Thesis Title

Ph.D Thesis

Author Name

Principal supervisor: First Last
Co-suprevisor: First Last

This thesis has been submitted to the Graduate School of Health and Medical Sciences, University of Copenhagen on Month Day, Year

Preface

The work in this thesis was carried out between month year and month year 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 funding for this PhD comes from the Novo Nordisk Foundation (grant no: NNFX).

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

Acknowledgements iii

English summary vi

Dansk resumé viii

List of publications x

Thesis content and overview xi

1 Background 1

1.1 The study of human gut biodiversity through DNA 2

1.2 Resolving uncultivated genomes in metagenomics 2

2 Methods 6

2.1 Machine learning 7

3 Research objectives 13

4 Description of research projects 16

4.1 Project I: Deep learning for binning and high resolution taxonomic profiling of microbial genomes 17

4.2 Project II: Genome binning of viral entities from bulk metagenomics data 18

4.3 Project III: Centenarians have a diverse population of gut bacteriophages that may promote healthy lifespan 19

4.4 Datasets overview 22

5 Summary of results and discussion 23

6 Conclusions and perspectives 31

7 Ethical and legal permits and approvals 33

8 Manuscripts **34**

8.1 Paper II: Genome binning of viral entities from bulk metagenomics data 35

9 Appendix **48**

9.1 Popular science article in Politiken 49

References **51**

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. **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.
In review, Nature Microbiology

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

PAPER I. 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*

PAPER II. 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.

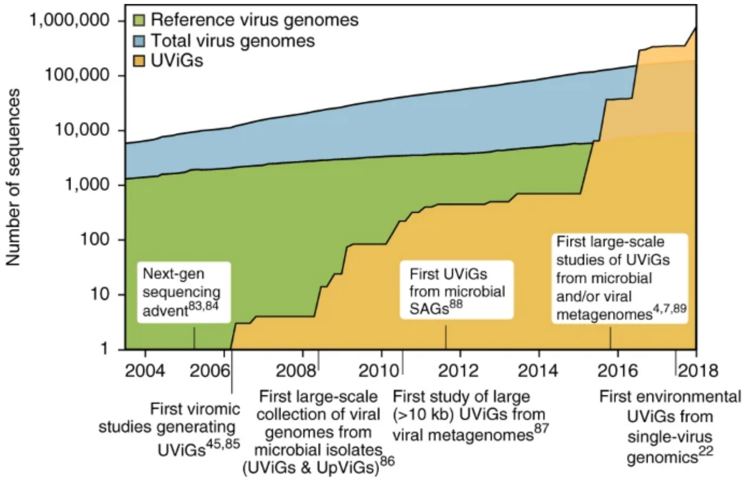


Figure 1.1. The figure illustrates a recent timeline going back to 2004 and the cumulative number of virus genomes uploaded to genome databases since then, including uncultured virus genomes (UViGs). In addition, major events related to virus discovery are notated across the timeline. Virus genomes cultured *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 bacteriophages 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.

2 Methods

2.1 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 [19, 20]. 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 [19]. 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 [21]. 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 [22]. For one of the publications in this dissertation, we leveraged the Random Forest (RF) model.

2.1.1 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 2.1). The randomness comes into play as the y-variables used to describe each observation are randomly sampled at each node in the decision trees.

