B.Sc.Eng. Thesis
Bachelor of Science in Engineering

DTU Bioengineering

Evaluation of phage encoded biosynthetic gene clusters

Signe Maite Conde Frieboes
and
Carina Nørgaard Kock Holst

Supervisor
Assoc Prof. Mikael Lenz Strube

Bachelor thesis 17.5 points
May 16, 2022

1 Abstract

The presence of specialized biosynthetical pathways often signifies that there is a selective pressure, and often imposes traits that induce fitness thus enhance resistance for pathogens. These can be found in biosynthetical gene clusters (BGCs), where the newfound BGCs in phages are called phage-encoded BGCs (pBGCs). pBGCs are an unexplored field of study, which can help lift the veil in understanding how phages can help mold microbial communities.
The topic of this paper is the occurrence of pBGCs in metagenomic data from the human gut, and looks into the different types of BGCs in phages.
In this study we gain further insight into the composition of phages and pBGCs in the human gut metagenome, as well as their geographical differences and similarities. Futhermore we attempt to group these phages according to genetic and protein similiarities as well as attempt to predict their potential hosts.

2 Resumé

I dette projekt har vi undersøgt tilstedeværelsen af fag-indkodede BGCs i den menneskelige tarms metagenom. Hvori at der ligges fokus af kompositionen af pBGCer, og der kigges på de kontinentale forskelle og ligheder. Her igennem opgaven vil pBGCs ekstraheres, hvorefter der fylogeni af fager of BGCs laves til at direkte sammenligne de genomiske sekvenser. Derudover vil der vil analyseres på genomers prøveudtagnings data, for at undersøge nogen skelsættende geografiske sammenhænge. Tilsidst blev der desuden lavet noget prædiktion af fager og deres værtsceller. På baggrund af disse undersøgelser findes der frem til at der faktisk findes fag-indkodede BGCer, og forskelle mellem disses regionelle abundans.

3 Acknowledgements

We would like to thank our friends and family for their patience and support during the making of this thesis.
We would also like to thank our supervisor Mikael Lenz Strube, for his continued support and guidance.
As for proofreading we would like to thank T.T. Branch, and David Lynch for their assistance in weeding out grammatical errors.

4 Notations

4.1 Acronyms

BGCs biosynthetic gene clusters

pBGCs phage-encoded BGCs

NRPS non-ribosomal peptide synthases

PKS polyketide synthases

RiPP ribosomally synthesized and post-translationally modified peptides

MGEs mobile genetic elements

ANI average nucleotide identity

GCFs gene cluster families

RRE RiPP recognition element

BH Benjamini-Hochberg

Contents

1 Abstract **1**

2 Resumé **2**

3 Acknowledgements **3**

4 Notations **4**
4.1 Acronyms 4

5 Project statement **6**

6 Introduction **7**
6.1 Components in the microbiome 7
6.2 The influence of phages on microbiomes 8
6.3 Secondary metabolites classification 8
6.4 The parts of our pipeline 9

7 Methods and materials **10**
7.1 Data collecting 10
7.2 Detection and analysis of pBGCs 11
7.3 Clusters and phylogeny 11
7.4 Statistical Analysis 12

8 Results and analysis **13**
8.1 Mining of pBGCs 13
8.2 Phylogeny 16
8.3 Phages and host bacteria 19
8.4 Regional pBGCs 22

9 Discussion **27**

10 Perspectivation **32**

11 Conclusion **33**

12 Availability **34**
12.1 Data availability 34
12.2 Code availability 34

5 Project statement

In this project we considered the potential existence of secondary metabolites in phages, more specifically in phages in the human gut genome.
We know that secondary metabolites are a rare occurrence in viral genomes, as they are thought of as a way for microorganisms living under extreme conditions to thrive. These secondary metabolites are encoded by what are called bio-synthetic gene clusters, which can give the host favorable fitness, thus helping the host organism surviving better than neighboring rival organisms.
Secondary metabolites in bio-synthetic gene clusters were relatively recently discovered to exist in some phages, leading to new considerations on how traits might be shared between microbial communities.
So we wanted to analyze the existence and specialization of secondary metabolites in phage metagenomes. Specifically we want to look at the prospect of phage-encoded BGCs (pBGC) within the gut system. An important aspect is also if there are any regional differences and similarities in the composition of the gut metagenome pBGCs. Based on the previous findings about pBGCs in whole phage genomes [12], we hypothesize that pBGCs also exist in metagenomes, specifically the human gut genome. To further explore this hypothesis we ask the following work questions:

- Are there pBGCs in the human gut metagenome, and are we able to identify any of these?
- Does the data allow for phylogenetic grouping of the pBGC containing phages?
- Are there global differences in the occurrence of pBGC types based on region?

To answer these hypotheses, we will look at a collection of phages genomes collected from the gut. We will screen the prophages collected for bio-synthetic gene clusters. We will use the program AntiSMASH, these results can then be used for the analysis. Then we will look into the composition and typing of the pBGCs. Another aspect is to look at the geographical locations of the samples to compare the pBGCs content and composition of the different continents. Alongside this we will use the program vContact2 to cluster the phages and look at any phylogeny in the samples. We also further analyse the data for hosts and attempt to identify relevant phages.
We will do this all in silico with a data set of already collected metagenomes.

6 Introduction

Microbiomes are a vital part of the world, they are communities of microorganisms both the symbiotic and parasitic. They exist both inside and on the surface of all multi-cellular organisms, and interact in a vast amount different ways. These microbiomes vary in composition from sample to sample based on factors such as location, composition of medium, and genetic variation.
In this paper we will look into the gut metagenome, where we already have vast amounts of research into the components of the microbiome. We already know that over 10^{14} microorganisms exist within the gut [15]. Furthermore if we take a look at the bacteria in here, we know that the overall most abundant bacteria types are the gram-positive Firmicutes and the gram-negative Bacteroidetes [46]. It is known that especially gram-negative bacteria often are antibiotics resistant [32]. Other bacteria which might be interesting to mention are Faecalibacterium. Faecalibacterium is a gram-negative non-spore-forming and rod-shaped bacteria, with only one validated species which is Faecalibacterium prausnitzii [13]. This species has a 5% proportion of the bacteria in faeces, making it an abundant bacteria in the human gut [4].
This gives us a degree of understanding into the communities we observe. But still there are a lot of mysteries still left unsolved when we take into account the interactions between the organisms.
One common communication pathways is quorum-sensing, which is the cell-to-cell, and cell-to-self, communication that is observed between bacteria. Quorum-sensing uses small chemical signalling molecules as a measure for cell density as well as cell states, based on which the processes can change. It is also seen used to communicate between different bacterial species [28].

6.1 Components in the microbiome

To further look into these interactions, we will look into the symbiosis between phages and bacteria. Phage, or bacteriophage, is the classification for viruses that reproduce via infection of bacteria. They are also the most abundant organisms in the world, with an estimated population of 10^{31} [10] in comparison the estimated abundance of bacteria which is 10^{30} [44], making phages a huge part of the genetic material in our biosphere.
Generally these viruses have two modes of existing within their host organism. This depends on whether the phages are in a lytic and lysogenic phase. In the lytic phase the genetic material of the virus is separate from the host cells genome, and uses the host cells replication tools to quickly multiply, which will eventually result in the host cell rupturing. This is therefore a fast cycle of existence, but phages can also be found in a lysogenic phase. Here the phages have incorporated their genome into the genetic material of the host cell, and exist dormant as prophages while being spread by the bacteria's own cell replication.
Phages are classified by their morpholical differences as well as their functions. We know from an article published in 2003 that out of over 5100 investigated

phages 96% are of the order Caudovirales [1]. Caudovirales are an order of tailed bacteriophages, where the structure of the tail separates the different families of Caudovirales [2]. The tail helps the injection of the phage genome into the host bacteria. Traditionally Caudovirales encompasses three families: Siphoviridae (tail is long and non-contractile), Podoviridae (has a short non-contractile tail), and Myoviridae (has a contractile tail). Now the order of Caudovirales encompasses more than 10 different families [3].

6.2 The influence of phages on microbiomes

While the prophages are dormant until the virus gets activated via chemical signal, they can still have influence on their host . Multiple studies have shown that they are known to alter the host metabolism [36], mold entire microbiome communities and lead microbial evolution [29]. Still there are a lot of unknowns in the bacteriophage's impact on the microbiomes.

6.3 Secondary metabolites classification

Bacteria can in some cases produce secondary metabolites. These can have multiple functions which can lead to a higher fitness of the organism that produces them. The metabolites are synthesised from biosynthetic gene clusters (BGCs), which are locally clustered genes, that have different structural and functional classes. These clusters can be identified based on their DNA-sequence.
A study from 2021 [22], describes the distribution of secondary metabolites a pool of ~190,000 bacterial genomes. The results show a distribution of 5 groups; $\sim$ 30% non-ribosomal peptide synthases (NRPS), which have a wide range of functions which include cytostatics, toxins, antibiotics and immunosupressants [14] [41]. $\sim$ 24% ribosomally synthesized and post-translationally modified peptides (RiPP) that have diverse structures with potential antimicrobial properties, where the bacteriocines in particular have antimicrobial toxins that inhibit cell growth [11]. [17]. $\sim$ 17% are polyketide synthases (PKS) whose products often have a range of bioactivities including antimicrobial, anticancer, or immunosuppressive abilities [16] [26]. Type I and II PKS are larger protein complexes [42], while type III PKS is smaller and produces a range of smaller aromatic molecules which are used as defense mechanisms in plants but also in microorganisms, giving potential antibacterial or antifungal properties [21].
$\sim$ 8% are terpenes which are one of the largest and most diverse groups of compounds often having biological activity, for example antimicrobial and anti-inflammatory abilities [39] [24]. Lastly $\sim$ 22% belong to "others", which as the name states, includes the vast variety of compounds that are not listed explicitly, or cannot accurately be categorised yet.

