Functional Phage Genomics of selected Taxa

Dissertation

for the award of the degree

"Doctor rerum naturalium" (Dr. rer. nat.)

of Georg-August-Universität Göttingen

within the doctoral program Biology

of the Georg-August University School of Science (GAUSS)

submitted by

Cynthia Maria Chibani

From Pointe-Claire, Canada

23 April Göttingen, 2019

Thesis Committee

Prof. Dr. Rolf Daniel, Department of General Microbiology, Institute for Microbiology and Genetics

Prof. Dr. Burkhard Morgenstern, Department of Bioinformatics, Institute for Microbiology and Genetics

Dr. Heiko Liesegang, Department of Genomic and Applied Microbiology, Institute of Microbiology and Genetics, Georg-August University Göttingen

Members of the Examination Board

Referee: **Prof. Dr. Rolf Daniel**, Department of General Microbiology, Institute for Microbiology and Genetics

2nd Referee: **Prof. Dr. Burkhard Morgenstern**, Department of Bioinformatics, Institute for Microbiology and Genetics

3rd Referee: **Dr. Johannes Soeding**, Computational Biology, Max Planck Institute for Biophysical Chemistry

Further members of the Examination Board

PD Dr. Michael Hoppert, Department of General Microbiology, Institute of Microbiology and Genetics, Georg-August University Göttingen

Prof. Dr. Joerg Stuelke, Department of General Microbiology, Institute of Microbiology and Genetics, Georg-August University Göttingen

Prof. Dr. Stefanie Poeggler, Department of General Microbiology, Institute of Microbiology and Genetics, Georg-August University Göttingen

Date of oral examination: 21.05.2019

List of publications

1. Wendling, C.C., Piecyk, A., Refardt, D., **Chibani, C.,** Hertel, R., Liesegang, H., Bunk, B., Overmann, J. and Roth, O., 2017. **Tripartite species interaction: eukaryotic hosts suffer more from phage susceptible than from phage resistant bacteria.***BMC evolutionary biology*,*17*(1), p.98.

2. **Cynthia Maria Chibani**, Anja Poehlein, Olivia Roth, Heiko Liesegang, and Carolin Charlotte Wendling. "**Draft Genome Sequence of Vibrio splendidus DSM 19640.**" *Genome announcements* 5, no. 48 (2017): e01368-17.

3. **Chibani, C.M.**; Farr, A.; Klama, S.; Dietrich, S.; Liesegang, H. **Classifying the Unclassified: A Phage Classification Method**. Viruses 2019, 11, 195.

4. **Chibani, C.M**.; Meinecke, F.; Farr, A.; Dietrich, S.; Liesegang **ClassiPhage 2.0 classification using ANN.** Bioarchive version.

Poster Presentations

1. **Evolved phage resistance in *Vibrio alginolyticus*K01M1 in response to two different temperate bacteriophages.**
   Wendling C.C., **Chibani C.M**., Poehlein A., Hertel R, Lange J., Goehlich H., Liesegang H., Roght O., Brockhurst M.
   VAAM Annual Conference 2018 - April 15 - 18, 2018, Wolfsburg, Germany.

2. **Novel phage isolates infecting Bacillus.**
   Cynthia Maria Chibani, Tobias Schilling, Heiko Liesegang, Robert Hertel.
   1st German Phage Symposium, 9-11 October 2017, Stuttgart, Germany.

3. **Co-genomics of Bacillus thuringiensis parasites and Tribolium castaneum hosts after experimental coevolution.**
   Barbara Milutinovic, Kevin Ferro, **Cynthia Chibani**, Stefani Diaz, Jacqueline Hollensteiner, Heiko Liesegang, Daniela Esser, Philip Rosenstiel, Hinrich Schulenburg, Joachim Kurtz
   ESEB, Second Joint Congress on Evolutionary Biology Montpellier 2018.

4. **Vibrio alginolyticus and Vibrio splendidus: prophage identification in the genome sequences of fish pathogens.**
   Chibani C.M., Hertel R., Roth O., Wendling C., Liesegang H.
   The 7th Vibrio conference, Station Biologique de Roscoff, France, 29 March – 1 April 2016.

5. **Vibrio alginolyticus and Vibrio splendidus – Inoviridae bacteriophages identification and classification in the genome sequences of fish pathogens.**
   C. M. Chibani, C. Wendling, R. Hertel, O. Roth, H. Liesegang
   The ProkaGENOMICS 2017 conference from 19–22 September 2017 in Göttingen/DE

6. **Comparative genome analysis of three Clostridioides difficile strains involved in a multiple and isochronal infection of a single patient.**
   E. Brzuszkiewicz , U. Groß , T. Riedel, B. Bunk, C. Spröer, A. Poehlein, **C. Chibani**, J.

Overmann, O. Zimmermann, R. Daniel, H. Liesegang

The ProkaGENOMICS 2017 conferencefrom 19–22 September 2017 in Göttingen/DE

7. **Comparative genomics of Bacillus thuringiensis biovar tenebrionis and its inducible phages.**

   J. Hollensteiner, **C. M. Chibani** , A. Poehlein , C. Spröer, M. Hoppert , J. Kurtz, R. Daniel , H. Liesegang

   The ProkaGENOMICS 2017 conference from 19–22 September 2017 in Göttingen/DE

List of Abbreviations

Acc	Accuracy
ANN	Artificial Neural Networks
ANOVA	Analysis Of Variance
AUC	Area Under Curve
BAVS	Bacterial and Archaeal Viruses Subcommittee
BLAST	Basic Local Alignment Search Tool
Cas	CRISPR associated protein
CDS	Coding DNA Sequence
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DNA	DeoxyriboNucleic Acid
DNN	Deep Neural Networks
Ds	Double Stranded
FN	False Negative
FP	False Positive
HGT	Horizontal Gene Transfer
HMM	Hidden Markov Model
HS-bacteria	Highly Susceptible
HTS	High-Throughput Sequencing
ICTV	International Committee on Taxonomy of Viruses
IS-bacteria	Intermediate Susceptible

Kbp	Kilo base pairs
ML	Machine Learning
MLSA	Multi Locus Sequence Alignment
MLST	Multi-Locus Sequence Typing
mRMR	minimal-redundancy-maximal-relevance
MSA	Multi Sequence Alignment
NCBI	National Center of Biotechnology Information
NGS	Next Generation Sequencing
Og	Orthologous Groups
ORF	Open Reading Frame
PCA	Principal Component Analysis
pVOG	Prokaryotic Virus Orthologous Groups
R-bacteria	Resistant
Relu	Rectified Linear Unit
RF	Random Forest
RM	Restriction Modification
RNA	RiboNucleic Acid
ROC	Receiver Operation Curve
Sn	Sensitivity
Sp	Specificity
Ss	Single Stranded
SVG	Scalable Vector Graph
SVM	Support Vector Machine

TEM	Transmission Electron Micrographs/Microscopy
TN	True Negative
TP	True Positive
UViGs	Uncultured Virus Genomes
WHO	World Health Organization
ZOT	Zona-Occludens Toxin

Table of Contents

CHAPTER I: Introduction

Phages: History and Description

Bacteriophages, viruses infecting prokaryotes, are one of the most abundant entities on earth. The amount of existing particles is estimated to be around 10^{31}(Whitman et al. 1998). They are ubiquitous and can be isolated from any ecological niche where their host is present (McNair et al. 2012; Brüssow & Hendrix 2002; Roux et al. 2016).

Frederick Twort(Frederick & Twort 1931) and Felix Hubert d'Herelle(Summers 2017)They were the first to independently describe phages in the early 1900s. While F.W. Twort failed to interpret his observation in 1915, d'Herelle published his discovery in 1917, where he described the bacteriophage as an obligate intracellular bacterial parasite (Summers 2017). Besides, d'Herelle examined the potential use of bacteriophages as therapeutic agents. He first discovered that phages clear dysentery in diseased patients (Salmond & Fineran 2015). Consequently, phages were widely used, in former Soviet countries up until the fall of the Soviet Union, in clinical studies and applications as antibacterial agents especially at the Eliava institute in Tbilisi, Georgia (Abedon et al. 2011).

Additionally, phages have enormously contributed to the field of molecular biology (Salmond & Fineran 2015). They led to the discovery of i) restriction-modification (RM) systems (Roberts 2005) and ii) the "clustered regularly interspaced short palindromic repeats" (CRISPR)-associated protein (cas) systems, which are defense systems used by the bacteria against phages (Sorek et al. 2013).

The interest in phage research was revived approximately in the year 2000, as a result of the genomic and metagenomic revolution, which highlighted phage diversity and abundance (Ofir & Sorek 2018).

Phages are termed as obligate intracellular parasites since they need their host's cellular machinery for their replication. They are known to have two life cycles once they infect a bacterial host: i) lytic or ii) lysogenic. A lytic phage replicates within the host and then lyses the cell at the end of the cycle for release to the environment. It has been estimated that over 20% of bacteria are lysed daily through bacteriophages infection in the ocean (Pan et al. 2018). On the other hand, a lysogenic phage, infects its host and can, either remain as an extra-chromosomal element or integrate its DNA into the host genome, replicating passively with the

replicating host. The latter is termed prophages (Fortier & Sekulovic 2013; Akhter et al. 2012; McNair et al. 2012). Under certain circumstances, a temperate phage can enter a lytic lifestyle after DNA damage caused by diverse stress factors (Fortier & Sekulovic 2013). Recently, Sorek et al.(2017) described that phages interact using a communication peptide, found in various versions in different phages, and trigger the switch between lytic and lysogenic life cycles(Sorek et al. 2017).

Moreover, phages were extensively explored for their potential to encode and confer virulence factors to their bacterial host. Accordingly, they convert their bacterial host into a pathogenic strain through lysogenic conversion. This further emphasized the interest in phage research (Fidelma Boyd & Brussow 2002). As a result, phages are an essential entity co-evolving with their bacterial host ever since the beginning of time (Iranzo, Krupovic, et al. 2016).

Taxonomy: Reasons and Importance

Virus taxonomy aims to describe viral evolutionary relationships and illustrate their remarkable genetic and structural diversity. Further interest is placed on their virulent lifestyle which has applications towards phage therapy (Aiewsakun et al. 2018; Housby & Mann 2009; McNair et al. 2012). Historically, phages have been characterized based on their morphology including shape, size, presence/absence of capsid, and on their genomic size and nature (whether they are ss/ds, DNA or RNA phages). Other criteria include the host genus and sequence similarities (Hans-W Ackermann 2011). The viral diversity is much more extreme than any other organism (Aiewsakun & Simmonds 2018),and their genomes can range from less than 2Kbp to more than 2,000 Kbp(Chow & Suttle 2015).

The International Committee on Taxonomy of Viruses (ICTV;https://talk.ictvonline.org/) is responsible for assigning viruses into hierarchical taxa, based on visualizing and resolving the phage morphology by electron microscopy. As of 2016, it consists of 8 orders, 122 families, 35 subfamilies, 735 genera and 4,404 species (Lefkowitz et al. 2018). Most known phages are classified into the order of *Caudovirales*, so-called the tailed phages. The order includes three phage families: The *Myo-*, *Podo-* and *Siphoviridae*; these three families describe the long contractile phages, the long non-contractile phages,and the short-tailed phages, respectively.

Viruses infecting archaea are classified into 13 families; those include *Ampullaviridae*, *Fuselloviridae*, and *Bicaudaviridae* which comprise bottle-shaped phages, spindle-shaped phages,and two-tailed-shaped phages, respectively (Aiewsakun et al. 2018).

Experimental identification and classification of bacteriophages remain a tedious and time consuming process, which fuels the demand for sequence-based computational methods to do so.

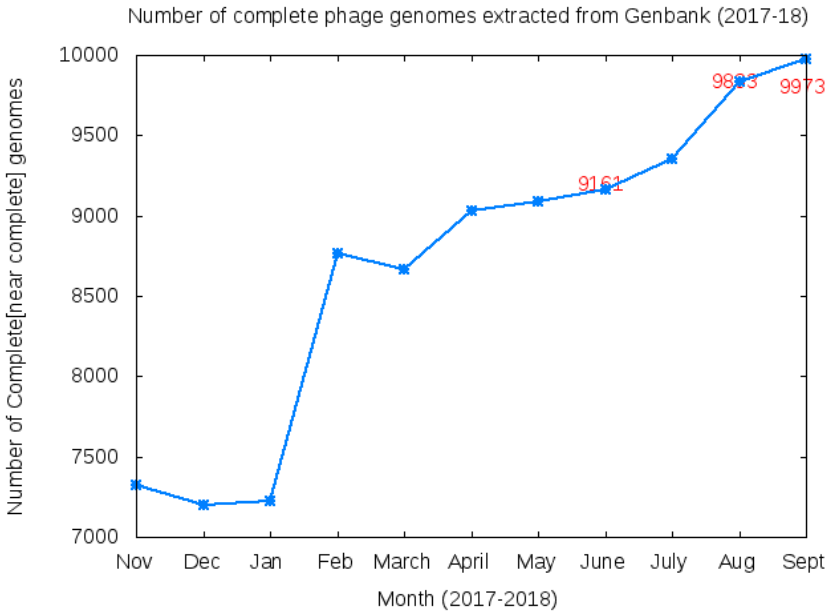


Figure 1: Number of complete phage nucleotide sequences deposited in public databases in the years 2017-2018. This figure is downloaded from the Millard lab webpage accessed on 03/14/2019 (http://millardlab.org/bioinformatics/bacteriophage-genomes/).

With the booming advances in High-Throughput Sequencing (HTS) technologies, metagenomic approaches, and the exploding amounts of sequenced data, the rate at which phage genomes are being sequenced (Figure1) surpasses that of isolation and culturing by order of magnitudes (Simmonds et al. 2017; McNair et al. 2012). Additionally, genome data from various environmental samples and human gut microbiome has unveiled the ubiquity of prophages, the sequences of which do not match any known sequences deposited in public databases (Aiewsakun & Simmonds 2018). As a result, a gap exists between bacteriophages sequences deposited in GenBank which are not classified by the ICTV according to their

classification procedure (Rohwer & Edwards 2002; Bolduc, Jang, Doulcier, Z. You, et al. 2017; Simmonds et al. 2017). This gap is envisaged to increase even more in the future (Manavalan & Lee 2017; Rohwer & Edwards 2002). Hence, there is a need that the ICTV expands their classification to include those more massive viral datasets (Aiewsakun & Simmonds 2018). Moreover, sequence data provides a reliable means of representing viral evolutionary relationships at high resolutions (Simmonds et al. 2017).

Sequence-based Taxonomy

To expand what is a morphology-based viral classification imposed by the ICTV, to viral sequences where phenotypic data cannot be obtained, scientists can profit from the relationship between phenotypic features used for family assignment and the corresponding genomic features (Aiewsakun & Simmonds 2018).

In the last decade, we have seen significant shifts towards sequence-based taxonomy of bacteriophages, and multiple approaches have been proposed, which proved robust as a guide for divergent and highly mosaic viruses (Aiewsakun & Simmonds 2018). Contrary to bacteria which have conserved genes and a 16S rRNA gene traditionally used for taxonomy, viruses lack such a marker gene to place them on the tree of life (Rohwer & Edwards 2002). As a result, different genes were used in an attempt to create viral phylogenies, such as the DNA polymerase, the major capsid proteins,and the ribonucleotide reductase; which are estimated to be found in over 90% of dsDNA viruses (Reyes & Gruber 2017). However, the mentioned proteins don’t share conserved sites explaining the limitation of their use (Novik et al. 2017).

Early on, the genetic complexity of viruses was recognized. Phages in the same taxonomic group might not have a similar nucleotide sequence, but share gene functionality (Lawrence et al. 2002). As a result, Rohwer and Edwards (2002) (Rohwer & Edwards 2002) described the Phage Proteomic Tree. The proposed method used translated genomes and showed high classification specificity for the viral dataset used. The techniquewas generated with a smalldataset and hence is not generally applicable (Bolduc, Jang, Doulcier, Z.-Q. You, et al. 2017).

Another approach is the use of pairwise alignments techniques (Meier-Kolthoff & Göker 2017; Merabishvili et al. 2011; Lavigne et al. 2008). However, it is only applicable to phage sequences similar to those in a reference database. This approach makes it impossible to classify distantly related phages without any prior knowledge (Bolduc, Jang, Doulcier, Z.-Q. You, et al. 2017).

Lastly, protein clustering techniques enabled classification of viral sequences with no prior knowledge (Lima-Mendez et al. 2008; Bolduc, Jang, Doulcier, Z.-Q. You, et al. 2017; Roux, Enault, et al. 2015). Monopartite gene sharing networks, as described by Lima-Mendez et al. (Lima-Mendez et al. 2008), correctly classified 95% of the 306 phage genomes available at that time. Recently, the performance of the method was re-evaluated and proved to be robust as it only failed in classifying only 1 in 4 dsDNA viruses (Bolduc, Jang, Doulcier, Z.-Q. You, et al. 2017). However, monopartite networks do not retain information about the encoded genes per virus. Contrastingly, bipartite gene sharing networks (Corel et al. 2016), consisting of two classes of nodes (homologous protein families and viral genomes), allow the identification of genes shared between and across genomes which have likely been exchanged via Horizontal Gene Transfer (HGT). As a result, they perform better for detecting mosaic genomes. Bipartite networks were successfully implemented in multiple studies (Roux, Hallam, et al. 2015; Iranzo, Krupovic, et al. 2016; Iranzo, Koonin, et al. 2016). Iranzo, Krupovic&Koonin(Iranzo, Krupovic, et al. 2016) revealed a module based structure of dsDNA viruses, while Iranzo, Koonin et al. (Iranzo, Koonin, et al. 2016) extended the method to archaeal viruses and related plasmids. Both networks showed the possibility of a genome-based viral taxonomy consistent with the ICTV accepted phage genera.

What is accepted by the ICTV for Classification

The Bacterial and Archaeal Viruses Subcommittee (BAVS), the subcommittee of the ICTV, have formally expressed their intention to include viruses based on their sequence information into their taxonomy. However,because those viruses lack the standard required phenotypic information, the sequence is used as an attribute to assist in the viral taxonomic assignment (Aiewsakun & Simmonds 2018). The BAVS currently approves the use of BLASTN for the comparison of closely related phages, and endorsed software such as CLANS (Asare et al.

2015), GEGENEES (Sundstro 2012) and VISTA (Frazer et al. 2004), which are based upon sequence similarities resulting from BLASTN. Nevertheless, little information is provided for parameters needed to assign divergent viruses into the taxonomic divisions based on nucleotide or protein information (Krupovic et al. 2016; Aiewsakun & Simmonds 2018).

The use of Hidden Markov Models as a basis for viral classification

Profile Hidden Markov Models (HMMs) represent a robust method for modeling viral sequence diversity which can detect, with high sensitivity, three times more remote homologs than conventional pairwise sequence-alignment methods (Grazziotin et al. 2017; Reyes et al. 2017; Barrett et al. 1998). Amino acids sequence divergence is, over time, much slower than nucleotide sequence divergence. Therefore, protein profile HMMs sensitivity can detect functionally related proteins even with low shared similarity, which enables virus detection without previous specific information (Alves et al. 2016; Ren et al. 2017).

It has been estimated that the number of HMMs compiled in databases, such as pVOG and vFam (http://derisilab.ucsf.edu/software/vFam/), represent less than 20% of viral protein sequences (Skewes-cox et al. 2014). This is due to uneven taxonomic sampling, as poorly characterized viral families with few members, especially archaeal viruses, display a low proportion of gene coverage (Figure2)(Grazziotin et al. 2017; Iranzo, Krupovic, et al. 2016; Reyes & Gruber 2017).

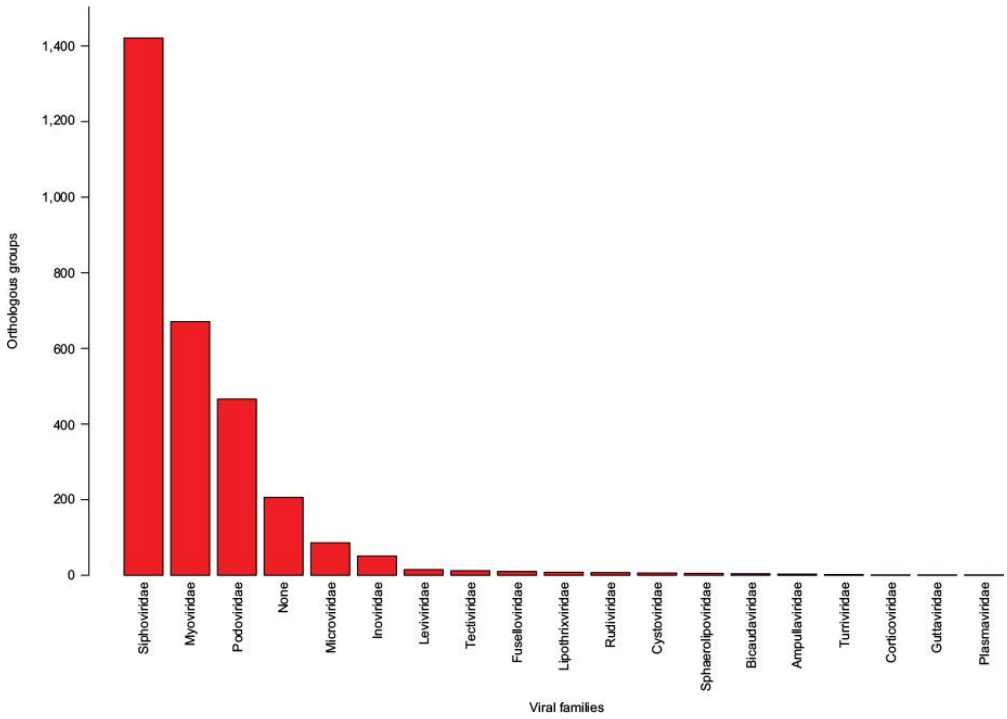


Figure 2: Number of Orthologous Groups (Ogs) sorted by viral families available on the pVOG database. The figure was taken from Reyes et al. (Reyes & Gruber 2017).

Nevertheless, the success of profile HMMs in dealing with divergence is proving to be an invaluable tool for viral identification (Reyes & Gruber 2017). Profile HMMs for viral detection and classification has been widely used in a considerable number of literature(Reyes & Gruber 2017; Grazziotin et al. 2017; Aiewsakun & Simmonds 2018; Lopes et al. 2014; Fouts 2006).

One of the most promising applications of viral HMMs is their use as seed for viral genomes reconstruction from metagenomic datasets along with their taxonomic assignment (Alves et al. 2016). It is worth mentioning that the combination of different profile HMMs is key since no single match is sufficient to assess true viral sequence diversity.

Phages in Metagenomes

Metagenomic data is expanding our understanding of viral diversity, thus challenging viral recognition, assembly and classification methods (Simmonds et al. 2017). It is proving to be instrumental in the identification of entirely new groups of viruses, as over 750,000

uncultivated viral genomes (UViGs) has been reported(Roux et al. 2019). UViGs make the evaluation of viral diversity possible in addition to the over-represented dsDNA viruses, thus addressing one of the significant issues concerning under-represented viral species in public databases. However, methods to clone and sequence ssDNA and RNA viruses still need further development (Simmonds et al. 2017).

In 2016 The ICTV expressed interest to incorporate identified viral groups from metagenomic datasets into their official taxonomy, even though they lack a direct correlation with biological characteristics (Lefkowitz et al. 2017; Simmonds et al. 2017). On account of sequence data providing essential information concerning evolutionary relationships, genome organization and other genomic features (Simmonds et al. 2017). Over the years multiple species and genera were assigned in the previously existing viral families, which were already set up based on phenotypic properties (Adams et al. 2017). However, the ICTV provides little or no systematic information on how divergent a virus has to be and what genomic features are to be considered for the taxonomic divisions (Aiewsakun & Simmonds 2018).

Future of Phage Classification

A consensus statement endorsed by the ICTV outlines a framework for the incorporation of metagenomic data into the standard ICTV taxonomy, where necessary checks for data integrity should be performed (Simmonds et al. 2017). It was proposed that i) the classification of UViGs into new taxa is possible, provided sequence relationships are comparable to those taxa already existing in that family; ii) When no relationship exists a new family can be assigned based on crucial variation in the genome organization and the inferred replication strategy; iii) Clustering and network analysis are to be used and critically evaluated for hierarchical taxonomic assignments; iv) Use the ICTV taxon nomenclature which is extendable to additional species; and lastly v) Procedure development to shorten the time needed by the ICTV to process newly submitted proposals and updating their “Master Species List”.