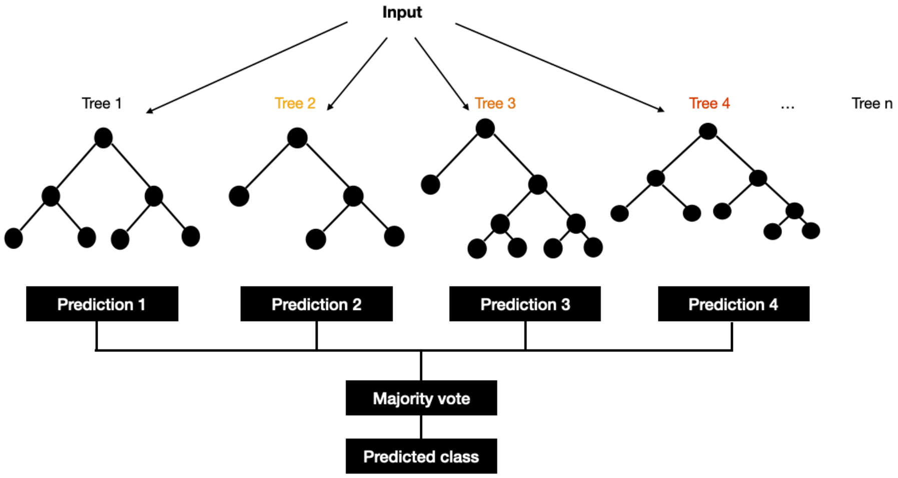


Figure 2.1. The Random Forest (RF) contains a predefined number of decision trees, which are randomly constructed resulting in n independent decision trees. A trained RF model works by receiving an input that is processed by each decision tree in the forest, which produces an independent prediction of 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

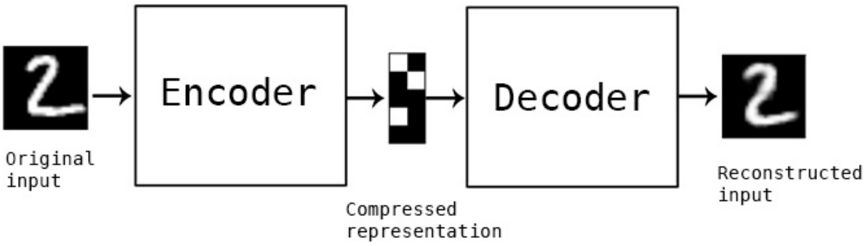


Figure 2.2. A classic example of Autoencoders is the usage for reconstructing images. In the example depicted, the original input (image of a "2") is encoded into a compressed representation and decoded into an output very similar to the original image. Illustration from: https://blog.keras.io/building-autoencoders-in-keras.html

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.

2.1.2 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 [23, 24]. 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 2.2).

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 representation z, which can be decoded into an approximation of the original input $\hat{y}$ (Figure 2.3), 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 $E(y)$ 2.3), 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.

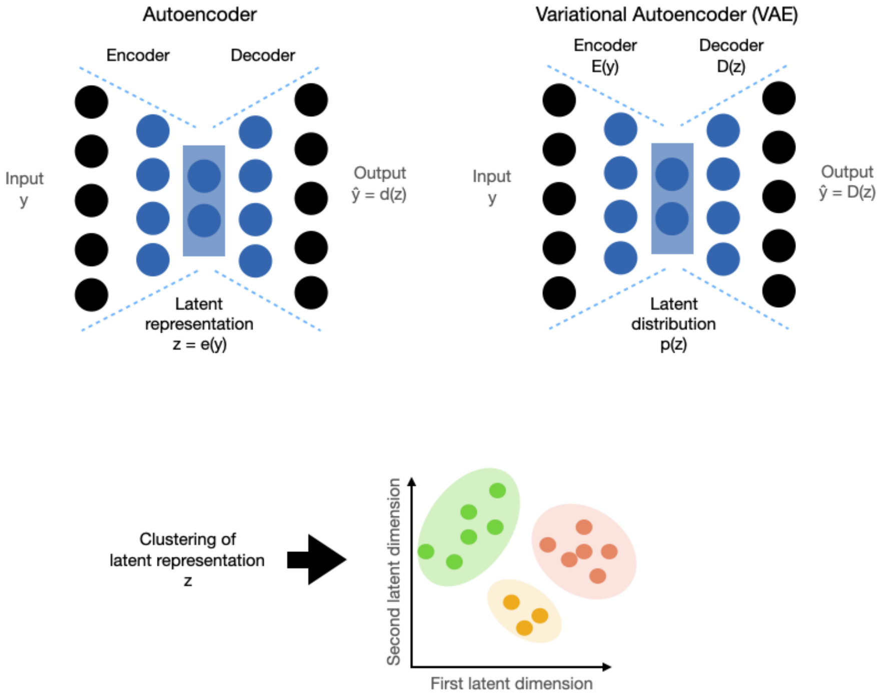


Figure 2.3. The Autoencoder and Variational Autoencoder (VAE) may have similar architectures as they both contain an input layer y and output $\hat{y}$ layer with same dimensions, an encoder and decoder layer and a latent representation (center layer). However, the VAE encodes the input into a latent 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 [25]; (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 arbitrary 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 [26]. 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)[27] and the European MetaHIT project [28]. 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 [29, 30], 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, 31], 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 [32] 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 [33]. 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 [34].

