://doi.org/10.19186/BC_2016.034)

[https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina\\_sequencing\\_introduction.pdf](https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf)

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* 19, 455–477. <https://doi.org/10.1089/cmb.2012.0021>

Compeau, P.E.C., Pevzner, P.A., Tesler, G., 2011. Why are de Bruijn graphs useful for genome assembly? *Nat. Biotechnol.* 29, 987–991. <https://doi.org/10.1038/nbt.2023>

Xu, Y., Zhao, F., 2018. Single-cell metagenomics: challenges and applications. *Protein Cell* 9, 501–510. <https://doi.org/10.1007/s13238-018-0544-5>

Yilmaz, S., Singh, A.K., 2012. SINGLE CELL GENOME SEQUENCING. *Curr. Opin. Biotechnol.* 23, 437–443. <https://doi.org/10.1016/j.copbio.2011.11.018>



# Insights into the dynamics between viruses and their hosts in a hot spring microbial mat

Jessica K. Jarett<sup>1,2,11</sup> · Mária Džunková<sup>1,2</sup> · Frederik Schulz<sup>1,2</sup> · Simon Roux<sup>1,2</sup> · David Paez-Espino<sup>1,2</sup> · Emiley Eloë-Fadros<sup>1,2</sup> · Sean P. Jungbluth<sup>1,2</sup> · Natalia Ivanova<sup>1,2</sup> · John R. Spear<sup>3</sup> · Stephanie A. Carr<sup>4</sup> · Christopher B. Trivedi<sup>5</sup> · Frank A. Corsetti<sup>6</sup> · Hope A. Johnson<sup>7</sup> · Eric Becraft<sup>8,9</sup> · Nikos Kyrpides<sup>1,2</sup> · Ramunas Stepanauskas<sup>9</sup> · Tanja Woyke<sup>1,2,10</sup>

Received: 13 January 2020 / Revised: 3 June 2020 / Accepted: 11 June 2020 / Published online: 13 July 2020  
© The Author(s) 2020. This article is published with open access

## Abstract

Our current knowledge of host–virus interactions in biofilms is limited to computational predictions based on laboratory experiments with a small number of cultured bacteria. However, natural biofilms are diverse and chiefly composed of uncultured bacteria and archaea with no viral infection patterns and lifestyle predictions described to date. Herein, we predict the first DNA sequence-based host–virus interactions in a natural biofilm. Using single-cell genomics and metagenomics applied to a hot spring mat of the Cone Pool in Mono County, California, we provide insights into virus–host range, lifestyle and distribution across different mat layers. Thirty-four out of 130 single cells contained at least one viral contig (26%), which, together with the metagenome-assembled genomes, resulted in detection of 59 viruses linked to 34 host species. Analysis of single-cell amplification kinetics revealed a lack of active viral replication on the single-cell level. These findings were further supported by mapping metagenomic reads from different mat layers to the obtained host–virus pairs, which indicated a low copy number of viral genomes compared to their hosts. Lastly, the metagenomic data revealed high layer specificity of viruses, suggesting limited diffusion to other mat layers. Taken together, these observations indicate that in low mobility environments with high microbial abundance, lysogeny is the predominant viral lifestyle, in line with the previously proposed “Piggyback-the-Winner” theory.

---

These authors contributed equally: Jessica K. Jarett, Mária Džunková

**Supplementary information** The online version of this article (<https://doi.org/10.1038/s41396-020-0705-4>) contains supplementary material, which is available to authorized users.

✉ Mária Džunková  
dzunkovam@gmail.com

✉ Tanja Woyke  
twoyke@lbl.gov

<sup>1</sup> Department of Energy Joint Genome Institute, Berkeley, CA, USA

<sup>2</sup> Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

<sup>3</sup> Civil and Environmental Engineering, Colorado School of Mines, Golden, CO, USA

<sup>4</sup> Hartwick College, Oneonta, NY, USA

## Introduction

Viruses shape phylogenetic and functional diversity of bacterial and archaeal communities [1, 2]. Our knowledge of viruses is rapidly increasing as a result of advances in computational methods for virus DNA/RNA sequence detection, which has enabled the development of large databases of complete viral genomes and viral proteins [3–6].

<sup>5</sup> German Research Centre for Geosciences GFZ, Potsdam, Germany

<sup>6</sup> University of Southern California, Los Angeles, CA, USA

<sup>7</sup> California State University Fullerton, Fullerton, CA, USA

<sup>8</sup> University of North Alabama, Florence, AL, USA

<sup>9</sup> Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA

<sup>10</sup> University of California, Merced, CA, USA

<sup>11</sup> Present address: AnimalBiome, Oakland, CA, USA

To fully understand the effects of viruses on local and global ecosystems, such as the control of host growth dynamics [7] or host-cell reprogramming through auxiliary metabolic genes [8, 9], it is critical to establish host–virus linkages for viruses found in a given environment.

Most viral sequences are currently not associated with any host, and for viruses with a known host, their full host range is usually unknown, hampering ecological and evolutionary insights [10]. Likewise, there are various bacterial and archaeal candidate phyla that have not yet been linked to any known viruses [6, 11]. Multiple computational approaches have been used to predict virus–host linkages in genome sequence data. For example, algorithms for detection of prophages in bacterial genomes [11–15] or protein sequence-based machine learning [16] enabled the detection of the first viruses for several uncultured bacterial phyla. However, these approaches are limited to viral sequences that are integrated into the host genome. Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) in microbial genomes store short sequences from previous viral infections and they can be inferred bioinformatically at the species level [17], but only a limited number of bacterial lineages use this virus-defense system [18]. In addition, tRNA sequences acquired by viruses during host infection [5], along with host and virus similarities of oligonucleotide signatures [19, 20], can be used to link viruses with uncultured bacterial or archaeal hosts, but with only limited accuracy. Recently, it has been suggested that similar DNA methylation patterns may also allow the assignment of viruses to hosts within metagenomes, although this approach remains to be validated [21]. While computational methods are powerful, the necessary sequence features are not always present in both virus and host genomes [5, 18], and predicted associations should still be validated experimentally when possible [22].

Studies of viruses in isolated microbes used to be the gold standard for providing experimental evidence of infective strategies [23], but many hosts cannot be cultivated [24], and even intensively studied bacteria, such as those found in the human gut, often lack any isolated viruses [10]. Single-cell genomics represents a unique opportunity to link viruses and hosts with experimental evidence, because a certain portion of cells collected directly from an environment contain viruses in the cell or attached to the cell [25, 26]. This approach has shed light on important aspects of viral biology, such as horizontal gene transfer [27], the ability of viruses to reprogram their host's energy metabolism [28] and micro-diversity within viral genomes [29]. Single-cell genomics in the context of studying viruses has been successfully applied to a variety of habitats, ranging from seawater [30] to hot springs [31].

It has been estimated that 40–80% of microbial cells on Earth reside in biofilms [32], but surprisingly they are

understudied by single-cell genomics. Biofilms are formed by aggregates of microorganisms in which cells are embedded in a self-produced matrix of extracellular polymeric substances that are adherent to each other and/or a surface [33]. Biofilms, which develop in a liquid–solid interface and contain layered organization of microorganisms, are called microbial mats [34]. Microbial mats are found across the planet in a variety of habitats [35–40]. Their laminated structure contains bacteria, archaea, and eukaryotes that work together, often symbiotically, sharing and cycling nutrients and energy [38, 41]. Often driven by photosynthesis at their surface and chemosynthesis at their base, microbial mat structures are a key component of the living world that provide us with a snapshot of how microorganisms work together in a complex, ordered community to propagate and ensure their survival. Microbial mats can also serve as potent biosignatures of life on Earth. As an example, stromatolites, laminated accretionary structures found throughout the rock record can be biogenic and are likely the mineralized, fossil record of microbial mats across time thus informing the evolution of life [39, 42, 43].

There are only a few single-cell genomic studies about bacteria in natural biofilms [44–46], in which viruses were not considered, leaving a knowledge gap in host–virus interactions in the microbiomes of these ecosystems. Viruses in biofilms have been typically studied separately from their hosts as purified viral particles, leaving CRISPR linkage as the only way of connecting them with their host [47]. The most common method for analyzing host–virus interactions in biofilms is computational modeling based on data from laboratory experiments performed with cultured bacteria-phage pairs [48–51], which complicates its application to complex natural biofilms. On the other hand, computational predictions have revealed important information about host–virus interactions in biofilms. Such work points out that host–virus interactions depend on a variety of factors, such as biofilm species composition, structural heterogeneity [52], and metabolic activity of bacteria in different layers of the biofilms [53]. It has also been suggested that viruses may enhance biofilm formation through induction of polysaccharide production [54, 55]. While computational predictions are the first step for interpreting dynamics of viral infection in biofilm microbial communities, experimental studies on naturally occurring biofilms are instrumental in order to both further our understanding of viral infections in complex natural biofilms and refine computational projections [56].