In addition a recent study from 2021 discovered the presence of phage-encoded BGCs (pBGCs), which are BGCs that are encoded in phage genomes and through these can be integrated into their bacterial hosts. These were found by screening 10,062 whole phage genomes in which 69 pBGCs were observed, thus

making them a rare occurrence most often occurring in prophages [12].
These secondary metabolites could be shared within bacterial communities, and then would be able to alter the dynamics within the microbiomes.

6.4 The parts of our pipeline

The focus of this study is to look at the potential existence of pBGCs in metagenomes, specifically the human gut metagenome which is chosen as a basis for our project. This viral metagenome is already available for further analysis, and the environment gives ground for medical applications as well as further research.

To explore this microbiome we can, on the basis of previous studies, use modern bioinformatical tools to detect and investigate pBGCs. We will briefly introduce these tools to give a better understanding of the process of our work.
Firstly we will be using antiSMASH, which will detect the presence of possible pBGCs, by using a hidden Markov model or hMM, which uses hidden internal states to predict a series of genes. With this it can identify gene clusters, which then indicate functions and potential products for the genes. An older study in 2011 has shown that antiSMASH had an accuracy of $\sim$ 97% [25]. Supporting the notion that we get a precise indication of the pBGCs in the metagenome, as well as their function.
We will also be using BIG-SCAPE, which by calculating sequence similarity networks, can group the pBGCs in gene cluster families (GCFs), linked by the presence of similar natural product chemotypes. A chemotype is a profile of the most produced chemicals in the organism, which can change drastically with only minor changes to the environment and epigenitics of the organism. This makes chemotypes very beneficial to detect BGC connections.
Hereafter we will be using vContact2, which clusters based on all protein clusters in the viral genomes, and compares them by inspecting their shared genes. This includes correcting for whether the connections between genes are coincidental. The precision was estimated to be $\sim$ 96% in predicting the taxonomy assignments [19].
We will also be applying BLAST (Basic Local Alignment Tool) which aligns sequences and calculates similarities, as well as compares them to databases to find the best match to help identify organisms, genes, or proteins [31].

7 Methods and materials

In the following section we will go through the pipeline of this study. These proceedings were run from February to May 2022, thus any updates to the of remaking of the results should check the versions of the programs. It should also be noted that dependencies for the programs might also have changed.

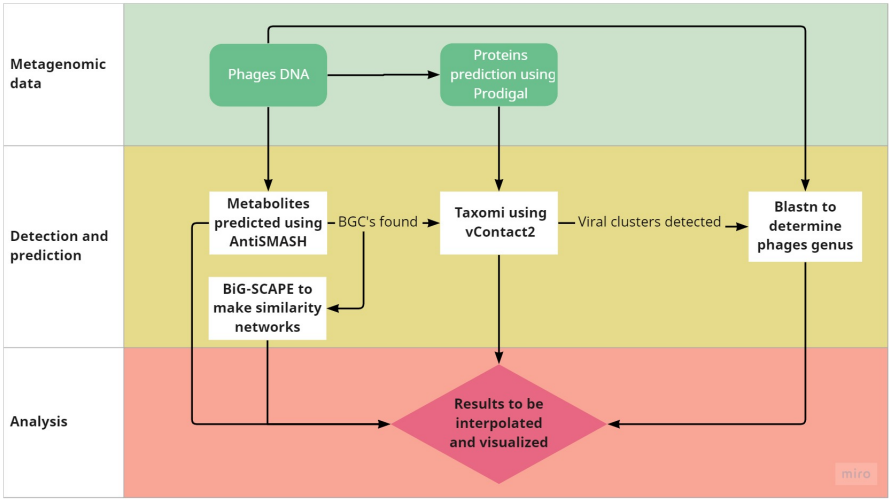


Figure (1) Flowchart [30] of the pipeline, with 3 distinct overall stages that our data undergoes. The arrows show our system's pathway, and how at each box the data is analyzed and processed. We want to thoroughly check and investigate the input, to get as many aspects compiled in our analysis. The green signifies the input, and the pink box symbolises the collection of outputs.

In figure 1 the overall work structure of the project is shown. We start with metagenomics data which is mined for pBGCs. The results are then further used to establish potential phylogenies and detect hosts, based on pre-established databases for reference. These results are then discussed and visualised.

7.1 Data collecting

In this study we used a previously sampled and sequenced metagenome[7].
This data has already been screened for viral sequences and quality tested. This particular study collected 28,060 human gut metagenomes sampled from around the world, as well as 2,898 isolated bacterial genomes. The sampled contigs were constructed using scaffolds using the software called SPAdes, which uses paired-end reads to assemble them.
Then VirFinder and VirSorter were used to distinguish and separate viral and bacterial contigs.
To provide quality control machine-learning classifiers were used to remove mobile genetic elements (MGEs). A criteria of 95 % average nucleotide

identity (ANI) was used to lessen redundancy and remove duplicates. From this $\sim$ 142,000 "non-redundant" viral genomes were detected [7].

7.2 Detection and analysis of pBGCs

These phage genomes were then read into antiSMASH 6.0.0 [5], to find the pBGCs across all of the genomes.
We used the database Prodigal [18] to identify genes, as the genomes are not annotated. AntiSMASH was run two times, first for all genomes with the minimal option, to lessen computational time. Then we ran the program with regular options for the genomes in which pBGCs were found, to get further information on placement and types of metabolites in the pBGCs.
For both the runs this process was optimized by parallelising singular genome analysis using GNU Parallel [38].
We then used three python scripts to clean the results. First we have a script, that deletes all folders that do not have any pBGCs, which will save storage and time. This is done by folder checking, and deleting all folders that contain only the automatically generated files. The next script is one which compiles all genomes containing the pBGCs into one file, this file is only used for input for other softwares. The last script creates a file containing a list of which phages have BGCs, and which pBGC type it is, by extracting the data from the Genbank files.
The results we gathered from antiSMASH were run through BiG-SCAPE. This took our predicted pBGCs, and calculated sequence similarity networks. With this we could look at different connections between the BGCs detected, and find possible distinct GCFs.

7.3 Clusters and phylogeny

To cluster for phylogeny and patterns in the phages we ran vContact2 0.9.19. We ran this software on all phage genomes annotated with pBGCs. To run this we first needed to generate the input files for vContact2. Initially to find the amino acid sequences in the genomes, we ran prodigal 2.6.3 [18], which ensures cohesive results with our antiSMASH results by using the same database. This is done, by the program calculating and finding the best scoring overlapping gene for each 3' end.
Secondly we ran vContact2's subscript vcontact2_gene2genome, which from the prodigal input creates the required mapping file, that associates genes to their original genome.
We chose not to use any database in the clustering, since it takes substantial less processing time, and we wanted to mainly look at connections between our found phages .
We then took the results from vContact2 and visualized the predicted clustering with Cytoscape 3.9.1 [35].

We used Nucleotide BLASTn [31], to determine the phages genus. We do this for phages in the viral clusters generated, as vContact2 should indicate strong taxonomic connections in these. To do this a phage database was needed, we used Millardlab april 2022 phage genomes [27].
From this we could predict the phages taxonomies on different levels, furthermore we could also attempt to predict the host microorganism, based on the same datatable from Millardlab. This could then be used to directly compared with the original studies detection of host bacteria, which was predicted with the detection of CRISPR spacers. Where these spacers are the immune system in bacteria, that store information on previous viral infections in spacers, that are separated in conserved regions [45]. We could also compare this to the study first discovering pBGCs and their hosts [12].

7.4 Statistical Analysis

When looking statistical analysis of our data we focused on analyzing regional differences. This was done using the original source data, which provides a summary of information for each sample, and which samples are used to construct the individual phage genomes.
To determine if these measured difference has any statistic value, we will lastly provide a two proportion Z-test, and measure the p-values for the possible combinations of these differences in the continents. For each two proportions compared we will set up the null hypothesis:

H_0 : The proportions of the two continents are the same.

We can by using this test compute p-values and if any p-value is bellow the threshold 0.05, we can reject the null hypothesis and accept the alternate hypothesis:

H_1 : The proportions of the two continents are statistically different.

To account for the high amount of hypothesis tested and to lower the false discovery rate, we have adjusted our p-values using Benjamini-Hochberg (BH) procedure.

8 Results and analysis

This section will strive to summarise our results as well as give some insight into the analyses that have been run on these results.

8.1 Mining of pBGCs

Using antiSMASH 872 BGCs were found from 852 phages, of them 53 % are prophages, according to predictions from the previous study via checkV [7]. This is an increase from the whole sample where prophages only constitute 23%, meaning that BGCs are much more commonly found in prophages than in freely existing viruses.

