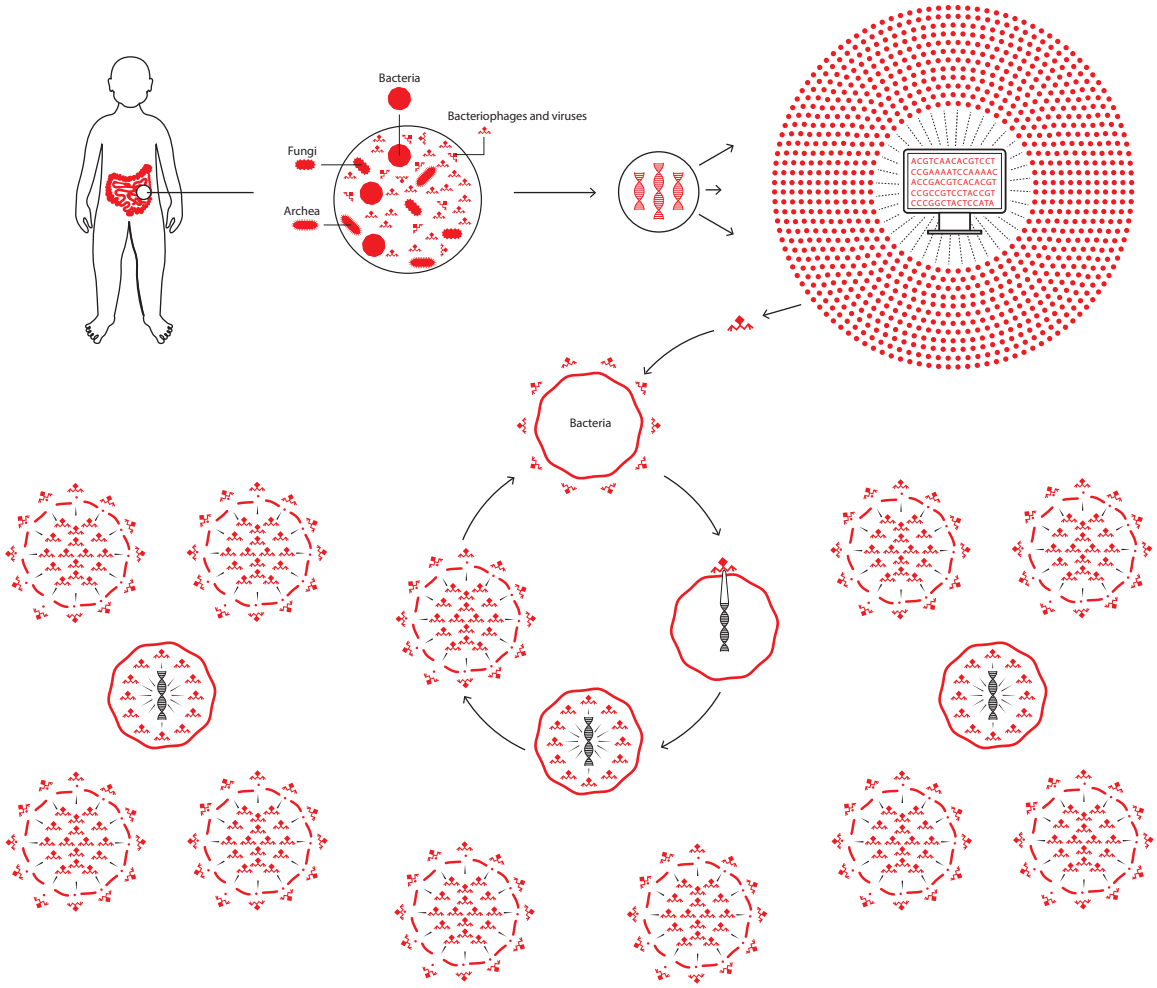
UNIVERSITY OF COPENHAGEN
FACULTY OF HEALTH AND MEDICAL SCIENCES

Unravelling phage genomics and biological-community structures in metagenomics

Ph.D. Thesis

Joachim Johansen

Principal supervisor: Simon Rasmussen
Co-supervisor: Søren Sørensen



This thesis has been submitted to the Graduate School of Health and Medical Sciences, University of Copenhagen on September 27, 2022

Preface

The work in this thesis was carried out between October 2019 and October 2022 in fulfillment of the requirements for acquiring a PhD degree. The work was mainly carried out at the Novo Nordisk Foundation Center for Protein Research under the Faculty of Health and Medical Sciences at the University of Copenhagen. The work was conducted under the supervision of Associate Professor Simon Rasmussen and Professor Søren Johannes Sørensen. In addition, part of this work was carried out during a four-month external stay at the Broad Institute of MIT and Harvard in Boston under the supervision of Professor Dr. Ramnik J. Xavier and Damian Rafael Plichta Ph.D. The funding for this PhD comes from the Novo Nordisk Foundation (grant no: NNF14CC0001). The external stay in the US was partly funded through personal travel scholarships from the Dansk Amerika Fondet, Knud Højgaard Fonden, Brorson legat, STIBOFONDEN and University of Copenhagen.

Copenhagen, September 2022
Joachim Johansen

Author
Joachim Johansen, MSc. cand. polyt.
Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen

Academic Supervisors
Associate Professor Simon Rasmussen, **Principal supervisor**
Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen

Professor Søren Johannes Sørensen, **Primary co-supervisor**
Section of Microbiology, Department of Biology, University of Copenhagen, Copenhagen, Denmark

Assessment Committee
Associate Professor Nicholas M I Taylor
Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen

Professor Mads Albertsen
Center for Microbial Communities, Aalborg University, Denmark

Dr Eduardo Rocha
Microbial Evolutionary Genomics, Institut Pasteur, France

Acknowledgements

Three years of computational science that worked out well has been condensed into this dissertation. For all the time that was not smooth sailing, I have been very lucky to have the support of many that I wish to acknowledge. First, a big thanks to Simon Rasmussen for all your support, energy and boundless enthusiasm for computational sciences and bioinformatics. Thanks to Søren Sørensen for co-supervision and for sharing your insights on microbiology. I would like to acknowledge all of my colleagues in the Rasmussen Lab who spent the entire or a great part of the time with me during this PhD. To the PhD-students of the group (Henry, Rosa, Marie Louise, Ricardo, Roman, Leonardo, Pau, Kirstine and Arnor) and postdocs (Jakob, Jonas, Lili, Katrine and Knud). Thanks for the many trips to the coffee machine, excessive cake meetings and amazing scientific retreat in Malta. Thanks to the CPR administration for supporting our scientific endeavors and to the SPA members for making all the hours at CPR more enjoyable with great events, beers and hygge.
I would also like to thank my collaborators and friends overseas. Thanks to Damian Plichta for many years of academic mentoring and collaboration, which started with running bash commands at Clinical Microbiomics and now a finished dissertation. Thanks to Hera Vlamakis who inspired me to take the final step and embark on the PhD. In addition, I am grateful to Ramnik J. Xavier for both his supervision and on-site and remote academic hosting during this PhD. Thanks to Thomas Pedersen and Eric Brown for sunny lunches and coffee breaks during my Autumn in Boston and also, by the way, doing three years of incredible lab work to verify my computational master thesis project and publish it. Furthermore, I wish to thank Kenya Honda and Koji Atarashi for their collaboration and inspiring work on human longevity and the microbiome. To my abroad-family in the US, Damian and Jesper, thanks for great hostmanship, superb company and slow mornings with several shots of espresso and long discussions on economy and technology.
Thanks to Christina for supporting me throughout all the ups and downs and painstakingly reading, editing, checking and *regulating* this dissertation. Finally, thanks to my family and friends for your unconditional support during this journey.

Table of contents

English summary

The human gut microbiome harbors several groups of residents including the bacterial, archeal, eukaryotic and viral kingdom. The bacterial kingdom is the most well studied and acknowledged for its significant role in metabolic processes and immune development important to the human host. The bacterial community is also a big contributor to the genetic pool and biomass of the gut which underscores its functional significance in the ecosystem. Yet, bacterial infecting viruses, known as bacteriophages, are suggested to equal or outnumber bacteria in the human gut. Due to the predatory mode of bacteriophages, they may exert a profound regulative role on bacterial constituents during health and disease.

In this thesis I explore computational frameworks using established methods from Artificial Intelligence and bioinformatics to mine and discover novel biological diversity including bacteria and viruses from human gut microbiomes. In the **first** article we present VAMB, a new method for binning. VAMB uses variational autoencoders to represent metagenomic sequences before the representation is clustered using a novel algorithm. We apply this method to a collection of synthetic metagenomes and thus demonstrate that Vamb creates more accurate bins than comparable software. By binning a large natural dataset with 1,000 human feces samples and almost 6 million assembled sequences, we demonstrate that Vamb can recreate bacterial strains with high phylogenetic resolution. In the **second** article we showcase how VAMB can be utilized for viral metagenomic binning in a framework we named PHAMB. Virus genomes present a different binning challenge compared to bacterial genomes as they are composed of smaller and more fragmented sequences in *de novo* assemblies. Even though VAMB was not originally designed with viruses in mind, our analysis shows that it successfully bins both bacteria and virus genomes in parallel, which facilitates downstream community analysis in metagenomic datasets. Importantly, we found that binning improves the total recovery and quality of virus genomes compared to single-sequence virus recovery across three different datasets. In the **third** article we apply viral genome binning to delineate viral populations in centenarian gut microbiomes to reveal novel viral diversity that may promote human longevity. Healthy aging seems to promote a rich and diverse virome that interacts with beneficial dominant bacterial hubs

in the microbiome. As bacteriophages represent a dynamic component of the microbiome, they may provide health promoting functional capabilities to the gut bacteria they infect. In support of this hypothesis, we discovered that centenarian bacteriophages encode key enzymes found in bacterial metabolic systems related to the conversion of sulfate to sulfide and methionine to homocysteine. Together with its bacterial part, the centenarian gut microbiome displayed increased potential for the conversion of sulfate to sulfide. A greater metabolic output of microbial hydrogen sulfide may in turn support mucosal integrity and resistance to pathobionts.

This thesis presents methodological frameworks for organizing bacteria and viruses in the human gut microbiome into biological meaningful entities and dissecting their potential impact on the human host. Until future generation sequencing technologies become cost effective, accurate and gradually replace current methods for studying metagenomics, computational methods such as metagenomics binning will be necessary for discovering and delineating bacterial and viral diversity and their intricate dynamics.

Dansk resumé

Det menneskelige tarmmikrobiom er vært for adskillige grupper af mikroorganismer, herunder bakterier, arkæer, eukaryoter og virus. Tarmbakterier er den mest undersøgte af disse og er anerkendt for dets rolle i tarmmetabolismen og immunsystemets udvikling, begge af stor betydning for den menneskelige vært. Desuden udgør bakterier en stor del af den genetiske arvemasse i tarmen som understreger deres funktionelle betydning i økosystemet. En mindre undersøgt gruppe i tarmens økosystem er bakterieinficerende virus, også kendt som bakteriofager. Bakteriofager er anslået til at udligne eller overstige antallet af bakterier i tarmen. Givet deres destruktive interaktion med bakterier, kan de have en stor betydning for bakteriebalancen under raske og sygdomsbetingede tilstande. I denne afhandling udforsker vi algoritmer og computerbaserede metoder med etablerede metoder fra kunstig intelligens og maskinlæring til at opdage og karakterisere tarmens biodiversitet, herunder bakterier og vira fra menneskelige tarmmikrobiomer. I den **første** artikel præsenterer vi VAMB, en computerbaseret algoritme til metagenomisk binning, hvilket kan oversættes til gruppering. VAMB er baseret på en variational autoencoder til at repræsentere metagenomiske sekvenser som nemmere kan splittes og grupperes til samlede bakterielle genomer. I artiklen har vi demonstreret at VAMB er bedre end tilsvarende værktøjer til at gruppere syntetiske metagenomer og genskabe bakteriestammer. Ligeledes har vi anvendt VAMB på et stort metagenomisk datasæt baseret på 1000 humane fæces prøver med næsten 6 millioner sekvenser og vist at VAMB kan genskabe bakteriestammer med høj fylogenetisk præcision. I den **anden** artikel beskriver vi, hvordan VAMB kan bruges til gruppering af tarmens virus genomer i en metode vi har kaldt PHAMB. Binning af virus genomer er betinget af andre udfordringer end bakterielle genomer, da deres genomer typisk er mindre og mere fragmenterede. Selvom VAMB ikke oprindeligt blev designet specifikt for virus, har vi vist at det kan anvendes til at genskabe både bakterielle og virale genomer parallelt, hvilket muliggør undersøgelse af begge biologiske domæner i metagenomiske datasæt. Desuden etablerer vi at at binning forbedrer den totale rekonstruktion og kvalitet af virusgenomer på tværs af tre forskellige datasæt. I den **tredje** artikel anvender vi begge metoder (VAMB og PHAMB) til at karakterisere og undersøge bakterielle og virale populationer i tarmmikrobiomer fra hundredårige samt

yngre kontrolgrupper. I studiet opdager vi først og fremmest ny viral diversitet der potentielt kan fremme menneskets levetid. Desuden, observerer vi at hundredåriges tarmmikrobiom huser et rigt og mangfoldigt virus system der interagerer med gavnlige bakterielle populationer. Det siges at bakteriofager kan bidrage med ekstra funktionel arvemasse til bakterierne de inficerer. Vi opdagede i forlængelse af denne hypotese at bakteriofager integreret i bakterier fra hundredårige bidrager med enzymer der faciliterer vigtige trin i bakterielle metaboliske systemer relateret til omdannelsen af sulfat til sulfid og methionin til homocystein. Tilsammen viste vi at hundredåriges tarmmikrobiom har et øget metabolisk potentiale for omdannelse af sulfat til sulfid, hvilket kan have stor betydning, da en øget mængde af svovlbrinte i tarmen kan understøtte tarmens integritet og resistens over for patogener. Afhandlingen beskriver computerbaserede metoder til at organisere bakterier og virus arter i det menneskelige tarmmikrobiom og en dissektion af deres potentielle indvirkning på den menneskelige vært. Indtil fremtidige sekventeringsteknologier bliver omkostningseffektive, nøjagtige og gradvist erstatter nuværende sekventeringsmetoder til at studere metagenomiske prøver, vil værktøjer baseret på binning være nødvendige for at etablere bakteriel og viral diversitet samt forstå deres indviklede dynamik.

List of publications

This thesis is based on the following three manuscripts:

PAPER I. Nissen, J.†; **Johansen, J**; Lundbye A.R.;, Kaae S.; C, Almagro A.; J, Grønbech; C., Jensen J.; L., Nielsen B.; H., Petersen N, T.; Winther, O.; Rasmussen, S.† (2021). *Improved metagenome binning and assembly using deep variational autoencoders*. Nature Biotechnology, 555-560. DOI: 10.1038/s41587-020-00777-4

PAPER II. **Johansen, J**; Plichta R.D.; Nissen, J.; Jespersen L. M; Shah A., S.; Deng, L.; Stokholm, J.; Bisgaard, H., Nielsen S.; D., Sørensen J; S., Rasmussen, S.† (2022). *Genome binning of viral entities from bulk metagenomics data*. Nature Communications, Article 965.

PAPER III. **Johansen, J**; Atarashi K., S.; Arai, Y.; Hirose, N.; Atarashi K.; Honda, K.; Sørensen J, Xavier R. J.†; Rasmussen, S.†; Plichta R.D.† (2022). *Centenarians have a diverse population of gut bacteriophages that may promote healthy lifespan.*
In review, Nature Microbiology

I have contributed to the following articles during my PhD, which are not included in this thesis:

PAPER I. Pedersen, T.K; Brown, E; Plichta R.D; **Johansen, J**; Twardus, W; Delorey, Toni; Lau, H; Vlamakis, Hera; Moon, J.J; Graham, D.B and Xavier, R.J† (2022) *T cell responses to immunodominant microbiome epitopes reflect dynamic transitions from tolerance to inflammation.* Cell Immunity. DOI: https://doi.org/10.1016/j.immuni.2022.08.016

PAPER II. Jespersen L. M; Munk, P.; **Johansen, J**; Kaas, R.; Webel, H.; Vigre, H.; Nielsen B., H; Rasmussen, S.; Aaastrup, F.† (2022) *Global within-species phylogenetics of sewage microbes suggest that local adaptation shapes geographical bacterial clustering*
In review, Communications Biology

PAPER III. Medina, H.R; Kutuzova, S; Nielsen, K.N; **Johansen, J**; Hansen, L.H; Nielsen, M; Rasmussen, S.† (2022) *Machine learning and deep learning applications in microbiome research.* Accepted, The ISME Journal.

PAPER IV. Líndez, P.P; **Johansen, J**; Sigurdsson, A.I; Nissen, J.N; Rasmussen,S.† *Adversarial and variational autoencoders improve metagenomic binning.* Manuscript in preparation.

PAPER V. Hauptmann, A†; **Johansen, J**; Stæger, F.F; Hansen, T; Rasmussen,S; Albreicht, A. GUTCYCLES consortium *The Heavy Metal Resistome of High Arctic Gut Microbiomes.*
Manuscript in preparation.

PAPER VI. Lundbye A.R.; Lundgaard, A.T; Medina, H.R; Aguayo-Orozco, A.; **Johansen, J**; Nissen, J; Brorsson, C.; ...; Banasik, K.; Rasmussen,S.; Brunak, S.†; IMI DIRECT Consortium. *Discovery of drug-omics associations in type 2 diabetes with generative deep-learning models.*
Accepted, Nature Biotechnology.

PAPER VII. Gallina, I.; Hendriks A. I.; Hoffmann S.; Larsen B. N.; **Johansen, J**, Colding-Christensen, C.; Schubert, L.; Sellés-Baige, S.; Kühbacher, U.; Gao, A.O; Räschle, M.; Rasmussen, S.; Nielsen L., M.; Duxin P.J† (2020) *The ubiquitin ligase RFWD3 is required for translesion DNA synthesis.* Molecular Cell. https://doi.org/10.1016/j.molcel.2020.11.029

PAPER VIII. Schubert, L; Hendriks, Ivo. A; Hertz, E; Wu, Wei; Sellés-Baige, S; Hoffmann, S; Viswalingam, K.S; Gallina, I; Pentakota, S; Benedict, B; **Johansen, J**; Apelt, K; Luijsterburg, M; Rasmussen, S; Lisby, M; Liu, Y; Nielsen, M.L; Mailand, N and Duxin P.J† (2022) *SCAI promotes error-free repair of DNA interstrand crosslinks via the Fanconi anemia pathway* EMBO reports. https://doi.org/10.15252/embr.202153639

† Corresponding author

Thesis content and overview

The central theme of this Ph.D. has been the development and application of methods to delineate bacterial and viral diversity from metagenomics. Successful organization of bacteria and viruses into ecological species hubs allows analysis of their interactions and the implications on the metazoan host like humans. The majority of microbiome studies that explored human gut biodiversity and its implications on human health have focused primarily on the bacterial constituents. The reason for this was twofold: (1) the virome (the collective of viruses in the environment) was mostly explored in viral-enriched metagenomic samples with *in vitro* isolated viral fractions and (2) the feasibility of exploring the virome from samples without viral preprocessing was barely described and problematic due to bacterial contamination. However, the growing wave of bulk/non-enriched metagenomic samples collected from the human gut, soil and marine environments is continuing to dwarf the number of collected viral-enriched metagenomic samples, which corresponds to metaviromes. Thus, there is a profound need for standardized methods to extract and explore the virome from bulk metagenomics due to their influence in the environment.

In **Chapter 1**, I provide a brief history on the discovery of bacterial and viral diversity in metagenomics and its current state and challenges. In addition, I outline a concise review on virus biology and dynamics with bacteria and altogether the possible implications to metazoan hosts like humans. In **Chapter 2**, I describe the bioinformatic methods and concepts with a focus on genome binning, genome annotation, machine learning and genome-driven ways to connect bacteria and viruses. In **Chapter 3**, I list and expand on the research objectives pursued during this Ph.D project. In **Chapter 4**, I describe the three major studies included in this thesis related to metagenomics binning of bacteria and viruses, and a study on the gut virome in humans with extreme longevity. In **Chapter 5**, I summarize and discuss the major results of the three studies included. In **Chapter 6**, I discuss future perspectives and suggestions for improved bacterial and virome analysis in metagenomics. **Chapter 7** contains ethical and legal permits and approvals required for the clinical and animal studies. **Chapter 8** includes the manuscripts included in this dissertation.

1 Background

1.1 The study of human gut biodiversity through DNA

Modern DNA sequencing technology has provided a way to read into the building blocks of life, from the first cultured bacteria to the first human genome [1, 2]. Today, we only need a biological specimen with DNA to sequence and read the genetic blueprint to determine who is there. Before modern sequencing technology brought us this far, we had to differentiate between bacteria by looking at them under a microscope based on their morphology, thanks to the inventions of Van Leeuwenhoek [3]. Culturing techniques have further improved this method by allowing single-culture isolates. However, In a mixed biological sample from soil, ocean water or feces that contains thousands of different bacterial species, isolating, culturing and determining every single one is an impossible task [4]. As not all bacteria are obligate aerobic bacteria, but obligate anaerobes like the trillions of bacteria in the human gut, many would not survive culturing on a petri dish.

A more precise determination of *who* a bacteria is can be made through its genetic blueprint of the 16S ribosomal RNA gene [5], which is an excellent phylogenetic marker for placing bacteria and archaea in the tree of life. The 16S rRNA gene is highly conserved between different bacteria and archaea due to its species specific signature of the letters A, T, C and G, which makes it useful for bacterial identification. Therefore, 16S rRNA sequencing technology has been an incredible tool for studying who is present in metagenomic samples from the human gut and allowed the first characterisations of commensal bacteria in the gut microbiome of hundreds of people. However, the metabolic and phenotypic traits of bacteria in an environment goes beyond the 16S rRNA marker gene. As a result, the desire to study the full genetic repertoire and *what* bacteria can do within an environment lead to the advent of modern shotgun sequencing.

1.2 Resolving uncultivated genomes in metagenomics

Shotgun sequencing ushered in an era of genome-resolved metagenomics leading to the recovery of genes from novel uncultivated organisms [6]. Simultaneously, the popularity of metagenomics exploded as it became abundantly clear that the human microbiome is strongly correlated with health and markedly changed

in disease states [7]. High-throughput Illumina shotgun sequencing produces only short random fragments of DNA but in vast quantities. Powerful genome assembly algorithms were designed to utilize small repetitive sequence fragments to identify sequence overlaps and establish long continuous sequences, which were coined contigs. The strategies to resolve multiple uncultivated genomes from a random "soup" of contigs included (1) basic sequence alignment to a database of cultured sequence isolates (2) contig-binning based on similar GC-frequencies (3) tetranucleotide frequencies and (4) differential read-coverage. The read-coverage strategy (4) was based on the idea that contigs from the same genome should display a roughly similar sequencing depth. This data-driven strategy represented a powerful concept that many modern binners were later designed to leverage for binning genes or contigs into metagenomic assembled genomes (MAGs)[8, 9, 10]. The initial strategy used to determine when a MAG corresponded to a complete genome was based on the presence of essential bacterial single-copy genes (bacterial markers) [8], which is an approach still used to some extent today by bioinformatic tools like CheckM [11].

The presence of universal markers like the 16S rRNA gene and other single-copy gene markers in bacteria has enabled their identification in metagenomics and fuelled an explosion of known bacterial diversity during the last decades. Thus, the availability of bacterial genomic blueprints in the form of uncultivated bacterial genomes tallies hundreds of thousands across the human microbiome(s), ocean and soil [12]. Meanwhile, the progress of developing databases containing genomes of other biotic constituents like fungi or viruses has been quite different. The reason why viruses were late to the party is due to several technical assembly challenges that will be addressed later, but most importantly they do not contain a universal virus marker gene. For the majority of people, viruses are considered obligate pathogens as we associate them to many types of diseases afflicted by human-infecting viruses, such as the Influenza virus. These types of viruses have represented the bulk of virus blueprints in databases for many years while the focus has been fixed on bacteria (Figure 1). As a result, most metagenomic sequences corresponding to an actual virus have resembled those in the current virus databases. From 2016 to 2018 (Figure 1.1) the number of uncultivated virus genomes exploded in the databases with 750,000 genomes when genomes mined from the first two large virus studies from ocean and soil were released [13, 14]. The success of these expansive viral studies could be attributed to the maturation of *in vitro* protocols for concentrating viral particles, which enabled a greater space for viral assembly and identification.

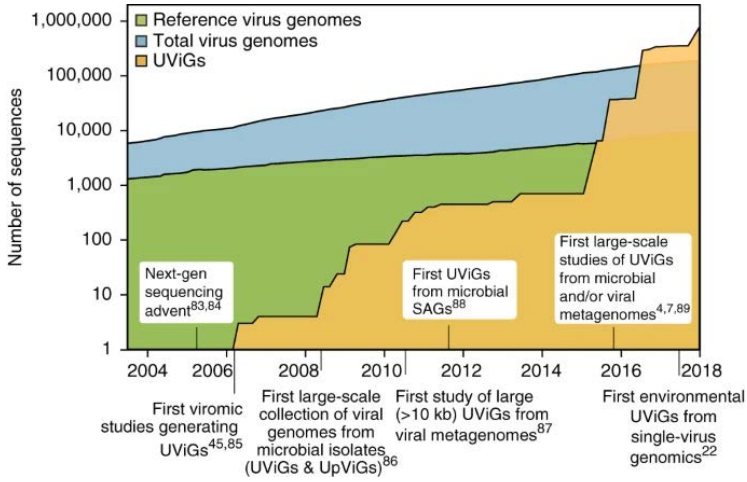


Figure 1.1. The figure illustrates a recent timeline going back to 2004 and the cumulative number of virus genomes uploaded to genome databases since then, including uncultured virus genomes (UViGs). In addition, major events related to virus discovery are notated across the timeline. Virus genomes cultured *in vitro* from isolates are depicted as blue and green, where the green corresponds to reference genomes at ncbi (`https://www.ncbi.nlm.nih.gov/nuccore`). The discovery of uncultured virus genomes (yellow) began in early 2006 and has since then exceeded the number of reference genomes many times over. Figure modified from *Minimum Information about an Uncultivated Virus Genome* [15].

We now recognise that virus particles can be identified almost anywhere where they sometimes massively outnumber other cells like bacteria [16]. To most people, it might be surprising that there are trillions of viral particles around us without the capacity to infect us as they prey on other microscopic organisms like bacteria. The group of viruses preying on bacteria are known as bacteriophages (phages for short). We are now recognising their impact and presence as early studies have unraveled a dynamic and changing virus community during disease like inflammatory bowel disease [17, 18]. Therefore, the search for potential viral culprits has begun in addition to beneficial bacterio-

phages with protective properties. In order to understand how bacteriophages could be implicated in disease as they do not infect humans, we have to consider the way phages interact with bacteria who directly influence the human immune system and metabolism.

1.3 Bacteriophages nature and lifestyle

Viral lifestyle has huge implications on the virus' influence on an ecological site like the human gut microbiome. Bacteriophages bind to a bacterial host and either propagate as a prophage by integrating into the host genome or by takeover of the bacterial replication machinery to produce new virions that are released by exploding and lysing the host [19]. These two canonical strategies represent the lysogenic and the lytic stages of bacteriophages, respectively, yet variations and exceptions to these strategies are continuously discovered (Figure 1.2). For instance, filamentous Inoviruses follow a chronic phage infection cycle and remain productive of new virions without lysing the host, either as an integrated prophage or in an episomal form [20]. Nevertheless, around 90% of bacteriophages described today are dsDNA viruses that follow a lytic lifestyle as a virulent phage or seamlessly switch between the two canonical strategies as a temperate phage [21] (Figure 1.2). In the lysogenic stage, a phage may actively replicate along the host bacteria by being passed down into daughter cells also described as a Piggybacking-the-Winner (PtW) dynamic. In the lytic stage, virus proliferation depends on the presence of a thriving host bacteria that can be infected and lysed for virion production, which is known as the Kill-the-Winner dynamic (KtW). Thus, the two canonical infection strategies have markedly different outcomes for the bacterial host, the right time and place for either strategy will be discussed later. However, not all notable families of phages follow these lifestyle dynamics.
The family of crAss-like viruses make an exception to the above mentioned lifestyle strategies as they do not encode the relevant functional genes for genome-integration, rather it is believed they exist in a carrier state. Even though crAss-viruses are described as intrinsically lytic-viruses they are known to play-nice with bacteria in the human microbiome, which have made them excellent long-term gut residents [22, 23]. Specific crAss-like viruses can even be detected in the same person after several years and thereby persist in the gut for a long duration [23]. In addition, crAss-viruses can make up a substantial proportion of the human virome in some individuals [24].

Whether phages in a community exhibit lysogeny or lysis strategies within an environment is suggested to depend on the spatial structure of the microbial community. In the human gut, commensal bacteria are spatially organized by an increasing gradient from the epithelium towards the lumen (Figure 1.3), hence bacterial abundance is higher in the outer layer of the mucosa. This gradient can be explained by the constant secretion of mucin molecules by the

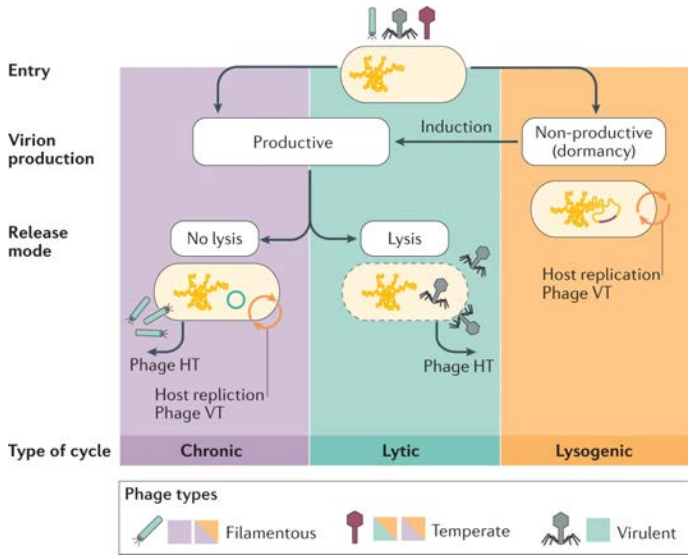


Figure 1.2. The figure depicts three well characterized phage life cycles; Chronic (purple-box), Lytic (teal-box) and Lysogenic (orange-box). (1) Phages in a chronic life-cycle produce new virions (self-replicates) without lysing their entered host-cell. Newly produced virions are transferred out of the host cell by horizontal transfer (HT). (2) Phages in a lytic life-cycle enter a host cell, replicate themselves and release new virions by lysing the host cell. (3) Phages in a lysogenic life-cycle can integrate into the host chromosome and reproduce in a dormant and non-productive state that follows host replication cycles. Dormant phages that are induced can start actively producing new virions by either a chronic or lytic life-cycle. Filamentous phages are able to enter both a chronic and lysogenic life-cycle. Temperate phages are known to switch between lytic and lysogenic or chronic and lysogenic. Virulent viruses are strictly replicating by a lytic life-cycle. Figure from *Interactions between bacterial and phage communities in natural environments* [25]

epithelium creating an highly inaccessible space due to high concentrations of mucin. Thus, the inner layer of the mucosa is virtually deprived of bacteria. The phage concentration peaks around the mid-mucosal layer and falls toward both the epithelium and the lumen [26] (Figure 1.3). Interestingly, high lytic activity by virulent phages in the mid mucosa layer changes to lysogeny driven by temperate phages in the outer mucin layer as the bacterial density and growth rate is highest here. Importantly, lysogeny and PtW seems to be the dominant strategy in the human gut in spaces where the concentration of bacteria is highest, which creates a favorable space for viruses to replicate passively with a bacterial host [27].

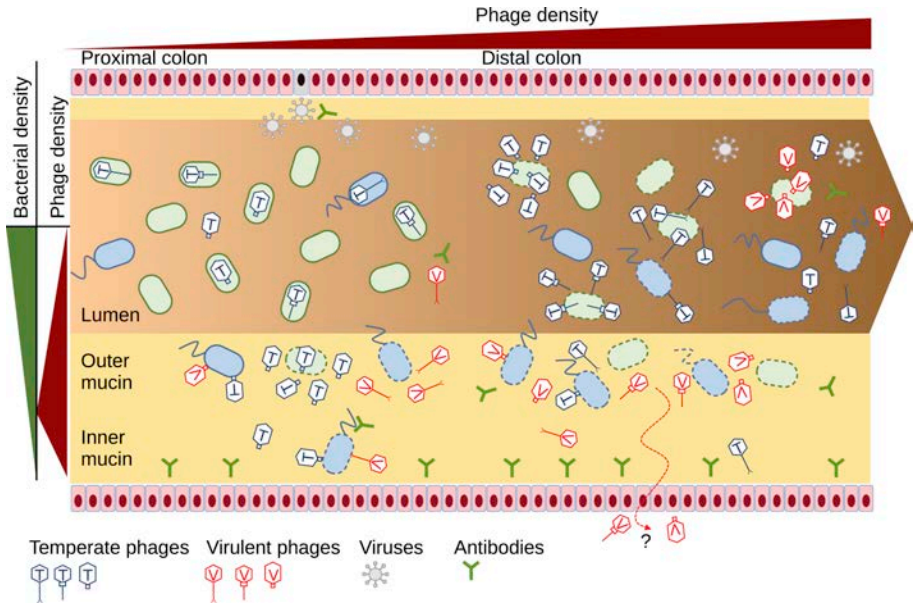


Figure 1.3. A figure illustrating phage and bacterial densities along the human gut colon from proximal to distal. In addition, different viruses and bacteria are depicted across the mucin layer and in the lumen. Generally, phage density increases toward the distal colon. Across the colon, phage density is lower in the lumen where predominantly temperate phages co-infect high densities of bacteria. Conversely, phage density is higher in the mucin where predominantly virulent phages infect and lysis bacteria. Figure from *Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome* [28]

The switch between lysogeny and lytic can also depend on cues from the environment such as environmental stressors like nutrient starvation [29], inflammatory molecules [30] or molecular messages from other phages, which can be sensed by the phage from within the host [31]. Thus, the availability of bacterial hosts and broadly the ecological context matters. Other factors such as the biological age of the human host also have a large impact on the virome community as illustrated by a range of infant gut studies. During infancy from 0 to 2 years of age, the viral community experiences dramatic shifts along the pioneering bacteria. In the early infant gut, a time with low bacterial density, viruses are mostly in the lytic stage and prey on pioneering bacteria leading to a high load of viral particles relative to bacteria, meaning that the virus-to-bacteria ratio is closer to or above 1 [32] (Figure 1.4). The first community of viruses in the infant gut are thought to originate from the pioneering bacteria, thus representing prophages that have been induced from their host [33]. These pioneering phages are accompanied by other lytic viruses acquired from environmental sources. In the subsequent months after birth, the bacterial density increases and higher frequency of integrated prophages lead to higher levels of lysogeny, in which the virus-to-bacteria ratio is reduced as the viral particle load decreases. At this point, the bacterial community contains a higher frequency of strains with integrated phages that protect them from lytic phage predation through superinfection immunity and prophage-encoded defense genes, which altogether increases community stability [34, 35]. Indeed as more phages integrate into a bacterial strain and the number of uninfected hosts are scarce, a molecular message “it is time to lay low” is spread by phages to promote staying in the lysogenic state [36]. Other abiotic factors like transit time may also be involved as it allows for more interaction between viruses and bacterial hosts.

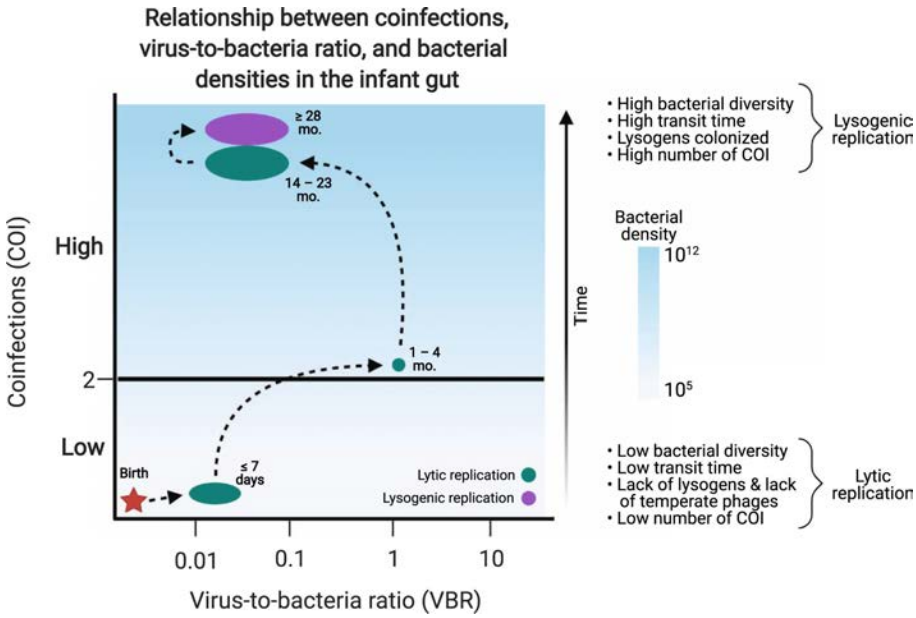


Figure 1.4. The figure depicts the current framework for understanding the relationship between virus-to-bacteria ratio (VBR) and viral co-infections (COI) in the infant human gut from birth. During the first 7 days after birth, the number of viruses is generally low, co-infection is low as replication is mostly lytic. Finally, the VBR is low and bacterial particles outnumber virus particles. Between 1 to 4 months after birth, viral particles start to equal or outnumber bacteria as the VBR reaches 1, which is driven by lytic replicating viruses. After the first 14 months after birth, the number of viral co-infections increases as the bacterial density and diversity is reaching higher levels. Simultaneously, bacteria outnumber viruses reflected by the lower VBR. Eventually, lytic replication is turned into predominantly lysogenic replication. Figure modified from *Phages in the infant gut: a framework for virome development during early life* [32].

As of today, exactly what happens in the adult gut virome and past into elderlyhood remains unexplored. Gregory et al. provided one of the only analyses on this and examined specific viral families like crAss-viruses that thrives into old age [37], albeit the study only included 20 sample points for modeling the virome in elderly (65+ year old subjects), meaning our insight into the age-dependent effect on the human gut virome is still very limited. Nevertheless,

infant gut studies have illustrated that the state of the virome is intrinsically linked with the bacterial community in the infant gut. By the age of 3-6 years, the bacterial density in the gut reaches a stable-community climax that is sustained through adulthood and only significantly disrupted by community disturbances such as changes to diet or medication [38]. Importantly, without major disturbances to the bacterial community, the dominant lysogenic phage community is assumed to be stable and preserved through adulthood. However, more research is needed on this topic as few virome studies have investigated the healthy virome community beyond infanthood.

For the individual phage, it would seem like piggybacking on the bacterial host as a passive passenger or conversely the infection and lysis of the bacteria, would only benefit the phage. However, on a bacterial community level, phages may add to the stability of the bacterial community via antagonistic coevolutionary mutualism or by chronic infection mechanisms.

1.4 Bacterial fitness and bacteriophages

Bacteriophages are shaping the microbial community by phage-mediated diversification. This can be explained by the constant arms race that selects for changes to the bacterial antiviral defenses and viral infectivity to ensure self-survival on both sides [39]. The arms race between phages and bacteria represents a type of antagonistic coevolution, which ensures diversity in both the viral and bacterial populations and is therefore mutually beneficial. In the end, the greatest benefit of the arms race is that a more diversified population stands a greater chance of survival in the case of inevitable bottleneck events that lead to dramatic reductions in ecological populations [40] (Figure 1.5). The phage and bacterial specific mechanisms employed in this neverending arms race are manyfold and will not be described in depth here. Bacteria are equipped with the anti-phage defense systems such as CRISPR-cas, restriction-modification enzymes, variable-surface receptor mechanisms and a range of abortive self-destruct processes during infection [41]. In the "revenge of the phages" [1] to counter anti-phage systems, phages readily modify their receptor-binding proteins to counter changes in surface receptors, encode antitoxins to bacterial self-destruct toxins and anti-antiphage systems to combat restriction-modification and CRISPR-cas systems [42]. The continuous

[1] Yes, this is the actual title of the cited review by Samson et al. 2013 in Nature Rev. Microbiology

evolution of these anti-phage and anti-antiphage systems in the human gut sustains the phage-mediated diversification of the bacterial community. In addition, the maintenance of integrated viruses in a bacterial population may create protection from other lytic phages by superimmunity exclusion but also through phage-encoded hotspots of antiviral weapons [35]. Bacterial populations may also weaponise integrated phages against competing phage-sensitive strains that challenge their position in the environment [43].

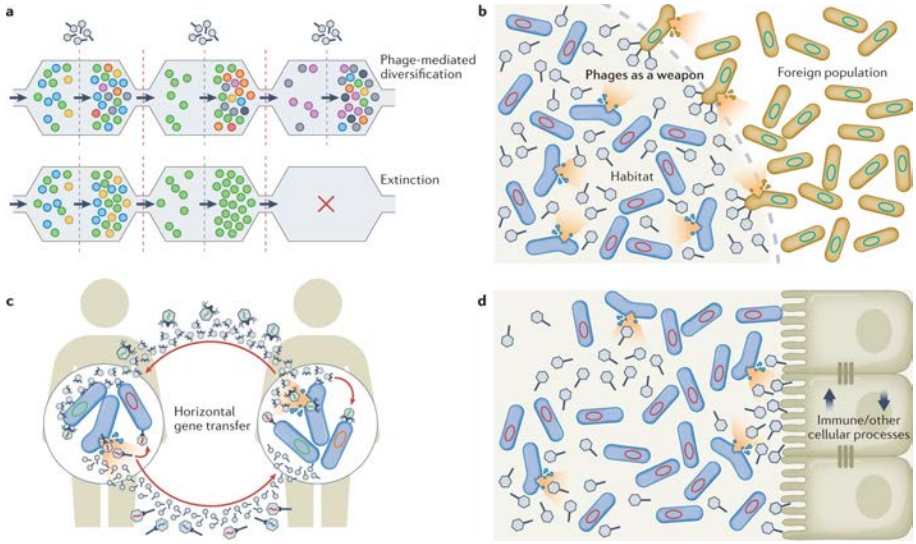


Figure 1.5. Panel (a) illustrates that phage-mediated diversification of bacterial populations (coloured balls) is crucial for greater bacterial phenotypic diversity and protects against population extinction by bottleneck events. Phages can also be weaponized against foreign bacterial populations that contest niches already colonized by bacterial populations (b). In panel (b), integrated phages (red circles) are induced and can lyse bacteria not containing an integrated copy of the phage. Phages can also mediate horizontal gene transfer between bacterial cells in the same environment or between bacterial cells in different environments, such as a different gut microbiome (c). Free phage particles that interact directly with the metazoan host via the epithelium barrier (d) are known to stimulate immune and other cellular processes. Figure from *Mutualistic interplay between bacteriophages and bacteria in the human gut* [44]

Integrated prophages often contain genes that increase the overall fitness of

the host that in return may promote phage survival [45]. Thus, high rates of viral lysogeny provides an extended genetic-reservoir to bacterial populations that protects their position in the ecosystem and in some cases gets them ahead of competitors. A bacterial strain in the human gut, *Bacteroides stercoris*, has been shown to harbor a prophage with an ADP-ribosyltransferase enzyme named bxa that can stimulate the release of inosine from the epithelium, which can be metabolized by its bacterial host (Brown et al. 2021). In addition, emerging evidence is adding another facet to phage-influence on microbial metabolism as integrated viruses encode auxiliary metabolic genes (AMGs) related to carbon, nitrogen and sulfur metabolism [46, 47, 48]. Through the process of horizontal gene transfer (HGT) phages provide a vehicle for effective dissemination of fitness associated genes in the bacterial community such as antibiotic resistance genes or virulence factors [49]. Bacterial pangenomes are a result of the fact that no single species can carry the entire genetic burden of all possibly beneficial accessory genes, which might be relevant in some instances [50]. Altogether, phages and other mobile genetic elements (MGEs) provide the machinery for genetic exchange of beneficial accessory genes between bacterial species in an environment. Phage-host specificity ensures the dissemination of relevant and useful accessory-genes that stay within the pangenome of a bacterial species, meaning that primarily bacteria of the same species will be able to access the same genetic bazaar.

Phage-influence on the bacterial community is profound and may also affect the human host in an indirect-manner as they influence the bacterial composition and the collective bacterial metabolism. However, recent investigations have provided new insights into mechanisms in which phages may also influence the gastrointestinal immune axis, both indirectly through bacteria but also directly as viruses come into contact with the human host barrier.

1.5 What do gut viruses mean to humans?

The intestine harbors a large body of immune cells that is continuously bombarded with immune stimuli by microbiome residents [51]. Commensal bacteria produce antigens that are recognised by immune cells in the gastrointestinal tract, which may be tolerated during homeostasis or initiate an immune response during inflammatory conditions [52]. These antigens are recognised by pattern recognition receptors (PPR) on intestinal epithelial cells (IEC) such as the Toll-like receptor (TLR) or NOD-like receptor [53]. In addition, the IEC constitutes the first line of defense acting as a self-maintaining physical barrier. The IEC also provides a second layer of defense against opportunistic pathogens in the gut, as it harbors goblet and paneth cells that together pump out a glycoprotein-rich source loaded with antibacterial peptides that forms the mucosa layer [54]. The mucosa also feeds clades of bacteria that further provides a living wall that hinders other bacteria in crossing the mucosal layer [55]. Briefly put, the IEC and mucosal barrier physically and chemically prevents translocation of bacteria to the underlying tissues. The state of the gut microbiome during manifestations of inflammatory bowel disease (IBD) such as Crohn's disease (CD) or ulcerative colitis (UC), is characterized by bacterial perturbations in the microbiome that affects the integrity of the mucosal layer and leads to higher exposure of the microbiome residents to the IEC [56, 57]. The weakening of the mucosal layer is a complex multi-facetted result and is not only explained by bacterial perturbations, nevertheless results in higher recruitment of immune cells to the epithelium barrier and increased immune activity to various bacteria and organisms now in closer contact. In addition to bacteria, viruses are also recognised at the exposed epithelium.

The DNA of eukaryotic viruses such as papillomaviruses, polyomaviruses and herpes viruses are also recognised by multiple of the PPRs located on immune cells and the IEC [58]. Infection of the epithelium by eukaryotic viruses causes the release of pro-inflammatory cytokines leading to inflammation [59]. In addition, phages in close contact with various types of mammalian cells can also be sensed and uptaken in some tissues thereby interacting with the host immune system [60]. As the colon epithelium in the healthy gut is protected by the mucosal layer, phage uptake is less likely in the gut epithelium but could be dramatically increased by a thinning mucus layer due to inflammation [61]. Thus, a sudden rise in abundance of a specific phage during IBD may aggravate the disease state. Do phages only provoke a pro-inflammatory

immune response? Recent research on viruses sampled from colon resections have shown that they exhibit a divergent immune stimulatory effect dependent on the state of the microbiome from which they were sampled. Viromes sampled from IBD colon resections stimulate a pro-inflammatory immune response through host immune cells while viromes from non-IBD colon resections stimulate an anti-inflammatory response [62]. In addition, virome depleted mice are protected against the inflammatory effects of Sodium trimethylsilylpropanesulfonate (DSS) when gavaged with non-IBD viromes while supplementation of UC or CD viromes aggravates an inflammatory state [62]. These experiments echo observations on human immune dendritic cells pulsed with UC and healthy viromes; only the former set of dendritic cells were able to activate CD4+T cells and induce INF-γ [30]. However, the ability of viromes to influence and aggravate a specific phenotype requires the presence of host-immune recognition receptors for immunomodulation.

Since phages are capable of driving a specific immune response, it is appealing to consider if phage populations can be identified and concentrated to either boost an anti-inflammatory environment or provide resistance to pathobionts like C. difficile [63]. Similar to the mechanism in which IBD-viromes transferred to a recipient may aggravate inflammation via fecal viral transplantation (FVT), viromes extracted from mice fed a low-fat diet and transferred to obese recepients can improve on markers of type II diabetes and obesity [64]. In a FVT study by Rasmussen et al., the virome diversity was not significantly different between mice fed a low-fat or high-fat diet, however, in mice fed a high-fat diet + FVT, the bacterial community in recipient mice was shifted to increasingly resemble the low-fat bacterial community [64]. Thus, phages influence host-physiology indirectly by restructuring of the microbial community. On the gut-brain axis, specific phages have been linked to improvement of memory and cognitive function via fecal transplantation by possibly promoting a range of lactic acid bacteria and their metabolism [65].

In terms of health and disease, improved viral discovery and analysis is important for delineating viral populations that aggravate a disease or conversely provide protection against it [66]. In addition, the development of homogenous bacterial and viral populations in the human gut represents a fragile environment. Transplantation of viruses and bacteria from a healthy gut may represent a key strategy for restoring the heterogeneity and stability of the gut community.

1.6 Maximizing the discovery of biological diversity in bulk metagenomics

The current state of virus identification and discovery is fueled by the product of *de novo* assembly of sequencing reads from metagenomic samples. Assembled contigs are categorized as viruses by sequence alignment to known reference databases or labeled viral based on i.e. enrichment of virus proteins. However, no universal marker exists for all viruses, thus virus annotation is limited to the space of known virus sequences. Furthermore, assembly of metagenomic sequencing reads into reconstructed phage genomes is challenged by (1) high levels of strain diversity and micropopulations, (2) genetic mosaicism and hypervariable regions (3) genomic repeats and (4) non-uniform sequence coverage [67], which altogether leads to higher degrees of fragmented viral genomes. One of the solutions to tackle these assembly challenges has been to perform viral enrichment, also known as virus-like particles (VLPs), which concentrates the amount of viral genetic material in a metagenomic sample prior to sequencing. The key idea of viral enrichment is to reduce the genetic noise from larger organisms such as bacteria and amplify virus reads to enable discovery and improved assembly of viral diversity (Figure 1.6). The removal of bacterial cells and DNA for VLP preparation can be performed in different ways including centrifugation, size-filtration and nucleases. Although the filtration method of choice does not seem to affect the total number of virus contigs discovere [37]. Nevertheless, metagenomic assemblers can be helped to some extent by viral enrichment to produce more reliable viral assemblies [67].
While viral enrichment methods have been improved to yield better viromes despite low volume limitations they are far from trivial, require optimized protocols, and do not capture the full breadth of viruses. Finally, viral enrichment adds an additional sample-preparation and sequencing step to the process and additional significant costs to a microbiome study [68]. Identification of viruses without enrichment, corresponding to searching for viruses in bulk metagenomes, overcomes the size-filtration step biases and enables identification of temperate and lytic viruses. Yet, the viruses identified with this approach may be biased towards viruses infecting dominant host cells and may miss out on rare viruses [15]. In addition, the technical virus assembly challenges are amplified without the filtration step and results in fragmented virus genomes. Recovery of virus genomes in bulk metagenomics could be significantly improved by the process of binning where the puzzle of fragmented virus sequences is solved computationally.