We characterized the host–virus linkages in a laminated microbial mat within the geothermal pool “Cone Pool” (Long Valley Caldera, CA, USA). We combined single-cell genomics with bulk shotgun metagenomics to predict different aspects of host–virus interactions, including viral host

range, viral genome copy number compared to its host, and diffusion of viruses across the mat layers. Taken together, our data point to a narrow host range and a low level of active viral replication in this microbial mat, indicative of a higher prevalence of temperate viruses in a lysogenic infection stage, and limited diffusion of viruses within the mat system. The Cone Pool mat has many functional similarities to other microbial mats, such as limited diffusion of viruses into deeper layers and diurnal vertical migration of redox gradients [57], which suggests that the results of this study might be more broadly applicable to other structured biofilms in different environments.

## Methods

### Sample collection and processing

Samples from a laminated microbial mat at a geothermal pool (“Cone Pool”), located in the Little Hot Creek geothermal spring area within the Long Valley Caldera, California (37.6905833° N, 118.844417° W), were collected on the 15<sup>th</sup> of August 2015, under the umbrella of the Microbial Dark Matter (MDM) Phase II study, an extension of the Genomic Encyclopaedia of Bacteria and Archaea MDM project (GEBA-MDM; [58]). An intact, submerged, dendrolitic cone, and the laminated mat beneath, was cored using a sterile drinking straw from the edge of the pool. The straw was shipped to the Colorado School of Mines, Golden, Colorado, on ice, and stored at 4 °C. Layers of the mat were delimited based on color and consistency, designated “A” through “I” by slicing through the straw using a sterile scalpel (Fig. 1). The layers “B” to “H” were extruded from the straw casing and divided for DNA extraction.

DNA for metagenomic and 16S rRNA gene amplicon sequencing was extracted from 0.25 g of each layer using the Xpedition™ Soil/Fecal DNA MiniPrep kit (Zymo Research Corp.), which uses bead beating based lysis previously shown to break capsids of dsDNA viruses [59]. The remainder of layer C was divided into replicates (~0.1–0.5 g each), submerged in 1 ml PBS, vortexed for 30 s, and centrifuged for 30 s at 2000 rpm. The 1 ml supernatant was then mixed with 100 µL of 0.2 µm filter sterilized GlyTE (20 ml 100× TE Buffer pH 8.0, 60 ml deionized water, 100 ml molecular grade glycerol), incubated at room temperature for 1 min, and then stored at –80 °C for downstream single-cell analysis.

### Bacterial/archaeal composition via amplicon sequencing

The 16S rDNA amplicons of the regions V4 and V5 were obtained using primers 515F-Y (5′-GTGYCAGCMGC

CGCGGTAA) and 926R (5′-CCGYCAATTYMTTTRAGTTT) [60] sequenced on an Illumina MiSeq sequencer in 2 × 300 bp run mode. Amplicons from the layer D were not sequenced due to poor DNA amplification. The sequence analysis was performed on  $3.6 \times 10^5$  ( $\pm 1.1 \times 10^5$ ) sequences per sample using the JGI iTagger v2.0 pipeline [61] that cluster sequences into operational taxonomic units (OTU) with 97% sequence similarity in the USEARCH software suite [62] and performs ecological analyses in QIIME [63] using RDP classifier v2.5 [64]. The purpose of 16S rDNA amplicon analysis was to select mat layers containing the highest abundance of understudied bacterial phyla [58] for subsequent shotgun metagenomic and single-cell genomic analysis.

### Bulk shotgun metagenome

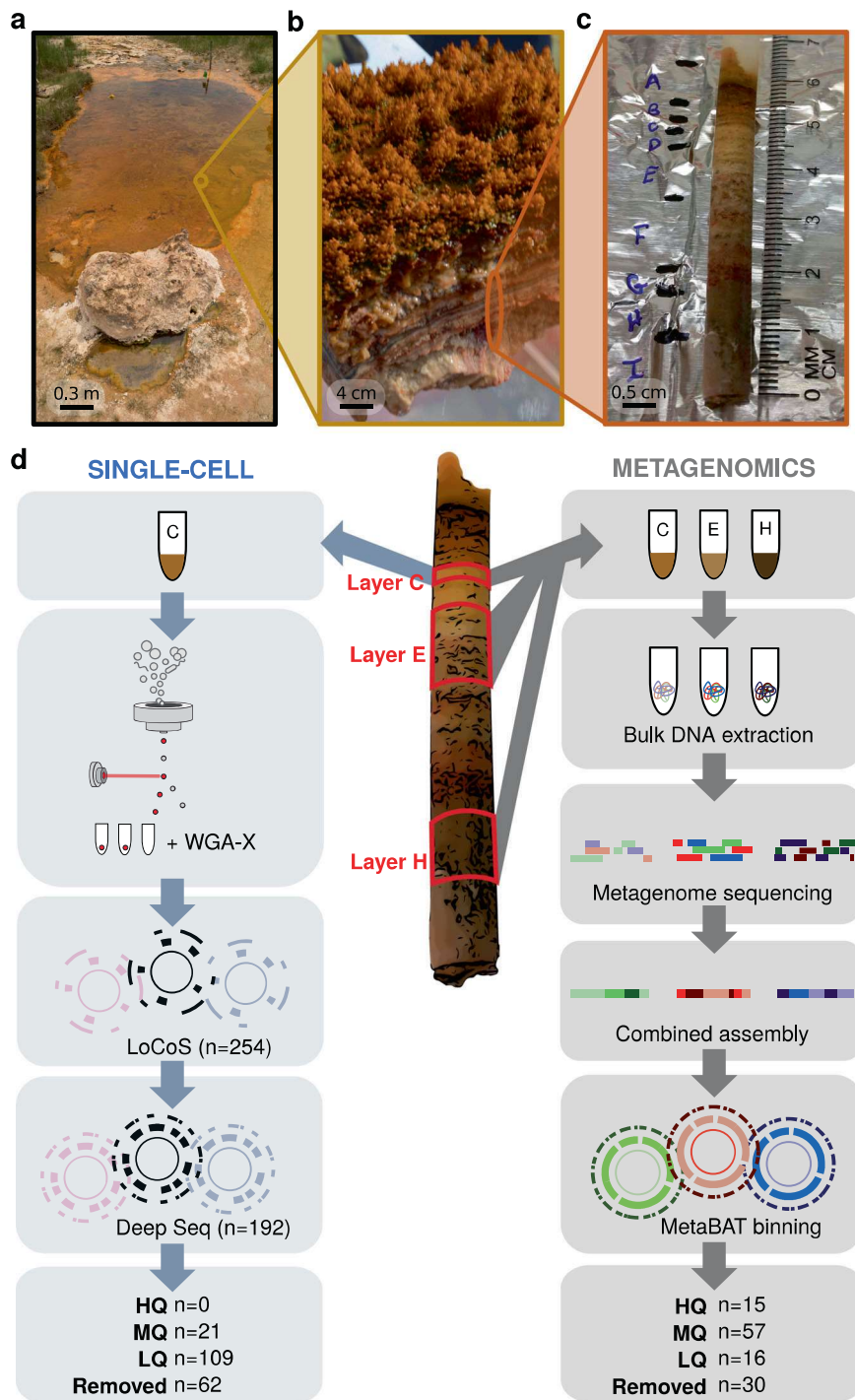
Genomic DNA from layers C, E, and H was sequenced on the Illumina HiSeq-2500 platform (libraries with 300 bp inserts) at the Department of Energy Joint Genome Institute in 2 × 150 bp mode resulting in  $101 \times 10^6$ ,  $63 \times 10^6$  and  $83 \times 10^6$  reads from layers C, E, and H, respectively (JGI sequencing project Gold IDs: Gp0147099, Gp0147100 and Gp0147101, respectively). Reads were trimmed and screened for common laboratory contaminants with BBTools v.37 (Bushnell [65], <http://bbtools.jgi.doe.gov>) and the sequencing errors were corrected by bfc v.181 [66] with the following parameters: “-s 10 g -k 21”. Mate-pair reads were assembled using SPAdes v.3.10.0 [67] with specified kmers 21, 33, 55, 77 and -meta flag. The co-assembly of all three metagenomic datasets was annotated using the IMG system [68, 69] and is publicly available under IMG taxon ID 3300022548. Using the contigs of the combined assembly, metagenome-assembled genomes (MAGs) were created by combining initial sets of genome bins from seven different binning approaches: (1) MaxBin v1.4.5 [70] using the universal 40 marker gene set and (2) the 107 marker gene set; (3) MaxBin v2.2.4 [71] with default parameters; (4) MetaBAT1 v0.32.5 [72] using the “super-specific” parameter and (5) “super-sensitive” parameter; (6) MetaBAT2 v2.12.1 [73] using default parameters; and (7) CONCOCT v0.4.0 [74] using default parameters. All binning methods used a minimum contig size of 3000 bp. Bins generated using the seven methods were used as input to DAS Tool v1.1.0 [75], which was run with default parameters to generate the final MAG set.