Prophage Prediction

It has been reported that viruses can infect 13 prokaryotic phyla (Roux, Enault, et al. 2015). Thus, numerous methods have been developed for recognizing integrated prophages in bacterial genomes. Those tools include PHAST (Zhou et al. 2011), which was later extended to PHASTER(Arndt et al. 2016) and then PHASTEST(Arndt et al. 2017); Phage_Finder(Fouts 2006), Prophinder(Lima-Mendez et al. 2008), PhiSpy(Akhter et al. 2012) and VirSorter(Roux, Enault, et al. 2015). Generally, bacterial genomes are scanned in a sliding window approach to finding regions with hits to known viral sequences (Ren et al. 2017). PhiSpy was the first tool ever described to include viral sequence features, which increased prophage prediction and outperformed the existing tools. Despite the added sequence derived features, these tools rely on finding homologous genes to known viral sequences, representing only a fraction of viral diversity. It has been estimated by Roux et al. (Roux, Hallam, et al. 2015)that known phage sequences are isolated from less than 15% of bacterial hosts. As a result, a gap still exists in generating a comprehensive reference free prophage finding tool. Lastly, VirSorter, a tool designed to detect viral sequences in genomic datasets as well as metagenomic assemblies, performs better for metagenomic and fragmented datasets since it does not consider additional prophage specific characteristics (Roux, Enault, et al. 2015).

Why use Machine Learning Algorithms?

To face the challenges resulting from the growing amount and complexity of phage sequenced data, Machine Learning (ML) algorithms and data mining techniques, have gained considerable interest and can be applied with little computational burden (Morota et al. 2018). They are expected to become instrumental for prediction and inference, due to their advantage in considering a large number of features simultaneously to identify a complex genomic element like a prophage(Manavalan, Tae H. Shin, et al. 2018; Manavalan & Lee 2017; Manavalan et al. 2014). They generally try to assign an outcome label to new samples given a list of input features the ML algorithm was trained on (Amgarten et al. 2018).

Phage Protein Prediction

To our knowledge, Seguritan et al. (2012) were the first to mention the use of Artificial Neural Networks (ANN) for the successful detection of phage structural proteins (Seguritan et al. 2012). What followed was an increasing number of studies, using various ML algorithms, for effectively predicting phage proteins (Ding et al. 2014; Manavalan, Tae H Shin, et al. 2018; Pan et al. 2018; Feng, Ding, et al. 2013). In 2013, Feng et al. (Feng, Lin, et al. 2013) used a Naïve Bayes approach which achieved an overall accuracy of 79.15%. The same dataset was used again in 2015, where first the analysis of variance (ANOVA) for selection of the most informative feature was performed. Accordingly, the selected featureswere used as an input for a support vector machine (SVM) classifier for the identification of phage proteins. This method, PVPred, achieved an overall accuracy of 85.02% (Ding et al. 2014). PVPred was outperformed by PVP-SVM (Manavalan, Tae H. Shin, et al. 2018), reaching an accuracy of 86.97 %, where they used a random forest (RF) algorithm for the feature selection process. Tan et al. proposed the use of a two-step feature selection process, using ANOVA and the minimal-redundancy-maximal-relevance (mRMR) method, reaching an accuracy of 87.95 % (Tan et al. 2018).

Recently, Pan et al.(Pan et al. 2018) generated a new method called PhagePred. It uses a g-Gap feature selection process and then feeding the most informative features to a Naïve Bayes classifier. PhagePred reached an exceptional 98.37% accuracy, outperforming the existing methods. Interestingly, all these methods use similar approaches, sometimes the same dataset for prediction-method development, a feature selection process, and two machine learning algorithms (Naïve Bayes and SVM), and finally showed promising results for phage proteins prediction.

All in all, the above mentioned tools focus on the identification and classification of single phage proteins, rather than the identification of complete phage genomes.

Phage Prediction in Metagenomic Bins

Amgarten et al. (2018) introduced a tool called MARVEL. It predicts phages in metagenomic bins using an RF algorithm and subsequently classifies *Caudovirales* phage families (Amgarten

et al. 2018). MARVEL achieved a much higher sensitivity when benchmarked against two state-of-the-art tools, VirSorter and VirFinder. They showed that three features were the most informative for bacteriophage prediction, (i) gene density, (ii) strand shifts and (iii) genes with significant hits against HMMs downloaded from the pVOG database.

In summary, MARVEL and VirSorter enabled the sorting of metagenomic assembled bins whether they belong to phages or not and subsequently classify the sorted sequences into taxonomic phage families. These approaches further reinforce the advantage of using ML algorithms as frameworks for solving pressing problems arising from the ever-increasing number of data and phage sequence diversity.

General Project Aims

Due to high viral diversity and the ever-increasing number of sequenced viral datasets, we aimed to describe a general approach that would still be pertinent whenever i) more phages are sequenced; ii) more phage families are represented and iii) regardless of how many bacterial genomes are sequenced. This study aimed to generate a method for complete prophage genome identification and subsequent taxonomical classification into the correct ICTV family. I intended to create phage models, when possible, for every phage family to taxonomically classify phage genome sequences on the family level. For confirmation, I tested the generated models, on a set of experimentally investigated and classified set of prophages. And lastly, I examined sequence derived features, to identify integrated prophages within bacterial genomes and potentially classify them using the generated models.

To generate a positive dataset for benchmarking purposes, I investigated:

- Temperate phage-bacteria interaction and *Inoviridae* as a driving force for *Vibrio* host evolution (Chapter II.1).
- Comparative genomics of experimentally proven *Inoviridae* prophages (Chapter II.2).

Regardless of the advances made for phage classification, no method exists that classifies phages based on whole genomes information. Therefore, the first aim of this project was to generate a phage classification method. I investigated:

- The use of HMM as a basis for phage classification, using vibriophages as a pilot project (Chapter II.3), a method we call ClassiPhage.
- The generation of specific profile HMMs per phage family of all available published phage genomes, and the generation of an input matrix that can be used to classify phages into phage families, using ANN (Chapter II.4), a method we call ClassiPhage 2.0.

Prophage identification has long been a topic of interest, and currently being dominated by software such as PHASTER and PhiSpy. PHASTER is based on gene annotations, and

BLASTp hits to a phage database while PhiSpy is based on different sequence features. However, it was constructed using a small dataset of closely related genomes. Therefore, the second aim of this project was to broadly identify integrated prophage regions in bacterial genomes based on sequence-derived features. I investigated and applied:

- DNA derived features and their use for prophage regions identification in bacterial genomes as input for a Deep Neural Network (DNN) classifier (Chapter II.5), a method we call IdentiPhage.

CHAPTER II: Manuscripts and Publications

II.1 Tripartite species interaction: eukaryotic hosts suffer more from phage susceptible than from phage resistant bacteria

Tripartite species interaction: eukaryotic hosts suffer more from phage susceptible than from phage resistant bacteria

Carolin C. Wendling*, Agnes Piecyk, Dominik Refardt, **Cynthia Maria Chibani**, Robert Hertel, Heiko Liesegang, Boyke Bunk, JörgOvermann,and Olivia Roth

Wendling et al. BMC Evolutionary Biology (2017) 17:98, **DOI** 10.1186/s12862-017-0930-2

Authors' contributions

OR and DR initiated this study and established the phage-bacteria system in the laboratory. AP and OR performed the phage-bacteria infection matrix. CCW did the multilocus genotyping of the bacteria strains. CCW and OR conducted the pipefish-bacteria infection experiment. Statistics of all laboratory experiments were done by CCW and AP. CCW, RH, HL, BB, and JO performed bacteria genome sequencing. **CC**, HL, and BB analyzed the bacterial genomes. CCW and OR coordinated the project and wrote the manuscript. All authors read and approved the final manuscript.

Wendling *et al. BMC Evolutionary Biology* (2017) 17:98
DOI 10.1186/s12862-017-0930-2

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Tripartite species interaction: eukaryotic hosts suffer more from phage susceptible than from phage resistant bacteria

Carolin C. Wendling[1*], Agnes Piecyk[1,2], Dominik Refardt[3], Cynthia Chibani[4], Robert Hertel[4], Heiko Liesegang[4], Boyke Bunk[5], Jörg Overmann[5] and Olivia Roth[1]

Abstract

Background: Evolutionary shifts in bacterial virulence are often associated with a third biological player, for instance temperate phages, that can act as hyperparasites. By integrating as prophages into the bacterial genome they can contribute accessory genes, which can enhance the fitness of their prokaryotic carrier (lysogenic conversion). Hyperparasitic influence in tripartite biotic interactions has so far been largely neglected in empirical host-parasite studies due to their inherent complexity. Here we experimentally address whether bacterial resistance to phages and bacterial harm to eukaryotic hosts is linked using a natural tri-partite system with bacteria of the genus *Vibrio*, temperate vibriophages and the pipefish *Syngnathus typhle*. We induced prophages from all bacterial isolates and constructed a three-fold replicated, fully reciprocal 75 × 75 phage-bacteria infection matrix.

Results: According to their resistance to phages, bacteria could be grouped into three distinct categories: highly susceptible (HS-bacteria), intermediate susceptible (IS-bacteria), and resistant (R-bacteria). We experimentally challenged pipefish with three selected bacterial isolates from each of the three categories and determined the amount of viable *Vibrio* counts from infected pipefish and the expression of pipefish immune genes. While the amount of viable *Vibrio* counts did not differ between bacterial groups, we observed a significant difference in relative gene expression between pipefish infected with phage susceptible and phage resistant bacteria.

Conclusion: These findings suggest that bacteria with a phage-susceptible phenotype are more harmful against a eukaryotic host, and support the importance of hyperparasitism and the need for an integrative view across more than two levels when studying host-parasite evolution.

Keywords: Temperate phages, Prophages, Bacteria-phage infection network, *Vibrio*, Tripartite interaction

Background

Infection of parasites by other parasites (i.e. hyperparasitism) plays an important role in the evolution of hosts and parasites. Micro-hyperparasites, for instance bacteriophages, are fundamental in determining the outcome of bacterial diseases [1]. To understand the ecology and evolution of bacterial diseases, it is necessary to extend the view of dual species interactions to tripartite interactions where the phage, its bacterial carrier and a eukaryotic host are involved. Such tripartite interactions have been well studied in systems using lytic phages, of which many demonstrate a trade-off between phage resistance and bacterial virulence (for a recent review see [2]). However, patterns of resistance and virulence between temperate phages, their bacterial carriers and the eukaryotic hosts are largely unexplored.

In contrast to lytic phages, temperate phages have two transmission modes. After infecting a bacterium they can either be transmitted horizontally through cell lysis (lytic cycle), or vertically as prophages, whereby the phage genome is integrated into the bacterial chromosome (lysogenic cycle). Indeed, prophages constitute up to 20% of the bacterial genome and are major contributors to the

* Correspondence: cwendling@geomar.de
[1]GEOMAR, Helmholtz Centre for Ocean Research, Evolutionary Ecology of Marine Fishes, Düsternbrooker Weg 20, 24105 Kiel, Germany
Full list of author information is available at the end of the article

large genomic and phenotypic variation among bacterial strains of the same species [3].

During lysogeny the fitness of the prophage and its bacterial carrier is aligned, which explains instances where prophages protect their hosts against superinfection [4] or provide them with genes that increase bacterial proliferation [3]. However, prophages have also been described as molecular time bombs [5] that, either spontaneously or in response to specific environmental conditions, kill their carriers through cell lysis and switch back to the lytic cycle [3, 5].

While bacteria are in constant coevolutionary interaction with their eukaryotic host, they simultaneously face selection by their micro-hyperparasites, i.e. lytic phages. For instance, evolution of resistance in *Pseudomonas aeruginosa* to lytic ΦPP/and ΦE79 resulted in an upregulation of virulence genes, which ultimately increased virulence against mammalian cells [6]. In contrast, resistance against lytic phages in *Flavobacterium columnare* reduced bacteria gliding motility and thus virulence against its eukaryotic host [7].

We here aimed to extend the existing knowledge of micro-hyperparasitism in phage – bacteria – eukaryotic host interactions using temperate phages. Specifically, the objective of the present study is to investigate resistance patterns to temperate phages in a natural temperate phage – bacterial interaction and its relationship to bacterial harm in an animal host. By using bacteria of the genus *Vibrio*, their derived prophages, and one of their eukaryotic hosts, the broad-nosed pipefish *Syngnathus typhle* as a model system, we addressed the question if bacterial resistance to temperate phages and bacterial harm to eukaryotic hosts can be linked.

While in a variety of human pathogenic strains, *Vibrio* virulence can be directly linked to the presence of prophage [8–10], we lack insight that goes beyond the knowledge about human pathogenic strains and addresses *Vibrio*-phage interactions covering a broader range of environmental isolates. Here, we present experimental data on the interaction between 75 environmental *Vibrio* isolates and their associated prophages as well as on the impact of a subset of these bacterial isolates to the natural eukaryotic host, the pipefish. We conducted fully reciprocal cross-infections between all bacteria and their derived phage lysates, and experimentally challenged pipefish with nine of the isolates that differed in phage resistance. Based on the relative gene expression of pipefish immune genes, we observed that phage resistant bacteria are less harmful than phage susceptible bacteria. Our results indicate that bacteria with a phage-susceptible phenotype are more virulent against their eukaryotic host and suggest that temperate phages are important in shaping bacterial virulence in the marine realm.

Methods

All *Vibrio* strains used in the present study had been isolated from nine healthy broad-nosed pipefish *Syngnathus typhle* collected in the Kiel Fjord during a previous study [11]. Labels were given according to the sampling area (first letter 'K' refers to the study site: Kiel), the fish individual (first number), the organ (second letter: 'E' referring to eggs, 'K' referring to gills, and 'M' referring to the whole intestines) and *Vibrio* colony number (second number). Healthy pipefish harbour a highly diverse community of bacteria of the genus *Vibrio spp.* that show a strong spatial diversification across Europe [11]. While most *Vibrio* are harmless, some are responsible for major disease outbreaks. For instance, several members of the *V. alginolyticus* and *V. splendidus* clade have been isolated from seahorses with signs of infections [12], while *V. harveyi* causes almost 90% mass mortalities in captive bred seahorses [13].

Vibrio phylogeny

To determine the genetic affiliation of each *Vibrio* isolate we used a multi locus sequence analysis (MLSA) approach based on partial DNA sequences of 3 different genes (16S rRNA, *recA* and *pyrH)*. Bacterial DNA was isolated from cell pellets of overnight cultures using the DNeasy 96 Blood & Tissue Kit (*Qiagen*) according to the manufacturers protocol. Amplification followed previously established protocols [14]. Primer details are listed in supporting information (Additional file 1: Table S1). PCR products were purified using ExoSAP (*Fermentas*) with 0.4 μl FastAP, 0.2 μl ExoI and 1.4 μl H_2O per 2 μl PCR Product. Sequences were obtained on an ABI 3130xl Genetic Analyser (*Applied Biosystems*) using standard Sanger sequencing with ABI BigDye Terminator v3.1 Cycle Sequencing kit (*Applied Biosystems*). The thermal program consisted of an initial denaturation step (60 s at 96 °C) followed by 25 cycles (10 s at 96 °C, 5 s at 55 °C, 5 min at 60 °C).

Whole genome sequencing

DNA for sequencing was isolated from cultures grown in Medium101 (Medium101: 0.5% (w/v) peptone, 0.3% (w/v) meat extract, 3.0% (w/v) NaCl in MilliQ water). The cultures were grown 16 h at 25 °C 250 rpm. High molecular weight DNA was prepared using Qiagen Genomic Tip/100 G from Qiagen, Hilden, Germany. SMRTbell™ template library was prepared according to the instructions from PacificBiosciences, Menlo Park, CA, USA, following the Procedure & Checklist - 10 kb Template Preparation Using BluePippin™ Size-Selection System. Briefly, for preparation of 15 kb libraries 8 μg genomic DNA was sheared using g-tubes™ from Covaris, Woburn, MA, USA. DNA was end-repaired and ligated overnight to hairpin adapters applying components from

the DNA/Polymerase Binding Kit P6 from Pacific Biosciences, Menlo Park, CA, USA. Reactions and BluePippin™ Size-Selection to 7 kb were performed according to the instructions of the manufacturer (Sage Science, Beverly, MA, USA). Conditions for annealing of sequencing primers and binding of polymerase to purified SMRTbell™ template were assessed with the Calculator in RS Remote, Pacific Biosciences, Menlo Park, CA, USA. SMRT sequencing was carried out on the PacBio *RSII* (Pacific Biosciences, Menlo Park, CA, USA) taking one 240-min movie for each SMRT cell using P6 chemistry. In total one SMRT cell per strain was run for eight selected *Vibrio alginolyticus* strains. Genome assembly was performed with the RS_HGAP_Assembly.3 protocol included in SMRT Portal version 2.3.0. The number of postfiltered reads and the average read length of the reads is summarized in Additional file 2: Table S5, as well as the number of contigs obtained after primary assembly. Each contig was trimmed and circularized to obtain the two typical *Vibrio* chromosomes as well as additional plasmids and artificial contigs were removed. Automated genome annotation was carried out using Prokka [15].

Phage-bacteria cross infection assay

We used standard spot-assays to construct a three-fold replicated, fully reciprocal phage-bacteria infection matrix [16].

Prophage induction

All *Vibrio* isolates were induced with mitomycin C (Sigma) as described in [4] with some modifications: bacteria were grown in liquid Medium101 (Medium101: 0.5% (w/v) peptone, 0.3% (w/v) meat extract, 3.0% (w/v) NaCl in MilliQ water) at 250 rpm and 25 °C overnight. Cultures were diluted 1:100 in fresh medium and grown for another 2.5 h at 250 rpm and 25 °C to bring cultures into exponential growth before adding mitomycin C at a final concentration of 0.5 μg/ml. Samples were incubated in an automated plate reader (TECAN infinite M200) for 4 h at 25 °C and mixed periodically. Bacterial lysis upon prophage induction was monitored via optical density at 600 nm (measured every other minute). We determined bacterial lysis time at induction as the time at which turbidity of the culture peaks (for details see [4]). After 4 h, lysates were centrifuged at 6000 *g* for 15 min and the supernatant was ten-fold diluted in TM buffer (modified from [17]): 50% (v/v) 20 mM $MgCl_2$, 50% (v/v) 50 mM Tris-HCl, pH 7.5).

Spot assay

To determine bacterial susceptibility to the different phage lysates we used standard spot assays, in which a lawn of host bacteria is grown in a medium overlaid on agar plates [16]. Phage lysates are spotted on the overlaid medium and may infect bacteria. Phage infection is visible as plaques, i.e. circular clear or turbid zones where a lytic infection has spread through the bacterial lawn. Overnight cultures of bacterial strains were diluted 1:10 in fresh medium and grown for 2 h before they were mixed with the overlay medium as follows: 200 μl of exponentially growing cells were added to 4 ml Medium101 soft agar (0.4%) at 41 °C. The medium was poured onto Petri dishes containing 20 ml Medium101 agar (1.5% (w/v)). After 30 min, 2 μl of each phage lysate were spotted onto the plates. Controls on every plate were 2 μl uninduced bacterial culture, Medium 101, Medium 101 with 0.5 μg/ml mitomycin C, and TM buffer. Plates were dried for 30 min before incubation at 25 °C for 20 h. Each bacterial strain was scored as either susceptible (plaque formation) or resistant (no plaque formation) to each of the phage lysates. Similarly, each phage lysate was scored as either infective (plaque formation) or non-infective (no plaque formation).

We are aware that plaque formation may be misinterpreted by thinning of the bacterial lawn, which can be associated with defective prophages [18], colicins [19] or other toxic components in the supernatant of mitomycin C treated cultures. To support that the supernatants do indeed contain phage particles, we used a serial dilution on a susceptible host ranging from 10^{-1} to 10^{-8} and only scored those isolates, where individual plaques were observed. In addition, we isolated viral DNA (MasterPure DNA Purification Kit, Epicentre) from the supernatants to perform a standard agarose gel electrophoresis with 0.8% agarose and a 1 kb GeneRuler (Fermentas) as marker. Based on these two approaches we could confirm that all mitomycin C treated culture supernatants contained viral particles, which have ssDNA genomes of ~6 kb.

Infection experiment

As the majority of our bacterial isolates (71 out of 75) could be assigned to the *Vibrio alginolyticus* clade, all subsequent analyses as well as the infection experiment are based on the *V. alginolyticus* isolates only.

Experimental procedure

Out of the 71 *Vibrio alginolyticus* strains we selected three strains that were highly susceptible to prophages (further named HS-bacteria), three strains that were intermediate susceptible to prophages (further named IS-bacteria) and three strains that were resistant to prophage infection (further named R-bacteria).

Pregnant male pipefish were randomly caught from the Kiel Fjord in July 2014 and transported to our laboratory facility in Kiel, Germany. Male pipefish were kept separately in 80-L aquaria and fed twice a day with live and frozen mysids. Immediately after birth, fathers were removed from the aquaria and juveniles were fed

twice a day with *Artemia salina* naupliae for another 3 weeks.

Selected bacteria were grown under agitation at 25 °C as described in [20]. After 24 h we adjusted the concentration of each strain to 5×10^8 cells/ml according to [14]. Prior to the start of the infection experiment we pooled 36 fish from nine different pregnant males and injured the skin of each fish with a sterile needle. Afterwards fish were kept separately in small 50-ml beakers, which either contained 10^6 cells/ml of each respective *Vibrio* isolate diluted in PBS or only PBS, which served as a control treatment. We infected nine fish per strain resulting in 108 fish in total. After 2 days all fish were killed with a lethal dose of MS222, immersed in RNA-later and stored at 4 °C until RNA-extraction. We considered 2 days as an optimal time point to end the experiment for two reasons: a) we wanted to give the immune system time to react to the infection, and b) we observed in previous studies that fish mortality during a controlled infection experiment starts on average 3 days after infection.

Gene expression

Expression of 44 target genes relative to two housekeeping genes was analysed using a Fluidigm BioMark™ as described in Beemelmanns and Roth [21]. Briefly, we used 22 target genes assigned to the innate immune system, three to the complement component system, seven target genes assigned to the adaptive immune system and 15 target genes assigned to gene silencing or activation through DNA and histone methylation/demethylation and histone acetylation/deacetylation. Details about function of genes, sequences and primer design can be found in [21] and are listed in Additional file 3: Table S2.

We extracted RNA from whole juvenile fish using an RNeasy 96 Universal Tissue Kit (Qiagen) according to the manufacturer's protocol. RNA concentration was adjusted to a total of 800 ng/μl per sample and subsequently transcribed into cDNA using the Quanti Tect Reverse Transcription Kit (Qiagen), which includes a genomic DNA (gDNA) digestion. After pre-amplification (for details see [21], samples and primers (two technical replicates per gene) were filled into specific inlets into the 96.96 dynamic array IFC (GE-chip) and measured in the BioMark™ system applying the GE fast 96.96 PCR protocol according to Fluidigm instructions. We included non-template controls (NTC), controls for gDNA contamination (-RT) and standard samples for inter-run calibration.

Infection intensity

To estimate the amount of viable *Vibrio* counts within infected pipefish we determined infection intensity, i.e. colony forming units (CFU) by plating 2 μl of the whole fish-suspension (which has been produced for total RNA extraction) on *Vibrio* selective Thiosulfate Citrate Bile Sucrose Agar (TCBS) plates (Fluka Analytica). Plates were incubated at 25 °C for 24 h. Afterwards CFU were counted for each fish.

Bacterial properties

Growth rate

We generated 24 h growth curves of all selected strains, to identify potential differences in growth rates between bacterial groups (HS, IS, R) that might result from increasing costs of phage resistance.

Twitching motility

We further determined bacterial twitching motility based on a standard motility assay to determine if resistance to phages can be assigned to pilus mutations. In brief, aliquots of equal number of cells were stab inoculated on petri dishes containing TCBS agar (Fluka Analytica) and incubated at 25 °C for 48 h. After incubation a hazy zone of growth at the interface between the agar and the polystyrene surface was observed and its surface area quantified using ImageJ. The surface area was calculated as follows: if the surface area is circular in shape, we used the formula $2r\pi$, where $r = 1/2$ the diameter. If the surface area is oval in shape, measures of the shortest and longest diameter were taken and the surface area calculated according to the formula, $\pi \times a \times b$, where $a = 1/2$ the longest and $b = 1/2$ the shortest diameter.

Statistical analysis

All statistics were performed in the R 3.1.2 statistical environment (R Foundation for statistical computing) unless otherwise stated.