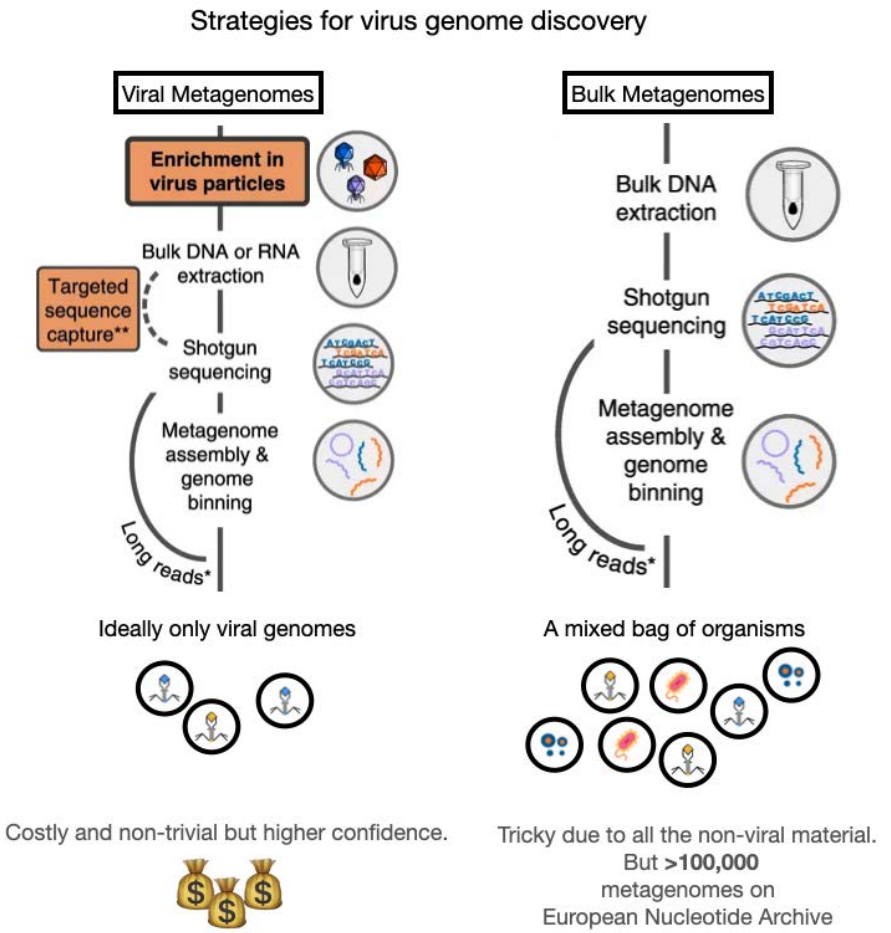


Figure 1.6. Depicted are two of the primary ways of generating uncultivated virus genomes from metagenomic sequencing, viral metagenomes and bulk metagenomes. The major step distinguishing the two methods is the presence of viral particle enrichment that filters out larger organisms like bacteria and eukaryotes. In both methods, extracted DNA or RNA are shotgun sequenced and assembled into contigs that can be binned into complete genomes. However, binning of contigs from bulk metagenomes produces a mixed bag of organisms including bacteria, eukaryotes and viruses. Viral enrichment ensures that more viral genomes are assembled with higher confidence, though it is a non-trivial and costly preparation step. Publicly available bulk metagenomes are in abundance but the mixed genomes require that viruses have to be isolated in silico. Figure modified from *Minimum Information about an Uncultivated Virus Genome* [15]

2 Methods

In this section, I will provide an overview of the most important bioinformatic concepts and methods for studying the known and the unknown microbiome in metagenomics today.

2.1 Metagenomic binners

Continuous discovery of novel biological diversity is currently very dependent on (1) metagenomic assemblers that can stitch together sequence reads into contigs and (2) metagenomic binners that resolve the contig puzzle produced by assemblers into whole genomes. The latter method has been crucial in recent expansions of novel biodiversity from metagenomics [12]. Metagenomic binners like MetaBAT2 [10], MaxBin2 [69] or Canopy [9] provides a computational framework for accurately clustering metagenomic assembled sequences into MAGs, an approach that is both efficient and most importantly reference free. Without metagenomic binners, the origin of every sequence would have to be determined by alignment, corresponding to matching each sequence with all blueprints of known diversity in the reference databases. Alignment faces two important drawbacks; (1) not all genomes are in the databases and (2) even if a sequence matches a reference genome, it might not be the same genome but a related one. So how do binners tell if two sequences originate from the same organism? Binners are primarily based on sequence composition and co-abundance clustering. In *sequence composition binning* (Figure 2.1), each nucleotide sequence is broken down into subsequences by a sliding-window of k nucleotides; subsequences are then enumerated. In this way, a k-mer spectra is determined for each sequence by calculating the frequency of each tetranucleotide. Species-specific signals in the tetranucleotide frequencies can then be used to connect contigs by origin [70] such that sequences with similar tetranucleotide frequencies are grouped together.

Co-abundance binning is based on the principle that sequences of the same genome should be present at similar levels of sequencing depth (Figure 2.1). For instance, if a genome is found twice in a sample and the assembler returns two different chromosomal fragments of that genome, the abundance of these sequences should equal the abundance of the genome, namely two. If the sequences of a genome are recovered across multiple samples, the sequence's abundance should be similar within each sample despite the genome being found in different levels across samples. This is unlikely to happen by sheer chance if the sequences are from distinct genomes and this is the co-

abundance principle, which can be leveraged for grouping genome fragments together. In practice, the contig abundance or depth (how many times a contig is covered by a read normalized by contig length) is calculated for each sample producing a Contig-depth x Sample matrix, which can be used as input to a metagenomic binner to group contigs displaying similar depth across samples. MetaBAT2, MaxBin2 and VAMB (featured in this dissertation) features computational models that utilize both sequence composition and co-abundance during binning.

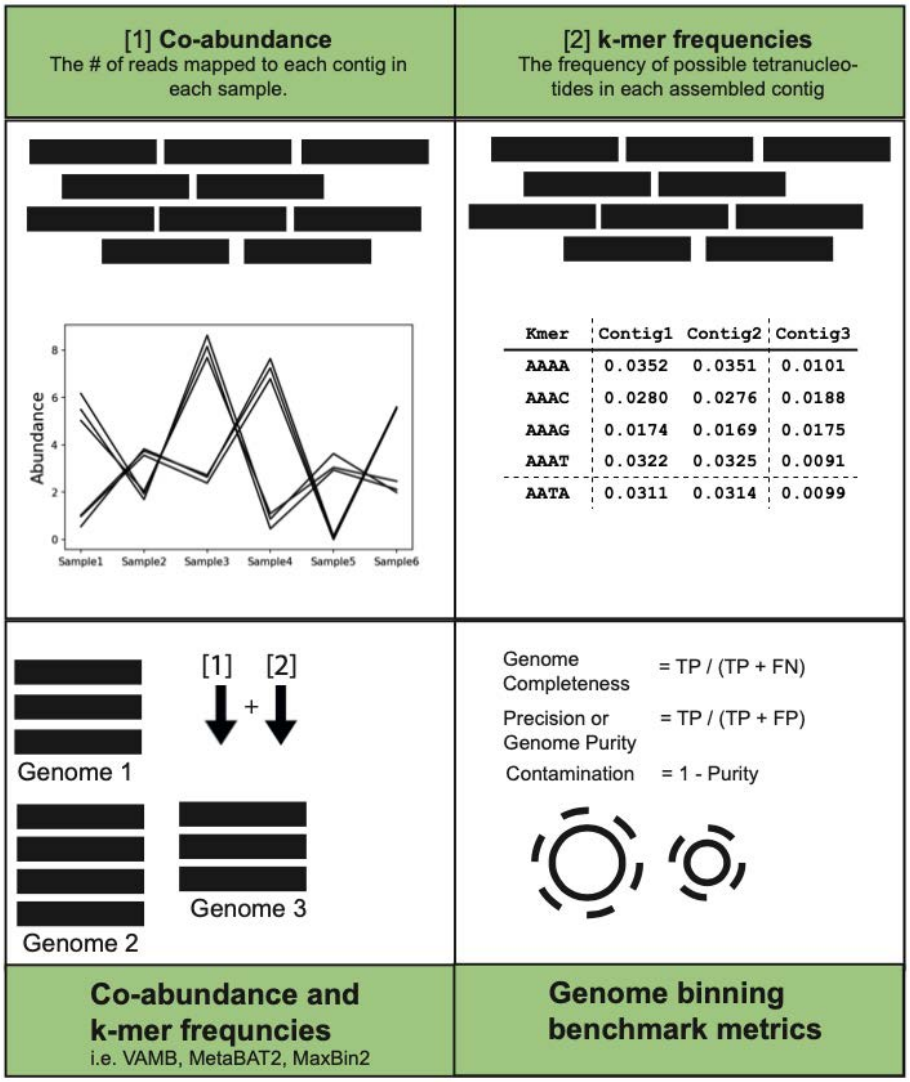


Figure 2.1. Panel 1 and 2 conceptualize the information provided to modern binners such as [1] co-abundance i.e. the number of reads (normalized by contig-length) mapped to each contig across samples and [2] tetranucleotide frequencies calculated in each contig. These two continuous data inputs can be combined to bin input contigs into genome-bins. Standardized metrics for evaluating a binner's accuracy and overall performance includes genome completeness, purity and contamination (1-purity). These metrics are computed for each metagenomic bin based on the closest match in the genome test-set. TP is the number of base pairs that are matched between the bin and genome, FP is the number of basepairs that matched other genomes and FN is the number of basepairs in the genome assigned to other bins. .

The performance of a computational binner can be assessed using standardized metrics and evaluation datasets from the CAMI consortium [71] (Figure 2.1). The metrics include genome completeness and genome purity, which is also described as recall and precision. Here, the validation datasets composed of chromosome fragments (contigs) labeled to their corresponding genome, allows computing the number of correctly and incorrectly assigned base pairs after binning contigs into genomes. A technology that may either improve binning or relieve the use of it is improved sequencing methods. Long-read sequencing methods, also known as 3rd generation sequencing, from Oxford Nanopore Technology (ONT) and Pacific Biosciences (PacBio) produce much longer sequence reads, which can be more efficiently and confidently merged into complete genomes due to higher sequence overlap. While 3rd generation methods have been maturing in terms of sequencing accuracy and price [72], 2nd generation sequencing with the Illumina Miseq as an example, has been the most practical and cost effective option to investigate the human microbiome by deep sequencing. As a result, the demand for metagenomic binners remains high as 2nd generation sequencing is still applied to metagenomic samples. However, even with the perfect binner, not every reconstructed genome can be readily taxonomically annotated and described in the context of known diversity. Thus, downstream delineation of microbial constituents depends on methods for annotating every genome produced by binning.

2.2 Annotation tools for microbiome residents

Bacterial genomes encode taxonomically informative markers like the 16S rRNA gene, which contain evolutionary conserved information and represent an important genomic trait for delineating prokaryotic species. The number of taxonomically informative bacterial markers has grown beyond the 16S rRNA gene and given rise to a number of tools for determining bacterial genome completeness and contamination like CheckM [11]. In addition, these markers have also created a basis for establishing the taxonomic identity of genomes by phylogeny aware placement with tools such as GTDB-TK [73]. These tools leverage the growing number of available genome sequences to refine the taxonomic annotation accuracy [74]. However, computational tools for determining the taxonomic identity of genomes beyond the bacterial tree of life are further behind in development.

Virus genomes in metagenomics have traditionally been difficult to annotate as they do not contain universal genetic viral markers, such as the 16S rRNA gene in bacteria. Viruses do encode a number of hallmark genes that encode structural proteins such as the capsid, tail or head, however, many of these viral-like genes can also be found in bacterial genomes as a product of genetic exchange from bacterial-virus interactions [75]. This inevitably creates uncertainty when evaluating sequences that fit into the viral-ish genome size. Virsorter was one of the first reliable and appreciated bioinformatic tools to provide a viral sequence probability score for a given sequence [76]. Virsorter deduces a viral confidence score as it slides through a sequence's gene-content and enumerates the number of viral, bacterial and uncharacterised genes. Then, based on known frequencies of gene-groups such as capsid, head or tail, it determines enrichment of viral-genes and depletion of bacterial affiliated genes. Basically, sequences are determined viral with higher confidence scores if they primarily contain proteins frequently annotated by viruses in reference databases. Thus, the validation of *bona fide* viral sequences is to some extent reference dependent despite the lack of a solid marker-gene and as the databases of uncultivated viruses grow the genomic-characteristics of virus genomes are increasingly refined. Other computational tools like Virfinder [77] produces a probability of viral-ness with less emphasis on gene content using machine-learning that have learned k-mer frequency patterns of known viruses and perform comparably to Virsorter.
The application of virus sequence predictors on metaviromics has been imper-

ative to grow the databases of uncultivated viruses that can be recycled for making better predictive tools. The recent CheckV tool takes advantage of the expanded virus databases and performs sequence evaluation by annotating viral genes and calculating amino acid similarity based on protein-coding genes to known viruses [78]. Thus, CheckV represents the first framework for massive quality-control of viruses to current databases of uncultured viruses combined with cultured viruses stored in the NCBI virus database (Figure 2.2). Computational frameworks that bridge the compilation of viral biodiversity with putative virus sequences helps to establish known and possible novel diversity.

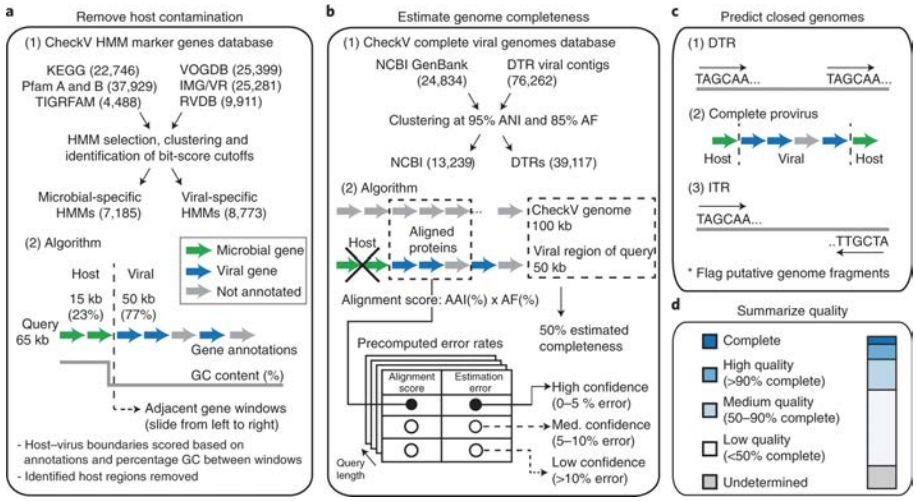


Figure 2.2. CheckV performs quality control and completeness estimates of sequences by annotating every gene in an input sequence to microbial and viral specific gene databases (a). The gene-based similarity of an input sequence to viruses in databases are then deduced based on viral-like stretches of sequence, while contaminating sequences of host-bacteria are removed (b). If the input sequence contains terminal repeats, which indicates a circular sequence, the genome is determined complete (c). If not, the similarity and size of the viral input sequence relative to reference-viruses are used to determine a completeness and quality score (b,d), Figure from *CheckV assesses the quality and completeness of metagenome-assembled viral genomes* [78]

The viral database boom and suite of viral prediction tools available have enabled virus sequence annotation with higher confidence (although limited by

the content of the databases) and provide an estimate of completeness in the format of Complete, high, medium and low genome quality. Before CheckV, VIBRANT was the first tool to provide a similar functionality and is still widely used [79]. Yet, the methods for predicting whether a sequence is viral or not does not provide a lot of information on virus biology. Next we will address methods to establish the viral host.

2.3 Connecting bacteria and viruses

2.3.1 Proviruses

The simplest approach to connect bacterial and phage populations is based on genomic evidence that the phage is integrated into the bacterial genome and exists as a provirus. Numerous computationals tools have been developed for identifying integrated phages, including CheckV and VIBRANT. These tools scan for viral sequences sandwiched between chromosomal sequences and identify the starting and end points of integration. Confidence in an annotated provirus is further increased by (1) the viral sequence being flanked by direct terminal repeats, which also indicates a complete circular virus genome, and (2) the presence of genes in the virus sequence encoding integrase enzymes that can facilitate the integration. Host-information for a single provirus can with some exceptions be transferred to similar viruses identified in a similar environment like the human gut. The host-range for many viruses is quite narrow. Viruses of the Podoviridae family are suggested to have a narrow host-range while Myoviridae viruses exhibit a broader host-range [80, 81]. There appears to be a trade-off between host-range and virulence competence; a very broad host-range may be detrimental to virus infection efficiency and helps to explain the higher prevalence of a narrow-host range observed in viruses which are specialized to constantly evolve and infect similar hosts [80].

2.3.2 Prokaryotic defense systems

The host-affiliation of freely assembled viruses is harder to determine. Fortunately, the discovery of prokaryotic defense systems such as clustered regularly interspaced short palindromic repeats (CRISPR) has greatly facilitated phage-host mapping. CRISPR systems serve as prokaryotic immune systems with a memory of past-infections by phage or other foreign mobile genetic elements (MGEs). The CRISPR array archives sequences from phages as sequence pro-

tospacers, which immunize against subsequent infections [82]. The CRISPR system leverages protospacers to recall, target and destroy a returning invader. Importantly, this memory-based immune system also provides a genomic context for learning the host-range of phages within an environment. As CRISPR spacers can be mined from bacterial isolate genomes and MAGs using a variety of computational CRISPR identifiers such as CRISPRcastyper[83], interacting phages can be identified by matching the protospacer sequence to phage genomes. However, protospacer matching may also result in false-positive interactions or conversely poor recall. If a protospacer matches a virus genome with a weak hit (corresponding to an impartial match) or too stringent hit criteria (perfect match only), false-positive or few connections between bacteria and viruses will be found [84]. The guidelines for using CRISPR spacers to associate phage and bacteria with maximum recall and precision is a work in progress, but initial benchmark studies have defined thresholds such as maximum two mismatches across 90% of the spacer-sequence, which yields up to 49% recall with 69% precision at genus level [85]. By applying the suggested thresholds for protospacers with a typical length of >20 nucleotides, the host range of phages to bacteria can be established with high confidence by using the natural prokaryotic immune system.

2.4 Predicting virus lifestyle

Virus genomes that do not encode the required genes for integrating into a bacterial host are by definition virulent phages as they are incapable of integrating into a potential host. Bioinformatic tools like BACPHLIP and Deephage provide computational methods to predict the viral lifestyle, temperate or virulent, as they search for the relevant genes necessary for integration [86, 87]. Deephage represents a convoluted neural network that has been trained to distinguish lytic phages from temperate phages based on local sequence features. BACPHLIP performs alignment of virus genes to a database of lysogeny marker genes. If no marker genes are detected, the outcome for both tools should be that the input genome is virulent as it does not encode the genetic machinery to facilitate integration. Therefore, completeness of a viral genome is an important variable for a confident prediction. If the genome is 100% complete and no lysogeny markers found, it must be virulent. If the genome is incomplete and no lysogeny markers found, the lifestyle cannot be determined as the missing genome could contain the genes. Conversely, if any lysogeny markers are found the virus is very likely an incomplete temperate phage.

2.5 Machine learning

Machine learning (ML) is a subgroup of artificial intelligence (AI), which encompasses a long range of computational models that learns from high quantities of data to classify and recognise patterns. The learning is either directed in a *supervised* way with labeled data structures or in an *unsupervised* way where the system aims to identify patterns without a predefined structure or truth. In both cases (supervised or unsupervised) the model is working to minimize a loss or cost function that captures the deviation of the predicted or generated output relative to the ground truth or input.

Within the realm of microbiome research, ML models have been applied in several cases for the purpose of classification. ML have been developed for host phenotyping using the microbiome composition and bacterial species abundance as the only information and can successfully stratify patients based on their microbial signature [88, 89]. Some ML models also include functions for assessing feature importance on model performance, which have been used to identify discriminative bacterial strains that exacerbate a disease phenotype [88]. The features represent measurable or categorical units in the dataset, such as the abundance of bacterial species or whether the sample is from a control or case patient. Microbial features can be combined with other omics such as metabolomics, gene-expression or host clinical data to increase the models ability to differentiate phenotypes [90]. The choice of ML method to use for each application largely depends on the data, method preference and importantly whether or not the data is complete with truth-labels to facilitate supervised learning of the model. Methods that have been used for supervised learning in microbiome studies are manyfold and include logistic regression, Linear Discriminant Analysis, support vector machines (SVMs), naive bayes classifiers and artificial neural networks [91]. For one of the publications in this dissertation, we leveraged the Random Forest (RF) model.

2.5.1 Random Forest

The RF model is an ensemble of multiple decision trees in which the performance is evaluated based on pre-labelled data. Each individual tree in the RF makes a class prediction and the majority vote across all trees becomes the final prediction, thereby employing the “wisdom of the crowds” instead of relying on a single classifier. For a given training dataset with x number of observations

characterized by y variables, the RF model constructs a prespecificied number of random decision trees (Figure 2.3). The randomness comes into play as the y-variables used to describe each observation are randomly sampled at each node in the decision trees.

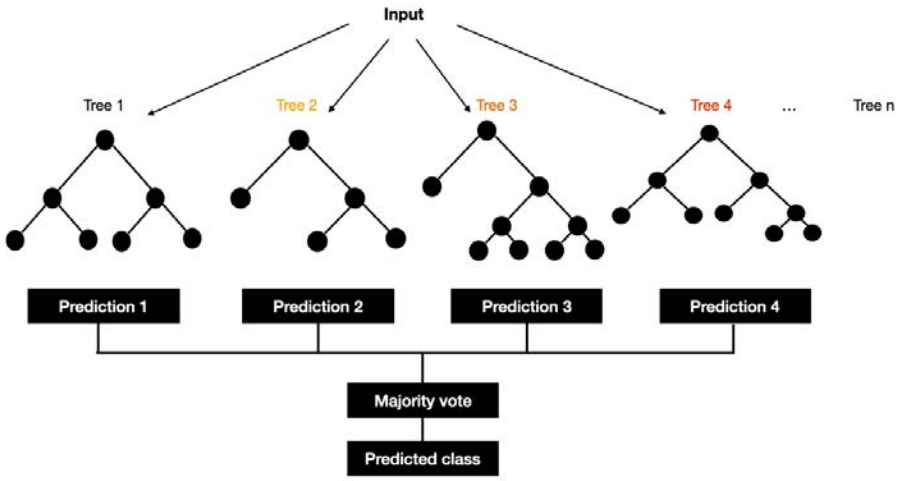


Figure 2.3. The Random Forest (RF) contains a predefined number of decision trees, which are randomly constructed resulting in n independent decision trees. A trained RF model works by receiving an input that is processed by each decision tree in the forest, which produces an independent prediction of the class based on the input. The final prediction is based on a majority vote across all trees to produce a final prediction.

The RF model is evaluated using a method called bagging as it randomly samples with replacements from the observations in the dataset. This also means that some observations are not sampled and therefore "bagged" in the out of bag (OOB) set, which can be used to test the accuracy of the final tree ensemble. Based on the OOB observations, an OOB-error is derived and provides an immediate estimate of the RF model's accuracy. When training other supervised ML methods like linear regression classifiers or SVMs that do not employ OOB, k-fold cross validation (CV) is an ideal approach for testing the accuracy of the model during training and ensuring that the accuracy is calculated based on observations not used in estimating the parameters of the

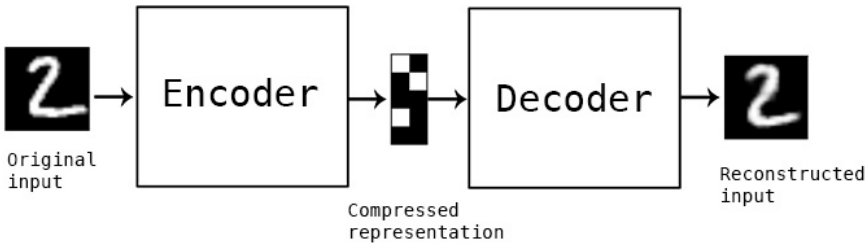


Figure 2.4. A classic example of Autoencoders is the usage for reconstructing images. In the example depicted, the original input (image of a "2") is encoded into a compressed representation and decoded into an output very similar to the original image. Illustration from: https://blog.keras.io/building-autoencoders-in-keras.html

model. With CV, the observations are split into k number of partitions; then the training is performed k times using one partition of the observations as the test dataset and the rest as training data. To achieve a final estimate on whether the ML model generalizes to new observations, the most important estimate of accuracy should ideally be calculated based on an independent dataset. An overfitted model may produce good results on the training data set but underperform on real-world data points.

2.5.2 Variational autoencoders

One of the artificial neural network methods that has received increased attention is the autoencoder that has proven useful for capturing structures in high dimensional data with many features per observation such as single-cell RNA-seq and multi-omics data [92, 93]. Autoencoders are designed to receive an input and encode it into a compressed representation coined the latent representation, which can be decoded to reconstruct the input. The autoencoder is trained to minimize the difference between the original input and decoded output by minimizing a loss function that captures this difference numerically (Figure 2.4).

Autoencoders consist of a chain of artificial neural network (ANN) layers with a distinct architecture in which the first layer (that receives an input) has the same dimension of the last layer (which recreates the input). In the process of encoding a high-dimensional input y into a low-dimensional compressed repre-

sentation z, which can be decoded into an approximation of the original input $\hat{y}$ (Figure 2.5), the network can capture important nonlinear structures in a dataset. The property for turning high-dimensional data into a low dimensional compressed representation z makes the autoencoder a dimensionality reduction method as it learns to store all the relevant information in a few explanatory variables. An issue with the traditional autoencoder architecture is the fixation on encoding and decoding an input with as little loss of information as possible, which can lead to overfitting. If the autoencoder is overfitted, the encoder may map any data point to an arbitrarily small segment of the latent space z as real numbers where the decoder can still recreate the input but the position in latent space is structurally meaningless. The autoencoders' mapping of an input y into the latent space z can therefore become completely arbitrary. This would effectively make the autoencoder produce gibberish when exposed to new data points. In addition, arbitrary mapping of input into latent space makes it impossible to cluster the latent representation z into anything meaningful if unrelated data points are positioned close in latent space.
With Variational autoencoders (VAE), the training is performed using a regularization technique to avoid overfitting and enforce structure to the data points in latent space. The key is to ensure that data points that are close in the input space are also close in the latent space, which makes clustering of the latent space more feasible. To achieve this, the input y is encoded as a Gaussian $\mathcal{N}(\mu, \sigma)$ distribution with a mean μ and standard deviation σ by a function $E(y)$ 2.5), instead of a low-dimensional data point. From the latent distribution $p(z)$, a latent representation z is sampled and decoded into $\hat{y}$, which makes the VAE a generative model.

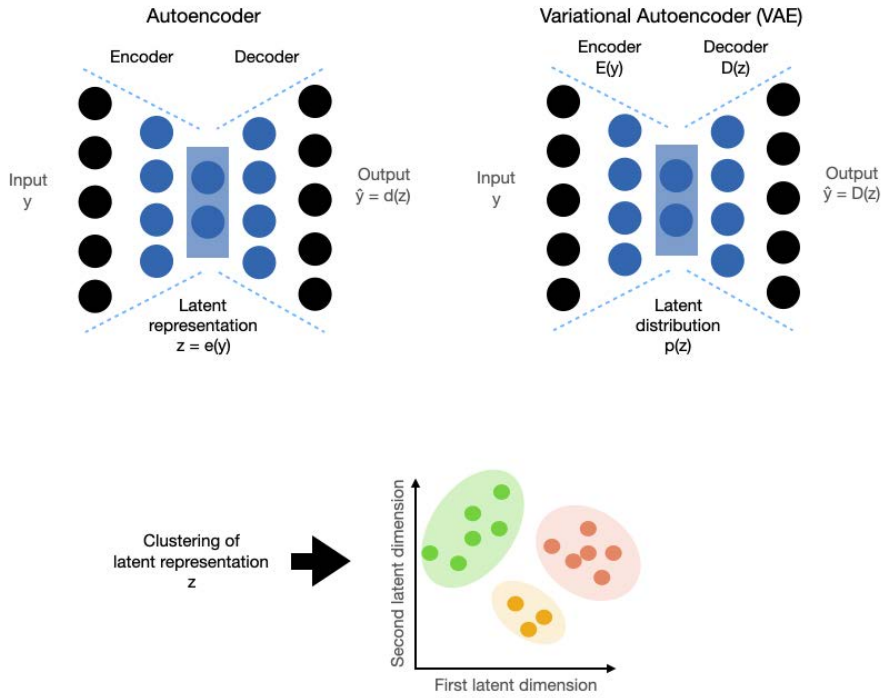


Figure 2.5. The Autoencoder and Variational Autoencoder (VAE) may have similar architectures as they both contain an input layer y and output $\hat{y}$ layer with same dimensions, an encoder and decoder layer and a latent representation (center layer). However, the VAE encodes the input into a latent distribution $p(z)$ function that can be sampled from to reconstruct the input. Both methods can produce a latent representation z for a group of high-dimensional data points that is more suitable for clustering and can be visualized in a lower dimensional space.

For training the VAE, the loss function is composed of two terms [94]; (1) a reconstruction error that represents the difference between the original input y and the decoded output $\hat{y}$ and (2) a regularization term on the latent layer that enforces the encoder to return a gaussian distribution $\mathcal{N}$. The regularization term is expressed as the Kulback-Leibler (KL) divergence function that captures how well the encoded Gaussian distribution approximates a standard Gaussian distribution $\mathcal{N}(0, 1)$ where μ is 0 and σ is equal to 1. These two terms

together make up The Evidence Lower BOund (ELBO) function also described as the variational lower bound. The ELBO can either be represented as the expected negative log likelihood plus the KL divergence, where the ELBO loss function is minimized during training of the VAE.

$$ELBO(y) = -\mathbb{E}_E[log(p(y|z))] - D_{KL}(E(y)\|p(z)) \tag{2.1}$$

Or as an expression where the ELBO is maximized by minimizing the KL divergence while maximizing the expected log-likelihood.

$$ELBO(y) = \mathbb{E}_E[log(p(y|z))] - D_{KL}(E(y)\|p(z)) \tag{2.2}$$

Ultimately, what we want to achieve using a VAE is to map similar values closely in the latent space. Compared to an ordinary autoencoder trained to decode arbritary data points into near exact recreations of the input, in the VAE the decoder samples from encoded Gaussian distributions of the input, which results in similar recreations of the input instead of near exact. The VAE's capacity for encoding high-dimensional observations into meaningful low-dimensional numeric vectors without any prior labeling of observations makes it an unsupervised and powerful method for dimensionality reduction and clustering. Biological samples from patients or natural environments are characterized today with thousands of features produced by multi-omics, thus the VAE is getting more recognised as an appropriate tool for differentiating anything from cell types to human pathologies.

3 Research objectives

In **Chapter 1**, I have provided a brief and non-exhaustive overview on how bacteria and viruses are studied in metagenomics. In addition, I have provided a basis for understanding why the interaction of bacteria and viruses can be profound to an environment like the human gut microbiome. Current and future metagenomic studies like the "Million Microbiome of Humans Project"[1] do not provide metaviromes as it would double the sequencing efforts and the costs. The development of methods for extracting the virome from bulk metagenomes are important and represent a desired approach to facilitate virome analysis in future metagenomic studies. We set out to develop and explore methods to face the lack of metaviromes and enable investigations into ecological hubs of viruses and bacteria in microbiomes. Based on ongoing work related to metagenomic binners, we chose binning as our starting point and have worked toward generating both bacterial and virus genomes from bulk metagenomics with this technique.

To pursue a framework for discovery of viral diversity in bulk metagenomics, I defined the following challenges:

1. Establish or adapt a method to bin virus genomes in parallel with bacteria and estimate the success of virus recovery relative to metaviromes.

2. Benchmark the framework from (i) for large scale discovery of viruses and bacteria in metagenomics, define how much viral diversity is captured and what diversity is missed.

3. Investigate and re-analyse gut viromes in cohorts without metaviromics to expand our understanding of viral and bacterial hubs in different cases of health and disease.

To expand on **Objective 1**, we planned to leverage paired metagenomics and metaviromics to evaluate and tune viral genome recovery in bulk metagenomic samples. We proposed to use the generative VAE model for phage binning, which was an ongoing project in the group. For developing and training the phage specific binning algorithm we used paired human gut microbiome and metavirome datasets from the COPSAC and Diabimmune (T1D) cohorts, where COPSAC is by far the largest paired metagenomic and metaviromic dataset produced as of this date. Using paired bulk metagenomic and metavirome data is crucial as the metavirome serves as a gold standard and corre-

[1] news.ki.se/first-project-to-create-atlas-of-human-microbiome

sponds to an estimated truth of the actual viruses in the environment. Importantly, the metavirome allowed us to define and annotate the presence of viral specimens that can be recovered in the corresponding complex metagenome. By exploring the intersection of virus genomes recovered from bulk metagenomics and metaviromes we could also estimate the degree of shared viruses from the two methods. This estimate is important to challenge current assumptions on the technical biases introduced by viral-enrichment, which selects for specific parts of the gut virome diversity such as virulent viruses at the expense of proviruses and temperate viruses [15].

Objective 2 was tightly connected to Objective 1 and involved large scale application of the framework to benchmark our methods in terms of the number of viruses recovered in metagenomic datasets. This included assembly, binning and viral identification, followed by quality and completeness estimations. As we had access to several metagenomic cohorts such as COPSAC, Diabimmune and HMP2, the virus genomes discovered as part of the objective could be leveraged for downstream microbiome community analysis.

In **Objective 3**, the aim was to apply our methods to published metagenomic cohorts and reanalyse datasets with a focus on viral and bacterial community analysis. Such analysis can provide an additional explanatory virome facet to groups of distinct microbiomes, which were originally investigated on the basis of bacteria only. Furthermore, insightful analysis on the bulk metagenome-derived virome may also serve as landmarks for future virome analysis. Viruses represent additional variables in microbiomes, association of viral communities to a phenotype of interest like a clinical variable can be applied to outline specific viral hubs of interest. As an example, we searched for viral-clades sustained in the microbiome of progressive IBD patients from the HMP2 IBD cohort. Knowledge about phage persistence and bacterial dynamics in the human gut microbiome may be used for developing diagnostic or medical therapeutic agents for different pathologies. In addition, we investigated the age-dependent effects on virome communities and bacteria assembled from a study of Japanese centenarians. This analysis helped to outline virus and bacterial hubs abundant in centenarians, which might be implicated in healthy aging and extreme longevity. In addition, we conducted a search into auxiliary metabolic genes from integrated proviruses that may influence bacterial metabolism. Investigations into the viral functional potential and the overlap with bacterial pangenomes will be key to understanding the viromes' influence on biological ecosystems.

4 Description of research projects

New advances in methods and analysis are needed to address the impact on macroecology by the thousands of viruses present in biotic environments such as the human gut [95]. The gut microbiota is tightly connected to human health and so far has been a major focus of research initiatives such as the American Human Microbiome Project (HMP)[96] and the European MetaHIT project [97]. There is a great desire to expand the knowledge sphere of gut ecology to less characterized segments of the gut community such as the viral kingdom. Bacterial infecting viruses (bacteriophages) are suggested to impact bacterial density and diversity, thus filing a profound niche in the environment. Gut viruses have largely been characterized in multiple studies using viral enrichment methods (Clooney et al. 2019; Shkoporov et al. 2019; Norman et al. 2015; Roux et al. 2016). This procedure greatly improves the metagenomic assembly and identification of gut viruses but also biases the types of virus studied by capturing a limited segment of virome diversity (Roux, Adriaenssens, et al. 2019; Gregory et al. 2020). Hence, improved methods for mining viral biodiversity in bulk metagenomic samples are needed to enable virome analysis without viral-enrichment and uncover the full spectrum of virome diversity in future metagenomic datasets.
Towards facilitating virome analysis on the growing number of metagenomic samples and enabling exploration into bacterial and viral communities in biotic sites like the human gut, I present and discuss key results of three major studies that have worked toward this aim. First, the metagenomic binning engine VAMB that has provided a fast and reliable framework for genome reconstruction. Second, our exploration and benchmark of viruses extracted from bulk metagenomics and paired metaviromes. Third, an application of our methods to delineate novel viral diversity in humans of extreme longevity and an analysis on the age-dependent impact on viral and bacterial interactions.

4.1 Project I: Deep learning for binning and high resolution taxonomic profiling of microbial genomes

Discovery of novel gut microbiome residents has been accelerated with computational methods such as metagenomic binning, which organize metagenomic assembled DNA sequences, corresponding to chromosome fragments, from the same organism into genome-bins [12]. Several attempts have been made to reconstruct thousands of microbial species from massive metagenomics datasets

of hundreds of people [98, 99], by independently assembling and binning each metagenomic sample into genomes. Single-sample binning allows massive parallel processing of samples but does not leverage co-abundance information across samples. Other methods that are developed to perform binning using co-abundance information from all samples deduplicate sequences before binning [10, 69], which may mask strain-level genome variation and produce intersample chimeric genomes. These chimeras do not represent real microbial genomes and it would be preferable to have such strains assembled per sample and enable functional strain comparison. The main difference between VAMB and existing binners including MetaBAT2, MaxBin2, Canopy and others is that VAMB utilizes unsupervised deep learning to encode contigs into lower-dimensional latent embeddings based on integrated information of co-abundance and sequence composition structure. In order to test VAMB's performance in reconstructing bacterial genomes, it was applied to (1) established simulated datasets for metagenomic binning benchmarks [100] and (2) a real metagenomic dataset comprising 1000 metagenomes [12].

4.2 Project II: Genome binning of viral entities from bulk metagenomics data

In the second project we explored genome binning of virus constituents in metagenomics using VAMB as our binning engine. One key feature of the method is, besides state-of-the-art binning performance, it learns to group genomes from the same organisms across samples. In other words, across a metagenomics dataset it learns which genomes are from the same species. We therefore hypothesized that besides bacterial genomes it could also bin and learn viral species despite their astounding diversity [101]. This would provide an important advancement to cataloging viral species that are notoriously difficult to separate due to the lack of conserved taxonomic markers [15]. Specifically, if the autoencoder framework effectively separates bacterial species on strain-level based on abundance and sequence composition, can it do the same for viruses? To evaluate VAMB's ability to capture individual viruses as bins, we had access to the largest dataset of deep-sequenced paired metagenome and metavirome samples from the human gut. This dataset encompassed 662 metagenomic and 662 metavirome paired samples obtained from infants at 1 year of age in the Danish COPSAC cohort [102].

Viruses from the metaviromes were assembled, quality-controlled and de-replicated to establish a ground truth set of viruses. With a set of labeled viruses, we looked up the origin of each sequence in a putative viral bin generated with assembly and binning of the paired bulk metagenomic samples, thus establishing whether a bin corresponded to a real virus. This enabled us to compute degrees of recall/completeness and contamination of viral bins, i.e. does every sequence in a viral bin map to the nearest reference virus in the truth set and does the bin correspond to a complete virus. From these calculations we established the completeness of viral bin recovered in bulk metagenomics and the viral overlap with viruses assembled from metaviromes based on viral enrichment. These efforts provided a list of annotated metagenomic viral bins that we leveraged for training a supervised viral prediction for bin classification in metagenomics. To create a training and validation set, viral bins were combined with bacterial bins corresponding to bacterial metagenome assembled genomes (MAGs). Key genomic features recorded for each bin such as bacterial and viral marker genes were used to train a Random Forest (RF) model to distinguish the two types of microbiome constituents. The RF model performance was evaluated on annotated metagenomic bins derived from the processing of the Diabimmune dataset containing 112 paired metagenomic and 112 metavirome samples [103]. To support the RF model performance on real-datasets, the model was further evaluated on simulated datasets containing virus, plasmids and bacteria generated with tools by the Critical Assessment of Metagenome Interpretation (CAMI) consortium [71].
Finally we applied our viral binning workflow Phages from metagenomics binning (PHAMB) to a massive public metagenomic dataset, the Human Microbiome Project 2 (HMP2) with IBD cases and controls longitudinally sampled [104], from which no virome characterisation had been described before. Virus populations derived from this dataset were used to establish longitudinal virome profiles, alpha and beta diversity estimates, separation of samples based on clinical dysbiosis scores and individual phage-dysbiosis associations.

4.3 Project III: Centenarians have a diverse population of gut bacteriophages that may promote healthy lifespan

In the third project we applied our combined VAMB and PHAMB approach to uncover and investigate the bacterial microbiome and virome in centenarians.

We investigated Japanese centenarians studied in collaboration with the Broad Institute, Boston, US and the Centre for Supercentenarian Medical Research, Keio University, Japan. Centenarians (Age > 100) and in particular supercentenarians (Age > 110) are examples of humans with exceptional longevity. Studies on centenarians have characterized their unique physiology with a low cardiometabolic risk, preferable lipid profiles and protective plasma biomarkers [105, 106]. In addition, centenarians show great resistance to aging-related diseases. One of the suggested components to contribute to their longevity is the gut microbiome [107]. An initial characterisation of centenarian microbiomes revealed enrichment of bacteria capable of producing novel secondary bile acids with antibiotic properties towards typical gut pathogens [108]. Altogether this suggested that centenarians likely exhibit greater resistance towards infectious diseases. Further bacterial and metabolomic analysis of centenarian microbiomes are needed to reveal other host-health related factors in the microbiome.

The cohort investigated by Sato et al. (2021), consisted of centenarians [n = 176 (172 individuals)], elderly (n = 133), young (n = 61) and represented by far the largest microbiome centenarian dataset published. As the gut virome of centenarians have not been described before, we delineated the virome by combining viral-binning and provirus search in bacterial MAGs. To establish the degree of novel viral diversity, we performed viral clustering of the discovered viral-bins and proviruses into viral operational taxonomic units (vOTUs) with the MGV database, which is the most representative DNA virus and phage database published [109]. In order to place the newly identified vOTUs in the context of known diversity, specific viral protein markers were identified in vOTUs and representative MGV genomes to build a phylogenetic viral tree for identifying branches and clades of novel viruses enriched in centenarians. The bacterial affiliation of viruses were determined by CRISPR-spacers, evidence of integration in bacteria and clustering with proviruses of isolated bacteria.

The way in which the virome interacts with the bacterial community has so far been studied in infants from birth up until 2 years of age, where the virome undergoes dramatic changes as the pioneering bacteria settle in the gut [33, 37]. In order to provide new insights into how the virome interacts with bacteria during the last stage of the human lifespan, we calculated viral-bacterial ratios of temperate viruses from young to centenarian microbiomes. Because we had developed a framework to establish the virome in bulk metagenomics, we could include two additional cohorts (infant and another young cohort) in

the analysis to establish that the calculated viral-bacterial ratios (VBRs) were reliable estimators of lysogenic activity between groups. We hypothesized that if the calculated VBR distributions captured overall trends or differences of lysogeny in the microbiome, we could compare general viral-bacterial interactions for different age-groups regardless of bacterial community composition. This analysis was limited to confidently annotated temperate viruses as these are capable of switching between a lytic and lysogenic lifestyle. Finally, as viruses are known to influence bacterial metabolism by infection [65], we characterized and investigated viral genes in search for auxiliary metabolic genes (AMG) related to metabolic systems and pathways in host-bacteria.

4.4 Datasets overview

The three projects featured in this dissertation are based on a wealth of different datasets. Here I provide an concise overview and description of each. In the overview I refer to bulk metagenomic samples from the human gut to human gut microbiomes. In addition, I refer to viral-enriched metagenomic samples as human viral metagenomes.

Dataset I. Almeida [12]. A cross sectional study of 11,850 human gut microbiomes from 75 different studies. From this study we sampled 1,000 metagenomic samples.

Dataset II. CAMI datasets [100]. Simulated metagenomic benchmark datasets.

Dataset III. COPSAC 2010 [102]. A cross sectional study of 647 healthy Danish infants. The dataset includes 647 paired human gut microbiomes and viral metagenomes.

Dataset IV. Diabimmune Type 1 Diabetes (T1D) [103]. Longitudinal study of 33 infants genetically predisposed to T1D. The dataset includes 220 paired human gut microbiomes and human viral metagenomes.

Dataset V. Human microbiome project 2 (HMP2) IBD [104]. Longitudinal study of 132 of participants with Crohn's disease (CD), Ulcerative colitis (UC) or no IBD (nonIBD). The dataset comprises 1337 human gut microbiomes.

Dataset VI. The Japanese centenarian cohort [108]. Cross sectional study of Japanese adults of three different age categories. The dataset comprises human gut microbiomes from 176 centenarians (>100 years old), 110 elderly (<100 years old) and 44 young (>18 and <55 years).

Dataset VII. The Sardinian centenarian cohort [110]. Cross sectional study of Sardinian adults of three different age categories. The dataset comprises human gut microbiomes from 19 centenarians (>100 years old), 23 elderly (<100 years old) and 17 young (>18 and <55 years).

Dataset VIII. EDIA cohort [111]. Longitudinal study of 142 infants and mothers from Finland, which were followed across the first year of the child's life. From this dataset we selected 668 bulk microbiomes of infants.

Dataset IX. Tanzania 300FG [112]. Cross sectional study of 315 adults from Tanzania. The dataset comprises 315 human gut microbiomes.

5 Summary of results and discussion

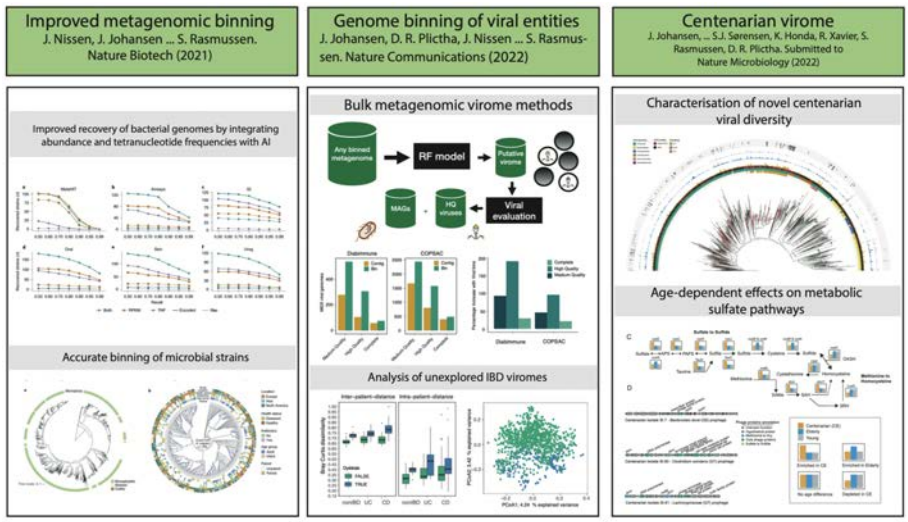


Figure 5.1. Visual summary of the three major studies included in this PhD dissertation.

Project I: Deep learning for binning and high resolution taxonomic profiling of microbial genomes

In this project we presented VAMB, a variational autoencoder (VAE) that integrates co-abundance and tetranucleotide frequencies for metagenomic contig binning. Benchmarking of the binner's performance on various CAMI datasets [100] revealed that VAMB is able to recover 29-98% additional near-complete genomes compared to state-of-the-art methods on simulated datasets. Furthermore, we showed that the improvement was significantly positively correlated with the complexity of the datasets, which was measured as the mean Shannon entropy of the genomes for each dataset. As a compelling application of the method to a real dataset, we applied VAMB to 1,000 human gut microbiome samples [12]. Importantly, for this dataset the original authors used state-of-the-art single sample binning, co-assembly and even an ensemble of multiple binners to recover and discover novel MAGs. VAMB outperformed all these approaches by binning 45% additional near-complete genomes from these samples. Furthermore, the improvement of our multi-sample strategy over the

single-sample approach was clear for all datasets tested, down to as few as six metagenomics samples. Therefore, almost all typical metagenomics experiments could benefit from using our method. At the very least, researchers that apply ensemble binning on a dataset to reach the best set of reconstructed genomes such as dRep [113], MetaWRAP [114] or DAS-tool [115] should include VAMB as it provides a higher number of unique NC genomes compared to MetaBAT2, which is the most widely used metagenomic binner.

5.0.1 High taxonomic resolution from binning

We found that VAMB binning of strain level genomes are grouped into taxonomically meaningful clusters. A major difficulty to metagenomics binning is strain separation, which we investigated with a mixed strain experiment with spike-in reads from various Salmonella genomes. VAMB was able to distinguish and reconstruct Salmonella genomes in samples even in the co-presence of genomes with 98-99.5% average nucleotide identity (ANI). A large-scale illustration of VAMB's strain-level binning performance was illustrated for a real dataset by using bacterial marker genes from 5036 near complete genomes, from which we constructed a phylogenetic tree and illustrated the consistent monophyletic structure of VAMB clusters (Figure 5.2).

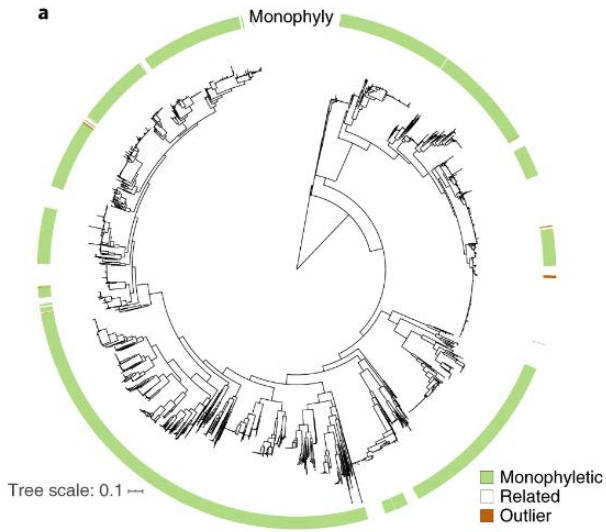


Figure 5.2. Phylogenetic tree of 5036 genomes based on amino acid bacterial marker genes.The majority of bins are monophyletic and placed in the tree adjacent to genomes of the same VAMB cluster and share highly similar markers (¿99% amino acid identity). White leafs correspond to genomes of a cluster with one or more outlier genomes. Orange leaf is an outlier compared to the medoid of the cluster.

Generally we found that VAMB-clusters comprise genomes with >=98 ANI based on the NC genomes recovered in the Almeida dataset from two distinct

Bacteroides species, as illustrated in (Figure 5.3). Curiously for clusters C546 and C94 the inter-cluster ANI was greater than 95 ANI, a frequently used ANI cutoff to distinguish species, but this could be explained for these particular species as they are known to be close phylogenetically and share high genomic similarity [116].