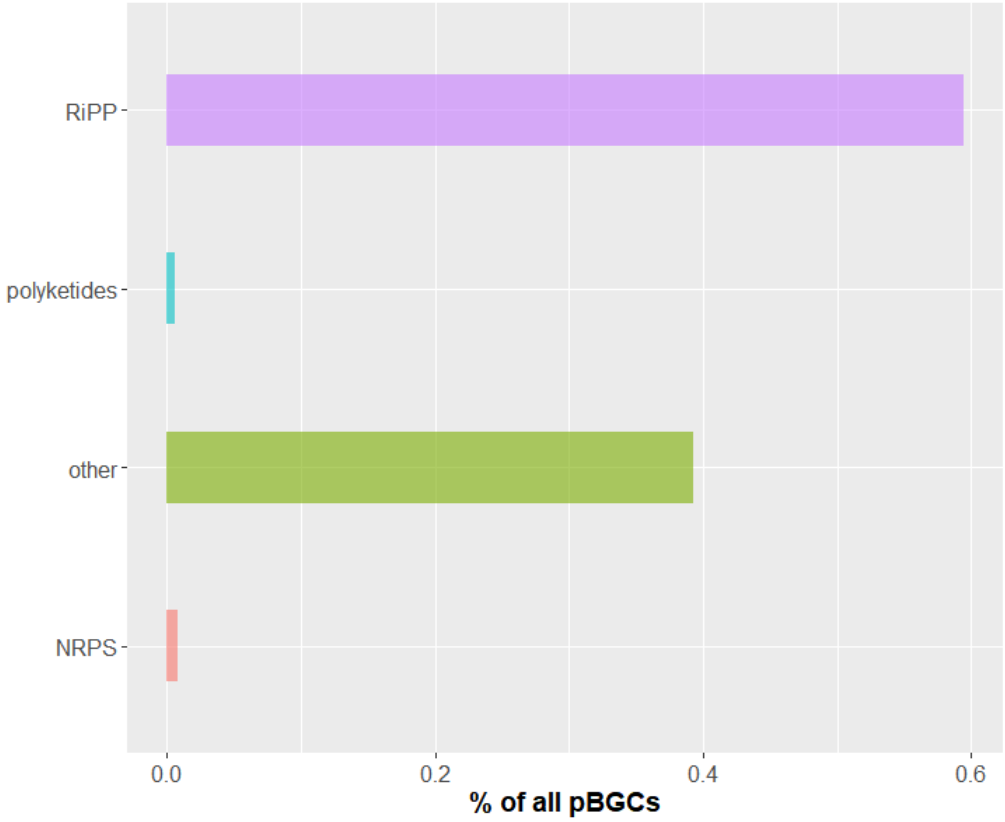


Figure (2) The relative distribution of BGCs in the gut phages. In total 872 BGCs were found, $\sim$ 60% of them are RiPP, $\sim$ 30% are "other" and the remaining 1% are split between polyketides and NRPs

From the detected pBGCs, figure 2 gives an overall look at the pBGCs clusters found in the gut phages. Here we can clearly see a stark difference between the

groups RiPP and "other" contra NRPs and polyketides, and terpenes were not detected at all.
Phage genomes are much smaller than their bacterial host's, often being 40-45 kb long [47], it is therefore not unexpected to see a distinct difference in the type of secondary metabolites that are found in pBGCs in comparison to the findings previously found in bacteria [22]. We see many RiPPs, which as stated are peptides, and cover a broad group of functions for metabolites. It is not surprising that these are the most abundant pBGCs.

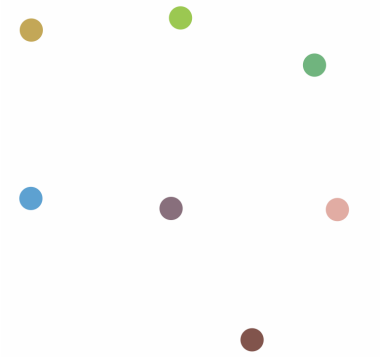
Table (1) Classification of the gut-phage's metabolites, all 872 detected pBGCs are spread across these four groups NRPS, polyketides, RiPPs and other. Here we can see the categorization of metabolites, and what subcategories of the pBGCs have been detected in our samples.

	Human gut flora
NRPs	
Nonribosomal peptide synthetase	5
NRPs like fragments	2
Polyketides	
Polyketide synthase type III	5
RiPPs	
Lantipeptide class I	4
Lantipeptide class II	2
Lassopeptide	3
Other unspecified RiPPs	389
Ranthipeptides	115
Sactipeptide	1
Thiopeptide	4
Others	
Other	9
Aryl polyene	21
Aminoglycoside/Aminocyclitol	1
β-lactone containing protease inhibitor	2
Cyclic lactone autoinducer peptides	267
Ladderane	3
Linear azol(in)e-containing peptides (LAP)	1
RRE-element containing	38
Terpenes	

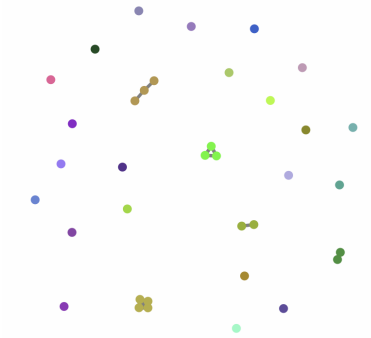
Table 1 is a compiled version of every pBGC that was detected, which gives a more detailed overview. The polyketides are our smallest sub-category, with only the occurence of the specific subclass of type III PKS.
NRPS, which we haven't found a lot of either, are larger enzymes which help

synthesize a variety of compounds.
RiPPs are our first larger category of our pBGCs, with the biggest sub-category found here being unspecified RiPPs. We know that RiPPs often display antimicrobial properties and show good stability from being modified, thus making this prevalence reasonable.
The other big group are the ranthipeptides, where studies have found that some ranthipeptides have quorum-sensing properties that help regulate cellular metabolism [9].
From the "others" group, besides those not identified, we also see a few frequent observations. We see a couple of aryl polyenes, which are fitness factors for host organisms, that can protect from oxidative stress and help biofilm formation [20], therefore serving a beneficial role for the host.
The RiPP recognition element (RRE) is a very important element, as they bind to RiPP precursor peptides, and directs them to their posttranslational modification enzymes, and thus essential with all the RiPPs observed [6].
Lastly we have another big group, the cyclic lactone autoinducer, which as the name indicates is an autoinducer. These are signal molecules, in this case cyclic lactones, that help control the gene expression in the quorum-sensing response pathways.

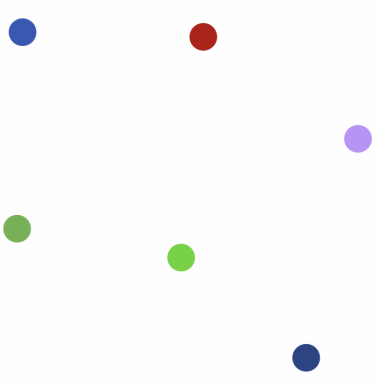
8.2 Phylogeny



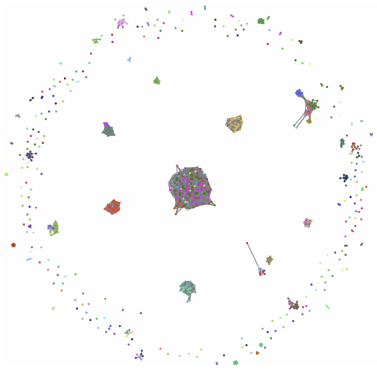
(a) GCF in our samples NRPS



(b) GCF in our samples other pBGCs types



(c) GCF in our samples polyketides



(d) GCF in our samples RiPPs

Figure (3) GCFs in each category of pBGCs, the categorization is different from table 1, thus the counts is not comparable. The colour of of each node in each plot indicate a family, and strings connecting indicate a clan.

From the results we can now look into the phylogeny between the pBGCs which done by creating a similarity measure using BIG-SCAPE, this can be seen in figure 3. Firstly it is important to note that BIG-SCAPE uses another way of categorizing the pBGCs, and there are therefore some distribution differences, as for example the cyclic-lactone autoinducers are classified as RiPPs.
For the 7 NRPs found (a), all where identified as their own separate family, and no linkage was therefore found. The same is true for the polyketides (c) where 6 pBGCs were found, each belonging in their own distinct family. From the other group (b) there were 36 pBGCs, with 27 families and of these 22 singletons, which as indicated makes these families really small (2-3 per family), making it difficult if not impossible, to see any patterns in the overall lineage.

For the RiPPs (d) we have 815 pBGCs distributed to 273 families, with only 188 being singletons. Here we see the first instances of multiple families in a clan, with the biggest GCF is for some RiPP-like pBGCs, which is distributed into 7 families and 199 individual pBGCs, and is much larger than the other families, which only has up to $\sim$ 30 pBGCs.

Figure (4) Cluster network of our pBGC containing phages, where every node represents a phage genome, and every edge is an identification of shared genes between genomes. The colour of the edges is an indication, of sequence similarity, based on the number of shared protein clusters, where yellow indicates low sequence identity, and red indicates high identity. Node sizes are a reflection of genome size. The colour of the nodes indicate the pBGC type, where those with more than one pBGC, only have one representational colour that indicates one of the types of pBGC.