Phylogenetic analysis

MLSA was performed as described in [14] with the following modifications: All sequences were manually edited and automatically assembled using CodonCode Aligner v3.7.1.2. Edited gene sequences were compared against published sequences in NCBI GenBank using BLASTN algorithm with default settings based on 99% sequence identity to assign *Vibrio* isolates to putative close phylogenetic relatives. After assembly and alignment of concatenated sequences (2507 bp) using MUSCLE [22], we constructed a phylogenetic tree using the Bayesian Markov chain Monte Carlo (MCMC) method as implemented in MrBayes version 3.2.5 [23, 24]. The generalised time reversible model plus invariant sites (GTR + I), as suggested by the Akaike information criterion (AIC) given by jModelTest [25], was used as statistical model for nucleotide substitution. The MCMC process was repeated for 10^6 generations and sampled every 5000 generations. The first 2000 trees were deleted as burn-in processes and the consensus tree was constructed from the remaining trees. Convergence was assured via the

standard deviation of split frequencies (<0.01) and the potential scale reduction factor (PSRF ~ 1). The resulting phylogenetic tree and associated posterior probabilities were illustrated using FigTree version 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/.

Whole genome analysis

We calculated the phylogenetic relationship between the eight fully sequenced *Vibrio* strains using a whole genome alignment phylogeny-based approach. The alignment was calculated using Mugsy [26], and only the relevant LCBs (local collinear blocks) aligned regions present in all analyzed strains were extracted using Phylomark. These regions were concatenated and positions with gaps removed [27]. A heuristic maximum-likelihood phylogenetic tree was calculated from the resulting core alignment (528,197 bp) using FastTree2 [28] and visualized using Interactive tree of life (iTOL) v3 [29]. We screened the sequenced genomes for selected common virulence factors, such as virulence islands and type 2 toxin-antitoxin system as wells as the presence of a CRISPR/Cas system and differences in methylation patterns. In detail, *Vibrio* Genomic islands were predicted using IslandViewer [30]. Type II TA modules were screened using TAfinder [31], a web-based tool to identify type II toxin-antitoxin (TA) loci in bacterial genomes. Potential toxin-like candidates were predicted using ClanTox [32]. SMRT sequencing data of all strains was mapped to the eight assembled genome sequences of *V. alginolyticus*, using the BLASR algorithm (Pubmed-ID 22988817) as implemented in Pacific Biosciences' SMRT Portal 2.3.0 within the "RS_Modification_and_Motif_Analysis.1" protocol applying default parameter settings.

Network analysis

After confirmation that the three infection matrices were not significantly different from each other (Mantel test; Monte-Carlo test observation based on 9999 permutations > 0.085; $p < 0.001$) we calculated a consensus matrix, in which we considered an infection to be positive if plaque formation was visible in at least two of the three replicates. Subsequent network analysis was performed on the consensus matrix using the bipartite package [33] and the Falcon interactive Mode for R [34]. Nestedness was calculated using the NODF index, which estimates nestedness based on overlap and decreasing fill. We used the SS null model to test for significance of the nestedness score.

Gene expression

A detailed description of our gene expression analysis is given in [21]. In short, we calculated the mean cycle time (ct) for each of the two replicates. We used qbase$^+$ (version 2.6.1 [35]) to calculate the optimal number of housekeeping genes and found that the combination of the two housekeeping genes ubiquitin (Ubi) and ribosome protein (Ribop) showed the highest stability (average geNorm M ≤ 0.5). After removal of samples with a coefficient of variation larger than 4% we calculated the geomean Ct of the two housekeeping genes to quantify the relative gene expression of each target gene by calculating - ΔCt. We used a multivariate analysis of variance (MANOVA) using the Pillai's trace statistics with - ΔCt values as dependent variable and bacterial group as well as strain nested within bacterial group as the independent variable. MANOVA was followed up by univariate analyses of the single genes. We further conducted a principal component analysis (PCA) using the ade4 package [36] to assess clustering according to the bacterial groups based on differences in expression patterns.

Viable Vibrio counts

We analysed the amount of CFU using a Kruskal-Wallis test for non-parametric data.

Bacterial growth rate

We used a linear mixed effect model with a Maximum likelihood error distribution using lme (package nlme) with bacterial group (HS, IS, R), time as well as their interaction as fixed effect and strains as random effect.

Twitching motility

We used a linear model to estimate differences in twitching motility based on differences in surface areas using bacterial group as fixed variable.

Lysis time

We defined lysis time as the time at which turbidity of the culture peaks [4]. According to the infectivity pattern of the derived prophages we grouped bacterial strains into three categories (HI: High infectivity, II: Intermediate infectivity, NI: No infectivity). We estimated the effect of these three bacterial groups on lysis time using a linear model (function: lm) followed by Tukey's HSD posthoc test (R-package lsmeans).

Results

Vibrio phylogeny

Vibrio phylogeny was constructed based on three concatenated housekeeping loci (16 s rRNA, *recA* and *pyrH*) representing 2,507 total nucleotides using the Bayesian Markov chain Monte Carlo (MCMC) method. The 75 isolates were separated into three major clusters, of which we could assign 71 strains to the *Alginolyticus* clade, three strains to the *Splendidus* clade and one strain to the *Vulnificus* clade (Fig. 1a). All strains belonging to the *Alginolyticus* clade had a 100% sequence

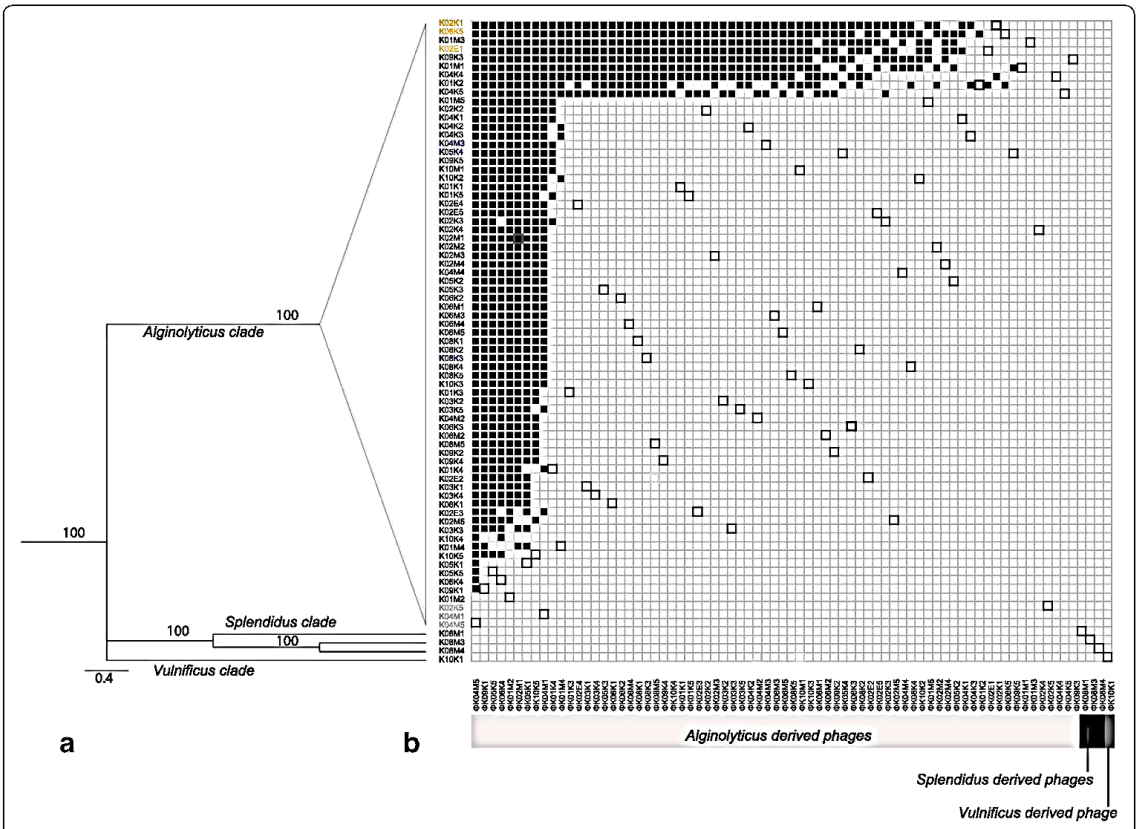


Fig. 1 **a** Collapsed *Vibrio* phylogeny of the 75 isolates based on three concatenated housekeeping loci (16S rRNA *recA* and *pyrH*, 2507 bp) using the Bayesian Markov chain Monte Carlo (MCMC) method. The 75 isolates were separated into three major clades, i.e. *Alginolyticus* clade, *Splendidus* clade and *Vulnificus* clade. Nodes are labelled with posterior probabilities. **b** *Vibrio*/phage cross-infection matrix. Rows and columns represent bacteria and phage lysates. *Black cells* indicate infection success. Framed cells depict cross-inoculation between a lysogen and its derived phage lysate. Strains used in the infection experiment are highlighted in *orange* (HS-bacteria), *blue* (IS-bacteria) and *grey* (R-bacteria)

identity based on the concatenated alignment and were therefore grouped by collapsing the internal branches within the *Alginolyticus* clade. However, based on a whole genome alignment of the selected eight strains, we could show that the isolates represent different strains (Additional file 4: Table S3 and Fig. 2b). These differences are mainly caused by different integrated prophages at different insertion sites, which might explain the observed distinct phenotypes.

Vibrio-phage cross infection network

We found inducible prophages in all *Vibrio* isolates. In 64 out of the 71 *Alginolyticus* isolates single plaques were visible at dilutions of 10^{-6} to 10^{-8}, however, they had fringed edges and were often overlapping and thus not clearly discernable making it impossible to count single PFUs. In addition, we could confirm the presence of prophages in the supernatants by DNA extraction and subsequent gel-electrophoresis, showing products of around ~6 kb, for all 71 isolates. A screening of the genomes of the complete sequenced strains confirmed the presence of several prophage loci within each of the genomes (Fig. 2a). The prophages include two that are shared by all strains (i.e. Φ1 and Φ2, Fig. 2a) as well as prophages which are unique within their encoding genome. On average less than half of all integrated prophages per strains are active of which the majority could be identified as *Inoviridae*. We found that R bacteria contain only one active prophage, while IS and HS bacteria contain on average two and three active prophages.

Based on all *Vibrio* strains and their induced prophages we generated a three-fold replicated 75 × 75 cross-infection matrix resulting in 16,875 inoculations. Among the 75 tested lysogens, 74 were homoimmune, i.e. immune to lytic infection by their own phage-lysates. The observed phage bacteria infection network (PBIN) is significantly nested: NODF nestedness score = 80.88; z-score = 126.63; $p < 0.001$ (Fig. 1b, for single matrices

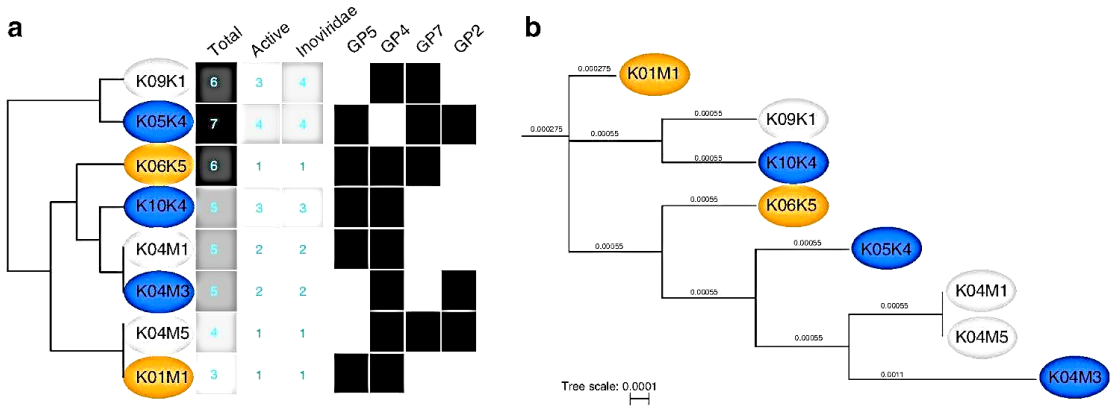


Fig. 2 **a** Clustering of the eight sequenced strains based on a similarity matrix containing the total number of prophages (Total), the total number of active prophages (Active), the total number of inoviridae (Inoviridae) and presence (*black square*)/absence (white square) of shared prophages that have been found in more than one genome (Φ1, Φ2, Φ6, Φ18) different colours correspond to bacterial groups: orange: HS bacteria, blue: IS bacteria and R: resistant bacteria; **b** Maximum likelihood tree based on whole genome alignment of the eight sequenced strains

see Additional file 5: Figure S1). Overall, 15.98% of the phage-bacteria combinations resulted in lytic infection success, which corresponds to a network connectance of 0.16. We observed that infections occurred only within the strains of the *Alginolyticus* clade, while the non-*alginolyticus* isolates could not get infected by any of the phage lysates nor could their phage suspensions infect any of the *V. alginolyticus* strains. Therefore we excluded non-*alginolyticus* bacteria from the rest of the analysis and the infection experiment on juvenile pipefish.

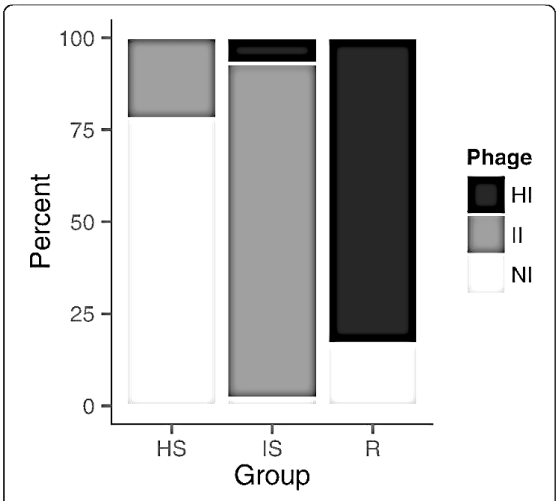


Fig. 3 Relative proportion of high infective (HI; *black*), intermediate infective (II; *dark grey*) and non-infective (NI; *light grey*) phage lysates per bacterial group (HS highly susceptible, IS: intermediate susceptible and R: resistant)

Most of the bacteria (82%) were susceptible to 13% of the phage lysates (thereafter called intermediate-susceptible (IS) bacteria), while 13% of the bacteria were highly susceptible to the majority (77%) of the phage lysates (thereafter called HS-bacteria). Approximately 5% of the bacteria were resistant against all phage lysates (thereafter called R-bacteria) whereas 10% of the phage lysates were not able to cause a visible lytic infection using a standard spot assay. Bacteria from these three phenotypic groups do not cluster based on their genotype (Fig. 2b). All three bacterial groups contained bacteria from diverse organs of different fish. Infection patterns could therefore not directly be linked to within population differentiation.

Within bacteria and phage lysates from the *Alginolyticus* clade we detected a significant infection pattern: five out of nine phage lysates from HS-bacteria were non-infectious, while the remaining four could infect other strains, which themselves were exclusively highly susceptible. In contrast, most phage lysates derived from R bacteria (3 out of 4) could infect the majority of the 71 *V. alginolyticus* strains, while only one phage lysate could not cause a lytic infection on any of the tested strains (Fig. 3).

Infection experiment

We used a controlled infection experiment on juvenile pipefish to directly test whether bacterial resistance to phages and bacterial harm to eukaryotic hosts can be linked. To control for clade effects all strains used in the infection experiment belonged to the *Alginolyticus* clade.

Viable Vibrio counts

Overall, the amount of CFU differed significantly between fish treated with PBS compared to fish infected with bacteria groups (Kruskal-Wallis test for non-parametric data: H = 11.96, $p < 0.001$, Additional file 6: Figure S2). However, there was no difference in CFU between all three bacterial groups (Kruskal-Wallis test for non-parametric data: H = 1.67, $p < 0.43$).

Gene expression

Bacterial group (HS, IS, R or control) significantly affected gene expression of infected juvenile pipefish, MANOVA (Pillai's trace = 2.2, Approx. $F_3 = 1.62$, $p = 0.01$). There was no difference in gene expression between sham-injected controls and pipefish infected with *Vibrio* strains resistant to phage infection (Fig. 4). However, gene expression differed significantly when pipefish were infected with *Vibrio* strains susceptible to phages. These observed differences in immune gene expression suggest that virulence on a eukaryotic host varies significantly between bacteria that have different phage-resistance phenotypes. Univariate ANOVAs revealed eleven genes that contribute to the observed significant group effect. Among these eleven genes, four genes belong to the innate and three to the adaptive immune system, while one gene belongs to the complement system and three genes are involved in gene silencing or deactivation (Additional file 7: Table S4).

Genome screening

We found no differences in the structure of the CRISPR/Cas system among the complete sequenced *V. alginolyticus* strains. A comprehensive screening for virulence factors revealed the presence of a gene that encodes a zona occludens toxin (ZOT) like protein in each genome. No other virulence factors were found in the genomes. All *V. alginolyticus* strains display nearly complete modification of the GATC motif (m6A, underlined is the methylated base) as of 99%. Around 20% of all CCAGCANY (m4C) motifs were modified additionally. Low methylation fractions of strain K05K4 cannot be taken into account as the coverage requirements of 50% were not met.

Bacterial properties

Growth rate

There was no difference in bacterial growth rate in Medium 101 over a 24 h period, linear mixed effect model, $F_{2,\ 6} = 3.81$, $p = 0.09$.

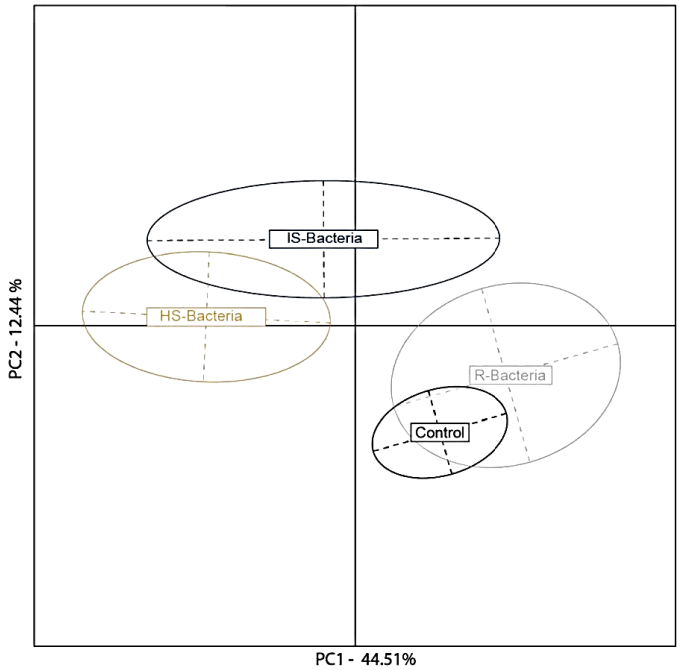


Fig. 4 Ordination of differentially expressed immune-genes between four different infection treatments. Juvenile pipefish were either sham injected with PBS (*black*), infected with bacteria resistant to phage infection (*grey*), intermediate susceptible bacteria (*blue*), or highly susceptible bacteria (*orange*). Note: all experiments have been performed with strains of *Vibrio alginolyticus*

Lysis time

Lysis time varied significantly between bacteria, which contain prophages that differ in their infection profile, linear model, $F_2 = 7.5$, $p = 0.001$. 'Highly-infective' phage lysates lysed their bacterial hosts on average after 81 min, while 'intermediate-infective' and 'non-infective' phage lysates took on average 105 and 100 min, respectively. Follow-up analysis revealed a significant difference in lysis time between bacteria that contain 'highly-infective' phage lysates and bacteria that contain 'non-infective' phage lysates $t_{62} = -2.68$, $p = 0.025$, as well as between 'highly-infective' phage lysates and 'intermediate-infective' phage lysates $t_{62} = -3.86$, $p < 0.001$. There was no significant difference in mean lysis time between 'intermediate-infective' and 'non-infective' phage lysates $t_{62} = 0.21$, $p = 0.98$.

Twitching motility

There was no significant difference in twitching motility between resistant and susceptible bacteria $F_{2,\ 63} = 0.098$, $p = 0.91$, indicating that pilus mutations, which could lead to reduced motility, were not the primary form of resistance against phages.

Discussion

Virulence shifts through a hyperparasite can change dual species interactions with profound implications on ecosystem dynamics and human health [37]. We empirically investigated a tripartite host-parasite interaction focusing on two players each, namely phage infectivity against bacteria as well as bacterial virulence against a eukaryotic host and found evidence that both two-way interactions are linked. We could induce prophages from all bacterial isolates indicating that lysogeny is common in the genus *Vibrio*. We then determined *Vibrio* resistance to each of the phage lysates and tested the virulence of nine selected *Vibrio* strains against their final eukaryotic hosts. Our results suggest that phage-resistant strains are less harmful to their eukaryotic host than phage-susceptible strains. These findings indicate that bacteria with a phage susceptible phenotype are associated with higher virulence against eukaryotic hosts.

Infectivity of phage lysates can be linked to bacterial resistance against superinfecting phages

The structure of phage-bacteria infection networks (PBINs) can range from random matrices over nearly diagonal matrices and nested structures to block-like matrices that exhibit high degrees of modularity [38, 39]. By generating a replicated 75 × 75 cross-infection matrix of *Vibrio* bacteria and phage lysates that were obtained from these bacteria by prophage induction, we found a clear-cut pattern between phage infection success and genetic distance of the host: *V. alginolyticus* genotypes were susceptible to phages from the same clade, but resistant to phages isolated from the *Splendidus* and the *Vulnificus* clade and vice versa. We are aware that the present PBIN comprises three different *Vibrio* clades with unequal sample sizes between clades and thus constrain the following discussion to the observed patterns within the *Alginolyticus* clade only.

Within the *Alginolyticus* clade we found a significantly nested structure (Fig. 1b). Nestedness results from sequences of gene-for-gene (GFG) coevolutionary adaptations and is the most common pattern in PBINs of natural communities [38, 39] but also in evolution experiments [40–42]. A nested structure results from cumulative GFG adaptations of bacterial resistance and phage infectivity, which confer resistance/infectivity against recently evolved phages/bacteria [39]. As a result, nested PBINs contain hierarchical interactions of phages and bacteria, which can be ordered according to the number of host genotypes/phage genotypes they can infect/resist. Likewise, according to their susceptibility to phages, bacteria from the present study can be grouped into three distinct categories: highly susceptible (HS), intermediate susceptible (IS) and resistant (R). This observed hierarchy indicates strong bacteria genotype by phage genotype interactions (GxG) and underlying GFG-like coevolutionary processes that characterize the present PBIN.

We found that 74 out of 75 bacterial isolates were immune to infection by their own lysate, indicating that homoimmunity is common for temperate filamentous *Vibriophages*. Indeed, most prophages immunize their host against their own kind and against phages of the same immunity group [4] for exceptions see [43]. We assume that a lytic infection in our spot assay is not possible if the superinfecting phage is homoimmune, i.e. it belongs to the same immunity group than the integrated prophage.

According to their infection pattern the 71 *alginolyticus* lysates could be grouped into 37 distinct groups, out of which 30 isolates had a unique infection profile. We further observed that most phage lysates isolated from HS-bacteria were non-infectious, while most phage lysates isolated from resistant bacteria could infect the majority of the tested bacteria isolates (Fig. 3). In addition, lysis time differed significantly between highly infective and non-infective as well as intermediate-infective phage lysates. Based on all these observed phenotypic properties we thus conclude that phage lysates of closely related host strains are different from each other.

Nevertheless, these phenotypic properties as well as the observed nestedness in the present infection matrix needs to be interpreted carefully by taking the potential multi-phage nature of the lysates into account. Whole genome sequencing of eight selected *Vibrio* strains indicates that resistant bacteria have more active prophages

than susceptible bacteria (Fig. 2a). We could not detect a clear-cut pattern between bacterial-resistance phenotypes and the presence of particular phages, which are shared across genomes (Fig. 2a) nor across *Vibrio* phylogeny (Fig. 2b). It is thus tempting to speculate that resistance to phages and infectivity of the lysate correlates with the number of active prophages. In the first case we assume that more phages protect the bacterium from infection by additional phages, for instance by actively eliminating the infecting phage. In the latter case we predict that the probability to infect any given strain is higher the more active phages a lysate contains. If this holds true, the observed nested structure of the present PBIN may not be exclusively the result of classical GFG evolution between bacterial genotypes and phage genotypes (GxG) but rather a complex combination of underlying coevolutionary processes between lysogens (bacterial genomes plus integrated phage genomes) and phages [(G + G)xG].