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 [35]. 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 [36].
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 [37], 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 [38, 39]. In addition, centenarians show great resistance to aging-related diseases. One of the suggested components to contribute to their longevity is the gut microbiome [40]. An initial characterisation of centenarian microbiomes revealed enrichment of bacteria capable of producing novel secondary bile acids with antibiotic properties towards typical gut pathogens [41]. 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 [42]. 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 [43, 44]. 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 [45], 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 [32]. Simulated metagenomic benchmark datasets.

Dataset III. COPSAC 2010 [34]. 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) [35]. 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 [37]. 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 [41]. 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 [46]. 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 [47]. 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 [48]. Cross sectional study of 315 adults from Tanzania. The dataset comprises 315 human gut microbiomes.

5 Summary of results and discussion

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 5.1). 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 5.1). 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?

5.0.1 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 [44], 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

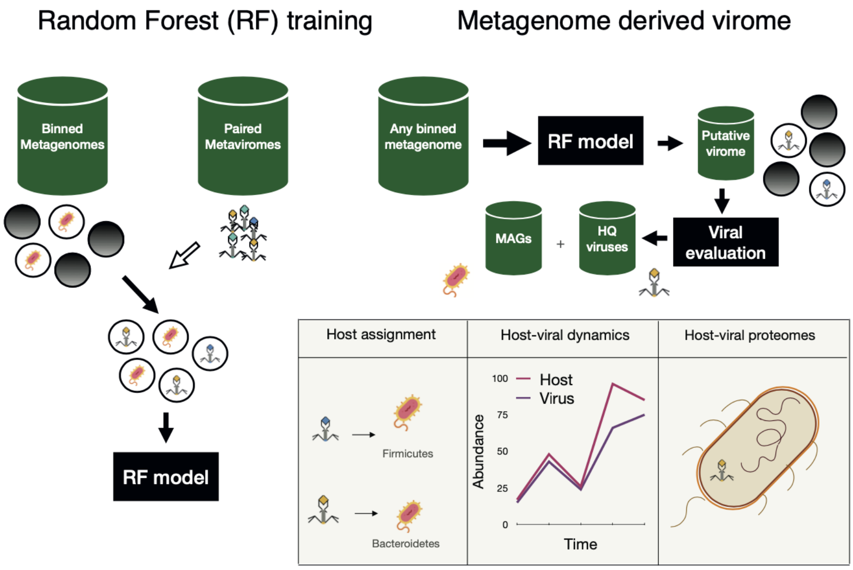


Figure 5.1. Random Forest (RF) modeling was performed on binned metagenomes paired with metaviromes. The trained RF model can be applied to any binned metagenome for predicting the putative virome in which high-quality (HQ) viruses can be identified and combined with bacterial MAGs. Identification of 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 [49]. 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 [49].

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 5.2). 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.

5.0.2 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

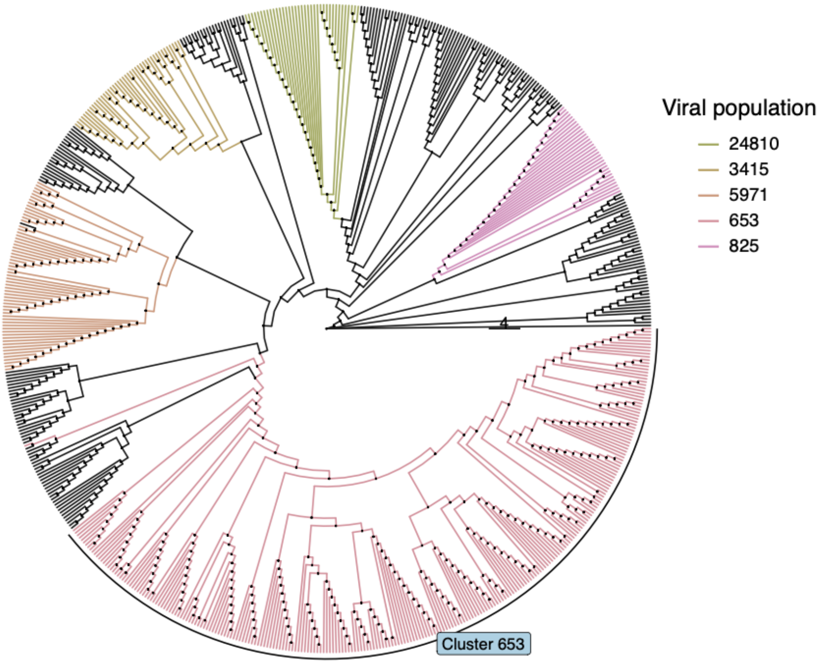


Figure 5.2. Cladogram based on a phylogenetic tree of crAss-like virus genomes colored and named by VAMB-cluster.