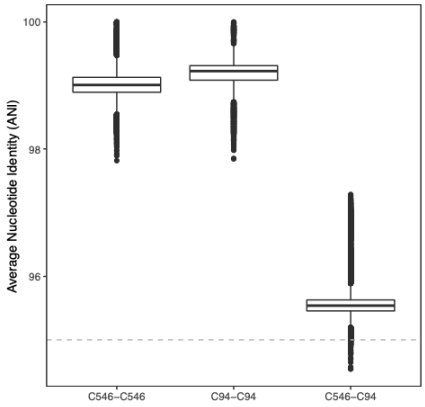


Figure 5.3. Boxplots showing the calculated Average Nucleotide Identity (ANI) between genomes in C546, C94 and between genomes of C546 (B. vulgatus genomes, N=255) and C94 (B. dorei, N=91). The grey line indicates an ANI of 95.

5.0.2 Clustering the VAE latent space

One of the reasons VAMB performs so well for metagenomic binning can be accredited to the VAE concept. Recall that a VAE is trained to map similar inputs closely in the latent space, in this case contigs originating from the same genome should be placed adjacent in the latent space. This is because the VAE is trained to infer the source genome likely to have generated the input contigs. How the contigs in the latent space are practically divided into accurate genome clusters is where the actual genome clustering is happening.
VAMB's custom clustering algorithm works through a series of clustering-steps on the inter-cosine distances of contigs generated in the VAE's latent representation (Figure 5.4). The function works by **(a)** establishing a medoid (center) contig and closely related neighbor contigs based on a small radius of 0.05 in cosine distance **(b)** optimizing the mediod by random sampling neighbors to

find the contig with most neighbors within the radius and **(c, d, e)** establishing an optimized cluster radius and the best center medoid based on a distance histogram from the medoid to the remaining contigs. The core idea is that a cluster of contigs represents an area of high density with small inter-contig cosine-distance in latent space, surrounded by an area of low density separating it from distant contigs of higher cosine-distance. Finally, a cluster is returned with all contigs, which are removed from the search, within the established radius of the medoid contig **(e)**. Then the search continues again based on the remaining contigs to be clustered **(f)**.

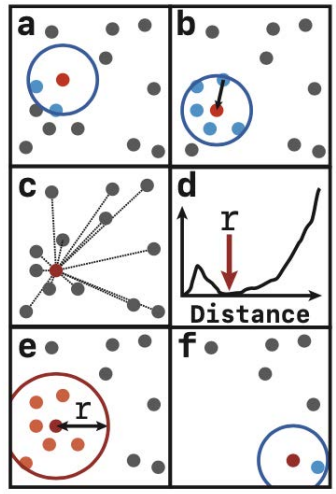


Figure 5.4. Visual illustration of VAMB's clustering algorithm (steps described in text). Each colored dot represents a single metagenomic contig.

The clustering algorithm is well suited for large datasets with millions of sequences and importantly can be accelerated with a graphical processing unit (GPU). Technically, any good clustering algorithm could be used on the latent representation and potentially improve upon the results. A close clustering competitor would be HDBSCAN [117], but VAMB's custom clustering algorithm was 20 times faster when tested with a dataset of 200000 contigs.

The clustering algorithm is inherently dependent on how the VAE infers the

source genome of each contig. An immediate issue in this case would be overfitting where the model learns to reconstruct every contig perfectly but to an extent where they are mapped too distantly in latent space. However, even after running VAMB for 3000 epochs (training iterations) on CAMI datasets (thereby increasing the risk of overfitting), we found no general decrease in the number of NC genomes reconstructed. Some details we did not investigate was how well the clustering algorithm dealt with small data or genomes represented by just a few contigs such as mobile genetic elements including phages, plasmids and horizontally transferred genomic islands. Large bacterial contigs should be the easiest to differentiate in the latent space from other contigs due to their size that provides a better estimation of sequence composition and abundance across samples, while the sequence composition of smaller contigs provide a worse representation. Therefore, a minimum of 2000 base pairs contigs is advised for running VAMB which produced the best results during benchmark.

5.0.3 Beyond bacterial binning

In this project we primarily focused on benchmarking the recovery of microbial genomes. As prokaryotic genomes also harbor other biological entities such as plasmids and viruses, improved bacterial reconstruction aids in the discovery of horizontal elements associated with the host. Based on a set of genomes reconstructed from the Almeida dataset by both MetaBAT2 and VAMB for which we had a NCBI reference genome, we found that VAMB reconstructed larger genomes than MetaBAT2. Within the additional genomic content in VAMB bins, we identified that 35% corresponded to phage-like contigs. In addition, we found an overall increased AT nucleotide content, which suggested horizontally transferred regions [118]. Therefore, VAMB recovered additional bacterial associated MGEs. This spiked our curiosity on whether binning with VAMB could also effectively bin and separate viral contigs simultaneously with larger entities such as bacteria. We conducted an initial alignment of contigs of smaller metagenomic bins to the NCBI virus database and found precise and consistent mapping (data not shown), which motivated our next study into other segments of the microbiome.

Project II: Genome binning of viral entities from bulk metagenomics data

In this project we expanded the scope of genome binning to viruses in metagenomics using VAMB as our binning engine. We trained a RF model using paired metagenomes and metaviromes to filter and extract putative viromes from metagenomics (Figure 5.5). Applying the trained RF model to any binned metagenome aids in the delineation of bacterial MAGs and viral MAGs (vMAGs), which can be followed up with exploration into host-viral dynamics and viral gene contributions and dissemination (Figure 5.5). Finally, we benchmarked viral genome recovery with a binning approach using synthetic CAMI generated datasets and three metagenomic datasets COPSAC, Diabimmune and HMP2 IBD.
One of the first facets of the paper to be discussed is the motivation for viral binning. Binning of bacterial MAGs has been in development for many years, why not viruses?

5.0.4 To bin or not to bin?

To achieve an absolute number on the improvement on viral genome quality gained with binning, we tallied the number of viruses recovered as single contigs and viral bins by genome quality tiers. We were able to recover up to 210% additional high-quality (HQ) viral genomes compared to using a single contig virus approach. To ensure a fair comparison, we investigated the exact same set of contigs in every dataset with and without binning. In addition, we found that binning enabled recovery of up to 36% of HQ viral populations found in the metavirome directly from paired bulk metagenomics data. Furthermore, 47% additional HQ viral populations were discovered in bulk metagenomics and not in the metavirome. Here, a likely explaining factor is viral sampling bias as a result of sample preparation, since metavirome preparation concentrates smaller viruses and predominantly viruses in a lytic stage and not integrated in bacteria [15]. The surprisingly high intersection (36%) of viruses in metagenomics with and without viral-enrichment , which has been estimated to 8.5-10% in another study [37], suggests a great potential to conduct virome analysis based on bulk metagenomics.

A major worry with binning is the risk of including contigs from other sources of species such as genome bin contaminants. We benchmarked the degree of

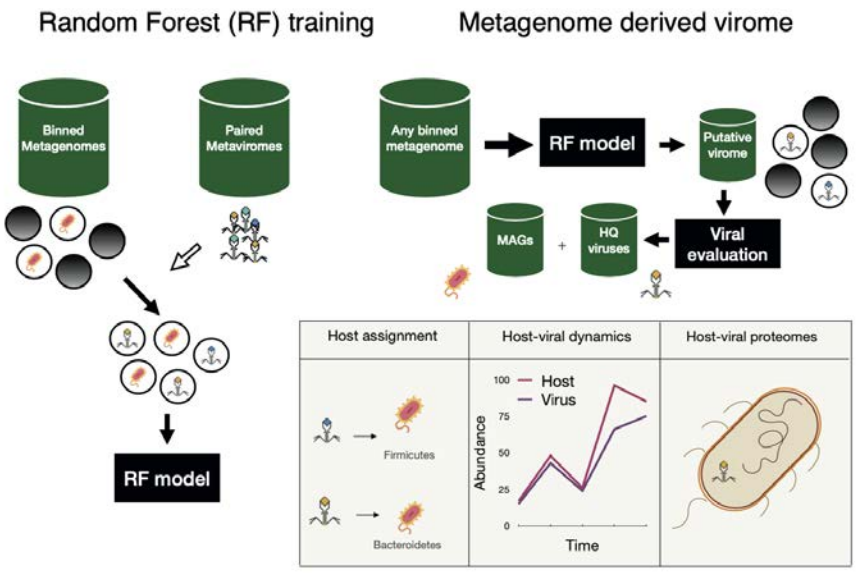


Figure 5.5. Random Forest (RF) modeling was performed on binned metagenomes paired with metaviromes. The trained RF model can be applied to any binned metagenome for predicting the putative virome in which high-quality (HQ) viruses can be identified and combined with bacterial MAGs. Identification of virus host affiliation enables analysis into host-viral abundance dynamics and the functional space shared between bacterial host and viruses.

contamination to an average of 2.55% of the viral bin genome in base pairs not aligning to the virus of reference, which equals a genome purity of 97.45% on average. This benchmark was also performed on synthetic datasets where the average genome purity was 94.5%. Evidently, viral binning with our methods is not a perfect process and does pose some risk for virus contamination, although contaminating sequences may be removed during bin post-processing. Binning is also a process which disregards the correct order of contigs in a genome as it merely groups contigs together that belong to the same genome. This might not be desirable if a researcher is interested in the order of virus gene transcription during infection of a bacterial host. For instance, phage encoded anti-defense proteins that counteract bacterial defense systems are suggested

to be phage early genes [119]. The correct order of phage genes annotated in a genome can therefore be critical for determining novel anti-defense genes that are expressed prior to defense system activators such as portal and terminase proteins, which are recognised and activate antiviral defense systems [119].

With the goal of virus discovery in mind on new or old metagenomic datasets, we have shown that binning improves identification of higher quality virus genomes. On the contrary, if a researcher's focus is to perform massive virus mining across hundreds of thousands of assembled metagenomic contigs on NCBI, corresponding to single-contig identifications, binning is not a technically feasible solution at this moment as it requires a contig x sample matrix. Including >20.000 of samples with several millions of contigs results in an astoundingly big matrix with intense demand for computational processing and memory allocation. However, for single datasets such as a new metagenomic dataset from a cohort of patients, the unsupervised clustering algorithm built into VAMB provides strain-like viral clusters simultaneously with bacterial MAGs. As such, the genome of an abundant virus in a patient can be tracked and compared across multiple samples, if longitudinally sampled, simultaneously with the predicted bacterial host. We illustrated for the HMP2 dataset that the VAMB-clusters produced for crAss-like virus genomes were accurately differentiated on a high taxonomic level, which we illustrated in a phylogenetic tree (Figure 5.6). Essentially, we found that viral binning across a cohort enables precise clustering of viral populations with high intra-VAMB-cluster ANI (>97.5%) that can be leveraged for longitudinal or cross-sectional viral genome comparison studies.

5.0.5 What we learned about the virus functional potential

Based on the current tools available for annotating protein domains, we established that the viral protein-coding genes in HMP2 exhibited high prevalence of core viral proteins related to structure (capsid, tail, head etc.) and integrase enzymes for integration into host chromosomes. In addition, we also found a high frequency of reverse-transcriptase (RT) domains in viral proteins. RT domains are increasingly identified in multiple gene configurations that go way beyond the RTs' role as a retro-viral transcriptase necessary for RNA-virus replication. RT domains have been identified in numerous prokaryotic anti-phage defense systems, such as restriction modification enzymes and abortive

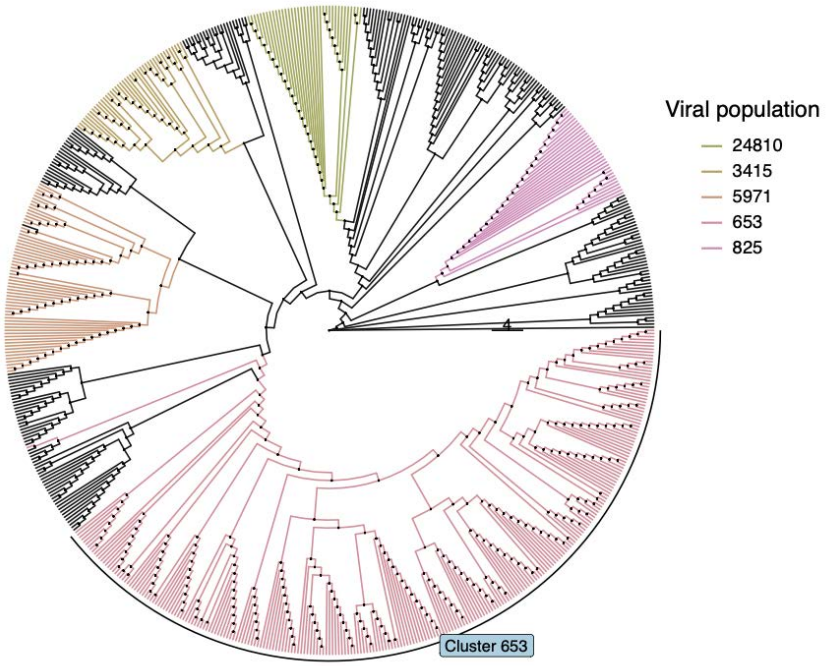


Figure 5.6. Cladogram based on a phylogenetic tree of crAss-like virus genomes colored and named by VAMB-cluster.

infection mechanisms [120] and in diversity generative regions (DGRs) [121]. In order to circumvent bacterial defense systems and exclude other viral competitors, phages also encode prokaryotic defense systems such as anti-viral systems [35]. Thus, the high frequency of annotated RT domains indicates the potential abundance and importance of these genetic systems in the bacterial-phage arms race. Future studies should leverage new bioinformatic tools to annotate the presence of anti-phage systems in bacteria [122] and phages to further understand intricate interactions in complex environments like the human gut. Furthermore, we also identified proteins in phage genomes with TonB plug and TonB receptor domains that encode established immune stimulating epitopes [52]. This finding underscores the presence and potential phage-driven distribution of epitopes that may stimulate host immune cells and contribute to

gut immune stimuli. A completely different perspective is cross-reactivity with phage-encoded epitopes that activate host T-cells through MHC-1 receptors via "molecular mimicry" [123]. Studies on the commensal epitope landscape should strive to recognise the presence of epitopes in bacteria and viruses, as the abundance of both entities may impact host immune activity during health and disease.

5.0.6 The future of virome analysis without viral enrichment

One of the major motivations for benchmarking virus recovery with binning in the first place was to investigate virome analysis for datasets where whole-virome sequencing is not available. To evaluate the methods' utility, we applied it to a massive public metagenomic dataset, the HMP2, from which no virome characterisation had been described before. Here we identified 3,625 viral populations consisting of 16,358 viral bins (Medium-quality or better). We have illustrated that virus-binning is feasible and quite valuable, thus future efforts in binner-development will likely improve upon these numbers by harnessing better computational models trained on better and larger datasets. Nevertheless, can we imagine a future without the need for binners?

Certainly, 3rd generation (long-read) sequencing technologies such as Nanopore and PacBio can produce long-reads which improve the assembly and binning of genomes from metagenomics [124, 125]. Furthermore, recent results have shown that combining short and long reads increased the number and genome-quality of viruses in marine environments compared to illumina sequencing (short read) alone [126]. These results may be a primer for the 3rd generation sequencing coming of age where long-reads can capture viruses in whole sequences, which ultimately alleviates assembly issues caused by repeat and low-coverage regions found in virus genomes [67]. However, the recurring issues with 3rd generation sequencing comprises higher base-calling error rate and frequency of insertions and deletions (indels) [127]. Combining long-reads and short-read sequencing has been the common strategy to deal with long-read errors, where short-reads are used to correct errors in assembled sequences [128]. Yet, recent Nanopore technology has shown to bridge the gap in terms of price and sequencing accuracy while also increasing the number and quality of recovered prokaryotic genomes [125]. So, where do we stand in terms of long-read sequencing of viruses in bulk metagenomics? Recent studies have ex-

plored the benefits of combining 2nd and 3rd generation sequencing [126, 129] and shown that long-reads captured additional viral diversity but also illustrated that long-reads with PacBio sequencing only captured few HQ viruses [126]. Thus, hybrid assembly combining short and long-reads seems to be a promising strategy for exploring viruses in bulk metagenomics at this point in time. At the very least, improved recovery of MAGs with long-read sequencing may strengthen identification of complete proviruses.

5.0.7 Frameworks for benchmarking virus completeness, where credit is due

In 2022, gut ecologists have access to reliable and trusted frameworks for identifying and annotating virus genomes, both *de novo* and by reference-based approaches. An assembled putative virus sequence is automatically gene-annotated using a wealth of finely curated virus marker databases and simultaneously aligned to a massive collection of viruses composed of hundreds of thousands of genomes [78]. Thus, phage-genomic research has indeed rocketed since the year of 2019 when we initiated the planning of a benchmark on viral genome binning based on bulk metagenomics assemblies. Programmes like Virsorter and Virfinder did exist but lacked features for referencing viruses in the space of known biodiversity or scoring genome-completeness to address whether a virus genome was complete or a fragment. Therefore we designed our benchmark strategy based on paired metaviromes with *bona fide* assembled viruses, which provided a sensible starting point for calculating one-to-one (viral-bin to virus) comparisons. Alas, two convenient bioinformatic tools were released in 2020, CheckV and VIBRANT, which brought new standardized measurements of virus quality such as completeness and contamination while simultaneously referencing a grand catalog of viral biodiversity. The extent to which binning can be used for recovering viruses in metagenomics could not have been explored so extensively without research efforts from other research-groups such as the Microbiome Data Science Group at JGI-DOE and Anantharaman-lab, to whom we are grateful. A background and blogpost about the article and research was further published in Nature's microbiology community forum [1]. In addition, we were fortunate that a science journalist at the danish newspaper Politiken (Appendix 9.1) found our article relevant for a story on combating

[1] https://microbiologycommunity.nature.com/posts/microbiome-analysis-of-viruses-is-more-accessible-t

antimicrobial resistance using phage cocktails [2].

In order to define future phage cocktails that can target and kill bacterial culprits, the relevant and specific viruses have to be discovered first, which is what our methods can support. The next question we phrased was: what metagenome and context would benefit from virus discovery and virome analysis? We had an established starting point for virome analysis in bulk metagenomics, but where to begin? A general topic of interest is the influence of age on the human microbiome and its development over time, which has been investigated extensively for bacteria [130, 131]. A recent study into the infant virome had revealed the turbulent development of the pioneering viral constituents and response to the maturing bacterial community [33], while another mapped the development of viral families over time from infant to elderly [37]. Interestingly, only a couple of studies have investigated the age-dependent effects on the virome and by no means the extreme end of human longevity.

[2] `https://politiken.dk/viden/Viden/art8735714/Praksissen-var-ellers-g\%C3\%A5et-i-glemmebogen-i-Vesteuropa-men-nu-har-forskere-for-alvor-f\%C3\%A5et-\%C3\%B8jnene-op-for-tarmbakterierne`

Project III: Centenarians have a diverse population of gut bacteriophages that may promote healthy lifespan

In this project, we studied a Japanese centenarian cohort that has previously been characterized with focus on bacteria and archaea and their unique bile acid signature, which has shown antibiotic properties against typical gut pathogens and protects against infectious diseases [108]. We applied computational viral binning and prophage search to characterize viral diversity in centenarian microbiomes and study the age-dependent effects on the virome and the viromes' role in aging and influence on bacterial constituents (Figure 5.7 A). We discovered 1,746 novel viral species (vOTUs), including many that were enriched and prevalent in centenarians compared to elderly and young controls. We also observed a viral lytic activity shift in centenarians compared to younger age groups, which we compared with viromes mined in two independent age-reference cohorts (Figure 5.7 B). We then outlined the functional implications of the diverse genetic reservoir encoded by phages (Figure 5.7 C) and identified that centenarian viruses were significantly enriched with genes encoding key enzymatic steps related to conversion of sulfate to sulfide and methionine to homocysteine. We further validated the presence of these genes in integrated prophages from centenarian bacterial isolates including *C. scindens*, *B. dorei* and *Lachnospiraceae spp.* This revealed that the centenarian microbiome configuration displayed increased potential for converting methionine to homocysteine, sulfate to sulfide and taurine to sulfide. Previously it has been shown that infection events in microbiomes lead to colonization resistance when taurine is converted by microbial taxa into hydrogen sulfide [132]. We therefore speculate that this could translate to increased levels of microbially derived sulfide leading to health promoting outcomes with protective effects against pathogens. Overall, healthy aging seems to promote a rich and diverse virome that interacts with dominant bacterial hubs in the microbiome. Furthermore, patterns of phage and host lysogeny suggest that prophage encoded genes contribute to bacterial metabolic dissimilatory sulfate reduction pathways in the microbiome. This contribution was more potent in centenarians and provides a machinery for greater metabolic output of microbial hydrogen sulfide which may in turn support mucosal integrity and resistance to pathobionts.

There were several interesting and important facets of this manuscript that should be discussed more in-depth. First I will address the process of estab-

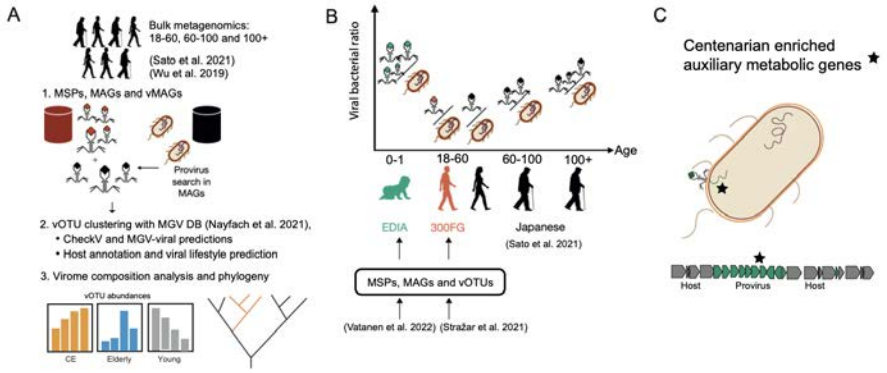


Figure 5.7. The three major arcs of analysis in the centenarian virome study. (a) Describes the assembly and binning of bacterial and viral genomes from metagenomics of different human age groups, which are used for virome compositional analysis and phylogenetic inference. In (b) the viral bacterial ratio is calculated for different age-groups in the Japanese cohort. Subsequently, the VBR is analyzed for two independent cohorts. The functional impact of virus encoded genes on bacterial metabolism is investigated in (c).

lishing viruses in a dataset that is highly underrepresented in the databases of virus genomes, since no one to this date has characterized the centenarian virome with metagenomics.

5.0.8 Characterizing viral novelty in the extreme end of a human lifespan

With the aim of characterizing novel viral diversity in centenarians we applied an array of published viral verification tools and the comprehensive MGV database [109]. All viral MAGs and proviruses extracted from MAG contigs were processed with CheckV, VIBRANT and the validation method used for the MGV database. In our viral binning paper (Paper II) all downstream viral analysis was based on HQ viruses according to CheckV. Here, HQ viruses correspond to viruses with a protein-coding gene content with high similarity and size to known viral genomes. For this manuscript, we had to identify viruses beyond HQ viruses (and MQ) as these viral-sequences mostly represent viral species already discovered in other publicly available metagenomic datasets.

Therefore, we included sequences containing viral genes but determined low-quality (<50% estimated completeness) virus genomes which could represent incomplete or novel virus genomes. The viral sequences were dereplicated into 5522 putative viral operational taxonomic units (vOTUs), corresponding to viral species clusters containing one to several homologous sequences. To trim this set into confident *bona fide* vOTUs, we kept those that satisfied the following criteria: The sequence should contain at least >=40% viral genes and <10% host genes or be determined viral by the MGV viral prediction pipeline [109]. This filtering step resulted in the final set of 4422 vOTUs. Subsequently, we employed VIBRANT to get an additional prediction which determined 4240/4422 (96%) of the vOTUs to be viral and increased our confidence in this subset. Admittedly, the criteria related to the percentage of viral content were rather conservative and we may have missed out on an additional 829 vOTUs that represent actual viral diversity. However, it is preferable to reduce the risk of including false-positive viruses that may confound the analysis. As new viruses are discovered, they can be uploaded to the CheckV and MGV database to increase the viral coverage for future studies. The next important step with impact on the downstream analysis is distinguishing novel vOTUs from known ones. To do this, we clustered vOTUs with the MGV database on species level (ANI>95), which contained >200.000 vOTUs from thousands of metagenomes stored at the JGI-DOE (`https://jgi.doe.gov/`). We would advise future researchers working on viral studies to follow a similar approach for annotating novel vOTUs by using the same clustering methods and parameters as the authors behind the MGV database, in order to achieve comparable results that are not confounded by the clustering method.

The viral clustering efforts showed that the smallest fraction of novel vOTUs were those detected only in samples of young microbiomes, in contrast to viruses detected in centenarian and elderly microbiomes which made up the biggest proportion of novel viruses (excluding novel viruses detected in all groups). This difference remained significant across several permutations where we downsampled the compared age-groups to match the smallest group (the young control). The large proportion of novel viral diversity from centenarian and elderly was not surprising, as these studies currently represent the biggest published metagenomic centenarian cohort. As expected, the small proportion of novel viruses detected exclusively in young microbiomes illustrate that the MGV coverage of this age group is comprehensive. On viral genus level, we found that many of the novel vOTUs expanded existing viral genera such as the genus we denoted G7 with multiple members annotated by CRISPR-spacers to

bacterial *Clostridia* species. We further identified proviruses in centenarian *C. scindens* isolates that also clustered into the G7 genera and further confirmed the taxonomic affiliation to the *Clostridia* family. Many of these viral genera of interest comprised bacteriophages.

5.0.9 The first piece on centenarian viral diversity

The technical differences between viromes extracted with and without viral-enrichment means that each preparation method captures different subsets of the virome in terms of size, rarity and biology. These differences have been described before and suggest that viromes from non-enriched samples (bulk metagenomics) may miss out on some rare viruses and be biased toward phages infecting dominant host-cells [15, 37]. The fact that bulk metagenomics captures different subsets of viral diversity was also evident from the analysis in our viral binning manuscript. We previously established that 51% of HQ viruses found in bulk metagenomics were not found in the corresponding samples prepared with viral enrichment [133]. The technical bias of viral enrichment may help to explain the low detection of Microviridae in the Japanese dataset. In the centenarian age group Microviridiae was only detected in 33% of the samples. In contrast, Microviridiae have been detected in every sample and acquired up to 90% of the VLP sequencing reads in other studies [24]. Thus, viral-enrichment is suggested to be biased toward virulent viruses including microviruses and therefore miss a lot of temperate bacteriophage diversity [15]. In conclusion, each sequencing approach has its own limitations and captures different subsets of the virome community, thus future studies in centenarians microbiomes should ideally leverage both approaches to capture the full picture of viral diversity.

A suggested benefit of using bulk metagenomics to study the virome is the abundance of viral diversity which interacts with bacteria like temperate bacteriophages [15]. This motivated us to investigate how the stable lysogenic community of phages acquired during early childhood develops throughout life from infant to old age.

5.0.10 Age and viral-bacterial interactions

In order to provide new insights into how the virome community develops during the last stage of the human lifespan and even into the extreme of human longevity, we calculated viral-bacterial ratios (VBR) of temperate viruses in

microbiomes of infants, young, elderly and centenarians. In the most classical sense, the VBR provides a value that reflects the measured viral particles (concentration or weight) relative to bacteria. This aggregated measurement has been used to establish our theoretical concepts of virus development in the infant gut and viral activity in marine environments [33, 134]. An environment where the concentration of viral particles outnumbers bacteria is reflected in a VBR higher than 1 and suggests higher lytic viral activity. Furthermore, VBRs have also become a popular proxy to determine whether a provirus in a bacterial genome is currently actively replicating beyond its host and corresponds to active prophage induction. In this case, the VBR is measured based on a higher provirus sequence depth relative to the surrounding bacterial chromosome [135, 136]. None of the datasets presented in this study was accompanied with absolute quantifications of viral or bacterial particles and thus we had to make an approximation of the VBR based on an aggregate of bacterial and viral abundance ratios. To calculate VBRs, we focused on temperate viruses as virulent viruses are by definition always actively replicating without a bacterial host. This VBR calculation is akin to estimating provirus lytic activity in bacterial isolates but much more subject to noise as the VBR is calculated across microbiomes of multiple different bacterial hosts and environments. Hence, some general caution is warranted for concluding whether a single vOTU is actively replicating and in a lytic state based on these calculations. However, we found that the VBR value reflected some expected trends: (1) on average it was within the expected human gut microbiome VBR range (0.01 to 1) [32] and (2) significantly higher in infants relative to adult controls [32] (Figure 5.8). In addition, we concluded that geography was an unlikely confounder as the VBRs derived from young Tanzanian microbiomes were comparable to young Japanese microbiomes. In summary, the VBR calculated using sequencing coverage of prophages and their bacterial hosts revealed age related trends that are supported by current theoretical frameworks.

5.0.11 How may phages contribute to bacterial-fitness

As bacteriophages represent a dynamic component of the microbiome, they have a potential to act as therapeutic agents that provide health promoting functional capabilities to the gut bacteria they infect. In this analysis we made an attempt to digest how aging shapes phage and bacterial communities and if that translates into functional enrichments with implications for human health. We found that phage genes are frequently annotated to sulfur metabolic pathways, which echoes annotation results on marine and human gut phages [137].

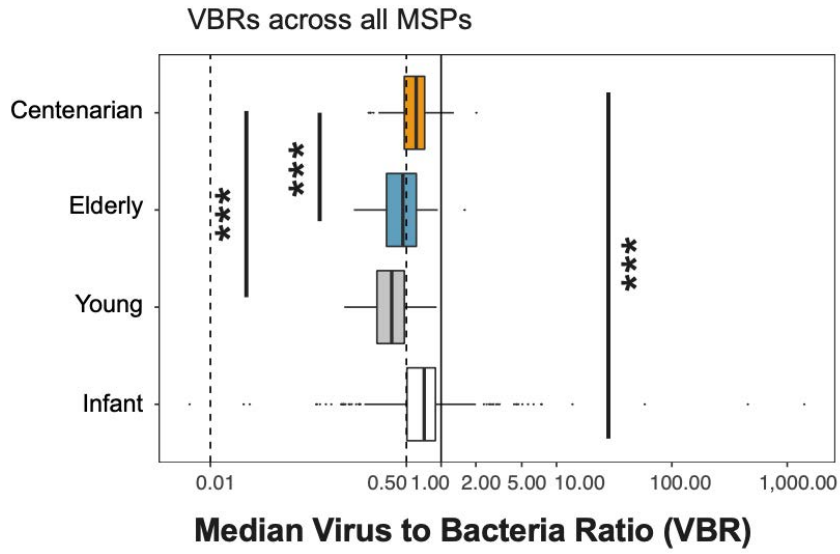


Figure 5.8. The virus to bacteria ratio (VBR) across different age categories. The two first dashed-lines represent the value 0.01 and 0.5, the solid line represents a VBR of 1.

Evidently, there might be a sulfur bias in terms of what we annotate in phages, note that on average 75% of phage protein coding genes cannot be functionally annotated. However, we found that centenarians exhibit higher abundance of these phages encoding sulfur-related genes compared to elderly and young, even when normalized to the abundance of bacteria encoding the cognate genes.

Our interpretation of phage-encoded genes with impact on bacterial fitness through bacterial metabolic pathways represents just one chapter of the complete story on phage-host influence. Brown *et al.* have showcased how phage encoded ADP-ribosyl transferases are secreted and can stimulate the release of epithelial inosine which can be metabolized by host bacteria [138]. Phages may also act as toxin-vectors and inhibit the growth of susceptible bacterial host-competitors by encoding polymorphic toxins with a MuF domain [139]. Phage-encoded toxins are frequently accompanied by an anti-toxin, and thus bacteria containing an integrated phage with an anti-toxin may confer protec-

tion while bacteria without it are susceptible to killing by the phage toxin. In this way, phage-harboring may turn into an advantage and also protect against secondary phage-infection by superinfection exclusion mechanisms [140]. Our analysis only covered a subset of viral functional domains with implications on bacterial fitness. Recently, the curious presence of sporulation sigma-factors in phages, which we also identified in proviruses discovered from centenarian bacterial isolates, have been shown to have significant impact on host sporulation-mechanisms [141]. The impact of virus integration on bacterial fitness and propagation is slowly being unraveled but needs support from experiments and improved protein function prediction methods.

5.0.12 Beyond phage bioinformatics and unannotated viral proteins

Bioinformatics in metagenomics research provides a neat framework for distilling community structures and enrichments which might be unique to a given environment, such as the centenarian gut microbiome. However, higher level community structures are not particularly granular and do not capture the intricate dynamics of bacteria and phages, except for higher level abundance-correlations. In order to delineate the mechanism and influence of provirus gene content, experimental setups with phage and host will be of crucial importance. To go beyond bioinformatics and reveal dynamics of centenarian bacteria and their phages, our collaborators in Japan embarked on isolating specific proviruses. K. Honda and K. Atarashi have been leading this effort with emphasis on isolating proviruses we identified in *Clostridium scindens*, which is one of the dominant secondary bile acid producing bacteria in centenarians. If the secondary bile acids produced by the community of *Clostridia* in centenarians provide effective protection from infectious disease, the community of prophages associated with these bacteria might be important for their host-protective properties or auxiliary metabolic genes. Although phages are primarily promoted for their potential as effective biological killing-cocktails against bacterial pathogens [142], they could also be harnessed as agents for boosting bacterial fitness and modulating the bacterial composition in favor of specific bacteria [64]. The bacterial-fitness effect of the identified prophages in *C. scindens* and other isolates could first be explored in bacterial isolate and phage co-culture experiments with transcriptomic readouts on selected genes related to bile acid conversion and metabolomic quantification of the enzymatic pathway products. This experiment could be followed up by animal

models supplemented with different combinations of isolated phage consortia. The downstream metagenomics should be studied with focus on specific bacterial persistence and the fecal metabolome to detect community changes in bile acid conversion. Experiments that examine the interaction of phages with bacterial species asserted to be health protagonists or disease antagonists in specific situations would greatly improve our understanding of bacteriophages' influence.

Finally, experimental studies allow us to dissect and functionally investigate viral protein functions, including unannotated protein-coding genes. However, the sheer magnitude of the unknown function of proteins in viruses presents a daunting challenge, which would benefit from computational support such as bioinformatics and protein modeling. Protein-sequence language models may be necessary to improve function prediction of unannotated sequences [143]. Currently, massive protein sequence databases represent a major resource for data-driven structural learning, which has been leveraged to develop programmes such as Alphafold2 [144]. Language models that are trained to infer structural relationships based on protein sequences will be necessary to describe the unannotated virus protein domains with little to no homology to proteins with a known function. Even sequences with <30% AA homology may encode the same functional proteins [119]. The practical importance of methods for deciphering the function of unannotated proteins cannot be understated for viruses. As an example, the ability to confidently annotate phage endolysins across thousands of phage genomes could enable systematic high-throughput evaluation of endolysins as species-specific antimicrobials that can be weaponized against antibiotic multi-resistant bacteria [145]. In summary, the thousands of uncultivated virus genomes in databases presents a rich but unannotated dataset that in time can be harnessed for important biotechnological applications.

6 Conclusions and perspectives

The space of known viral biodiversity is increasing at such a pace that the official viral taxonomy structure struggles to keep up [146], yet the degree of viral genomic diversity and variation between biotic environments is suggesting that only a fraction of viral diversity have been identified [147]. A great proportion of the established human gut viruses originate from the first series of metagenomic studies using viral-enrichment strategies that selects for a limited subset of the virome [17, 18, 148, 30, 24]. The viral-enrichment strategy has been imperative to face the technical challenges involved in virus assembly and identification from metagenomics due to the wealth of genetic remnants from other biological organisms, but also impose restrictions on the type of viral diversity studied [15, 37]. The costs and non-trivial implementation of *in vitro* viral enrichment is a strong motivator for alternative strategies to identify viruses in the growing number of metagenomic samples produced to study biodiversity in biotic and abiotic environments [149].

In this thesis, I have presented computational methods based on deep-learning frameworks that improve the recovery of both prokaryotic, viral and potentially other MGE genomes from bulk metagenomics **(Paper I and Paper II)**. Importantly, these methods can be applied across metagenomes collected from different environments and allow investigations into dominant hubs of viruses and bacteria in disease and co-evolutionary dynamics of bacteria and viruses. Our study on the gut microbiomes of people with extreme longevity illustrate an important strength of these methods as they can be applied to various metagenomic cohorts and facilitate combined bacterial and viral analysis to answer biological questions such as the human age-dependent impact on ecological viral communities **(Paper III)**. It is worth noting that the viromes characterized from bulk metagenomics without viral enrichment does not seem to capture the entirety of virome diversity and may be biased towards viruses infecting dominant host cells. RNA-viruses can be abundant in the human gut during disease [150], but their discovery is dependent on metatranscriptomics and construction of cDNA libraries [151]. Identification of Microviridae viruses might also be better captured with viral-enrichment, which however could be biased toward micro viruses and virulent viruses but miss larger bac-

teriophages and integrated proviruses [15, 152, 37]. Ideally, the microbiome should be studied using a combination of both approaches to capture the best picture of the entire virome simultaneously with studying larger organisms like bacteria. However, viral enrichment adds additional costs to a study with focus on the entire microbiome community as a result of further preparation and sequencing expenses. Therefore, a less costly compromise is a greater focus on maximizing virus discovery from bulk metagenomics, which has also been suggested to yield a comparable number of viral contigs to VLP preparations [37]. The extent to which virus genome quality and discovery in metagenomes can be improved using long-read technologies is an interesting topic which deserves more attention. Especially since long-read technologies have become a more cost-effective approach to study prokaryotic biodiversity in metagenomics [124, 125].

For future virome analysis in bulk metagenomes we propose a combined short and long-read sequencing approach to improve assembly and binning of bacterial MAGs (including integrated proviruses), viral MAGs and MGEs. In terms of the functional influence of viruses in an ecological space, there is a dire need for new computational models to explore the unannotated viral genomes. Fortunately, there is an increased adoption of deep language models on protein sequences [143], which could help accelerate the annotation process of the growing bulk of virus protein-coding genes. Computational models for ab initio structure modeling of virus proteins are available [144]. In addition, deep learning language models can be used to distill informative statistical embeddings of unannotated virus sequences which can be connected to functionally annotated proteins [153]. Altogether, improved annotation of virus gene-content should increase our understanding of the virus influence on bacterial constituents through predation mechanisms or by contribution of auxiliary metabolic genes in a provirus or episomal state, which have profound implications on the environment and host [48, 65]. In addition, there might be many complex mechanisms of bacterial and viral teamwork not discovered yet, such as how gut *Bacteroides spp.* benefit from proviruses that induce the release of inosine [138]. Our understanding of the human gut virome and its interplay with bacterial constituents is still in its infancy [68], but recent and new computational methods and databases will help to fuel future discoveries in metagenomic datasets.

7 Ethical and legal permits and approvals

All studies – encompassing cohorts of human subjects both healthy and diseased were granted the legal and ethical approvals for conducting the experiments, collecting samples and analyzing metagenomic samples, which are listed in the previous publications that describe each cohort for the first time. Here is a copy of ethical information for the main cohorts applied across project in the thesis.

1. "The HMP2 study was reviewed by the Institutional Review Boards at each sampling site: overall Partners Data Coordination (IRB 2013P002215), MGH Adult cohort (IRB 2004P001067), MGH Paediatrics (IRB 2014P001115); Emory (IRB IRB00071468), Cincinnati Children's Hospital Medical Center (2013-7586), and Cedars-Sinai Medical Center (3358/CR00011696). All study participants gave written informed consent before providing samples." [104]

2. "The COPSAC study was conducted in accordance with the guiding principles of the Declaration of Helsinki and was approved by the Capital Region of Denmark Local Ethics Committee (H-B-2008-093), and the Danish Data Protection Agency (2015-41-3696). Both parents gave oral and written informed consent before enrolment." [102]

3. For the Japanese centenarian study, "Fecal samples and blood tests from Japanese young and older participants, centenarians, and lineal relatives of centenarians were obtained following a protocol approved by the Institutional Review Board of Keio University School of Medicine (code 20150075 for young healthy donors; 20160297 for older cohorts (as part of the Kawasaki Ageing and Wellbeing project); and 20022020 for centenarians and lineal relatives of centenarians (as part of The Japan Semi-supercentenarian Study1)." [108]

8 Manuscripts

8.1 Paper I: Improved metagenome binning and assembly using deep variational autoencoders

Nissen, J.†; **Johansen, J**; Lundbye A.; R., Kaae S.; C, Almagro A.; J, Grønbech; C., Jensen J.; L., Nielsen B.; H., Petersen N, T.; Winther, O.; Rasmussen, S.† (2021). *Improved metagenome binning and assembly using deep variational autoencoders.* Nature Biotechnology, 555-560. DOI: 10.1038/s41587-020-00777-4

VAMB can be accessed at github via the following the link: `https://github.com/RasmussenLab/vamb`

nature biotechnology

LETTERS

https://doi.org/10.1038/s41587-020-00777-4

Improved metagenome binning and assembly using deep variational autoencoders

Jakob Nybo Nissen[1,2], Joachim Johansen[2], Rosa Lundbye Allesøe[2], Casper Kaae Sønderby[3], Jose Juan Almagro Armenteros[1], Christopher Heje Grønbech[3,4], Lars Juhl Jensen[2], Henrik Bjørn Nielsen[5], Thomas Nordahl Petersen[6], Ole Winther[3,4,7] and Simon Rasmussen[2]✉

Despite recent advances in metagenomic binning, reconstruction of microbial species from metagenomics data remains challenging. Here we develop variational autoencoders for metagenomic binning (VAMB), a program that uses deep variational autoencoders to encode sequence coabundance and *k*-mer distribution information before clustering. We show that a variational autoencoder is able to integrate these two distinct data types without any previous knowledge of the datasets. VAMB outperforms existing state-of-the-art binners, reconstructing 29–98% and 45% more near-complete (NC) genomes on simulated and real data, respectively. Furthermore, VAMB is able to separate closely related strains up to 99.5% average nucleotide identity (ANI), and reconstructed 255 and 91 NC *Bacteroides vulgatus* and *Bacteroides dorei* sample-specific genomes as two distinct clusters from a dataset of 1,000 human gut microbiome samples. We use 2,606 NC bins from this dataset to show that species of the human gut microbiome have different geographical distribution patterns. VAMB can be run on standard hardware and is freely available at https://github.com/RasmussenLab/vamb.

Metagenomic binning is the process of grouping metagenomic sequences by their organism of origin[1,2]. In metagenomic studies, binning allows the reconstruction of known and unknown genomes, enabling a broad description of the community and creating a starting point for further analysis of the organisms[3]. We developed a binning tool that uses deep learning in the form of variational autoencoders (VAE)[4,5] that integrates coabundance[6] and *k*-mer composition[7] data from metagenomics de novo assemblies and clusters the resulting latent representation into genome clusters and sample-specific bins. Our approach leverages multiple samples while simultaneously avoiding between-sample chimeras. It outperforms commonly used single-sample binning approaches by reconstructing 29–98% more NC genomes from simulated datasets, as well as 45% more NC genomes from a dataset of 1,000 human gut microbiome samples. Furthermore, our clustering method automatically groups per-sample bins into clusters with high taxonomic consistency, allowing precise strain-resolution taxonomic profiling.

Earlier work on metagenomics binning has mainly relied on the principles that DNA sequences originating from the same organism will have high covariance of their abundance signal across samples (coabundance) and that they share similar patterns of *k*-mer usage in their DNA (for example, 2–5 mer)[7–15]. Several attempts have been made to reconstruct thousands of microbial species from massive metagenomics datasets[16–18], independently assembling and binning each sample into genomes. These simple workflows allow for parallel analysis of samples, but do not leverage coabundance. Typical workflows using coabundance deal with sequence redundancy by either coassembling distinct samples or deduplicating sequences before binning[6,8–12]. This leads to intersample chimeric genomes that do not exist, which is especially problematic when strain-level variation can have important biological implications[19]. Furthermore, none of the existing methods leverage deep learning.

The main difference between our method, VAMB, and others is that it utilizes an unsupervised deep learning approach known as a VAE[10,18]. Second, our approach clusters the combined contig dataset from all samples without any preclustering or homology reduction and applies a strategy for splitting genome clusters after clustering (Supplementary Figs. 1–4). When applying this approach, which we term 'multisplit', each cluster should correspond to an organism and each bin in a cluster to a per-sample representation of the genome of that organism. To demonstrate the performance of VAMB compared to other binners, we benchmarked VAMB, Canopy[6], MetaBAT2 (ref. [12]) and MaxBin2 (ref. [11]) on five synthetic datasets from Critical Assessment of Metagenomic Interpretation (CAMI)[20] and one semisynthetic dataset from MetaHIT[21] samples (Supplementary Table 1). We assessed binning performance by counting the number of NC (>90% recall and >95% precision) genomes reconstructed as done in previous work[18,22]. VAMB reconstructed 29–98% more NC genomes at strain level compared to any of the other three binners (Fig. 1a and Supplementary Table 2). Interestingly, the increased performance of VAMB correlated (Pearson correlation coefficient = 0.90, linear regression $P = 0.035$) with the difficulty of the CAMI2 datasets, which we measured as the entropy of the genomes (Fig. 1b). Similarly, we found that VAMB reconstructed more genomes compared to MetaBAT2 at all levels of genome difficulty (Supplementary Fig. 5 and Supplementary Table 3). Additionally, we compared VAMB to ensemble binning, where bins from multiple programs are combined. Using DAS Tool[23], we tried combinations of the other binners and found VAMB to be better compared to all others (Fig. 1c). The addition of VAMB bins to DAS Tool improved the output of DAS Tool by up to 14% compared to VAMB only, but decreased performance on the Airways and MetaHIT datasets. VAMB and MetaBAT2 agreed on 39 NC genomes on average across the datasets and generated 49 and

[1]Department of Health Technology, Technical University of Denmark, Lyngby, Denmark. [2]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [3]Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [4]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark. [5]Clinical-Microbiomics A/S, Copenhagen, Denmark. [6]National Food Institute, Technical University of Denmark, Lyngby, Denmark. [7]Center for Genomic Medicine, Copenhagen University Hospital, Copenhagen, Denmark. ✉e-mail: simon.rasmussen@cpr.ku.dk

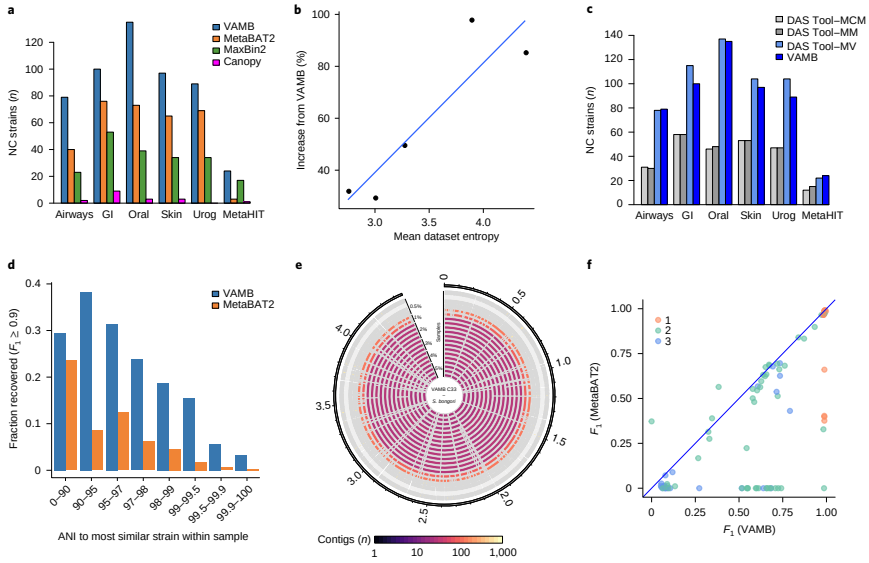


Fig. 1 | Performance of VAMB. a, Number of distinct NC strains recovered from the six benchmark datasets for VAMB (blue), MetaBAT2 (orange), MaxBin2 (green) and Canopy (magenta). **b**, Number of NC strains recovered by VAMB relative to MetaBAT2 as a function of mean sample entropy per dataset. Sample entropy was calculated as the Shannon entropy, with each contig an observation and each strain a class, and was used as a proxy for dataset complexity. **c**, Number of NC strains recovered when using the ensemble binner DAS Tool. We used the binning output from MetaBAT2, MaxBin2 and Canopy (DAS Tool–MCM, light gray), MetaBAT2 and MaxBin2 (DAS Tool–MM, gray), MetaBAT2 and VAMB (DAS Tool–MV, light blue) and VAMB (blue). **d**, Number of genomes recovered with F1 ≥ 0.9, stratified by the ANI to the most similar strain in the same sample across CAMI2 datasets. Blue, VAMB; orange, MetaBAT2. **e**, Alignment of sample-specific genome bins from VAMB cluster 33 to the *S. bongori* reference genome. The rings are ordered according to the number of *S. bongori* reads spiked into the HMP gut microbiome sample from 5% (inner) to 0.5% (outer), and colored according to the number of contigs in the particular sample. **f**, F1 of reconstructed genomes with VAMB and MetaBAT2 in the mixed-strain *Salmonella* spike-in experiment. A total of ten different *Salmonella* genomes were used, and between one and three genomes were added per HMP sample. Each dot represents F1 of a sample–genome pair and the color indicates how many *Salmonella* genomes were added to the particular sample: orange, 1; teal, 2, blue, 3. GI, gastrointestinal; urog, urogenital.

16 unique NC genomes on average, respectively. However, only a few NC genomes were unique to the combination (average, 1.6) and more NC genomes (average, 13) were lost (Supplementary Fig. 6).

To show that the superior performance of VAMB on strains was due to better binning, and not merely that VAMB defaults to a precision–recall tradeoff that happens to fit strain-level binning in our datasets, we tested the performance of VAMB at the species and genus levels. Here, VAMB on average reconstructed 14% more species than the second-best binner, MetaBAT2, which outperformed VAMB on only the CAMI2 Urogenital dataset. At genus level, VAMB and MetaBAT2 had similar performance with the former 4% better across all datasets, but MetaBAT2 outperformed VAMB on the CAMI2 Skin and Urogenital datasets (Supplementary Fig. 7 and Supplementary Tables 4 and 5). Furthermore, we tried to subsample the number of reads used for binning and found that VAMB performed well even with as few as 200,000 read pairs from each dataset (Supplementary Fig. 8).