The number of integrated prophages is not the sole factor that can influence bacterial resistance. Such an infection pattern could additionally be impacted by the presence of bacteriocins, e.g. colicin, which can also confer homoimmunity [44], the restriction modification system [45] or the involvement of the CRISPR/Cas system, which provides acquired immunity against mobile genetic elements by targeting invasive DNA in a sequence specific manner [46]. Based on the eight fully sequenced genomes we could not detect any differences in virulence factors, neither in the CRISPR/Cas system nor in the methylome of those strains. Mutations on specific cell surface components were assigned as an alternative mechanism explaining resistance to phages, for instance pili, which represent the main entry site for filamentous phages [47]. However, follow up experiments detected no difference in twitching motility between IS, HS and R bacteria, rejecting the hypothesis that R bacteria are resistant to superinfecting phages due to a pilus deficient mutant.

Phage susceptible bacterial phenotypes may be associated with higher virulence against eukaryotic hosts

While it is acknowledged that prophages play an important role in bacterial virulence and evolution [48], the coupling between bacterial virulence against eukaryotic hosts and bacterial resistance against temperate phages has received little attention. Using a controlled infection experiment with selected strains that vary in their resistance to temperate phages, we tested whether bacterial resistance to phages and bacterial harm to eukaryotic hosts can be linked. While the amount of CFU in infected pipefish did not differ among treatment groups, host transcriptional response, notably expression of immune genes differed significantly between phage resistant and phage susceptible bacteria. We suggest that this observed difference in immune gene expression is linked to differences in virulence between phage resistant and phage susceptible strains. This indicates that the harm to the eukaryotic host and thus the virulence of a phage-resistant strain is significantly lower compared to the harm by a phage-susceptible strain.

The observed resistance-virulence trade-off has been frequently observed with lytic phages [6, 49, 50], for a recent review see [2] but has to our knowledge never been described for temperate phages. Common mechanisms/theories from studies using lytic phages explaining this trade-off in gram-negative bacteria are modifications of cell wall receptors, such as outer membrane proteins (OMPs) and Lipopolysaccharides (LPS) or bacterial appendices, such as flagellae or pili [2]. As known, filamentous phages enter the bacterium via the pilus [51], and no difference in twitching motility between phage susceptible and phage resistant strains could be detected, which would have suggested pilus-deficient mutants, the above mentioned mechanisms cannot explain the observed pattern. So far, we lack insight into the exact mechanism that couples virulence against eukaryotic hosts and resistance to temperate phages. The major difference between those closely related isolates is due to different prophages at different insertion sites, which can explain the distinct phenotypes. Thus we assume, that temperate phages are involved in mediating bacterial virulence and resistance.

There are different ways how prophages can contribute to the success of their bacterial hosts during infection. On the one side, prophages, and in particular filamentous phages are capable of influencing the virulence and evolution of their host by lysogenic conversion (for a recent review see [52]), with the most prominent example being the *Vibrio cholerae* CTXΦ phage carrying the cholera toxin gene [10]. However, in the case of prophages that do not contribute a clear phenotype such as virulence genes [53], their contribution to the fitness of the bacterial host is still unknown. In this context, we found that virulent strains (HS- and IS- bacteria) contain on average less active prophages than non-virulent strains (R-bacteria). In addition, the harm of selected strains did not depend on the presence of specific active prophages. Thus, our results suggest that (1) filamentous vibriophages do not always increase bacterial virulence but can also have opposite effects and (2) prophages may have more subtle effects on bacterial virulence apart from providing specific virulence toxins.

Conclusion

Based on an empirical approach that goes beyond a classical dual host-parasite interaction, we show that phage-resistant bacteria strains harm their eukaryotic

host less than phage-susceptible bacteria strains. These results illustrate the importance of hyperparasitism and that dual host-parasite interactions should not be studied in isolation. Ecological and evolutionary outcomes predicted by classical pairwise interactions differ profoundly, if we take additional players into account [54–56]. However, multiplayer interactions are only beginning to be explored [55], and are mostly limited to host-plant interactions as reviewed in [56], while studies using animal hosts are rare.

Phages are the most abundant entity in aquatic systems [57, 58] and their ecological importance in the marine environment has gained much attention in the last decade; for detailed reviews see [59–62]. Especially prophages have become recognized as important components of the marine environment through their ability to manipulate bacterial properties, such as pathogenicity. Our experimental results demonstrate that if we are to understand the spread and evolution of prophage-mediated diseases, it is paramount to take an integrative view across more than two levels by considering the interaction between all species involved.

Additional files

Additional file 1: Table S1. MLSA Primer information. (DOCX 14 kb)

Additional file 2: Table S5. The number of postfiltered reads and the average read length of the reads. (DOCX 14 kb)

Additional file 3: Table S2. Fluidigm Primer information. (DOCX 20 kb)

Additional file 4: Table S3. Average nulceotide identity between complete sequenced *Vibrio alginolytcus* genomes. (DOCX 14 kb)

Additional file 5: Figure S1. Original sorted and nested sorted matrices of each replicate of the qualitative assays. Rows and columns represent bacteria and phages. A black square indicates an interaction, i.e. infection success as determined by plaque formation. White cells refer to no infection, i.e. absence of plaques. (PDF 4077 kb)

Additional file 6: Figure S2. Number of colony forming units in infected pipefish differentiated by bacterial group as well as non-infected pipefish (PBS control). (PDF 117 kb)

Additional file 7: Table S4. Univariate ANOVAs of each immune gene of pipefish infected with R-, IS-, and HS bacteria. Bacterial group was treated as a fixed factor and each single strain was nested in its bacterial group. Significant *p*-values are presented in boldface. (DOCX 34 kb)

Abbreviations

CFU: Colony forming units; HI-, II-, NI-phages: Highly infective, intermediate infective, non infective phages; HS-, IS-, R-bacteria: High susceptible, intermediate susceptible, resistant bacteria; MLSA: Multi locus sequence analysis; PBN: Phage bacteria infection network

Acknowledgements

We thank Katja Trübenbach, Maude Poirier, Svenja Köpper, Tanita Wein, Marie Küter and Henrieke Wunderow for their support during infection experiments as well as in the laboratory. We thank Cathrin Spröer, Nicole Heyer and Simone Severitt for technical assistance in complete genome sequencing of the eight *Vibrio* isolates. We further thank Lasse Riemann, Hinrich Schulenburg and Thorsten Reusch for helpful discussions and input during the development of this research project and Mike Brockhurst for valuable comments on a previous version of the manuscript.

Funding

This study was supported by a grant from the Volkswagen Programme "Evolutionary Biology" given to OR and a start-up grant from the Cluster of Excellence "The Future Ocean" given to OR and CCW.

Availability of data and materials

Sanger sequences for the 16s rRNA, *recA* and *pyrH* locus are available at GenBank (16S rRNA: KY747252-KY747325, recA: KY771247 - KY771320, pyrH: KY771174 - KY771246). Whole *Vibrio* genomes are available at GenBank (CP017889 - CP017919). All other data sets are available at PANGAEA under (doi: 10.1594/PANGAEA.873510).

Authors' contribution

OR and DR initiated this study and established the phage-bacteria system in the laboratory. AP and OR performed the phage-bacteria infection matrix. CCW did the multilocus genotying of the bacteria strains. CCW and OR conducted the pipefish-bacteria infection experiment. Statistics of all laboratory experiments were done by CCW and AP. CCW, RH, HL, BB and JO performed bacteria genome sequencing. CC, HL and BB analysed the bacteria genomes. CCW and OR coordinated the project and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Approval for using pipefish during infection experiments was given by the Ministerium für Landwirtschaft, Umwelt und ländliche Räume des Landes Schleswig-Holstein.

Publisher's Note

Author details

[1]GEOMAR, Helmholtz Centre for Ocean Research, Evolutionary Ecology of Marine Fishes, Düsternbrooker Weg 20, 24105 Kiel, Germany. [2]Present address: Max Planck Institute for Evolutionary Biology, Department of Evolutionary Ecology, August-Thienemann-Straße 2, 24306 Plön, Germany. [3]Institute of Natural Resource Sciences, Zurich University of Applied Sciences, School of Life Sciences and Facility Management, Campus Grüental, CH-8820 Wädenswil, Switzerland. [4]Institute for Microbiology and Genetics, Georg-August University Goettingen, Grisebachstr. 8, 37077 Goettingen, Germany. [5]Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, Inhoffenstr. 7B, 38124 Braunschweig, Germany.

Received: 3 March 2017 Accepted: 9 March 2017
Published online: 11 April 2017

References

1. Parratt SR, Laine AL. The role of hyperparasitism in microbial pathogen ecology and evolution. Isme J. 2016;10:1815–822.
2. Leon M, Bastias R. Virulence reduction in bacteriophage resistant bacteria. Front Microbiol. 2015;6:343.
3. Casjens S. Prophages and bacterial genomics: what have we learned so far? Mol Microbiol. 2003;49(2):277–300.
4. Refardt D. Within-host competition determines reproductive success of temperate bacteriophages. Isme J. 2011;5(9):1451–60.
5. Paul JH. Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? Isme J. 2008;2(6):579–89.
6. Hosseinidoust Z, van de Ven TG, Tufenkji N. Evolution of *Pseudomonas aeruginosa* virulence as a result of phage predation. Appl Environ Microbiol. 2013;79(19):6110–6.
7. Laanto E, Bamford JKH, Laakso J, Sundberg LR. Phage-driven loss of virulence in a fish pathogenic bacterium. Plos One. 2012;7(12):e53157.
8. Lan SF, Huang CH, Chang CH, Liao WC, Lin IH, Jian WN, Wu YG, Chen SY, Wong HC. Characterization of a new plasmid-like prophage in a pandemic *Vibrio parahaemolyticus* O3:K6 Strain. Appl Environ Microb. 2009;75(9):2659–67.

9. Wagner PL, Waldor MK. Bacteriophage control of bacterial virulence. Infect Immun. 2002;70(8):3985–93.
10. Waldor MK, Mekalanos JJ. Lysogenic conversion by a filamentous phage encoding cholera toxin. Science. 1996;272(5270):1910–4.
11. Roth O, Keller I, Landis SH, Salzburger W, Reusch TB. Hosts are ahead in a marine host-parasite coevolutionary arms race: innate immune system adaptation in pipefish *Syngnathus typhle* against *Vibrio phylotypes*. Evolution. 2012;66(8):2528–39.
12. Balcazar JL, Gallo-Bueno A, Planas M, Pintado J. Isolation of *Vibrio alginolyticus* and *Vibrio splendidus* from captive-bred seahorses with disease symptoms. Antonie Van Leeuwenhoek. 2010;97(2):207–10.
13. Alcaide E, Gil-Sanz C, Sanjuan E, Esteve D, Amaro C, Silveira L. *Vibrio harveyi* causes disease in seahorse, *Hippocampus sp.* J Fish Dis. 2001;24(5):311–3.
14. Wendling CC, Batista FM, Wegner KM. Persistence, seasonal dynamics and pathogenic potential of Vibrio communities from Pacific oyster hemolymph. PLoS One. 2014;9(4):e94256.
15. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–9.
16. Clokie MRJ, Kropinski AM. Bacteriophages: methods and protocols, volume 1: isolation, characterization, and interactions. New York: Humana Press; 2008.
17. Sen A, Ghosh AN. Physicochemical characterization of vibriophage N5. Virol J. 2005;2:27.
18. Asadulghani M, Ogura Y, Ooka T, Itoh T, Sawaguchi A, Iguchi A, Nakayama K, Hayashi T. The defective prophage pool of Escherichia coli O157: prophage-prophage interactions potentiate horizontal transfer of virulence determinants. PLoS Pathog. 2009;5(5):e1000408.
19. Hardy KG, Meynell GG. "Induction" of colicin factor E2-P9 by mitomycin C. J Bacteriol. 1972;112(2):1007–9.
20. Wendling CC, Wegner KM. Relative contribution of reproductive investment, thermal stress and Vibrio infection to summer mortality phenomena in Pacific oysters. Aquaculture. 2013;412–413:88–96.
21. Beemelmanns A, Roth O. Biparental immune priming in the pipefish Syngnathus typhle. Zoology. 2016;119(4):262–72.
22. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
23. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012;61(3):539–42.
24. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001;17(8):754–5.
25. Posada D, Buckley TR. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst Biol. 2004;53(5):793–808.
26. Angiuoli SV, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics. 2011;27(3):334–42.
27. Sahl JW, Matalka MN, Rasko DA. Phylomark, a tool to identify conserved phylogenetic markers from whole-genome alignments. Appl Environ Microbiol. 2012;78(14):4884–92.
28. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5(3):e9490.
29. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics. 2007;23(1):127–8.
30. Dhillon BK, Laird MR, Shay JA, Winsor GL, Lo R, Nizam F, Pereira SK, Waglechner N, McArthur AG, Langille MG, et al. IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. Nucleic Acids Res. 2015; 43(W1):W104–8.
31. Shao Y, Harrison EM, Bi D, Tai C, He X, Ou HY, Rajakumar K, Deng Z. TADB: a web-based resource for Type 2 toxin-antitoxin loci in bacteria and archaea. Nucleic Acids Res. 2011;39(Database issue):D606–11.
32. Naamati G, Askenazi M, Linial M. ClanTox: a classifier of short animal toxins. Nucleic Acids Res. 2009;37(Web Server issue):W363–368.
33. Dormann C, Gruber B, Fründ J. Introducing the bipartite package: analysing ecological networks. R News. 2008;8(2):8–11.
34. Beckett SJ, Boulton CA, Williams HT. FALCON: a software package for analysis of nestedness in bipartite networks. F1000Res. 2014;3:185.
35. Hellemans J, Mortier G, De Paepe A, Speleman F, Vandesompele J. qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. Genome Biol. 2007;8(2):R19.
36. Dray S, Dufour AB. The ade4 package: implementing the duality diagram for ecologists. J Stat Softw. 2007;22(4):1–20.
37. Tinsley CR, Bille E, Nassif X. Bacteriophages and pathogenicity: more than just providing a toxin? Microbes Infect. 2006;8(5):1365–71.
38. Flores CO, Meyer JR, Valverde S, Farr L, Weitz JS. Statistical structure of host-phage interactions. Proc Natl Acad Sci U S A. 2011;108(28):E288–97.
39. Weitz JS, Poisot T, Meyer JR, Flores CO, Valverde S, Sullivan MB, Hochberg ME. Phage-bacteria infection networks. Trends Microbiol. 2013;21(2):82–91.
40. Luria SE. Mutations of bacterial viruses affecting their host range. Genetics. 1945;30(1):84–99.
41. Meyer JR, Dobias DT, Weitz JS, Barrick JE, Quick RT, Lenski RE. Repeatability and contingency in the evolution of a Key innovation in phage Lambda. Science. 2012;335(6067):428–32.
42. Poullain V, Gandon S, Brockhurst MA, Buckling A, Hochberg ME. The evolution of specificity in evolving and coevolving antagonistic interactions between a bacteria and its phage. Evolution. 2008;62(1):1–11.
43. Faruque SM, Bin Naser I, Fujihara K, Diraphat P, Chowdhury N, Kamruzzaman M, Qadri F, Yamasaki S, Ghosh AN, Mekalanos JJ. Genomic sequence and receptor for the Vibrio cholerae phage KSF-1phi: evolutionary divergence among filamentous vibriophages mediating lateral gene transfer. J Bacteriol. 2005;187(12):4095–103.
44. Herschman HR, Helinski DR. Comparative study of events associated with colicin induction. J Bacteriol. 1967;94(3):691–9.
45. Tock MR, Dryden DT. The biology of restriction and anti-restriction. Curr Opin Microbiol. 2005;8(4):466–72.
46. Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. Science. 2010;327(5962):167–70.
47. Rakonjac J, Bennett NJ, Spagnuolo J, Gagic D, Russel M. Filamentous bacteriophage: biology, phage display and nanotechnology applications. Curr Issues Mol Biol. 2011;13(2):51–76.
48. Canchaya C, Fournous G, Brussow H. The impact of prophages on bacterial chromosomes. Mol Microbiol. 2004;53(1):9–18.
49. Addy HS, Askora A, Kawasaki T, Fujie M, Yamada T. Loss of virulence of the phytopathogen *Ralstonia solanacearum* through infection by phiRSM filamentous phages. Phytopathology. 2012;102(5):469–77.
50. Hosseinidoust Z, Tufenkji N, van de Ven TG. Predation in homogeneous and heterogeneous phage environments affects virulence determinants of Pseudomonas aeruginosa. Appl Environ Microbiol. 2013;79(9):2862–71.
51. Mai-Prochnow A, Hui JG, Kjelleberg S, Rakonjac J, McDougald D, Rice SA. 'Big things in small packages: the genetics of filamentous phage and effects on fitness of their host'. Fems Microbiol Rev. 2015;39(4):465–87.
52. Ilyina TS. Filamentous bacteriophages and their role in the virulence and evolution of pathogenic bacteria. Mol Genet Microbiol. 2015;30(1):1–9.
53. Davies EV, Winstanley C, Fothergill JL, James CE. The role of temperate bacteriophages in bacterial infection. Fems Microbiol Lett. 2016;363(5): fnw015.
54. Stanton ML. Interacting guilds: moving beyond the pairwise perspective on mutualisms. Am Nat. 2003;162(4 Suppl):S10–23.
55. Goodnight CJ. Evolution in metacommunities. Philos Trans R Soc Lond B Biol Sci. 2011;366(1569):1401–9.
56. Strauss SY, Irwin RE. Ecological and evolutionary consequences of multispecies plant-animal interactions. Annu Rev Ecol Evol S. 2004;35:435–66.
57. Bergh O, Borsheim KY, Bratbak G, Heldal M. High abundance of viruses found in aquatic environments. Nature. 1989;340(6233):467–8.
58. Proctor LM, Fuhrman JA. Viral mortality of marine-bacteria and Cyanobacteria. Nature. 1990;343(6253):60–2.
59. Breitbart M. Marine viruses: truth or dare. Annu Rev Mar Sci. 2012;4:425–48.
60. Rohwer F, Thurber RV. Viruses manipulate the marine environment. Nature. 2009;459(7244):207–12.
61. Suttle CA. Viruses in the sea. Nature. 2005;437(7057):356–61.
62. Suttle CA. Marine viruses - major players in the global ecosystem. Nat Rev Microbiol. 2007;5(10):801–12.

Supplementary information

Supplementary information for this manuscript can be found at the "BMC Evolutionary Biology" website under the following address:

https://bmcevolbiol.biomedcentral.com/articles/10.1186/s12862-017-0930-2#Bib1

Additionally, supplementary figures and tables are provided along with the electronic version of this thesis (on DVD), under the following paths:

Additional Figures:

Additional file 5: SupplementaryMaterial/ChapterII/ChapterII.1/Figure S1.pdf

Additional file 6: SupplementaryMaterial/ChapterII/ChapterII.1/Figure S2.pdf

Additional Tables:

Additional file 1: SupplementaryMaterial/ChapterII/ChapterII.1/Table S1.docx

Additional file 2: SupplementaryMaterial/ChapterII/ChapterII.1/Table S5.docx

Additional file 3: SupplementaryMaterial/ChapterII/ChapterII.1/Table S2.docx

Additional file 4: SupplementaryMaterial/ChapterII/ChapterII.1/Table S3.docx

Additional file 7: SupplementaryMaterial/ChapterII/ChapterII.1/Table S4.docx

II.2 Comparative genomic analysis of *Vibrio alginolyticus* reveals that the dynamics lie within the mobilome

Comparative genomic analysis of *Vibrio alginolyticus* reveals that the dynamics lie within the mobilome

Cynthia Maria Chibani, Robert Hertel, Olivia Roth, Michael Hoppert, Heiko Liesegang, Carolin Charlotte Wendling

Comparative genomic analysis of *Vibrio alginolyticus* reveals that the dynamics lie within the mobilome.

Cynthia Maria Chibani[1], Robert Hertel[1], Olivia Roth[2], Michael Hoppert[3], Heiko Liesegang[1], Carolin Charlotte Wendling[2].

Institutional Affiliation

1. Institute of Microbiology and Genetics, Department of Genomic and Applied Microbiology, Georg-August-University, 37077 Göttingen, Germany. e-mail: cchiban@gwdg.de, rhertel@gwdg.de, hlieseg@gwdg.de
2. GEOMAR, Helmholtz Centre for Ocean Research, Evolutionary Ecology of Marine Fishes,DuesternbrookerWeg 20, 24105 Kiel, Germany. e-mail: cwendling@geomar.de, oroth@geomar.de
3. Institute for Microbiology and Genetics, University of Göttingen, Göttingen, Germany. E-mail: mhoppet@gwdg.de

Corresponding author: Dr. Carolin Charlotte Wendling; cwendling@geomar.de

Abstract

Vibrio alginolyticus is a ubiquitous Gram-negative halophilic opportunistic pathogen, causing mass mortalities in shellfish, shrimps,and fish resulting in worldwide economic losses. The organism is considered as an independent species since 1980 and is closely related to the *Harveyi*clade, a group of seven species within the genus *Vibrio*. *V. alginolyticus*, as a species, comprise strains that are adapted to live as commensal as well as pathogenic bacteria within

habitats provided by a host organism. Considering the closely related species within the harveyi clade, this study targets the question of what genetic elements contribute to making *V. alginolyticus* a species. This study was performed to especially elucidate the contribution of mobile genetic elements, i.e., phages and plasmids, to the adaptation of *V. alginolyticus* strains to their host, their niche as well as to the switch between a commensal and a pathogenic lifestyle.

Here we present a comparative genomic analysis of nine sequenced *Vibrio alginolyticus* isolates with a focus on infecting *Inoviridae* phages. We show that those infecting phages encode a toxin similar to the closely related CTX-phage known to infect various *V. cholera* strains. Altogether, our analysis revealed that genomic fluidity reflected by the presence of extra-chromosomal phages, prophages,and plasmids specific for the habitat facilitates the understanding of the phylogenetic diversity as well as the emergence of virulence of the various studied strains.

Keywords

Vibrio alginolyticus- comparative genomics - mobile genetic elements - mega-plasmids - *Inoviridae*- ssDNA phages - vibriophage - bacteriophages - phage activity - pathogenicity.

1. Introduction

Vibrio alginolyticus is a ubiquitous marine opportunistic pathogen can cause mass mortalities in shellfish, shrimp, and fish, resulting in severe economic losses worldwide (Zhang et al., 2014; González-Escalona, Blackstone, & DePaola, 2006; Lee, Yu, Yang, Liu, & Chen, 1996). Additionally, wound infections and fatal septicemia in immunocompromised patients caused by *V. alginolyticus* have been reported in humans (Hörmansdorfer, Wentges, Neugebaur-büchler, & Bauer, 2000). *Vibrio* pathogenicity is a complex interaction of abiotic and biotic

factors (Defoirdt, 2014), including high temperatures (Harvell, Altizer, Cattadory, Harrington, & Weil, 2009), low salinities, host and bacterial genotypes (Le Roux et al., 2015) and the presence of virulence encoding prophages, termed 'vibriophages' (Lan et al., 2009; Wagner & Waldor, 2002). The contribution of vibriophages to *Vibrio* virulence is a well-studied phenomenon and best described for *V. cholera* and the filamentous phage CTX, which encodes the cholera toxin (CT). Upon integration into the *V. cholera* chromosome, the CTX phage can transform an avirulent*V. cholera* strain into a deadly pathogen (Sarkar, Chakrabarti, Sarkar, & Dutta, 2016; Waldor & Mekalanos, 1996).

Vibriophages are generally specific for a single *Vibrio* species or even specific to a single strain within a species (Maxwell, 2019). The CTX-phages, as well as the observed filamentous phages infecting other *Vibrionaceae*, have been classified as *Inoviridae* encoding genomes with a size from 4.5 Kbp to 12.4 Kbp (International Committee on Taxonomy of Viruses & King, 2012). *Inoviridae* can enter a lysogenic cycle by integrating their entire genome into their host genome followed by a passive replication by the host replication apparatus during cell division. Alternatively, they can have a lytic cycle where the virus genome replicates independently by a rolling-circle mechanism and hijacks the bacterial resources to produce the phage proteins and assemble new phage particles (Mai-Prochnow et al., 2015). In contrast to other lytic phages, who kill their host to release the free phage particles, the filamentous phage replication process results in a constant production of phage particles without killing the host cell, which is a distinct characteristic of filamentous phages (Mai-Prochnow et al., 2015). Most *Inoviridae* carry genes, which encode toxins, change their host's phenotype using lysogenic conversion (Waldor & Mekalanos, 1996). This process enables their host bacterium to exploit a eukaryotic host and ultimately, to adapt to and colonize new habitats (Wendling et al., 2017).