infection mechanisms [50] and in diversity generative regions (DGRs) [51]. In order to circumvent bacterial defense systems and exclude other viral competitors, phages also encode prokaryotic defense systems such as anti-viral systems [52]. 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 [53] 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 [54]. 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” [55]. 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.

5.0.3 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 [56, 57]. 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 [58]. 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 [59]. However, the recurring issues with 3rd generation sequencing comprises higher base-calling error rate and frequency of insertions and deletions (indels) [60]. 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 [61]. 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 [57]. So, where do we stand in terms of long-read sequencing of viruses in bulk metagenomics? Recent studies have explored the benefits of combining

2nd and 3rd generation sequencing [58, 62] and shown that long-reads captured additional viral diversity but also illustrated that long-reads with PacBio sequencing only captured few HQ viruses [58]. 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.

5.0.4 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 [63]. 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 [64, 65]. A recent study into the infant virome had revealed the turbulent development of the pioneering viral constituents and response to the maturing bacterial community [43], while another mapped the development of viral families over time from infant to elderly [44]. 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

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 [66], yet the degree of viral genomic diversity and variation between biotic environments is suggesting that only a fraction of viral diversity have been identified [67]. 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, 68, 69, 70]. 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, 44]. 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 [71].

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 [72], but their discovery is dependent on metatranscriptomics and construction of cDNA libraries [73]. 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 bacteriophages and integrated proviruses [15, 74, 44]. 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 [44]. 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 [56, 57].

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 [75], 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 [76]. 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 [77]. 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 [78, 45]. 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 [79]. Our understanding of the human gut virome and its interplay with bacterial constituents is still in its infancy [80], 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.” [37]

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.” [34]

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).” [41]

8 Manuscripts

8.1 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

Check for updates

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-

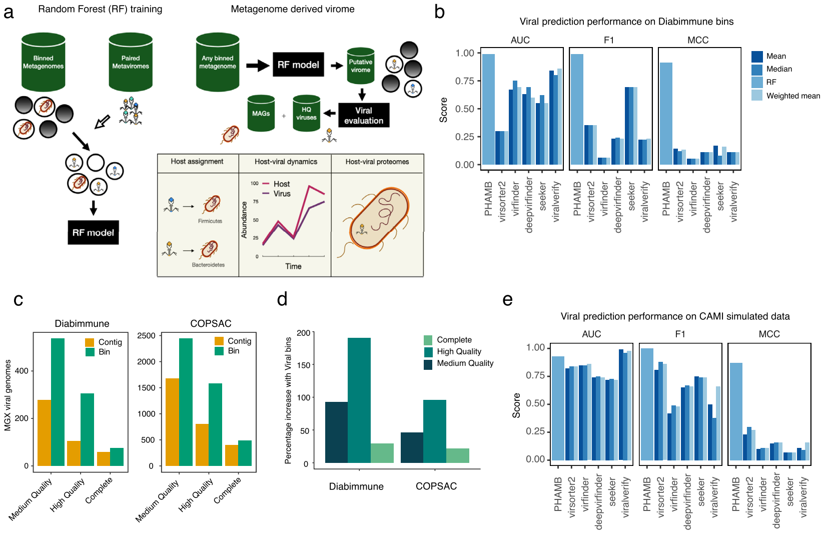


Fig. 1 A framework to bin and assemble viral populations from metagenomics data. a Illustration of workflow to explore viruses from binned metagenomes. First, the RF model was trained on binned metagenomes; bacterial bins were identified using reference database tools and viruses were identified using 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, $T = 34.02$, $\mathrm{CI} = 24.4;27.4$, $P = 2.2\mathrm{e}{-16}$) (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 $P = 0.017$, $t = -2.47$, $\mathrm{CI} = -0.01;-0.13$) and CD subjects (*T*-test, two-sided, $P = 0.023$, $t = -2.3$, $\mathrm{CI} = -0.12;-0.01$) (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