One particularly difficult aspect of metagenomics binning is when multiple strains are present in a sample simultaneously. We therefore revisited the CAMI2 datasets, which contain a mixture of different strains and community complexities (Supplementary Fig. 9), and analyzed two simulated datasets originally created by Cleary et al.[24]. For the CAMI2 datasets we found that VAMB was able to bin more genomes with a weighted recall and precision ($F1 > 0.9$) across all intervals of strain abundances (Supplementary Fig. 10). Similarly, VAMB showed better performance across all intervals when assessing ANI than the most similar strain in the same sample. Here VAMB reconstructed 38% of all genomes as NC when the most similar strain had between 90–95% ANI, and even 15.5% of all genomes when the most similar strain had 99.0–99.5% ANI (Fig. 1d). For the Cleary et al. spike-in datasets, we first investigated a single spike-in with *Salmonella bongori* to a background of human gut microbiome samples[25]. Here VAMB created a single cluster (C33) where each sample-specific bin had $F1 = 1$ when 200,000 or more read pairs were added. Additionally, no contigs were assigned from any sample where we did not spike-in *S. bongori* reads, highlighting the ability of VAMB to group related strains across samples into a single cluster of sample-specific bins (Fig. 1e,

NATURE BIOTECHNOLOGY LETTERS

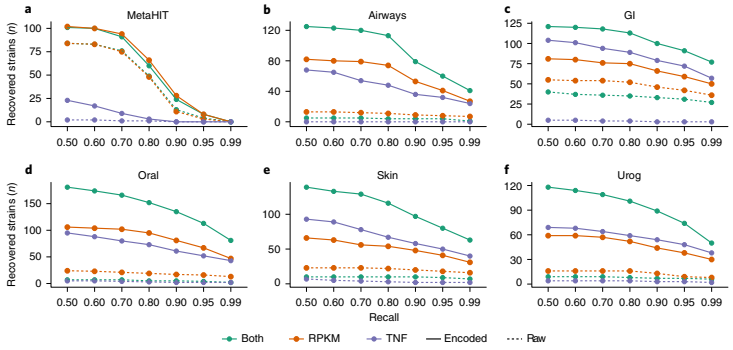


Fig. 2 | Performance of clustering different inputs. a–**f**, VAMB can effectively integrate coabundance and *k*-mer information (teal solid lines) to a clusterable representation that yields more reconstructed genomes than other combinations of the data, or using raw compared to encoded data. **a**, MetaHIT dataset. **b**, CAMI2 Airways dataset. **c**, CAMI2 Gastrointestinal (GI) dataset. **d**, CAMI2 Oral dataset. **e**, CAMI2 Skin dataset. **f**, CAMI2 Urogenital (Urog) dataset. Purple. *k*-mer frequency (TNF); orange, coabundance (RPKM); teal, concatenation of both *k*-mer and coabundance). Dashed lines, raw data input; solid lines, latent representation from the variational autoencoder in VAMB; *y* axis, number of distinct strains recovered at precision >0.95; *x* axis, increasing recall threshold of genomes.

Supplementary Fig. 11 and Supplementary Table 6). We then performed a second experiment where we spiked-in reads from ten different *Salmonella* genomes, with up to three *Salmonella* genomes per gut microbiome sample (Supplementary Data 1). Here, VAMB and MetaBAT2 were able to reconstruct 19 and 12 *Salmonella* strain–sample pairs, respectively, with $F1 > 0.9$ (Fig. 1f and Supplementary Data 2). As above, we quantified the ability of VAMB to distinguish between within-sample *Salmonella* genomes as a function of ANI. Here 14 genomes (78%) could be reconstructed ($F1 > 0.6$) when the other *Salmonella* genome had 90–91% ANI, eight genomes (57%) at 93–94% ANI and four (27%) at 98–99.5% ANI (Supplementary Fig. 12). Taken together, VAMB is able to distinguish between mixed strains at even 98–99.5% ANI, although the accuracy may be limited by the de novo assembly process for very similar genomes.

To test our hypothesis that the performance of VAMB stemmed in part from the VAE integrating information from both coabundance and *k*-mer composition, we compared the number of NC genomes produced by clustering of the raw coabundance data, raw *k*-mer composition or both raw datasets concatenated. Further, we compared to the bins produced by clustering their VAE latent spaces. For five of the six datasets, clustering the concatenation of raw data did not yield better results than the abundance or *k*-mer composition. However, for all datasets apart from MetaHIT, encoding of the concatenation gave the best results of all six input combinations, yielding 27 and 67% more NC genomes for our two validation datasets compared to the second-best combination (Fig. 2). Integrating the two data types with the VAE therefore results in a latent representation that is more informative than either of the inputs alone, and more amenable to clustering than the simple concatenation of the two raw data types. Furthermore, we investigated the effect of using different sizes of *k*-mers for encoding ($k = 2$–5) and, in line with previous work[7,10,26,27], found that $k = 4$ gave the best performance in three of the datasets (Supplementary Fig. 13). To test the importance of the probabilistic VAE encodings in VAMB, we tested a version of VAMB with the VAE replaced by a deterministic autoencoder. Here we found worse performance for all datasets, with the number of NC genomes from our two validation datasets dropping by 43 and 39%, respectively (Supplementary Fig. 14). We visualized the input space and latent encodings and, in line with our hypothesis, found that the VAE encoding appears to have genomes more clearly separated (Supplementary Fig. 15). Finally, we tested using *k*-means clustering rather than VAMB's iterative medoid clustering method. We found that VAMB's clustering algorithm was superior when using VAE-encoded data, and had the best overall performance compared to any combination of *k*-means clustering (Supplementary Fig. 16).

Single-sample binning workflows are popular because they are trivial to parallelize and inherently prevent intersample chimeras. We therefore tested the performance of VAMB on single samples of the CAMI2 datasets compared to MetaBAT2 and MaxBin2. While VAMB reconstructed most genomes on average, the differences were not significant (Wilcoxon rank-sum tests, two-tailed, $P > 0.05$) (Supplementary Fig. 17). We then compared the performance of the single-sample and multisplit approaches. Here, we found for all datasets that the multisplit approach was superior because the number of NC genomes rose from 1 to 24 for the MetaHIT dataset and increased by 28–105% for the five CAMI2 datasets (Supplementary Fig. 18). Importantly, using VAMB in multisplit mode was significantly better when measured across all datasets for as few as four samples (Supplementary Table 7 and Supplementary Data 3). Furthermore, we repeated the benchmark after discarding all 'easy' genomes (fewer than five contigs) and found that the improvement gain from multisplit was even more pronounced (109–282%; Supplementary Fig. 19). Because most metagenomics studies compare multiple samples of similar microbial communities with highly fragmented genomes, we expect that much higher-quality genomes can be recovered using VAMB and the multisplit approach.

One advantage of single-sample binning workflows is that they are inherently parallel and allow binning of large datasets[16–18]. To test the scalability of VAMB, we ran it on the entire benchmark dataset of Almeida and coworkers[18], consisting of 1,000 randomly selected human gut microbiome samples and a total of 5.9 million contigs (Supplementary Data 4). We used a single graphical processing unit (GPU) and ran VAMB in 12.4 h, 27 times faster than

running MetaBAT2 in single-sample mode (Supplementary Table 8 and Supplementary Fig. 20). To compare the quality of the resulting bins to those obtained by Almeida et al. using MetaBAT2, we estimated genome completeness and contamination and counted the number of bins estimated as NC. Using VAMB with default parameters, we obtained 5,036 NC bins compared to 3,480 for MetaBAT2, an increase of 1,556 NC bins (45%). We additionally tested VAMB with different hyperparameters and found that a slight decrease in the network yielded 5,288 NC bins, an increase of 252 NC bins (7.2% additional increase). For a fair comparison, we focused on the results of the default run and found that 2,517 of the NC bins were found by both methods whereas 1,019 and 1,500 NC bins generated by VAMB were medium quality (MQ) or missing from MetaBAT2, respectively. Similarly, MetaBAT2 reconstructed 480 and 483 NC bins that were MQ or missing from VAMB, respectively. Additionally, VAMB generated more MQ bins (5,169 versus 4,221) and therefore a larger number of MQ and NC bins in total (10,205 versus 7,701), as well as significantly more NC bins per sample (Wilcoxon signed-rank test, two-tailed, $V = 209{,}930$, $P = 7.1 \times 10^{-92}$) (Supplementary Fig. 21). Additionally, for the common set of 6,017 MQ or NC bins, median completeness, contamination and $F1$ were consistently better for VAMB ($F1 = 0.96$) compared to MetaBAT2 ($F1 = 0.94$) ($F1$, Wilcoxon signed-rank test, two-tailed, $V = 10{,}535{,}000$, $P = 7.8 \times 10^{-110}$) (Supplementary Fig. 22). However, because estimates of completeness based on conserved genes can be overestimated[28], we compared the sequence length of the common set of bins but found no significant difference (Wilcoxon signed-rank test, two-tailed, $V = 9{,}152{,}000$, $P = 0.20$) (Supplementary Fig. 22). Moreover, we compared NC bins that had an assembled genome at the National Center for Biotechnology Information (NCBI), and found that VAMB and MetaBAT2 bins were 10.5 and 14.3% shorter on average, respectively (Supplementary Fig. 23). For these bins we investigated the functional potential of extra contigs binned only by VAMB and found these to be highly enriched in Gene Ontology terms including 'Translation', 'Metal ion binding', 'Transposases' and more. Furthermore, we predicted phage-like contigs in 338 of 959 (35%) of these and identified a significantly increased AT nucleotide content compared to the entire bin (Wilcoxon rank-sum test, one-tailed, $W = 515{,}500$, $P = 2.4 \times 10^{-8}$) (Supplementary Fig. 23). Higher AT content is consistent with previous findings of horizontally transferred regions[29], and presumably reflects the ability of VAMB to recruit mobile genetic elements to the bins. Finally, we investigated taxonomic annotations and found a large overlap between the two sets. However, VAMB bins represented a larger taxonomic diversity from genus to genome level and, on average, reconstructed 97 more NC bins per phylum (Supplementary Fig. 24, Supplementary Tables 9 and 10 and Supplementary Data 5 and 6). Importantly, while VAMB is clearly better than MetaBAT2, Almeida et al. found no difference when using MetaBAT2 in single-sample mode, in coassembly mode or using the information from three different binners combined with MetaWRAP[30], replicating results from DAS Tool on the benchmark datasets.

As mentioned previously, another advantage of the multisplit approach is that a single cluster represents a particular organism across multiple samples. To test the phylogenetic consistency of clusters, we used 40 bacterial marker genes from the 5,036 NC bins to create a phylogeny (Fig. 3a). Here we found that 93.2% of clusters were monophyletic and that for 98.7% of the bins all leaves were extremely close to the cluster's central leaf, corresponding to >99% amino acid identity. Similar to the example with *S. bongori*, this implies that bins split from the same cluster are very closely related and represent different strains of the same species observed across samples. Zooming in on microdiversity, we analyzed the largest cluster, cluster 546, that contained 255 NC, 94 MQ and 115 low-quality bins. We found high taxonomic consistency with 92% of all contigs assigned to *B. vulgatus* and 5.7% to other *Bacteroides* species with slightly lower identity (Supplementary Data 7). These bins therefore represent 349 (NC and MQ) different individually de novo reconstructed *B. vulgatus* genomes in 349 human gut microbiomes. If we compare this to using an approach based on ANI > 95%, such as used in other large-scale, single-sample binning studies[17,18], *B. vulgatus* and *B. dorei* would have been merged into one species (Supplementary Fig. 25) rather than clusters 546 and 94, respectively.

One advantage of VAMB reconstructing more NC bins is increased statistical power when investigating associations with metadata. We therefore reconstructed the phylogeny of the 255 NC bins from the *B. vulgatus* cluster (Fig. 3b and Supplementary Fig. 26). We verified the phylogeny by considering samples ($n = 18$) with multiple sequencing runs ($n = 40$) and found the samples to be placed either monophyletically or with very short distances between them (Fig. 3b and Supplementary Fig. 26). When comparing phylogenetic placement of *B. vulgatus* strains to the recorded metadata, we found phylogenetic distance to be significantly associated with the geographical location of the sample (permutational multivariate analysis of variance (PERMANOVA), adjusted $P = 0.007$, $F = 2.26$, degrees of freedom = 216), although only at a low coefficient of determination ($R^2 = 0.02$). European and North American samples did not cluster exclusively, and Asian *B. vulgatus* strains were interspersed throughout the entire tree. Previous work comparing North America and Europe found a similar trend for *B. vulgatus*[31], although another study investigating Ca. *Cibiobacter qucibialis* found a clade associated with Chinese samples[17]. Furthermore, previous work based on sample taxonomy and community structure has shown a clear association with geographical location of the sample[17,32–34]. We therefore expanded our analysis to all clusters with 20 or more NC bins ($n = 52$ and $n = 2{,}606$ NC bins in total) and found significant association with geographical location for 34 of the 52 clusters (adjusted $P < 0.05$) (Fig. 3c and Supplementary Data 8). However, the effect of geographical location (R^2) was markedly different between the clusters and was not associated with whether they corresponded to known or unknown species. We found clear differences between the families (Fig. 3d) and Bacteroidaceae, which included six different *Bacteroides* species, had the lowest overall association (median $R^2 = 0.03$) (Supplementary Fig. 27). On the contrary, taking species of Bifidobacteriaceae and Lachnospiraceae as examples, these showed much higher variance in R^2, from 0.06 to 0.50 and 0.08 to 0.44, respectively (Fig. 3d and Supplementary Fig. 27). These results indicate that strains of certain gut microbiome species are ambiguously distributed whereas others are geographically restricted. This could be due to either diet or to how well a species adapts to differences in host genetics, but could also be influenced by difference in transmission mode—for instance, vertical transmission (mother–child inheritance)[35,36].

Here, by combining metagenomics binning with unsupervised deep learning, we show improvements compared to state-of-the-art methods across datasets of different types and sizes. We also show that the VAE automatically learns how to integrate two distinct data types—in this case, coabundance and k-mer composition—and that the resulting latent representation clusters better than either of the inputs. This is, in principle, not limited to two input data types and it is possible to add additional data as input to the VAE. For VAMB we avoided using more complex models such as, for example, Gaussian mixture VAEs[37,38], and designed our method to be feasible for standard users. For instance, a standard laptop without GPU acceleration could process the six benchmark datasets each in <6 h with 1 GB random-access memory (Supplementary Table 8). Finally, we believe that the importance of our findings is not limited to the field of microbiome and metagenomics, because data integration is a central process in many fields of life science research. Future discoveries within precision medicine will be greatly enhanced by

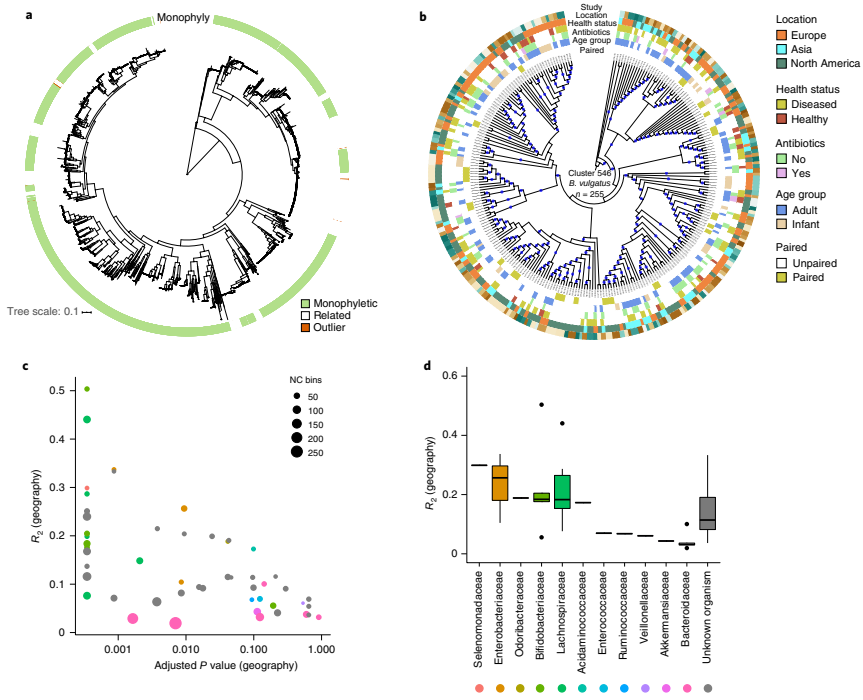


Fig. 3 | Phylogeny of bins across 1,000 human gut microbiome samples. a, Amino acid-level bacterial marker gene maximum likelihood tree for all 5,036 NC bins. Despite VAMB having no phylogenetic information, a large majority of bins are monophyletic and misplaced bins are generally very similar to neighboring clades. Green, leaf is in a monophyletic or extremely closely related bin (>99% aa identity); white, leaf is in a cluster with one or more outliers; orange, leaf is an outlier compared to the medoid of the cluster. **b**, Cladogram of ASTRAL species tree generated from 2,433 gene trees from cluster 546 containing 255 NC bins of *B. vulgatus*. ASTRAL local posterior probabilities branch support is indicated as a blue circle when support is >0.95. The tree is rooted on sample SRR341600, which is the most basal *B. vulgatus* in the CheckM tree (**a**). Rings, from inner to outer: 1, paired samples are in the same clade (green), not paired (white) or paired and not in same clade (red); 2, age group—infant (beige) or adolescent/adult (blue); 3, individual used antibiotics (pink) or not (green); 4, individual has a disease (green) or is healthy (red), 5, geographical origin: Europe (orange), Asia (turquoise), Americas (teal); 6, study origin (multiple colors). White in rings 2–6 (no color) indicates missing data. **c**, Association of phylogenetic distance and geographical location of sample. PERMANOVA *P* values adjusted using Benjamini–Hochberg for geography are shown on the *x* axis, and the coefficient of determination (R^2) is shown on the *y* axis. Point sizes indicate the number of NC bins in the cluster, and colors indicate family (see **d** for color coding). Unannotated clusters (<50% annotated with BLAST) are set as unknown (gray). Values left-most on the *x* axis indicate adjusted $P < 4 \times 10^{-4}$. **d**, We summarized R^2 from **c**, showing that geography explains different amounts of variation for families. Color coding for each family is shown as points below the *x* axis. **c**,**d**, Exact *P* values, *F*-statistic, degrees of freedom and number of observations (*n*) are available in Supplementary Data 8. The lower and upper hinges correspond to the first and third quartiles (25th and 75th percentiles). The upper and lower whiskers extend from the hinge to the highest and lowest values, respectively, but no further than 1.5× interquartile range (IQR) from the hinge. IQR is the distance between the first and third quartiles. Data beyond the ends of whiskers are outliers and are plotted individually.

data integration across several omics datasets. To achieve this, deep learning methods such as VAEs or other models represent promising approaches.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-020-00777-4.

Received: 6 December 2019; Accepted: 17 November 2020;
Published online: 4 January 2021

References

1. Turaev, D. & Rattei, T. High definition for systems biology of microbial communities: metagenomics gets genome-centric and strain-resolved. *Curr. Opin. Biotechnol.* **39**, 174–181 (2016).
2. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
3. Wang, J. & Jia, H. Metagenome-wide association studies: fine-mining the microbiome. *Nat. Rev. Microbiol.* **14**, 508–522 (2016).
4. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. Preprint at https://arxiv.org/abs/1312.6114 (2014).
5. Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *Proc. Mach. Learn. Res.* **32**, 1278–1286 (2014).
6. Nielsen, H. B. et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
7. Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. & Glöckner, F. O. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.* **6**, 938–947 (2004).
8. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
9. Albertsen, M. et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).
10. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ.* **3**, e1165 (2015).
11. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
12. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ.* **7**, e7359 (2019).
13. Plaza Oñate, F. et al. MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics* **35**, 1544–1552 (2019).
14. Lin, H. H. & Liao, Y. C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci. Rep.* **6**, 24175 (2016).
15. Chatterji, S., Yamazaki, I., Bai, Z. & Eisen, J. A. CompostBin: A DNA Composition-Based Algorithm for Binning Environmental Shotgun Reads. in Research in Computational Molecular Biology (eds. Vingron, M. & Wong, L.) 17–28 (Springer, 2008).
16. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
17. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662 (2019).
18. Almeida, A. et al. A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
19. Brooks, B. et al. Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. *Nat. Commun.* **8**, 1–7 (2017).
20. Sczyrba, A. et al. Critical Assessment of Metagenome Interpretation – a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
21. Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
22. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
23. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
24. Cleary, B. et al. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat. Biotechnol.* **33**, 1053–1060 (2015).
25. Huttenhower, C. et al. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
26. Saeed, I., Tang, S.-L. & Halgamuge, S. K. Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. *Nucleic Acids Res.* **40**, e34 (2012).
27. Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. & Blaser, M. J. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res.* **13**, 145–156 (2003).
28. Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).
29. Daubin, V., Lerat, E. & Perrière, G. The source of laterally transferred genes in bacterial genomes. *Genome Biol.* **4**, R57 (2003).
30. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
31. Schloissnig, S. et al. Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).
32. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
33. Deschasaux, M. et al. Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. *Nat. Med.* **24**, 1526–1531 (2018).
34. He, Y. et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat. Med.* **24**, 1532–1535 (2018).
35. Asnicar, F. et al. Studying vertical microbiome transmission from mothers to infants by strain-level metagenomic profiling. *mSystems* **2**, e00164–16 (2017).
36. Ferretti, P. et al. Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell Host Microbe* **24**, 133–145 (2018).
37. Grønbech, C. H. et al. scVAE: variational auto-encoders for single-cell gene expression data. Bioinformatics 36, 4415–4422 (2020).
38. Dilokthanakul, N. et al. Deep unsupervised clustering with Gaussian mixture variational autoencoders. Preprint at https://arxiv.org/abs/1611.02648 (2017).

Methods

Overview of VAMB. The input to the VAMB pipeline is (1) a catalog of metagenomic sequences to be binned and (2) their abundances. The VAMB pipeline consists of three major steps (Supplementary Fig. 1). For each sequence in the catalog, the per-sequence tetranucleotide frequencies (TNF) for all possible canonical tetramers are calculated and the abundance of each sequence is estimated based on read mappings. These tables are concatenated and used to train a VAE tabula rasa (Supplementary Fig. 2). After training, the DNA sequences and coabundance information of the sequence catalog are encoded to the mean of their latent distributions. This latent representation is then clustered through an online iterative medoid clustering algorithm that dynamically estimates clustering threshold in cosine distance space. VAMB can be run in three different workflows: single-sample approach, where each sample is binned independently, a multisample approach on a coassembly or a multisplit approach (Supplementary Fig. 3). For the single-sample approach, normalized abundances are used whereas intersample abundance ratios (coabundance) are used for the multisample approaches. Finally, in the multisplit approach each cluster is split into sample-specific bins of the particular organism (Supplementary Fig. 4).

Computation of abundance and TNF. For each sequence, the frequencies of each tetramer not containing ambiguous bases were calculated to obtain TNFs[7]. TNFs were projected into a 103-dimensional orthonormal space as done in other work[39]. Thus, for n sequences the output was a table, $n \times 103$. To determine abundance, we counted the number of individual reads mapped to each sequence. If a read was mapped to n sequences, it counted $1/n$ towards each. The read counts were normalized by sequence length and total number of mapped reads, such that abundance was given in reads per kilobase sequence per million mapped reads (RPKM). With s samples and n sequences, the abundance output was a table, $n \times s$. Abundance values were normalized across samples to sum to 1, mimicing a probability distribution that was reconstructed from the final VAE by applying softmax to the abundance output neurons. Finally, TNFs were normalized by z-scaling each tetranucleotide across the sequences to increase the relative intersequence variance.

Architecture of the VAE. Each sequence was input to the VAE as an abundance vector $\mathbf{A}_{\mathbf{in}}$ of length s and a TNF vector $\mathbf{T}_{\mathbf{in}}$ of length 103 (Supplementary Fig. 2). These were concatenated to a vector of length $s+103$ before being passed through the hidden encoding layers consisting of two fully connected layers, each using batch normalization[40] and dropout[41] ($P = 0.2$). The output of the last layer was passed to two different, fully connected layers of length N_l, termed the μ and σ layers. The latent layer, l, is of length N_l obtained by sampling the Gaussian distribution using the μ and σ layers as parameters—for example, $l_i \sim N(\mu_i, \sigma_i)$ for each neuron $i = 1,...,N_l$. The sampled latent representation was then passed through the hidden decoding layers, identical in size to the hidden encoding layers except arranged in reverse order. Finally, the last hidden decoding layer was connected to a $s+103$ fully connected layer, which was split into two output vectors, $\mathbf{A}_{\mathbf{out}}$ and $\mathbf{T}_{\mathbf{out}}$, of length s and 103, respectively. We used leaky rectified linear units[42] as activation functions except for the μ and σ layers, which used linear and softplus activation, respectively. Furthermore, for the last layer generating the reconstructions we used softmax activation for generating $\mathbf{A}_{\mathbf{out}}$ mimicking a probability distribution and linear activation for $\mathbf{T}_{\mathbf{out}}$. After training, the input sequences were encoded by passing them through the VAE and extracting the values of the μ layer. The VAE models were trained using the Adam optimizer[43] and one Monte Carlo sample of the Gaussian latent representation. The VAE was implemented using PyTorch[44] (v.1.2.0), and CUDA (v.10.1.243) was used when running on a GPU.

Loss function. When training the VAE with s samples and N_l hidden neurons, the failure to reconstruct the input was penalized by the reconstruction error, consisting of an abundance error (E_{ab}) and a TNF error (E_{TNF}), defined as

$$E_{ab} = \sum \ln\left(\mathbf{A}_{\mathbf{out}} + 10^{-9}\right)\mathbf{A}_{\mathbf{in}}, \; E_{TNF} = \sum\left(\mathbf{T}_{\mathbf{out}} - \mathbf{T}_{\mathbf{in}}\right)^2$$

that is, using cross-entropy (CE) and the sum of squared errors (SSE), respectively. When running on a single sample, E_{ab} was defined using SSE, because CE on a single normalized value trivially is zero. To regularize the model, the distribution given by the μ and σ layers was constrained by a prior $N(0,I)$, by penalizing the deviance from this distribution with the Kullback–Leibler divergence:

$$D_{KL}(\text{latent}|\text{prior}) = -\sum \frac{1}{2}\left(1 + \ln(\sigma) - \mu^2 - \sigma\right)$$

Finally, the combined model loss was then

$$L = w_{ab}E_{ab} + w_{TNF}E_{TNF} + w_{KLD}D_{KL}$$

where the weighting terms are defined as $w_{ab} = (1-\alpha)\ln(s)^{-1}$, $w_{TNF} = \alpha/103$ and $w_{KLD} = (N_l\beta)^{-1}$. The parameters α and β were set to 0.15 and 200, respectively. For values of loss, E_{ab}, E_{TNF} and D_{KL} represent the six benchmark datasets (Supplementary Fig. 20).

Clustering. Clustering of the latent space was done using an iterative medoid clustering algorithm inspired by Nielsen et al.[8] based on cosine distances between encodings. The algorithm works in two steps (Supplementary Fig. 4): (1) an arbitrary point is chosen to be medoid. The medoid's 'neighbors' are defined as any points within a distance of 0.05 in cosine distance space. VAMB then randomly samples points from the neighbors and, if any point has more neighbors than the medoid, this becomes the new medoid. When VAMB has futilely sampled 25 neighbors in a row or tried all neighbors, go to step 2. (2) The distances from the medoid to all other points are calculated and a histogram is created. A heuristic function checks whether the histogram is composed of a 'near' peak of close points and a 'far' peak of further points separated by a deep valley with fewer points in intermediate distance from the medoid. 'Deep' is initially defined as the valley minimum being <0.1× the maximum of the small peak. If a deep valley is found, all points closer than the valley minimum are removed as a cluster; if not, the medoid is ignored. VAMB checks how often a medoid has been ignored: if >185 of the last 200 tries, the definition of 'deep' is increased by 0.1; if 'deep' is already 0.6, VAMB will ignore the valley's minimum and instead remove all points within an adaptive cosine distance as a cluster. This distance is determined as the median distance from all previous clusters. The method was implemented for both central processing unit (CPU) and GPU usage.

Benchmarking datasets. We used four training and two holdout datasets. One training dataset was the MetaHIT 'error-free' dataset ($n = 264$) originally created by Kang and coworkers[45] while the other three were datasets from CAMI[20], where we used the sample-specific assemblies from three of the five CAMI2 'toy' human short-read datasets: CAMI2 Airways ($n = 10$), CAMI2 Oral ($n = 10$) and CAMI2 Urogenital ($n = 9$). Our holdout datasets were the other two, CAMI2 Skin ($n = 10$) and CAMI2 Gastrointestinal ($n = 10$). We originally also tested VAMB on the CAMI High dataset ($n = 5$) but, due to an unrealistic contig size distribution (Supplementary Fig. 28) influencing both abundance and TNF estimation, we discarded the dataset (see Supplementary Table 1 for an overview of the datasets). For all datasets we used only contigs >2,000 base pairs (bp) as input to VAMB. For the MetaHIT error-free dataset we used the abundance table supplied from Kang and coworkers (originally created using the script jgi_summarize_bam_contig_depths from MetaBAT) and the contigs as input to VAMB with default parameters. For each of the CAMI2 datasets we aligned the synthetic short paired-end reads from each sample using bwa-mem (v.0.7.15)[45] to the concatenation of per-sample contigs from the particular dataset. BAM files were sorted using samtools (v.1.7)[46] and abundances calculated using jgi_summarize_bam_contig_depths from MetaBAT2 (v.2.10.2)[12]. The jgi-abundance table and contig sequences were input to VAMB and run using default parameters with bin splitting enabled.

Benchmarking. When benchmarking a set of bins against a set of genomes, we matched each bin with each genome and defined the number of nucleotides in the genome covered by any contig from the bin as true positives. The total number of covered nucleotides of other genomes from contigs in that bin represented the false positives, and number of nucleotides in the genome that were covered by any contig in the dataset, but not by any contig in the bin, represented the false negatives. A genome was considered recovered at a particular recall–precision threshold pair if any bin matched with the genome reached or exceeded those precision and recall thresholds. For the CAMI2 datasets we used their definitions of strain, species and genus taxonomic levels[20]. Sczyrba et al. aligned extracted marker genes for each genome and aligned these to a 16S RNA alignment, clustered the alignment and then assigned a taxonomy based on that clustering (see Supplementary Note 1 of Sczyrba et al.'s paper for details)[20]. For the MetaHIT dataset, strain was defined as the individual reference genomes that were used to create the dataset, while species and genus levels were defined using NCBI taxonomy of the given reference genome. When comparing the performance of VAMB with other binners, we used Canopy from the original version (published in 2014) and ran it with default parameters. MetaBAT2 (v.2.10.2)[12] was run with default parameters, except setting–minClsSize=1, so that it would not discard small but accurate bins. MaxBin2 (v.2.2.4)[11] was run with default parameters. For all runs we used default parameters of VAMB as determined in the hyperparameter searches. For DAS Tool (v.1.1.1)[13] we used a combination of bins as input and default parameters. Benchmarking of subsampling reads for the CAMI2 datasets was done by randomly sampling reads for each sample to between 200,000 and 10 million read pairs and then running VAMB and benchmarking as described above. For comparison to k-means we used minibatch k-means[47] implemented in scikit-learn: 'MiniBatchKMeans (n_clusters=750, random_state=0, batch_size=4096, max_iter=25, init_size=20000, reassignment_ratio=0.02)'.

Hyperparameter search. To identify the best hyperparameters of the VAE, we developed it using four training datasets (MetaHIT, CAMI2 Oral, CAMI2 Airways and CAMI2 Urogenital) and used two other datasets (CAMI2 Skin and CAMI2 Gastrointestinal) as held-out test sets. We first varied each hyperparameter while keeping the others fixed and assessed the resulting bins. In the second round of optimization, we tested various hyperparameter combinations to select the final, best-performing values (Supplementary Figs. 29–32 and Supplementary Data 9). For single-sample analyses we used 256 neurons in two hidden layers with

32 latent neurons, no dropout, minibatch size of 128 and doubling after 25, 75, 150 and 300 epochs to a final total of 2,048, a learning rate of 10^{-3} and trained for 500 epochs. For the multisample approaches we similarly used default parameters, which were 512 neurons in two hidden layers with 32 latent neurons, 0.2 dropout, minibatch size of 128 that doubled after 25, 75, 150 and 300 epochs to a final total of 2,048, a learning rate of 10^{-3} and trained for 500 epochs. For the effect of the number of epochs, see Supplementary Fig. 33.

Strain-mixing datasets. We replicated the two *Salmonella* spike-in simulation experiments from Cleary et al.[24]. For the spike-in of a single genome we used *S. bongori* NCTC 12419 (NC_015761), where we simulated Illumina paired-end reads using ART (v.2.5.8)[48]. The reads were simulated as 100-nt error-free pairs with 300-nt insert size and a standard deviation of 10 nt. The reads were added in amounts of between 100,000 and 1 million read pairs (0.5–5% of total reads) to 30 different Human Microbiome Project (HMP) human gut microbiome samples, so that the total was 20 million read pairs (Supplementary Table 6). For the mixed-strain spike-in dataset we used three *Salmonella enterica* subsp. *enterica* serovar *Typhimurium* genomes (NC_003197, NC_016810, NC_022544), five non-Typhimurium *S. enterica* subsp. *enterica* serovar genomes (NC_010067, NC_011094, NC_021812, NC_021902, NC_022221) and two *S. bongori* genomes (NC_015761, NC_021870). We simulated read pairs as described above and added them to 19 million read pairs from 50 different HMP human gut microbiome samples (Supplementary Data 1). For both experiments, each sample was de novo assembled individually using SPAdes (v.3.9.0)[49] with the --meta flag, and scaffolds from each sample >2,000 bp were added to a combined scaffold set for each dataset. Reads were then mapped to the combined scaffold set using Minimap2 (v.2.15r905)[50], sorted using samtools (v.1.7)[46] and abundances calculated using jgi_summarize_bam_contig_depths from MetaBAT2 (v.2.10.2)[12]. These were combined into one file and used as input to VAMB, with default parameters and bin splitting enabled. MetaBAT2 (v.2.10.2)[12] was run in single-sample mode using default parameters on all samples. To assign scaffolds to reference genomes we used blat (v.385)[51] with default parameters and the *Salmonella* genomes as database and accepted all hits of length ≥500 nt, and with ≥99.5% identity for the single-strain spike-in experiment and 99.9% identity for the mixed-strain spike-in. When assessing whether *Salmonella* genomes could be reconstructed, we used $F1 > 0.9$ and 0.6 for the single- and mixed-strain spike-in experiments, respectively. We used the lower threshold for the mixed-sample spike-in to account for de novo assembly, creating chimeric sequences when strains were very similar. To assign true positives we used the length of the blat hit and, to determine false positives, we used the entire length of the contig (Benchmarking). The plot of *Salmonella* alignments to the reference genome was done using Circos (v.0.69.9)[52]. For genome comparisons of the *Salmonella* and CAMI2 datasets we used FastANI (v.1.1)[53] with default parameters to calculate ANI. Within each sample we then determined the distance to the most similar genome. For abundance of each genome we used that supplied with the CAMI2 datasets. The two *Salmonella* datasets (30 and 50 gut microbiome samples) were run in 4 and 8 h, respectively, using 24 CPU cores.

Calculation of dataset entropy and genome difficulty. We determined entropy of the datasets by calculating the Shannon diversity (SD) of individual samples in each of the CAMI2 datasets. Here SD was calculated based on the number of contigs per strain in a sample. We defined the entropy of a dataset as the mean SD across all its samples. MetaHIT was excluded because no per-sample annotation was available (coassembly data). Genome difficulty was determined as the minimum number of contigs needed to reconstruct the genome at 90% recall.

Binning a large dataset of the human gut microbiome. We obtained de novo assemblies of 1,000 human gut microbiome samples from Almeida et al.[18]. These samples had been randomly selected across datasets in the European Nucleotide Archive (ENA), and we obtained the exact assemblies that were used in that particular work. The assemblies had been created using SPAdes (v.3.10.0)[49] with the flag --meta. Similar to their approach, we used only contigs >2,000 bp and accepted only bins ≥200 kb. We downloaded the reads from each sample from ENA and verified that we had precisely the same number of reads as reported for each sample in Almeida et al. Hereafter we used Minimap2 (v.2.15r905)[50] to map reads from each sample to the pooled set of contigs from all samples and sorted the alignments using samtools (v.1.7)[46]. We then calculated the abundances of each sample with jgi_summarize_bam_contig_depths from MetaBAT2 (v.2.10.2)[12] and combined the abundance information into one file. This abundance information was used as input to VAMB, together with the combined fasta of the contigs and run with default settings. Training and clustering were done on a NVIDIA Tesla V100 GPU. When running with a smaller network than default, we used 24 latent neurons and 384 hidden neurons. We used CheckM (v.1.0.18)[54] to estimate the completeness and contamination of each bin and compared these to the results of Almeida et al. (Supplementary Data 10). Because Almeida et al. did not include archaeal bins in their data, we downloaded the samples where VAMB had produced an archaeal genome of any quality ($n = 27$) and ran MetaBAT2 using the same parameters as in Almeida et al. This yielded 11 NC and nine medium-quality (MQ) archaeal bins, which we added to the MetaBAT2 set; VAMB generated 15 NC and six MQ archaeal bins from the same samples. For comparison we used the definition of NC bins from their work as >0.9 completeness and <0.05 contamination, MQ as >0.5 completeness and <0.1 contamination and we defined low-quality bins as those not passing NC or MQ criteria. $F1$ was calculated as $2\times(\text{precision}\times\text{recall})/(\text{precision}+\text{recall})$, where recall was set as CheckM completeness and precision as 1 – CheckM contamination. Annotation of cluster 546 was done using ncbi-blastn (v.2.8.1)[55] against the nonredundant nucleotide database (nt), and filtered for 90% identity and 500-nt alignment length. Abundance of each cluster was determined from jgi_summarize_bam_contig_depths, calculated above from the alignments. First the weighted average was determined for each bin in a cluster weighting the read abundance with contig length. Hereafter, the abundance of the clusters was determined as the sum of each bin in the particular cluster. The abundance matrix, CheckM results and bins in fasta format for all clusters are available for download (Data Availability). To identify which bins overlapped each other from VAMB and MetaBAT2 runs, we used MASH (v.2.0)[56] with 10,000 sketches per bin to compare the two sets. We then assigned corresponding bins between VAMB and MetaBAT2 from MASH distance ≤0.01 and confirmed that the bins were from the same sample. Using this approach, we could match 6,017 bins between the two datasets. NCBI-assembled genome lengths were obtained from ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/ASSEMBLY_REPORTS/ANI_report_bacteria.txt. For the analysis of functional differences between VAMB and MetaBAT2 bin pairs, we predicted protein sequences using Prodigal (v.2.6.3)[57] and annotated them using InterProScan (v.5.36–75.0)[58]. We counted gene ontologies in the common and unique sets of each bin and used a two-sided Fisher's test with Benjamini–Hochberg correction[59] to determine VAMB unique contig-enriched GOs (adjusted $P < 0.005$). We scanned for phages using CheckV (v.0.4.0)[60] and accepted a hit if CheckV completeness was >40%. Additionally, we used DeepVirFinder (v.1.0)[61] and accepted hits with $P < 0.01$. Annotation of bins using GTDB (release 89)[62] was done using GTDB-TK (v.1.1.0)[63] and the function classify_wf. We estimated run time for MetaBAT2 on the Almeida dataset by running it in single-sample mode on 50 random samples and extrapolating to 1,000 samples.

Phylogeny. For the provisional taxonomic assignment of the Almeida et al. dataset clusters, contigs were aligned with ncbi-blastn (v.2.8.1)[55] against nonredundant nt_v5, retaining for each contig the best hit with >90% nucleotide identity over 500 nucleotides. The cluster was assigned the species with most hits. For the marker gene tree, we concatenated the core gene amino acid alignments created by CheckM (v.1.0.18)[54] and ran IQ-TREE (v.1.6.8)[64] with the LG model and one partition per gene. For the individual trees of cluster 546 (*B. vulgatus*) and the 51 other clusters with at least 20 NC bins, we first used Prodigal (v.2.6.3)[57] to infer genes of the bins and then used SonicParanoid (v.1.3.0)[65] on protein sequences using the 'fast' mode to identify orthologous groups per cluster. Here we accepted all orthologous groups of proteins when they had the same number of proteins as the number of seed orthologous down to 90% of the number of NC bins in that cluster. In other words, for cluster 546 with 255 NC bins we accepted all orthologous groups of proteins when there were 255–230 proteins and the same number of seed orthologous in the group. We then extracted the DNA sequence for the genes and aligned each gene using MAFFT (v.7.453)[66] and the '—auto' option. We then reconstructed a tree for each gene using IQ-TREE (v.1.6.8)[64] using automated model selection for each gene[67]. For each cluster we then used all gene trees as input to ASTRAL-III (v.5.7.3)[68] to build a species tree, and calculated branch lengths on the tree using IQ-TREE (v.1.6.8)[64] where the ASTRAL tree was input as constrained topology. Here the gene sequences were concatenated into a supermatrix and we set a partition for each gene with automated model selection. Bootstrap support were calculated by UFBoot2 (ref. [69]) using ASTRAL. Trees were visualized using ITOL (v.5.2)[70]. Association between phylogenetic placement and metadata (location and study) was done using PERMANOVA implemented in the R package vegan (v.2.5–6)[71] using the function adonis2. If multiple NC strains originated from different sequencing runs of the same sample, one was randomly selected as the representative. Leaf distances were extracted from each phylogenetic tree using the R package ape (v.5.3)[72] with the function 'cophenetic.phylo', and the model used for adonis2 was '$d \approx$ Location + Study', where d is the phylogenetic distance, Location is Asia, North America and Europe and Study is given in Supplementary Table 1 from Almeida et al.

Statistics. For our analyses we used Wilcoxon signed-rank tests (paired data) and Wilcoxon rank-sum tests (unpaired data) implemented in R as the function wilcox.test to test for statistical significance between distributions. When investigating for GO enrichment we used a two-tailed Fisher's test implemented in R as the function fisher.test. Furthermore, we used PERMANOVA with 9,999 permutations implemented in the R package vegan as the function adonis2. Adjustment for multiple testing was done using Benjamini–Hochberg correction implemented in R and the function p.adjust. Sample sizes (n) and test statistics for all tests are given either in the text or the respective Supplementary figure, table or data.

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

Data availability

The sequence data used in this study are publicly available from either the respective studies or ENA. The semisynthetic MetaHIT dataset was downloaded from https://portal.nersc.gov/dna/RD/Metagenome_RD/MetaBAT/Files/ as the files depth.txt.gz and assembly-filtered.fa.gz. The simulated CAMI High and CAMI2 datasets were downloaded from https://data.cami-challenge.org/participate from 'Toy Test Dataset High_Complexity' and '2nd CAMI Toy Human Microbiome Project Dataset', respectively. The de novo assemblies of the Almeida dataset were obtained through personal communication with A. Almeida and R. D. Finn, and the reads downloaded from ENA as specified in their publication. The data and results of binning the MetaHIT, CAMI2 and Almeida datasets, as well as the source data for Figs. 1–3, are available on figshare at https://figshare.com/projects/VAMB/72677. A CodeOcean capsule of VAMB v.3.0.1, including the six training and test datasets for reproducing benchmarking results, is available from https://doi.org/10.24433/CO.2518623.v1. Source data are provided with this paper.

Code availability

All code can be found on GitHub at https://github.com/RasmussenLab/vamb and is freely available under the permissive MIT license. All analyses were performed using VAMB v.3.0.1. Additionally, code are available as a CodeOcean capsule at https://doi.org/10.24433/CO.2518623.v1.

References

39. Kislyuk, A., Bhatnagar, S., Dushoff, J. & Weitz, J. S. Unsupervised statistical clustering of environmental shotgun sequences. *BMC Bioinform.* **10**, 316 (2009).
40. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Preprint at https://arxiv.org/abs/1502.03167 (2015).
41. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. Preprint at https://arxiv.org/pdf/1207.0580.pdf (2012).
42. Maas, A. L., Maas, A. L., Hannun, A. Y. & Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. Preprint at https://arxiv.org/pdf/1207.0580.pdf (2013).
43. Kingma, D. P. & Ba, J. L. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2017).
44. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 8026–8037 (2019).
45. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997?upload=1 (2013).
46. Li, H. et al. The sequence alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
47. Sculley, D. Web-Scale k-Means Clustering. in *Proc. 19th International Conference on World Wide Web* 1177–1178 (ACM Press, 2010).
48. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).
49. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. MetaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
50. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
51. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
52. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
53. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat. Commun. 9, 5114 (2018).
54. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
55. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
56. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
57. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).
58. Mitchell, A. L. et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).
59. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B* **57**, 289–300 (1995).
60. Nayfach, S., Pedro Camargo, A., Eloe-Fadrosh, E. & Roux, S. CheckV: assessing the quality of metagenome-assembled viral genomes. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.06.081778 (2020).
61. Ren, J. et al. Identifying viruses from metagenomic data using deep learning. *Quant. Biol.* **8**, 64–77 (2020).
62. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
63. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).
64. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
65. Cosentino, S. & Iwasaki, W. SonicParanoid: fast, accurate and easy orthology inference. *Bioinformatics* **35**, 149–151 (2018).
66. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
67. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
68. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinform.* **19**, 153 (2018).
69. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation.*Mol. Biol. Evol.* **35**, 518–522 (2018).
70. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
71. Oksanen, J. et al. Package 'vegan'. Community Ecology Package v.2.5-6. R Package version 3.4.0 1–296. https://cran.r-project.org/src/contrib/Archive/vegan/vegan_2.5-6.tar.gz (2019).
72. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2018).

Acknowledgements

We thank A. Almeida and R. D. Finn for sharing de novo assemblies of the 1,000 gut microbiome samples that we used for benchmarking VAMB. We thank C. Titus Brown for his source code contribution to the VAMB software package. J.N.N., J.J., R.L.A., L.J.J. and S.R. were supported by the Novo Nordisk Foundation (grant NNF14CC0001). S.R. was supported by the Jorck Foundation Research Award.

Author contributions

S.R. conceived the study and guided the analysis. J.N.N., S.R., J.J. and R.L.A. performed the analyses. J.N.N. wrote the software. C.K.S., J.J.A.A., C.H.G., T.N.P., L.J.J., H.B.N. and O.W. provided guidance and input for the analysis. J.N.N., L.J.J. and S.R. wrote the manuscript with contributions from all coauthors. All authors read and approved the final version of the manuscript.

Competing interests

H.B.N. is employed at Clinical-Microbiomics A/S. The remaining authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-020-00777-4.

Correspondence and requests for materials should be addressed to S.R.

Reprints and permissions information is available at www.nature.com/reprints.

8.2 Paper II: Genome binning of viral entities from bulk metagenomics data

Johansen, J; Plichta R.D.; Nissen, J.; Jespersen L. M; Shah A., S.; Deng, L.; Stokholm, J.; Bisgaard, H., Nielsen S.; D., Sørensen J; S., Rasmussen, S.† (2022). *Genome binning of viral entities from bulk metagenomics data*. Nature Communications, Article 965.

VAMB can be accessed at github via the following the link: `https://github.com/RasmussenLab/phamb`

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-28581-5 OPEN

Genome binning of viral entities from bulk metagenomics data

Joachim Johansen[1,2], Damian R. Plichta[2], Jakob Nybo Nissen[1,3], Marie Louise Jespersen[1,4], Shiraz A. Shah[5], Ling Deng[6], Jakob Stokholm[5,6], Hans Bisgaard[5], Dennis Sandris Nielsen[6], Søren J. Sørensen[7] & Simon Rasmussen[1✉]

Despite the accelerating number of uncultivated virus sequences discovered in metagenomics and their apparent importance for health and disease, the human gut virome and its interactions with bacteria in the gastrointestinal tract are not well understood. This is partly due to a paucity of whole-virome datasets and limitations in current approaches for identifying viral sequences in metagenomics data. Here, combining a deep-learning based metagenomics binning algorithm with paired metagenome and metavirome datasets, we develop Phages from Metagenomics Binning (PHAMB), an approach that allows the binning of thousands of viral genomes directly from bulk metagenomics data, while simultaneously enabling clustering of viral genomes into accurate taxonomic viral populations. When applied on the Human Microbiome Project 2 (HMP2) dataset, PHAMB recovered 6,077 high-quality genomes from 1,024 viral populations, and identified viral-microbial host interactions. PHAMB can be advantageously applied to existing and future metagenomes to illuminate viral ecological dynamics with other microbiome constituents.

[1] Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [2] Infectious Disease and Microbiome Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3] Statens Serum Institut, Viral & Microbial Special diagnostics, Copenhagen, Denmark. [4] National Food Institute, Technical University of Denmark, Kongens Lyngby, Denmark. [5] Copenhagen Prospective Studies on Asthma in Childhood (COPSAC), Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen, Denmark. [6] Section of Food Microbiology and Fermentation, Department of Food Science, Faculty of Science, University of Copenhagen, Copenhagen, Denmark. [7] Section of Microbiology, Department of Biology, University of Copenhagen, Copenhagen, Denmark. ✉email: simon.rasmussen@cpr.ku.dk

The human gut microbiota is tightly connected to human health through its massive biological ecosystem of bacteria, fungi, and viruses. This ecosystem has been profoundly investigated for discoveries that can lead to diagnostics and treatments of gastrointestinal diseases such as inflammatory bowel disease (IBD) and colon cancer as well as type 2 diabetes (T2D)[1–3]. In IBD, multiple studies have compiled a list of keystone bacterial species undergoing microbial shifts between inflamed and non-inflamed tissue sites[4,5] and there are strong indications that the gut virome plays a role in disease aetiology[6–8]. Now, the influence of bacteria-infecting viruses, known as bacteriophages, are increasingly studied and their role in controlling bacterial community dynamics in the context of gastrointestinal pathologies is slowly being unravelled[9]. Several studies have presented evidence of temperate *Caudovirales* viruses increasing in Crohn's disease (CD) and ulcerative colitis (UC) patients[6,8,10,11]. However, it has been left unanswered if this phage expansion was due to alterations in host-bacterial abundance, thus viral-host dynamics remains another unexplored facet of the gut virome in diseases such as IBD[12].

Today, the virome is studied through metagenomics where high-throughput sequencing is computationally processed to construct genomes of uncultivated viruses de novo. Viral assembly is a notoriously difficult computational task and is known to produce fragmented assemblies and chimeric contigs[13] especially for rare viruses with low and uneven sequence coverage[14,15]. For better viral assemblies, metaviromes are prepared with extra size-filtration to increase the concentration of viral particles[16,17]. However, identification of viruses without enrichment from bulk metagenomics, is increasingly utilised and overcomes the size-filtration step biases while enabling identification of primarily temperate but also lytic viruses[18]. Currently, several approaches for identifying viral sequences in metagenomics data exist and have helped in supersizing viral databases of uncultivated viral genomes (UViGs) over the last few years[19–21]. These tools are often based on sequence similarity[22], sequence composition[23–28], and identification of viral proteins or the lack of cellular ones[27,28]. A common denominator for these tools is their per-contig/sequence virus evaluation approach that is not optimal for addressing fragmented multi-contig virus assemblies.