Turner et al. (2018) underline in their analysis that *Vibrios*related to the Harveyi clade are highly similar concerning the chromosomes (Turner et al., 2018). Thisis reflected by a shared core-genome consisting of ~ 4,800 chromosomally encoded genes, which is approximately 80% of the gene content of an average *V. alginolyticus* genome. The most significant share of strain-specific chromosomally encoded genes are located within mobile genetic elements such as plasmids and prophages. These prophages include *Inoviridae* to which the CTX infecting phage belongs to and other members of the *Caudovirales*phage family (Castillo et al., 2018).

In this study, we investigate the genomic sequences of nine different *V. alginolyticus* genomes, which have been isolated from the pipefish *Syngnathus typhle* at the Kiel Fjord (Wendling et al., 2017; Roth, Keller, Landis, Salzburger, & Reusch, 2012). We focus on the habitat-specific genes encoded on plasmids exclusively found in strains isolated in the Kiel Fjord and on prophages. We show that from the identified prophages solely the *Inoviridae* closely related to the CTX-phage concerning genome size, gene order and the presence of a toxin gene have been found actively producing phage particles.

2. Materials and Methods

Bacterial genome data

We compared all replicons from nine *V. alginolyticus* strains to 159 closed *Vibrio* replicons sequences downloaded from NCBI nucleotide database; date of accession 12.06.2018 (Table S7). The nine strains were described in an earlier study and were phylogenetically previously with multi-locus sequence analysis (MLSA) based on partial DNA of 3 different genes (16S rRNA, recA and pyrH) (Wendling et al., 2017).

DNA isolation, whole genome sequencing, assembly, and annotation

Using a combination of PacBio and Illumina sequencing, we generated eight closed *V. alginolyticus* genomes and one permanent draft as described in the following sections.

Prophage induction and sequencing

Prophages were induced from all nine *V. alginolyticus* strains using mitomycin C (Sigma) as described in Wendling et al. (2017) (Wendling et al., 2017) with minor modifications: bacteria were grown in liquid Medium101 (Medium101: 0.5% (w/v) peptone, 0.3% (w/v) meat extract, 3.0% (w/v) NaCl in MilliQ water) at 250 rpm and 25 °C overnight. Cultures were diluted 1:100 in fresh medium and grown for another 2.5 h at 250 rpm and 25 °C to bring cultures into exponential growth before adding mitomycin C at a final concentration of 0.5 µg/ml. Samples were incubated in an automated plate reader (TECAN infinite M200) for 4 h at 25 °C and mixed periodically. Bacterial lysis upon prophage induction was monitored via optical density at 600 nm (measured every other minute). We determined bacterial lysis time at induction as the time at which turbidity of the culture peaks.After 4 h, lysates were centrifuged at 6000 *g* for 15 min. The supernatant was sterile filtered using 0.45 µm pore size filter (Sarstedt, Nümbrecht, Germany) and consequently supplemented with lysozyme from chicken egg white (10µg/ml, SERVA Heidelberg, Germany) was added to the filtered supernatant to disrupt the cell walls of potentially remaining host cells. RNAse A (Quiagen, Hilden, Germany) and DNAse I (Roche Diagnostics, Mannheim, Germany) were added to a final concentration of 10µg/ ml each incubated at 25ºC for overnight (16 hours) to remove free nucleic acids and remaining host cells as described in Hertel et al.(Hertel et al., 2015). The supernatant was subsequently used for phage precipitation.

Ultracentrifugation

After the enzymatic removal of free nucleic acids, the phage particles were sedimented by ultracentrifugation using a Sorvall Ultracentrifuge OTD50B with a 60Ti rotor applying 200,000 g for 2 hours. The supernatant was discarded, and the pellet was solved in 200 µl TMK buffer, and stored at 4ºC or directly used for DNA isolation.

DNA Extraction

The DNA isolation was performed using a MasterPure DNA Purification kit from Epicenter (Madison, WI, USA). 200 µl 2x T&C-Lysis solution containing 1µl Proteinase K was added to the phage suspensions and incubated for 10 min at 10,000 g. The supernatant was transferred to a new tube, mixed with 670 µl cold isopropanol and incubated for 10 min at – 20ºC. DNA precipitation was performed by centrifugation for 10 min at 17,000 g and 4ºC. The DNA pellet was washed with twice with 150 µl 75% Ethanol, air-dried and re-suspended in DNase free water.

Next-generation sequencing

dsDNA for library construction was generated from viral ssDNA in a 50 µl reaction. The reaction was supplemented with 250 ng viral ssDNA dissolved in water, 1µM final concentration random hexamer primer (#SO142, Thermo Scientific), 10 units Klenow Fragment (#EP0051, Thermo Scientific) and 200 µM dNTPs final concentration each (#R0181, Thermo Scientific) and incubated for 37ºC for 2 hours. The reaction was stopped by adding 1µlof a 0.5M EDTA pH 8 solution. The generated DNA was precipitated by adding 5 µl of a 3M NaAcetate pH 5.2 and 50 µl100% Isopropanol to the DNA solution, gently mixing and chilling for 20 min at -70ºC. DNA was pelleted by centrifugation at 17,000 g, 4ºCand 10 min. Pellet was washed twice with 70% Ethanol and re-solved in 40ºCof pure water.

Remaining primers and viral ssDNA were removed in a 50 µl reaction using 10 units S1 nuclease (#EN0321, Thermo Scientific) for 30 min at 25ºC. S1 nuclease was inactivated through adding 1µM 0.5M EDTA pH 8 and incubated for 10 min at 70ºC. Consequently, dsDNA was precipitated as described above and resolved in pure water. Presence of dsDNA was verified via TAE gel electrophoresis in combination with an ethidium bromide staining and visualization via UV-light. NGXS phage DNA libraries were generated with the NexteraXT DNA Sample Preparation Kit (Illumina, San Diego, USA), and the sequencing was performed on an Illumina Gaii sequencer (Illumina, San Diego, USA).

Transmission electron microscopy

Electron microscopy was carried out on a Jeol 1011 electron microscope (Peabody, USA). Negative staining and transmission electron microscopy (TEM) were performed as described previously (Willms et al. 2017). Phosphotungstic acid dissolved in pure water (3%; pH 7) served as staining solution.

Average nucleotide identity and orthologous proteins

Average nucleotide identity (ANI) analysis of the different 159 *Vibrio* replicons was performed in ANIm mode which uses MUMmer (https://github.com/widdowquinn/pyani). Briefly, nucleotide sequences were extracted from each GenBank file using Biopython (https://biopython.org/) and subsequently used as input for pyani for genome sequence alignment.

To identify orthologous genes between the closest selected genomes from the pyani analysis and the nine sequenced *V. alginolyticus* strains, Proteinortho(Lechner et al., 2011)was used. Proteinortho cutoffs parameters used were an E-value of 1e-10 and protein sequence of 80% coverage and 50% identity. The nine sequenced strains and closest selected *Vibrio* genomes

were scanned with PATRIC (https://patricbrc.org/) for the detection of potential virulence factors (Wattam et al., 2014).

Determination of active phages

Resulting PacBio reads were *de novo* assembled using the HGAP 2.0 assembly pipeline (Chin et al., 2013) with further analysis using SMRT Portal (v2.3.0) to generate *V. alginolyticus* references genomes (https://www.pacb.com/support/software-downloads/). Resulting Illumina sequence reads from i) whole genome sequencing and ii) induced phage sequencing were mapped using Bowtie2 (Langmead & Salzberg, 2012) to the corresponding reference *V. alginolyticus* genome.

The generated mapping files were analyzed using TraV(Dietrich, Wiegand, & Liesegang, 2014) to identify the genomic location and context of the phage particle provided DNA. Peaks of coverage that mapped to genome region encoding phage genes were used as an indication for active prophages.

Prediction of phage loci and comparative analysis

All genomes were scanned with PHASTER (http://phaster.ca/) (Arndt et al., 2016) to identify additional non-induced prophages. Easyfig(Sullivan, Petty, & Beatson, 2011) with the BLASTn mode was used for pairwise phage sequence comparisons and synteny comparisons with an *E*-value cutoff of 1e−10.

Statistical analysis and visualization graphs

All statistics and visualization graphs were performed using ggplot2 (Wickham, 2011) library in R 3.1.2 unless otherwise stated.

3. Results and discussion

Genome sequencing

Nine strains of *Vibrio alginolyticus* isolated from pipefish at the Kiel Fjord(Roth et al., 2012) have been genome sequenced using PacBio long and Illumina short read technology. The assembly resulted in eight closed genome sequences of the nine *V. alginolyticus*strains. All *V. alginolyticus* genomes contain a ~3.47 Mbp chromosome 1 and a ~ 1.88 Mbp chromosome 2 (Table 1) as has been found previously for the genus *Vibrio* as well as for the species *V. alginolyticus* (Wang, Wen, Li, Zeng, & Wang, 2016; Okada, Iida, Kita-Tsukamoto, & Honda, 2005).

Table 1: *Vibrio alginolyticus* genomes used in this study

Strain	GC%	Replicon	Size[bp]	CDS	Ref	Genbank
K01M1	44.60	chromosome 1	3,468,303	3,206	This study	CP017889.1
		chromosome 2	1,883,748	1,668	This study	CP017890.1
		pL9064	9,064	8	This study	CP028135.1
K04M1	44.31	chromosome 1	3,473,127	3,213	This study	CP017891.1
		chromosome 2	1,870,775	1,660	This study	CP017892.1
		pL19	19,690	28	This study	CP017893.1
		pL280	280,614	305	This study	CP017894.1
		vK04M1*	7,079	11	This study	CP017895.1

K04M3	44.31	chromosome 1	3,476,174	3,219	This study	CP017896.1
		chromosome 2	1,903,830	1,708	This study	CP017897.1
		pL294	294,086	325	This study	CP017898.1
K04M5	44.31	chromosome 1	3,470,916	3,211	This study	CP017899.1
		chromosome 2	1,900,618	1,688	This study	CP017900.1
		pL294	294,721	320	This study	CP017901.1
K05K4	44.34	chromosome 1	3,473,579	3,218	This study	CP017902.1
		chromosome 2	1,875,554	1,670	This study	CP017903.1
		pL289	289,065	315	This study	CP017904.1
		vK05K4_1*	21,012	34	This study	CP017905.1
		vK05K4_2*	13,327	23	This study	CP017906.1
K06K5	44.31	chromosome 1	3,471,297	3,213	This study	CP017907.1
		chromosome 2	1,879,729	1,662	This study	CP017908.1
		pL29	29,688	20	This study	CP017909.1
		pL291	291,285	322	This study	CP017910.1
K08M3	44.32	chromosome 1	3,468,139	3,214	This study	CP017913.1

		chromosome 2	1,886,577	1,675	This study	CP017914.1
		pL300	300,425	331	This study	CP017915.1
K09K1	44.61	chromosome 1	1,897,210	3,209	This study	CP017918.1
		chromosome 2	3,465,619	1,704	This study	CP017919.1
K10K4	44.60	chromosome 1	3,494,647	3,231	This study	CP017911.1
		chromosome 2	1,894,531	1,682	This study	CP017912.1
ATCC 33787	44.48	chromosome 1	3,362,673	3,190	(Wang et al., 2016)	CP013484.1
		chromosome 2	1,851,538	1,674	(Wang et al., 2016)	CP013485.1
		pMBL128	128,112	144	(Wang et al., 2016)	CP013486.1
		pMBL287	286,750	301	(Wang et al., 2016)	CP013487.1
		pMBL96	95,866	109	(Wang et al., 2016)	CP013488.1
ZJ-T	44.67	chromosome 1	3,535,128	3,301	(Deng, YiqinChen, Zhao, Huang, Ding, & Yang, 2016)	CP016224.1
		chromosome 2	1,870,966	1,657	(Deng, YiqinChen, Zhao, Huang, Ding, & Yang, 2016)	CP016225.1

NBRC 15630	44.70	chromosome 1	3,334,467	3,128	(Liu, Cao, Zhang, Chen, & Hu, 2015)	CP006718.1
		chromosome 2	1,812,170	1,640	(Liu, Cao, Zhang, Chen, & Hu, 2015)	CP006719.1

*circular phage replicons

Seven isolates contain extra-chromosomal replicons including plasmids as has been found for strain ATCC33787 (Wang et al., 2016). The isolates K04M1 and K04K5 contain circular replicons encoding *Inoviridae* phages which fits to the observation that *Inoviridae* can replicate as extra-chromosomal circular molecules in a rolling circle replication mode (Wawrzyniak, Plucienniczak, & Bartosik, 2017; Székely & Breitbart, 2016; Mai-Prochnow et al., 2015) without killing their hosts by switching into the lytic lifestyle. In case of strain K09K1, the chromosomes 1 and 2 have been assembled into a single contig due to a multiple repeats that contained as several copies of integrated *Inoviridae* prophages. The replicon boundaries could not be resolved experimentally based on PCR; thus the *V. alginolyticus* K09K1 genome has been assigned "permanent draft" status.

Species definition and phylogenetic relationships

To elucidate the taxonomy of the nine genomes and 150 *Vibrio* replicons from closed sequenced genomes available at the time of analysis (for details see Table S7), the average nucleotide identity was performed using pyani with the ANIm option (https://github.com/widdowquinn/pyani) (Figure 1).

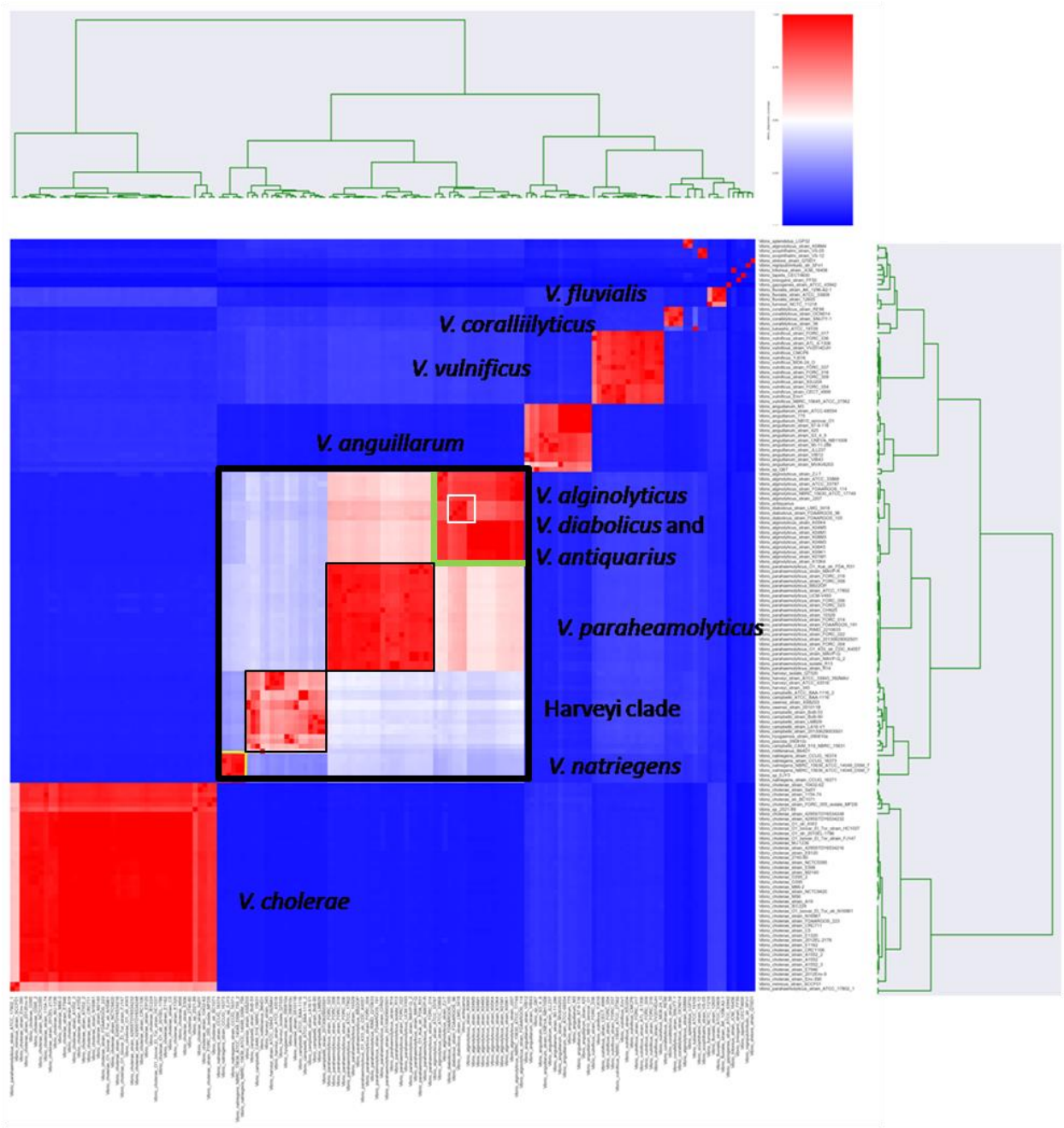


Figure 1: Average nucleotide identity percentage analysis (ranging from 0 to 50% colored in blue, and higher than 50% ANI in red, up to 100% ANI dark red) of closed *Vibrio* genomes.

ANI analysis based on MUMmer alignment of the genome sequences was performed and visualized using PYANI. All *V. alginolyticus* cluster with *V. diabolicus*and *V. antiquaries*(green box) on ANI similarity values

below the species level, the latter two form a subcluster (white box). The extended Harveyi group (black box) forms a cluster of four distinct species groups including (i) the Harveyi clade sensustricto (blue box), (ii) the parahaemolyticus group (light blue box), the (iii) natriegens group (yellow box) and the alginolyticus group (green box).

The species *V. alginolyticus*, *V. diabolicus*and *V. antiquaries* exhibited ANIm values between 96 to 100% (dark red color), which is above species threshold (Yoon, Ha, Lim, Kwon, & Chun, 2017; Goris et al., 2007). However, a distinct block within the alginolyticus/diabolicus group that contains *V. diabolicus* exclusivelyand *V. antiquaries* strains to indicate that they are more similar to each other than to *V. alginolyticus*. The analysis of the ANI clustering shows that members of the Harveyi clade (Ke et al., 2018) of the genus *Vibrio*, consisting of the species V. *harveyi, V. campbellii, V. hyugaensis, and V. owensii*, forms a close group with *V. jasidica*, *V. natriegens, V. rotiferianus*. This group can be clearly separated from *V. diabolicus*/*V. alginolyticus* cluster (Turner et al., 2018) and the *V. parahaemolyticus* cluster (Ghenem, Elhadi, Alzahrni, & Nishibuchi, 2017). Our data confirm the close taxonomic proximity of these species (Turner et al., 2018). The analysis confirms that the nine new genomes belong to the species *V. alginolyticus* and form with *V. diabolicus* and*V. antiquaries* a distinct species group. It is a species group that is related but distinct to the harveyi-clade and *V. parahaemolyticus*.

Pan/core genomes *Vibrio alginolyticus*

To elucidate how the nine genomes from *V. alginolyticus* strain from the Kiel Fjord are related to *V. alginolyticus* strains isolated from other habitats we determined the pan/core genomes (Land et al., 2015) of the new genomes with the *V. alginolyticus* type strain NBRCC15630 (isolated in Japan, (Liu et al., 2015)) and the strain ZJ-T (isolated in Zhanjiang, Guangdong Province, China(Deng, YiqinChen et al., 2016)) and ATCC33787 (isolated from sea-water near Oahu, USA (Wang et al., 2016)). The analysis included in total 53,893 proteins sequences

encoded in chromosomes, plasmids, extra-chromosomal phage replicons as well as in integrated prophages. The core genome calculate based on these 12 strains comprises 3,861 orthologous groups (Figure 2, for details, see Table Figure 2 and Table S6).

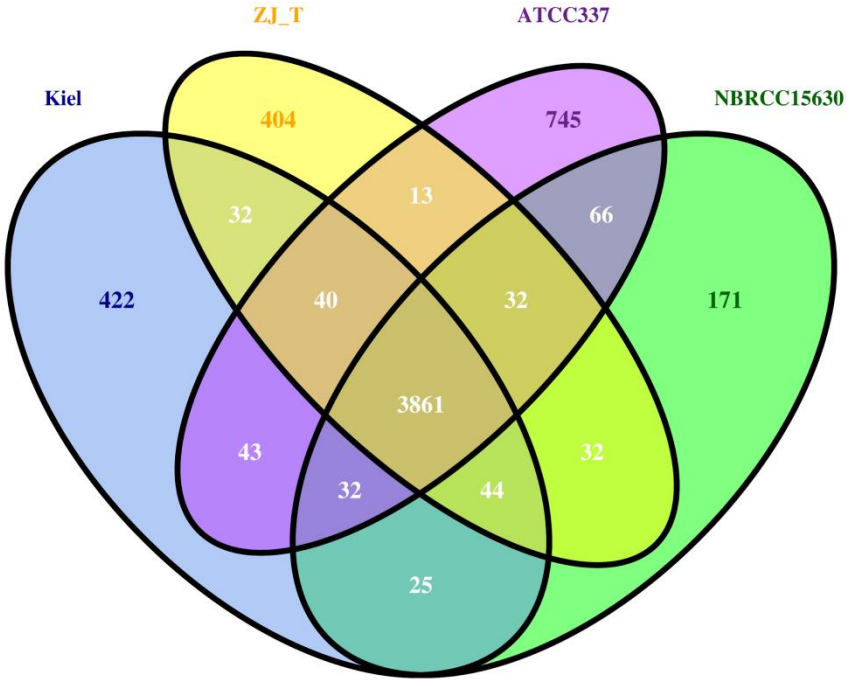


Figure 2: Pan/Core genome analysis of *V. alginolyticus* strains isolated from four different habitats. Note the Kiel habitat represents all genes shared by all 9 strains isolated from the Kiel Fjord. The other habitats are represented by strains ZJ-T (isolated from *Epinepheluscoioides* in Zhanjiang, Guangdong Province, China), ATCC337 (sea water near Oahu 20.3N 157.3 W) and NBRCC15630 by single isolates. Matches to a group of paralogs have been counted once per orthologous group. Notably, the number of genes shared by the different habitats is low compared to the number of habitat-specific genes.

The Proteinortho analysis indicates that the species defining core genome of *V. alginolyticus* includes approximately 79% of the genes in each sequenced genome, which is close to the 77%

of the core genome identified in *V. parahaemolyticus* (Gonzalez-Escalona, Jolley, Reed, & Martinez-Urtaza, 2017). The number of specific genes is varying between 171 and 422, which represent the adaption of strains to their particular habitat. The Venn diagram indicates what is shared between the *Vibrio* genus while the singletons indicate what is particular to the source of the strain isolation and related to the niche adaptation. Singletons, here defined as genes that are exclusively found in one habitat, have been checked for their genomic location (Figure 3, for details, see Table Figure 3).