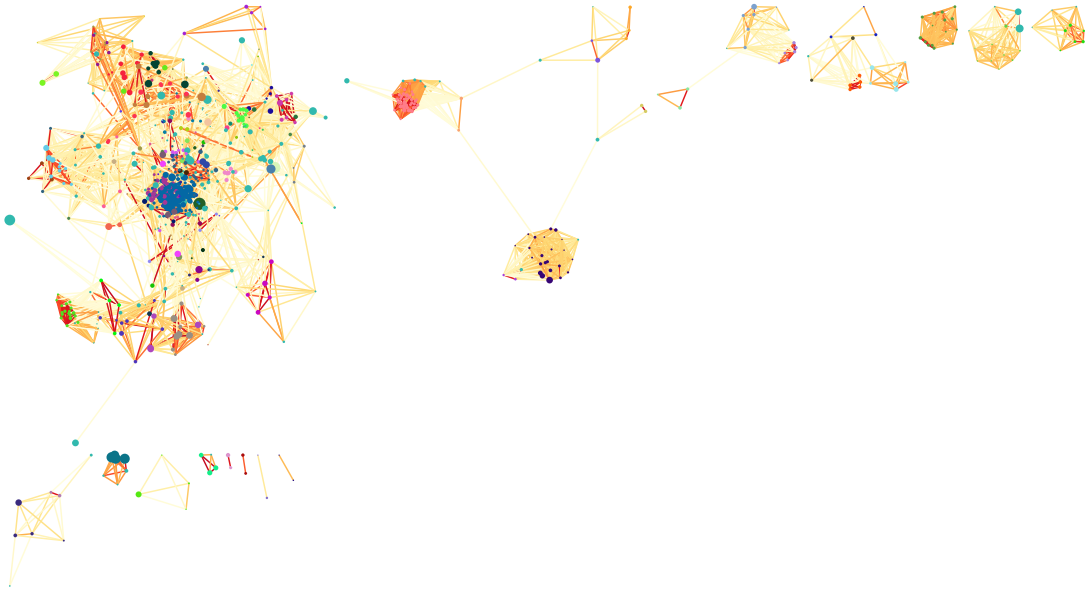


Figure (5) Same overall structure as figure 4, but now the nodes are coloured by clustering done by vContact2.

When we then look at the sequence identities between the phages, we should be able to identify the different identities in our samples. From running vContact2 we gathered 98 unique viral clusters. We have chosen to look into the initial viral cluster and not the sub-clusters, here the clusters should, with the intentions of vContact2, indicate the same genus or the same family for the genomes.
From figure 4 and 5, we have 662 genomes predicted to belong to any cluster, where we overall see one big group of connected nodes, and some smaller interconnected pools of nodes. We can then, by comparing the two networks, see that what connects these smaller groups of nodes is often their BGC type. Because when we look at figure 5 compared to 4, we see that for the viral clusters, not all the nodes are always the same colour, meaning they don't belong to the same viral cluster according to the vContact2 clustering. It could very well be that they do share some genes, as they do produce similar metabolites, but there isn't enough evidence to prove that members of a cluster belong to the same phage family.
In the broader perspective, we do still see correlation between viral group and pBGCs produced, where the nodes of similar pBGC type, can often be seen grouped together in a viral cluster.
Additionally we also see a lot of pale yellow edges, which indicate weaker connection between genomes, but there are also some darker edges between clusters which could very well indicate that we observe some clusters connected at genus level. We can also see that those with very red edges, often are

connected between nodes of similar size, and thus we don't see a big or bias based on size.

8.3 Phages and host bacteria

The following segment highlights our results from BLASTn. We have taken all the phages that vContact2 identified as being part of a viral cluster, and run their sequences through BLAST. It is important to note that this doesn't cover all of our phages, but only those where we would already assume have some relations. Furthermore we filtered the BLAST results, so that only alignment coverage over 2 kb was used, as to try to exclude as many false positive results. This is of course also dependent on genome length, but for simplicity we have only looked into the coverage length. Because of this we will not look to much in what genus type blast predicts, and mainly look at the higher taxonomic levels.

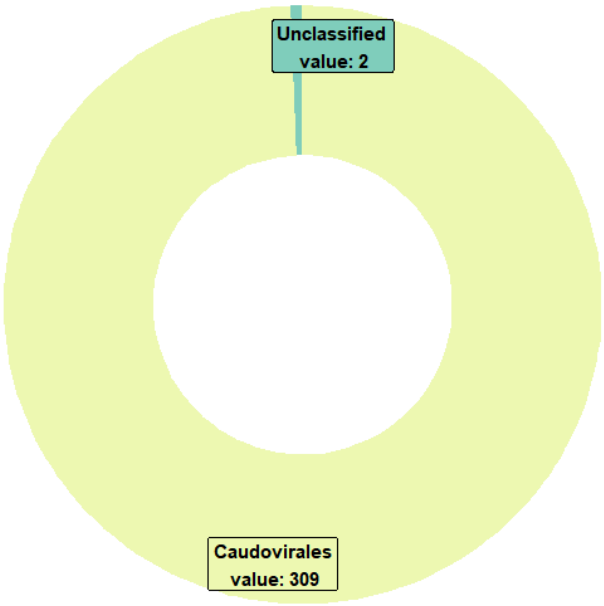


Figure (6) Order classification of phages from the viral clusters predicted, the vast majority being part of the order Caudovirales.

We can see in figure 6 that out of the phages in our viral clusters, the order Caudovirales is clearly the most abundant, while less than one percent of them were unable to be classified. As Caudovirales is such a predominant order seen

in phages, we see some validation of our results.

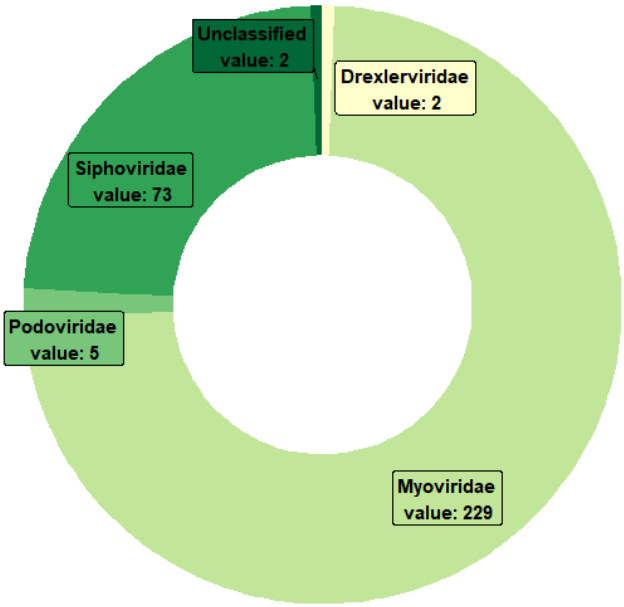


Figure (7) Expansion on figure 6, here we see all the same phages sorted by families. It can be seen that we have two dominating families, as well as multiple families that account for the rest.

In figure 7 we can now see the families in our viral cluster samples.
We can see that the largest families of vira are Siphoviridae $\sim$ 23 % and Myoviridae $\sim$ 74 %, with some few groups of other families. Myoviridae often having a larger head and with bigger genomes than other Caudovirales [23], making them likely to be able to carry bigger BGCs.
As previously stated, we wont go to much into genus type, but it is worth to note that 56% of the BLAST hits for these genomes as shown in figure 6 and 7, all came out as the genus Brigitvirus which is of the Myoviridae family.

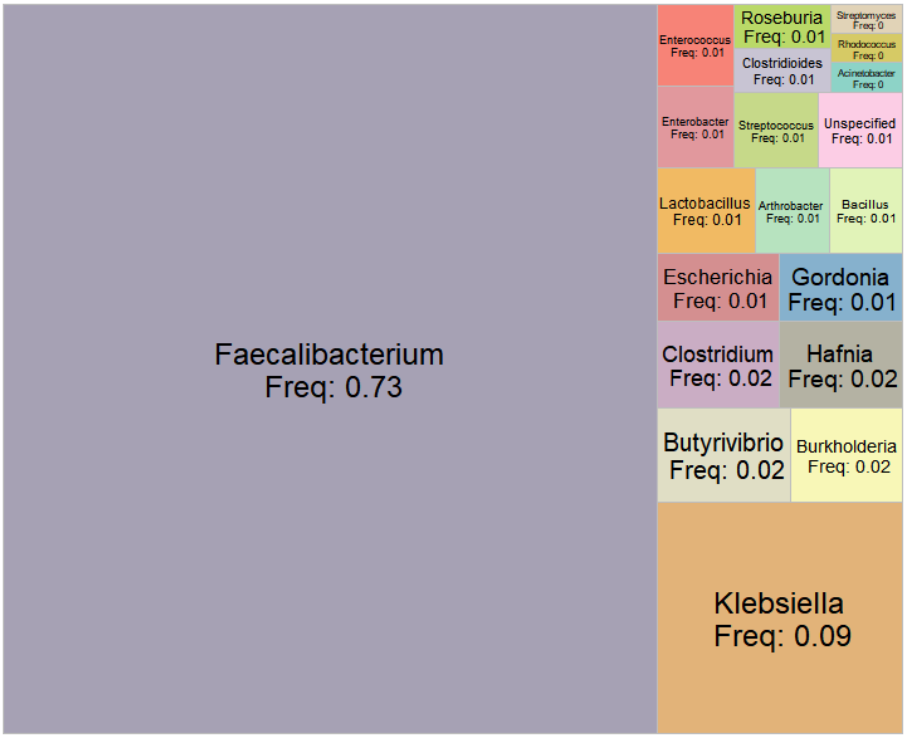


Figure (8) Predicted host of phages using BLAST results. Where the larger the frequency the bigger the box is, with the percentage of the total occurrences is presented as well.