### Single amplified genomes

Single amplified genomes (SAGs) were generated from layer C at the Single Cell Genomics Center at Bigelow Laboratory for Ocean Sciences. Briefly, single cells were

**Fig. 1 Experimental workflow.**

**a** Overview of Cone Pool hot spring; **b** section through microbial mat, showing dendritic cones and layers of the mat; **c** delineation of layers of the mat; and **d** Sequencing workflow for single-cell genomics (left, layer C) and shotgun metagenomics (right, layers C, E, and H). The numbers at the bottom of the figure indicate the number of resulting high quality (HQ), medium quality (MQ) and low quality (LQ) genomes, as based on MISAG/MIMAG standards.



isolated with fluorescence-activated cell sorting (FACS), lysed with a combination of freeze-thawing and alkaline lysis, and genomic DNA was amplified with WGA-X [76]. Barcoded libraries were created using Nextera XT (Illumina) following standard protocols. Low-coverage shotgun sequencing (LoCoS) and de novo assembly was carried out at Bigelow Laboratory for Ocean Sciences as previously described [76]. SAGs which had at least 50,000 trimmed reads, 1000 normalized reads, and a total

assembly size of at least 50 kb from LoCoS were selected for deeper sequencing ( $n = 192$ ). Deep sequencing of libraries was performed at the DOE Joint Genome Institute on the Illumina NextSeq platform in  $2 \times 150$  bp mode. Raw reads were filtered for quality and contamination with BBTools v.37 (Bushnell [65], <http://bbtools.jgi.doe.gov>), then BBTools components BBNorm and Tadpole were used for read normalization and error correction prior to assembly with SPAdes (v3.9.0; --phred-offset 33 --sc -k

22,55,95 --12) [77]. After assembly, 200 bp was trimmed from contig ends, contigs <2 kbp in length or with read coverage <2 were discarded, and annotation was carried out according to IMG standard protocols [68, 69].

### Taxonomic classification of the host genomes

The overall quality of SAGs and MAGs was categorized according to published standards [78]. CheckM v1.0.8 using lineage-specific workflow [79] was used to estimate completeness and contamination. Genomes with more than 10% estimated contamination ( $n = 62$ ) were excluded from further analysis.

A set of 56 universal single-copy marker proteins [80, 81] was used to build a phylogenetic tree for the newly generated SAGs and MAGs and a representative set of bacteria and archaea based on all publicly available microbial genomes in IMG/M ([68]; genomes accessed in April 2018) and about 8000 MAGs from the Genome Taxonomy Database (GTDB, [82], downloaded 18 October 2017). Marker proteins were identified with *hmmsearch* (version 3.1b2, *hmmer.org*) using a specific HMM for each of the markers. For every marker protein, alignments were built with *MAFFT* (v7.294b, [83]) and subsequently trimmed with *BMGE* using *BLOSUM30* [84]. Single protein alignments were then concatenated resulting in an alignment of 10,866 sites. Maximum likelihood phylogenies were inferred with *FastTree2* using the options: `-spr 4 -mlacc 2 -slownni -lg` [85]. In the following step, a subtree was built that employed the above described methods but included only query genomes and reference genomes from query-genome containing clades in the initial tree. The genomes were classified at the phylum level using the naming system of the National Center for Biotechnology Information taxonomy [86].

In addition, all genomes were classified with the GTDB Toolkit v0.1.0 [87] according to the GTDB taxonomy, which was created to standardize microbial taxonomy according to genomic information [88].

Finally, all MAGs and SAGs were clustered by *Mash* v1.1 [89] using a 95% average nucleotide identity (ANI) cutoff to approximate a species-level resolution.

### Putative viral sequence detection

*VirSorter* [3] and comparison to the IMG/VR database [6] were used to detect viral sequences in SAGs, MAGs, and the unbinned fraction of metagenomic contigs.

*VirSorter* was used on all contigs at least 2 kb in length, retaining predictions from categories 1 and 2 (fully viral contigs), and 4 and 5 (integrated viruses). Contigs in which bitscores of *pfam* hits were higher than bitscores of viral hits were not considered as viral contigs.

In the next step, all sequences were queried against the full IMG/VR database (version IMG\_VR\_2018-01-01\_3) with *BLAST* [90]. Hits where alignment length was at least 70% of the query or subject sequence length (whichever was shorter) were retained (at the “detection” threshold, [91]), and all overlapping and adjacent hit regions of each query contig were merged into consensus coordinate range (s) with the R package *plyranges* [92]. Single best hits with at least 90% identity and 75% alignment coverage to IMG/VR were assigned with the IMG/VR Viral Cluster ID (“assignment” threshold, [91]); otherwise the contigs were annotated as a novel virus.

Results from the two prediction approaches were combined to extract viral sequences. The coordinate range(s) predicted as viral by each method were merged with the R package *plyranges* [92], to obtain the most inclusive estimates of viral sequence. The final set of viral sequences was clustered with *MUMmer* v3.23 [93], requiring at least 95% ANI over at least 85% of the length of the shorter of the two sequences to add sequences to a cluster, in accordance to community standards [16].

An additional curation step was performed for viral contigs within MAGs in order to remove viral contigs erroneously binned together with a bacterial genome. For that reason, only integrated viruses containing flanking bacterial sequences (*VirSorter* categories 4 and 5) and viral contigs clustering with other integrated viruses were retained; otherwise they were assigned to the unbinned fraction which represented 40% of the reads.

### CRISPR-based host–virus linking

The host–virus pairings detected in SAGs and MAGs were tested for their consistency with clustered regularly interspaced short palindromic repeat (CRISPR)-based linking prediction [94]. CRISPRs in host genomes were identified using *CRT* v1.1 [17] with script modifications as used in the IMG/M [68], *CRISPRCasFinder* [95] and *CRISPRDetect* [96] and 100% identical spacers detected by these three programs in the genomes belonging to the same hosts species (95% ANI) were de-replicated by *cd-hit* v4.8.1 [97] and matched against all viral genomes found in the Cone Pool using *blastn* [90], where only identical hits over the complete length of a CRISPR spacer were scored as a positive match.

### Host–virus ratios

The host–virus genome coverage ratio was assessed by mapping the metagenomic reads from the layers C, E and H to a reference database containing host and virus genomes obtained in the previous steps by *BBMap* (Bushnell

[65], <https://sourceforge.net/projects/bbmap/>) using the default settings. From the viruses that formed sequence-similarity clusters (see above), only the longest viral contig was selected as a cluster representative for mapping. From the host genomes grouped by ANI (see above), the genome with the highest genome completeness and the lowest contamination (estimated by CheckM) was selected as the cluster representative for mapping. The number of reads per sample used for mapping was normalized to the lowest number of reads per sample obtained for the three layers. Only the genomes with reads distributed across more than 75% of their genome length (“assignment” threshold, [91]) were considered as positive hits.

## Results and discussion

### Cone Pool microbial community data through amplicons, SAGs and metagenomes

To first assess the overall microbial community structure of the Cone Pool hot spring microbial mat, 16S rRNA gene amplicon analysis was performed on the mat layers B, C, E, F, G and H (Fig. 1), exclusive of the cone tip (layer A), which has previously been published [57]. The amplicon data yielded 440 OTUs, of which 24 had average abundances higher than 1% (Supplementary Fig. S1). The six mat layers differed remarkably in their microbial composition. Layer B (Fig. 1) was dominated by aquatic thermophiles, as was the cone tip previously analyzed from the same sampling site [57]. In contrast, the lower layers had a higher proportion of candidate phyla, including *Aminicenantes* (OP8), *Microgenomates* (OP11), and *Edwardsbacteria* (AC1). Layer C was selected for single-cell genomics analysis to capture some of the candidate phyla representatives while minimizing any potential challenges that might occur during cell sorting due to accumulation of calcium carbonate in the lower layers of the mat. Layers C, E, and H, providing three reference points across the mat, were selected for shotgun metagenomic sequencing.

Single-cell genomics from layer C generated 254 SAGs; 192 SAGs met LoCoS selection criteria for deep sequencing, from which 130 SAGs passed our minimum genome quality thresholds of genome completeness, contamination and taxonomic classification as described in the “Methods” (Fig. 1, Supplementary Table S1). The 130 SAG assemblies averaged  $1 \pm 0.5$  Mbp in size with a  $34.1 \pm 16.0\%$  estimated genome completeness (range 2.7–74.9%), totaling 21 medium quality and 109 low quality genomes, as based on MISAG standards [78]. Binning contigs from the co-assembly of metagenomes