Therefore, we developed a framework (PHAMB) based on contig binning to discover viral genome bins directly from bulk metagenomics data (MGX). For this, we utilised a recently developed deep-learning algorithm for metagenomic binning (VAMB)[29] that is based on binning the entire dataset of assembled contigs. Altogether, we reconstructed 2676 viral populations from bulk metagenomes corresponding up to 36% of the paired metavirome dataset (MVX), based on two independent datasets with paired MGX and MVX. A key development in our method is a classifier that can classify non-phage bins from any dataset with very high accuracy (93–99%) compared to existing virus prediction tools such as DeepVirFinder (69–74%)[25], Virsorter2 (30–84%)[30] and viralVerify (86–98%)[31]. Our approach enables identification and reconstruction of viral genomes directly from metagenomics data at an unprecedented scale with up to 6077 viral populations with at least one High-Quality (HQ) genome by MIUViG standards[18] in a single dataset. In addition, we show an increase of up to 210% of HQ viral genomes extracted by combining contigs into viral bins. Using this method to extract viruses from the microbial metagenomes of the HMP2 cohort we were able to delineate both viral and bacterial community structures. This allowed us to investigate viral population dynamics in tandem with predicted microbial hosts for instance identifying 123 and 230 viral populations infecting *Faecalibacterium* and *Bacteroides* genomes, respectively.

Results

A framework to bin and assemble viral populations from metagenomics data. To generate the metagenomics bins we used VAMB that has the advantage of both binning microbial genomes, and grouping bins across samples into subspecies or conspecific clusters. This has proven useful for the investigation of bacterial and archaeal microbiomes, but the approach has even more potential within viromics as viruses are much less conserved, more diverse, and harder to identify without universal genetic markers such as those found in bacterial organisms[32]. Clusters of conspecific viral genomes would enable straightforward identification and tracking of populations across a cohort of samples (Fig. 1a). To develop our framework we used two Illumina shotgun sequencing-based datasets with paired metagenome and metavirome available. The Copenhagen Prospective Studies on Asthma in Childhood 2010 (COPSAC) dataset consisted of 662 paired samples (refs. [33,34]) and the Diabimmune dataset contained 112 paired samples[35]. Each of the two datasets included a list of curated viral species, 10,021 and 328 respectively, that we used here as our gold standard for training and testing our tool. Compared to COPSAC, Diabimmune metaviromes had low viral enrichment (Supplementary Fig. 1), we, therefore, used the average amino acid identity (AAI) model of CheckV[28] to stratify the genomes of the metaviromes into quality tiers ranging from Complete, High-Quality (HQ), Medium-Quality (MQ), Low-Quality (LQ) and Non Determined (ND) to establish a comparable viral truth.

Viral binning is more powerful compared to single-contig approaches. The output of binning metagenomic samples can be hundreds of thousands of bins and we therefore first developed a Random Forest (RF) model to distinguish viral-like from bacterial-like genome bins. The RF model takes advantage of the cluster information from binning and aggregates information across sample-specific bins to form subspecies clusters. Here, we found that the RF model was able to separate bacterial and viral clusters very effectively with an Area Under the Curve (AUC) of 0.99 and a Matthews Correlation Coefficient (MCC) of 0.91 on the validation set (Fig. 1b and Supplementary Table 1). Compared to single-contig-evaluation methods, the RF model was superior as other methods achieved an AUC of up to 0.86 and MCC up to 0.16. This difference in performance is likely explained by the RF model evaluating on bin-level where one sequence with a low viral score does not lead to a misprediction of the whole bin. For instance, we achieved an increase of 200 (190%) and 771 (95%) HQ bins recovered for the Diabimmune and COPSAC datasets compared to using single-contig-evaluation according to CheckV (Fig. 1c, d). Based on the single-contig CheckV evaluations, we found that 97.7 and 95.3% of HQ contigs were binned into HQ bins in COPSAC and Diabmmune, respectively. This means that a small percentage of the HQ contigs, up to 2.3 and 4.7%, are lost in the binning process at the expense of a net increase in genome recovery but can be recovered by parallel single-contig evaluations. Finally, we observed a significantly greater number of viral hallmark genes per virus when using viral bins in both datasets (T-test, two-sided, $t = 16.85$, $P < 0.0005$), while the length and viral fraction were largely comparable (Supplementary Fig. 2).

High viral binning performance on simulated viromes. We then investigated the viral binning performance of VAMB and the prediction performance with simulated datasets including two pure viral and one mixed dataset containing bacteria, plasmids and viruses. The two pure viral datasets comprised 80 crAss-like viruses and 50 small-genome (<6000 bp) randomly sampled from the MGV database[20]. To establish the mixed dataset, the crAss-

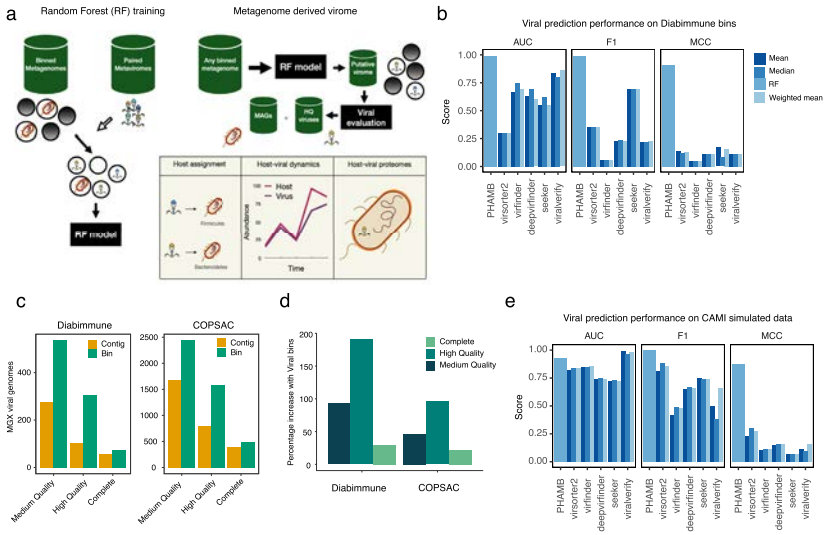


Fig. 1 A framework to bin and assemble viral populations from metagenomics data. a Illustration of workflow to explore viruses from binned metagenomes. First, the RF model was trained on binned metagenomes; bacterial bins were identified using reference database tools and viruses were identified using assembled viruses from paired metaviromes. Viral and bacterial labelled bins were used as input for training and evaluating the RF model. Bins from any metagenome such as human gut, soil or marine can be parsed through the RF model to extract a space of putative viral bins that are further validated for HQ viruses using dedicated tools like CheckV. Binned MAGs and viruses can then be associated in a host assignment step. Host-viral dynamics can be explored in longitudinal datasets to establish temperate phages and the contribution of viruses to Host pangenomes **b** AUC, F1-score and Matthews correlation were calculated for prediction results on viral bins from Diabimmune. These performance scores were calculated based on probability scores from the trained RF model and summarised viral bin-scores of various viral prediction tools. For all tools except the RF model, genomes were labelled viral if the summarised viral score across all contigs, calculated either as a mean, median or contig-length weighted mean passed a threshold. The following thresholds used were 7, 0.5, 0.9, 0.9, 0.9 for viralVerify, Seeker, Virsorter2, Virfinder and DeepVirfinder, respectively. **c** The number of viral genomes recovered from bulk metagenomes, counted at three different levels of completeness in Diabimmune or COPSAC cohorts, evaluated as either single-contigs or viral bins from bulk metagenomes. Evaluation of genome completeness was determined using CheckV here shown for MQ ≥ 50%, HQ ≥ 90%, Complete = Closed genomes based on direct terminal repeats (DTR) or inverted terminal repeats. **d** The percentage-increase of viral genomes found in Diabimmune or COPSAC cohorts using our approach relative to single-contig evaluation. The increase is coloured at three different levels of completeness determined using CheckV, corresponding to the ones used in (**c**). **e** Similar to (**b**) prediction performance scores were calculated for the trained RF model and various viral predictors but on prediction results of CAMI simulated viral genomes from the mixed genome set including bacteria, viruses and plasmids. MAGs metagenome-assembled genomes, HQ high-quality, MQ medium-quality and AUC area under curve.

like and small-genome datasets were combined with an additional 150 random virus genomes, 8 bacterial genome isolates and 20 plasmids (see methods). On the mixed dataset, VAMB outperformed MetaBAT2 on bins with high >0.9 recall and >0.9 precision with a total of 144 vs 134 bins, corresponding to just above 50% (144/280) of all simulated virus genomes (Supplementary Fig. 3a). Furthermore, we found that VAMB binned increasingly a higher number of bins at lower recall (>0.5) and increasing precision levels. Regarding plasmids, both tools were comparable and binned up to 10/20 plasmids with >0.5 recall and >0.95 precision (Supplementary Fig. 3b). Next, we addressed how binning performance could be influenced by virus genome size and highly-similar viruses. For this we sampled smaller virus genomes (<6000 bp, $n = 50$) and viruses of the same family (crAss-like, $n = 80$). A total of 48/50 and 70/80 genomes were binned with >0.99 recall and >0.99 precision for the small-virus and same family-virus set, respectively (Supplementary Fig. 4ab). The ease of binning small viruses was confirmed in the mixed dataset where VAMB captured the majority of small viruses with high recall and precision (F1 > 0.9) (Supplementary Fig. 4c), indicating that genome size was less confounding to binning performance. Finally, to further validate the RF model, we compared the performance in predicting if a bin was viral or bacterial to single-contig viral predictors (Fig. 1e). Using the mixed simulated dataset the single-contig methods displayed much lower discriminatory performance compared to the RF model. For instance, multiple single-contig viral predictors with a high AUC (up to 0.98) displayed low MCC scores meaning that the prediction was not very accurate at the given threshold (Fig. 1e and Supplementary Figs. 5, 6). We then tried to optimise the decision threshold for each of the single-contig viral predictors (Supplementary Figs. 5, 6) which improved the MCC slightly. For instance, viralVerify achieved an AUC of 0.98 on the simulated data, showing that it was effective in separating bacterial and viral

genomes, however with an overlap in the bacterial and viral score distributions. Therefore, even with an optimised threshold, viralVerify displayed an MCC of 0.39. In contrast, the RF model displayed both high AUC (0.93) and MCC (0.87). Thus, we found the RF model, followed by viralVerify, to be the best-suited method on bin-level in mixed-organism assembly datasets. While the RF model predicts plasmids incorrectly as viral, we found that the downstream use of CheckV helped in making a final confident evaluation as plasmid bins contain multiple bacterial-origin genes and are typically classified as 'NA' or picked up by the less precise HMM-model (Supplementary Fig. 7).

Binning the metagenome identifies viral genomes not identified from the metavirome. When applying our method of binning with VAMB and the RF model we obtained 4,480 and 916 viral bins with an MQ or HQ representative bin across the COPSAC and Diabimmune datasets, respectively. We then considered all VAMB clusters as 'viral populations' and thus obtained 2428 and 534 viral populations with at least 1 MQ or better viral bin. After comparing the viral populations obtained from the metagenomics datasets to the respective metaviromes we recovered 17–36% of HQ viruses (corresponding to 527 and 2676 metaviromic viral populations) established in the metaviromes on species (ANI > 95) level and 9–28% on strain (ANI > 97) level (Fig. 2a). The fraction of viruses in the metavirome recovered in the metagenome was considerably higher than more recent estimates[36], which estimated 8.5–10%. This was interesting since the deeply sequenced metavirome may capture multiple low abundant viruses typically not found in metagenomes. Additionally, we found that 46–69% of the HQ metagenome viral populations, corresponding to 124 in Diabimmune and 839 viral populations in COPSAC, were not found in the metavirome, suggesting that a significant part of the virome may be lost during viral enrichment or not represented in induced forms as they are integrated prophages (Fig. 2b). However, we also found that 65–83% of the HQ viral populations in the metavirome were not found in the metagenome data (total 197 in Diabimmune and 2589 in COPSAC) suggesting the reverse to be true as well. For a subset of the viruses found in the COPSAC bulk and metavirome, we estimated higher mean completeness with viral bins (*T*-test, two-sided, $T = 34.02$, CI = 24.4;27.4, $P = 2.2e\text{-}16$) (Fig. 2c). Altogether we found that a great proportion of the gut viral populations can be reconstructed from the metagenomics data and retrieved with even higher completeness compared to the metavirome counterparts.

Viral bins have low contamination. Lastly, we wanted to investigate the occurrence of technically 'misbinned' and contaminating contigs that could inflate viral genome size and influence evaluation and downstream analyses. Based on the viral bins ($n = 1705$) that were highly similar to metavirome viruses in the COPSAC dataset (see Methods), we found in 91.4% of all cases, each bin contained no unrelated contigs (Fig. 2d). Considering only multi-contig bins ($n = 570$) we calculated an average bin-purity of 97.4% in base pairs (median 100%), meaning that on average 2.55% of the genome was not aligning to the corresponding MVX virus. This indicates contamination or, alternatively, a more complete virus in the bulk metagenomic dataset. We further investigated the extent of contamination based on simulated data where 87.6% of the viral bins had a precision of 1 (Supplementary Fig. 8a). For multi-contig bins, we calculated an average bin-purity of 94.5% (median 100%) supporting the results on real data that the majority of bins have low contamination. In summary, our combined binning and machine learning approach improves identification and recovery of viral genomes from metagenomics data and outlines the possibility of binning both fragmented and complete viruses directly from human gut microbiome samples with low degrees of contamination.

Reconstructing the virome of the HMP2 IBD gut metagenomics cohort. We then applied our method to the HMP2 IBD cohort consisting of 27 healthy controls, 65 CD, and 38 UC patients[37]. These samples were gathered in a longitudinal approach and consisted of between 1–26 samples per patient. Importantly, no characterised metaviromics data is available from this cohort and using our approach we were able to identify bacterial and viral populations in the cohort and explore their dynamics in IBD using only metagenomics data. From the cohort, we recovered 577 Complete, 6077 HQ, 9704 MQ (Fig. 3a) and 122,107 LQ viral bins corresponding to 263 Complete, 1024 HQ, 2238 MQ and 44,017 LQ viral populations. We also observed an increase in genome completeness for larger viruses/jumbo viruses with a genome size >200 kbp[38] compared to a single-contig evaluation (Supplementary Fig. 9). Across all the datasets we observed 54 binned putative jumbo viruses (Supplementary Data 1). In addition, we observed that similar viral length distributions for viruses recovered as a single-contig and as viral bins, both correlated with CheckV quality tiers (Fig. 3b).

Viral population taxonomy is highly consistent. We then investigated the taxonomic consistency of our viral populations and found this to be very high as the median intra-cluster Average Nucleotide Identity (ANI) for MQ to Complete viral clusters was 97.3–99.3% (Supplementary Fig. 11). Even in clusters with over 100 sample-specific viral bins the intra-cluster median ANI was consistently high (median = 97.1–98.5%) (Fig. 3c). Inter-cluster ANI was much lower in the 91.7–92.8% range closer to the genus level. Therefore, our approach was able to identify and cluster near strain-level viral genomes across samples. For example, in the HMP2 dataset, we identified 50 different viral populations for a total of 916 MQ or better crAss-like viral bins. Here, viral population *653* corresponded to the prototypic crassphage[39] and accounted for 253 of the 916 crAss-like genomes discovered in the HMP2 dataset. We then used all of these 916 bins to generate a phylogenetic tree based on the large terminase subunit (TerL) and found the highly consistent placement of the viral genomes according to their binned viral population (Fig. 3d and Supplementary Fig. 12). Viral population 653 formed one monophyletic clade except for one bin while all the other crAss-like clusters were monophyletic. The division of the crAss-like genomes into the binned clusters therefore likely represents actual viral diversity. Taken together, this shows that our reference-free binning produces taxonomically accurate viral clusters, thus aggregating highly similar viral genomes across samples.

The metagenomic virome is personal and highly stable in healthy subjects. Several metavirome studies have reported the presence of stable, prevalent and abundant viruses in the human gut[7,40]. We found that the gut virome in the HMP2 cohort[37] was highly personal and stable over time in nonIBD subjects, which was reflected by the lower Bray–Curtis dissimilarity between samples from nonIBD subjects compared to UC (*T*-test, two-sided $P = 0.017$, $t = -2.47$, CI = −0.01;−0.13) and CD subjects (*T*-test, two-sided, $P = 0.023$, $t = -2.3$, CI = −0.12;−0.01) (Fig. 4a, b). In addition, the dysbiotic samples, as defined by Price et al. (2019)[37], could be clearly separated with a principal component analysis (PCoA), where the virome explained 4.2 and 3.4% of the variation (Fig. 4c). This was confirmed with a PERMANOVA test on viral ($P < 10 - 3$, $R^2 = 1.6\%$, $F = 9.51$, permutations = 999) and bacterial abundance profiles ($P < 10 - 3$, $R^2 = 3.0\%$, $F = 11.97$) and shows dysbiosis affecting both the virome and bacteriome. Alpha-diversity metrics supported this as

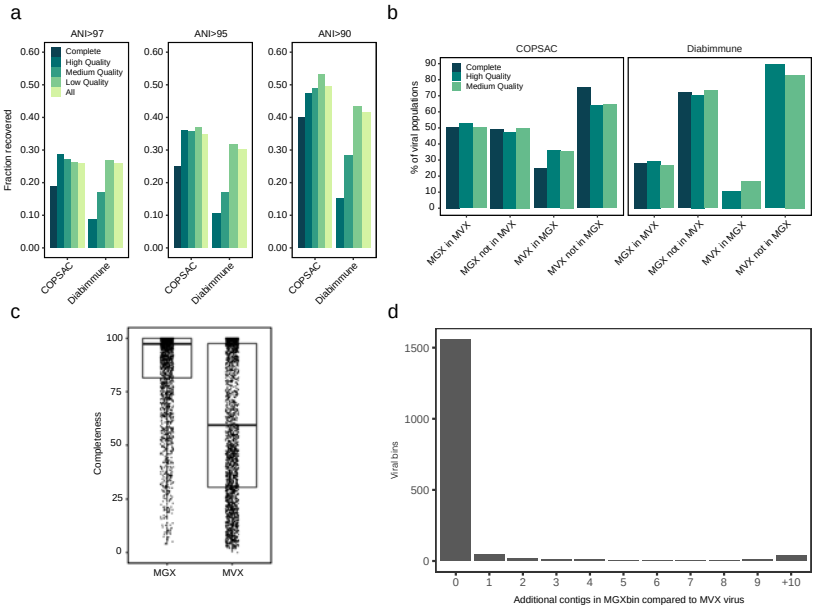


Fig. 2 Binning the metagenome identifies viral genomes not identified from the metavirome. a The fraction of metavirome viruses in COPSAC and Diabimmune coloured at different levels of completeness or all together determined with CheckV, identified in VAMB bins from bulk metagenomics of the same cohorts. We defined a metavirome virus to be recovered if the aligned fraction was at least 75% and ANI was >90, >95 or >97.5 to a VAMB bin based on FastANI. **b** The percentage of viral populations, at different levels of completeness determined with CheckV, identified in both metaviromes (MVX) and bulk metagenomics (MGX) or unique to either dataset. Shared populations are identified with a minimum sequence coverage of 75% and ANI above 95%. (1) MGX in MVX: % of Viral populations found in MGX also found in MVX. (2) MGX not in MVX: % of Viral populations unique to MGX i.e. not found in MVX. (3) MVX in MGX: % of Viral populations found in MVX are also found in MGX. (4) MVX not in MGX: % of Viral populations unique to MVX i.e. not found in MGX. **c** Viral genome completeness estimated for $n = 2646$ viruses found both in metaviromes and bulk metagenomics sharing the same nearest reference in the CheckV database. **d** The number of contigs in viral bins from bulk metagenomics that do not align to the closest viral reference in the metavirome. In the majority of viral bins, all contigs align to the nearest reference. ANI average nucleotide identity.

Shannon-Diversity (SD) was higher in nonIBD subjects compared to both UC and CD (*T*-test, two-sided, $P = 0.000155$, $t = -3.79$ and $P = 7.9e\text{-}09$, $t = -5.81$) while dysbiosis affected every patient group resulting in a significantly reduced SD. In accordance, viral richness was lower in UC (two-sided *T*-test, $P = 1.44e\text{-}15$, $t = -8.09$, CI = −12.40;−19.80) and CD (two-sided *T*-test, $P = <2e\text{-}16$, $t = -9.39$, CI = −12.91;−19.50) patients and further exaggerated in dysbiotic samples (Fig. 4d, e). These viral alpha-diversity trends were also observed in the bacteriome, suggesting that the viruses follow the expansion or depletion of their bacterial host during dysbiosis (Supplementary Fig. 14). Indeed, we identified 250 likely temperate viruses out of 348 differentially abundant viruses that expanded with increasing dysbiosis (linear-mixed-effect model, adj. $P < 0.005$, FDR-corrected). This observation acknowledges earlier results showing an increase in temperate viruses in UC and CD[6,10]. Further analysis on the longitudinal abundance profiles of virus and predicted bacterial host reaffirmed the synchronised expansion theory (Supplementary Fig. 15).

Viral–host interactions can be explored from viral populations and MAGs. A unique feature of performing the analysis on metagenomics data is that both the bacterial and viral populations are binned simultaneously. Therefore, we were able to estimate the abundance of both the viral and bacterial compartments of the microbiome and explore the viral host range in silico using the MAGs. In total from the HMP2 dataset, we obtained 3130 and 3819 Near-Complete (NC) and Medium-Quality (MQ) MAGs[41]. Based on MAG-derived CRISPR spacers we found spacer hits to 464 (45.3%) to viral populations with at least one HQ representative. To further expand our viral-host prediction we conducted an all-vs-all alignment search between the MAGs and viral populations for prophage signatures. Then by combining the CRISPR spacer and prophage search we connected 93.6, 74.4, 82.5 and 65.0% of MAGs from *Bacteroidetes, Firmicutes, Actinobacteria, and Proteobacteria phylum*, respectively, with at least one virus (Supplementary Fig. 16). We estimated host-prediction purities to be 94.5 and 75.6% on species rank for the CRISPR spacer and prophage signature (Supplementary

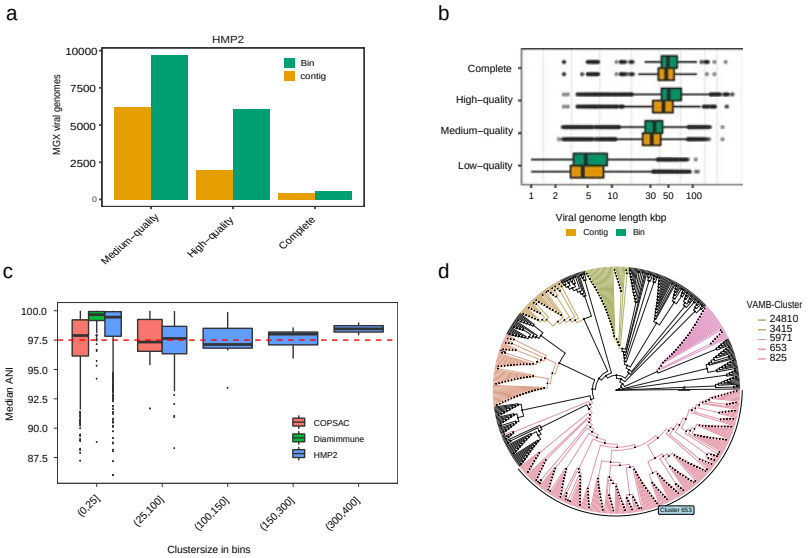


Fig. 3 Reconstructing the virome of a human gut metagenomics cohort. a The number of viral genomes with three different levels of completeness in HMP2, evaluated as either single-contigs or viral bins from bulk metagenomes. Evaluation of genome completeness was determined using CheckV here shown for medium-quality ≥50% (MQ), high-quality ≥90% (HQ), Complete = closed genomes based on direct terminal repeats or inverted terminal repeats. **b** The sequence length distribution in kbp of viral genomes at four different levels of completeness in HMP2, evaluated as either single-contigs ($n = 215{,}009$) or viral bins ($n = 138{,}367$) from bulk metagenomes. Shown for low-quality (LQ) <50%, MQ, HQ and Complete. **c** Median ANI based on pairwise ANI genome measurements between bins within the same VAMB cluster. Median ANI is consistently above 97.5 in small VAMB clusters with 0–25 bins and in larger VAMB clusters with 300–400 bins. **d** Cladogram of an unrooted phylogenetic tree with crAss-like bins based on the large terminase subunit protein (TerL). Five different VAMB clusters have been coloured and illustrate high monophyletic relationships. The phylogenetic tree was constructed using IQtree using the substitution model VT + F + G4. ANI average nucleotide identity %, DTR direct terminal repeats, ITR inverted terminal repeats, Kbp kilobase pairs.

Fig. 17B). Therefore, we confirmed that most gut phages have a primarily narrow host range[42]. MAGs belonging to the genera *Faecalibacterium* and *Bacteroides* seemed to be viral hotspots since 99.7 to 98.7% could be associated with a HQ viral bin, corresponding to 123 and 230 distinct viral populations, respectively (Fig. 5a). For instance, in abundant commensals like *Bacteroides vulgatus* (cluster 216) we observed consistent prophage signals over time for multiple viruses across several samples (Fig. 5b). Interestingly, because the host range of crAss phages are not well understood we investigated CRISPR spacer hits to the MAGs in our databases. Even though we could host-annotate an overall of 45.3% of all HQ viral populations to a MAG, only 74 of the 916 crAss-like bins could be associated with any of the 3306 *Bacteroidetes* bins in our dataset using CRISPR spacers. This was despite having assembled CRISPR arrays (with confidently predicted subtypes) for 998/3306 (~30%) of the *Bacteroidetes* bins. When we performed a similar search to a comprehensive CRISPR spacer database[43] of 580,383 bacterial genomes we could annotate 512 of the 916 crAss-like bins to Bacteroidetes bacteria. These findings suggest that crAss-like phages are not frequently targeted by CRISPR spacers extracted from *Bacteroidetes* CRISPR-Cas systems within the same environment.

The binned viral populations are enriched in proteins found in temperate phages. Another topic of interest was viral-host complementarity, in particular, what functions bacteriophages could provide to the host and how the viral proteome differs with respect to host taxonomy. Using our map of viral-host connections and through characterisation of viral protein sequences, we ranked protein annotations stratified by their predicted host genera. Overall, the proteins were highly enriched for annotations related to viral structural proteins such as baseplate, portal, capsid, head, tail/tail-fibre and tail tape measure but also viral integrase enzymes and Lambda-repressor proteins (Supplementary Data 2). For instance, Lambda-repressor proteins were found in up to ~60% of all viruses suggesting that our dataset was enriched with temperate phages (Fig. 6a). Interestingly, we also identified virally encoded protein domains, which are known to function as viral entry receptors[44], to be enriched within a group of viral populations infecting *Bacteroides* and *Alistipes* such as the TonB plug and TonB-dependent receptor domains (PF07715 and PF00593, Fisher's exact test, adj. $P < 0.05$, FDR-corrected) (Supplementary Data 3). Furthermore, the TonB domains also encode an established immunodominant epitope[45] suggesting that viral populations carry immunogenic entry receptors when expressed

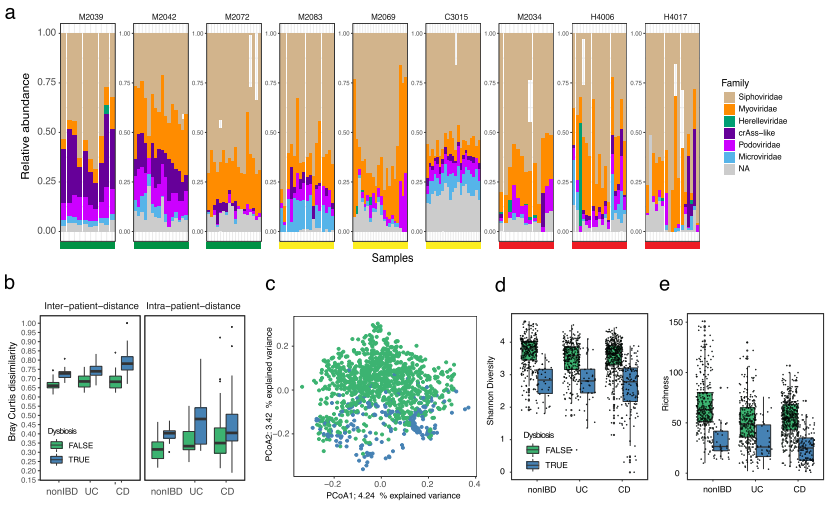


Fig. 4 The metagenomics estimated virome is personal and highly stable in healthy controls. a Longitudinal virome compositions for three nonIBD (green bar), three UC (yellow bar) and three CD (red bar) diagnosed subjects. Each panel represents a subject where the virome composition was organised according to the total relative abundance according to the taxonomic viral family, where 'NA' populations are coloured grey. **b** Dissimilarity boxplots based on Bray–Curtis distance (BC) function between samples from different subjects (first panel inter-patient-distance) and between samples from the same subject (second panel intra-patient-distance). The BC distances are shown for samples from nonIBD ($n = 326$), UC ($n = 323$) and CD ($n = 573$) diagnosed subjects. Furthermore, BC distances are coloured according to dysbiosis (blue, UC = 39 samples, CD = 133 samples, nonIBD = 38 samples) or not (green, UC = 284 samples, CD = 425 samples, nonIBD = 286 samples). **c** Principal component analysis (PCoA) of Bray–Curtis distance matrix calculated from the viral abundance matrix in HMP2. Each point is coloured according to diagnosed dysbiosis as in (**b**). **d** Shannon-diversity estimates of metagenomics derived viral populations and coloured according to dysbiosis as in (**b**). **e** Per sample viral population richness based on the number of viral populations detected (abundance >0) in the samples. Coloured according to dysbiosis as in (**b**). nonIBD: healthy control, UC ulcerative colitis, CD Crohn's disease.

by their host. Finally, Reverse Transcriptase (RT, PF00078) proteins were also highly detected, in agreement with recent results[20] and shared by all viral populations irrespective of the predicted host (Supplementary Fig. 18A). These proteins are known modules in bacteriophage diversity generating regions that cause hypervariability in specific viral genes[46].

Exploring the dark-matter metavirome. Finally, we investigated the part of the RF predicted bins that did not resemble any of the known genomes, i.e. metagenomics 'dark-matter'. These were defined as populations without at least one HQ or MQ viral bin. Such populations, therefore, represent a part of the microbiome that are not classified as bacterial, archaeal and not alike known viral genomes. Since dark-matter populations were numerous (97.6% of all RF predicted VAMB clusters) we suspected many of these to be fragmented viruses or unknown viruses. Dark-matter populations larger than 10 kbp with at least one viral hallmark gene displayed lower viral prediction scores compared to HQ-MQ viral bins, while bins targeted by CRISPR spacers displayed a significantly higher prediction score (*T*-test, two-sided, CI = 0.05:0.067, $P = 2.2e\text{-}16$), thus we annotated these as 'viral-like' (Fig. 6b and Supplementary Fig. 19). When stratifying read abundance on these groups (HQ-MQ, viral-like, dark-matter) we found them to explain on average 2.77, 2.04 and 17.7% of total read abundance across samples, respectively (Fig. 6c). Furthermore, we found that 5% HQ and 3.7% viral-like populations were detected in at least 40% of the patients across disease states. For instance, HQ viral populations cluster 653 were observed in 41% of the cohort (Fig. 6d). Simultaneously, a viral-like population of 1338 was observed in 98% of individuals but displayed a low similarity to any reference genome (Fig. 6e). However, caution should be taken with labelling dark-matter bins as viruses since these are possibly incomplete, contaminated or contain other types of mobile genetic elements that encode proteins shared with viruses such as integrases, polymerases and toxin-antitoxin addiction modules[47,48].

Discussion

Because of the current challenges facing the viral assembly process, which results in partial and fragmented viral genome recovery[13,15], viral communities have traditionally been notoriously difficult to study. Metavirome datasets have been crucial for identifying a broad scope of viruses, in particular virulent ones. However, the paucity and difficulties in creating metavirome datasets combined with the fact that bulk metagenomes are

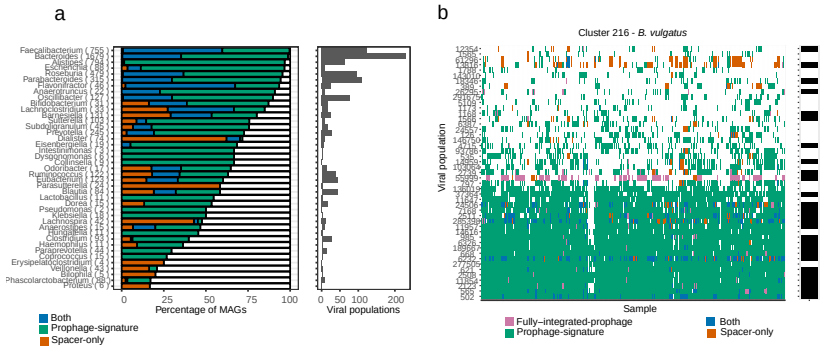


Fig. 5 Viral–host interactions can be explored from viral populations and MAGs. a Bacterial MAGs and viral relations. Each MAG was connected to the viral bins using either sequence alignment of the virus to MAG (green), CRISPR spacer alignment (orange) or both (blue). The right panel shows the percentage of MAGs, grouped by genera, that was annotated with the virus via alignment or CRISPR spacer. The number of distinct viral populations associated with a MAG genus based on either of the following: sequence alignment of the virus to a MAG within the given genera, CRISPR spacer alignment or both. **b** Viral association to all MAGs of VAMB cluster 216 (*B. vulgatus*) in the HMP2 dataset. For instance, viral population 502 was associated with the *B. vulgatus* across the vast majority of samples where *B. vulgatus* was present.

produced in abundance, calls for more methods to efficiently extract the viromes found therein. Here we present an improved framework for exploring metavirome directly from bulk metagenomics datasets.

Using our map of viral and bacterial connections we wanted to associate and study the human gut virome along highly abundant gut bacteria such as *Bacteroides* and *Faecalibacterium*. Several of these genera represent not only highly abundant gut commensals but also hotspots for viruses as we have shown by connecting 230 and 123 viral populations to *Bacteroides* and *Faecalibacterium*, respectively. Viral hotspots could be partially explained by factors such as their absolute numbers and genome sequencing depth, which may allow for a more complete assembly of CRISPR-cas systems. A large part of these connections was also made via prophage signatures, i.e. shared genomic elements between bacteria and phage (Fig. 5). Prophage signatures could be the result of increased rates of lysogeny and coinfection as higher microbial densities and phage adsorption rates provide favourable conditions for multiple phages to 'piggyback' highly productive hosts and exchange genetic material[49]. In agreement with other results[11], we found that *F. prausnitzii* genomes are rich in prophages and were able to annotate one for 99.7% of the bacterial bins in HMP2. In the HMP2 cohort, we identified 250 likely temperate *Caudovirales* viruses expanding in a synchronised manner with bacterial hosts following increasing gut dysbiosis[6,10]. However, more work is needed to outline the intricate virus-host dynamics that can explain the degree of viral influence on bacterial perturbations observed in IBD related to dysbiosis such as 'Piggyback-the-Winner' or 'Kill-the-Winner' dynamics[50] with carefully calculated correlations[51].

Based on the viral proteomes it is clear that a majority of HQ viruses extracted in the bulk metagenomes are likely temperate as we have found integrase proteins in 46% of the viral populations and Lambda-repressor proteins in 60% of viruses infecting *Faecalibacterium* bacteria. This adds to the expectation that the non-enriched viromes can be biased toward viruses that infect the dominant host cells in the sample[18]. Interestingly, we found examples of viruses encoding proteins with immunodominant epitopes such as the TonB plug domain (PF07715) and TonB-dependent beta-barrel (PF00593)[45] in hundreds of viral proteomes extracted from viruses infecting members of *Bacteroidetes* such as *Bacteroides* and *Alistipes*. A recent study has shown that common structural phage proteins such as the tail length tape measure protein (TMP) also harbour immunodominant epitopes that cross-react to cause antitumour immunity[52]. It is therefore interesting to investigate the extent to which viral organisms can influence the human host-microbiota immune balance through horizontal transfer and expression of immunogenic proteins.

Metavirome studies have until now been the primary source for exploring viral diversity in microbiomes. Now, viral populations are increasingly uncovered in bulk metagenomes and we showed that more complete viral genomes can be identified via viral binning across three different cohorts, similar results were found in a recent paper focused on binning of sequenced viral particles[53]. Our approach allowed precise clustering of both viral and bacterial populations in three cohorts that enabled direct investigation into viral-host interactions and discovery of new diversity. We believe that future studies can greatly leverage this approach to conduct virome analyses and investigate the viral influence of the intricate microbiome ecosystem that governs human health.

Methods

Datasets. The Copenhagen Prospective Studies on Asthma in Childhood 2010 (COPSAC) dataset consisted of 662 paired samples obtained at age 1 year from an unselected childhood cohort (refs. [33,34]). The COPSAC study was conducted in accordance with the guiding principles of the Declaration of Helsinki and was approved by the Capital Region of Denmark Local Ethics Committee (H-B-2008-093), and the Danish Data Protection Agency (2015-41-3696). Both parents gave oral and written informed consent before enrolment. The Diabimmune dataset contained 112 paired samples from controls and type 1 diabetes patients. The Human Microbiome Project 2 cohort consisting of 1317 metagenomic samples were downloaded from https://ibdmdb.org/tunnel/public/summary.html.

Processing of metagenomics and metaviromics datasets. Metagenomic samples of infants en route T1D recruited to the Diabimmune study were downloaded from https://pubs.broadinstitute.org/diabimmune (October 2019). Metagenomic samples were quality-controlled and trimmed for adaptors using kneaddata

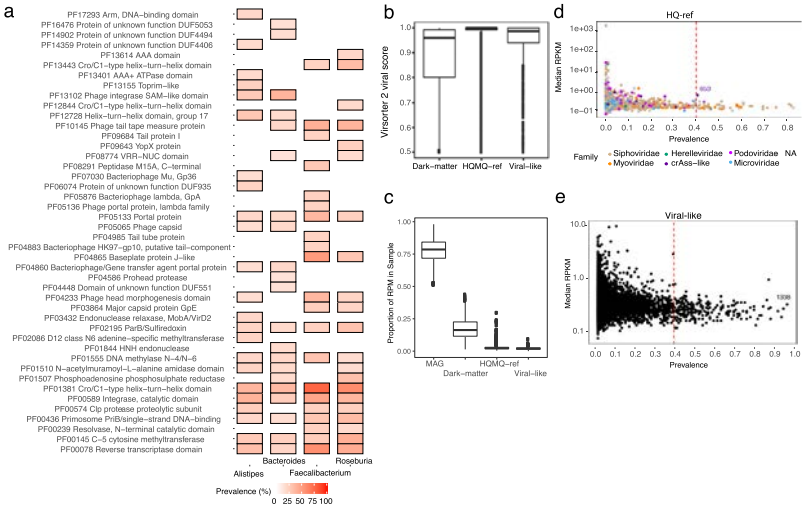


Fig. 6 Viral proteins and the dark-matter metavirome. a The percentage of HQ viruses, associated with four bacterial host genera; Alistipes, Bacteroides, Faecalibacterium and Roseburia, which encode top-20 prevalent PFAM domains. **b** Virsorter2 viral prediction scores for all viral bins with at least one viral hallmark gene. Completeness was estimated using CheckV and the bins were grouped as (1) HQ-MQ-ref when completeness ≥50% or high-quality ≥90% ($n = 45{,}983$ bins), (2) bins with less than 50% completeness were annotated as Dark-matter ($n = 392{,}226$ bins), and (3) dark-matter bins with confident CRISPR spacers against a bacterial host were annotated as Viral-like ($n = 43{,}695$ bins). **c** The distribution of sample RPM of bacterial MAGs, HQ-MQ-ref viral populations, Dark-matter and Viral-like populations as defined in (**b**). The majority of sample reads were mapped to MAGs but on average 17.7% of all reads mapped to Dark-matter bins. **d** The abundance in RPKM of rare and highly prevalent viruses with an HQ genome in HMP2. Each point represents a viral population coloured according to the viral taxonomic family. The progenitor-crAssphage is indicated as cluster 653. **e** As in (**d**) but with viral-like populations like cluster 1338 showing that many are low abundant, but highly prevalent. RPM read per million, RPKM read per kilobase million.

(https://github.com/biobakery/kneaddata) and trimmomatic (v.0.36)[54] settings: ILLUMINACLIP: NexteraPE-PE.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:100. Each metagenomic sample was assembled individually using metaspades (v. 3.9.0)[55] using the parameters '--meta, -k 21,33,55,77,99' and filtered for contigs with minimum length of 2000 base pairs. Mapping of reads to contigs was done using minimap2 (v.2.6)[56] using '-N 50' and filtered with samtools (v.1.9)[57] using '-F 3584'. Contig abundances were calculated using jgi_summarize_bam_contig_depths from MetaBAT2 (v.2.10.2)[58]. Metagenomic bins were defined using VAMB (v. 3.0.1)[29] to cluster the metagenomic contigs into putative MAGs and viruses. Initially, the contents of all bins were searched for viral proteins with hmmsearch (v. 3.2.1)[59] against VOGdb (v. 95) (https://vogdb.csb.univie.ac.at/). The presence of bacterial hallmark genes were determined using both CheckM (v.1.1.2)[60] and hmmsearch against the miComplete bacterial marker HMM database (v.1.1.1)[61]. A viral score of each contig was computed using DeepVirFinder (DVF v.1.0)[25]. We initially assessed the metaviromes of the COPSAC and Diabimmune datasets using ViromeQC[62] and found 5.1 and 0.21 times viral enrichment of the two datasets, respectively (Supplementary Fig. 1).

Training the random forest to predict viral bins. First we established an initial viral truth set in the metagenomic assembly for the random forest classification. For each metagenomics bin, we computed the fraction of contigs mapping to a set of non-redundant viral sequences (Gold standard) using blastn (v. 2.8.1)[63] with a minimum sequence identity of 95% and query coverage of 50%. Gold standard viral contigs of the paired metaviromics datasets were provided by the authors of the Diabimmune and COPSAC studies (https://doi.org/10.5281/zenodo.5821973). Metagenomic bins with ≥95% of contigs matching with the above criteria were annotated as Viral bins. For annotating bacterial bins, MAGs were identified using CheckM (v.1.1.2). MAGs with a completeness score of 10% or above and contamination ≤30% were added to the training and validation set labelled as bacteria. For training, we used COPSAC and validated using the Diabimmune dataset. Thus the model was trained to distinguish confidently labelled bacterial and viral bins produced by VAMB, this provided an RF model highly effective at removing non-viral bins and providing a highly enriched candidate set of viral bins that could be further evaluated using dedicated validation tools. In the RF model we included features such as bin size, the number of distinct bacterial hallmark genes, the number of different PVOGs in a bin divided by the number of contigs in the bin, viral prediction DVF score (median DVF score for a bin) defined by DeepVirFinder. The Random Forest model was implemented in Python using *RandomForestClassifier* (sklearn v. 0.20.1) with 300 estimators and using the square root of the number of features as the number of max features. The model was trained on the COPSAC dataset using 40% of observations for training and 60% for validation. Subsequently ROC/AUC, recall and precision was calculated using the Diabimmune recovered viruses as an evaluation set. We ran viral predictions on contigs of minimum 2,000 bp using Virsorter2 (v. 2.2.3)[30], viralVerify (v.1.1)[31], Seeker (v.1.0)[64], Virfinder (v.1.1)[26] and DeepVirfinder (v. 1.0), all on their default settings. In order to calculate single-contig viral prediction performance, a contig was labelled viral if the prediction score was above 7, 0.5, 0.9, 0.9 and 0.9 viralVerify, Seeker, Virfinder, DeepVirFinder and Virsorter2, respectively. Genome level predictions (bacterial or viral) for each of the aforementioned tools were done with the same cutoffs mentioned above but based on the aggregated bin-score. The bin-scores were aggregated as a contig-length weighted mean, mean and median.

Virus binning and prediction performance on simulated datasets. We compared the viral binning performance of VAMB and MetaBAT2 using the official CAMISIM method to create assemblies and metagenome profiles[65]. To this end we generated three different metagenome compositions with up to 308 reference genomes; one mixed with bacteria, plasmids and viruses to test binning in complex samples i.e. high diversity (1), one with only crass-like viruses to test binning with highly similar viruses i.e. high relatedness (2) and a set of small viruses (<6000 bp) including members of the Microviridae family to address the bias of size (3). Bacterial genomes were pulled from NCBIs refseq genome repository 2021, plasmids from the PLSDB database (v. 2021_06_23)[66] and viral genomes from the recent MGV database[20] (Supplementary Data 4). Fragmented genome assemblies

were generated for each metagenome composition using CAMISIMs (v.1.1.0) metagenome simulation-pipeline with default settings for ten samples[65]. In order to test genome recovery via binning, abundance of the simulated contigs were calculated by mapping of reads to contigs with minimap2 (v.2.6) using '-N 50' and filtered with samtools (v.1.9) using '-F 3584'. Then the abundances were calculated using jgi_summarize_bam_contig_depths from MetaBAT2 and used as input for VAMB and MetaBAT2 that were run with default parameters on the simulated contigs of a minimum of 2000. Furthermore, we ran viral predictions on contigs of minimum 2000 bp using Virsorter2 (v. 2.2.3)[30], viralVerify (v.1.1)[31], Seeker (v.1.0)[64], Virfinder (v.1.1)[26] and DeepVirfinder (v. 1.0), all on their default settings. In order to calculate single-contig viral prediction performance, a contig was labelled viral if the prediction score was above 7, 0.5, 0.9, 0.9 and 0.9 viralVerify, Seeker, Virfinder, DeepVirFinder and Virsorter2, respectively. Genome level predictions (viral or non-viral) for each of the tools were done with the same cutoffs mentioned above on the aggregated bin-score. The bin-scores were aggregated as a contig-length weighted mean, mean and median. The RF model was run as intended where information about each contig was aggregated and parsed by the model to produce a viral/non-viral label. Optimised and overfitted bin/genome-score thresholds were determined by inspection of genome-score distributions (Supplementary Fig. 5) for each viral prediction method. These thresholds were −1.3, 0.75, 0.9, 0.5 and 0.5 for viralVerify, Seeker, Virsorter2, DeepVirFinder and Virfinder, respectively.

Intersection of viruses in MGX and MVX data. In order to identify the number of viruses assembled and binned in the metagenomic (MGX) datasets we searched the metavirome (MVX) viruses in all-vs-all search and calculated genome-to-genome average nucleotide identity (ANI) and genome coverage as an aligned fraction (AF). Here we defined species level above 95% ANI and strain-level above 97% ANI. Overlapping or also described as highly-similar viruses between the paired MGX and MVX datasets were those fulfilling the ANI >95% and >75% AF criteria. This search was conducted using FastANI (v.1.1, '-fragmenlen 500 -minimumfrag 2 -minimum 80% ANI')[67] with genome coverage ≥50% (bidirectional fragments / total fragments). We note that hits with less than 80% ANI were not included. We expected that we might be able to find fragmented/incomplete viruses assembled in the metavirome but were more curious about near-complete viruses, thus we quality controlled all MVX viruses using CheckV (v0.4.0, default settings, database v. 0.6)[28] to achieve a completeness estimate for each. By labelling the quality of each MVX virus we organised the success of genome recovery into the four CheckV levels (low-quality ≤50%, medium-quality ≥50%, high-quality ≥90%, Complete = closed genomes based on direct terminal repeats (DTR) or inverted terminal repeats). Furthermore, we also quality controlled the putative viruses assembled and binned in the MGX to ask the reverse question, i.e. to what extent do we find complete viruses with no similarity to viruses in the MVX.

Completeness of viruses recovered in metavirome and bulk metagenomes. To standardise our viral recovery performance across different datasets, we used the guidelines on Minimum Information about an Uncultivated Virus Genome (MIUViG)[18]. The viral completeness of viruses from metaviromics data was assigned using CheckV described as above. CheckV was used to conduct a benchmark on virus genome completeness by evaluating single-contig assemblies against the use of viral bins (also described as viral MAGs). To this end, we based our analysis solely on AAI-model predictions. As the authors of CheckV note, the method was not designed by default to accommodate viral MAGs and may not deal properly with contaminants from bacterial or viral sources[28]. This became clear as we observed a majority of HMM-model predicted viruses consisting of sequences with close to zero percent viral sequence (Supplementary Fig. 20). We suspect that this was to be expected since the HMM-model is designed for single-contig viral assemblies. Thus, the model could not deal properly with cases where a viral marker gene was identified in a single-contig of the bin and contaminating sequences inflate the total bin size to randomly fit into the reference size range of viruses encoding the same viral marker. Hence to avoid including false-positive viral bins, we defined a viral population as HQ-ref when at least one bin in the VAMB cluster contained an HQ evaluation based on AAI-evaluation. All viral bins with a CheckV computed genome copy number ≥1.25 were removed to control for 'concatemers'. Finally, viral bins with an estimated completeness >120% (over-complete-genomes) were removed as well to control for highly contaminated bins. We found that the frequency of HQ genomes, which according to MIUViG standards[18,19] were 'overcomplete-genomes' (estimated completeness >120%), was between 7.9–14.2% for the viral bins and 3.8–6.1% for single-contig evaluation (Supplementary Table 2). Hence, the binning approach generates more over-complete-genomes, although these can be identified and removed using for instance CheckV, which we highly advise. We found that after removal of over-complete-genomes, VAMB mainly produces viral bins with low contamination and high purity. Contamination and purity in this case was calculated according to a reference/ground truth. Example: for a viral bin with a total size of 90,000 and 8000 bp not aligned to the corresponding ground truth genome, contamination is 8000/90,000 = 8.8% and purity is 100–8.8% = 91.2%. The remaining populations without a single HQ or MQ bin within their VAMB cluster were described as dark-matter. For identifying viruses in 'dark-matter' populations, we ran Virsorter2 (v.2.0)[30] and considered sequences or bins with a prediction score >0.75, at least one viral hallmark and a minimum size of 10 kbp as a putative virus. In this subset of putative viruses, we defined 'viral-like' dark-matter when they were targeted with a CRISPR spacer by a bacterial MAG (see 'Viral-host prediction').