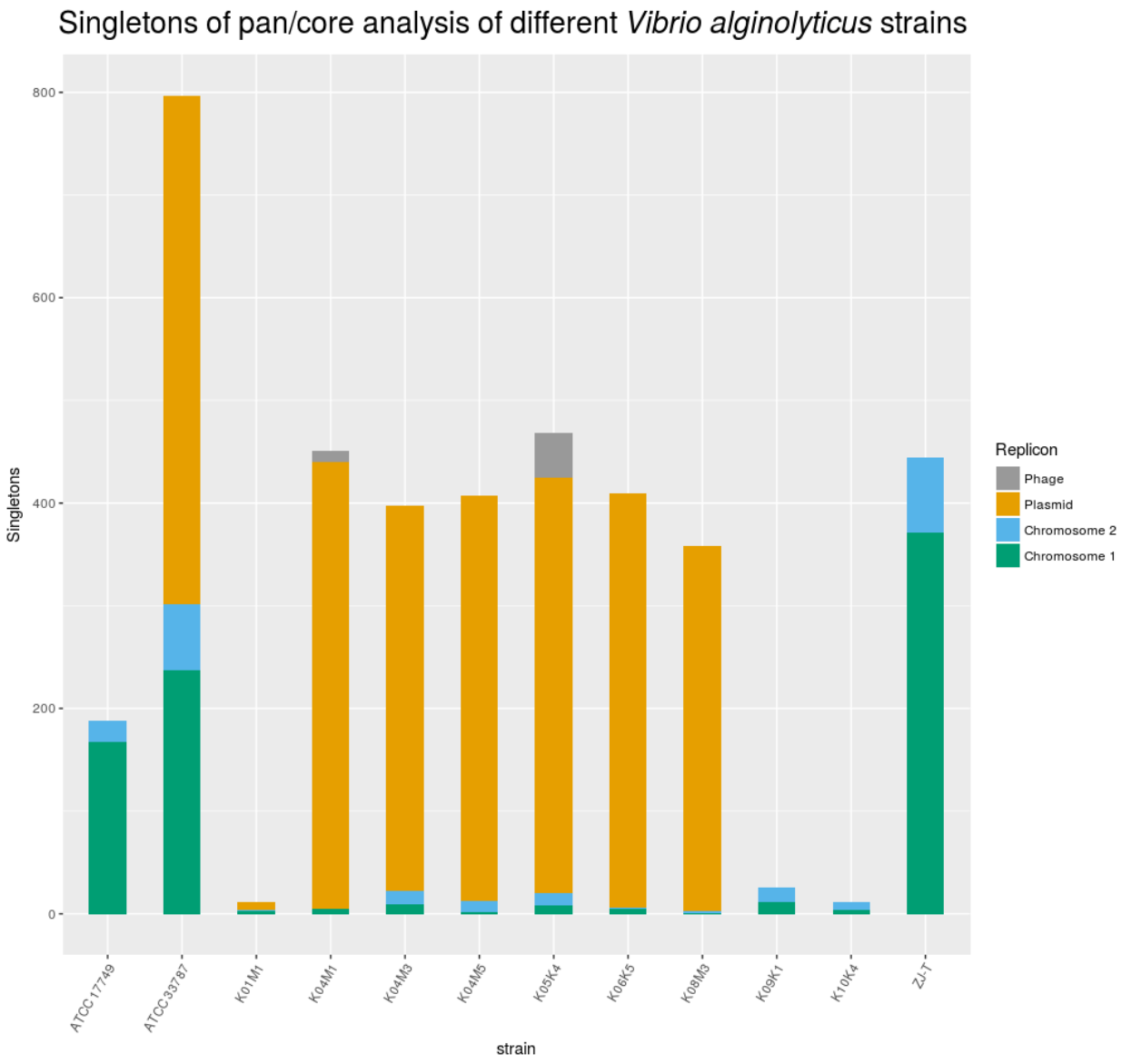


Figure 3: Singletons per habitat displayed in a stacked bar plot resulting from Core/Pan genome analysis of 12 *Vibrio alginolyticus* strains. The number of genome-specific singletons is depicted per replicon. Orthologs/Paralogs/Singletons detection was done with blastp and the Proteinortho software with a similarity cutoff of 50% and an *E*-value of 1e−10.

This analysis revealed that i) chromosome 1 of *V. alginolyticus* genomes has between ~190 to ~270 depending on the habitat while Kiel specific genomes share the same orthologous habitat

specific genes (422 Ogs including plasmid and prophage singletons) which are not visible by this analysis; However ii) the vast majority of the singletons are encoded on plasmids or episomal phages.

PATRIC analysis

Virulence factors prediction via PATRIC of the 12 *V. alginolyticus* genomes did not reveal a discernible niche specific pattern. However the highest number of potential virulence factors for NBRC 15630 strain were corroborated by Turner et al.'s investigation (Turner et al., 2018)(Figure 4, for details, see Table Figure 4).

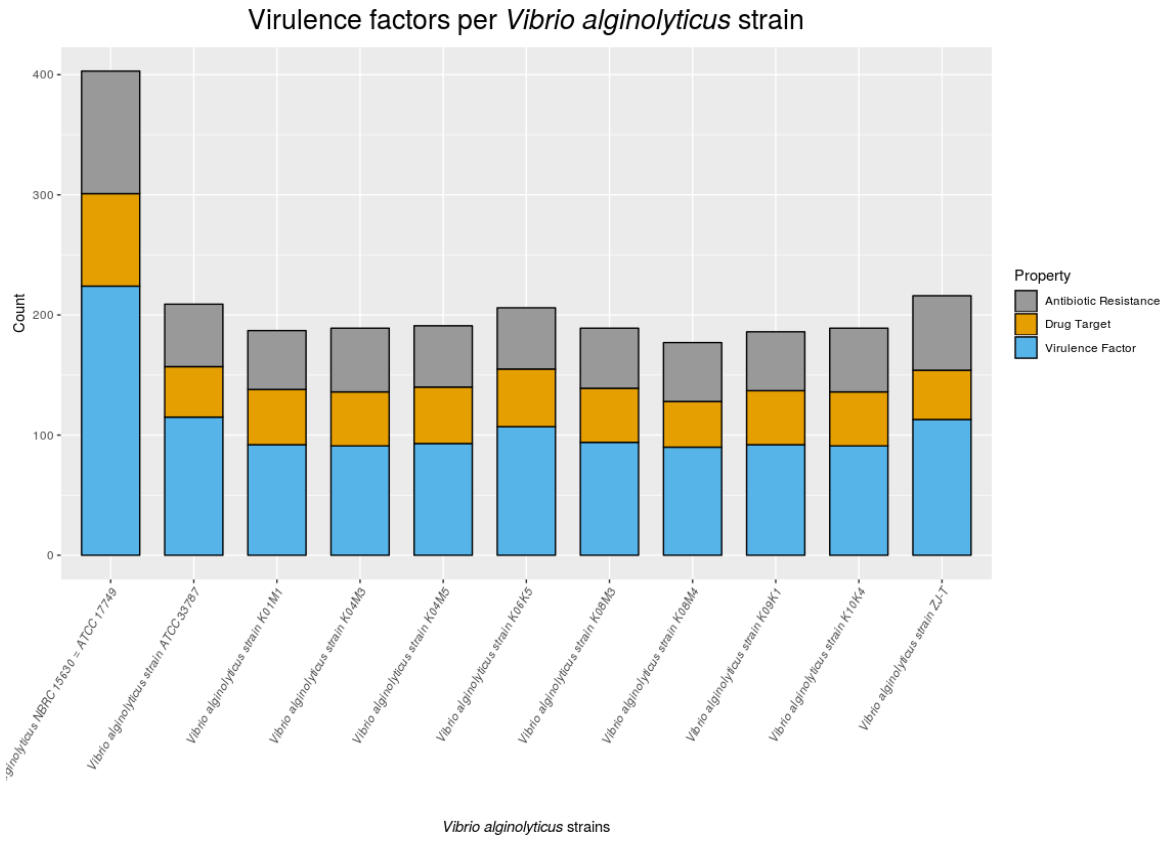


Figure 4: Virulence factors predictions displayed in a stacked bar plot resulting from the PATRIC analysis of 12 *Vibrio alginolyticus* strains.

The Kiel Fjord adaptations

The isolates of the Kieler Fjord share 422 genes that are exclusively present in these strains. The majority of these genes are located within mobile genetic elements including plasmids, prophages and genomics islands. Interestingly six of the nine Kiel isolates contain closely related plasmids ranging between 291 and 300 Kb in size and sharing over 90% nucleotide identity. A plasmid pan/core analysis revealed that the closely related plasmids encode 297 orthologous genes (Figure 5, for details, see Table Figure 5 and Table S1), which represent the main part of the 422 genes exclusively found in this habitat.

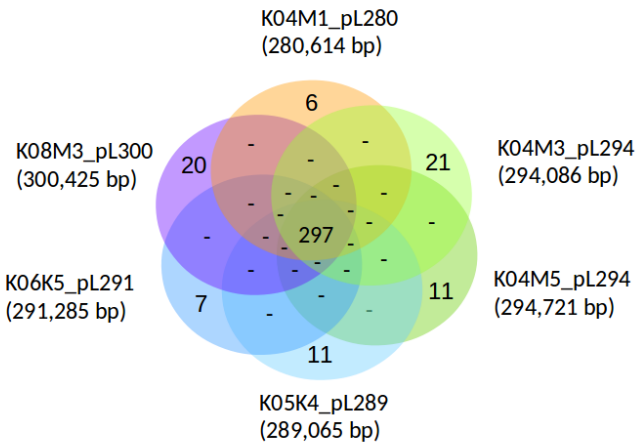


Figure 5: Analysis of orthologous genes (Ogs) on six plasmids found in six *V. alginolyticus* strains. The number of genome-specific Ogsis depicted in the respective ellipse. Ortholog detection was done with the Proteinortho software setting a similarity cutoff of 50% and an *E*-value of 1e−10.

V. alginolyticus ATCC 33787 contains as well three plasmids including the 287 Kb plasmid pMBL287 (Wang et al., 2016). However, a comparison of ATCC 33787 plasmids revealed no sequence similarity to any of the plasmids from the Kiel strains. In addition to the six related plasmids, three smaller plasmids (ranging between 9 -19 kb) without any similarity to the bigger plasmids or the plasmids from ATCC 33787 and three extra-chromosomal *Inoviridae*

phage replicons VK05K4_1, VK04K5_2 both from strain K05K4 and VK04M1_1 from strain K04M1 have been identified (See section Induced phages).

Induced phages

In many organisms integrated prophages can be induced by mitomycin C (Hertel et al., 2015), a stress-inducing compound. Liquid cultures of all nine *V. alginolyticus* strains were treated with mitomycin C in a phage induction experiment. A cell-free supernatant was investigated with transmission electron micrograph (TEM) and revealed filamentous structures in all strain derived supernatants. The TEM image from one shown isolate enabled us to classify the induced phages as *Inoviridae* (Figure 6).

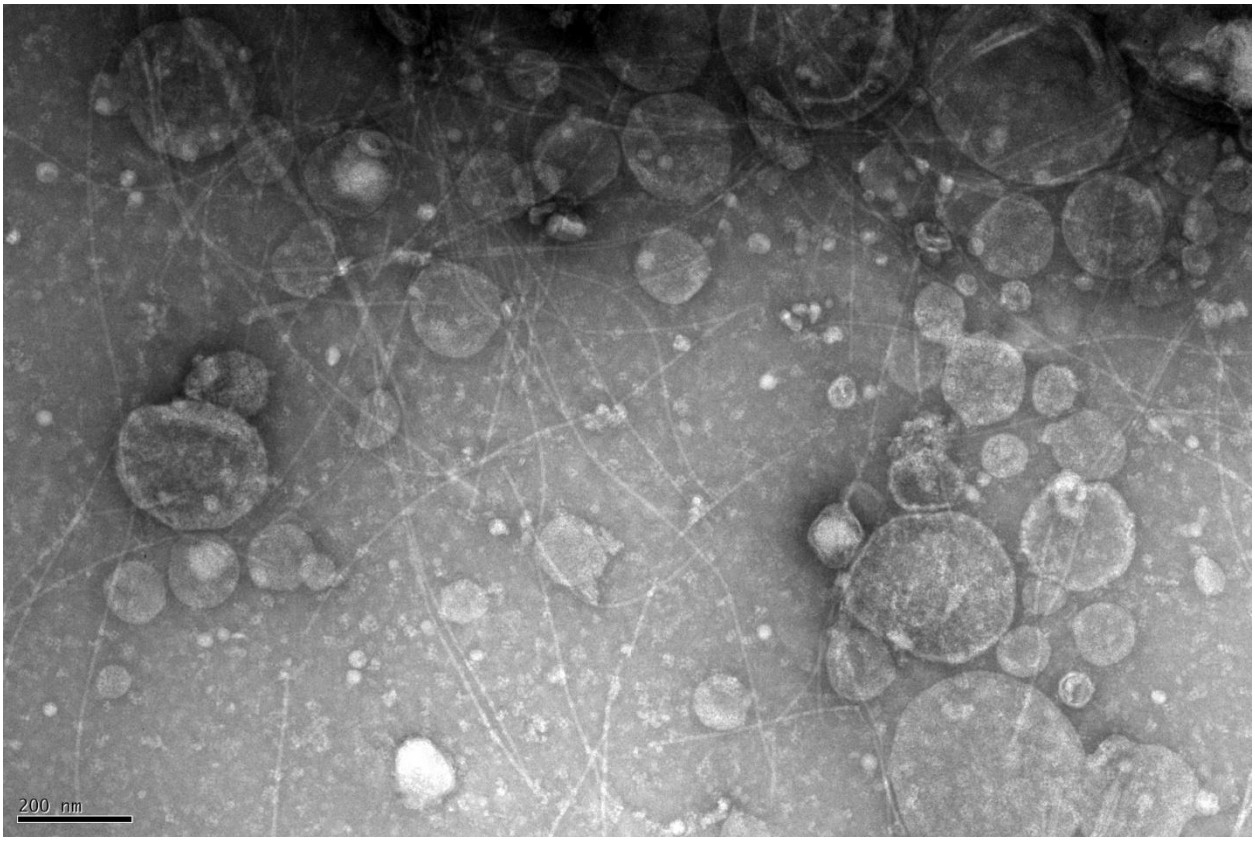


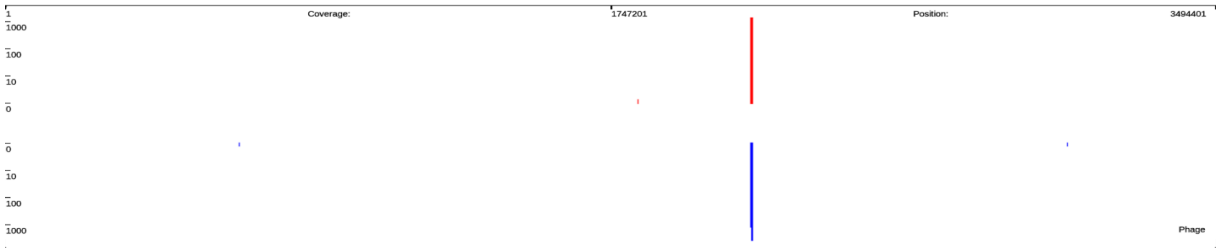
Figure 6: TEM image of the induced *Inoviridae* phages.

The TEM result indicates that we have filamentous particles in all samples we did not find any *Caudovirales*-like particles. Even with a series of mitomycin C induction experiments we never found any *Caudovirales*-like phage particles.

Identification of active phage loci

To locate the exact positions of the induced prophages, we performed a PHAGE-seq experiment (Hertel et al., 2015). In control experiments, the complete procedure has been applied without mitomycin C where the reference genomes were sequenced using Illumina technology. Both experiments revealed an increased coverage at *Inoviridae* loci (Figure 7). This indicates that induced and non-induced cultures produce comparable amounts of particles encoded by the same *Inoviridae* prophage.

a)



b)

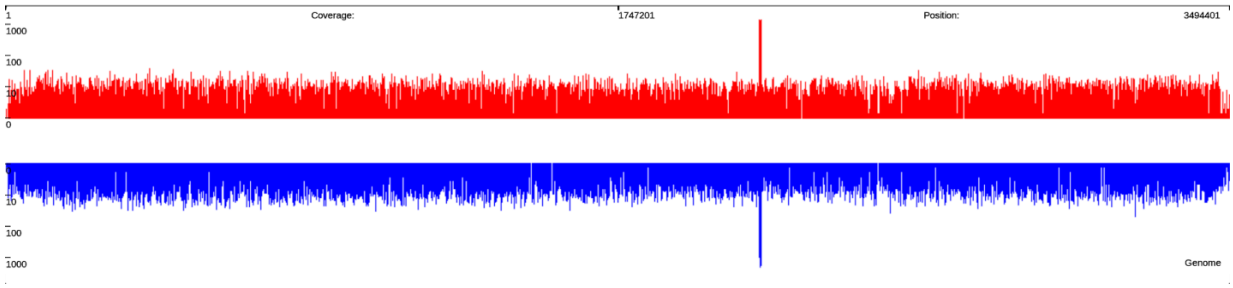


Figure 7: Phage-seq results of induced and non-induced *V. alginolyticus* K10K4 strain culture. a) Visualization of phage particle protected DNA to the corresponding reference genome. b) Visualization of complete bacterial genomic DNA.

The position of the mapped prophage DNA prophage, indicative for phage activity, enabled us to locate the exact positions of the integrated *Inoviridae* prophages (Table S4). Interestingly both samples generated mappings to the same loci with comparable coverage. This indicates that the *Inoviridae* phages derived from the nine *V. alginolyticus* are constitutively active with and without mitomycin C induction. As a further control, total DNA without DNase A treatment resulted in a coverage increased by the factor of 100 at the phage loci compared to the average chromosomal coverage (Figure 7b). The cultures produced a permanent amount of phage particle protected DNA. Within the nine sequenced *V. alginolyticus* isolates we found exclusively 19 active *Inoviridae*. None of the *Caudovirales* resulted in phage particle protected DNA. In case of the active *Inoviridae*, 16 were integrated on chromosome 2,and three exist as extra-chromosomal replicating replicons.

Prophages

PHASTER was used to investigate whether predicted phage loci correlates to the DNA within phage particles of the different cultures and to search for the complete set of predictable prophages. All replicons of the nine *V. alginolyticus* genomes were scanned with PHASTER where in total 45 prophages were predicted (Table2, for details, see Table S2), including at least one *Inoviridae* per genome. The presence of *Inoviridae* in each genome, each of them encoding a version of the ZOT-toxin confirms the importance of temperate members of this phage family for *V. alginolyticus* as a member of the genus *Vibrio* (Castillo et al., 2018; Kalatzis et al., 2017; Naser et al., 2017; Mai-Prochnow et al., 2015).

Table 2: Prophages predicted in *V. alginolyticus* strains

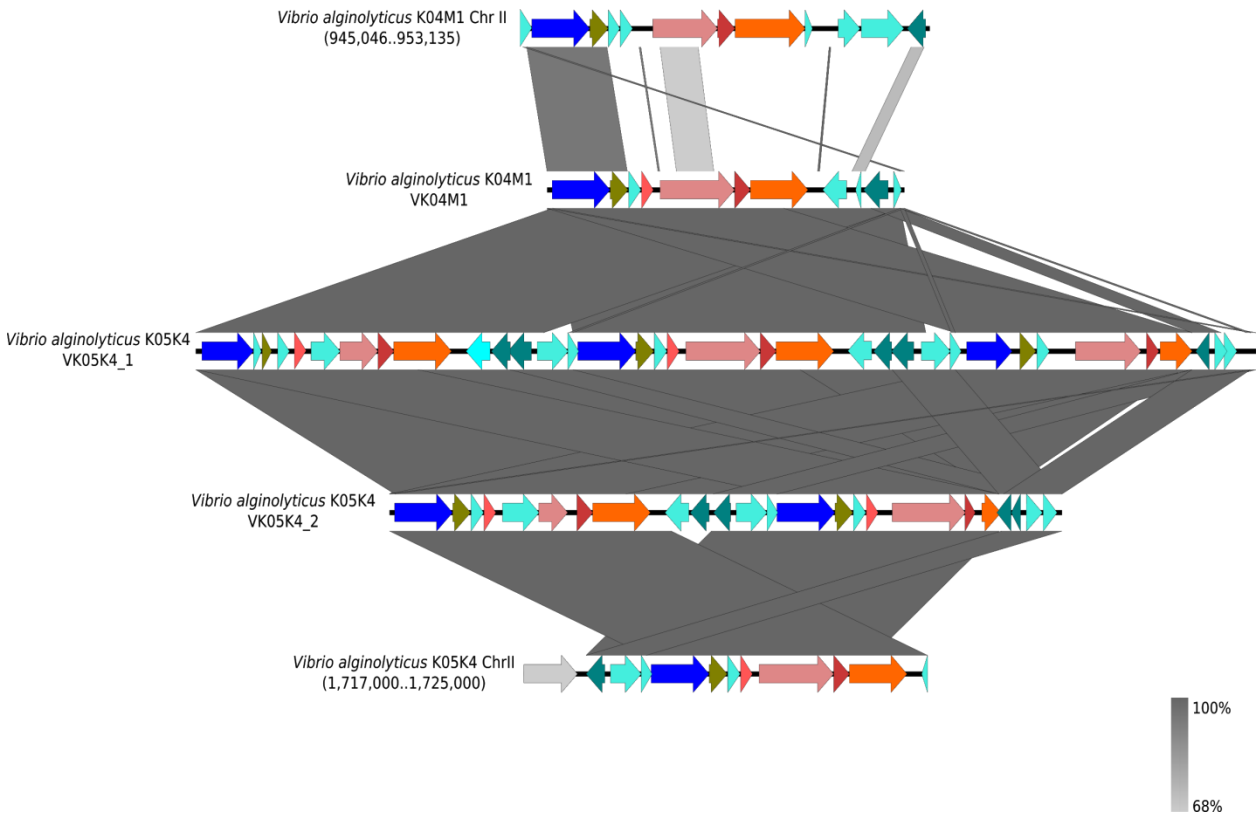
Genome	Phages
K01M1	2 *Caudovirales*, 1 *Inoviridae*
K04M1	3 *Caudovirales*, 2 *Inoviridae*
K04M3	3 *Caudovirales*, 2 *Inoviridae*
K05M5	3 *Caudovirales*, 2 *Inoviridae*
K05K4	3 *Caudovirales*, 6 *Inoviridae*
K06K5	3 *Caudovirales*, 1 *Inoviridae*
K08M3	3 *Caudovirales*, 1 *Inoviridae*
K09K1*	2 *Caudovirales*, 4 *Inoviridae*
K10K4	2 *Caudovirales*, 3 *Inoviridae*

*Due to the draft status of the genome the number of *Inoviridae* prophages is preliminary.

In addition to the expected *Inoviridae* prophages, 24 prophages containing key genes of the *Caudovirales*phage families have been predicted. Integrated prophages are in a lysogenic state, thus replicating via the hosting replicon. An in-depth analysis of the *Caudovirales*to one of the three subfamilies *Myo*-, *Podo*- or *Siphoviridae*was not possible due to the lack of the required morphological data.

Extra-chromosomal phages

Within the assembly of K04M1 and K05K4 strains, three closed circular contigs have been identified that consist of complete *Inoviridae* genomes. This indicates the presence of free phage replicons in two out of nine *V. alginolyticus* genomes. A sequence comparison of the three extra-chromosomal contigs to one another and the prophages integrated into chromosome 2 of both strains (Figure 7) confirmed that all of these phages are related. The i) annotation, the ii) TEM (Figure 8) vizualisation of the induced phages and the iii) genome comparison to published *Inoviridae* phages (for details see Figure S2 and Table S3) identified them as *Inoviridae* (Mai-Prochnow et al., 2015; International Committee on Taxonomy of Viruses & King, 2012).



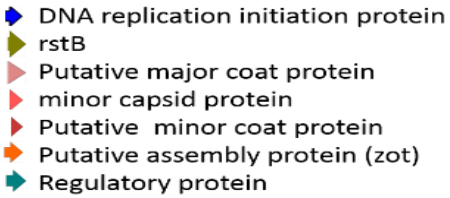


Figure 8: Genome comparison of extra-chromosomal phage genomes and prophages from *V. alginolyticus* strains K04M1 and K05K4. Annotated genes are color-coded. Visualization was done with the program Easyfig with an *E*-value cutoff of 1e−10.

The two extra-chromosomal phages VK05K4_1 and VK05K4_2 were compared to the integrated *Inoviridae* prophage located at 1,717,000-1,725,000 bp on chromosome 2 of the *V. alginolyticus*K05K4 strain. This comparison unveils that VK05K4_2 consists of two K05K4 prophages and that VK05K4_1 consists of three K05K4 prophages. An additional comparison reveals that VK04M1 is syntenic to K05K4 prophage; however K04M1 prophage located at 930,000-990,000 of chromosome 2 shares no sequence similarities with VK04M1 but shares gene functionalities. A Proteinortho analysis of the three extra-chromosomal phages (Table S5) revealed the shared orthologs between these phages. Considering the observation that *Inoviridae* of the genus *Vibrio* can multiply by the rolling circle replication (RCR) (Wawrzyniak et al., 2017; Mai-Prochnow et al., 2015; International Committee on Taxonomy of Viruses & King, 2012) suggests the hypothesis that the two extra-chromosomal circular contigs represent RCR intermediates of the phage. However, to confirm or falsify this hypothesis experiments have to be performed that are beyond the scope of this project. In contrast, the comparison of VK04M1 to the integrated K04M1 *Inoviridae* prophage confirms that the extra-chromosomal phage that distinct from the strains own prophage but very close to the phages of strain K05K4.