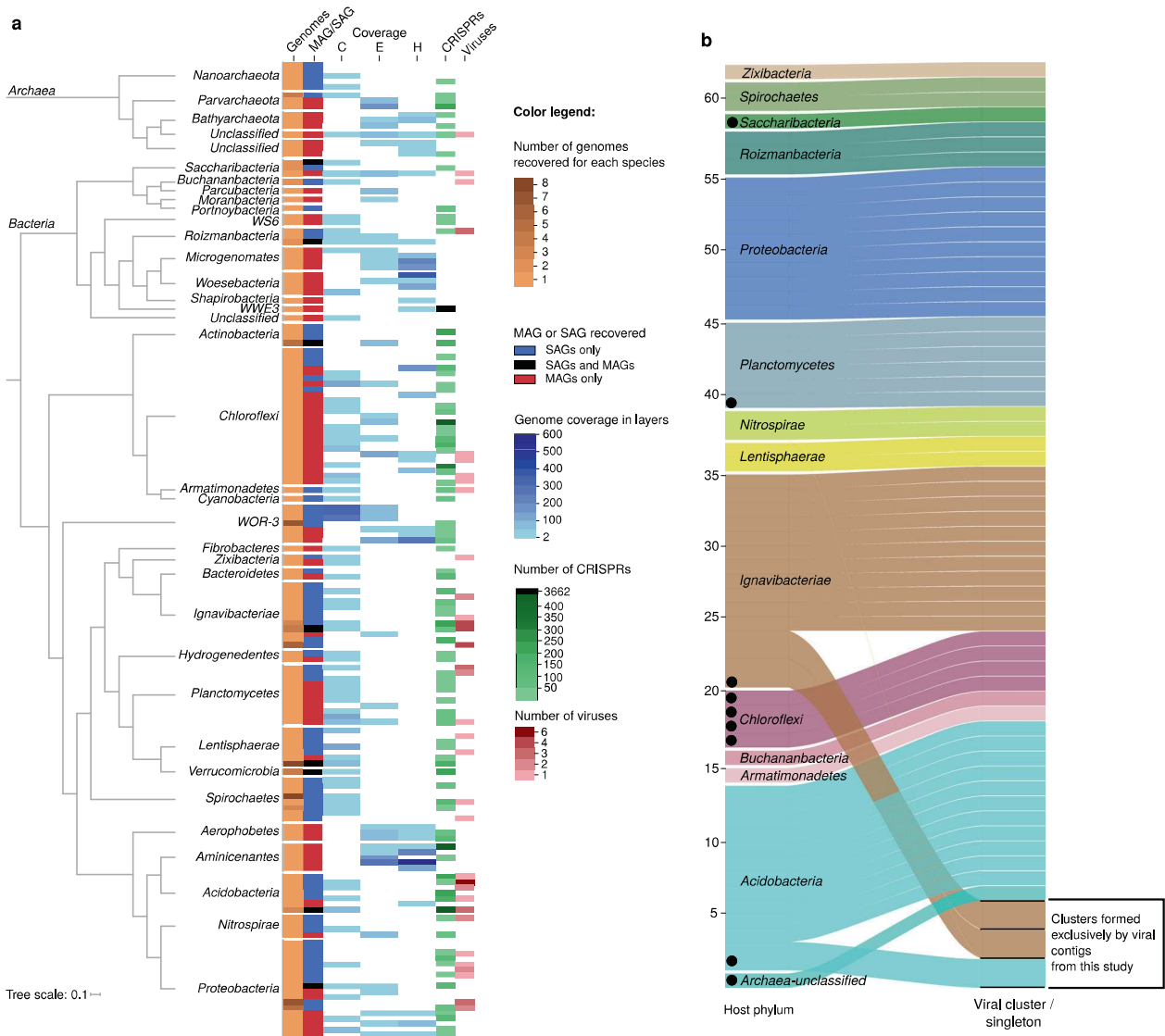
from layers C, E, and H resulted in 88 MAGs which averaged  $2 \pm 1.2$  Mbp in size with  $73 \pm 19\%$  estimated genome completeness (range 25.4–99.0%), consisting of 15 high-quality, 57 medium-quality, and 16 low-quality genomes, based on MIMAG standards ([78]; Fig. 1, Supplementary Table S1). ANI-based clustering (>95% ANI, [98]) of the 130 SAGs and 88 MAGs resulted in 144 bacterial and 15 archaeal ANI-based, nominal species-level groups, distributed across 36 phyla, mostly *Proteobacteria*, *Chloroflexi*, *Ignavibacteriae* and *Planctomycetes* (genome classification based on 56 markers genes, see “Methods”, Fig. 2a, Supplementary Table S1).

### Host–virus links in the Cone Pool mat microbiome

To make inferences about different aspects of host–virus interactions in the Cone Pool mat, virus detection tools were applied to SAGs and metagenome contigs. Using two different approaches for viral sequence detection (VirSorter and comparison to IMG/VR), 385 putative viral contigs were identified: 53 viral contigs (14%) were detected in 34 SAGs, 9 viral contigs were found (2%) in 9 MAGs, and 323 (84%) in the unbinned assembled metagenomes (Supplementary Table S2, Supplementary Fig. S2). Eighty-eight out of 385 detected viruses (23%) matched viruses from IMG/VR forming 52 groups with up to 5 viruses (Supplementary Fig. S3). The low number of matches to IMG/VR (at the time of analysis containing ~730,000 viruses) highlights the uniqueness of this sampling site, with a paucity of existing relatives in the database. The matched IMG/VR contigs mainly came from different thermal spring environments with some similarities to Cone Pool (Supplementary Table S2), but none of them were associated with a host in IMG/VR. Remarkably, the single-cell genomes from this study uncovered hosts for four IMG/VR contigs that had no previous host information.

Twenty-six percent of SAGs in this study contained 1–6 viral contigs (originating from attached/intracellular virions or integrated viruses), which was similar to infection rates reported for SAGs from surface ocean bacterioplankton [30]. When considering the results across host taxa, viruses were detected in 14 out of 36 host phyla, and in 21% of host species (95% ANI) detected in this study (Fig. 2a). *Acidobacteria* was the phylum in which viruses were detected most frequently (7 out of 9 genomes had a virus). In contrast, no viruses were associated with *Aminicenantes*, *Nanoarchaeota*, and WOR-3, despite these taxa being represented by multiple SAGs or MAGs. To summarize, we detected a total of 62 viruses in 43 different SAGs or MAGs, which sequence similarity-based clustering reduced to 59 viruses and 34 host species (Fig. 2b). These included 56 singletons and three clusters





**Fig. 2 Summary of microbes and viruses found in this study.** **a** Bacterial and archaeal species recovered by metagenomics and single-cell analysis. Phylum-level cluster representatives are displayed in a phylogenetic tree based on concatenated alignment of 56 universal single copy marker proteins. Each row represents a species based on 95% ANI. The first two columns represent the number of genomes in each species cluster and the source of the genomes of the given species (SAGs only, MAGs only or both MAGs and SAGs). The following three columns show read coverage of each species in metagenomics

samples from layers C, E and H. The last two columns indicate the number of CRISPR spacers and of viral contigs detected for each species. **b** Alluvial plot of virus and host connection. The left panel represents host species, colored by phylum, and the right panel shows viral clusters separated by horizontal black lines and singletons. The black dots indicate viruses detected on MAGs, while other viruses were detected on SAGs. Full results of this analysis, including the viruses with unknown host information, are shown in Supplementary Fig. S3.

of two sequences each, for which both members were always linked to the same host species (Fig. 2b). The small size of these viral clusters is not sufficient to assess the virus–host range in Cone Pool mat, but the wide diversity of detected viruses suggests that there is no dominant virus targeting multiple host species. One might hypothesize that a laminated microbial mat has the potential to select for broader host ranges of resident viruses, since viruses and bacteria exist in proximity over the long-term with limited diffusion between the layers

[52]. However, our single-cell genomic data from Cone Pool does not support this scenario.

### CRISPR-based linking

The detection of viral contigs in SAGs is evidence of ongoing viral infection in the collected single cells. Additional host–virus links can be obtained by analyzing the presence of CRISPRs in the host genomes that can indicate past viral infections. CRISPRs were detected in 43% of host

genomes in this study (Fig. 2a). From the 385 viruses detected in the Cone Pool, only three viruses were linked to their hosts by CRISPR. These hosts belonged to *Aminicenantetes*, *Lentisphaerae*, and *Portnoybacteria* phyla and their CRISPR-associated viruses were found in the unbinned fraction. The 62 viruses which were detected on 43 SAGs or MAGs were not targeted by CRISPRs found in these genomes, which suggests that these hosts have not yet built defenses against these viruses. The low CRISPR-based links in the Cone Pool highlights the utility of using single-cell genomics for linking viruses with their hosts.

### Lack of active replication of viruses associated with single cells

To gain insights into the putative interactions of viruses and hosts within the microbial mat environment, we were interested in the level of active viral replication. Viral contigs without flanking bacterial sequences in a SAG do not necessarily represent attached or intracellular virions during active infection. Due to the fragmentation of genome assembly from short reads, inactive integrated viruses might be found without the flanking bacterial sequences [10]. In the present study, only four SAGs contained integrated prophages with flanking bacterial sequences (Supplementary Table S2). To assess active viral replication in the Cone Pool microbial mat, we applied two different methodologies.