Viral taxonomy and function. While the databases of viral genomes continue to grow, taxonomy is still a challenge for viral genomes with little similarity to the International Committee on Taxonomy of Viruses (ICTV) annotated genomes. Viral proteins were predicted using prodigal (v.2.6.3)[68] using '-meta'. All proteins were annotated using viral protein-specific databases such as VOG (http://vogdb.org) or viral subsets of TrEMBL used in the tool Demovir (v.1.1.0) (https://github.com/feargalr/Demovir). Viral taxonomy was assigned to each bin using the plurality rule described before in Roux et al. (ref. [19]): (1) taxonomy was assigned to genomes with at least two PVOG proteins using a majority vote (≥50% else NA) on each taxonomic rank based on the last common ancestor (LCA) annotation from the PVOG entries. (2) The CheckV VOGclade taxonomy was transferred if available from the best viral genome match in the CheckV database. In order to annotate 'crAss-like' viruses, predicted proteins were aligned using blastp (v. 2.8.1)[63] to the large subunit terminase (TerL) protein and DNA polymerase (accessions: YP_009052554.1 and YP_009052497.1) of the progenitor-crassphage using already described cutoffs[69]. When investigating taxonomic annotations, considering only MQ-Complete viral bins, the most dominant viral family annotated was Siphoviridae accounting for 53.5% of the viral bins (Supplementary Figure 9). Furthermore, we could assign Myoviridae 14.57%, Podoviridae 8.59%, Microviridae 8.30%, crAss-like 3.61%, CRESS 2.52%, Herelleviridae 1.37% and Inoviridae 0.58%. Finally, 6.93% of viruses could not be confidently assigned any viral taxonomy. Similar distributions of taxonomic annotations were also observed for Diabmmune and COPSAC (Supplementary Table 3).

For viral proteomes, we utilised CheckV's contamination detection workflow to extract proteins encoded only in viral regions to avoid host contamination. These viral proteins were analysed with interproscan (v. 5.36-75.0)[70] using the following databases: PFAM, TIGRFAM, GENE3D, SUPERFAMILY and GO-annotation. For each annotated functional domain in viruses predicted to infect a given host genus enriched proteins were identified using Fisher's exact test using the function *phyper* in base R. P-values were adjusted using false discovery rate (FDR) correction[71]. Viral reverse transcriptase enzymes were grouped into DGR-clades by querying each protein sequence against a database of RT DGR clade HMM models while DGR target genes were identified using the methods and pipeline provided[72].

Phylogenetic tree of crAss-like viruses. A phylogenetic tree was constructed for crAss-like viruses identified in the HMP2 dataset based on proteins annotated as the large terminase subunit protein (the TerL gene). First, viral bins annotated as 'crAss-like' were determined as described above. 'crAss-like' proteomes were aligned to a terminase large subunit protein (accession: YP_009052554.1) and also against VOGdb hmmsearch (v. 3.2.1, hmmscore ≥ 30)[59] against VOGdb (v. 95) (https://vogdb.csb.univie.ac.at/). The VOG entries corresponding to the terminase large subunit:VOG00419, VOG00699, VOG00709, VOG00731, VOG00732, VOG01032, VOG01094, VOG01180 and VOG01426, were identified using a bash command on a VOGdb file: 'grep -i terminase vog.annotations.tsv'. An alignment file was produced for proteins, annotated as terminase large subunit, using MAFFT (v. 7.453)[73] and Trimal (v. 1.4.1)[74] and converted into a phylogenetic tree using IQtree (v. 1.6.8 -m VT + F + G4 -nt 14 -bb 1000 -bnni)[75].

Viral-host prediction. Viral genomes were connected to hosts using a combination of CRISPR spacers and sequence similarity between viruses categorised as HQ-ref and MAGs. CRISPR arrays were mined from COPSAC and HMP2 MAGs using CrisprCasTyper (v.1.2.3)[76] with '--prodigal meta' and all spacers were blasted with blastn-short (v. 2.8.1)[63] against all viral genomes to identify protospacers. CRISPR spacer matches with ≥95% sequence identity over 95% of spacer length and maximum of two mismatches were kept. In order to identify the host of viruses, viral bins were aligned to MAGs using FastANI (v.1.1, '--fragLen 5000 --minFrag 1')[67] and blastn megablast (v. 2.8.1)[63] with a minimum ANI ≥90% and sequence identity ≥90, respectively. We followed the approach described by Nayfach et al. (ref. [42]) to calculate host-prediction consensus and accuracy. The viral host was defined using a plurality rule at each taxonomic rank based on the lineage of bacteria connected using either CRISPR spacer or alignment to the given virus. The cutoffs described above were selected after benchmarking the alignment approach with FastANI and blastn at various thresholds. We observed an increased host-prediction consensus and accuracy at the species rank using the threshold described above with FastANI with ANI ≥90% based on at least one 5000 bp fragment, compared to blastn thresholds described by Nayfach et al. (ref. [42]). We evaluated the agreement of our two host prediction methods and found up to 58% consensus on host taxonomy on species rank (Supplementary Fig. 11A). We further benchmarked host-prediction purity by calculating the most common host for each viral population according to (1) CRISPR spacer and (2) alignment independently.

Viruses were annotated as temperate virus if (1) the virus was found to be integrated into a MAG with ≥80% query coverage and ANI ≥90% or (2) an integrase protein-annotation could be found in the viral proteome. Integrase

proteins were determined by searching for *integrase* in the InterPro entry description of each interproscan protein-annotation (see Viral taxonomy and function for details).

Differential abundance of viral populations and MAGs. Sample abundance of each viral population was calculated as a mean read per kilobase million (RPKM) of all contigs with at least 75% coverage belonging to a VAMB cluster. Differential abundance analysis of all viruses was tested using the Linear-mixed-effect model R-function *lmer* (lme4 package v. 1.1-26)[77]. The model used was 'Virus ~ dysbiosis_index + diagnosis + sex + (1|Subject)'. Subjects were included as random effects to account for the correlations in the repeated measures (denoted as (1 | subject)) and the log-transformed relative abundance of each virus was modelled as a function of diagnosis (a categorical variable with nonIBD as the reference group) and the dysbiosis index (continuous covariate) while adjusting for subjects age as a continuous covariate and sex as a binary variable.

Definition of boxplots. The lower and upper hinges correspond to the first and third quartiles (25th and 75th percentiles). Centre corresponds to the median. The upper and lower whiskers extend from the hinge to the highest and lowest values, respectively, but no further than 1.5 × interquartile range (IQR) from the hinge. IQR is the distance between the first and third quartiles. Data beyond the ends of whiskers are outliers and are plotted individually. This definition is used for all main and supplementary figures displaying a boxplot.

Data availability

The Diabimmune dataset and HMP2 datasets are available from the European Nucleotide Archive with the accessions PRJNA387903 and PRJNA398089. The COPSAC metagenomics and metaviromics datasets are available with the accessions PRJNA715601 and PRJEB46943, respectively. Gold standard virus genomes for COPSAC and Diabimmune were provided by Shiraz Shah and Tommi Vatanen, respectively, and are available on Zenodo: https://doi.org/10.5281/zenodo.5821973. A CodeOcean capsule of PHAMB v.1.0, including a dataset of 3,000 contigs from 5 HMP2 samples, is available at CodeOcean (https://doi.org/10.24433/CO.4597219.v1). Furthermore, the capsule includes a Dockerfile encoding required databases, Python modules, Snakemake and DeepVirFinder dependencies. Genomes used in the viral CAMISIM benchmark have been uploaded to Zenodo and are available here: https://doi.org/10.5281/zenodo.5821973. Simulated genomes are listed in Supplementary Data 4, entries were collected from the PLSDB database (v. 2021_06_23), MGV database (2021), NCBI Refseq (May 2021). Source data is provided with this paper. Source data are provided with this paper.

Code availability

The VAMB code is available at https://github.com/RasmussenLab/vamb and the PHAMB workflow is available at https://github.com/RasmussenLab/phamb.

Received: 6 August 2021; Accepted: 28 January 2022;
Published online: 18 February 2022

References

1. Kostic, A. D., Xavier, R. J. & Gevers, D. The microbiome in inflammatory bowel disease: current status and the future ahead. *Gastroenterology* **146**, 1489–1499 (2014).
2. Tanoue, T. et al. A defined commensal consortium elicits CD8 T cells and anti-cancer immunity. *Nature* **565**, 600–605 (2019).
3. Gurung, M. et al. Role of gut microbiota in type 2 diabetes pathophysiology. *EBioMedicine* **51**, 102590 (2020).
4. Schirmer, M., Garner, A., Vlamakis, H. & Xavier, R. J. Microbial genes and pathways in inflammatory bowel disease. *Nat. Rev. Microbiol.* **17**, 497–511 (2019).
5. Chen, L. et al. Gut microbial co-abundance networks show specificity in inflammatory bowel disease and obesity. *Nat. Commun.* **11**, 1–12 (2020).
6. Norman, J. M. et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160**, 447–460 (2015).
7. Manrique, P. et al. Healthy human gut phageome. *Proc. Natl Acad. Sci. USA* **113**, 10400–10405 (2016).
8. Gogokhia, L. et al. Expansion of bacteriophages is linked to aggravated intestinal inflammation and colitis. *Cell Host Microbe* **25**, 285–299.e8 (2019).
9. Maronek, M., Link, R., Ambro, L. & Gardlik, R. Phages and their role in gastrointestinal disease: focus on inflammatory bowel disease. *Cells* **9**, 1013 (2020).
10. Clooney, A. G. et al. Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. *Cell Host Microbe* **26**, 764–778.e5 (2019).
11. Cornuault, J. K. et al. Phages infecting Faecalibacterium prausnitzii belong to novel viral genera that help to decipher intestinal viromes. *Microbiome* **6**, 65 (2018).
12. Adiliaghdam, F. & Jeffrey, K. L. Illuminating the human virome in health and disease. *Genome Med.* **12**, 66 (2020).
13. Smits, S. L. et al. Assembly of viral genomes from metagenomes. *Front. Microbiol.* **5**, 714 (2014).
14. García-López, R., Vázquez-Castellanos, J. F. & Moya, A. Fragmentation and coverage variation in viral metagenome assemblies, and their effect in diversity calculations. *Front. Bioeng. Biotechnol.* **3**, 141 (2015).
15. Sutton, T. D. S., Clooney, A. G., Ryan, F. J., Ross, R. P. & Hill, C. Choice of assembly software has a critical impact on virome characterisation. *Microbiome* **7**, 12 (2019).
16. Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* **4**, 470–483 (2009).
17. Castro-Mejía, J. L. et al. Optimizing protocols for extraction of bacteriophages prior to metagenomic analyses of phage communities in the human gut. *Microbiome* **3**, 64 (2015).
18. Roux, S. et al. Minimum information about an uncultivated virus genome (MIUViG). *Nat. Biotechnol.* **37**, 29–37 (2019).
19. Roux, S. et al. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res.* **49**, D764–D775 (2021).
20. Nayfach, S. et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).
21. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098–1109.e9 (2021).
22. Jurtz, V. I., Villarroel, J., Lund, O., Voldby Larsen, M. & Nielsen, M. MetaPhinder-identifying bacteriophage sequences in metagenomic data sets. *PLoS ONE* **11**, e0163111 (2016).
23. Abdelkareem, A. O., Khalil, M. I., Elbehery, A. H. A. & Abbas, H. M. Viral sequence identification in metagenomes using natural language processing techniques. Preprint at *bioRxiv* https://doi.org/10.1101/2020.01.10.892158 (2020).
24. Sirén, K. et al. Rapid discovery of novel prophages using biological feature engineering and machine learning. *NAR Genom. Bioinform.* **3**, lqaa109 (2020).
25. Ren, J. et al. Identifying viruses from metagenomic data using deep learning. *Quant. Biol.* **8**, 64–77 (2020).
26. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).
27. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
28. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
29. Nissen, J. N. et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-020-00777-4 (2021).
30. Guo, J. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
31. Antipov, R. OUP accepted manuscript. *Bioinformatics* (2020).
32. Sullivan, M. B. Viromes, not gene markers, for studying double-stranded DNA virus communities. *J. Virol.* **89**, 2459–2461 (2015).
33. Shah, S. A. et al. Manual resolution of virome dark matter uncovers hundreds of viral families in the infant gut. Preprint at *bioRxiv* https://doi.org/10.1101/2021.07.02.450849 (2021).
34. Redgwell, T. A. et al. Prophages in the infant gut are largely induced, and may be functionally relevant to their hosts. Preprint at *bioRxiv* https://doi.org/10.1101/2021.06.25.449885 (2021).
35. Zhao, G. et al. Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. *Proc. Natl Acad. Sci. USA* **114**, E6166–E6175 (2017).
36. Gregory, A. C. et al. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. *Cell Host Microbe* **28**, 724–740.e8 (2020).
37. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
38. Yuan, Y. & Gao, M. Jumbo bacteriophages: an overview. *Front. Microbiol.* **8**, 403 (2017).
39. Dutilh, B. E. et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5**, 4498 (2014).
40. Shkoporov, A. N. et al. The human gut virome is highly diverse, stable, and individual specific. *Cell Host Microbe* **26**, 527–541.e5 (2019).
41. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).

42. Nayfach, S. et al. A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2020).
43. Dion, M. B. et al. Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. *Nucleic Acids Res.* **49**, 3127–3138 (2021).
44. Nobrega, F. L. et al. Targeting mechanisms of tailed bacteriophages. *Nat. Rev. Microbiol.* **16**, 760–773 (2018).
45. Graham, D. B. et al. Antigen discovery and specification of immunodominance hierarchies for MHCII-restricted epitopes. *Nat. Med.* **24**, 1762–1772 (2018).
46. Benler, S. et al. A diversity-generating retroelement encoded by a globally ubiquitous Bacteroides phage. *Microbiome* **6**, 191 (2018).
47. Mruk, I. & Kobayashi, I. To be or not to be: regulation of restriction–modification systems and other toxin–antitoxin systems. *Nucleic Acids Res.* **42**, 70–86 (2013).
48. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Comprehensive comparative-genomic analysis of type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. *Biol. Direct* **4**, 19 (2009).
49. Luque, A. & Silveira, C. B. Quantification of lysogeny caused by phage coinfections in microbial communities from biophysical principles. *mSystems* **5**, e00353 (2020).
50. Knowles, B. et al. Lytic to temperate switching of viral communities. *Nature* **531**, 466–470 (2016).
51. Alrasheed, H., Jin, R. & Weitz, J. S. Caution in inferring viral strategies from abundance correlations in marine metagenomes. *Nat. Commun.* **10**, 1–4 (2019).
52. Fluckiger, A., Daillere, R., Sassi, M., Sixt, B. S. & Liu, P. Cross-reactivity between tumor MHC class I–restricted antigens and an enterococcal bacteriophage. *Science* **369**, 936–942 (2020).
53. Arisdakessian, C. G., Nigro, O., Steward, G., Poisson, G. & Belcaid, M. CoCoNet: an efficient deep learning tool for viral metagenome binning. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btab213 (2021).
54. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
55. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
56. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
57. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
58. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
59. Potter, S. C. et al. HMMER web server: 2018 update. *Nucleic Acids Res.* **46**, W200–W204 (2018).
60. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
61. Hugoson, E., Lam, W. T. & Guy, L. miComplete: weighted quality evaluation of assembled microbial genomes. *Bioinformatics* **36**, 936–937 (2020).
62. Zolfo, M. et al. Detecting contamination in viromes using ViromeQC. *Nat. Biotechnol.* **37**, 1408–1412 (2019).
63. Johnson, M. et al. NCBI BLAST: a better web interface. *Nucleic Acids Res.* **36**, W5–W9 (2008).
64. Auslander, N., Gussow, A. B., Benler, S., Wolf, Y. I. & Koonin, E. V. Seeker: alignment-free identification of bacteriophage genomes by deep learning. *Nucleic Acids Res.* **48**, e121 (2020).
65. Fritz, A. et al. CAMISIM: simulating metagenomes and microbial communities. *Microbiome* **7**, 17 (2019).
66. Galata, V., Fehlmann, T., Backes, C. & Keller, A. PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res.* **47**, D195–D202 (2019).
67. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
68. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* **11**, 119 (2010).
69. Guerin, E. et al. Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. *Cell Host Microbe* **24**, 653–664.e6 (2018).
70. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
71. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
72. Roux, S. et al. Ecology and molecular targets of hypermutation in the global microbiome. *Nat. Commun.* **12**, 3076 (2021).
73. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
74. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
75. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
76. Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S. A. & Sørensen, S. J. CRISPRCasTyper: automated identification, annotation, and classification of CRISPR-Cas Loci. *CRISPR J.* **3**, 462–469 (2020).
77. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **64**, 1–48 (2015).

Acknowledgements

We thank Mani Arumugam, Eduardo Rocha, Nicolas Rascovan, Ramnik Xavier and Hera Vlamakis for fruitful discussions. J.J., J.N.N. and S.R. were supported by the Novo Nordisk Foundation (grant NNF14CC0001). COPSAC authors were supported by The Lundbeck Foundation (Grant no R16-A1694); The Ministry of Health (Grant no 903516); Danish Council for Strategic Research (Grant no 0603-00280B) and The Capital Region Research Foundation have provided core support to the COPSAC research center.

Author contributions

S.R. conceived the study and guided the analysis. J.J. wrote the software, performed the analyses and wrote the manuscript. S.A.S., J.S., L.D., and D.S.N. generated metavirome data and created the viral gold standard for COPSAC data. S.J.S. and J.S. generated COPSAC metagenome data. D.R.P. and J.N.N. guided the analyses. J.J., S.R., M.L.J. and D.R.P. wrote the manuscript with contributions from all co-authors. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-28581-5.

Correspondence and requests for materials should be addressed to Simon Rasmussen.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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

8.3 Paper III: Centenarians have a diverse population of gut bacteriophages that may promote healthy lifespan

Johansen, J; Atarashi K., S.; Arai, Y.; Hirose, N.; Atarashi K.; Honda, K.; Sørensen J, Xavier R. J.†; Rasmussen, S.†; Plichta R.D.† (2022). *Centenarians have a diverse population of gut bacteriophages that may promote healthy lifespan.*
In review, Nature Microbiology

Supplementary files
`https://www.dropbox.com/sh/mig7ff8ss8hw4uf/AAAO5Mz-8Rt06py77QOMF8DSa?dl=0`

Associated manuscript files
`https://zenodo.org/record/6579480#.Yo3xHZNBweY`

Centenarians have a diverse population of gut bacteriophages that may promote healthy lifespan

Author list

Joachim Johansen[1,2], Koji Atarashi[3], Yasumichi Arai[4], Nobuyoshi Hirose[4], Søren J. Sørensen[5], Tommi Vatanen[2,6,7], Mikael Knip[7,8,9], Kenya Honda[3], Ramnik J. Xavier[2], Simon Rasmussen[1], Damian R. Plichta[2]

Affiliations

[1] Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

[2] Infectious Disease and Microbiome Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[3] Department of Microbiology and Immunology, Keio University School of Medicine, Tokyo, Japan

[4] Center for Supercentenarian Medical Research, Keio University School of Medicine, Tokyo, Japan

[5] Section of Microbiology, Department of Biology, University of Copenhagen, Copenhagen, Denmark

[6] Liggins Institute, University of Auckland, Auckland, New Zealand

[7] Research Program for Clinical and Molecular Metabolism, Faculty of Medicine, University of Helsinki, Helsinki, Finland

[8] New Children's Hospital, Helsinki University Hospital, Helsinki, Finland

[9] Tampere Center for Child Health Research, Tampere University Hospital, Tampere, Finland

Author List Footnotes

Correspondence: Ramnik J. Xavier (rxavier@broadinstitute.org), Simon Rasmussen (simon.rasmussen@cpr.ku.dk), Damian R. Plichta (damian@broadinstitute.org).

Abstract

Distinct gut microbiome ecology may be implicated in the prevention of aging related diseases as it influences systemic immune function and resistance to infections. Yet, many facets of the microbiome throughout different stages in life remain unexplored including viruses and their interactions with bacterial constituents. Here we present a characterisation of the centenarian gut virome and outline the effect of age on viral hubs of importance in tandem with bacterial communities. We identified enrichment of novel viral diversity in centenarians, including viral genera associated with *Clostridia*, and a population shift towards higher lytic activity. Finally, we investigated phage encoded auxiliary functions that influence bacterial physiology by supporting key steps in sulfate metabolic pathways. Phage and bacterial constituents in the centenarian microbiome displayed an increased potential for converting methionine to homocysteine, sulfate to sulfide and taurine to sulfide. A greater metabolic output of microbial hydrogen sulfide in centenarians may in turn support mucosal integrity and resistance to pathobionts.

Main

The healthy centenarian gut microbiome is rich in bacteria, exhibits a unique microbial signature and metabolic fingerprint characterized by novel secondary bile acids[1,2]. As a result, the bacterial community may be implicated in warding off infections and governing gut homeostasis thereby promoting healthy aging. Few studies have yet to address the effect of human aging on the gut virome community nor in the context of immunosenescence experienced by elderly populations[3]. In fact, a recent study by Gregory *et al.* represents one of the only established gut virome comparisons between elderly (+65) and younger subjects, which suggested an overall decline in viral richness and an increase of crAss-phages with age[4]. As such, the effect of healthy aging on the gut virome and its implications on microbial ecology remains largely unexplored and begs for further investigation. Indeed the virome's influence on the microbiome is likely profound as viruses represent a great reservoir of genetic diversity and can offer a myriad of selective advantages to bacterial hosts to sustain dominance within an environment and sustain the phage itself. By complementing bacterias influence during various conditions such as sustaining gut homeostasis or potentially aggravating disease, viruses are targets for understanding health and disease states[5].

Viruses infecting bacteria, known as bacteriophages, depend on the presence of a suitable host bacteria in the environment. Bacteriophages bind to a host and either propagate as a prophage by integrating into the host or by bacterial takeover, virion production and lysis of the host[6]. These two canonical strategies represent the lysogenic and the lytic stages of bacteriophages respectively, yet variations and exceptions to these strategies are continuously discovered. Viruses in a microbiome can be studied by enrichment and sequencing of viral particles or by direct mining of bulk metagenomic data. While viral-enrichment may be biased towards rare and virulent viruses, bulk metagenomics emphasizes temperate viruses associated with dominant host bacteria[7]. In the healthy gut, phages are known to persist for up to 4 years[8,9], though this varies by community context, such as recent antibiotic usage or severe inflammation[10,11]. Some families of viruses belonging to the Crassvirales order were found to show high stability in the human gut when sampled several

years apart[8]. With aging follows increasing inflammaging that creates a persistent pro-inflammatory environment[12]. Inflammatory molecules are known drivers of prophage induction as they invoke bacterial SOS response leading to lytic phage cycles[13]. As the presence of free phages are suggested to stimulate a host immune response by the activation of epithelial Toll-like receptors (TLRs)[11], the degree of lytic activity in the viral community could be a factor in the inflammaging paradigm of aging.

Current models based on infant gut studies, describe that the viral community experiences dramatic shifts during infancy. In the early infant gut, the viral bacterial ratio (VBR) exceeds 1 due to a high load of viral particles that is largely dominated by the induction of prophages responding to environmental cues at a time with low bacterial density[14,15]. Subsequently follows months of bacterial colonization and stabilization that in turn leads to higher viral lysogeny and a lower VBR that is sustained into adulthood[16,17]. Community lysogeny as a preferred lifestyle might be the product of optimal survival conditions for both bacteriophages and bacteria as integrated phages may protect the bacteria from virulent phages by superinfection immunity mechanisms[18]. During the lysogenic state, viruses may also augment bacterial host metabolism by encoding auxiliary metabolic genes (AMG)[19]. For instance, phages encode genes implicated in sulfur metabolism have recently been identified in viruses sampled from human environments[20,21]. In addition, a recent report highlighted the influence of bacteriophages on the microbiome-brain axis where supplementation of an assembled phageome translated into improved memory and cognition in a mice model, which also altered bacterial metabolism[22]. Yet, additional work is needed to establish how viral encoded AMG may influence bacterial metabolism in the human gut and its further implications on human health.

In this study we present an investigation into gut virome in Japanese centenarians[2] and comparisons to younger adults (>18 years) and elderly (>60 years). We applied a virome discovery approach to bulk metagenomics data using *de novo* assemblies of viruses and prophages mined with state of the art computational methods **(Figure 1a)**. The uncovered viral biodiversity consisted of 4422 viral operational taxonomic units (vOTUs), including 1746 novel vOTUs and greatly expanded known viral genera associated with *Clostridia* bacteria. The prevalence of these viral genera

were further investigated in gut metagenomes of Sardinian centenarians[23]. We also observed a shift towards a viral lytic activity in centenarians**,** which we compared with gut viromes mined in two independent age-reference cohorts **(Figure 1b)**. Finally, we found that the centenarian virome contributes in a higher degree to key steps in sulfur metabolic pathways (especially conversion of methionine to homocysteine [Hcy]) mediated by gut bacteria **(Figure 1c)**. This in turn revealed that the microbiome of centenarians harbors a microbiome configuration with greater potential for converting methionine to homocysteine, sulfate to sulfide and taurine to sulfide.

Results

Establishing reference of viral species in centenarians

Since viruses in centenarians gut microbiome have not been characterized before, we delineated the virome by combining viral-binning using a Variational Autoencoder Metagenomic for Binning (VAMB)[24], phage metagenomic binning (PHAMB)[25] and provirus search in bacterial MAGs. As such, we mined *de novo* assembled viruses from bulk metagenomics without viral particle specific enrichment (VLPs) in previously published stool samples from centenarians[2,23] [n = 195 (172 individuals), predominantly female], elderly (n = 133), young adults (n = 61)] **(Figure 1a)**. Bacterial species were characterized as Metagenomic Species (MSPs) and MAGs. Following a series of strict viral annotation rules (see methods), we reduced the list of assembled virus contigs and viral bins into 4422 dereplicated vOTUs (viral OTUs) across samples representing all age groups. Whole genome species-level clustering (ANI>95% & AF>85%) with the entire viral metagenomic gut virus (MGV) database[26] revealed that 1746 of 4422 vOTUs were novel as they did not form a vOTU cluster with any MGV reference. In order to place these novel vOTUs within the context of known viral diversity, we used a set of taxonomically informative viral-protein markers and identified 2,388 (54%) vOTUs with at least 3 of the 77 markers necessary for classification using a concatenated protein tree method[27]. Thus, we established a viral phylogenetic tree **(Figure 1d)** using the 2,388 *de novo* assembled vOTUs and the 37,946 MGV vOTUs, of which the majority made it into the final tree. Performing viral genus level clustering with the MGV resulted in massive expansion

of a viral genus G7 containing 228 MGV Clostridia;Ruminococcus viruses, adding 389 vOTUs (304 novel) to a total of 532 vOTUs (**G7**, **Figure 1d, Supplementary file 1**). Hence, G7 represents a viral genera enriched with novel vOTUs present in virtually every Japanese centenarian (94%) and a large portion of Sardinian centenarians (52%)**.** In addition, we added 108 (78 novel) vOTUs to an phylogenetically adjacent genus in the tree G78 containing 33 MGV Clostridia;Ruminococcus viruses, thereby expanding the genus to 111 vOTUs (**G78**, **Figure 1d, Supplementary file 1**). In addition, species-level clustering of vOTUs with integrated prophages of centenarian bacterial isolates were used to link viruses from *Lachnospiraceae* (st 61+62)*, C. scindens (*st 59+60), *C. innocuum* (st 51) and *C. symbiosum* (st 65+66) to G7 (**Sup. Figure 1A**, **Supplementary file 2**), and further supported G7 as a viral genus associated with Clostridia sp. Combined, hundreds of novel vOTU were added to the G7 and G78 genus associated with Clostridia;Ruminococcus. Conversely, 137 vOTUs (21 novel) were added to a phylogenetically distant genus G2 containing 855 Clostridia;*Faecalibacterium prausnitzii* viruses to a total of 872 (**G2, Figure 1d, Supplementary file 1**), showing that few novel *F. prausnitzii* vOTUs were discovered from the Centenarian and elderly microbiomes. Interestingly 741/1517 Low-quality (LQ) vOTUs (median size of 14.3 kbp and max 114 kbp) contained at least 3 informative viral-protein markers, hence several of these viruses may represent fragmented but also full sized viruses too distinct from viral species in the CheckV database to properly estimate high completeness. Altogether we unraveled a large trove of novel viral diversity that could be used for investigating the age-related differences in the centenarian and control viromes.

Centenarians display a more diverse and richer virome compared to young and elderly

To characterize the virome configuration in the centenarians, elderly and young adults, we computed and analyzed vOTU abundance profiles. We focused the analysis on the Japanese cohort (unless stated otherwise) to not confound potential age-dependent effects with geographic virome differences. Overall, 49% of all vOTUs were detected in at least one sample of every age group, while 12% were exclusively found in centenarians **(Sup. Figure 1B)**. A principal coordinate analysis (PCoA) revealed separation of the individual gut viromes over an age gradient where

the first and second principal component explained 7.85% and 4.18% of the variance, respectively **(Figure 1e)**. In agreement, centenarian viromes were clearly differentiated based on vOTU abundance profiles from young and elderly subjects using PERMANOVA (P=0.001, R2=2.17%, F=3.67), where age-group explained more variance than bacterial community types established by Sato *et al* (P=0.001, R2=1.33%, F=4.49). Nevertheless, while centenarian microbiomes harbored viruses also present in other age groups, they also carried a set of unique viral populations. Furthermore, alpha-diversity metrics also showed more diverse (T-test, two sided, P-value<3.22e-07) and rich (T-test, two sided, P-value<4.37e-09, FDR corrected) centenarian viromes compared to both younger and elderly subjects **(Figure 1f+g).** In fact we observed an increasing virome diversity with age, which is in contrast to a previous characterisation of elderly gut microbiomes that observed an overall depletion in viral diversity in elderly gut viromes[4]. In accordance with the literature[4], we found the crAss-phage and Microviridae family at relatively higher abundance (Wilcoxon rank sum test, two sided, P-value<0.05, FDR corrected) in centenarians compared to the young control **(Sup. Figure 1C)**, thereby providing additional evidence that these viral families thrive well into the extreme end of human age. CrAss-phages infecting *Alistipes shahii* were twice as prevalent in centenarians and indicates a shift in crAss-phage host preference associated with aging **(Sup. Figure 1D).**

The centenarian viral signature is partially explained by novel vOTUs

We further evaluated the influence of the new viral populations in the centenarian microbiomes. We found that the observed age-gradient in PCoA space could be partially explained by the novel vOTUs as they accounted for 30%, 20% and 12% of the virome abundance in the centenarian, elderly and young samples, respectively **(Figure 1h)**. Thus, a much larger proportion of mapped reads in the centenarian samples originated from novel vOTUs. In addition, we found that increasing viral diversity coincided with a loss of core-virus abundance (Cor -0.29, P= 9.57e-08) specifically viruses with more than 10% prevalence across all age groups **(Sup. Figure 1E)**. Therefore core-viruses make up a smaller part of the total virome in elderly and centenarian **(Figure 1i, Sup. Figure 1G)**. Furthermore, the total core-virus abundance displayed strong correlation with the two principal components derived in the PCoA analysis (PC1 Cor 0.33, P=1.07e-09; PC2 Cor 0.38, P=9.18e-

13) **(Sup. Figure 1F)** suggesting that the overall increase in viral diversity and the relative depletion of core-viruses characterizes the dynamics of gut virome in extreme longevity**.** Based on the abundance profiles, we also identified 483 vOTUs enriched in centenarians (Wilcoxon rank sum test test, CE vs young P<0.05 or CE vs elderly P<0.05, FDR corrected) where more than half (257/483) corresponded to novel vOTUs. In summary, the centenarian virome is populated by abundant novel viruses and exhibits greater viral diversity. Altogether, this signature may be attributed to the robust and healthy centenarian microbiome.

Phage signatures correlate with the unique centenarian bacterial communities

To further characterize the ecological hubs of bacteria and viruses that are central to the centenarian microbiota, we analyzed the viral composition in each age group according to their predicted bacterial host based on CRISPR annotation and prophage integration events. Using viral host-labels, we identified several overabundant (Wald's test, two sided, Q-value<0.01) groups of viruses that were associated with Alistipes, Parabacteroides, Clostridium, Eggerthella, Ruminococcus and Akkermansia on species level (**Figure 2a**). For instance, viruses associated with *Clostridium scindens* were more prevalent and abundant in the centenarian microbiome compared to elderly and young controls (**Figure 2b)**. Conversely we found a significant depletion (Wald's test, two sided, Q-value<0.01) of *Bacteroides* and *Faecalibacterium* viruses in the centenarian microbiomes (**Figure 2a+c)** that coincided with the relative depletion of these commensal bacteria described already[2]. We hypothesized that the trend could reflect the abundance profile of the viral bacterial hosts and thus correlated viral profiles of predicted temperate viruses to their predicted host. Here we found the profile of overabundant bacteria such as *Clostridium scindens* and its associated viruses were significantly correlated (Cor= 0.30, P=4.49e-08). A similar trend was found for *Akkermansia muciniphila* (Cor=0.25, P=6.23e-11), *Enterocloster bolteae* (Cor=0.51, P=4.82e-66) and *Parabacteroides distasonis* (Cor=0.35, P=6.00e-11) (**Figure 2d)**. These results suggested that the viral profile corresponded well to the state of the bacterial host and pointed to the "Piggybacking the Winner" (PtW) dynamics between hosts and viruses; this correlation was further supported by a permutation test on abundance (**Sup. Figure 2A-C)**. Interestingly, for some bacterial hosts like *Alistipes shahii* (Cor=0.20, P=1.83e-04), PtW dynamics did not seem to apply to the same extent

after further correlation evaluation and suggested that the viral lifestyle of *Alistipes* temperate viruses may be in a state of higher lytic activity (**Sup. Figure 2A-C)**.

Lysogeny in the virome dominates from young to old age in the healthy microbiome

We hypothesized that the preference for a lytic interaction between the temperate viruses and their host, as observed for the aging associated *A. shahii*[2], may be increased in the centenarian microbiome. To evaluate it, we derived community viral-bacterial ratios (VBR) for vOTUs predicted as temperate viruses. VBR values above 1 suggest lytic community activity and below suggest lysogeny between bacteria and temperate viruses. Centenarian microbiomes displayed a higher VBR relative to elderly (Wilcoxon rank sum test, two-sided, P=7.774e-09, W=5751) and young adults (Wilcoxon rank sum test, two-sided, P=2.2e-16, W=10286) (**Figure 3a)**, which suggests a potential shift driven by temperate viruses adopting a lytic lifestyle. Furthermore, the relative abundance of temperate viruses decreased in centenarians and could indicate a disruption of the lysogenic community (**Sup. Figure 3A)**; this development could not be explained by sequencing depth as they were comparable between age groups (**Sup. Figure 3B-C)**.

We included a cohort of healthy infants from Finland profiled from birth until one year of age (EDIA[28], n=562 samples) to compare how the VBR trends early in life compare to those in centenarians **(Figure 1b)**. We assumed that if the calculated VBR distributions captured overall trends or differences of lysogeny in the microbiome, we could outline cohort specific viral-bacterial interactions for different age-groups regardless of bacterial community composition. The early infant gut virome has been previously characterized as dominated by lytic activity by temperate viruses[14,15] and we accordingly found that the VBR in the analyzed infants peaked around 3-4 months of age and then decreased until 12 months of age **(Sup. Figure 3D)**. This fits with the current understanding that increasing bacterial densities during gut maturation promotes lysogenic maintenance of temperate viruses and decreasing VBR[16]. Furthermore, the relative abundances of temperate viruses significantly expanded from 6 to 12 months of age (**Sup. Figure 3E).** Compared with centenarians, the community VBR in infants was significantly higher (Wilcoxon rank sum test, two-sided, P=3.039e-05, W=57563) (**Figure 3a)**, which could be explained

by the low number of viral coinfections and spontaneously induced temperate phages as the bacterial community is establishing itself in the infant gut[15,29]. Finally, we included additional healthy young adults from geographically distinct population (Tanzania[30], n=234 samples, >18 and <60 years, **Figure 1b**) and observed that their VBR distribution resembled that of the young adults from Japan (Wilcoxon rank sum test, two-sided, P=0.1409, W=4427) (**Sup. Figure 3F)**. This suggests that the calculated VBR captures an estimate of microbiome lysogeny that is more explained by age than geographic differences. Altogether our analysis confirmed that temperate viruses expand over time in the infant gut and manifest into a stable lysogenic community throughout adulthood (**Sup. Figure 3D-E)**. We further identified a return to an increased lytic community in very old individuals.

The age-related shift into more lytic activity of temperate phages observed in centenarians could be influenced by other abiotic factors such as slower transit-time that promotes more viral and bacterial interactions in the colon or low-grade inflammation that induces prophages[16]. We examined the relationship of gut inflammatory markers and the VBR-shift, by correlating VBR estimates in centenarians with C-reactive protein (Spearman rank, Cor=-0.015, P=0.79) and lipocalin measurements (Spearman rank, Cor=-0.043, P=0.58) but found no correlation. Instead, we focused on dissecting the VBR distribution to identify the underlying prevalent bacterial hosts and viruses that might explain the VBR difference between the age groups. Host species where the viral abundance exceeded that of its host (VBR>1) included *B. intestinihominis, B. caccae, E. boltea, F. prausnitzii, Acetatifactor sp900066365, A. shahii* and *P. merdae* (**Figure 3b)**.

We also looked at intrinsically virulent crAss-like viruses[31] and found a significant VBR distribution shift for *A. shahii* (Wilcoxon rank sum test, two-sided, P=0.04, W=10577) and *P. merdae* (Wilcoxon rank sum test, two-sided, P=0.03, W=115949) compared to remaining viruses (**Figure 3c),** which indicates independent host-interactions from other vOTUs**.** In addition, viruses associated with bacterial species depleted in the centenarians like *F. prausnitzii* displayed subgroups of vOTUs with very different VBRs. vOTUs enriched in centenarian samples (5 virulent and 2 temperate vOTUs) were negatively correlated with *F. prausnitzii* **(Figure 3d)** while the remaining viruses contained 131 temperate viruses and only 36 virulent ones.

Here the temperate viruses were positively correlated with *F. prausnitzii* (**Figure 3d),** which illustrates the complexity of the bacterial viral interactions in the microbiome where viruses associated with the same host species can persist with different phage-lifestyles (**Sup. Figure 3G+H)**.

The centenarian virome supports a greater host-metabolic conversion of microbial sulfide

An avenue of viral host dynamics is the viral contribution of auxiliary metabolic genes (AMG), corresponding to viral encoded genes homologous to host genes in major bacterial metabolic networks. AMGs have been shown to influence host-metabolic enzymes and indirectly support viral fitness in marine bacteria[32]. The most frequently annotated AMG across all vOTU were related to sulfur metabolic pathways, for instance the most widespread AMG, dcm, could be found in 47% of AMG encoding vOTUs (**Sup. Figure 4A)**. Next we wanted to further investigate how the gut virome contributes to sulfate reductive metabolic pathways on a population level relative to bacteria and how this is affected by aging. In total we found that 25% of all vOTUs (n=1,050) encoded at least one established AMG (**Sup. Figure 4A)**. In comparison, it has been reported before that 6% of gastrointestinal viruses encode 1 or more AMG[21]. Known AMGs only made up 3% of viral proteomes and as 70-75% of viral proteomes remain unannotated (**Sup. Figure 4B-C)**, we focused the analysis on this snippet of auxiliary genes including sulfur related enzymes.

When the abundance of AMGs were summarized on population level, we found differential higher viral levels of metK (*S*-adenosylmethionine [SAM] synthase, converts methionine to SAM), dcm (DNA-cytosine methyltransferase, methylates SAM to *S*-adenosyl-L-homocysteine [SAH]), cysH (Phosphoadenosine 5'-phosphosulfate [PAPS] reductase, converts PAPS to sulfite), mec (CysO-cysteine peptidase, hydrolyzes CysO-cysteine adduct) and iscS (Cysteine desulfurase, catalyzes removal of sulfur from cysteine) in Japanese centenarians relative to young adults (**Figure 4a, Supplementary file 3)**. Compared to the cognate genes encoded by bacteria, viruses contributed on average with >7.5% of the total abundance for cysH, metK and mec and almost 50% for the dcm gene across all age groups (**Figure 4a)**. Thus, the virome contributes in a higher degree to backbone enzymes in the sulfur metabolic pathway and especially in centenarians for

355 enzymatic steps related to methionine conversion to homocysteine (Hcy).
356
357 In addition, high prevalence (>90%) of viruses encoding AMGs for dcm and cysH
358 was found for every age group (**Figure 4b).** Interestingly, since the virome
359 contributed with almost 50% of the dcm gene abundance, we sought to determine
360 the conservation of the cognate genes in bacterial genomes and found that 80%
361 (218/272) were annotated as accessory and the remaining 20% as core bacterial
362 genes (**Figure 4c)**. This illustrated that the majority of bacterial encoded dcm can be
363 found in the flexible bacterial pangenome and may be subject to transfer via phages
364 or other mobile genetic elements (MGEs). Similarly, cysH was also largely identified
365 (45%) as an accessory gene in MSPs, which is not surprising if this enzyme has a
366 dual role in phages like the dcm methylase that support SAM to SAH conversion in
367 bacteria but is also suggested to protect the phage against the host restriction
368 modification system during viral infection[33]. For the remaining bacterial genes, 80%
369 were typically annotated as MSP core genes, thus more conserved in bacterial
370 genomes. In addition, we found the viral gene abundance positively correlated with
371 the MSP counterpart for dcm, mec, metK and cysH (Cor=0.16-0.42, Adj. P-
372 value<0.05) suggesting that these genes were encoded by lysogenic viruses
373 expanding in tandem with the host and not via lytic expansion. In contrast, nadE
374 (Cor=-0.16, CI=-0.05;-0.26, Adj. P-value<0.05) and iscS (Cor=-0.24, CI=-0.15;-0.34,
375 Adj. P-value<0.05) were negatively correlated and more likely encoded by viruses in
376 a lytic state. Since the virome supported a higher metabolic potential of key
377 enzymatic steps in methionine and sulfate conversion, we expanded the analysis to
378 include additional steps not limited to viruses and explore the full microbial sulfate-
379 reduction potential in centenarians.
380
381 **The centenarian gut microbiome configuration is primed for effective**
382 **utilization of taurine, sulfate and methionine.**
383 After expanding the sulfate metabolism analysis with additional enzymes related to
384 dissimilatory sulfate reduction (**Figure 5a)**, including viral encoded AMG, we found
385 that Japanese centenarians displayed a higher microbiome-encoded potential for
386 converting methionine to Hcy (**Figure 5b)** as well as sulfate to sulfide and taurine to
387 sulfide (**Figure 5c)**. These trends could be replicated in the Sardinian cohort that
388 displayed an overall concordance (**Sup. Figure 5A+B).** Importantly, the dsrC

(Dissimilatory sulfite reductase subunit C) enzyme that converts either taurine or sulfate derived sulfite to sulfide was highly enriched in centenarians vs young adults (T-test, P=4.01e-05, t=-4.164, CI=-16.68;-5.97). In addition, we observed a decrease in measured taurine conjugated bile acids in centenarians compared to other age groups, which might have been mediated by *Escherichia flexneri*, *Klebsiella pneumoniae* and *Desulfovibrio fairfieldensis* that encode TauD or dsrC (**Sup. Figure 5C+D)**.

Furthermore, levels of taurine conjugated bile acids were negatively correlated with the total abundance of microbial encoded TauD (T-DCA, Cor -0.23, P=3.97e-03; T-UDCA, Cor -0.27, P=5.34e-03, T-CDCA, Cor -0.28, P=2.20e-04; T-CA, Cor -0.3, P= 1.83e-04, FDR corrected) and dsrC (T-UDCA, Cor -0.26, P=1.73e-02; T-CA, Cor -0.2, P=3.20e-02, FDR corrected) (**Sup. Figure 5E)**. The enzymes cysD, cysC, cysH and cysJ that govern the enzymatic pathway from sulfate to sulfide were all enriched in centenarians as well (T-test, Adj. P<0.01, **Figure 5a, Supplementary file 4**). However, further conversion of sulfide to cysteine by cysM/cysK was comparable between centenarians and young (T-test, Adj. P=0.06) but enriched in elderly. The first two enzymes that govern the conversion of methionine into SAH, dcm and metK, were enriched in centenarians and strengthened by viral AMGs (**Figure 4a).** This AMG reservoir is most likely supporting the host through viruses in their lysogenic state, such as the integrated viruses in *Clostridia* isolates encoding dcm and metK and in *Bacteroides dorei* encoding cysH and cysD **(Figure 5d**). Interestingly, the final conversion of SAH into Hcy seemed to be mostly mediated by ahcY in centenarians while the alternative path LuxS that converts S-ribosylhomocysteine into Hcy was significantly depleted compared to young and elderly **(Figure 5b, Supplementary file 4**). Thus, greater abundance of ahcY and depletion of LuxS suggests higher metabolic support of the direct pathway from SAH to Hcy in centenarians. In essence, we found that the microbiome features a higher metabolic potential for sulfide and homocysteine production in centenarians that is supported by AMGs in centenarian viruses.

Discussion

As centenarians represent a surprisingly robust population with a decreased susceptibility to age-related diseases and infection compared to people decades younger, the key to their longevity remains a topic of interest. Here we presented the first characterization of the gut virome of Japanese centenarians and dissected their ecology in tandem with their unique microbial signature. The human gut virome in centenarians represents a rich and diverse community with unique viral populations that interacts with enriched bacterial taxa identified in centenarians[2]. Furthermore, we found that the centenarians microbiome configuration is highly optimal in terms of sulfur metabolic activity where a large portion of viruses (25%) encode AMGs that support key enzymatic steps in sulfate reducing metabolism.

We next expanded viral biodiversity with 1746 of the vOTUs as they did not resemble genomes of the MGV database[26], of which 462 novel vOTUs were exclusively identified in centenarian microbiomes. Hundreds of novel viruses could be attributed to *Clostridia* bacteria and resulted in a great expansion of an existing viral genus (G7) with 40% of the vOTUs also detected in centenarians residing in Sardinia. We confirmed that viruses from genus G7 comprise prophages of *C. scindens, C. innocuum*, *C. symbosium* and *Lachnospiraceae* in bacterial isolates of a centenarian. Considering that viral populations are known to display long term stability in the gut environment[8,9], centenarian gut microbiomes could represent an archeological site of bacterial and viral ecology with unique populations that may have persisted for almost a century. Even though our analysis presents a bulk of viral diversity that expands existing viral genera, it is important to note that strict viral-filtering might have left out additional viral diversity. In our conservative approach, we filtered out almost 900 vOTUs due to lack of consensus between CheckV, VIBRANT and the MGV viral discovery pipeline. In addition, while viromes extracted with and without viral-enrichment capture overlapping viral diversity[4,25], the latter may miss out on some rare viruses and be biased toward phages infecting dominant host-cells[4,7]. The technical bias may be implicated in the low detection of Microviridae in this analysis and the high detection in studies with viral-enrichment[9].

While the question on bacterial richness and its increase with age is a disputed topic[34], healthy centenarians stand out as a separate population with high bacterial and viral richness as presented in our analysis. On this, we found the diversity and signature of the virome highly correlated with the bacteriome as viral populations interact with thriving bacterial taxa in the environment exemplified by *Clostridium sp.* and *Alistipes sp.* in centenarians. In agreement with the most recent study on the aging virome[4], we found that the crAss-like family generally increases in abundance and prevalence in elderly subjects (>60 years old). Furthermore, we highlighted that the expansionist trend of crAss-like viruses remain strong into the centenarian age. Finally, we added a new branch of the *Bacteroidetes* phylum, *Alistipes shahii,* to the crAss-like host range. As several *Bacteroides* species, which represents a typical crAss virus host, were significantly depleted in centenarians while *Alistipes* on the contrary were enriched, it illustrates that the broad host tropism in the crAss-like virus family is a crucial factor for crAss-like viruses to exhibit a life-long membership in the human gut virome[8,35].

Another surprising observation was the clear separation of VBR distributions by age groups. We found that the gut microbiome of healthy younger individuals represents a state of maximal lysogeny between viruses and bacteria that gradually converges toward more lytic activity with increasing age. Infant virome studies have illustrated that the high viral load in the early infant gut is primarily driven by induced integrated viruses, which peak around 3-4 months but then decline due to increasing bacterial density[14,15]. We found a similar trend in the infant EDIA cohort profiled over one year with an eventual increase in temperate virus abundance. Curiously, the VBR distribution for centenarians was higher than younger controls (excluding infants) and exhibited a drop in the abundance of temperate viruses. This suggested a development in the opposite direction of the maturing infant microbiome. Factors such as age-related mucus-thinning could be involved as it leads to inflammation that stimulates prophage induction[13,36]. It could be postulated that low-grade inflammation related to aging combined with nutrient starvation or other environmental cues leads to an incremental induction of the temperate viral reservoir and higher lytic activity in the centenarian gut[2,37]. In turn, higher abundance of free phages could be a factor in low-grade inflammation observed in the aging gut by stimulating a pro-inflammatory immune response via epithelial TLRs[11].

The microbiome represents a key component in the healthy aging of centenarians, which have been previously exemplified by their unique bile acid composition[2]. Here we provide another facet to the centenarian microbiome ecology by outlining the viral contribution to dissimilatory sulfate pathways that revealed key observations; centenarian viruses contribute in a higher degree to key enzymatic steps related to conversion of sulfate to sulfide and methionine to homocysteine. This in turn revealed that the centenarian microbiome configuration exhibits an increased potential for converting methionine to homocysteine, sulfate to sulfide and taurine to sulfide. These trends were supported in the Sardinian centenarians. *Desulfovibrio sp*. are known sulfate reducers[38], these species and their associated viruses were also enriched in centenarians and may partly explain the greater metabolic capacity. For instance, viruses encoding dcm and metK are widely prevalent in centenarians and contribute up to 50% of the combined viral and bacterial abundance. Similar viral-host contribution ratios have been observed in marine environments for dissimilatory sulfur metabolism genes[20]. We speculate that in centenarians this translates to increased levels of microbially derived sulfide, which may lead to health promoting outcomes[39]. A higher microbial output of H2S promotes colonization resistance and protection against aerobic pathogens, as shown by Stacy *et al.* where infection events lead to colonization resistance when taurine is converted by microbial taxa into hydrogen sulfide[40]. Beyond sulfur metabolism, we expect that other viral AMGs influence bacterial physiology in the centenarian microbiomes. As the majority (~75%) of viral proteomes remains functionally unannotated, improved viral databases and annotation methods will be key for capturing the bigger picture of viral host complementary. Additional sequencing efforts of both bulk and VLP samples combined with long-read sequencing technology will be of immense value for continuing the delineation of viral populations that may benefit human longevity.

Methods

Datasets

The Centenarian Japanese dataset comprises fecal samples of 176 centenarian (>100 years old) 110 elderly controls (<100 years old) and 44 young controls (max. 55 years old) shotgun metagenomic sequenced[2]. We also processed the Sardinian

Centenarian dataset with shotgun samples from 19 centenarians, 23 elderly controls and 17 young controls[23]. The Tanzanian dataset also known as 300 Functional Genomics (300FG) (www.human functional genomics.org) included 315 samples from adults (18-65 years old), of which we selected 234 metagenomic samples from healthy subjects[30]. The infant dataset EDIA is a paired longitudinal study of mothers and infants followed across the first year of the child's life. From this dataset we processed 668 samples of 142 infant subjects[41]. Briefly, we processed and analyzed shotgun metagenomic samples from the cohorts Japanese, Sardinia, Tanzania 300FG and EDIA (infant control) with the same workflow but separately as described further below in the methods section. Multiple cohorts from distinct geographic regions were included in this analysis. However, the cohorts were used in different parts of the manuscript; The Japanese cohort was included in all parts of the manuscript. Metagenomic assemblies from the Sardinian cohort were included in (1) establishing viral (vOTUs) and bacterial diversity (MSPs and MAGs) for the age groups (centenarian, elderly and young) and (2) sulfur metabolic enzyme abundance (see Sup. Figure 5A+B). The EDIA and Tanzania cohort were only included in the analysis related to viral-bacterial ratios (see “Lysogeny in the virome dominates from young to old age in the healthy microbiome”). Fecal bile acids were only available for the Japanese cohort and measured using LC-MS/MS, a thorough description of sample treatment and equipment for measuring is described by Sato *et al.*[2].