In figure 8 we can see that Faecalibacterium accounts for 73% of the predicted hosts. When we compare our findings with the previous study's predictions of host organisms[7], for which they used CRISPR spacers, we do see some similarities. Because of the limitations of CRISPR spacers, only 29 of their predicted hosts, overlapped with what which phages we predicted the host for, as the remainder of their predictions were not for pBGCs encoding phages. But of those 29, 23 ($\sim$ 79%) gave the same predicted host bacteria as our predictions, validating our results, even though the sample overlap is small.

8.4 Regional pBGCs

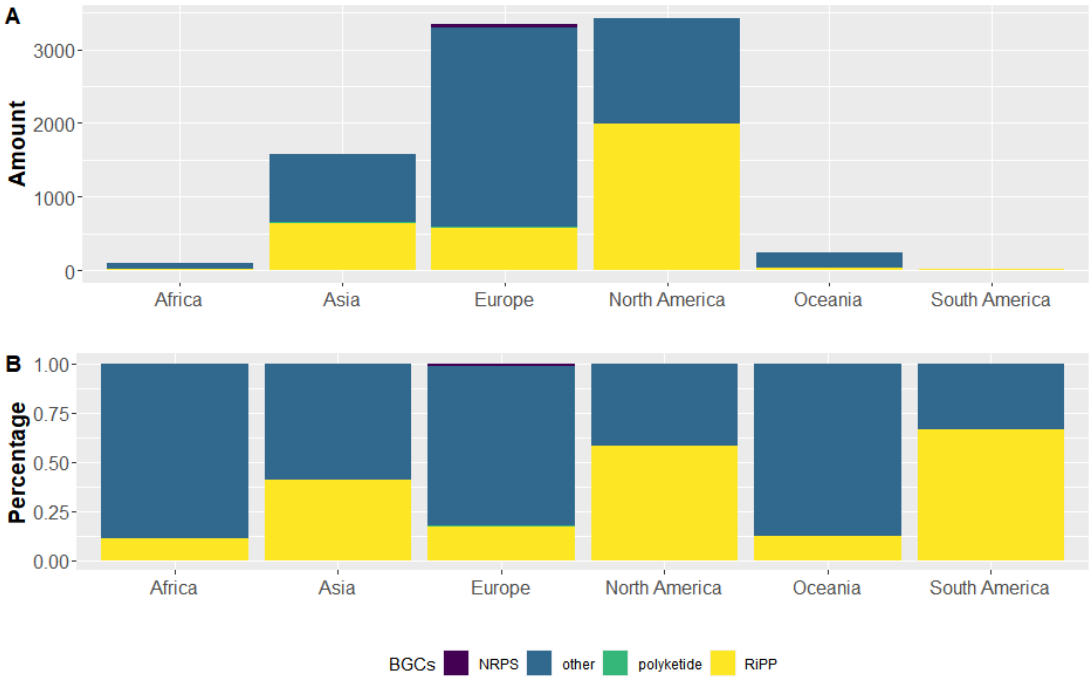


Figure (9) We can here observe the pBGCs found by continent, this does include phages that were found in multiple countries on the same continent, as well as multiple instances of the same phages found in samples in the same country. Thus the occurrences far proceeds the number of phages actually detected. The lower of the two graphs shows the fractional distribution of pBGCs per continent

In the metagenome that we work from, some genomes are not given any country of origin, this covers approximately 13% of them. The other genomes are found in up to 20 different countries within a lot of samples, although it should be noted we have no samples from Antarctica. This is observable in figure 9 (A) , the amount of phages here is skewed by the amount of samples taken. This still gives us a good overview of the origins of phages though. The lower graph (B) of the same figure gives us a better outlook into the similarities and differences between continents.
From graph (B) we cant see the existence of NRPs and polyketides, but they do exist, just on a much smaller scale. For NRPs we only observed 51 instances, where 48 of them coming from Europe.
Polyketides have only 24 instances observed sampled from Europe and Asia.

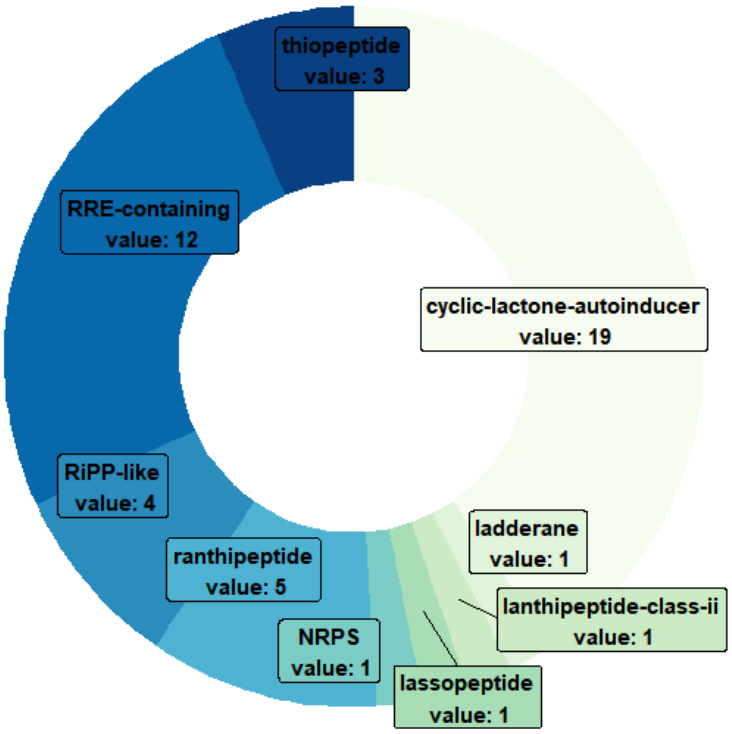


Figure (10) The most wide spread pBGCs types across the border, seen in 5 or more different countries, meaning pBGCs occuring in the same country multiple times only count as one occurrence.

The results in figure 10 show which pBGC types are most internationally conserved.
This only shows 47 pBGCs, because it is only these 47 pBGCs are found across 5 or more countries. We can see that the most conserved pBGCs are cyclic-lactone autoinducer and RRE pBGC. Although if we sum up the smaller categories, we observe a lot of peptides, meaning many RiPP's are also conserved globally.
We will then look at whether we actually measure any statically significant difference in figure 9. Here we will look at both whether there is a significance in the occurrence of pBGCs and whether there is a difference in the distribution of the types pBGCs found.

Table (2) The distribution of phages collected in region that contained pBGCs contra the amount of phages found in the same region. We then compare this difference, so that the element in each cell signifies the proportion factor to the continent of each row. A value > 1, means that the columns continent has a that much higher frequency of pBGCs than the region in the row.

	Africa	Asia	Europe	North A.	Oceania	South A.
World $_{bgc}$	**0.40**	**0.87**	0.93	**1.49**	**0.33**	**0.40**
Africa $_{bgc}$		**2.03**	**2.17**	**3.45**	NS*	NS
Asia $_{bgc}$	**0.49**		1.07	**1.70**	**0.38**	0.46
Europe $_{bgc}$	**0.46**	0.93		**1.59**	**0.36**	**0.43**
North A. $_{bgc}$	**0.29**	**0.59**	**0.63**		**0.22**	**0.27**
Oceania $_{bgc}$	NS	**2.61**	**2.80**	**4.45**		NS
South A. $_{bgc}$	NS	2.17	**2.32**	**3.69**	NS	

Notes: Abbreviations used in table.
* NS = non significant (p-value > 0.05)
BOLD Very strong evidence (p-value < 0.001)

Table (3) The distribution of NRPs detected in region contra the amount of pBGCs in that region. We then compare this difference, so that the element in each cell signifies the proportion factor that the continent of each row. So that a value > 1, means that the columns continents has that cell higher frequency of NRPs than the region in the row.

	Africa	Asia	Europe	North A.	Oceania	South A.
Africa $_{nrp}$		NS*	NS	NS	ND**	ND
Asia $_{nrp}$	NS		**11.33**	NS	NS	NS
Europe $_{nrp}$	NS	**0.09**		**0.02**	NS	NS
North A. $_{nrp}$	NS	NS	**49.28**		NS	NS
Oceania $_{nrp}$	ND	NS	NS	NS		ND
South A. $_{nrp}$	ND	NS	NS	NS	ND	

Notes: Abbreviations used in table.
** ND = non detected (NRPs where not detected in either continent)
* NS = non significant (p-value > 0.05)
BOLD Very strong evidence (p-value < 0.001)

Table (4) The distribution of RiPPs detected in region contra the amount of pBGCs in that region. We then compare this difference, so that the element in each cell signifies the proportion factor that the continent of each row. So that a value > 1, means that the columns continents has that cell higher frequency of RiPPs than the region in the row.