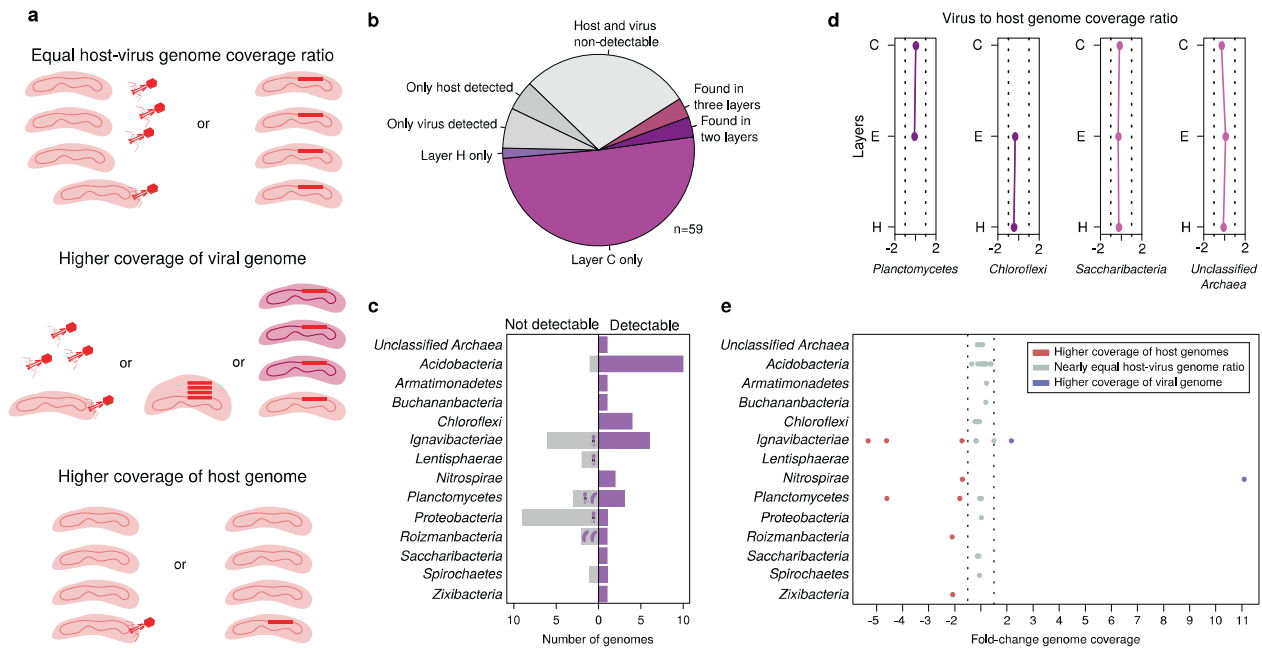
First, we applied an approach proposed by Labonté et al. [30], which relies on the correlation between the amount of DNA template and the speed of whole-genome amplification (WGA). If a SAG contains an actively replicating virus, WGA proceeds relatively fast (expressed as low crossing point (Cp) values of real-time WGA kinetics), but results in low host genome completeness, which is a consequence of: (a) large fraction of DNA available to WGA being viral, and (b) partial degradation of host genome by the lytic infection. The identification of SAGs with actively replicating viruses is based on comparison with SAGs with no viruses from the same experiment. In the Cone Pool data, none of the SAGs containing viral contigs had significantly lower Cp values of the WGA-X reaction, nor significantly lower genome coverage compared to the SAGs without viral signal (Supplementary Fig. S4), indicating that the viruses in our single-cell dataset were not actively replicating in the host cell.

Second, according to Schulz et al. [99], genomes present in multiple copies in a collected MDA-enriched sample, in theory, could be identified (despite the WGA bias) by having read coverages that were hundreds of times higher than other contigs in the sample. This trend was not observed in our SAG dataset, which further suggests that the collected single cells did not contain actively replicating viruses (Supplementary Fig. S5).

### Low level of active viral replication across the mat layers

We also analyzed whether viruses were actively replicating in all samples from the Cone Pool microbial mat layers. Replication of viruses can be assessed by mapping metagenomic reads to host–virus pairs detected in SAGs obtained from the same sample, followed by the comparison of host and virus genome coverage [10]. Typically, metagenomic samples contain DNA from bacterial cells as well as from highly abundant intracellular viral particles (if their capsids are opened during DNA extraction). Moreover, active replication of viruses inside the bacterial cells can also be detected in the metagenome data [100]. Because DNA of hosts and viruses in metagenomic reads is not amplified by WGA, it provides a more robust estimate of genomic DNA copies of both viruses and their hosts on the community level. By applying such read recruitment, we assumed that if a virus has the same genome coverage as its host, it is not actively replicating (Fig. 3a). In contrast, actively replicating virus would have a higher genome coverage than their host, due to the additional viral genome copies either free in the cytoplasm of a virocell [101–103] or in newly formed viral particles released from the host cell (Fig. 3a). We mapped the metagenomic reads from Cone Pool microbial mat layers C, E and H to de-replicated pairs of 59 viruses and their 34 hosts. It was possible to calculate the host–virus genome coverage ratios for 35 of these inferred pairs; the remaining 24 pairs had coverages below the detection threshold (>75% of the genome length covered, Fig. 3b), from which three pairs were not detectable due to the low coverage of viruses and four pairs due to the low coverage of the host (Fig. 3c), but the genome coverages of their detectable counterparts were not exceptionally high compared to the average in the Cone Pool. The pairs below the detection threshold involved hosts from seven host phyla (Fig. 3c). Given that assessing host–virus relationships directly from metagenomic data in complex communities is typically limited to highly abundant bacteria or archaea and viruses with known predation patterns [7, 104], our ability to calculate host–virus coverage ratios for so many pairs highlights the utility of single-cell genomics for capturing rare host–virus pairings.

From the 35 detectable host–virus pairs, only four pairs were found across multiple layers. Genome coverage of hosts was similar to the genome coverage of their viruses (<1.5× fold difference), and this ratio was conserved across different layers (Fig. 3d). In total, 76% of all detectable pairs involved hosts and viruses with nearly equal genome coverages, indicating a low level of active viral replication, while also suggesting that induction of integrated viruses or active infection in the mat is not



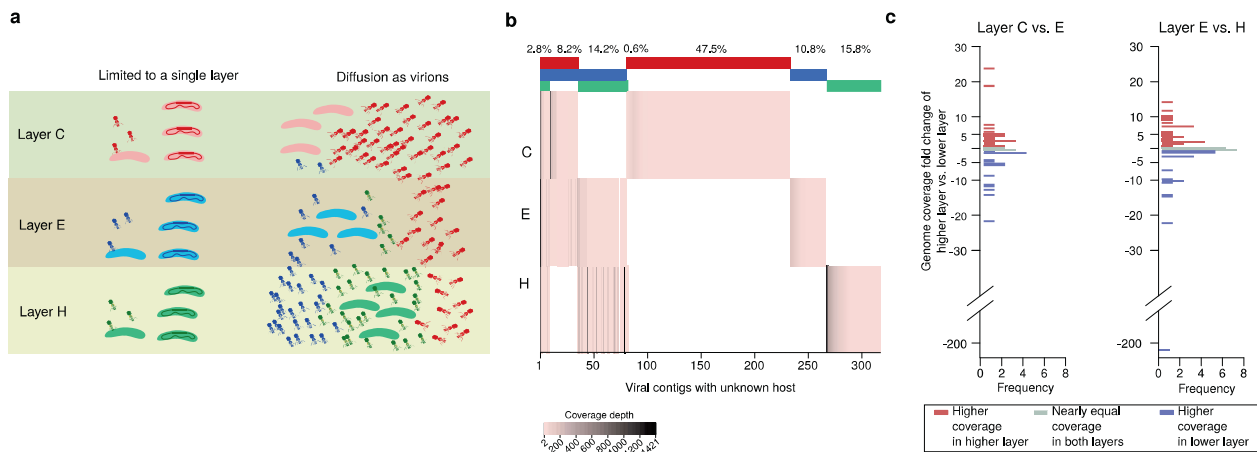
**Fig. 3** Host–virus genome read coverage ratios. **a** Possible scenarios for the interpretation of genome read coverage results. If there is a low rate of viral replication, we expect the genome coverage ratio of the virus and host to be nearly the same in a metagenome. Higher viral coverage could result from a higher number of virions compared to host cells, or more copies of the viral genome in each infected cell, but could also mean that a lysogenic virus has more than the single predicted host species. Higher coverage of a bacterial genome suggests that single-cell genomics captured a very rare infection event or that the virus infected only a subset of cells (i.e. only certain strains). **b** Detection of the 59 de-replicated host–virus pairings in the three layers. The gray portions indicate the pairs in which virus, host, or both genomes were below the detection threshold. For 35 pairings

(purple), detection was possible in at least one of the layers (>75% of the genome length covered). **c** Number of host genomes in each phylum for which the host–virus genome coverage ratio could (Detectable) or could not (Not detectable) be calculated. Virus and host icons indicate which one from the host–virus pair was above the detection threshold. **d** Fold-change of host and virus genome coverage for the four pairs detected in two or three layers. The dashed line indicates the 1.5× fold-change range. Dots positioned on the right from the host-baseline in the middle indicate higher coverage of the viral genome, while on the left indicate lower coverage of the viral genome compared to the host. **e** Fold-change of host and virus genome coverage of all 35 host–virus pairs (dots) grouped by host phyla. The distribution of points relative to the *x*-axis is described in **(d)**.

common, or was not occurring at the time of sampling (Fig. 3e). Alternatively, if some viruses were actively replicating, they did not form large numbers of progeny or only infected a small fraction of the available host cells, such that the host-to-virus genome coverage ratio looked similar at the population level in a given layer (Fig. 3a). This is in accordance with previous studies on microbial mats that were based on the counting of viral particles and bacterial cells by fluorescent microscopy [105, 106]. The study of Carreira et al. [106] performed on a photosynthetic microbial mat showed microscale (mm) and seasonal variation of the viruses-to-bacteria ratio, but viruses never outnumbered bacteria by orders of magnitude as is usually reported for the marine environment [107]. Our metagenomic read mapping analysis revealed eight viruses that had genome coverages lower than their hosts (up to 5.3× times lower, Fig. 3e). This suggests that these viruses were present only in a subset of cells of a given host species (Fig. 3a), e.g., in a susceptible strain or in a subpopulation of cells which has transiently lost immunity [108]. Lower genome coverage of a virus

compared to its host could also be explained by host genome polyploidy [109]. However, polyploidy is common only in extremely large bacterial cells and these were not targeted by FACS in this study. Only two viruses had a genome coverage higher than that of their host (>1.5×). The coverage of one of these viruses was 11 times higher than its *Nitrospirae* host (Fig. 3e), indicating induction of integrated viruses, plasmid-like replication, lytic lifestyle, or existence of additional hosts, which remain uncovered (Fig. 3a).