Establishing assemblies and gene catalogues

For each of the three datasets Japanese (including Sardinia), Tanzanian 300FG and EDIA infant cohort, shotgun metagenomic samples were quality-controlled, cleaned for human DNA contamination, trimmed for adaptors and trimmed for low-quality sequences (HEADCROP:15, SLIDINGWINDOW:1:20) using Kneaddata to a minimum read length of minimum 50 base pairs (https://github.com/biobakery/kneaddata) and Trimmomatic (v.0.7.2). Each metagenomic sample was assembled individually into contigs using MegaHIT (v.1.2.9, default settings)[42] followed by an open-reading-frame prediction using Prodigal (v.2.6.3, settings -p meta)[43], only predicted genes with a start and stop codon were retained. A non-redundant gene catalogue was constructed by clustering predicted genes based on sequence similarity at 95% identity and 90% coverage of

the shorter sequence using CD-HIT (v.4.8.1)[44]. Reads were then mapped to the gene catalogue with BWA-mem (v.0.7.17)[45] requiring confident mapping to contigs with at least 95% sequence identity over the length of the read, counted (count matrix) and normalized to transcripts per kilobase per million to form a TPM matrix using in-house scripts. The count matrix was processed with MSPminer (v.2.0) using default settings[46] that groups genes of the gene catalogue into metagenomic species (MSPs) pan-genomes (core and accessory genes).

Establishing bacterial MAGs and vOTUs

Contig binning and downstream processing to establish MAGs and vOTUs was conducted separately for each dataset [Japanese (including Sardinia), Tanzanian 300FG and EDIA] with the same computational workflow. Prior to binning, concatenated assemblies were filtered for contigs of minimum 2000 base pairs long. Reads were mapped to contigs with minimap2 (v.2.6)[47] using '-N 50' and filtered with Samtools (v.1.9)[48] using the flag '-F 3584'. Contig abundance profiles for each sample were calculated using the jgi_summarize_bam_contig_depths module from MetaBAT2 (v.2.10.2)[49] and combined to form a jgi-depth matrix for all contigs. Metagenomic bins were established running VAMB (v. 3.1) that applies a deep-learning framework to cluster the metagenomic contigs into putative biological entities using the jgi-depth matrix as input and tetranucleotide-frequencies derived from input sequences. Bacterial bins (MAGs) were identified using the lineage-wf of CheckM (v.1.1.2)[50] and bins with a completeness of >= 50% and contamination <=10%, corresponding to MQ and HQ bins, were retained for further analysis. The taxonomy of each bacterial bin was determined with the *classify-wf* of GTDBK-TK (v.1.7.0)[51] using the GTDBTK-database (release 202).

Viral operational taxonomic units (vOTUs) were obtained through a series of steps to remove potential contaminants and avoid false positive viral inclusion. First, all VAMB bins were parsed with the viral binning method PHAMB[25] then evaluated with the workflow "end_to_end" in CheckV (v0.8.1)[52] and CheckV-database (v.1.0) to establish viral MAGs (vMAGs). In addition, we conducted a provirus search by scanning contigs of all MAGs with CheckV to extract provirus contigs. Second, the CheckV genome result tables were carefully parsed to only vMAGs and viral contigs

who were annotated as Complete (based on direct terminal repeats and inverted terminal repeats), High-quality (HQ) and Medium-quality (MQ) viruses with >=75% average amino acid identity (AAI) across >30% of all encoding proteins and with less than 15% absolute genome-size difference to their closest CheckV database reference. By including viruses with higher AAI and minimum size deviation to references in the CheckV database we basically excluded all viruses predicted by the CheckV HMM-based model. Pro-viruses were also included if they were of at least MQ determined with CheckVs AAI-model using >=75% AAI across >30% of encoded proteins. Third, In order to include potential novel viruses, including Low-quality (LQ) viruses, with little similarity to references in the CheckV database, we included putative viruses with >=10 genes where >=40% of genes were viral-annotated and <10% host-annotated. Fourth, the resulting set of vMAGs and pro-virus contigs were then clustered and dereplicated at >=95% ANI across >=85% of the shorter sequence to form vOTUs using the longest representative sequence resulting in 5252 vOTUs. To get comparable species cluster delineation with the MGV genome database (especially for comparison with the MGV genome repository, see “Novel vOTUs and taxonomic annotation”), clustering was performed using the same all-versus-all local alignments method employed by the authors of the MGV manuscript [26]. Then, all 5251 vOTUs were analyzed with the viral-detection pipeline provided with the MGV database[26] to get an additional viral prediction for each vOTU.

Finally, we selected (1) vOTUs of at least MQ determined by CheckV with the confident AAI-model (2) vOTUs annotated as viral by the MGV viral-detection pipeline (3) and those whose gene-content contained >=40% viral genes and <10% host genes, which resulted in a subset of 4422 vOTUs. Furthermore, we also parsed vOTUs using VIBRANT (v.1.2.1, default settings)[53] and found a 4240/4422 (96%) viral prediction agreement. Completeness estimates derived using CheckV corresponded to a total of 241 Complete (53% novel), 683 HQ (39% novel), 109 MQ (44% novel) and 3389 LQ (45% novel) vOTUs. We note that as contigs were derived from bulk metagenomic samples, RNA viruses are less likely to be assembled and detected.

Abundance profiles of vOTUs were calculated across all samples in the following way; First, viral genes in the gene catalogue were determined by aligning all the predicted genes of each vOTUs using blastn (v. 2.8.1)[54] to the gene catalogue, only hits with a sequence similarity of at least >=95% identity across >=80% of gene length were accepted. Then, viral abundance profiles were calculated as a median TPM based on the confidently aligned genes to the gene catalogue for each vOTU using the depth-normalized TPM matrix (see 'Establishing assemblies and gene catalogues'). Bacterial abundance profiles were also calculated based on the gene catalogue, as the abundance profiles across samples for each MSP were calculated as a median TPM for the 30-top representative core genes reported by MSPminer. In order to investigate potential technical artifacts of gene catalogue derived virus abundance, the viral abundance and genome coverage was also calculated using CoverM (v0.6.1, https://github.com/wwood/CoverM) accepting only reads mapping with >=95% sequence identity over the length of the read.

Novel vOTUs and taxonomic annotation

In order to taxonomically annotate all vOTUs on species level we clustered all viruses with the entire MGV database[26]. This was done separately for each dataset by combining the purified and dereplicated set of vOTUs with all deposited MGV genomes (containing at least MQ viruses) by >=95% ANI across >=85% of the shorter sequence. All vOTUs that clustered with a MGV reference (MGVref) were described as "known viruses" (similar to a MGV genome on species level), for these we transferred taxonomy from the MGV database metadata if available, vOTUs in a cluster without a MGVref were described as novel. Furthermore, to improve the characterization of *Microviridae* vOTUs, we performed clustering with a dedicated Microvirus database[55] at >=90% ANI across >=85% the shorter sequence. To identify genus and family level vOTUs relations, we calculated all-vs-all proteome similarity between all vOTUs and the MGV database followed by clustering using cutoffs used to establish MGV vOTUs. First, viral proteins were predicted of all vOTUs and MGV reference genomes using Prodigal (v.2.6.3, settings -p meta) then aligned all-vs-all using diamond blastp (v.2.0.14)[56] keeping hits with a >50% protein sequence similarity across >50% of the sequence and an evalue <1e-5. At the genus level, we filtered edges between genomes with <40% AAI or <20% genes shared

and used an inflation factor of 2 to get genus level vOTU clusters. At the family level, we filtered edges between genomes with <20% AAI or <10% gene sharing and used an inflation factor of 1.2.

Virus host prediction

Bacterial MQ and HQ bins that were retained after running CheckM were CRISPR-cas typed and mined for CRISPR-spacers and arrays using CrisprCasTyper (v.1.2.3)[57]. In addition, CRISPR-spacers were mined with CRT (v.1.2 -minNR 2)[58]. The combined set of CRISPR-spacers mined from MAGs were then mapped to all vOTUs using blastn (v. 2.8.1)[54]. Subsequently, spacer-hits were processed such that only CRISPR spacer matches with >=80% sequence identity over >=90% of spacer length and maximum 2 mismatches were retained to prevent false-positive host predictions. Retained spacer hits from CrisprCasTyper and CRT were integrated and used to calculate consensus taxonomy at each taxonomic rank based on the lineage of the predicted host bacteria. Consensus was obtained using a plurality rule such that the most frequent host annotation was used if it corresponded to >=40% of all hits, else it was determined “NA”.

Viral lifestyle prediction

In order to get viral lifestyle prediction (temperate or virulent) we used BACPHLIP [59] that searches for lysogeny marker genes in virus genomes such as integrase enzymes. If no lysogeny marker genes are found, the input genome is assumed to be virulent. Thus, completeness of a viral genome is important to keep in mind to get a confident prediction. First, if vOTUs clustered on species level with a MGV reference genome, the predicted MGV genome lifestyle (also by BACHPHLIP) was transferred to the vOTU. Second, novel vOTUs predicted *de novo* as temperate were annotated as such, due to presence of marker lysogeny marker genes in the genome. Third, *de novo* virulent predictions were applied only to vOTUs with at least 90% completeness (High-quality) according to CheckV. vOTUs with a completeness <90% predicted *de novo* as virulent were not assigned a lifestyle because the sequence missing might have contained lysogenic markers. In summary, viral lifestyle annotation (temperate, virulent) resulted in 1832 *temperate* vOTUs (41%), 977 *virulent* vOTUs (22%) and 1613 NA vOTUs (36%). Within the subset of *virulent*

vOTUs, we found that the *virulent*-prediction transferred from MGV genomes to vOTUs showed 90% agreement with *de novo* BACPHLIP predictions on vOTU, which increased our confidence that the vOTUs were not predicted virulent as a result of i.e. incompleteness. In order to evaluate real correlation between bacterial MSPs and associated viruses that might indicate “Piggybacking the Winner” relationship, we applied a bootstrapped randomisation test between bacterial MSP and viral abundance profiles using only associated viruses established with CRISPR spacer-alignment[60]. For each bacterial species (detected in at least >10 samples) we permuted the virus abundance profiles without replacement while keeping the host abundances fixed. We then compared the random-distribution of computed linear regression coefficients from the randomisation test to the original correlation coefficient. If the correlation coefficient was positive and above or below the 95% randomized confidence intervals we assumed PtW dynamic to be generally supported by the correlation.

Viral-bacterial ratios (VBRs) were calculated on the basis of TPMs for vOTUs and associated host-bacteria on MSP species level. In each sample, every MSP has a calculated abundance in TPM based on the top-30 core-marker genes defined by MSPminer. However, there may be multiple detected vOTUs associated with a MSP in a sample. First, we calculated the VBR for every virus based on its abundance and the abundance of the MSP. Second, to get a combined VBR for viruses associated with a MSP we summarized the average VBR in each sample. In order to capture a composite/microbiome aggregate VBR for each sample, which goes across all bacteria and viruses, we calculated the median VBR across MSPs.

Viral phylogenomic tree

We followed the phylogenomic approach described by Low *et al*.[27] to construct a viral genome tree, which was based on 77 selected viral taxonomic informative protein markers that can be used to phylogenetically organize most tailed double-stranded DNA viruses such as members of Caudovirales. The 77 markers were originally established by profiling 2232 complete viral genomes from NCBI Refseq (May 2017) with VOGdb (http://vogdb.org; accessed 1 April 2017). The first 48 VOG markers were selected because they were present at least 10% bacterial or archeal

709 virus genomes and with average copy number ≤1.2 and average length above 100
710 amino acids[27]. The top-10 VOG markers present in at least 10% of viruses within the
711 respective viral families Siphoviridae, Myoviridae and Podoviridae yielded an
712 additional 29 markers to a total of 77 viral protein single-copy markers[27]. Using the
713 77 single-copy protein markers, we annotated 2388/4422 (54%) of vOTUs and
714 37946/65535 (57%) of MGV representative genomes with at least 3 distinct viral
715 markers using HMMER (v.3.1b2, –noali, e-value<1e-5)[61]. A multiple sequence
716 alignment (MSA) was built for each marker, then trimmed using trimAl (v.1.4, -gt
717 0.5)[62] retaining positions with less than 50% gaps. Alignments were then
718 concatenated for each vOTU genome containing at least 3 annotated markers and
719 less than <95% gaps in the concatenated alignment. Subsequently, the viral
720 phylogenetic tree was built using the concatenated protein phylogeny as input to
721 IQtree[63] (v.1.6.8, LG4M model and -bb 1000) and visualised using iToL[64].

722 Whole isolate prophage detection

723 Bacterial isolates (n=68) of a centenarian were purified and sequenced as described
724 by Sato et al[2] (**File S5**). Genomic DNA of 68 isolated strains were assembled using
725 whole-genome shotgun strategy with PacBio Sequel and Illumina MiSeq
726 sequencers. Integrated prophages in bacterial isolates from the centenarian CE91
727 were detected by identifying prophage regions with VIBRANT (v.1.2.1, default
728 settings). Likely prophage sequences were subsequently quality evaluated with
729 CheckV's (v0.8.1) workflow "end_to_end". Prophages were linked to vOTUs
730 assembled from the metagenomic samples by clustering and dereplication at >=95%
731 ANI across >=85% of the shortest target sequence. In total, 63% of the bacterial
732 isolates contained pro-viruses with co-clustered vOTUs. The viral genus of each
733 prophage were determined by species clustering with the initial set of vOTUs already
734 organized into genera; prophages linked to a vOTU by clustering acquired the same
735 genus. Selected prophage genomes were visualized using gggenomes (v.
736 0.9.5.9000, https://github.com/thackl/gggenomes).

737 Age-dependent effects on viral communities

738 To investigate the effect of age and microbial derived community types on the vOTU
739 abundance profiles, we applied permutational multivariate analysis of variance

(PERMANOVA, permutations = 999) implemented in the adonis function from the R-package vegan (v.2.5.7). To derive microbial derived community types, Sato *et al.* clustered each sample based on MSP abundances using a Dirichlet multinomial mixtures algorithm and refined samples into their appropriate community type using a lowest Laplace approximation score [2]. In order to characterize differential viral populations associated with a given host species, we applied generalized linear and mixed modeling functions (centenarians versus young or older controls) using the R-package MaAslin2 (v.1.7.3)[65]. Prior to modeling, log-transformation and Total Sum Scaling (TSS) was applied to abundances summed for host-associated viruses in each sample. Furthermore, in the analysis of differential abundance, we restricted the analysis to host-associated viruses present in at least 15% of samples and accepted significant hits with a Q-value < 0.01 (FDR corrected). Modeling included fixed-effect covariates: cohort information (centenarian, older or young); random effects included participant information to account for more than one sample among a few centenarians.

Microbiome functional and sulfur pathway annotation

The entire gene catalogue was annotated with the PFAM database (v. 34) using HMMER (v.3.1b2, –noali, e-value<1e-2) mapping the domains of each protein sequence of the gene catalogue. Phage encoded auxiliary metabolic genes (AMG) were annotated using VIBRANT (v.1.2.1, default settings) that identifies viral proteins matching KEGG-orthologs (KO) that are mapped to their cognate metabolic pathway in the KEGG database. In addition, the PFAM-ids associated with each AMG enzyme, provided by VIBRANT, could then be used to map and calculate the abundance of the bacterial encoded orthologs of phage AMG for calculating MSP and viral abundance ratios. In order to complete the enzyme map of the full sulfate and methionine metabolic pathways, we manually curated a list of PFAM-ids mapped to each enzymatic step querying the Uniprot database (January 2022, https://www.uniprot.org/).

Statistics and quantifications

The ‘‘n’’ number reported in each figure legend refers to biological replicates for human metagenomic studies i.e. metagenomic samples. Statistical analysis was

performed using R-Studio (v.4.1.2), all statistical details of experiments can be found in the figure legends. Statistical significance was given as * P-value < 0.05; NS (not significant) P-value > 0.05, unless specified otherwise. Regarding boxplots in main figures and supplementary figures. The lower and upper hinges correspond to the first and third quartiles (25th and 75th percentiles). Centre corresponds to the median. The upper and lower whiskers extend from the hinge to the highest and lowest values, respectively, but no further than 1.5 × interquartile range (IQR) from the hinge. IQR is the distance between the first and third quartiles. Data beyond the ends of whiskers are outliers and are plotted individually. This definition is used for all main and supplementary figures displaying a boxplot.

Definition of boxplots

The lower and upper hinges correspond to the first and third quartiles (25th and 75th percentiles). Centre corresponds to the median. The upper and lower whiskers extend from the hinge to the highest and lowest values, respectively, but no further than 1.5 × interquartile range (IQR) from the hinge. IQR is the distance between the first and third quartiles. Data beyond the ends of whiskers are outliers and are plotted individually. This definition applies used for all main and supplementary figures displaying a boxplot.

Data availability

This paper analyzes existing, publicly available data. The raw sequencing files used in this study for Centenarian, Sardinian, Tanzanian 300FG and EDIA microbiome datasets are available at NCBI database under accession numbers PRJNA675598, PRJEB25514, PRJNA686265, and PRJNA707065, respectively. The viral genomes used to establish novel viral biodiversity are available from the Metagenomic Viruses Database (https://portal.nersc.gov/MGV). Additional files including vOTU genomes, phylogenetic tree, VOG marker table and master annotation table for vOTUs can be accessed at Zenodo (https://zenodo.org/record/6579480#.Yo3xHZNBweY)

Code availability

All original code has been deposited at Github and is publicly available as of the date of publication. DOIs are listed in the key resources table. Workflows and supporting code can be accessed at the following repository: https://github.com/RasmussenLab/vCentenarian. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

References

1. Odamaki, T. *et al.* Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. *BMC Microbiol.* **16**, 90 (2016).
2. Sato, Y. *et al.* Novel bile acid biosynthetic pathways are enriched in the microbiome of centenarians. *Nature* **599**, 458–464 (2021).
3. Goronzy, J. J. & Weyand, C. M. Understanding immunosenescence to improve responses to vaccines. *Nat. Immunol.* **14**, 428–436 (2013).
4. Gregory, A. C. *et al.* The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host Microbe* **28**, 724–740.e8 (2020).
5. Adiliaghdam, F. & Jeffrey, K. L. Illuminating the human virome in health and disease. *Genome Med.* **12**, 66 (2020).
6. Sutton, T. D. S. & Hill, C. Gut Bacteriophage: Current Understanding and Challenges. *Front. Endocrinol.* **10**, 784 (2019).
7. Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* **37**, 29–37 (2019).
8. Gulyaeva, A. *et al.* Discovery, diversity, and functional associations of crAss-like phages in human gut metagenomes from four Dutch cohorts. *Cell Rep.* **38**, 110204 (2022).
9. Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host Microbe* **26**, 527–541.e5 (2019).
10. Maier, L. *et al.* Unravelling the collateral damage of antibiotics on gut bacteria. *Nature* **599**, 120–124 (2021).
11. Gogokhia, L. *et al.* Expansion of Bacteriophages Is Linked to Aggravated Intestinal Inflammation and Colitis. *Cell Host Microbe* **25**, 285–299.e8 (2019).

12. Franceschi, C., Garagnani, P., Parini, P., Giuliani, C. & Santoro, A. Inflammaging: a new immune–metabolic viewpoint for age-related diseases. *Nat. Rev. Endocrinol.* **14**, 576–590 (2018).
13. Diard, M. *et al.* Inflammation boosts bacteriophage transfer between Salmonella spp. *Science* **355**, 1211–1215 (2017).
14. Lim, E. S. *et al.* Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat. Med.* **21**, 1228–1234 (2015).
15. Liang, G. *et al.* The stepwise assembly of the neonatal virome is modulated by breastfeeding. *Nature* **581**, 470–474 (2020).
16. Shamash, M. & Maurice, C. F. Phages in the infant gut: a framework for virome development during early life. *ISME J.* (2021) doi:10.1038/s41396-021-01090-x.
17. Khan Mirzaei, M. *et al.* Bacteriophages Isolated from Stunted Children Can Regulate Gut Bacterial Communities in an Age-Specific Manner. *Cell Host Microbe* **27**, 199–212.e5 (2020).
18. Bondy-Denomy, J. *et al.* Prophages mediate defense against phage infection through diverse mechanisms. *ISME J.* **10**, 2854–2866 (2016).
19. Howard-Varona, C. *et al.* Phage-specific metabolic reprogramming of virocells. *ISME J.* **14**, 881–895 (2020).
20. Kieft, K. *et al.* Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. *Nat. Commun.* **12**, 3503 (2021).
21. Kieft, K. *et al.* Virus-associated organosulfur metabolism in human and environmental systems. *Cell Rep.* **36**, 109471 (2021).
22. Mayneris-Perxachs, J. *et al.* Caudovirales bacteriophages are associated with improved executive function and memory in flies, mice, and humans. *Cell Host & Microbe* vol. 30 340–356.e8 (2022).

23. Wu, L. *et al.* A Cross-Sectional Study of Compositional and Functional Profiles of Gut Microbiota in Sardinian Centenarians. *mSystems* **4**, (2019).
24. Nissen, J. N. *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-020-00777-4.
25. Johansen, J. *et al.* Genome binning of viral entities from bulk metagenomics data. *Nat. Commun.* **13**, 965 (2022).
26. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat Microbiol* (2021) doi:10.1038/s41564-021-00928-6.
27. Low, S. J., Džunková, M., Chaumeil, P.-A., Parks, D. H. & Hugenholtz, P. Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales. *Nat Microbiol* **4**, 1306–1315 (2019).
28. Vatanen, T. *et al.* Rarely transmitted maternal species shape the infant gut microbiome through lateral gene transfer. *Submitted* (2022).
29. Redgwell, T. A. *et al.* Prophages in the infant gut are largely induced, and may be functionally relevant to their hosts. (2021).
30. Stražar, M. *et al.* Gut microbiome-mediated metabolism effects on immunity in rural and urban African populations. *Nat. Commun.* **12**, 4845 (2021).
31. Shkoporov, A. N. *et al.* ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nat. Commun.* **9**, 4781 (2018).
32. Thompson, L. R. *et al.* Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proc. Natl. Acad. Sci. U. S. A.* **108**, E757–64 (2011).

33. Murphy, J., Mahony, J., Ainsworth, S., Nauta, A. & van Sinderen, D. Bacteriophage orphan DNA methyltransferases: insights from their bacterial origin, function, and occurrence. *Appl. Environ. Microbiol.* **79**, 7547–7555 (2013).

34. Santoro, A. *et al.* Gut microbiota changes in the extreme decades of human life: a focus on centenarians. *Cell. Mol. Life Sci.* **75**, 129–148 (2018).

35. Guerin, E. *et al.* Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host Microbe* **24**, 653–664.e6 (2018).

36. Sovran, B. *et al.* Age-associated Impairment of the Mucus Barrier Function is Associated with Profound Changes in Microbiota and Immunity. *Sci. Rep.* **9**, 1437 (2019).

37. Biagi, E. *et al.* Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PLoS One* **5**, e10667 (2010).

38. Heidelberg, J. F. *et al.* The genome sequence of the anaerobic, sulfate-reducing bacterium Desulfovibrio vulgaris Hildenborough. *Nat. Biotechnol.* **22**, 554–559 (2004).

39. Buret, A. G., Allain, T., Motta, J.-P. & Wallace, J. L. Effects of Hydrogen Sulfide on the Microbiome: From Toxicity to Therapy. *Antioxid. Redox Signal.* **36**, 211–219 (2022).

40. Stacy, A. *et al.* Infection trains the host for microbiota-enhanced resistance to pathogens. *Cell* **184**, 615–627.e17 (2021).

41. Vatanen, T. *et al.* Transcription shifts in gut bacteria shared between mothers and their infants. *Sci. Rep.* **12**, 1276 (2022).

42. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct

908 de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
909 43. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation
910 site identification. *BMC Bioinformatics* **11**, 119 (2010).
911 44. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for
912 clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682
913 (2010).
914 45. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–
915 Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
916 46. Plaza Oñate, F. *et al.* MSPminer: abundance-based reconstitution of microbial
917 pan-genomes from shotgun metagenomic data. *Bioinformatics* **35**, 1544–1552
918 (2019).
919 47. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*
920 **34**, 3094–3100 (2018).
921 48. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics*
922 **25**, 2078–2079 (2009).
923 49. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and
924 efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359
925 (2019).
926 50. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W.
927 CheckM: assessing the quality of microbial genomes recovered from isolates,
928 single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
929 51. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit
930 to classify genomes with the Genome Taxonomy Database. *Bioinformatics*
931 (2019) doi:10.1093/bioinformatics/btz848.
932 52. Nayfach, S. *et al.* CheckV assesses the quality and completeness of

933 metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
934 53. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery,
935 annotation and curation of microbial viruses, and evaluation of viral community
936 function from genomic sequences. *Microbiome* **8**, 90 (2020).
937 54. Johnson, M. *et al.* NCBI BLAST: a better web interface. *Nucleic Acids Res.* **36**,
938 W5–9 (2008).
939 55. Kirchberger, P. C., Martinez, Z. A. & Ochman, H. Organizing the Global
940 Diversity of Microviruses. *MBio* **13**, e0058822 (2022).
941 56. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using
942 DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
943 57. Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S. A. & Sørensen, S. J.
944 CRISPRCasTyper: Automated Identification, Annotation, and Classification of
945 CRISPR-Cas Loci. *CRISPR J* **3**, 462–469 (2020).
946 58. Bland, C. *et al.* CRISPR recognition tool (CRT): a tool for automatic detection of
947 clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209
948 (2007).
949 59. Hockenberry, A. J. & Wilke, C. O. BACPHLIP: Predicting bacteriophage lifestyle
950 from conserved protein domains. *Cold Spring Harbor Laboratory*
951 2020.05.13.094805 (2020) doi:10.1101/2020.05.13.094805.
952 60. Alrasheed, H., Jin, R. & Weitz, J. S. Caution in inferring viral strategies from
953 abundance correlations in marine metagenomes. *Nat. Commun.* **10**, 501 (2019).
954 61. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**,
955 e1002195 (2011).
956 62. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for
957 automated alignment trimming in large-scale phylogenetic analyses.

Bioinformatics **25**, 1972–1973 (2009).

63. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

64. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–5 (2016).

65. Mallick, H. *et al.* Multivariable association discovery in population-scale meta-omics studies. *PLoS Comput. Biol.* **17**, e1009442 (2021).

Acknowledgements

We thank Elizabeth Heppenheimer for editorial assistance with text and figure preparation.

Funding

J.J. and S.R. were supported by the Novo Nordisk Foundation (grant NNF14CC0001).

Author Contributions

Conceptualization, J.J, D.R.P and R.J.X; Methodology, J.J, D.R.P; Software, J.J; Formal Analysis, J.J, D.R.P; Investigation, J.J; Resources, R.J.X, S.R; Data Curation, J.J, D.R.P, K.H, K.A, Y.A, N.H, T.V, M.K,; Visualization, J.J; Writing - Original draft, J.J; Writing - Review, J.J, D.R.P, R.J.X, S.R, K.H, S.J.S, K.A; Supervision, D.R.P, R.J.X, S.R.

Competing interests

R.J.X. is a co-founder of Celsius Therapeutics and Jnana Therapeutics, a member of the Scientific Advisory Board of Nestle, as well as a member of the Board of Directors at Moonlake Immunotherapeutics.

Supplementary information

Supplementary Figures:

Supplementary Figure 1: Viral family abundances and core-virus structures
Supplementary Figure 2: Viral bacterial interaction inference
Supplementary Figure 3: Inspection of host interaction differences between viruses enriched in centenarians and control groups
Supplementary Figure 4: Viral proteomes and auxiliary metabolic genes
Supplementary Figure 5: Microbiome metabolic potential in Sardinian cohort and Taurine conjugated bile acids levels

Supplementary Files:

File S1: vOTU genus counts and host annotation
File S2: Integrated prophage summary table
File S3: Test-statistics for viral bacterial abundance ratios between age groups
File S4: Test-statistics for dissimilatory sulfate gene abundance between age groups
File S5: CE91 isolate information including taxonomy and genome quality

Extra files (not referenced in main text, uploaded to Zenodo):
File 1: VOG Markers in vOTUs and MGV genomes
File 2: Viral Tree file
File 3: All vOTUs genomes
File 4: Master table annotation of vOTUs
File 5: Centenarian bacterial isolate prophages

Figure titles and legends

Figure 1. Healthy centenarians display a more diverse and rich Virome compared to young and elderly
The three main arcs of analysis (a,b,c) beginning with **a)** depict the workflow for viral and bacterial delineation for three distinct age groups as MSPs+MAGs and vOTUs,

respectively. vOTUs established by clustering were annotated by novelty (based on MGV genomes), host and lifestyle, followed by compositional and phylogenetic analysis to reveal age-dependent differences **b)** Viral bacterial ratios were outlined by age and included two independent reference cohorts **c)** Analysis of auxiliary metabolic genes encoded in vOTUs and proviruses. **d)** A phylogenomic tree of vOTUs based on 1,415 *de novo* assembled and 29,057 MGV vOTUs, created with IQtree and visualised using iToL (https://itol.embl.de/). Predicted host-taxonomic phylum is indicated for each vOTU as either Firmicutes (green), Actinobacteria (orange), Bacteroidetes (yellow), Proteobacteria (blue), Verrucomicrobia (pink) and Desulfobacteriota (purple). Host color indications can be seen in the third ring around the tree. The *de novo* assembled vOTUs across age groups were clustered at the viral species level with the entire MGV genome database to establish vOTU novelty. vOTUs not clustered with an MGV genome are labeled novel vOTU else as MGV indicating that a highly similar viral genome exists, indicated as red and black leaves, respectively and shown in the second ring. Viral genera of interest, G7 and G78, which were expanded with hundreds of novel vOTU, are indicated in the first ring. Novel viruses added to G2 are also marked. The prevalence (0-1) of each vOTU in the microbiome of centenarian, elderly and young is indicated as orange, blue and gray bars (from the 4th ring). Finally, significant vOTU enrichment (Wilcoxon rank sum test, one-sided, FDR corrected) on median TPM is indicated as a black rectangle between Centenarian vs Young, Centenarian vs Elderly and Elderly vs Young. **e)** Principal coordinate analysis based vOTU Bray-Curtis dissimilarity shows separation of centenarian samples from elderly and young. **f-g)** Boxplot of viral richness (the number of detected vOTUs in sample; **f**) and diversity (Shannon diversity; **g**) shows significantly increased richness and diversity in centenarians (Wilcoxon rank sum test, two-sided, $P<0.05$). **h)** The proportion of relative abundance, summed for novel vOTUs and vOTUs similar to a MGV genome, indicates that a larger proportion of read-mapping signal originates from vOTUs in centenarians (pairwise T-test, $P<0.05$). **i)** The proportion of relative abundance in a sample summed by core-virus (prevalence >10% across age groups) and individual viruses (prevalence <10% across age groups) displays a gradual age-dependent depletion of core-virus abundance. Centenarian [n = 176 (153 individuals)], elderly (n = 110), young (n = 44). Abbreviations, MSP: Metagenomic species of MSPminer,

(v)MAG: (viral) Metagenomic assembled genome, vOTU: Viral operational taxonomic unit. TPM: Transcript per million. MGV: Metagenomic Gut Virus (database).

Figure 2. Phage signatures correlate with the unique centenarian bacterial communities

a) Changes in the relative viral abundance of vOTU populations grouped by bacterial host (on species level) between centenarian (CE), elderly, and young subjects (MaAsLin2 analysis). The heatmap shows differentially abundant vOTU populations, with Q-value < 0.01 indicated by an asterix (*). Only vOTU populations, such as those associated with *Clostridium scindens* or *Faecalibacterium prausnitzii*, that were found enriched or depleted in either CE vs Elderly or CE vs Young were included in the final heatmap. Color scale represents the beta-coefficient from the general linear model in MaAslin2 and indicates the degree of enrichment (red) or depletion (blue). vOTU are sorted by their prevalence (horizontal bar panels) where viruses associated with *Clostridium scindens* or *Akkermansia muciniphila* are detected in >40% of CE samples and <10% in young adult samples. **(b-c)** *Clostridium scindens* viruses also display higher relative abundance in centenarians while *Faecalibacterium prausnitzii* are relatively depleted despite their prevalence in centenarians. **d)** Several enriched temperate vOTUs displayed PtW abundance (log10 scale) trends with MSP host abundance such as the abundance of *Enterocloster boltea* vOTUs showing a strong correlation of 0.51 (P-value: 4.82e-66), *Clostridium scindens* (P-value: 4.49e-08) and *Parabacteroides distasonis* (P-value: 6.00e-11).

Figure 3. Lysogeny in the phageome dominates from young age to centenarian in the healthy microbiome

a) Viral-bacterial ratios (VBRs) distributions estimated using TPM of MSPs and temperate vOTUs, displayed as boxplots, shows age-driven separation of microbiomes. Triple asterix (***) indicates significant difference of pairwise VBR distributions (box plots) in the specified comparison at P < 0.001 (Wilcoxon rank sum test, two sided). **b)** The average VBR calculated on MSP level, based on all associated temperate vOTUs to a given MSP, are illustrated for all centenarian samples. The VBR distribution for some MSPs such as *Clostridium scindens* and *Alistipes putredinis* were close to ~0.5 suggesting the presence of viruses in a

lysogenic state, while the VBR distribution of *Alistipes shahii* and *Barnesiella intestinihomis* was primarily above 1, indicating lytic activity. **c)** The VBR distribution of viruses known to be intrinsically virulent, such as crass-like viruses infecting *A. shahii* and *P. merdae*, could be separated from other viruses with a predetermined lifestyle. **d)** Complex virus-host interactions were found on population level for viruses associated with *Faecalibacterium prausnitzii* with viruses enriched in centenarians (CE_enriched) displaying negative abundance (TPM) correlations (Pearson correlation test) with different *F. prausnitzii* MSPs, while non-enriched viruses displayed a positive correlation suggesting lysogeny interaction. Centenarian [n = 176 (153 individuals)], elderly (n = 110), young (n = 44) and infants [n = 688 (142 individuals)] subjects. Each dot represents one sample.

Figure 4. The centenarian virome supports a greater host-metabolic conversion of microbial sulfide.
a) The average total abundance (TPM) of an auxiliary metabolic gene (AMG) was calculated by summing the abundance of all genes encoded by bacterial MSPs (first panel) and viruses (second panel) then an average was calculated across samples. Using the mean MSP and viral abundance of a given AMG, the mean viral proportion of abundance seen in the third panel could be derived (Viral TPM / Viral + MSP TPM). For the AMGs dcm, cysH and metK, >10% of the total abundance comes from genes encoded in viruses AMGs in centenarians. All error bars are derived using standard error mean **b)** Viruses encoding AMGs such as dcm and cysH are highly prevalent in every age group, while those encoding, for instance, metK or iscS, were more prevalent in centenarians and elderly relative to young. **c)** Based on the bacterial pangenomes derived with MSPminer it was shown that some AMGs like cysH and dcm are encoded to a higher degree in the flexible part of pangenomes. **d)** The total viral and MSP abundance is correlated in some AMGs suggesting these to be encoded by primarily lysogenic viruses while negative correlations indicate signal from lytic viral encoders. Centenarian [n = 176 (153 individuals)], elderly (n = 110) and young (n = 44) subjects.

Figure 5. The centenarian gut microbiome configuration is primed for effective utilization of taurine, sulfate and methionine.
a) Box plots of the total microbiome abundance (log10 scale) of genes encoding

enzymes in the sulfate to sulfide conversion pathway and taurine to sulfide, shown for each age group. **b)** Similarly, the abundance (log10 scale) of genes encoding enzymes relevant for methionine to homocysteine conversion. Both (a) and (b) collectively show higher abundance of the relevant enzymes in centenarians indicated with an asterix (*) if significant (FDR corrected T-test, P<0.05) else not-significant (NS). **c)** Metabolic diagram of genes involved in the conversion of sulfate to sulfide, taurine to sulfide and methionine to homocysteine. Age specific gene enrichment, if any, is indicated for each enzymatic step. **d)** Genomic visualization of integrated prophages identified in bacterial isolates of a centenarian (CE91) including a virus from *Bacteroides dorei, Clostridium scindens* and a *Lachnospiracae.* Coding gene segments of prophages are colored according to annotated function including domain of unknown function (orange), hypothetical proteins (gray), typical structural phage proteins and enzymes (green), proteins corresponding to enzymes found in the pathway of methionine to homocysteine (blue) and sulfate to sulfide conversion (yellow). Each dot in box plots (A and B) represents one sample, for the comparison and statistical tests the number of samples from Japanese Centenarian [n = 176 (153 individuals)], elderly (n = 110) and young (n = 44) subjects.

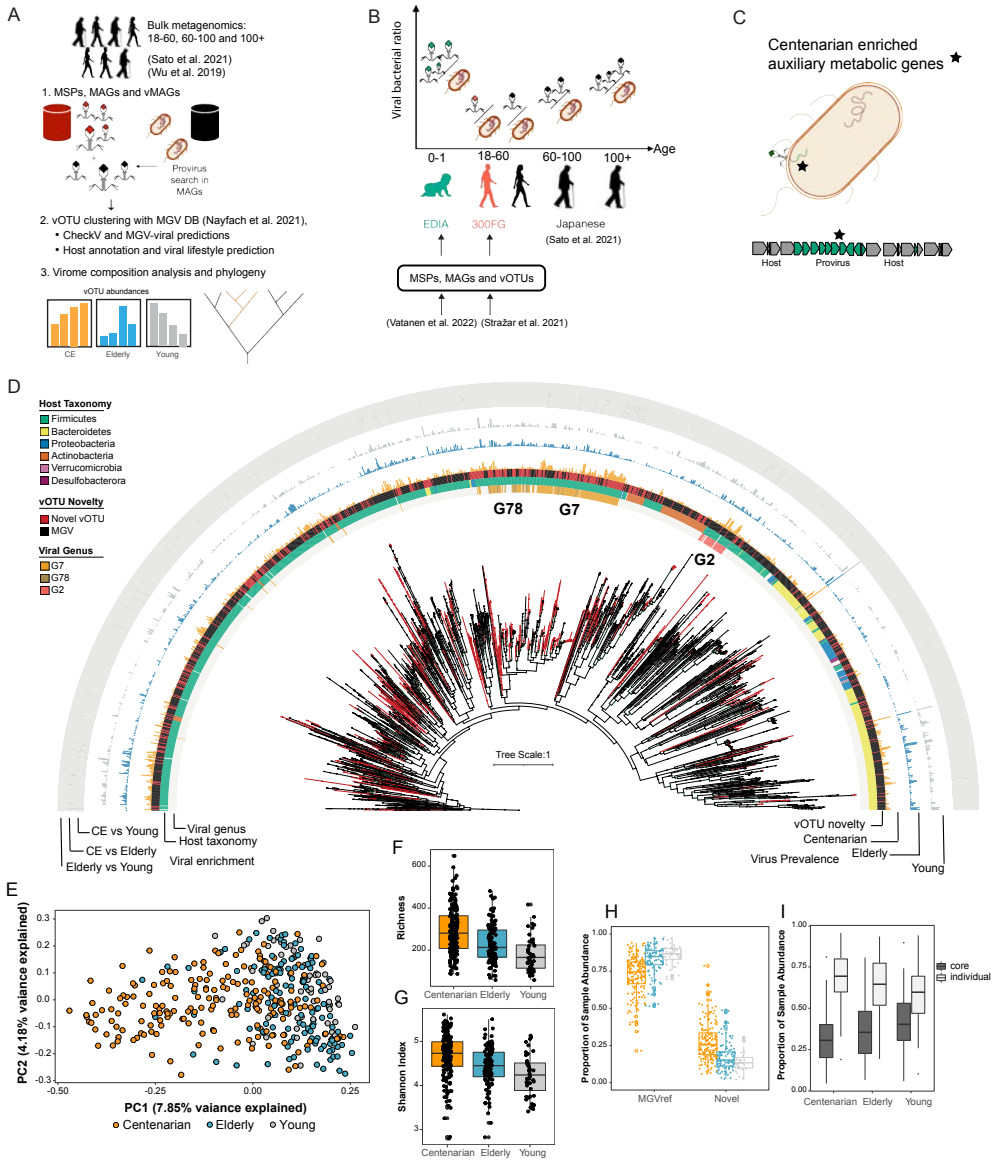

A
Bulk metagenomics:
18-60, 60-100 and 100+
(Sato et al. 2021)
(Wu et al. 2019)
1. MSPs, MAGs and vMAGs
Provirus search in MAGs
2. vOTU clustering with MGV DB (Nayfach et al. 2021),
• CheckV and MGV-viral predictions
• Host annotation and viral lifestyle prediction
3. Virome composition analysis and phylogeny
vOTU abundances
CE
Elderly
Young
B
Viral bacterial ratio
Age
0-1
18-60
60-100
100+
EDIA
300FG
Japanese
(Sato et al. 2021)
MSPs, MAGs and vOTUs
(Vatanen et al. 2022) (Stražar et al. 2021)
C
Centenarian enriched
auxiliary metabolic genes
Host
Provirus
Host
D
Host Taxonomy
Firmicutes
Bacteroidetes
Proteobacteria
Actinobacteria
Verrucomicrobia
Desulfobacterora
vOTU Novelty
Novel vOTU
MGV
Viral Genus
G7
G78
G2
G78
G7
G2
Tree Scale:1
CE vs Young
CE vs Elderly
Elderly vs Young
Viral genus
Host taxonomy
Viral enrichment
vOTU novelty
Centenarian
Elderly
Young
Virus Prevalence
E
PC2 (4.18% valance explained)
PC1 (7.85% vaiance explained)
Centenarian
Elderly
Young
F
Richness
Centenarian Elderly Young
G
Shannon Index
Centenarian Elderly Young
H
Proportion of Sample Abundance
MGVref
Novel
I
Proportion of Sample Abundance
Centenarian Elderly Young
core
individual

Figure 2

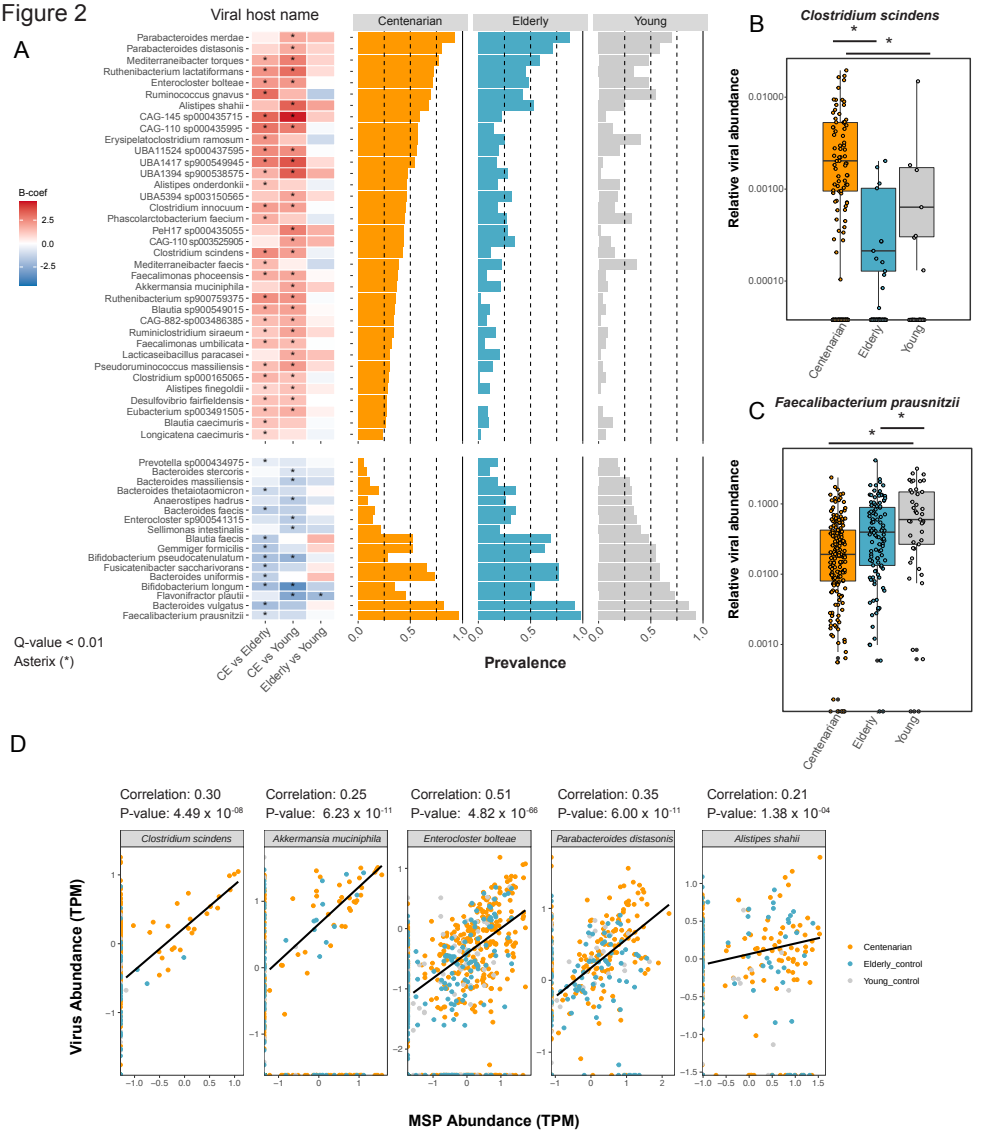

A
Viral host name
Centenarian
Elderly
Young
B-coef
2.5
0.0
-2.5
Parabacteroides merdae
Parabacteroides distasonis
Mediterraneibacter torques
Ruthenibacterium lactatiformans
Enterocloster bolteae
Ruminococcus gnavus
Alistipes shahii
CAG-145 sp000435715
CAG-110 sp000435995
Erysipelatoclostridium ramosum
UBA11524 sp000437595
UBA1417 sp900549945
UBA1394 sp900538575
Alistipes onderdonkii
UBA5394 sp003150565
Clostridium innocuum
Phascolarctobacterium faecium
PeH17 sp000435055
CAG-110 sp003525905
Clostridium scindens
Mediterraneibacter faecis
Faecalimonas phoceensis
Akkermansia muciniphila
Ruthenibacterium sp900759375
Blautia sp900549015
CAG-882-sp003486385
Ruminiclostridium siraeum
Faecalimonas umbilicata
Lacticaseibacillus paracasei
Pseudoruminococcus massiliensis
Clostridium sp000165065
Alistipes finegoldii
Desulfovibrio fairfieldensis
Eubacterium sp003491505
Blautia caecimuris
Longicatena caecimuris
Prevotella sp000434975
Bacteroides stercoris
Bacteroides massiliensis
Bacteroides thetaiotaomicron
Anaerostipes hadrus
Bacteroides faecis
Enterocloster sp900541315
Sellimonas intestinalis
Blautia faecis
Gemmiger formicilis
Bifidobacterium pseudocatenulatum
Fusicatenibacter saccharivorans
Bacteroides uniformis
Bifidobacterium longum
Flavonifractor plautii
Bacteroides vulgatus
Faecalibacterium prausnitzii
CE vs Elderly
CE vs Young
Elderly vs Young
0.0
0.5
1.0
Prevalence
Q-value < 0.01
Asterix (*)
B
Clostridium scindens
Relative viral abundance
0.01000
0.00100
0.00010
Centenarian
Elderly
Young
C
Faecalibacterium prausnitzii
Relative viral abundance
0.1000
0.0100
0.0010
Centenarian
Elderly
Young
D
Correlation: 0.30
P-value: 4.49 x 10-08
Correlation: 0.25
P-value: 6.23 x 10-11
Correlation: 0.51
P-value: 4.82 x 10-66
Correlation: 0.35
P-value: 6.00 x 10-11
Correlation: 0.21
P-value: 1.38 x 10-04
Clostridium scindens
Akkermansia muciniphila
Enterocloster bolteae
Parabacteroides distasonis
Alistipes shahii
Virus Abundance (TPM)
MSP Abundance (TPM)
Centenarian
Elderly_control
Young_control