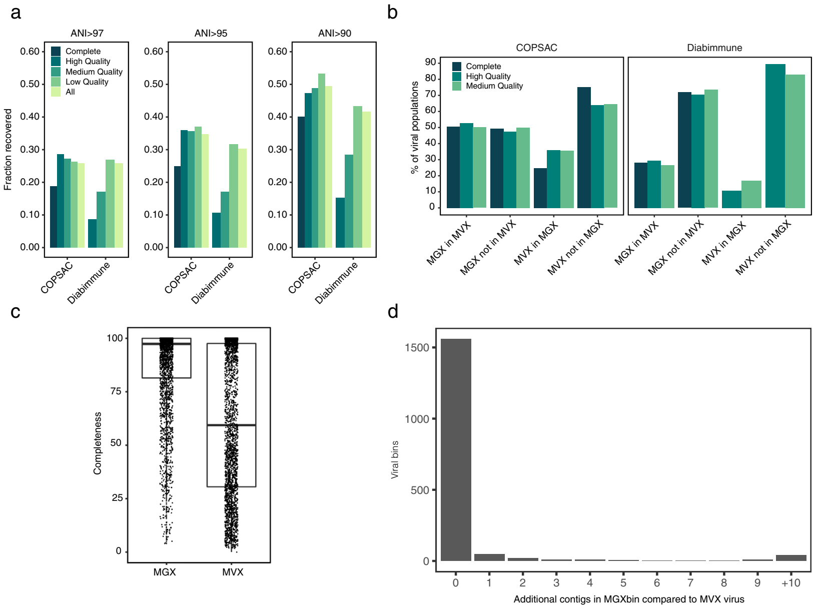


Fig. 2 Binning the metagenome identifies viral genomes not identified from the metavirome. **a** The fraction of metavirome viruses in COPSAC and Diabimmune coloured at different levels of completeness or all together determined with CheckV, identified in VAMB bins from bulk metagenomics of the same 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

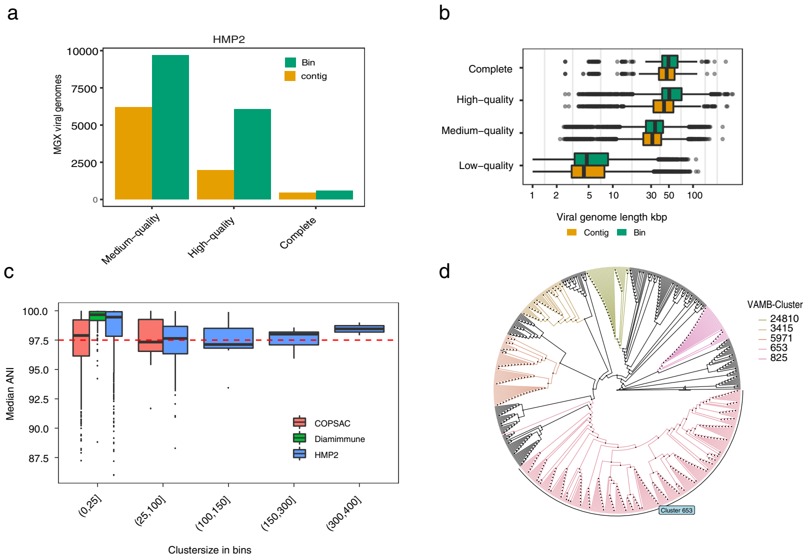


Fig. 3 Reconstructing the virome of a human gut metagenomics cohort. a The number of viral genomes with three different levels of completeness in HMP2, evaluated as either single-contigs or viral bins from bulk metagenomes. Evaluation of genome completeness was determined using CheckV here shown for 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