Interestingly, each of the 35 inferred viruses was only detectable in those layers where its host was present, which indicates that these viruses likely resided in proximity to and/or within host cells. It is possible that integrated viruses were spontaneously induced in a small portion of cells to enhance biofilm integrity, and thus the resulting viral particles remained near the host cells from which they originated [110, 111]. However, this does not exclude the possibility of released virions penetrating adjacent layers, where they might perish without their hosts and thus remain undetectable [112].



**Fig. 4 Predicted virus diffusion across the mat layers.** **a** Possible scenarios of virus diffusion across the layers. Given that the layers differ by bacterial composition, a layer-specific viral composition suggests limited diffusion across the layers. If the viruses can move across layers, their abundances would vary across layers. **b** Genome coverage of the 323 virus sequences with no host information across

the mat layers. The red, blue, and green stripes above the heatmap indicate whether the viruses were detected in one, two, or three layers, and the percentage above the stripes indicate the proportion of each of these groups. **c** Histograms of genome coverage fold-change of viruses detected in upper layer compared to the lower layer.

### Mat viruses exhibit high layer specificity

The analysis of the viral distribution in the mat layers described above was limited to the 35 viruses with identified hosts detectable in the metagenomic reads, pointing to only four viruses distributed across multiple mat layers (Fig. 3d). However, there were another 323 viral contigs detected in the unbinned fraction that could not be linked to any host (Supplementary Fig. S2). Given that the bacterial distribution in the Cone Pool mat is generally layer-specific (Supplementary Fig. S1, Fig. 2a), these viruses with unknown hosts could also be used for a more robust analysis of the viral composition and diffusion across the mat layers (Fig. 4a).

We mapped the metagenomic reads from the layers C, E, and H to the viral contigs from the unbinned fraction. We found that 75% of them were detected only in one layer (Fig. 4b), which confirms the high layer-specific viral composition of the Cone Pool microbial mat. Similar spatial distribution of viruses on a microscale was reported for the soil sampled in 1-cm resolution [113] and from the surface and inside of plant leaves [114].

The highest diversity of viruses was detected in layer C (Fig. 4b), which might be a consequence of this layer being close to the mat surface where it would be exposed to viruses from outside of the mat. This observation is in accordance with reports of higher viral counts on a microbial mat surface compared to the mat interior [53, 106]. Because there was no excessive accumulation of virions in the lower layers compared to upper layers in the Cone Pool microbial mat (Fig. 4c), gravity is not likely the principal force of virion diffusion in this mat; rather it is the tight virion/host-cell association in the spatial

zonation of the mat that determines virion abundance. In addition, it has been shown that biofilm maturation and structural complexity are critical for protecting the bacteria against a continuous flux of phages from outside of the biofilm [115].

Interestingly, the metagenomic read mapping showed that the genome coverage of 93% of the viruses from the unbinned fraction was lower than the genome coverage of both MAGs and SAGs in this study. This likely means that the Cone Pool mat has only a small portion of actively replicating viruses. Taken together, it is likely that the low viral progeny number compared to host cells is an effective strategy for viral survival in the layers of the Cone Pool mat, as there is a limited supply of new hosts and a limited possibility to diffuse to other layers. This is in accordance with the “Piggyback-the-Winner” theory of viral infection stating that lysogeny is the predominant viral lifestyle in low mobility environments with high microbial abundance, where integrated viruses help their hosts to prevent infection by closely-related viruses which increases their ability to compete with other host species [116]. While integration of a virus into the host chromosome generally represents an extra energetic expense for the host, such expenses are probably insignificant in nutrient-rich and protective environments, such as a microbial mat [117]. Taken together, the predicted host–virus interactions in the Cone Pool mat are quite similar to the gut microbiome, which is generally characterized by a high concentration of bacterial cells and a high prevalence of lysogeny [118]. In comparison, aquatic environments with higher mobility of biomass are generally reported to have dynamic virus–host ratios [119], but this dynamics can decrease with sampling depth [120]

or can have seasonal variations when switches between lytic and lysogenic cycles occur [121].

## Conclusions

By detecting viral contigs in both SAGs and MAGs from the Cone Pool microbial mat, we linked 59 viruses with 34 hosts, many of them belonging to taxonomic groups with no cultured representatives. Due to the complex bacterial/archaeal and viral composition of this mat and the limited direct detection of integrated viruses in host genome assemblies from this mat metagenomes, we infer that most of the host–virus pairings in this study could not have been obtained if not for the employed single-cell genomics methodology. This highlights the utility of the technique for linking viruses with their hosts to provide a deeper understanding of mat microbial ecology.

Single-cell genomics paired with metagenomic read recruitment provided insights into viral host range and distribution across the mat layers, as well as predicted viral lifestyle. While previous studies on bacterial biofilms have focused on computational predictions and were based on laboratory experiments with a limited number of cultivated phages and bacteria, this is the first study that reports detailed infection dynamics in a complex natural microbial mat for host–virus pairs with known identity. Our results point to a low rate of active viral replication in each layer and a limited spread of viral particles across the mat layers. This hints to different factors shaping the mat layers microbial composition, such as bacterial predation by nematodes [122] or seasonal variation in bacterial/archaeal metabolism [123]. While our work shed light on diversity of dsDNA and ssDNA viruses in microbial mats, further work might involve investigation of ssRNA viruses, as these have been found in abundance in other environments [124, 125] and could play a role in biofilm formation [54]. We believe that our observations can be expanded to other natural biofilms and contribute to the development of novel microbial dynamics prediction models for biofilms.

## Data availability

The sequences and the genome assemblies are accessible on Integrated Microbial Genomes and Microbiomes website <https://img.jgi.doe.gov/> with IDs listed in Supplementary Table S1.

**Acknowledgements** We wish to thank the course participants of the International Geobiology Course (2015 and 2016) for assistance with sampling and geochemical measurements at Cone Pool. The United States Forest Service Inyo District, Bishop, California, granted a research permit for field work and sampling (#MLD15053) at Cone Pool. We also thank Janey Lee for preparing the illustrations for Fig. 1,

and Heather Maughan for proofreading and comments on the final version of the manuscript. We thank the staff of the Bigelow Laboratory Single Cell Genomics Center for SAG generation.

**Funding** The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231. JRS is supported by the NASA Astrobiology Institute Rock Powered Life grant. RS is supported by NSF grants DEB-1441717, OCE-1335810, and OIA-1826734.

**Author contributions** JKJ and MD wrote the manuscript, analyzed data, and prepared figures, with input from all authors. FAC, HAJ, and JRS initiated sampling at the Cone Pool site and assisted with field work. SAC, CBT, and JRS assisted with field work, sample processing, and provided subject area expertise. EE-F. performed combined assembly of metagenomic samples. SPJ generated MAGs. EB and RS generated SAGs. FS classified host genomes. DP-E. provided subject area expertise in viral detection by IMG/VR. SR provided subject area expertise in viral detection by VirSorter and viral sequences clustering. NI and NK provided subject area expertise. TW supervised work. All authors read and approved the final manuscript.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Suttle CA. Marine viruses-major players in the global ecosystem. *Nat Rev Microbiol.* 2007;5:801–12.
2. Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasić L, Thingstad TF, Rohwer F, et al. Explaining microbial population genomics through phage predation. *Nat Rev Microbiol.* 2009;7:828–36.
3. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. *PeerJ.* 2015;3: e985.
4. Graziotin AL, Koonin EV, Kristensen DM. Prokaryotic virus orthologous groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res.* 2017;45: D491–8.
5. Paez-Espino D, Eloë-Fadrosch EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth's virome. *Nature* 2016;536:425–30.