Figure 3

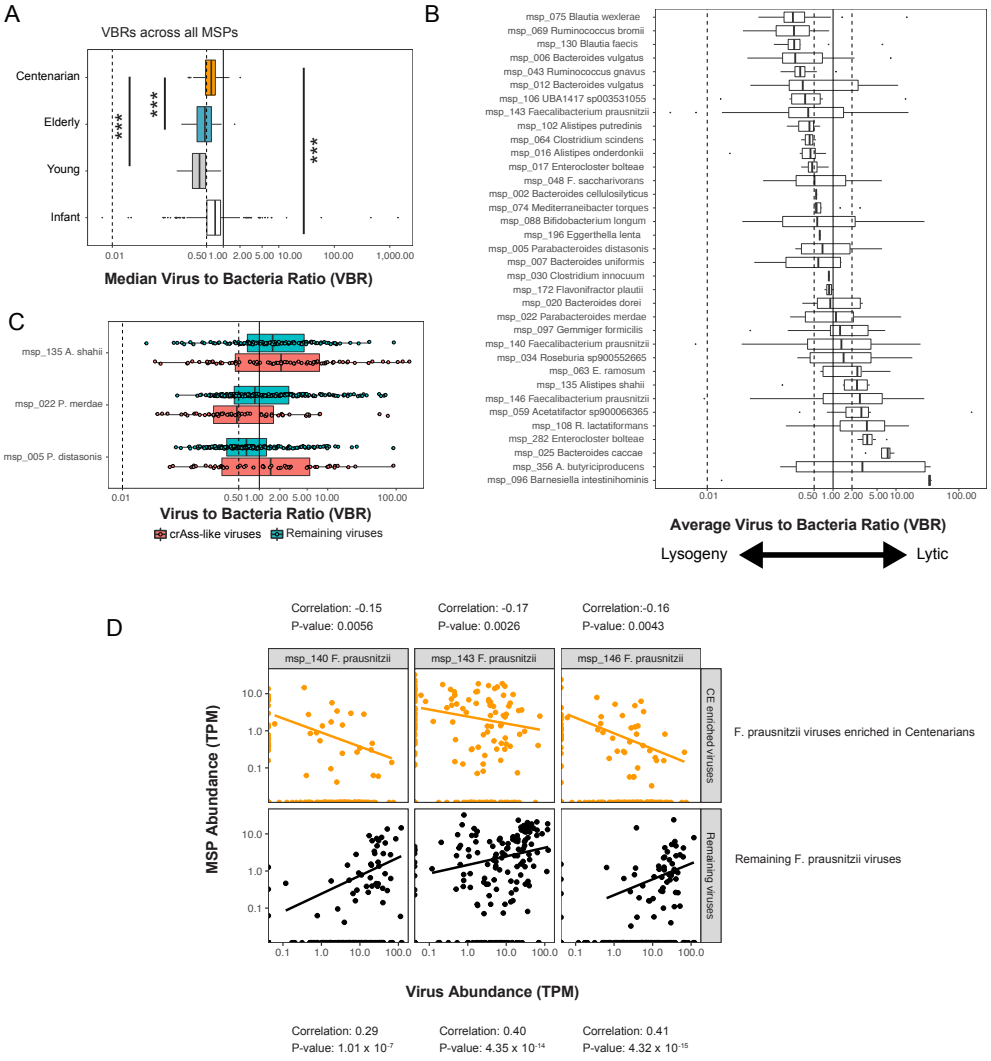

A
VBRs across all MSPs
Centenarian
Elderly
Young
Infant
***
***
***
0.01
0.50 1.00 2.00 5.00 10.00
100.00
1,000.00
Median Virus to Bacteria Ratio (VBR)
B
msp_075 Blautia wexlerae
msp_069 Ruminococcus bromii
msp_130 Blautia faecis
msp_006 Bacteroides vulgatus
msp_043 Ruminococcus gnavus
msp_012 Bacteroides vulgatus
msp_106 UBA1417 sp003531055
msp_143 Faecalibacterium prausnitzii
msp_102 Alistipes putredinis
msp_064 Clostridium scindens
msp_016 Alistipes onderdonkii
msp_017 Enterocloster bolteae
msp_048 F. saccharivorans
msp_002 Bacteroides cellulosilyticus
msp_074 Mediterraneibacter torques
msp_088 Bifidobacterium longum
msp_196 Eggerthella lenta
msp_005 Parabacteroides distasonis
msp_007 Bacteroides uniformis
msp_030 Clostridium innocuum
msp_172 Flavonifractor plautii
msp_020 Bacteroides dorei
msp_022 Parabacteroides merdae
msp_097 Gemmiger formicilis
msp_140 Faecalibacterium prausnitzii
msp_034 Roseburia sp900552665
msp_063 E. ramosum
msp_135 Alistipes shahii
msp_146 Faecalibacterium prausnitzii
msp_059 Acetatifactor sp900066365
msp_108 R. lactatiformans
msp_282 Enterocloster bolteae
msp_025 Bacteroides caccae
msp_356 A. butyriciproducens
msp_096 Barnesiella intestinihominis
0.01
0.50 1.00 2.00 5.00 10.00
100.00
Average Virus to Bacteria Ratio (VBR)
Lysogeny
Lytic
C
msp_135 A. shahii
msp_022 P. merdae
msp_005 P. distasonis
0.01
0.50 1.00 2.00 5.00 10.00
100.00
Virus to Bacteria Ratio (VBR)
crAss-like viruses
Remaining viruses
D
Correlation: -0.15
P-value: 0.0056
Correlation: -0.17
P-value: 0.0026
Correlation:-0.16
P-value: 0.0043
msp_140 F. prausnitzii
msp_143 F. prausnitzii
msp_146 F. prausnitzii
CE enriched viruses
Remaining viruses
MSP Abundance (TPM)
10.0
1.0
0.1
F. prausnitzii viruses enriched in Centenarians
Remaining F. prausnitzii viruses
0.1
1.0
10.0
100.0
Virus Abundance (TPM)
Correlation: 0.29
P-value: 1.01 x 10-7
Correlation: 0.40
P-value: 4.35 x 10-14
Correlation: 0.41
P-value: 4.32 x 10-15

Figure 4

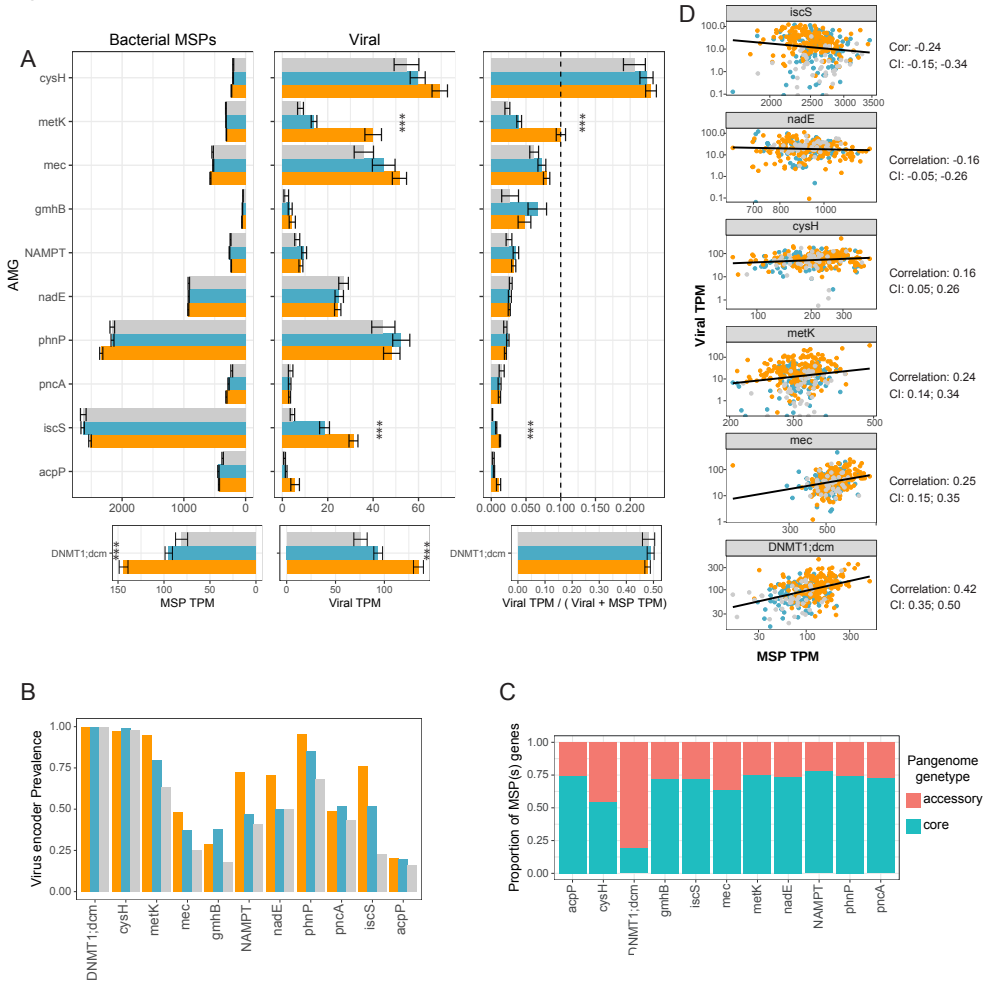
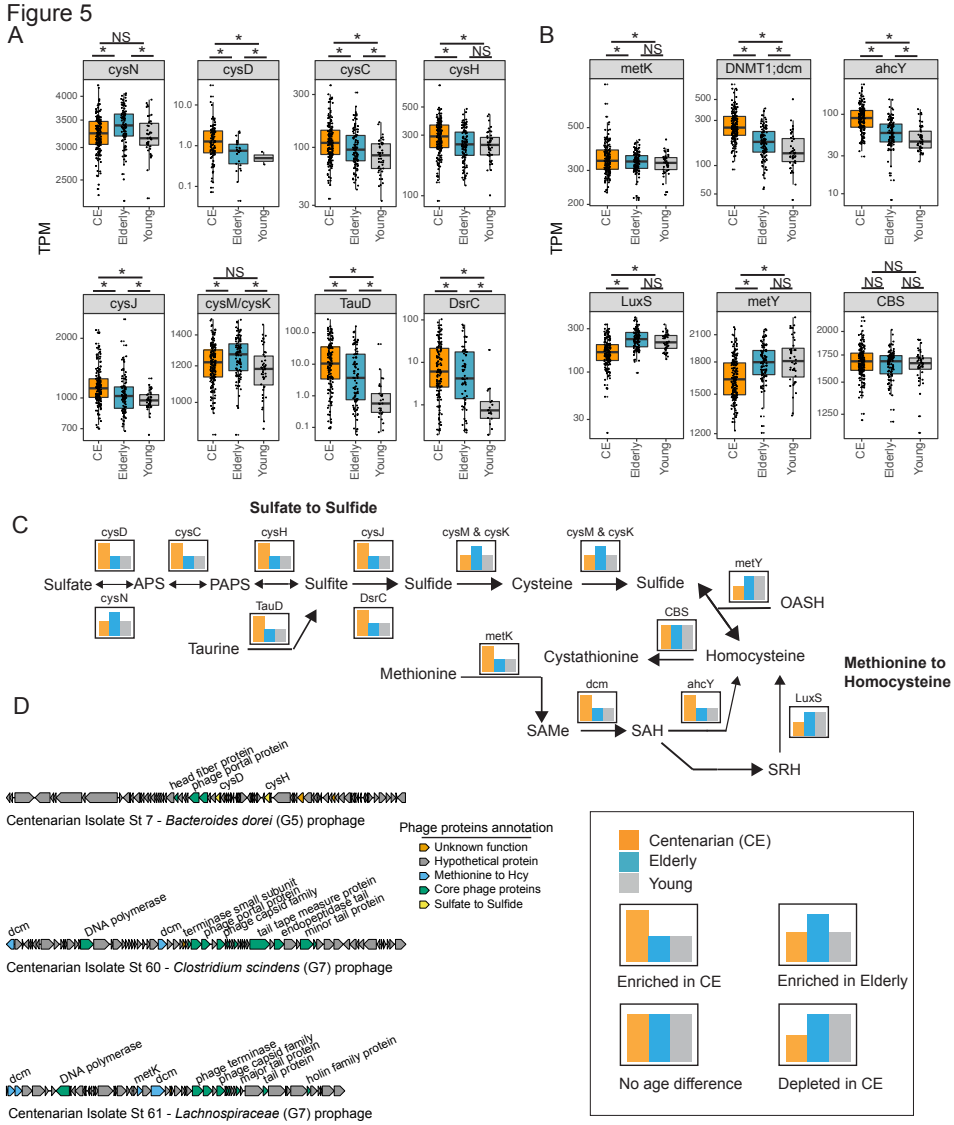

A
Bacterial MSPs
Viral
AMG
cysH
metK
mec
gmhB
NAMPT
nadE
phnP
pncA
iscS
acpP
DNMT1;dcm
MSP TPM
Viral TPM
Viral TPM / (Viral + MSP TPM)
B
Virus encoder Prevalence
C
Proportion of MSP(s) genes
Pangenome genetype
accessory
core
D
Viral TPM
MSP TPM
Cor: -0.24
CI: -0.15; -0.34
Correlation: -0.16
CI: -0.05; -0.26
Correlation: 0.16
CI: 0.05; 0.26
Correlation: 0.24
CI: 0.14; 0.34
Correlation: 0.25
CI: 0.15; 0.35
Correlation: 0.42
CI: 0.35; 0.50

Figure 5


A
B
TPM
cysN
cysD
cysC
cysH
cysJ
cysM/cysK
TauD
DsrC
metK
DNMT1;dcm
ahcY
LuxS
metY
CBS
CE
Elderly
Young
NS
C
Sulfate to Sulfide
Sulfate
APS
PAPS
Sulfite
Sulfide
Cysteine
Taurine
Methionine
Cystathionine
Homocysteine
SAMe
SAH
SRH
OASH
Methionine to Homocysteine
D
Centenarian Isolate St 7 - Bacteroides dorei (G5) prophage
Centenarian Isolate St 60 - Clostridium scindens (G7) prophage
Centenarian Isolate St 61 - Lachnospiraceae (G7) prophage
Phage proteins annotation
Unknown function
Hypothetical protein
Methionine to Hcy
Core phage proteins
Sulfate to Sulfide
Centenarian (CE)
Elderly
Young
Enriched in CE
Enriched in Elderly
No age difference
Depleted in CE

9 Appendix

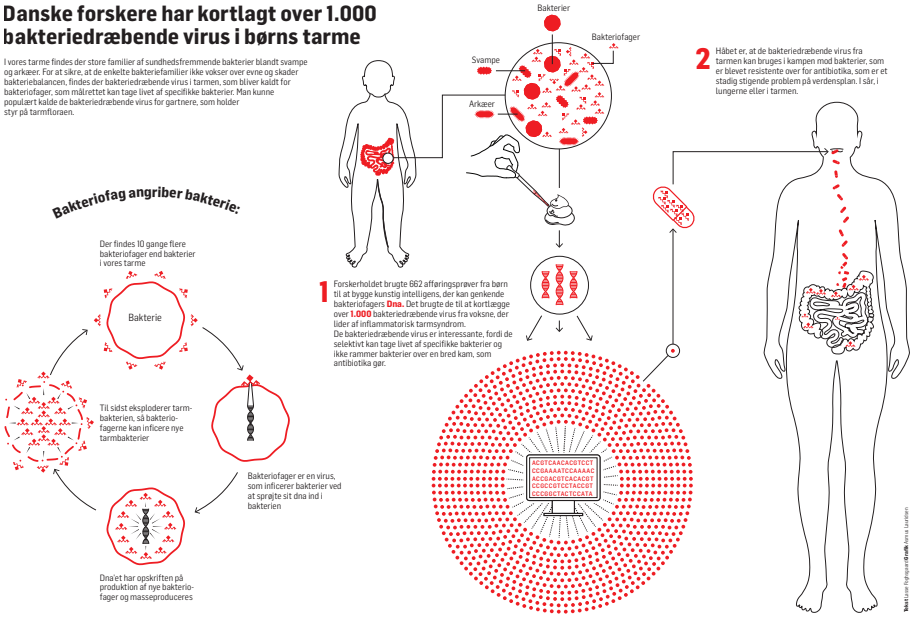
9.1 Popular science article in Politiken

POLITIKEN · Søndag 1. maj 2022

Siden er redigeret/layoutet af:
Dorrit Saietz/Tomas Østergren

Danske forskere har kortlagt over 1.000 bakteriedræbende virus i børns tarme

I vores tarme findes der store familier af sundhedsfremmende bakterier blandt svampe og arkæer. For at sikre, at de enkelte bakteriefamilier ikke vokser over evne og skader bakteriebalancen, findes der bakteriedræbende virus i tarmen, som bliver kaldt for bakteriofager, som målrettet kan tage livet af specifikke bakterier. Man kunne populært kalde de bakteriedræbende virus for gartnere, som holder styr på tarmfloraen.



Din tarm kan bære på det næste supervåben mod sejlivede bakterier

Et dansk forskerhold har fundet over 1.000 bakteriedræbende virus i vores tarme, som holder styr på bakteriebalancen, der er vigtig for vores sundhed. Håbet er, at den nye viden kan bruges i kampen mod antibiotika-resistente bakterier.

LASSE FOGHSGAARD

Det er nok de færreste, der tænker over, at kuren mod antibiotika-resistente bakterier måske gemmer sig i den brune afføring, som man hver dag skyller ud af toilettet.

Men det gør de to forskere Joachim Johansen og Simon Rasmussen fra Københavns Universitet, som sammen med deres forskerkollegaer har offentliggjort et opsigtsvækkende resultat i tidsskriftet Nature Communications, hvor de har fundet og kortlagt over 1.000 bakteriedræbende virus, som sørger for at opretholde en sund bakteriebalance i vores tarme.

De bakteriedræbende virus, som man også kalder for bakteriofager, lever i en tæt symbiose med tarmbakterierne og kan for eksempel dræbe specifikke tarmbakterier, så de holdes i et passende antal og dermed ikke udkonkurrerer andre arter af gode tarmbakterier, som også er vigtige for vores sundhed.

Man kunne populært kalde bakteriofagerne for tarmens gartnere, som sørger for, at græsset ikke vokser for højt, hvis tarmbakterierne var en græsplæne.

»Vi har fundet over 1.000 bakteriofager, som er med til at opretholde en sund bakteriebalance i vores tarme, og de kan endda tage livet af fremmede og sygdomsfremkaldende bakterier. Vi er med vores forskning blevet klogere på, hvad det er for nogle specifikke bakteriofager, som findes i vores tarme, som i samspil med tarmbakterierne har stor indflydelse på vores krop som for eksempel immunforsvaret, hormonproduktionen, fordøjelsen, energi, humør og opfattelse af smerte, siger lektor Simon Rasmussen.

Har studeret børns afføring

Forskerne er kommet frem til resultaterne ved i første omgang at studere dna fra afføringsprøver fra 662 danske børn, der var cirka 1 år gamle ved prøvetagningen, og som indgår i det danske projekt COPSAC, hvor man undersøger, om børnenes økosystem af tarmbakterier spiller en rolle for udvikling af allergi og astma senere i livet.

Ved at systematisere data ved hjælp af kunstig intelligens er det lykkedes forskerne at identificere de mere end 1.000 bakteriofager, som vi alle bærer rundt i vores tarme og holder vores tarmflora sund.

»Vi har nu med vores nye viden udviklet en teknik, hvor vi meget hurtigt fra en afføringsprøve kan analysere, hvad det er for nogle bakteriofager, den enkelte bærer i sine tarme. Vi har blandt andet testet, at det virker på patienter, som lider af kronisk tarmbetændelse, hvor vi har fundet enormt mange af de bakteriofager, som er tæt knyttet til de tarmbakterier, som man også ser en stor overrepræsentation af i forbindelse med sygdommens, siger ph.d. Joachim Johansen fra Københavns Universitet, som er førsteforfatter på det aktuelle studium.

Det danske forskerhold er ikke alene om kaste lys over, hvad der findes af bakteriofager i vores tarme. For eksempel offentliggjorde et engelsk forskerhold for et par måneder siden et katalog over 140.000 forskellige bakteriofager, der er kortlagt fra mere end 28.000 afføringsprøver fra mennesker i 28 lande verden over.

Man skønner, at menneskets tarme indeholder mindst 10 gange så mange bakteriofager sammenlignet med bakterier, fortæller professor Oluf Borbye Pedersen fra Novo Nordisk Fondens Metabolisme-center på Københavns Universitet, som ikke selv har deltaget i den aktuelle undersøgelse.

»Mine danske kolleger har meldt sig ind i jagten på tarmens bakteriofager og har udviklet og afprøvet nye metoder, der gør det muligt at analysere det komplekse samspil, der er mellem bestemte bakteriofager og bestemte bakteriestammer i tarmens økosystem. Der er allerede nu forskning, der peger på forstyrrelser i det samspil ved for eksempel kronisk tarmbetændelse og anoreksi. En dybere forståelse af disse mulige sygdomsmekanismer kan få betydning for den fremtidige behandling af flere kroniske lidelser, siger Oluf Borbye Pedersen.

Våben mod antibiotika-resistens

Der er en særlig god grund til at interessere sig for de potentielt bakteriedræbende virus i vores tarme. I dag dør cirka 700.000 mennesker på verdensplan på grund af antibiotika-resistens. Verdenssundhedsorganisationen WHO anslår, at det tal vil stige til 10 millioner mennesker årligt i 2050. Det er flere, end der i dag dør af kræft.

De danske forskere håber, at kortlægningen af de mange bakteriofager i vores tarme vil kunne hjælpe i kampen mod bakterier, når klassisk medicin i form af penicillin og antibiotika må give op og ikke kan dræbe de sejlivede mikroorganismer.

»Bakteriofager er fra naturens side designet til at inficere bakterier og kan derfor potentielt være det ultimative våben i kampen mod de sygdomsfremkaldende bakterier, som antibiotika ikke kan hamle op med. Med kortlægningen af tusindvis af bakteriofager fra vores tarme er vores håb, at nogle af dem vil kunne bruges til at nedkæmpe antibiotika-resistente bakterier uden at tage livet af tarmens gode bakterier, som er en uhelbig bivirkning ved traditionelle antibiotikas, siger Simon Rasmussen.

Tanken om at behandle kroniske infektioner med bakteriofager er langtfra ny. Blot to år efter, at den engelske bakteriolog Frederick Twort opdagede eksistensen af bakteriofager med sit mikroskop i 1915, isolerede den fransk-canadiske mikrobiolog Félix D'Herelle bakteriofager fra afføringen hos patienter, der var ved at komme sig efter dysenteri. I 1931 udgav Félix D'Herelle en videnskabelig artikel, hvor han beskrev, hvordan bakteriofager kunne bruges medicinsk til at behandle akutte infektioner vmed stor succes.

»Mens der har været en lang tradition for at behandle kroniske infektioner med bakteriofager i lande som Georgien, Polen og Rusland, så gik behandlingsprincipper i glemmebogen i de vesteuropæiske lande fra 1940'erne og næsten frem til i dag. Det er først nu med Belgien i spidsen, at man i Vesteuropa for alvor er begyndt at eksperimentere med behandling med bakteriofager i stor stils, siger Joachim Johansen.

Forskere fra Gentofte og Herlev Hospital samt Københavns Universitet med professor Dennis Sandris Nielsen i spidsen, som også er medforfattere på den aktuelle videnskabelige artikel, har også eksperimenteret med at give transplantationer med bakteriofager fra raske mennesker ind i tarmen på patienter med type-2 diabetes.

»De foreløbige resultater tyder på, at det kan hjælpe patienter med type 2-diabetes og virker lige så godt, som hvis man lavede en fæces-transplantation med den fulde tarmflora. Det giver et forsigtigt håb om, at bakteriofager måske kan bruges i behandlingen af diabetes i fremtiden, siger Simon Rasmussen.

I dag behandler man den livstruende diarré-sygdom forårsaget af bakterien *Clostridium difficile* med afføring fra en rask donor.

»Behandlingen med afføring vil sandsynligvis om nogle år blive erstattet med specifikke bakteriofager, der som et missil kan dræbe den sygdomsfremkaldende bakteries, spår Oluf Borbye Pedersen.

> **Behandlingen med afføring vil sandsynligvis om nogle år blive erstattet med specifikke bakteriofager, der som et missil kan dræbe den sygdoms-fremkaldende bakteries**
>
> Oluf Borbye Pedersen
> professor, Københavns Universitet

Jagt på oldinges hemmelighed

På offentligt tilgængelige databaser på internettet ligger der data på afføringsprøver taget fra 100.000 mennesker. De danske forskere glæder sig nu til at gå på jagt i databaserne med deres nye metode for at finde ud af, hvad det er for nogle bakteriofager, der gemmer sig i afføringsprøverne.

Joachim Johansen er allerede i gang med et projekt i samarbejde med amerikanske og japanske forskere, hvor de undersøger, hvad det er for nogle bakteriofager, som sunde og raske japanere over 100 år bærer i deres tarme, som indtil nu har været et mysterium.

»Vi har brugt vores nye metode til at undersøge, hvordan økosystemet i tarmen ser ud hos de ældgamle japanere over 100 år, og vi er begyndt at se nogle klare mønstre på, hvad det er for nogle tarmbakterier, der i samspil med bakteriofager holder japanerne kernesunde i en høj alder, siger Joachim Johansen.

Grøntsager eller piller

Det drejer sig for eksempel om den store familie af tarmbakterier Clostridia, som i samspil med bakteriofager blandt andet producerer galdesyre, som dels er med til at etablere et anti-inflammatorisk miljø i tarmen og dels har en antibiotisk effekt mod sygdomsfremkaldende bakterier hos de japanske oldinge, forklarer forskerne.

»Det vækker håb om, at hvis man etablerer et økosystem i tarmene, som minder om det, som findes hos de japanske hundredårige, kunne man sikre en sund aldring og et langt liv, siger Simon Rasmussen.

Man ved, at en kost rig på grøntsager kan sikre en sund bakteriebalance i tarmen, men man kunne ifølge forskerne også forestille sig en fremtid, hvor man på pilleform forsøger at kolonisere vores tarme med de bakterier og de bakteriofager, som har sikret de gamle japanere en god tarmsundhed.

»De gamle japanere bærer rundt på en utrolig biologisk fabrik i deres tarme, som producerer kemiske stoffer, der virker beskyttende mod sygdomsfremkaldende bakterier. Det ville være super, hvis vi allesammen kunne få nogle af de bakterier og deres bakteriofager. Det ville være godt for fællesundhedens, siger Joachim Johansen.

lasse.foghsgaard@pol.dk

References

[1] E V Koonin, A R Mushegian, and K E Rudd. “Sequencing and analysis of bacterial genomes”. en. In: *Curr. Biol.* 6.4 (Apr. 1996), pp. 404–416.

[2] E S Lander et al. “Initial sequencing and analysis of the human genome”. en. In: *Nature* 409.6822 (Feb. 2001), pp. 860–921.

[3] Nick Lane. “The unseen world: reflections on Leeuwenhoek (1677) ’Concerning little animals’”. en. In: *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370.1666 (Apr. 2015), p. 20140344.

[4] J T Staley and A Konopka. “Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats”. en. In: *Annu. Rev. Microbiol.* 39 (1985), pp. 321–346.

[5] Pace Nr. “Analyzing natural microbial populations by rRNA sequences”. In: *ASM News* 51 (1985), pp. 4–12.

[6] J Craig Venter et al. “Environmental genome shotgun sequencing of the Sargasso Sea”. en. In: *Science* 304.5667 (Apr. 2004), pp. 66–74.

[7] Jason Lloyd-Price, Galeb Abu-Ali, and Curtis Huttenhower. “The healthy human microbiome”. en. In: *Genome Med.* 8.1 (Apr. 2016), p. 51.

[8] Mads Albertsen et al. “Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes”. en. In: *Nat. Biotechnol.* 31.6 (June 2013), pp. 533–538.

[9] H Bjørn Nielsen et al. “Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes”. In: *Nat. Biotechnol.* 32.8 (Aug. 2014), pp. 822–828.

[10] Dongwan D Kang et al. “MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies”. en. In: *PeerJ* 7 (July 2019), e7359.

[11] Donovan H Parks et al. “CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes”. en. In: *Genome Res.* 25.7 (July 2015), pp. 1043–1055.

[12] Alexandre Almeida et al. “A new genomic blueprint of the human gut microbiota”. In: *Nature* 568.7753 (Apr. 2019), pp. 499–504.

[13] Simon Roux et al. “Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses”. en. In: *Nature* 537.7622 (Sept. 2016), pp. 689–693.

[14] David Paez-Espino et al. “Uncovering Earth’s virome”. en. In: *Nature* 536.7617 (Aug. 2016), pp. 425–430.

[15] Simon Roux et al. “Minimum Information about an Uncultivated Virus Genome (MIUViG)”. en. In: *Nat. Biotechnol.* 37.1 (Jan. 2019), pp. 29–37.

[16] Mya Breitbart et al. “Phage puppet masters of the marine microbial realm”. en. In: *Nat Microbiol* 3.7 (July 2018), pp. 754–766.

[17] Jason M Norman et al. “Disease-specific alterations in the enteric virome in inflammatory bowel disease”. en. In: *Cell* 160.3 (Jan. 2015), pp. 447–460.

[18] Adam G Clooney et al. “Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease”. en. In: *Cell Host Microbe* 26.6 (Dec. 2019), 764–778.e5.

[19] Thomas D S Sutton and Colin Hill. “Gut Bacteriophage: Current Understanding and Challenges”. en. In: *Front. Endocrinol.* 10 (Nov. 2019), p. 784.

[20] Simon Roux et al. *Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth’s biomes*. 2019.

[21] Moı.ra B Dion, Frank Oechslin, and Sylvain Moineau. “Phage diversity, genomics and phylogeny”. en. In: *Nat. Rev. Microbiol.* 18.3 (Mar. 2020), pp. 125–138.

[22] Andrey N Shkoporov et al. “ΦCrAss001, a member of the most abundant bacteriophage family in the human gut, infects Bacteroides”. en. June 2018.

[23] Anastasia Gulyaeva et al. “Discovery, diversity, and functional associations of crAss-like phages in human gut metagenomes from four Dutch cohorts”. en. In: *Cell Rep.* 38.2 (Jan. 2022), p. 110204.

[24] Andrey N Shkoporov et al. “The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific”. en. In: *Cell Host Microbe* 26.4 (Oct. 2019), 527–541.e5.

[25] Anne Chevallereau et al. “Interactions between bacterial and phage communities in natural environments”. en. In: *Nat. Rev. Microbiol.* 20.1 (Jan. 2022), pp. 49–62.

[26] Jeremy J Barr et al. “Subdiffusive motion of bacteriophage in mucosal surfaces increases the frequency of bacterial encounters”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 112.44 (Nov. 2015), pp. 13675–13680.

[27] Cynthia B Silveira and Forest L Rohwer. *Piggyback-the-Winner in host-associated microbial communities*. 2016.

[28] Andrey N Shkoporov and Colin Hill. “Bacteriophages of the Human Gut: The “Known Unknown” of the Microbiome”. In: *Cell Host Microbe* 25.2 (Feb. 2019), pp. 195–209.

[29] Amos B Oppenheim et al. “Switches in bacteriophage lambda development”. en. In: *Annu. Rev. Genet.* 39 (2005), pp. 409–429.

[30] Lasha Gogokhia et al. “Expansion of Bacteriophages Is Linked to Aggravated Intestinal Inflammation and Colitis”. en. In: *Cell Host Microbe* 25.2 (Feb. 2019), 285–299.e8.

[31] Zohar Erez et al. “Communication between viruses guides lysis–lysogeny decisions”. en. In: *Nature* 541.7638 (Jan. 2017), pp. 488–493.

[32] Michael Shamash and Corinne F Maurice. “Phages in the infant gut: a framework for virome development during early life”. en. In: *ISME J.* (Aug. 2021).

[33] Guanxiang Liang et al. “The stepwise assembly of the neonatal virome is modulated by breastfeeding”. en. In: *Nature* 581.7809 (May 2020), pp. 470–474.

[34] Joseph Bondy-Denomy et al. “Prophages mediate defense against phage infection through diverse mechanisms”. en. In: *ISME J.* 10.12 (Dec. 2016), pp. 2854–2866.

[35] François Rousset et al. “Phages and their satellites encode hotspots of antiviral systems”. en. In: *Cell Host Microbe* 30.5 (May 2022), 740–753.e5.

[36] Elie Dolgin. “The secret social lives of viruses”. en. In: *Nature* 570.7761 (June 2019), pp. 290–292.

[37] Ann C Gregory et al. “The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut”. en. In: *Cell Host Microbe* 28.5 (Nov. 2020), 724–740.e8.

[38] Cameron Martino et al. “Microbiota succession throughout life from the cradle to the grave”. en. In: *Nat. Rev. Microbiol.* (July 2022).

[39] Luisa De Sordi, Marta Lourenço, and Laurent Debarbieux. ""I will survive": A tale of bacteriophage-bacteria coevolution in the gut". In: *Gut Microbes* 10.1 (Jan. 2019), pp. 92–99.

[40] Christophe Fraser et al. "The bacterial species challenge: making sense of genetic and ecological diversity". en. In: *Science* 323.5915 (Feb. 2009), pp. 741–746.

[41] Simon J Labrie, Julie E Samson, and Sylvain Moineau. "Bacteriophage resistance mechanisms". en. In: *Nat. Rev. Microbiol.* 8.5 (May 2010), pp. 317–327.

[42] Julie E Samson et al. "Revenge of the phages: defeating bacterial defences". en. In: *Nat. Rev. Microbiol.* 11.10 (Oct. 2013), pp. 675–687.

[43] Jakob Haaber et al. "Bacterial viruses enable their host to acquire antibiotic resistance genes from neighbouring cells". en. In: *Nat. Commun.* 7 (Nov. 2016), p. 13333.

[44] Andrey N Shkoporov, Christopher J Turkington, and Colin Hill. "Mutualistic interplay between bacteriophages and bacteria in the human gut". en. In: *Nat. Rev. Microbiol.* (June 2022).

[45] Xiaoxue Wang et al. "Cryptic prophages help bacteria cope with adverse environments". en. In: *Nat. Commun.* 1 (2010), p. 147.

[46] Kristopher Kieft et al. "Virus-associated organosulfur metabolism in human and environmental systems". en. In: *Cell Rep.* 36.5 (Aug. 2021), p. 109471.

[47] Karthik Anantharaman et al. "Sulfur oxidation genes in diverse deep-sea viruses". en. In: *Science* 344.6185 (May 2014), pp. 757–760.

[48] Luke R Thompson et al. "Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 108.39 (Sept. 2011), E757–64.

[49] Marie Touchon, Jorge A Moura de Sousa, and Eduardo Pc Rocha. "Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer". en. In: *Curr. Opin. Microbiol.* 38 (Aug. 2017), pp. 66–73.

[50] Matthew S Fullmer, Shannon M Soucy, and Johann Peter Gogarten. "The pan-genome as a shared genomic resource: mutual cheating, cooperation and the black queen hypothesis". en. In: *Front. Microbiol.* 6 (July 2015), p. 728.

[51] June L Round and Sarkis K Mazmanian. “The gut microbiota shapes intestinal immune responses during health and disease”. en. In: *Nat. Rev. Immunol.* 9.5 (May 2009), pp. 313–323.

[52] Daniel B Graham et al. “Antigen discovery and specification of immunodominance hierarchies for MHCII-restricted epitopes”. In: *Nat. Med.* 24.11 (Nov. 2018), pp. 1762–1772.

[53] Lora V Hooper, Dan R Littman, and Andrew J Macpherson. *Interactions Between the Microbiota and the Immune System*. 2012.

[54] Malin E V Johansson, Henrik Sjövall, and Gunnar C Hansson. “The gastrointestinal mucus system in health and disease”. en. In: *Nat. Rev. Gastroenterol. Hepatol.* 10.6 (June 2013), pp. 352–361.

[55] Bjoern O Schroeder. *Fight them or feed them: how the intestinal mucus layer manages the gut microbiota*. 2019.

[56] Jian Fang et al. “Slimy partners: the mucus barrier and gut microbiome in ulcerative colitis”. en. In: *Exp. Mol. Med.* 53.5 (May 2021), pp. 772–787.

[57] Malin E V Johansson. “Mucus layers in inflammatory bowel disease”. en. In: *Inflamm. Bowel Dis.* 20.11 (Nov. 2014), pp. 2124–2131.

[58] Rebecca N Metzger, Anne B Krug, and Katharina Eisenächer. “Enteric Virome Sensing-Its Role in Intestinal Homeostasis and Immunity”. en. In: *Viruses* 10.4 (Mar. 2018).

[59] Breck A Duerkop and Lora V Hooper. “Resident viruses and their interactions with the immune system”. en. In: *Nat. Immunol.* 14.7 (July 2013), pp. 654–659.

[60] Marion C Bichet et al. “Bacteriophage uptake by mammalian cell layers represents a potential sink that may impact phage therapy”. en. In: *iScience* 24.4 (Apr. 2021), p. 102287.

[61] Sjoerd van der Post et al. “Structural weakening of the colonic mucus barrier is an early event in ulcerative colitis pathogenesis”. en. In: *Gut* 68.12 (Dec. 2019), pp. 2142–2151.

[62] Fatemeh Adiliaghdam et al. “Human enteric viruses autonomously shape inflammatory bowel disease phenotype through divergent innate immunomodulation”. In: *Science Immunology* 7.70 (2022), eabn6660.

[63] Katherine R Hargreaves and Martha R J Clokie. “Clostridium difficile phages: still difficult?” en. In: *Front. Microbiol.* 5 (Apr. 2014), p. 184.

[64] Torben Sølbeck Rasmussen et al. “Faecal virome transplantation decreases symptoms of type 2 diabetes and obesity in a murine model”. en. In: *Gut* (Mar. 2020).

[65] Jordi Mayneris-Perxachs et al. *Caudovirales bacteriophages are associated with improved executive function and memory in flies, mice, and humans*. 2022.

[66] Barr, Auro, Furlan, et al. “Bacteriophage adhering to mucus provide a non–host-derived immunity”. In: *Proc. Estonian Acad. Sci. Biol. Ecol.* ().

[67] Thomas D S Sutton et al. “Choice of assembly software has a critical impact on virome characterisation”. In: *Microbiome* 7.1 (Jan. 2019), p. 12.

[68] Sanzhima Garmaeva et al. “Studying the gut virome in the metagenomic era: challenges and perspectives”. In: *BMC Biol.* 17.1 (Oct. 2019), p. 84.

[69] Yu-Wei Wu, Blake A Simmons, and Steven W Singer. “MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets”. en. In: *Bioinformatics* 32.4 (Feb. 2016), pp. 605–607.

[70] David T Pride et al. “Evolutionary implications of microbial genome tetranucleotide frequency biases”. en. In: *Genome Res.* 13.2 (Feb. 2003), pp. 145–158.

[71] Adrian Fritz et al. “CAMISIM: simulating metagenomes and microbial communities”. en. In: *Microbiome* 7.1 (Feb. 2019), p. 17.

[72] Shanika L Amarasinghe et al. “Opportunities and challenges in long-read sequencing data analysis”. en. In: *Genome Biol.* 21.1 (Feb. 2020), p. 30.

[73] Pierre-Alain Chaumeil et al. “GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database”. en. In: *Bioinformatics* (Nov. 2019).

[74] Philip Hugenholtz et al. “Prokaryotic taxonomy and nomenclature in the age of big sequence data”. en. In: *ISME J.* 15.7 (July 2021), pp. 1879–1892.

[75] Shahana S Malik et al. *Do Viruses Exchange Genes across Superkingdoms of Life?* 2017.

[76] Simon Roux et al. “VirSorter: mining viral signal from microbial genomic data”. en. In: *PeerJ* 3 (May 2015), e985.

[77] Jie Ren et al. “VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data”. In: *Microbiome* 5.1 (July 2017), p. 69.

[78] Stephen Nayfach et al. “CheckV assesses the quality and completeness of metagenome-assembled viral genomes”. en. In: *Nat. Biotechnol.* 39.5 (May 2021), pp. 578–585.

[79] Kristopher Kieft, Zhichao Zhou, and Karthik Anantharaman. “VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences”. en. In: *Microbiome* 8.1 (June 2020), p. 90.

[80] Patrick A de Jonge et al. “Molecular and Evolutionary Determinants of Bacteriophage Host Range”. en. In: *Trends Microbiol.* 27.1 (Jan. 2019), pp. 51–63.

[81] Matthew B Sullivan, John B Waterbury, and Sallie W Chisholm. “Cyanophages infecting the oceanic cyanobacterium Prochlorococcus”. en. In: *Nature* 424.6952 (Aug. 2003), pp. 1047–1051.

[82] Frank Hille et al. “The Biology of CRISPR-Cas: Backward and Forward”. en. In: *Cell* 172.6 (Mar. 2018), pp. 1239–1259.

[83] Jakob Russel et al. “CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci”. en. In: *CRISPR J* 3.6 (Dec. 2020), pp. 462–469.

[84] Robert A Edwards et al. “Computational approaches to predict bacteriophage–host relationships”. en. In: *FEMS Microbiol. Rev.* 40.2 (Dec. 2015), pp. 258–272.

[85] Moıra B Dion et al. “Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter”. en. In: *Nucleic Acids Res.* 49.6 (Apr. 2021), pp. 3127–3138.

[86] Adam J Hockenberry and Claus O Wilke. “BACPHLIP: Predicting bacteriophage lifestyle from conserved protein domains”. en. May 2020.

[87] Shufang Wu et al. “DeePhage: distinguishing virulent and temperate phage-derived sequences in metavirome data with a deep learning approach”. en. In: *Gigascience* 10.9 (Sept. 2021).

[88] Edoardo Pasolli et al. “Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights”. en. In: *PLoS Comput. Biol.* 12.7 (July 2016), e1004977.

[89] Alexander Statnikov et al. “A comprehensive evaluation of multicategory classification methods for microbiomic data”. en. In: *Microbiome* 1.1 (Apr. 2013), p. 11.

[90] Wenyu Zhou et al. “Longitudinal multi-omics of host–microbe dynamics in prediabetes”. en. In: *Nature* 569.7758 (May 2019), pp. 663–671.

[91] Laura Judith Marcos-Zambrano et al. “Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment”. en. In: *Front. Microbiol.* 12 (Feb. 2021), p. 634511.

[92] Tianle Ma and Aidong Zhang. “Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE)”. en. In: *BMC Genomics* 20.Suppl 11 (Dec. 2019), p. 944.

[93] Juexin Wang et al. “scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses”. en. In: *Nat. Commun.* 12.1 (Mar. 2021), p. 1882.

[94] Carl Doersch. “Tutorial on Variational Autoencoders”. In: (June 2016). arXiv: `1606.05908 [stat.ML]`.

[95] Colin J Carlson et al. “Global estimates of mammalian viral diversity accounting for host sharing”. en. In: *Nat Ecol Evol* 3.7 (July 2019), pp. 1070–1075.

[96] Peter J Turnbaugh et al. “The human microbiome project”. en. In: *Nature* 449.7164 (Oct. 2007), pp. 804–810.

[97] S Dusko Ehrlich. “MetaHIT: The European Union Project on Metagenomics of the Human Intestinal Tract”. In: *Metagenomics of the Human Body*. Ed. by Karen E Nelson. New York, NY: Springer New York, 2011, pp. 307–316.

[98] Donovan H Parks et al. *Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life*. 2017.

[99] Edoardo Pasolli et al. “Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle”. en. In: *Cell* 176.3 (Jan. 2019), 649–662.e20.

[100] Alexander Sczyrba et al. “Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software”. en. In: *Nat. Methods* 14.11 (Oct. 2017), pp. 1063–1071.

[101] Varun Aggarwala, Guanxiang Liang, and Frederic D Bushman. "Viral communities of the human gut: metagenomic analysis of composition and dynamics". en. In: *Mob. DNA* 8 (Oct. 2017), p. 12.

[102] Shiraz A Shah et al. *Manual resolution of virome dark matter uncovers hundreds of viral families in the infant gut.*

[103] Guoyan Zhao et al. "Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 114.30 (July 2017), E6166–E6175.

[104] Jason Lloyd-Price et al. "Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases". In: *Nature* 569.7758 (May 2019), pp. 655–662.

[105] Takumi Hirata et al. "Associations of cardiovascular biomarkers and plasma albumin with exceptional survival to the highest ages". en. In: *Nat. Commun.* 11.1 (July 2020), p. 3820.

[106] N Barzilai, G Atzmon, and C Schechter. *Unique lipoprotein phenotype and genotype associated with exceptional longevity.* 2004.

[107] Tomasz Wilmanski et al. "Gut microbiome pattern reflects healthy ageing and predicts survival in humans". en. In: *Nat Metab* 3.2 (Feb. 2021), pp. 274–286.

[108] Yuko Sato et al. "Novel bile acid biosynthetic pathways are enriched in the microbiome of centenarians". en. In: *Nature* 599.7885 (Nov. 2021), pp. 458–464.

[109] Stephen Nayfach et al. "Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome". en. In: *Nat Microbiol* (June 2021).

[110] Lu Wu et al. "A Cross-Sectional Study of Compositional and Functional Profiles of Gut Microbiota in Sardinian Centenarians". en. In: *mSystems* 4.4 (July 2019).

[111] Tommi Vatanen et al. "Transcription shifts in gut bacteria shared between mothers and their infants". en. In: *Sci. Rep.* 12.1 (Jan. 2022), p. 1276.

[112] Martin Stražar et al. "Gut microbiome-mediated metabolism effects on immunity in rural and urban African populations". en. In: *Nat. Commun.* 12.1 (Aug. 2021), p. 4845.

[113] Matthew R Olm et al. “dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication”. en. In: *ISME J.* 11.12 (Dec. 2017), pp. 2864–2868.

[114] Gherman V Uritskiy, Jocelyne DiRuggiero, and James Taylor. “MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis”. en. In: *Microbiome* 6.1 (Sept. 2018), pp. 1–13.

[115] Christian M K Sieber et al. “Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy”. en. In: *Nat Microbiol* 3.7 (July 2018), pp. 836–843.

[116] Mohammad Abdul Bakir et al. “Bacteroides dorei sp. nov., isolated from human faeces”. en. In: *Int. J. Syst. Evol. Microbiol.* 56.Pt 7 (July 2006), pp. 1639–1643.

[117] Leland McInnes, John Healy, and Steve Astels. “hdbscan: Hierarchical density based clustering”. In: *J. Open Source Softw.* 2.11 (Mar. 2017), p. 205.

[118] Vincent Daubin, Emmanuelle Lerat, and Guy Perrière. “The source of laterally transferred genes in bacterial genomes”. en. In: *Genome Biol.* 4.9 (Aug. 2003), R57.

[119] Linyi Alex Gao et al. “Prokaryotic innate immunity through pattern recognition of conserved viral proteins”. en. In: *Science* 377.6607 (Aug. 2022), eabm4096.

[120] Nicolás Toro et al. “Multiple origins of reverse transcriptases linked to CRISPR-Cas systems”. en. In: *RNA Biol.* 16.10 (Oct. 2019), pp. 1486–1493.

[121] Simon Roux et al. “Ecology and molecular targets of hypermutation in the global microbiome”. en. Apr. 2020.

[122] Leighton J Payne et al. “Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types”. en. In: *Nucleic Acids Res.* 49.19 (Oct. 2021), pp. 10868–10878.

[123] A Fluckiger et al. “Cross-reactivity between tumor MHC class I–restricted antigens and an enterococcal bacteriophage”. In: (2020).

[124] Eli L Moss, Dylan G Maghini, and Ami S Bhatt. “Complete, closed bacterial genomes from microbiomes using nanopore sequencing”. en. In: *Nat. Biotechnol.* 38.6 (June 2020), pp. 701–707.

[125] Mantas Sereika et al. “Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing”. en. In: *Nat. Methods* 19.7 (July 2022), pp. 823–826.

[126] Asier Zaragoza-Solas et al. “Long-Read Metagenomics Improves the Recovery of Viral Diversity from Complex Natural Marine Samples”. en. In: *mSystems* 7.3 (June 2022), e0019222.

[127] Clara Delahaye and Jacques Nicolas. “Sequencing DNA with nanopores: Troubles and biases”. en. In: *PLoS One* 16.10 (Oct. 2021), e0257521.

[128] Ryan R Wick et al. “Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads”. en. In: *PLoS Comput. Biol.* 13.6 (June 2017), e1005595.

[129] Koji Yahara et al. “Long-read metagenomics using PromethION uncovers oral bacteriophages and their interaction with host bacteria”. en. In: *Nat. Commun.* 12.1 (Jan. 2021), p. 27.

[130] Christopher J Stewart et al. “Temporal development of the gut microbiome in early childhood from the TEDDY study”. en. In: *Nature* 562.7728 (Oct. 2018), pp. 583–588.

[131] Tarini Shankar Ghosh, Fergus Shanahan, and Paul W O’Toole. “The gut microbiome as a modulator of healthy ageing”. en. In: *Nat. Rev. Gastroenterol. Hepatol.* 19.9 (Sept. 2022), pp. 565–584.

[132] Apollo Stacy et al. “Infection trains the host for microbiota-enhanced resistance to pathogens”. en. In: *Cell* 184.3 (Feb. 2021), 615–627.e17.

[133] Joachim Johansen et al. “Genome binning of viral entities from bulk metagenomics data”. eng. In: *Nat. Commun.* 13.1 (Feb. 2022), p. 965.

[134] Wei Wei et al. “Reduced bacterial mortality and enhanced viral productivity during sinking in the ocean”. en. In: *ISME J.* 16.6 (June 2022), pp. 1668–1675.

[135] Mirjam Zünd et al. “High throughput sequencing provides exact genomic locations of inducible prophages and accurate phage-to-host ratios in gut microbial strains”. en. In: *Microbiome* 9.1 (Mar. 2021), p. 77.

[136] Robert Hertel et al. “Genome-based identification of active prophage regions by next generation sequencing in Bacillus licheniformis DSM13”. en. In: *PLoS One* 10.3 (Mar. 2015), e0120759.

[137] Kristopher Kieft et al. “Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages”. en. In: *Nat. Commun.* 12.1 (June 2021), p. 3503.

[138] Eric M Brown et al. “Gut microbiome ADP-ribosyltransferases are widespread phage-encoded fitness factors”. en. In: *Cell Host Microbe* 29.9 (Sept. 2021), 1351–1365.e11.

[139] Anne Jamet et al. “A widespread family of polymorphic toxins encoded by temperate phages”. en. In: *BMC Biol.* 15.1 (Aug. 2017), p. 75.

[140] J W Obringer. “The functions of the phage T4 immunity and spackle genes in genetic exclusion”. en. In: *Genet. Res.* 52.2 (Oct. 1988), pp. 81–90.

[141] D A Schwartz, B K Lehmkuhl, and J T Lennon. “Phage-encoded sigma factors alter bacterial dormancy”. en. Nov. 2021.

[142] Sara Federici et al. “Targeted suppression of human IBD-associated gut microbiota commensals by phage consortia for treatment of intestinal inflammation”. en. In: *Cell* 185.16 (Aug. 2022), 2879–2898.e24.

[143] Tristan Bepler and Bonnie Berger. “Learning the protein language: Evolution, structure, and function”. en. In: *Cell Syst* 12.6 (June 2021), 654–669.e3.

[144] John Jumper et al. “Highly accurate protein structure prediction with AlphaFold”. en. In: *Nature* 596.7873 (Aug. 2021), pp. 583–589.

[145] Ellen Murray et al. “The Advantages and Challenges of Using Endolysins in a Clinical Setting”. en. In: *Viruses* 13.4 (Apr. 2021), p. 680.

[146] Evelien M Adriaenssens et al. “Taxonomy of prokaryotic viruses: 2018-2019 update from the ICTV Bacterial and Archaeal Viruses Subcommittee”. en. In: *Arch. Virol.* 165.5 (May 2020), pp. 1253–1260.

[147] Dance. “The incredible diversity of viruses”. In: *Nature* ().

[148] Pilar Manrique et al. “Healthy human gut phageome”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 113.37 (Sept. 2016), pp. 10400–10405.

[149] John C Wooley and Yuzhen Ye. “Metagenomics: Facts and Artifacts, and Computational Challenges”. In: *J. Comput. Sci. Technol.* 25.1 (Jan. 2010), pp. 71–81.

[150] Tao Zuo et al. “Depicting SARS-CoV-2 faecal viral activity in association with gut microbiota composition in patients with COVID-19”. en. In: *Gut* 70.2 (Feb. 2021), pp. 276–284.

[151] J Callanan et al. “Expansion of known ssRNA phage genomes: From tens to over a thousand”. en. In: *Sci Adv* 6.6 (Feb. 2020), eaay5981.

[152] Marcos Parras-Moltó et al. “Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses”. en. In: *Microbiome* 6.1 (June 2018), p. 119.

[153] Ethan C Alley et al. “Unified rational protein engineering with sequence-based deep representation learning”. en. In: *Nat. Methods* 16.12 (Dec. 2019), pp. 1315–1322.