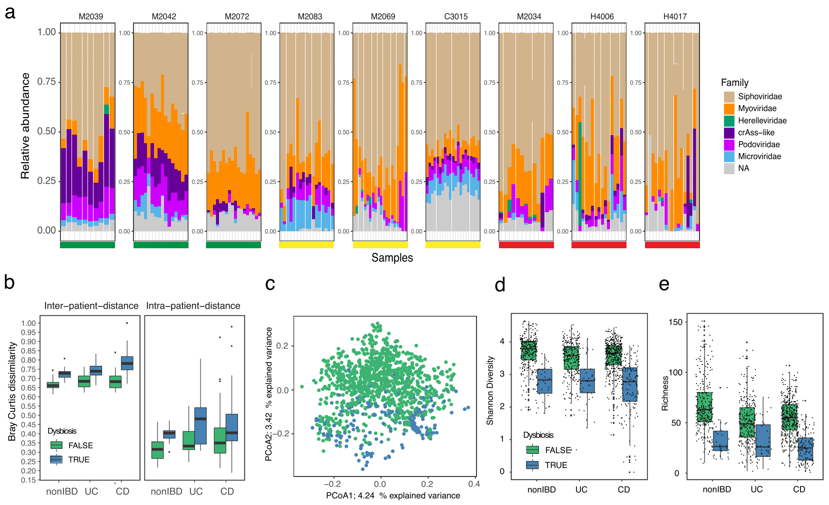


Fig. 4 The metagenomics estimated virome is personal and highly stable in healthy controls. a Longitudinal virome compositions for three nonIBD (green bar), three UC (yellow bar) and three CD (red bar) diagnosed subjects. Each panel represents a subject where the virome composition was organised according 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

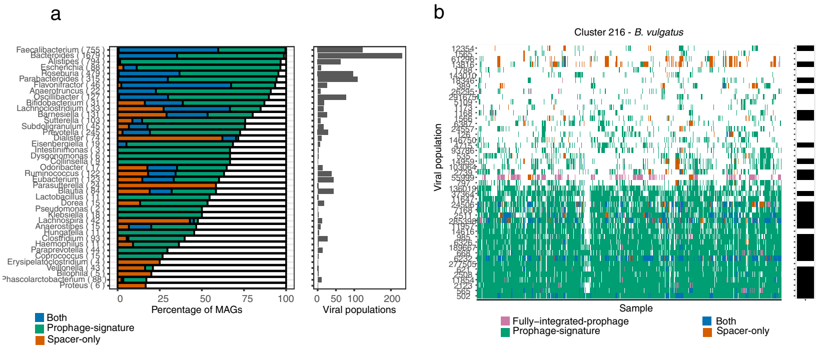


Fig. 5 Viral-host interactions can be explored from viral populations and MAGs. a Bacterial MAGs and viral relations. Each MAG was connected to the viral bins using either sequence alignment of the virus to MAG (green), CRISPR spacer alignment (orange) or both (blue). The right panel shows the percentage 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

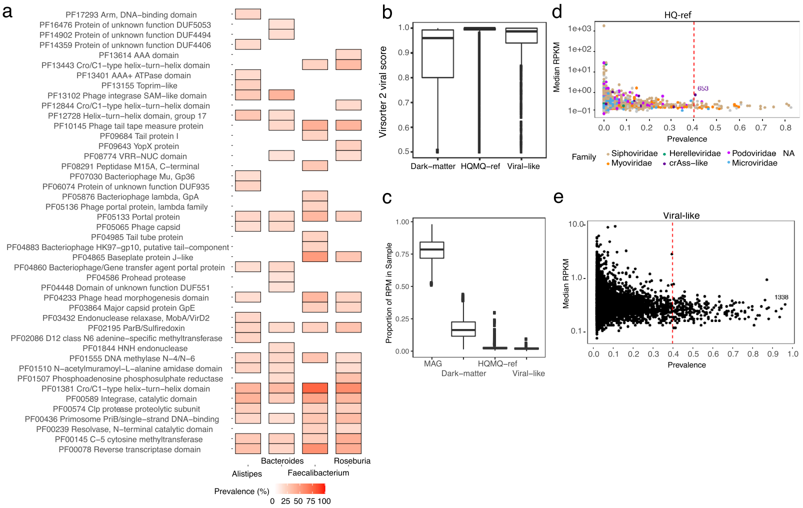


Fig. 6 Viral proteins and the dark-matter metavirome. a The percentage of HQ viruses, associated with four bacterial host genera; Alistipes, Bacteroides, Faecalibacterium and Roseburia, which encode top-20 prevalent PFAM domains. **b** Virsorter2 viral prediction scores for all viral bins with at least 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 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.