6. Paez-Espino D, Chen IA, Palaniappan K, Ratner A, Chu K, Szeto E, et al. IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. *Nucleic Acids Res.* 2017;45: D457–65.
7. Alrasheed H, Jin R, Weitz JS. Caution in inferring viral strategies from abundance correlations in marine metagenomes. *Nat Commun.* 2019;10:501.
8. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature.* 2016;537:689–93.
9. Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang HB, et al. Host-linked soil viral ecology along a permafrost thaw gradient. *Nat Microbiol.* 2018;3:870–80.
10. Džunková M, Low SJ, Daly JN, Deng L, Rinke C, Hugenholtz P. Defining the human gut host-phage network through single-cell viral tagging. *Nat Microbiol.* 2019;4:2192–203.
11. Roux S, Hallam SJ, Woyke T, Sullivan MB. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *eLife.* 2015;4:e08490.
12. Fouts DE. Phage\_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res.* 2006;34:5839–51.
13. Akhter S, Aziz RK, Edwards RA. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res.* 2012;40: e126.
14. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics.* 2008;24:863–5.
15. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 2016;44:W16–21.
16. Roux S, Krupovic M, Daly RA, Borges AL, Nayfach S, Schulz F, et al. Cryptic Inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nat Microbiol.* 2019;4:1895–906.
17. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinform.* 2007;8:209.
18. Burstein D, Sun CL, Brown CT, Sharon I, Anantharaman K, Probst A, et al. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nat Commun.* 2016;7:10613.
19. Ahlgren NA, Fuchsmann CA, Rocap G, Fuhrman JA. Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. *ISME J.* 2019;13:618–31.
20. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. Alignment-free d\*2 oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.* 2017;45:39–53.
21. Beaulaurier J, Zhu S, Deikus G, Mogno I, Zhang XS, Davis-Richardson A, et al. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat Biotechnol.* 2018;36:61–9.
22. Shkoporov AN, Khokhlova EV, Fitzgerald CB, Stockdale SR, Draper LA, Ross RP, et al.  $\Phi$ CrAss001 represents the most abundant bacteriophage family in the human gut and infects *Bacteroides intestinalis*. *Nat Commun.* 2018;9:4781.
23. Flores CO, Meyer JR, Valverde S, Farr L, Weitz JS. Statistical structure of host-phage interactions. *PNAS.* 2011;108:E288–97.
24. Lloyd KG, Steen AD, Ladau J, Yin J, Crosby L. Phylogenetically novel uncultured microbial cells dominate Earth microbiomes. *mSystems.* 2018;3:e00055–18.
25. Woyke T, Doud DFR, Schulz F. The trajectory of microbial single-cell sequencing. *Nat Methods.* 2017;14:1045–54.
26. Dang VT, Sullivan MB. Emerging methods to study bacteriophage infection at the single-cell level. *Front Microbiol.* 2014;5:724.
27. Labonté JM, Field EK, Lau M, Chivian D, Van Heerden E, Wommack EK, et al. Single cell genomics indicates horizontal gene transfer and viral infections in a deep subsurface Firmicutes population. *Front Microbiol.* 2015;6:349.
28. Roux S, Hawley AK, Torres Beltran M, Scofield M, Schwientek P, Stepanauskas R, et al. Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *eLife.* 2014;3:e03125.
29. Martinez-Hernandez F, Fornas Ö, Lluésma Gomez M, Garcia-Heredia I, Maestre-Carballea L, López-Pérez M, et al. Single-cell genomics uncover *Pelagibacter* as the putative host of the extremely abundant uncultured 37-F6 viral population in the ocean. *ISME J.* 2019;13:232–6.
30. Labonté JM, Swan BK, Poulos B, Luo H, Koren S, Hallam SJ, et al. Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton. *ISME J.* 2015;9:2386–99.
31. Munson-McGee JH, Field EK, Bateson M, Rooney C, Stepanauskas R, Young MJ. Nanoarchaeota, their sulfobales host, and Nanoarchaeota virus distribution across Yellowstone National Park hot springs. *Appl Environ Microbiol.* 2015;81:7860–8.
32. Flemming HC, Wuertz S. Bacteria and archaea on Earth and their abundance in biofilms. *Nat Rev Microbiol.* 2019;17:247–60.
33. Flemming HC, Wingender J, Szewzyk U, Steinberg P, Rice SA, Kjelleberg S. Biofilms: an emergent form of bacterial life. *Nat Rev Microbiol.* 2016;14:563–75.
34. Prieto-Barajas CM, Valencia-Cantero E, Santoyo G. Microbial mat ecosystems: structure types, functional diversity, and biotechnological application. *Electron J Biotechnol.* 2018;31:48–56.
35. Spear JR, Barton HA, Robertson CE, Francis CA, Pace NR. Microbial community biofabrics in a geothermal mine adit. *Appl Environ Microbiol.* 2007;73:6172–80.
36. Varin T, Lovejoy C, Jungblut AD, Vincent WF, Corbeil J. Metagenomic profiling of Arctic microbial mat communities as nutrient scavenging and recycling systems. *Limnol Oceanogr.* 2010;55:1901–11.
37. Kunin V, Raes J, Harris JK, Spear JR, Walker JJ, Ivanova N, et al. Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. *Mol Syst Biol.* 2008;4:198.
38. Ley RE, Harris JK, Wilcox J, Spear JR, Miller SR, Bebout BM, et al. From the cover: unexpected diversity and complexity from the Guerrero Negro hypersaline microbial mat. *Appl Environ Microbiol.* 2006;72:3685–95.
39. Pepe-Ranney C, Berelson WM, Corsetti FA, Treants M, Spear JR. Cyanobacterial construction of hot spring siliceous stromatolites in Yellowstone National Park. *Environ Microbiol.* 2012;14:1182–97.
40. Robertson CE, Spear JR, Harris JK, Pace NR. Diversity and stratification of archaea in a hypersaline microbial mat. *Appl Environ Microbiol.* 2009;75:1801–10.
41. Prieto-Barajas CM, Alcaraz LD, Valencia-Cantero E, Santoyo G. Life in hot spring microbial mats located in the trans-Mexican volcanic belt: a 16S/18S rRNA gene and metagenomic analysis. *Geomicrobiol J.* 2018;35:704–12.
42. Berelson WM, Corsetti FA, Pepe-Ranney C, Hammond DE, Beaumont W, Spear JR. Hot spring siliceous stromatolites from Yellowstone National Park: assessing growth rate and laminae formation. *Geobiology.* 2011;9:411–24.
43. Mata SA, Harwood CL, Corsetti FA, Stork NJ, Eilers K, Berelson WW, et al. Influences of gas production and filament orientation on stromatolite microfabric. *Palaios.* 2012;27:206–19.

44. Marshall IP, Blainey PC, Spormann AM, Quake SR. A Single-cell genome for *Thiovulum* sp. *Appl Environ Microbiol*. 2012;78:8555–63.
45. McLean JS, Lombardo MJ, Ziegler MG, Novotny M, Yee-Greenbaum J, Badger JH, et al. Genome of the pathogen *Porphyromonas gingivalis* recovered from a biofilm in a hospital sink using a high-throughput single-cell genomics platform. *Genome Res*. 2013;23:867–77.
46. McLean JS, Lombardo MJ, Badger JH, Edlund A, Novotny M, Yee-Greenbaum J, et al. Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *PNAS*. 2013;110:E2390–9.
47. Sharma A, Schmidt M, Kiesel B, Mahato NK, Cralle L, Singh Y, et al. Bacterial and archaeal viruses of Himalayan hot springs at Manikaran modulate host genomes. *Front Microbiol*. 2018;9:3095.
48. Simmons M, Krescher K, Nadell CD, Bucci V. Phage mobility is a core determinant of phage-bacteria coexistence in biofilms. *ISME J*. 2018;12:531–43.
49. Simmons M, Bond MC, Koskella B, Drescher K, Bucci V, Nadell CD. Biofilm structure promotes coexistence of phage-resistant and phage-susceptible bacteria. *BioRxiv*. 2019; <https://doi.org/10.1101/552265>.
50. Eriksen RS, Svenningsen SL, Sneppen K, Mitarai N. A growing microcolony can survive and support persistent propagation of virulent phages. *PNAS*. 2018;115:337–42.
51. Heilmann S, Sneppen K, Krishna S. Coexistence of phage and bacteria on the boundary of self-organized refuges. *PNAS*. 2012;109:12828–33.
52. Abedon ST. Spatial vulnerability: bacterial arrangements, microcolonies, and biofilms as responses to low rather than high phage densities. *Viruses*. 2012;4:663–87.
53. Abedon ST. Bacteriophage exploitation of bacterial biofilms: phage preference for less mature targets? *FEMS Microbiol Lett*. 2016;363:fnv246.
54. Secor PR, Sweere JM, Michaels LA, Malkovskiy AV, Lazzareschi D, Katznelson E, et al. Filamentous bacteriophage promote biofilm assembly and function. *Cell Host Microbe*. 2015;18:549–59.
55. Fernández L, Rodríguez A, García P. Phage or foe: an insight into the impact of viral predation on microbial communities. *ISME J*. 2018;12:1171–9.
56. Sutherland IW, Hughes KA, Skillman LC, Tait K. The interaction of phage and biofilms. *FEMS Microbiol Lett*. 2004;232:1–6.
57. Bradley JA, Daille LK, Trivedi CB, Bojanowski CL, Stamps BW, Stevenson BS, et al. Carbonate-rich dendrolitic cones: insights into a modern analog for incipient microbialite formation, Little Hot Creek, Long Valley Caldera, California. *NPJ Biofilms Microbiomes*. 2017;3:32.
58. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*. 2013;499:431–7.
59. Trubl G, Roux S, Solonenko N, Li YF, Bolduc B, Rodríguez-Ramos J, et al. Towards optimized viral metagenomes for double-stranded and single-stranded DNA viruses from challenging soils. *PeerJ*. 2019;7:e7265.
60. Parada A, Needham DM, Fuhrman JA. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time-series and global field samples. *Environ Microbiol*. 2015;18:1403–14.
61. Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T, et al. Primer and platform effects on 16S rRNA tag sequencing. *Front Microbiol*. 2015;6:771.
62. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010;26:2460–1.
63. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. 2010;7:335–6.
64. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucl Acids Res*. 2014;42:D633–42.
65. Bushnell B. BMAP: A Fast, Accurate, Splice-Aware Aligner. 2014. <http://btools.jgi.doe.gov>.
66. Heng L. BFC: correcting Illumina sequencing errors. *Bioinformatics*. 2015;31:2885–7.
67. Nurk S, Meleshk D, Korobeynikov A, Pevzner P. metaSPAdes: a new versatile metagenomic assembler. *Genome Res*. 2017;27:824–34.
68. Chen IA, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res*. 2019;47:D666–77.
69. Huntemann M, Ivanova NN, Mavromatis K, Tripp HJ, Paez-Espino D, Tennesen K, et al. The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4). *Stand Genom Sci*. 2016;11:17.
70. Wu YW, Tang YH, Tringe SG, Simmons BA, Singer SW. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*. 2014;2:26.
71. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*. 2016;32:605–7.
72. Kang D, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*. 2015;3:e1165.
73. Kang D, Li F, Kirton ES, Thomas A, Egan RS, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*. 2019;7:e27522v1.
74. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. *Nat Methods*. 2014;11:1144–6.
75. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol*. 2018;3:836–43.
76. Stepanauskas R, Fergusson EA, Brown J, Poulton NJ, Tupper B, Labonté JM, et al. Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nat Commun*. 2017;8:84.
77. Nurk S, Bankevich A, Antipov D, Gurevich AA, Korobeynikov A, Lapidus A, et al. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J Comput Biol*. 2013;20:714–37.
78. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol*. 2017;35:725–31.
79. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. 2015;25:1043–55.
80. Yu FB, Blainey PC, Schulz F, Woyke T, Horowitz MA, Quake SR. Microfluidic-based mini-metagenomics enables discovery of novel microbial lineages from complex environmental samples. *Elife*. 2017;6:e26580.
81. Elie-Fadrosh EA, Paez-Espino D, Jarett J, Dunfield PF, Hedlund BP, Dekas AE, et al. Global metagenomic survey reveals a new