	Africa	Asia	Europe	North A.	Oceania	South A.
Africa $_{ripp}$		0.66	NS*	**0.47**	NS	NS
Asia $_{ripp}$	1.51		**1.38**	**0.71**	**1.49**	NS
Europe $_{ripp}$	NS	**0.73**		**0.52**	NS	NS
North A. $_{ripp}$	**2.12**	**1.40**	**1.93**		**2.09**	NS
Oceania $_{ripp}$	NS	**0.67**	NS	**0.48**		NS
South A. $_{ripp}$	NS	NS	NS	NS	NS	

Notes: Abbreviations used in table.
* NS = non significant (p-value > 0.05)
BOLD Very strong evidence (p-value < 0.001)

Table (5) The distribution of "Others" detected in region contra the amount of pBGCs in that region. We then compare this difference, so that the element in each cell signifies the proportion factor that the continent of each row. So that a value > 1, means that the columns continents has that cell higher frequency of "Others" than the region in the row.

	Africa	Asia	Europe	North A.	Oceania	South A.
Africa $_{other}$		**3.71**	NS*	**5.28**	NS	**6.06**
Asia $_{other}$	**0.27**		**0.42**	**1.42**	**0.31**	NS
Europe $_{other}$	NS	**2.38**		**3.39**	NS	**3.89**
North A. $_{other}$	**0.19**	**0.70**	**0.29**		**0.22**	NS
Oceania $_{other}$	NS	**3.25**	NS	**4.63**		**5.31**
South A. $_{other}$	**0.17**	NS	**0.26**	NS	**0.19**	

Notes: Abbreviations used in table.
* NP = non significant (p-value > 0.05)
BOLD Very strong evidence (p-value < 0.001)

Table (6) The distribution of polyketides detected in region contra the amount of pBGCs in that region. We then compare this difference, so that the element in each cell signifies the proportion factor that the continent of each row. So that a value higher than 1, means that the columns continents has that cell higher frequency of polyketides than the region in the row.

	Africa	Asia	Europe	North A.	Oceania	South A.
Africa $_{polyke}$		NS*	NS	ND**	ND	ND
Asia $_{polyke}$	NS		NS	0***	NS	NS
Europe $_{polyke}$	NS	NS		0	NS	NS
North A. $_{polyke}$	ND	Inf***	Inf		ND	ND
Oceania $_{polyke}$	ND	NS	NS	ND		ND
South A. $_{polyke}$	ND	NS	NS	ND	ND	

Notes: Abbreviations used in table.

*** 0, Inf as only one of the two continets contain polyketides the distribution factor is equal to infinity.

** ND = non detected (Polyketides where not detected in either continent)

* NP = non significant (p-value > 0.05)

BOLD Very strong evidence (p-value < 0.001)

The previous tables show us proportional differences between continents, all proportions shown are calculated to be statistically significant. Although one should still remember that this is still somewhat skewed by the amounts of data collected in each continent.
This is especially clear when taking into account the small sample sizes from Africa, Oceania and South America, where we have not sampled any polyketides or NRPs. Furthermore for the polyketides we do not see any valid factor, as we only see polyketides in Europe and Asia, where there is no significant difference. But although the factor is not one to hold as the precise difference, there is seen a significant different between the two continents and North America, as we here have no polyketides observed.

9 Discussion

We can from our AntiSMASH results clearly see, that while not particularly abundant in our collection of phages, we definitely do have pBGCs in the human gut metagenome and therefore theoretically in each and every one of us. As mentioned we have more pBGCs in prophages than in free virusses, this is because not all vira have a lysogenic stage, and it could be assumed that the ones that the ones that do not, have no interest in furthering the survival of their host. This in itself is already very intriguing to know, because it leads into a lot of potential considerations, as well as research that we could barely even begin to outline in this paper.
But we can look closer into the specifics of the pBGCs that we did find, and consider what implications they have, both in their existence in the microbiome and in what evolutionary benefit it might be to the phages that carry them. Starting from the smaller groups, we found 7 NRPS see table 1 which isn't a lot out of our samples.
One consideration that is initially important is the fact that our phage genomes are somewhat limited in size, and NRPS can be very large. It could therefore be contemplated whether some of these are in fact false positives and not actually phage encoded BGCs. When individually inspecting these NRPS and the NRPS like fragments we can see that they range in size from 2 kb to around ∼13 kb. The lower end of those is very plausible to be phage-encoded, and even the higher end could theoretically be encoded in some of the larger Myoviridae. But why would a phage even carry around a huge NRPS domain?
While this isn't a question that we can fully answer without more detailed insight into the NRPS found in phages, or without a bigger sample of these, we could reasonably consider whether there is a connection with the synthesis of beta-lactam antibiotics, which require NRPS and would add significant evolutionary advantage to the host. Beta-lactam antibiotics primarly work against gram-positive bacteria, which again leads us to the fact that the majority of our phages' hosts are gram-negative [43].
While it is interesting that we have detected the presence of polyketides in our metagenome, there isn't a lot to analyse from this, as they aren't very abundant, and not all that much is even known about the effects of microbial type III PKS's. But it is interesting that they do exist, and with further research into the field it might become clearer why phages would carry them around, and whether they truly do give ground to antifungal or protective properties[21].
RiPPs are the first of our larger categories, and as it was previously mentioned, these cover many a property including many antimicrobials. As the majority of our RiPPs are unspecified we can not go into the specifics of these. But this type of secondary metabolite is probably so well established within our sample because it conveys a potentially strong evolutionary advantage to the host, for example by equipping the bacteria with weapons against other species [12]. The second largest subtype of RiPPs that we observed are Ranthipeptides, these are assumed to be quorum sensing type molecule. It is interesting the we observe the quorum sensing molecules because it raises the question whether these can

be shared between species and colonies through phages, potentially leading to inter-species communication in microbial communities [9].
Within the "others" category of pBGCs we again observed some different sub-categories the majority to of them being in one way or another connected to fitness factors or anti-oxidants which directly improve the conditions for the host, therefore making good sense to find in our pBGCs. A smaller, but yet present, sub-category is the RRE-containing element which interacts with RiPPs. It could be interesting to look more closely into these in the future to analyse the interaction for the potential to manipulate hosts, both in regards to host composition but also in regards to output RiPPs. We also observed some quorum-sensing pathway molecules, namely cyclic-lactone autoinducers.
This again reiterates that the phages have some sort of advantage in controlling the proliferation rate, changing the behaviour of their hosts, or supporting inter-bacterial communication. Whatever the reason, considering how common this motif seems to be, there must be a clear biological advantage for the phages to be carrying these around. This could give grounds for phage based manipulation of microbial communities, whether it be for disease prevention, or creation of advantageous microbial gut communities.
We then looked into the relations between these phages we first looked into using BiG-SCAPE. According to this there were no relations in the two smaller categories of pBGCs that we observed, but we had a little bit of relations in the "others" and quite a lot of relations in RiPPs. For example in figure 3 (d) we can see the large central cluster of related phages, which contain RiPP-like pBGCs. It is interesting that these in particular seem to be so interconnected when there is less interconnection between other subgroups of RiPPs.
But we didn't base our analysis of familial relations purely on BiG-SCAPE. From the vContact2 clustering we can see that in the two versions of clustering, figure 5 and figure 4, that the relations seem to be similar, both when clustering after viral identity or pBGC type. This seems to indicate that phages expressing similar types of BGCs might be evolutionary closer related to each other than to the others. But to have proper conclusion on this it would need further analysis into the phylogeny and species of the phages.
When we looked further into the phylogeny based BLAST we can see that the majority of our phages, with enough sequence coverage to be considered, were observed to be Caudovirales. Looking towards the families we get a little more of split, where we see a majority of Myoviridae. This observation makes a lot of sense considering that these phages have the size to contain larger domains and therefore shouldn't have trouble containing even some of the bigger pBGCs that we've observed. This consideration is furthermore made likely by the knowledge that Myoviridae in general are more rare then Siphoviridae, meaning that they in our data are overrepresented, probably due to their increased size and therefore ability to transport pBGCs [2]. While there is not much known about it, the majority of our phages looked to be in the genus Brigitvirus, which might be interesting to further investigate in connection with hosts for the phages.

The majority of the hosts were predicted to be in Faecalibacterium, which makes sense considering our biome. We did not have a lot of overlap in which phages we could predict hosts for when comparing to the original study [7], but out of these the majority resulted in the same host. Considering that we used very different prediction methods, this lends some credibility to our results.
We also compared our hosts to the 2021 study first describing pBGCs [12]. When comparing here we can see that the predictions vary, although there is some overlap, for example Escherichia, Lactobacillus and Bacillus being represented quite a bit more in this study than in our predictions. There are also quite a few species not represented in both of the samples. This is probably caused by the variation of the data origin with our study being based on a metagenome and therefore containing a lot of unidentified phages with hypothetical hosts, and the origin study using complete genomes. Another reason for the differences could be the origin of the metagenome being the gut, and therefore containing many bacteria more specific to this environment, which naturally will be over-represented.

It is very interesting that we predict so many Faecalibacterium as our hosts, as these are bacteria which seem to have positive effects on the human gut, providing a certain degree of protection against different diseases of the bowel, such as IBD(Irritable Bowel Disease) and ulcerative collitis [40][37] [34] [33]. This leads to us to consider possibilities in the prevention of such diseases by manipulating the population of phages that use these bacteria as hosts, to give the hosts more favourable traits, such as competitive bacteriocines or quorum sensing molecules that could increase growth rates or communication with neighbouring similar communities. Similar considerations could also be made for some of the other beneficial bacterial hosts, such as the Lactobacillus.
The second most frequent host is Klebsiella, which is a pathogenic gram-negative bacteria, causing many dangerous infections [8]. Surprisingly many of the predicted hosts are pathogenic, including Clostridium, Hafnia, Arthrobacter and Streptococcus. It could be debated whether some of these rarer and very dangerous bacteria are overreprented as hosts, because they become more pathogenic through their pBGCs. Quorum sensing pBGCs could further communication between colonies of any of these, or help trigger pathogenic lifecycles. Not to speak of the fact that antibiotics resistance might be spread through phages as well. This is also interesting, because many of these are gram-negative as well, which are known to be very hardy towards antibiotics[32].