4. Conclusions

We performed a comparative genome analysis of nine isolated *V. alginolyticus* strains and could show that the strains of the habitat share 422 genes specific for their shared habitat. The majority of 297 genes are encoded by a set of six closely plasmids whereas, the remaining Kiel habitat specific genes are encoded by prophages. These results show the importance of these mobile functions in shaping the *V. alginolyticus* genome.

A bioinformatic scan for prophages predicts *Inoviridae* and *Caudovirales* prophages. Surprisingly induction experiments with and without mitomycin C exclusively induced *Inoviridae* prophages to produce particles. This confirms the dominance of ssDNA *Inoviridae* in the isolates. Our experiments show that the ability to produce *Inoviridae* particles is not dependent on the induction with mitomycin C. All predicted prophages encode genes for the Ace protein and the ZOT assembly proteins which are orthologous to pathogenicity factors of related fish pathogenic *Vibrio* species.

5. References

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., & Wishart, D. S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Research, 44(W1), W16–W21. http://doi.org/10.1093/nar/gkw387

Castillo, D., Kau, K., Hussain, F., Kalatzis, P., Rørbo, N., Polz, M. F., & Middelboe, M. (2018). Widespread distribution of prophage-encoded virulence factors in marine Vibrio communities, (June), 2–10. http://doi.org/10.1038/s41598-018-28326-9

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., … Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods, 10(6), 563–569. http://doi.org/10.1038/nmeth.2474

Defoirdt, T. (2014). Virulence mechanisms of bacterial aquaculture pathogens and antivirulence therapy for aquaculture. Reviews in Aquaculture, 6(2), 100–114. http://doi.org/10.1111/raq.12030

Deng, YiqinChen, C., Zhao, Z., Huang, X., Ding, X., & Yang, Y. (2016). Complete Genome Sequence of Vibrio alginolyticus ZJ-T. Genome Announcements, 4(5), 4–5. http://doi.org/10.1128/genomea.00912-16

Dietrich, S., Wiegand, S., & Liesegang, H. (2014). TraV: A genome context sensitive transcriptome browser. PLoS ONE, 9(4). http://doi.org/10.1371/journal.pone.0093677

Ghenem, L., Elhadi, N., Alzahrni, F., & Nishibuchi, M. (2017). Vibrio Parahaemolyticus: A Review on Distribution, Pathogenesis, Virulence Determinants and Epidemiology. Saudi Journal of Medicine and Medical Science, 167–171. http://doi.org/10.4103/sjmms.sjmms

González-Escalona, N., Blackstone, G. M., & DePaola, A. (2006). Characterization of a Vibrio alginolyticus strain, isolated from Alaskan oysters, carrying a hemolysin gene similar to the thermostable direct hemolysin-related hemolysin gene (trh) of Vibrio parahaemolyticus. Applied and Environmental Microbiology, 72(12), 7925–7929. http://doi.org/10.1128/AEM.01548-06

Gonzalez-Escalona, N., Jolley, K. A., Reed, E., & Martinez-Urtaza, J. (2017). Defining a Core Genome Multilocus Sequence Typing Scheme for the Global Epidemiology of Vibrio parahaemolyticus. Journal of Clinical Microbiology, 55(6), 1682–1697. http://doi.org/10.1128/jcm.00227-17

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. International Journal of Systematic and Evolutionary Microbiology, 57(1), 81–91. http://doi.org/10.1099/ijs.0.64483-0

Harvell, D., Altizer, S., Cattadory, I. M., Harrington, L., & Weil, E. (2009). Climate change and wildlife diseases: When does the host matter the most? Ecology, 90(4), 906–912. http://doi.org/10.1890/08-0730.1

Hertel, R., Rodríguez, D. P., Hollensteiner, J., Dietrich, S., Leimbach, A., Hoppert, M., … Volland, S. (2015). Genome-based identification of active prophage regions by next generation sequencing in Bacillus licheniformis DSM13. PLoS ONE, 10(3), 1–18. http://doi.org/10.1371/journal.pone.0120759

Hörmansdorfer, S., Wentges, H., Neugebaur-büchler, K., & Bauer, J. (2000). Isolation of Vibrio alginolyticus from seawater aquaria, 175, 169–175.

International Committee on Taxonomy of Viruses, & King, A. M. Q. (2012). Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses. Elsevier, 9, 375–383. http://doi.org/10.1016/B978-0-12-384684-6.00036-7

Kalatzis, P. G., Rørbo, N., Castillo, D., Mauritzen, J. J., Jørgensen, J., Kokkari, C., … Middelboe, M. (2017). Stumbling across the same phage: Comparative genomics of widespread temperate phages infecting the fish pathogen Vibrio anguillarum. Viruses, 9(5), 1–19. http://doi.org/10.3390/v9050122

Ke, H.-M., Ogura, Y., Tsai, I. J., Urbanczyk, H., Liu, D., & Hayashi, T. (2018). Tracing Genomic Divergence of Vibrio Bacteria in the Harveyi Clade . Journal of Bacteriology, 200(15), 1–10. http://doi.org/10.1128/jb.00001-18

Lan, S.-F., Liao, W.-C., Wong, H. -c., Huang, C.-H., Chang, C.-H., Chen, S.-Y., … Wu, Y.-G. (2009). Characterization of a New Plasmid-Like Prophage in a Pandemic Vibrio parahaemolyticus O3:K6 Strain. Applied and Environmental Microbiology, 75(9), 2659–2667. http://doi.org/10.1128/aem.02483-08

Land, M., Hauser, L., Jun, S. R., Nookaew, I., Leuze, M. R., Ahn, T. H., … Ussery, D. W. (2015). Insights from 20 years of bacterial genome sequencing. Functional and Integrative Genomics, 15(2), 141–161. http://doi.org/10.1007/s10142-015-0433-4

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–9. http://doi.org/10.1038/nmeth.1923

Le Roux, F., Kirschner, A., Baker-Austin, C., Wendling, C. C., Osorio, C. R., Cava, F., … Amaro, C. (2015). The emergence of Vibrio pathogens in Europe: ecology, evolution, and pathogenesis (Paris, 11–12th March

2015). Frontiers in Microbiology, 6(August), 1–8. http://doi.org/10.3389/fmicb.2015.00830

Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P. F., & Prohaska, S. J. (2011). Proteinortho: Detection of Co-orthologs in large-scale analysis. BMC Bioinformatics, 12(1), 124. http://doi.org/10.1186/1471-2105-12-124

Lee, K. K., Yu, S. R., Yang, T. I., Liu, P. C., & Chen, F. R. (1996). Isolation and characterization of Vibrio alginolyticus isolated from diseased kuruma prawn, Penaeus japonicus. Letters in Applied Microbiology, 22(2), 111–114. http://doi.org/10.1111/j.1472-765X.1996.tb01121.x

Liu, X.-F., Cao, Y., Zhang, H.-L., Chen, Y.-J., & Hu, C.-J. (2015). Complete Genome Sequence of Vibrio alginolyticus ATCC 17749 T. Genome Announcements, 3(1). http://doi.org/10.1128/genomeA.01500-14

Mai-Prochnow, A., Hui, J. G. K., Kjelleberg, S., Rakonjac, J., McDougald, D., & Rice, S. A. (2015). "Big things in small packages: The genetics of filamentous phage and effects on fitness of their host." FEMS Microbiology Reviews, 39(4), 465–487. http://doi.org/10.1093/femsre/fuu007

Maxwell, K. L. (2019). Phages Tune in to Host Cell Quorum Sensing. Cell, 176(1–2), 7–8. http://doi.org/10.1016/j.cell.2018.12.007

Naser, I. Bin, Hoque, M. M., Abdullah, A., Bari, S. M. N., Ghosh, A. N., & Faruque, S. M. (2017). Environmental bacteriophages active on biofilms and planktonic forms of toxigenic Vibrio cholerae: Potential relevance in cholera epidemiology. PLoS ONE, 12(7), 1–15. http://doi.org/10.1371/journal.pone.0180838

Okada, K., Iida, T., Kita-Tsukamoto, K., & Honda, T. (2005). Vibrios commonly possess two chromosomes. Journal of Bacteriology, 187(2), 752–757. http://doi.org/10.1128/JB.187.2.752-757.2005

Roth, O., Keller, I., Landis, S. H., Salzburger, W., & Reusch, T. B. H. (2012). Hosts are ahead in a marine host-parasite coevolutionary arms race: Innate immune system adaptation in pipefish syngnathus typhle against vibrio phylotypes. Evolution, 66(8), 2528–2539. http://doi.org/10.1111/j.1558-5646.2012.01614.x

Sarkar, B. L., Chakrabarti, A. K., Sarkar, S., & Dutta, S. (2016). Vibriophage and Cholera Disease, 19(2), 9–20.

Sullivan, M. J., Petty, N. K., & Beatson, S. A. (2011). Easyfig: A genome comparison visualizer. Bioinformatics, 27(7), 1009–1010. http://doi.org/10.1093/bioinformatics/btr039

Székely, A. J., & Breitbart, M. (2016). Single-stranded DNA phages: From early molecular biology tools to recent revolutions in environmental microbiology. FEMS Microbiology Letters, 363(6), 1–9. http://doi.org/10.1093/femsle/fnw027

Turner, J. W., Tallman, J. J., Macias, A., Pinnell, L. J., Elledge, N. C., Azadani, D. N., … Strom, M. S. (2018). Comparative genomic analysis of Vibrio diabolicus and six taxonomic synonyms: A first look at the distribution and diversity of the expanded species. Frontiers in Microbiology, 9(AUG), 1–14. http://doi.org/10.3389/fmicb.2018.01893

Wagner, P. L., & Waldor, M. K. (2002). MINIREVIEW Bacteriophage Control of Bacterial Virulence. Society, 70(8), 3985–3993. http://doi.org/10.1128/IAI.70.8.3985

Waldor, M. K., & Mekalanos, J. J. (1996). Lysogenic Conversion by a Filamentous Phage Encoding Cholera Toxin. Science, 272(5270), 1910–1914. http://doi.org/10.1126/science.272.5270.1910

Wang, P., Wen, Z., Li, B., Zeng, Z., & Wang, X. (2016). Complete genome sequence of Vibrio alginolyticus

ATCC 33787Tisolated from seawater with three native megaplasmids. Marine Genomics, 28, 45–47. http://doi.org/10.1016/j.margen.2016.05.003

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., … Sobral, B. W. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Research, 42(D1), 581–591. http://doi.org/10.1093/nar/gkt1099

Wawrzyniak, P., Plucienniczak, G., & Bartosik, D. (2017). The different faces of rolling-circle replication and its multifunctional initiator proteins. Frontiers in Microbiology, 8(NOV), 1–13. http://doi.org/10.3389/fmicb.2017.02353

Wendling, C. C., Piecyk, A., Refardt, D., Chibani, C., Hertel, R., Liesegang, H., … Roth, O. (2017). Tripartite species interaction: eukaryotic hosts suffer more from phage susceptible than from phage resistant bacteria. BMC Evolutionary Biology, 17(1). http://doi.org/10.1186/s12862-017-0930-2

Wickham, H. (2011). Ggplot2. Wiley Interdisciplinary Reviews: Computational Statistics, 3(2), 180–185. http://doi.org/10.1002/wics.147

Willms, I., Hoppert, M., & Hertel, R. (2017). Characterization of Bacillus subtilis viruses vB_BsuM-Goe2 and vB_BsuM-Goe3. Viruses, 9(6), 146.

Yoon, S. H., Ha, S. min, Lim, J., Kwon, S., & Chun, J. (2017). A large-scale evaluation of algorithms to calculate average nucleotide identity. Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology, 110(10), 1281–1286. http://doi.org/10.1007/s10482-017-0844-4

Zhang, J., Cao, Z., Xu, Y., Li, X., Li, H., Wu, F., … Jin, L. (2014). Complete genomic sequence of the Vibrio alginolyticus lytic bacteriophage PVA1. Archives of Virology, 159(12), 3447–3451. http://doi.org/10.1007/s00705-014-2207-z

Supplementary Data

Figure S1: Venn diagram of extra-chromosomal phages shared orthologs and singletons.

Table S1: Singletons encoded on plasmids shared between V. alginolyticus strains.

Table S2: PHASTER prophage regions identified in the 9 sequenced Vibrio alginolyticus strains.

Table S3: Vibrio Inoviridae reference phages used for comparisons.

Table S4: Inoviridae sequence mapping on V. alginolyticus reference genome.

Table S5: Annotation of shared Inoviridae encoded proteins from the extra-chromosomal phages.

Table S6: Proteinortho output of CDS derived from 13 Vibrio strains.

Table S7: Vibrio genomes accessed from NCBI 12.06.2018used for PyANI analysis.

Additional Files

Table Figure 2: Input information of unique and shared Orthologs between the 10 Vibrio Kiel strains and Reference Vibrio alginolyticus strains.

Table Figure 3: Input information on the number of singletons per Vibrio replicon.

Table Figure 4: Input information of the number of genes associated to antibiotic resistance, drug target and virulence factor (PATRIC output).

Table Figure 5: Proteinortho output of CDS derived from Vibrio Megaplasmids.

Supplementary information

Additionally, supplementary figures and tables are provided along with the electronic version of this thesis (on DVD), under the following paths:

Additional Figures:

Figure S1: SupplementaryMaterial/ChapterII/ChapterII.2/Figure S1.png

Additional Tables:

Table S1: SupplementaryMaterial/ChapterII/ChapterII.2/TableS1.xlsx

Table S2: SupplementaryMaterial/ChapterII/ChapterII.2/TableS2.xlsx

Table S3: SupplementaryMaterial/ChapterII/ChapterII.2/TableS3.xlsx

Table S4: SupplementaryMaterial/ChapterII/ChapterII.2/TableS4.xlsx

Table S5: SupplementaryMaterial/ChapterII/ChapterII.2/TableS5.xlsx

Table S6: SupplementaryMaterial/ChapterII/ChapterII.2/TableS6.xlsx

Table S7: SupplementaryMaterial/ChapterII/ChapterII.2/TableS7.xlsx

Additional Data:

Table Figure 2: SupplementaryMaterial/ChapterII/ChapterII.2/Table_Fig2.xlsx

Table Figure 3: SupplementaryMaterial/ChapterII/ChapterII.2/Table_Fig3.xlsx

Table Figure 4: SupplementaryMaterial/ChapterII/ChapterII.2/Table_Fig3.xlsx

Table Figure 5: SupplementaryMaterial/ChapterII/ChapterII.2/Table_Fig3.xlsx

II.3 Classifying the unclassified: A phage classification method

Viruses

Classifying the unclassified: A phage classification method

Cynthia Maria Chibani, Anton Farr, Sandra Klama, Sascha Dietrich, Heiko Liesegang

Chibani et al. Viruses (2019) 11:195 **DOI:**https://doi.org/10.3390/v11020195

Authors' contributions

CC performed research, designed algorithm, wrote program, performed data analysis, wrotemanuscript,

AF generated Markov Models, wrote program, performed data analysis,

SK compared and visualized phage genomes,

SD designed algorithm,

HL designed research, designed algorithm.

viruses

MDPI

Article

Classifying the Unclassified: A Phage Classification Method

Cynthia Maria Chibani, Anton Farr, Sandra Klama, Sascha Dietrich and Heiko Liesegang *

Institute for Microbiology and Genetics, Georg-August University Goettingen, Grisebachstr. 8, 37077 Goettingen, Germany; cchiban@gwdg.de (C.M.C.); anton.farr@stud.uni-goettingen.de (A.F.); sandra.klama@gwdg.de (S.K.); sascha.dietrich@uni-wuerzburg.de (S.D.)

* Correspondence: hlieseg@gwdg.de

Received: 21 December 2018; Accepted: 20 February 2019; Published: 24 February 2019

Abstract: This work reports the method ClassiPhage to classify phage genomes using sequence derived taxonomic features. ClassiPhage uses a set of phage specific Hidden Markov Models (HMMs) generated from clusters of related proteins. The method was validated on all publicly available genomes of phages that are known to infect *Vibrionaceae*. The phages belong to the well-described phage families of *Myoviridae*, *Podoviridae*, *Siphoviridae*, and *Inoviridae*. The achieved classification is consistent with the assignments of the International Committee on Taxonomy of Viruses (ICTV), all tested phages were assigned to the corresponding group of the ICTV-database. In addition, 44 out of 58 genomes of *Vibrio* phages not yet classified could be assigned to a phage family. The remaining 14 genomes may represent phages of new families or subfamilies. Comparative genomics indicates that the ability of the approach to identify and classify phages is correlated to the conserved genomic organization. ClassiPhage classifies phages exclusively based on genome sequence data and can be applied on distinct phage genomes as well as on prophage regions within host genomes. Possible applications include (a) classifying phages from assembled metagenomes; and (b) the identification and classification of integrated prophages and the splitting of phage families into subfamilies.

Keywords: Hidden Markov Models; *Vibrionaceae*; vibriophages; *Inoviridae*; *Myoviridae*; *Podoviridae*; *Siphoviridae*; phages; classification; protein coding sequences

1. Introduction

Phages, defined as viruses that infect bacteria, are the most abundant biological entities known so far [1,2]. The taxonomic classification of viruses and naming of virus taxa is maintained by the International Committee on Taxonomy of Viruses (ICTV) [3] and the Bacterial and Archaeal Subcommittee (BAVS) within the ICTV that focuses on phages. The system is based on the evaluation of a variety of phage properties including the molecular composition of the virus genome (ss/ds, DNA, or RNA), the structure of the virus capsid and whether or not it is enveloped, the host range, pathogenicity, and sequence similarity [4,5]. Based upon these different properties the ICTV established a highly valuable and widely accepted Virus taxonomy. Considering the complexity of features that contribute to the taxonomy of a phage a comprehensive guideline has been published by Adriaenssens and Brister [6]. However, due to the availability of Next Generation Sequencing (NGS)-technologies an increasing amount of genomic and metagenomic sequence data is available that include complete as well as fragments of so far unknown phage genomes [7,8]. Unfortunately, a systematic classification of these genomes into the ICTV scheme is impossible due to the lack of corresponding biological and experimental data [4,9,10]. So for that matter, a taxonomic characterization based on the phages genome sequence information has become indispensable [5].

Many attempts at creating viral phylogenetic trees have failed due to the lack of a universal marker gene and the high mosaic structure of phages [11]. Sequence-based phylogenetic analysis procedures like 16S and multi locus sequence typing (MLST) [12–14] are based on the existence of an orthologous marker molecule shared among monophyletic entities. The finding that phages are a polyphyletic group of biological entities results in the finding that orthologous markers are available only within the monophyletic subgroups of phages [15]. Consequently, sequence alignment and similarities based approaches using single selected marker molecules have been designed for phage classifications restricted to closely related phage taxa [16]. Clustering techniques for viral classification have been applied by several authors and confirmed that comparative sequence analysis is effective [11,17–19]. Deschavanne et al. [20] demonstrated that genomic signatures based on oligomer composition are effective to determine the phylogenetic distance of closely-related phages and their hosts, as well as within the phages preying on related hosts. The investigation revealed that in the case of temperate phages, the amelioration process [21] interferes with the calculation of phylogenetic distances between phages. Rohwer and Edwards used the presence and the similarity of shared proteins to generate a phage proteomic tree using 105 complete sequenced genomes [22]. This approach is robust towards dynamic changes in the nucleotide composition. However, proteomic trees are limited in cases where the BLASTP based similarity determination is challenged by distantly related protein sequences. Bolduc et al. [23] introduced vConTACT, a tool that uses protein clusters and bipartite network-based distances to assign a given dsDNA phage genome to a taxon.Aiewsakun et al. [24] demonstrated that the Genome Relationship Applied to Virus Taxonomy (GRAViTy) software platform, which is designed for eukaryotic virus genomes, performs well on monophyletic subfamilies of viruses that infect bacteria and archaea. GRAViTy uses composite generalized Jaccard (CGJ) distances based on shared genomic features to determine the genetic relatedness of a given set of virus genomes.

The development of bioinformatics methods to recognize and characterize genomics elements is strongly supported if a well-described sample dataset is available. In the case of our project, we selected vibriophages, i.e., phages that infect *Vibrionaceae*, as a training dataset. Vibriophages are known as an important driving force of the evolution of *Vibrionaceae*, contributing to the emergence of virulence and the ecological success of this genus [25]. In the case of *Vibrio cholera*, the causative of the pandemic disease cholera (WHO newsletter 2018), the virulence of the bacterium is encoded by viral genes of the phage. Due to its medical importance, it is a well investigated example of how phages contribute to the evolution and the virulence of bacterial hosts [26–28]. *Vibrionaceae* include in addition a number of important fish pathogens, where integrated prophages have been shown to contribute to the virulence of the strains, and thus leading to great economic losses [29]. *Inoviridae*, which comprises the CTX-phage of *V. cholera* [30], as well as the filamentous M13 phage [31], are among the best-investigated phages that have been studied for more than VI decades [28]. Due to the medical and economic importance and the in detail molecular biological knowledge on *Inoviridae*, a substantial amount of sequencing data on *Vibrionaceae* and Vibriophages is available. Castillo et al. [21] have recently estimated that there exist 5674 prophage-like elements within 1874 published *Vibrio* genome sequences, and that 45% of the strains harbor prophages of the family *Inoviridae*, that contribute by lysogenic conversion, with the Zonaoccludens toxin (Zot), to the virulence of *Vibrionaceae*. Multiple studies have shown the presence of *Caudovirales* in addition to *Inoviridae* phages in *Vibrio* species [32–34]. This makes this group an excellent test case for a sequence based characterization method and a potential identification of phages.

Hidden Markov Model (HMM) based search and clustering methods proved to be efficient for the characterization of protein families, as well for the taxonomic characterization of corresponding genes [15,35]. Here, we present a case study that investigates profiles of combined HMMs derived from related dsDNA and ssDNAphage genomes, and their efficiency characterize and potentially identify members of four well-described families of vibriophages. We demonstrate that a method based exclusively on genome sequences achieves a classification of phages that is consistent with the ICTV standards. Furthermore, a genomic analysis of the profile HMM characterized genomes, reveals details and relation of phages corresponding to their phylogenetic distance and their host range.

2. Materials and Methods

2.1. Data Sources

2.1.1. Phages

Various phage datasets have been used in this study. Firstly, publicly available Vibriophages sequence datasets were downloaded from NCBI nucleotide database by keyword search; date of accession 13 February 2018. In total numbers, 159 phage genomes out of which 58 were unclassified, 19 *Inoviridae*, 37 *Myoviridae*, 42 *Podoviridae*, and 15 *Siphoviridae* genomes infecting *Vibrio* genomes were downloaded (Table S1). This dataset was split into a classified for the generation of HMM models and an unclassified dataset to which the HMMs were applied for classification purposes.

Secondly, a set of 19 experimentally proven and sequenced *Inoviridae* phages derived from a genome sequencing project on 9 *V. alginolyticus* strains and 1 *V.typhli* strain isolated from Pipefish [29], was used for validation of the *Inoviridae* generated HMMs.

Lastly, in order to test the limitations of the method, sequence datasets of the four phage families were downloaded from the Millard lab database (http://millardlab.org/bioinformatics/bacteriophage-genomes/); date of accession March 2018. In total numbers, 119 *Ino-*, 1766 *Myo-*,1066 *Podo- and* 3466 *Siphoviridae* were downloaded (Table S2).

2.1.2. Host Genomes

In order to test the generated HMMs for phage identification, 154 closed *Vibrio* genomes publicly available (Table S3) were downloaded from NCBI by keyword search; date of accession 18.06.2018. In total numbers, 154 *Vibrio* genomes out of which 39 were *V. cholerae*, 22 *V. parahaemolyticus*, 15 *V. vulnificus*, 13 *V. alginolyticus*, 13 *V. anguillarum*, 9 *V. campbellii*, 5 *V. natriegens*, 4 *V. harveyi*, 4 *V. coralliilyticus*, and other *Vibrio* species were downloaded.

2.2. Data Preparation

For each genbank file, a multi-FASTA file containing all annotated coding sequences was created. The collected protein sequences were concatenated and clustered with the Markov clustering algorithm (MCL) [36]. CD-hit [37] (V4.5.4) was used to remove redundant proteins. In addition, information on classification, host, phage size, isolation source was extracted from each genbank file.

2.3. Profile HMM Construction

Produced multi-sequence alignment files were used to build profile HMMs [38], using the "hmmbuild" command available as part of the HMMER (v3.1b1) package. Subsequently, sensitive profile HMMs were created out of a minimum of five clustered proteins. Removed proteins were stored for later refinement steps. The command "hmmpress" was used to create binary compressed data files (.h3m, .h3i, .h3f, and .h3p) from a profile HMM. These binary files were used to look for orthologous protein hits in the scanned dataset. The scanned input dataset was used to map hit to the phage family proteins they were derived from. The function "hmmemit" was used to create a consensus sequence from a generated profile HMM. This consensus sequence is closest in similarity to the majority of sequences used to create the respective HMM.