- bacterial candidate phylum in geothermal springs. *Nat Commun.* 2016;7:10476.
82. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol.* 2017;2:1533–42.
  83. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013;30:772–80.
  84. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol.* 2010;10:210.
  85. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE.* 2010;5:e9490.
  86. Federhen S. The NCBI Taxonomy database. *Nucleic Acids Res.* 2012;40:D136–43.
  87. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 2019;36:btz848.
  88. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol.* 2018;36:996–1004.
  89. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 2016;17:132.
  90. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215:403–10.
  91. Paez-Espino D, Pavlopoulos GA, Ivanova NN, Kyrpides NC. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. *Nat Protoc.* 2017;12:1673–82.
  92. Lee S, Cook D, Lawrence M. Plyranges: a grammar of genomic data transformation. *Genome Biol.* 2019;20:4.
  93. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. *Genome Biol.* 2004;5:R12.
  94. Sorek R, Kunin V, Hugenholtz P. CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol.* 2008;6:181–6.
  95. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, et al. CRISPRCasFinder, an update of CRISPRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.* 2018;46:W246–51.
  96. Biswas A, Staals RHJ, Morales SE, Fineran PC, Brown CM. CRISPRDetect: a flexible algorithm to define CRISPR arrays. *BMC Genomics.* 2016;17:356.
  97. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics.* 2006;22:1658–9.
  98. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. *PNAS.* 2005;102:2567–72.
  99. Schulz F, Alteio L, Goudeau D, Ryan EM, Yu FB, Malmstrom RR, et al. Hidden diversity of soil giant viruses. *Nat Commun.* 2018;9:4881.
  100. Luo E, Aylward FO, Mende DR, DeLong EF. Bacteriophage distributions and temporal variability in the ocean's interior. *MBio.* 2017;8:e01903–17.
  101. Forterre P. The virocell concept and environmental microbiology. *ISME J.* 2013;7:233–6.
  102. Howard-Varona C, Lindback MM, Bastien GE, Solonenko N, Zayed AA, Jang H, et al. Phage-specific metabolic reprogramming of virocells. *ISME J.* 2020;14:881–95.
  103. Rosenwasser S, Ziv C, Graff van Creveld S, Vardi A. Virocell metabolism: metabolic innovations during host–virus interactions in the ocean. *Trends Microbiol.* 2016;24:821–32.
  104. Reyes A, Wu M, McNulty NP, Rohwer FL, Gordon JI. Gnotobiotic mouse model of phage–bacterial host dynamics in the human gut. *PNAS.* 2013;110:20236–41.
  105. Yoshida-Takashima Y, Nunoura T, Kazama H, Noguchi T, Inoue K, Akashi H, et al. Spatial distribution of viruses associated with planktonic and attached microbial communities in hydrothermal environments. *Appl Environ Microbiol.* 2012;78:1311–20.
  106. Carreira C, Piel T, Staal M, Stuetz JBW, Middelboe M, Brussaard CPD. Microscale spatial distributions of microbes and viruses in intertidal photosynthetic microbial mats. *Springerplus.* 2015;4:239.
  107. Haro-Moreno JM, Rodriguez-Valera F, López-Pérez M. Prokaryotic population dynamics and viral predation in a marine succession experiment using metagenomics. *Front Microbiol.* 2019;10:2926.
  108. Cenens W, Makumi A, Govers SK, Lavigne R, Aertsen A. Viral transmission dynamics at single-cell resolution reveal transiently immune subpopulations caused by a carrier state association. *PLoS Genet.* 2015;11:e1005770.
  109. Soppa J. Polyploidy in archaea and bacteria: about desiccation resistance, giant cell size, long-term survival, enforcement by a eukaryotic host and additional aspects. *J Mol Microbiol Biotechnol.* 2014;24:409–19.
  110. Carrolo M, Frias MJ, Pinto FR, Melo-Cristino J, Ramirez M. Prophage spontaneous activation promotes DNA release enhancing biofilm formation in *Streptococcus pneumoniae*. *PLoS ONE* 2010;5:e15678.
  111. Gödeke J, Paul K, Lassak J, Thormann KM. Phage-induced lysis enhances biofilm formation in *Shewanella oneidensis* MR-1. *ISME J.* 2011;5:613–26.
  112. Jończyk E, Kłak M, Międzybrodzki R, Górski A. The influence of external factors on bacteriophages—review. *Folia Microbiol.* 2011;56:191–200.
  113. Vos M, Birkett PJ, Birch E, Griffiths RI, Buckling A. Local adaptation of bacteriophages to their bacterial hosts in soil. *Science.* 2009;325:833.
  114. Koskella B, Thompson JN, Preston GM, Buckling A. Local biotic environment shapes the spatial scale of bacteriophage adaptation to bacteria. *Am Nat.* 2011;177:440–51.
  115. Vidakovic L, Singh PK, Hartmann R, Nadell CD, Drescher K. Dynamic biofilm architecture confers individual and collective mechanisms of viral protection. *Nat Microbiol.* 2018;3:26–31.
  116. Silveira CB, Rohwer FL. Piggyback-the-Winner in host-associated microbial communities. *NPJ Biofilms Microbiomes.* 2016;2:16010.
  117. Bobay LM, Touchon M, Rocha EPC. Pervasive domestication of defective prophages by bacteria. *PNAS.* 2014;111:12127–32.
  118. Sausset R, Petit MA, Gaboriau-Routhiau V, De Paepe M. New insights into intestinal phages. *Mucosal Immunol.* 2020;13:205–15.
  119. Breitbart M, Bonnain C, Malki K, Natalie A, Sawaya NA. Phage puppet masters of the marine microbial realm. *Nat Microbiol.* 2018;3:754–66.
  120. Kavagutti VS, Andrei AS, Mehrshad M, Salcher MM, Ghai R. Phage-centric ecological interactions in aquatic ecosystems revealed through ultra-deep metagenomics. *Microbiome.* 2019;7:135.
  121. Brum JR, Hurwitz BL, Schofield O, Ducklow HW, Sullivan MB. Seasonal time bombs: dominant temperate viruses affect Southern Ocean microbial dynamics. *ISME J.* 2016;10:437–49.



122. Feazel LM, Spear JR, Berger AB, Harris JK, Frank DN, Ley RE, et al. Eucaryotic diversity in a hypersaline microbial mat. *Appl Environ Microbiol.* 2008;74:329–32.
123. Hegler F, Lösekann-Behrens T, Hanselmann K, Behrens S, Kappler A. Influence of seasonal and geochemical changes on the geomicrobiology of an iron carbonate mineral water spring. *Appl Environ Microbiol.* 2012;78:7185–96.
124. Callanan J, Stockdale SR, Shkoporov A, Draper LA, Ross RP, Hill C. Expansion of known ssRNA phage genomes: from tens to over a thousand. *Sci Adv.* 2020;6:eaay5981.
125. Creasy A, Rosario K, Leigh BA, Dishaw LJ, Breitbart M. Unprecedented diversity of ssDNA phages from the family Microviridae detected within the gut of a protochordate model organism (*Ciona robusta*). *Viruses.* 2018;10:404.