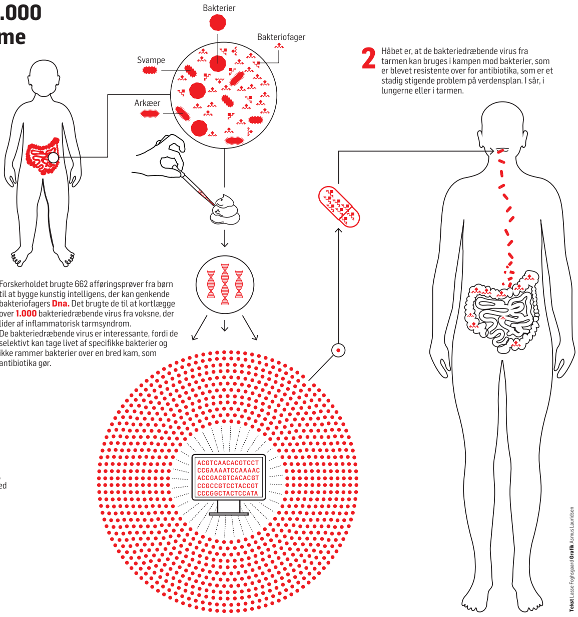
9 Appendix

9.1 Popular science article in Politiken

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

Danske forskere har kortlagt over 1.000 bakteriedræbende virus i børns tarme

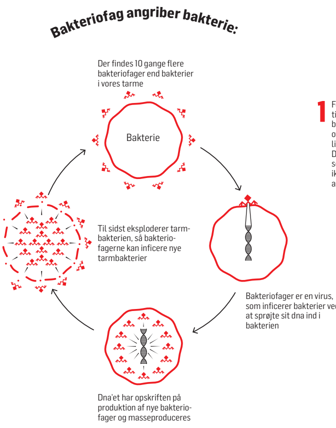
I vores tarme findes der store familier af sundhedsfremmende bakterier blandt svampe og arkæer. For at sikre, at de enkelte bakteriefamilier ikke vokser over evne og skader bakteriebalancen, findes der bakteriedræbende virus i tarmen, som bliver kaldt for bakteriofager, som målrettet kan tage livet af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



1 Forskerholdet brugte 662 afføringsprøver fra børn til at bygge kunstig intelligens, der kan genkende bakteriofagers **Dna.** Det brugte de til at kortlægge over **1.000** bakteriedræbende virus fra voksne, der lider af inflammatorisk tarmsyndrom.
De bakteriedræbende virus er interessante, fordi de selektivt kan tage livet af specifikke bakterier og ikke rammer bakterier over en bred kam, som antibiotika gør.

2 Håbet er, at de bakteriedræbende virus fra tarmen kan bruges i kampen mod bakterier, som er blevet resistente over for antibiotika, som er et stadig stigende problem på verdensplan. I sår, i lungerne eller i tarmen.

Bakteriofag angriber bakterie:



Der findes 10 gange flere bakteriofager end bakterier i vores tarme

Bakteriofager er en virus, som inficerer bakterier ved at sprøjte sit dna ind i bakterien

Dna'et har opskriften på produktion af nye bakteriofager og masseproduceres

Til sidst eksploderer tarmbakterien, så bakteriofagerne kan inficere nye tarmbakterier

Tekst Lasse Foghsgaard **Grafik** Asmus Lauridsen

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 forskerkolleger 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 sygdommen«, 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 Metabolismecenter 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 uheldig bivirkning ved traditionelle antibiotika«, 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'Hérelle bakteriofager fra afføringen hos patienter, der var ved at komme sig efter dysenteri. I 1931 udgav Félix D'Hérelle en videnskabelig artikel, hvor han beskrev, hvordan bakteriofager kunne bruges medicinsk til at behandle akutte infektioner »med 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 behandlingsprincippet 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 stil«, 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 bakterie«, spår Oluf Borbye Pedersen.

> **“ Behandlingen med afføring vil sandsynligvis om nogle år blive erstattet med specifikke bakteriofager, der som et missil kan dræbe den sygdomsfremkaldende bakterie**
>
> 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 folkesundheden«, 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] 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.

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

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

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

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

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

[35] 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.

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

[37] 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.

[38] 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.

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

[40] 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.

[41] 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.

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

[43] 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.

[44] 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.

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

[46] 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).

[47] 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.

[48] 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.

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

[50] 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.

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

[52] 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.

[53] 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.

[54] 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.

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

[56] 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.

[57] 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.

[58] 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.

[59] 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.

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

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

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

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

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

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

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

[77] 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.

[78] 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.

[79] 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.

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