We also looked at the distributions of our phages in a geographical sense, with this being skewed by the sampling rate at the different locations. In figure 9 we can see that it is either RiPPs or ”other” that dominate the distribution of pBGCs in an individual continent. This makes sense considering that these were the most commonly observed categories. For Europe, Africa, and Oceania ”other” were more abundant, while Asia, and both the Americas have a higher occurrence of RiPPs. The considerable frequency of RiPPs is probably because of their variety of pro-host properties making it very reasonable to find in so

many different versions as well as having regional variants. This would makes sense when looking at how interconnected the familial relations of our RiPPs were.
We can further prove this point by looking at figure 10 where we can see that over half of the most conserved pBGCs across multiple countries are either peptides which are RiPPs, RiPP-like, or RRE-containing which is an " others", meaning that they all concern RiPPs in some way, even when categorised as "others". The remainder of the most conserved pBGCs are cyclic-lactone-autoinducers which as previously mentioned are quorum sensing molecules. It could be contemplated whether these are preserved across many countries as they are specific to particular types of hosts and therefore are similar if not the same, in countries where their host occurs. In connection to these its worthwhile to yet again debate whether manipulation of quorum sensing molecules could have potential in the treatment and prevention of gut based diseases by controlling their host populations behaviour.

When looking at the differences between the different continents, their pBGCs, and the world in general, we can draw some conclusions from the proportions that we calculated and their statistical significance. In table 2 we can see there are significant differences between the amount of pBGCs found between each continent, as well as each continent and the world, with the exception of the comparisons between the 3 continents with the smallest sample sizes. From this it would seem that sample sizes still have a big impact on this. But again it appears that there is a statistically significant difference in the amount of pBGCs found in different continents. With a closer look at this, it looks like the proportion of pBGCs in the phages for the continents ranges from 1.8% to 0.4% in the order NA, EU, As, AF, SA, OC. The high percentage of pBGC containing phages found in especially in Northern America, Europe and Asia is also interesting because, apart from the amount being statistically significantly higher than the rest of the world, it also is a higher percentage of pBGCs found when comparing to the previous paper which first described pBGCs [12]. This makes us consider whether pBGCs might be overrepresented in the gut compared to other microbiomes, but also that many unidentified phages contain pBGCs.
Looking more closely at the individual pBGCs in the continents we can see that for the two rare pBGC types we can observe statistically significant differences for the 3 continents with the biggest sample sizes. In the case of the polyketides this is because Asia and Europe did find polyketides, while Northern America did not. Still considering the rarity of polyketides it would be more reasonable to get a bigger sample-size before concluding that polyketides don't exist in Northern America, but the tendency can be observed. A similar situation can be observed with the NRPS but here it is Europe being the outlier in comparison to the other two, with vastly more NRPS, although all three did detect NRPS. Again it would be difficult to conclude something about these without a bigger sample size, but the tendency for a difference in composition is there, as Europe has the vast majority of found NRPS.

Now looking at our two abundant groups, we can observe some statistically significant differences which would lead us to conclude that there is in fact a difference in not only the amount of pBGCs found in different continents but also in their composition, further cemented by our findings that RiPPs were more common in Northern America for example, while "others" were more frequent in Europe. This leads us to the conclusion that, at least within the bounds of our data, the composition of pBGCs is statistically different from continent to continent, probably based on the environmental differences in the human gut.

10 Perspectivation

In this project we mainly worked with trying to uncover pBGCs in the gut metagenome, where we mainly wanted to look into the compositional elements and whether any patterns could be observed either in the genomes or regionally. As previously mentioned this is an extremely new phenomenon, not much data is known about pBGCs. Thus this research gives a brief introduction, and works as springboard for further examination of the sequence structure of phages' secondary metabolites, to broaden the understanding of the interactions in our microbiomes.
To accomplish our goals we have made a lot of direct and indirect choices, and of such there is much room to explore other options. Furthermore there were restrictions on some of our choices that where determined based on our limitations in time and resources, and of course our own focus in this report.
Some of the things we did not get to do, for example, was to look into the domains in the pBGCs, to further understand what the specific pBGCs are apart from their overall classifications. This could lead to some interesting discoveries, but would also require some extensive work.
Another options we didn't get to do was adding the phage database used in our BLASTn, in our vContact2, which would have caused our analysis of the viral clusters to be directly associated with the phages. This could lead to some interesting dynamics in our samples, as we then would see if we would get the same distribution of viral clusters. Comparing the clustering based only on sequence similarity to clustering based on sequence similarity to the phages in the database.
We would also have liked to compare our gut metagenome phages to other microbiome's metagenomes. With this, and enough time and data, we could start linking pBGCs to their respective environments and read deeper into their effects on the microbiome. But alas we could not, within the timeframe, find another sizeable viral metagenome to analyse, nor did we have the time to compose one ourselves. We would also like to further increase the sample size in the human gut metagenome, and with that further validate or disprove our results.
Another direction we would have liked to go further into is the potential applications of our research within the manipulation of the human gut microbiome for medicinal reason, as well as exploration of the hosts and the effects of the pBGCs on these. One example of this would be to compare obesity and diabetes statistics in different continents to the pBGC contents and see whether we can observe a connection or pattern.
With this said, these options would be supplementary to our research, had we had more time, as we did reach the answers we were grasping for. It would be interesting to further deepen the understanding of this topic in the future though, and it is our hope that this research helps pave the way to that.

11 Conclusion

Through the process of our work we have learned many things, we did in fact discover pBGCs within the human gut metagenome. These pBGCs turned out to be quite rare, only occurring in a very low percentage of phages, but covered a wide array of different properties and types. Some of the types of pBGCs that we found were way more common than others, but this differentiated from region to region of the world. In future work one could look more closely into these regional differences and try to supplement the findings with more data. We were also able to establish some preliminary networks of these phages, based on both protein similarity and pBGC type. These networks, although somewhat different have a lot similarities which gave base to a rudimentary understanding of potential phylogenies. The fact that these connections exist could lead to research based on identifying phages, their familial connections, and investigations into the pBGCs connected to particular genii.
We also got to start finding the phylogenies of some of the phages, leading to an understanding of the type of vira carrying pBGCs. Therefore giving us directions in which we could research in the future, as well as considering which sorts of hosts we could observe and therefore which organisms are most affected by the phages. These predictions of hosts gave us the opportunity to consider potential effects and applications of pBGC encoding phages in connection with gut health. So within the scope of our work, we did find pBGCs, we did rudimentary grouping of them and identified potential families and hosts. We also found statistically significant differences in the compositions of pBGCs found in the different continents, and have therefore found some very interesting points to further investigate in the yet unknown world of pBGCs.

12 Availability

12.1 Data availability

The gut microbiome phages database, that we have used can be accessed by the link: http://ftp.ebi.ac.uk/pub/databases/metagenomics/genome_sets/gut_phage_database/. Where we specifically chose the file GPD_sequences.fa.gz, that has the entire sequences of all the phage genomes and GPD_metadata.tsv that contain information for each genomes used in plots.
The phage database used is found under 1Apr2022_genomes.fa at: http://millardlab.org/bioinformatics/bacteriophage-genomes/phage-genomes-mar2022/. Here we also used the file phage_data.tsv that is used to convert phages accession codes to the actual phage types.

12.2 Code availability

Here are the programs and the accession of software's used, several different packages was also used, which was dependencies for the programs listed below.
To separate the genomes UCSC Fasplit was used: https://anaconda.org/bioconda/ucsc-fasplit. For the BGCs mining we used antiSMASH 6.0 which can be accessed though: https://github.com/antismash/antismash.
The gene prediction was run with prodigal v 2.6.3: https://github.com/hyattpd/Prodigal
The pBGCs phylogeny was made with BIG-SCAPE: https://bigscape-corason.secondarymetabolites.org/index.html
The taxonomic classification of phages though vContact2 0.9.1 can be downloaded through: https://bitbucket.org/MAVERICLab/vcontact2/src/master/
Multiple process were run with GNU Parallel: https://www.gnu.org/software/parallel/
The full code used in the making of this study is found at the following site: https://github.com/CarinaNK/BGCs-evaluation

References

[1] H.-W. Ackermann. Bacteriophage observations and evolution. *Research in Microbiology*, 154(4):245–251, 2003.

[2] H.-W. Ackermann. Tailed bacteriophages: The order caudovirales. *Advances in Virus Research*, 51:135–201, 2008.

[3] E. M. Adriaenssens, M. B. Sullivan, M. Krupovic, P. Knezevic, L. J. van Zyl, B. L. Sarkar, B. E. Dutilh, P. Alfenas-Zerbini, M. Łobocka, Y. Tong, J. R. Brister, A. I. M. Switt, J. Klumpp, R. K. A. J. Barylski, J. Uchiyama, R. A. Edwards, A. M. Kropinski, N. K. Petty, M. R. J. Clokie, A. I. Kushkina, V. V. Morozova, S. Duffy, A. Gillis, J. Rumnieks, İpek Kurtböke, N. Chanishvili, L. Goodridge, J. Wittmann, R. Lavigne, H. B. Jang, D. Prangishvili, F. Enault, D. Turner, M. M. Poranen, H. M. Oksanen, and M. Krupovic. Taxonomy of prokaryotic viruses: 2018-2019 update from the ictv bacterial and archaeal viruses subcommittee. *Archives of Virology*, 164:1253–1260, 2020.

[4] M. Arumugam, J. Raes, E. Pelletier, D. L. Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W. M. de Vos, S. Brunak, J. Doré, M. Consortium, J. Weissenbach, S. D. Ehrlich, and P. Bork. Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180, 2011.

[5] K. Blin, S. Shaw, A. M. Kloosterman, Z. Charlop-Powers, G. P. van Weezel, M. H. Medema, and T. Weber. antismash 6.0: improving cluster detection and comparison capabilities, 2021.

[6] B. J. Burhart, G. A. Hudson, K. L. Dunbar, and D. A. Mitchell. A prevalent peptide-binding domain guides ribosomal natural product biosynthesis. *Nature Chemical Biology*, 11:564–570, 2015.

[7] L. F. Camarillo-Guerrero, A. Almeida, G. Rangel-Pineros, R. D. Finn, and T. D. Lawley. Massive expansion of human gut bacteriophage diversity. *Cell*, 184(4):1098–1109, 2021.

[8] Centers for Disease Control and Prevention. Klebsiella pneumoniae in healthcare settings, Nov 2010.

[9] Y. Chen, Y. Yang, X. Ji, R. Zhao, G. Li, Y. Gu, A. Shi, W. Jiang, and Q. Zhang. The sciff-derived ranthipeptides participate in quorum sensing in solventogenic clostridia. *Biotechnology Journal*, 15(10):2000136, 2020.

[10] A. M. Comeau, G. F. Hatfull, H. M. Krisch, D. Lindell, N. H. Mann, and D. Prangishvili. Exploring the prokaryotic virosphere. *Research in Microbiology*, 159(5):306–313, 2008.

[11] P. D. Cotter, C. Hill, and R. P. Ross. Bacteriocins: developing innate immunity for food. *Nature Reviews Microbiology*, 3(10):777–788, 2005.

[12] A. Dragoš, A. J. Andersen, C. N. Lozano-Andrade, P. J. Kempen, Ákos T. Kovács, and M. L. Strube. Phages carry interbacterial weapons encoded by biosynthetic gene clusters. *Current Biology*, 31(16):3479–3489, 2021.

[13] S. H. Duncan, G. L. Hold, H. J. M. Harmsen, C. S. Stewart, and H. J. Flint. Growth requirements and fermentation products of fusobacterium prausnitzii, and a proposal to reclassify it as faecalibacterium prausnitzii gen. nov., comb. nov. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTION MICROBIOLOGY*, 52(6):2141–2146, 2002.

[14] R. Finking and M. A. Marahiel. Biosynthesis of nonribosomal peptides1. *Annu Rev Microbiol*, 58(1):453–488, 2004.

[15] S. R. Gill, M. Pop, R. T. DeBoy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson. Metagenomic analysis of the human distal gut microbiome. *Science*, 312(5778):1355–1359, 2006.

[16] J. B. Harborne and C. A. Williams. Advances in flavonoid research since 1992. *Phytochemistry*, 55(6):481–504, 2000.

[17] G. A. Hudson and D. A. Mitchell. Ripp antibiotics: Biosynthesis and engineering potential. *Current opinion in microbiology*, 45:61–69, 2018.

[18] D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(119), 2010.

[19] H. B. Jang, B. Bolduc, O. Zablocki, J. H. Kuhn, S. Roux, E. M. Adriaenssens, J. R. Brister, A. M. Kropinski, M. Krupovic, R. Lavigne, D. Turner, and M. B. Sullivan. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature Biotechnology*, 37(6):632–639, 2019.

[20] I. Johnston, L. J. Osborn, R. L. Markley, E. A. McManus, A. Kadam, K. B. Schultz, N. Nagajothi, P. P. Ahern, J. M. Brown, and J. Claesen. Identification of essential genes for escherichia coli aryl polyene biosynthesis and function in biofilm formation. *npj Biofilms and Microbiomes*, 7(56), 2021.

[21] Y. Katsuyama and Y. Ohnishi. Chapter sixteen - type iii polyketide synthases in microorganisms. In D. A. Hopwood, editor, *Natural Product Biosynthesis by Microorganisms and Plants, Part A*, volume 515 of *Methods in Enzymology*, pages 359–377. Academic Press, 2012.

[22] S. A. Kautsar, K. Blin, S. Shaw, T. Weber, and M. H. Medema. Big-fam: the biosynthetic gene cluster families database. *Nucleic Acids Research*, 49(1):490–497, 2021.

[23] A. M. King, M. J. Adams, E. B. Carstens, and E. J. Lefkowitz. Part i. introduction. In *Virus Taxonomy*, pages 1–20. Elsevier, San Diego, 2012.

[24] N. A. Mahizan, S.-K. Yang, C.-L. Moo, A. A.-L. Song, C.-M. Chong, C.-W. Chong, A. Abushelaibi, S.-H. E. Lim, and K.-S. Lai. Terpene derivatives as a potential agent against antimicrobial resistance (amr) pathogens. *Molecules*, 24(14), 2019.

[25] M. H. Medema, K. Blin, P. Cimermancic, V. de Jager, P. Zakrzewski, M. A. Fischbach, T. Weber, E. Takano, and R. Breitling. antismash: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic acids research*, 39(2):339–346, 2011.

[26] E. Middleton, C. Kandaswami, and T. C. Theoharides. The effects of plant flavonoids on mammalian cells:implications for inflammation, heart disease, and cancer. *Pharmacological Reviews*, 52(4):673–751, 2000.

[27] Millardlab. Phage genomes – mar2022. http://millardlab.org/bioinformatics/bacteriophage-genomes/phage-genomes-mar2022//.

[28] M. B. Miller and B. L. Bassler. Quorum sensing in bacteria. *Annual Review of Microbiology*, 55(1):165–199, 2001. PMID: 11544353.

[29] S. Mills, F. Shanahan, C. Stanton, C. Hill, and A. C. R. P. Ross. Movers and shakers: influence of bacteriophages in shaping the mammalian gut microbiota. *Gut microbes*, 4(1):4–16, 2013.

[30] Miro. Miro: The visual collaboration platform for every team. https://miro.com. Accessed: 2022-04-28.

[31] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 44(D1):D7–D19, 2016.

[32] J. Oliveria and W. C. Regaert. Gram negative bacteria. https://www.ncbi.nlm.nih.gov/books/NBK538213/. Accessed: 2022-05-09.

[33] ScienceDirect. Burkholderia. Accessed: 2022-05-12.

[34] ScienceDirect. Hafnia alvei. Accessed: 2022-05-12.

[35] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[36] J. D. Smet, M. Zimmermann, M. Kogadeeva, P.-J. Ceyssens, W. Vermaelen, B. Blasdel, H. B. Jang, U. Sauer, and R. Lavigne. High coverage metabolomics analysis reveals phage-specific alterations to pseudomonas aeruginosa physiology during infection. *The ISME Journal*, 10(8):1823–1835, 2016.

[37] Statens Serum Institut. Clostridioides difficile-infektion (cdi). Accessed: 2022-05-12.

[38] O. Tangen. Gnu parallel 2018. *In GNU Parallel 2018*, page 112, 2018.

[39] R. J. S. Vega, N. C. Xolalpa, A. J. A. Castro, C. P. González, J. P. Ramos, and S. P. Gutiérrez. *Terpenes from Natural Products with Potential Anti-Inflammatory Activity*. IntechOpen, Rijeka, 2018.

[40] T. Vuckovic. Beneficial bacteria: A focus on faecalibacterium prausnitzii, May 2021.

[41] K. J. Weissman and P. F. Leadlay. Biosynthesis of nonribosomal peptides1. *Nature Reviews Microbiology*, 3(12):925–936, 2005.

[42] J.-K. Weng and J. P. Noel. Chapter fourteen - structure–function analyses of plant type iii polyketide synthases. In D. A. Hopwood, editor, *Natural Product Biosynthesis by Microorganisms and Plants, Part A*, volume 515 of *Methods in Enzymology*, pages 317–335. Academic Press, 2012.

[43] M. J. Wheadon and C. A. Townsend. Evolutionary and functional analysis of an nrps condensation domain integrates β-lactam, d-amino acid, and dehydroamino acid synthesis. *Synthetic and Systems Biotechnology*, 7(2):677–688, 2022.

[44] W. B. Whitman, D. C. Coleman, and W. J. Wiebe. Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences*, 95(12):6578–6583, 1998.

[45] F. Wimmer and C. L. Beisel. Crispr-cas systems and the paradox of self-targeting spacers. *Frontiers in Microbiology*, 10, 2020.

[46] E. G. Zoetendal, M. Rajilic-Stojanovic, and W. M. de Vos. High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. *Gut*, 57(11):1605–1615, 2008.

[47] N. Zrelovs, A. Dislers, and A. Kazaks. Motley crew: Overview of the currently available phage diversity. *Frontiers in Microbiology*, 11, 2021.