2.4. Profile HMM Refinement

Using "BLASTP" to align each protein of a cluster against the consensus sequence, and by specifying the output table to feature the coverage of each sequence compared to the consensus, the coverage was compared with the user-specified threshold(standard <50%). Proteins not reaching the threshold were removed. Created profile HMMs were used to scan the original master-FASTA. Proteins were refined according to hits of (a) proteins removed due to redundancies, (b) proteins used to create the HMMs themselves, and (c) not yet assigned proteins. Proteins which are hit and have not

yet been assigned were added to the profile HMM. Proteins that were used to create the HMM and were not hit were removed from the profile HMM. Proteins that are hit but were removed previously due to redundancies were not added. Whenever multiple HMMs hit the same sets of proteins as well as their inputs, they were merged. Otherwise, HMMs were not merged. Refined HMMs were used to rescan the input master-FASTA and if needed refinement steps of merging were repeated until no changes occured.

2.5. CDS Prediction and Additional HMM Refinement

Nucleotide sequences between predicted coding sequences (CDS) were extracted from each genbank file and were translated into an amino acid sequence. Generated refined HMMs were used to scan the translated regions. A 50% alignment coverage, a negative bit-score value and an E-value over 1.5×10^{-8} were used as cut-offs to filter the generated hmmscan output. Hits passing the filtered cut-offs were integrated in the multiple sequence alignment (MSA) input per HMM and HMMs were rebuilt with the updated MSA. The regenerated HMMs were used to rescan the input phage master-FASTA files in order to compare HMMs performance when generated based on the original genbank files and the HMMs generated based on improved genomes. The generated HMMs can be downloaded at http://appmibio.uni-goettingen.de/index.php?sec=sw.

2.6. Software Tools

PHASTER was used to scan all 154 *Vibrio* gbk files (Table S3) for the identification of integrated phages. Visualization was performed using R version 3.2.3 in Rstudio version 1.1.383 and using the R package "ggplot2" version 3.0.0 unless stated otherwise.

3. Results and Discussion

3.1. Phage Protein Families and Profile HMMs

To generate the initial set of HMMs, the protein sequences of all 110 available genomes known to infect *Vibrionaceae* were extracted. The data consists of the proteins from 19 *Ino-*, 35 *Myo-*, 42 *Podo-*, and 14 *Siphoviridae* phages. To ensure the internal model diversity, redundant sequences were removed and the remaining protein sequences were clustered with the *Markov cluster algorithm* (MCL) [37]. Models generated from clusters of five or more diverse sequences per protein family were evaluated for their taxonomic specificity (Table 1). In cases where models generated significant better hits against proteins of the phage taxon from which they have been encoded, the HMMs were considered as taxonomic indicators of the phage family.

Table 1. Phage family specific HMMs *.

	No of Genomes (Size in Kbp)	No of Proteins	Proteins after MCL	HMMS with >5 Proteins	Positive Evaluated HMMs
Siphoviridae	14 (37.3–128.6)	1497	414	94	54
Podoviridae	42 (38.4–112.1)	2641	490	233	96
Myoviridae	35 (33.1–250)	5915	921	634	242
Inoviridae	19 (6.3–21)	241	39	12	9
Total	110	10,294	1864	973	401

* Details on the complete calculation of the models are in supplementary Table S1.

The procedure resulted in 401 HMMs representing taxonomic indicative profile HMMs. In total 9 HMMs specific for *Ino-*, 242 for *Myo-*, 96 for *Podo-*, and 54 for *Siphoviridae* were identified as taxonomic indicators. The proteins used to generate refined HMMs per phage family are summarized in Supplemental Tables S5–S8. Note that, due to the lack of a sufficient number of diverse protein sequences, for some protein families no profile HMMs has been generated.

3.2. Taxon Specificity of the Protein Family Models

To evaluate the discriminative power of a protein family based taxonomy, the profile HMMs were applied on three different data sets. (I) HMM scan to classify genomes of bacteriophages, known to prey on *Vibrionaceae*, into taxonomic groups consistent to the rules defined by the ICTV. (II) A scan of all proteins encoded by host genomes to investigate the, performance of the method to classify as well as potentially identify integrated prophages. In this test, host genomes with known biologically active vibriophages were used as proof of principle. (III) Scan of proteins of all known phage genomes from the taxa *Ino-*, *Myo-*, *Podo-*, and *Siphoviridae*.

3.3. Consistency of Taxon-Specific HMMs

The refined profile HMMs, derived out of the four phages families, were used in scans against all 4630 proteins encoded by the 110 phage genomes (Figure 1).

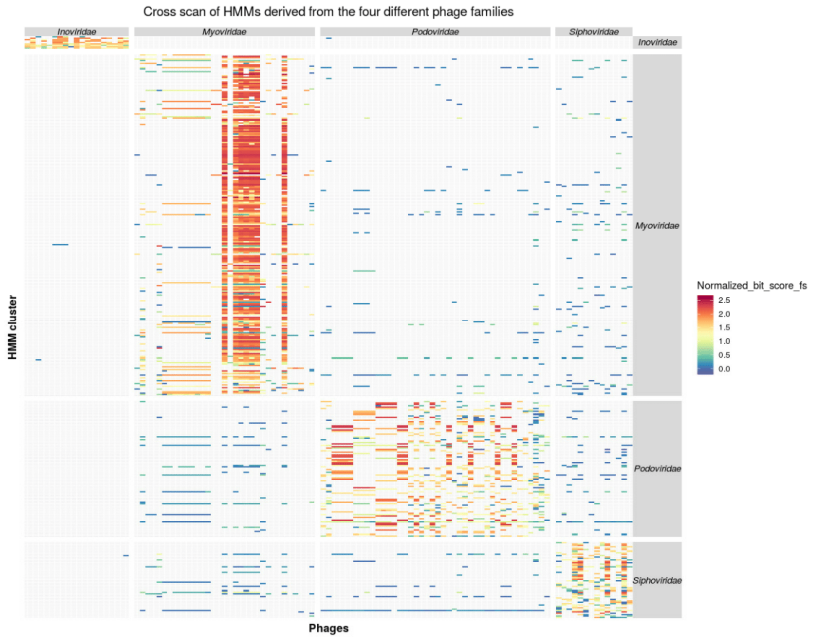


Figure 1. Markov Models (HMM) scan of phage family derived models own input "CDS" and coding sequences of other families. The scan of the protein sequences derived from *Ino-*, *Myo-*, *Podo-*, and *Siphoviridae*, was conducted by the profile HMMs. The names of all phages grouped into phage-families are marked at the bottom of heatmap. The bit-score of the HMM matches was normalized by the size (in bp) of the HMM's consensus sequence (data see Table S9). The results are color-coded from blue (low-score) to red (high-score).

An application of the HMM profiles on the input phage proteome sequences revealed that the vast majority of the proteins (83.45%) match exclusively the taxon specific HMMs from the corresponding phage family. However, there was a number of 16.37% cross matches between the different families within the *Caudovirales* models, which indicates that the investigated phage genomes might represent a monophyletic group within the *Caudovirales* [39]. In contrast, 0.17% cross-matches occurred between *Caudovirales* and *Inoviridae* and thus support the hypothesis that there is gene exchange between these not monophyletic taxa [39].

3.3.1. Inoviridae

In the case of the *Inoviridae* HMMs scanning *Caudovirales* proteomes, all HMMs match exclusively proteins encoded in *Inoviridae* genomes, except of one case which has an e-value of 2.9×10^{-6} to a protein annotated as "putative streptomycin biosynthesis operon regulatory protein (YP_009021749.1)". While *Caudovirales* HMMs scanning *Inoviridae* proteomes, all HMMs match exclusively proteins encoded in *Caudovirales* except in seven cases where the e-value ranged between 1.5×10^{-8} and 7.8×10^{-5} to proteins annotated as "hypothetical protein" and "RstR" (Table S10). The low number of cross matches between *Inoviridae* and *Caudoviridae* is due to the phenotypical unique features of filamentous phages in contrast to tailed phages [40,41]. However, cross match hits may as well reflect genes that have been exchanged between *Inoviridae* and *Caudovirales* by a horizontal gene transfer (HGT) event [11]. Under this condition, the lower quality of the match score would reflect the time that the proteins evolved after the HGT-event within their separate viral host genomes.

3.3.2. Caudovirales

In case of *Caudovirales*, scans of HMMs against their encoded proteins lead to a considerable number of cross matches (16.37%, 758 out of 4630). The proteins are related to basic phage functionality that are expected to be encoded by genomes of tailed phage like DNA polymerase, DNA replication initiation protein, ribonucleases, helicases, endonucleases, ligases, terminase, and phage tail proteins, as well as hypothetical proteins (Table S10). However, the taxon derived models scored better against taxon encoded proteins. The type of the proteins and the correlation of HMM scores indicate that the matches are due to the shared genes with a common phylogenetic history of the tailed phages [11] and not to false positive recognition event of the HMMs.

To further explore vibriophages of the three *Caudovirales* groups, genome alignments were performed revealing that the virus genomes have a host specific diversity (Figure 2).

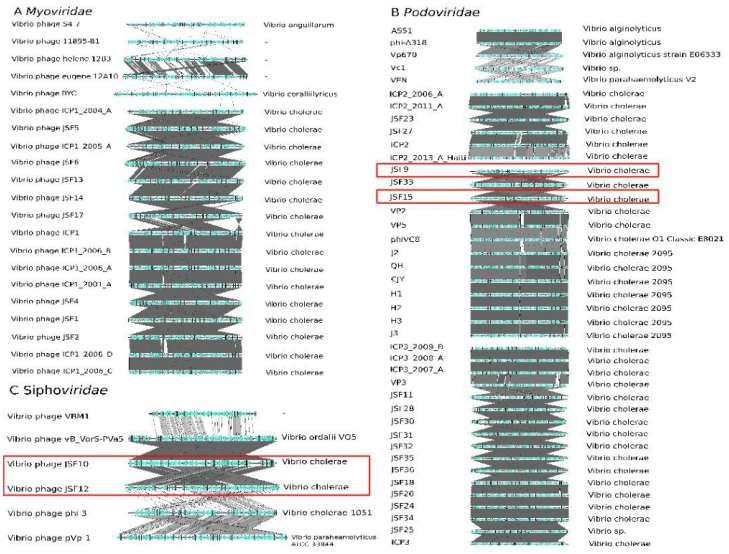


Figure 2. Alignment of *Caudovirales* genomes. (**A**) *Myoviridae*, (**B**) *Podoviridae*, and (**C**) *Siphoviridae*. Genomes of phages that have not yet been assigned by ICTV are marked in pink. Four phages JSF9, JSF10, JSF12, and JSF15 are boxed in red. JSF12 has been assigned to *Podoviridae* based on transmission electron micrographs (TEM) the complete genome alignment indicates a close relation to the *Siphoviridae* phage JSF10. The data has been visualized with Easyfig.

Viruses **2019**, *11*, 195

Genome alignments of the three *Caudovirales* in most of the cases revealed extended sequence similarities according to the BLAST algorithm within members of the taxonomic groups. Note that the BLAST algorithm is considerably less sensitive to identify distant similar sequences in comparison to the profile HMM [15]. This reduced sensitivity is the reason why BLAST based algorithms miss the taxonomic proximity of the *Myoviridae* phages 54-7, I1895-B1, and helene 12B3 as well as between the *Myoviridae* Eugene 12A10, RYC, and ICP1_2004_A (Figure 2). However, in most of the cases of *Myo*- and *Siphoviridae*, all members exhibit different degrees of similarities over the complete genome sequences and thus support the statement that the families are monophyletic [24,42]. However, within the *Podoviridae*, the comparison revealed four subgroups that did not show pronounced sequence.

3.4. Classification of Unclassified Phages

To examine the generated profile HMMs with regard to their application as a means of genome sequence based classification of bacteriophages, HMMs derived out of the four different bacteriophage families were used on to scan the proteomes of 58 published but taxonomically unclassified *Vibrio*-phages (Figure 3, Table S1). The details of the HMM scan are summarized in Table S11.

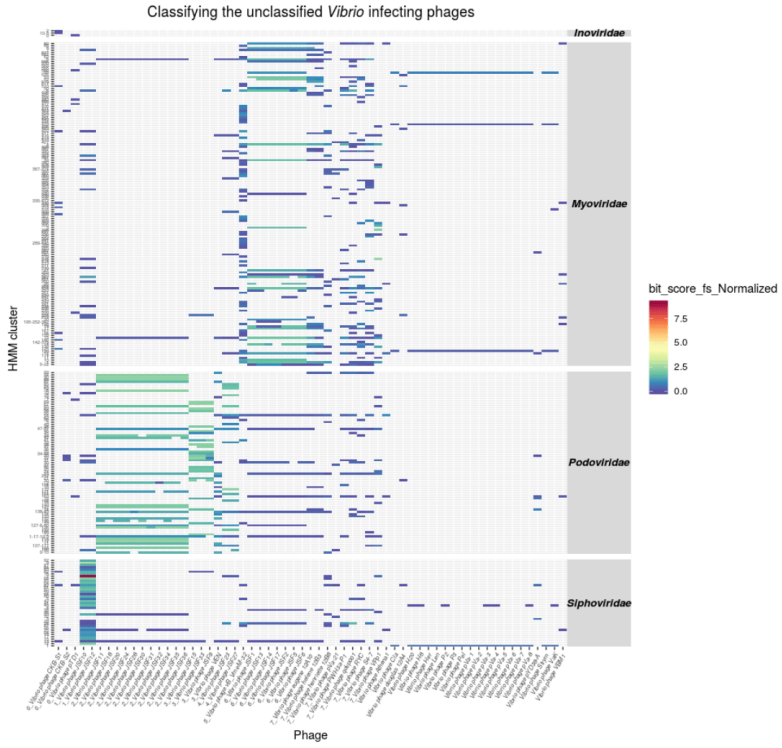


Figure 3. Taxonomic classification of vibriophages. This heatmap shows a profile HMM scan on the proteins of 58 unclassified bacteriophages genomes. Forty-one unclassified genomes generated sufficient with enough hits to be assigned to a taxonomic group. The HMMs have been integrated in the heatmap (*x*-axis). The HMMs are grouped (on the *y*-axis) into the respective phage families. The indicator for the quality of a hit is color coded to the normalized bit-score assigned for the respective match by hmmscan.

Taxonomic Assignments of Tailed Phages

For some of the investigated phage genomes, a taxonomic assignment based on experimental data is available. Zahid et al. [43] classified the vibriophages JSF9, JSF12 and JSF15 as *Podo*- and vibriophage JSF10 as *Siphoviridae*. Our data supports the assignment of phages JSF9, JSF15, and JSF 10. However, JSF12 according to the profile HMM hits should be classified as *Siphoviridae* (Figure 2A,B).

Whole genome alignments revealed that all phages that have been assigned by the ClassiPhage method to an ICTV taxon comprise large genome regions that can be aligned to corresponding classified reference genomes. However, in some cases, the overall coverage of the alignable parts of the phage genomes to reference genomes is sparse. In the case of phage JSF12, experimental data indicates an assignment to *Podoviridae* while the alignment reveals a higher similarity to reference genomes from the *Siphoviridae*. The latter result is in accordance with the results of the profile HMM scan. Both sequences have been aligned and closely inspected using ACT where no missing ORF was observed.

The application of the method on the unclassified vibriophages dataset explored the capabilities of ClassiPhage, where transmission electron micrographs (TEM) images confirm the generated classification. The HMMs of the different families demonstrated a high specificity, meaning that when a phage genome is specifically targeted by HMMs of one family, the HMMs of other families show only insignificant numbers of HMM/protein matches. This specificity further supports the idea that it is possible to use the generated HMMs as a means of classification as discussed by [15].

The generated *Vibrio* derived profiles scanning the proteomes of the phages of the four families gave us the unique opportunity for a Markov based classification, and sometimes subclassification of distantly related phages, independently of shared molecular markers or pairwise alignment, but still in accordance with the ICTV classification scheme.

3.5. Inoviridae Taxonomy Phages and Profile HMMs

The nine HMMs specific for *Inoviridae* infecting *Vibrionaceae* were used to scan proteins encoded by all known *Inoviridae*. Profile HMMs scan resulted in a number of positive matches (Table S12) reflecting that the *Inoviridae* phages infecting *Ralstonia*, *Enterobacteria*, *Pseudomonas*, *Xanthomonas*, and *Stenotrophomonas* encode proteins of the same families as the *Inoviridae* infecting *Vibrionaceae* (Figure 4).

Four out of the nine vibriophage generated HMMs had hits only to *Inoviridae* infecting *Vibrio* hosts proteomes. The rest matched to proteins from non-*VibrioInoviridae*. Although all investigated *Inoviridae* genome encodes more than one *vibrio Inoviridae* like protein, not a single protein family was present in all phages. The most commonly shared protein family members are zot-like proteins, which have been found in 95% of all phages [28]. According to Mai-Prochnow et al. [28] the genomes of *Inoviridae* range within a size of 4 Kbp to 12 Kbp which gives spaces to encode up to 11 genes. The *Inoviridae* profile HMMs generated within this work contain 19 protein families which explains why not each HMM finds a protein in each *Inoviridae* genome supporting the contribution to virulence of the phage class [44]. However, what is indicative for a member of *Inoviridae* is the set of proteins that are found exclusively in members of this phage family [28].

The generated *Inoviridae Vibrio* derived profiles scanning the proteomes of all *Inoviridae* phages gave us the unique opportunity to explore the extent to which proteins are shared between *Inoviridae* infecting different bacterial hosts.

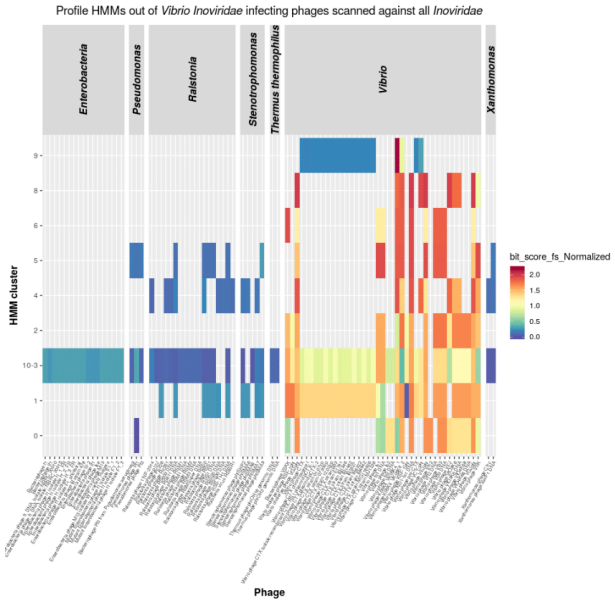


Figure 4. HMM scan results of all *Inoviridae* phages.This heatmap shows an *Inoviridae* derived profile HMM (*y*-axis) scan on the proteins of 119 *Inoviridae* genomes grouped by host genome (*x*-axis). HMMs ranged from hits specific to *Inoviridae* infecting *Vibrionaecea* to general hits for *Inoviridae* infecting other hosts. The indicator for the quality of a hit is color coded to the normalized bit-score assigned for the respective match by hmmscan.

3.6. Taxonomy of Podoviridae

To elucidate the taxonomic relation of *Podoviridae* identified by profile HMMs, an extended scan with the *Vibrio Podoviridae* models were performed against a set of *Podoviridae* that infect other bacterial hosts (Figure 5).

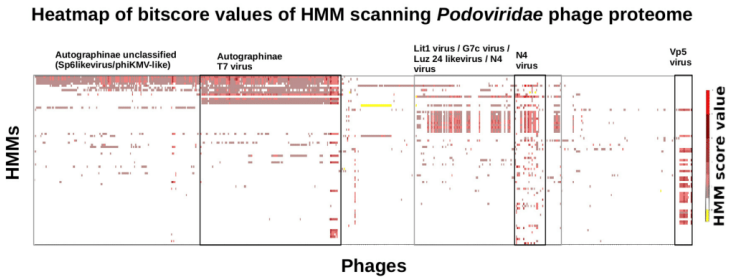


Figure 5. Profile HMM scan of *Podoviridae* HMMs from *Vibrionaceae* versus genomes from *Podoviridae* phages infecting non-vibrio hosts. This heatmap shows a profile HMM scan on the proteome of 1066 *Podoviridae* genomes. Sufficient hits were generated to discriminate four groupings of *Podoviridae*. The HMMs have been integrated in the heatmap (*y*-axis). The HMMs are grouped (on the *x*-axis) into general *Podoviridae* subclassifications. The indicator for the quality of a hit is color coded to thenormalized bit-score assigned for the respective match by hmmscan. The generated hmmscan output was visualized using matplotlib library in Python 3.5.

The *Podoviridae* profile HMMs from vibriophages exhibited, as in the case of the *Inoviridae*, hits to multiple proteins out of all published *Podoviridae* phages (Table S13). *Podoviridae* represent a much more complex and diverse class of phages compared to *Inoviridae*. The genomes of *Podoviridae* from *Vibrionaceae* comprise 96 distinct protein families. However, when grouped by shared HMM hits and hosts the results of the scan display a degree of specificity and sensitivity that may be useful to subclassify the taxon. It is no surprise that phages that prey on the same host share proteins. However, the scores of the HMM hits reflect the degree of similarity shared by the single proteins. Thus, the heatmap shows the diversity of the different protein classes and thus gives us an idea of the phylogenetic history of the proteins.

3.7. CDS Prediction and Additional HMM Refinement

The genome annotation of public available phages is the product of gene prediction programs with different sensitivity [45–48]. This results in genomes where some CDS have not been annotated. To examine the value of HMMs to identify such missing phage CDS, the intergenic regions of each phage genbank file used in this study was scanned using the profile HMMs. In total, 234 nucleotide regions were identified encoding gene products that align to one of the protein families modelled by the HHMs (Table S14). Indeed, profile HMMs can be used to identify missing CDS.

To investigate whether these new proteins may improve the profile HMMs, we generated refined HMMs using the original proteins plus the new identified CDS as described in the material and methods section. An evaluation of the refined HMMs identified exactly the same proteins per HMM with slightly moderated hit scores. The test revealed that the refinement of the HMMs did not yield better performing HMMs. The sensitivity of HMMs is correlated much stronger to the diversity than to the number of the proteins used in the initial alignment step. We concluded that our original profile HMMs already contain sufficient diverse proteins to model the protein families and thus the model's predictive power is already close to saturation.

3.8. Identification and Classification of Prophages within Bacterial Genomes

Scan of Positive Dataset of Vibrio Genomes

Apart from phage genomes generated from phage particles that have been experimentally confirmed to infect bacteria, host genomes themselves contain in many cases integrated prophages derived from old infection events [49]. To examine the reliability of the profile HMMs with regards to their ability to identify and support the classification of bacteriophages integrated within a bacterial genome, a scan of 10 sequenced *Vibrio* strains with experimentally proven active *Inoviridae* prophages [29] was performed. The bacteriophage family specific HMMs were used to search for matches within the complete protein sets of nine *Vibrio alginolyticus* and one *Vibrio typhli* genome. The same strains have been scanned using PHASTER for phage identification. Whenever HMM hits co-localized and matched a prophage region predicted by PHASTER, they were represented in a separate facet (Figure 6).

In each of the genomes, the profile HMMs hits indicate the presence of genes encoding putative phage proteins. In the case of the strains *V. alginolyticus* K04M1 and K05K4 two complete replicons are present as extra-chromosomal phages [29] where the nine refined HMMs had matches. In all nine *V. alginolyticus* strains, *Inoviridae* derived profile HMMs match to a single locus on chromosome 2 of eight strains, and two other loci on the K09K1 strain.In some instances we could identify two distinct prophages that integrated in close proximity within the host chromosome [29] and was reflected by multiple hits of the same HMM in the same region. While strains K06K5 and K10K4 had an additional *Inoviridae* integrated at the same locus on chromosome 1. For the *V. alginolyticus* strains it has been shown by the Phage-seq method [50] that the corresponding genome regions express biological active *Inoviridae* particles encoding the protein sequences that match the profile HMM.

The scan of a positive data set of 9 *V. alginolyticus* and 1 *V. typhli* genome confirmed several hits for *Inoviridae* proteins, where the integrated prophages were located and experimentally confirmed as well as on three extra-chromosomal *Inoviridae* phages supporting the reliability of the method.

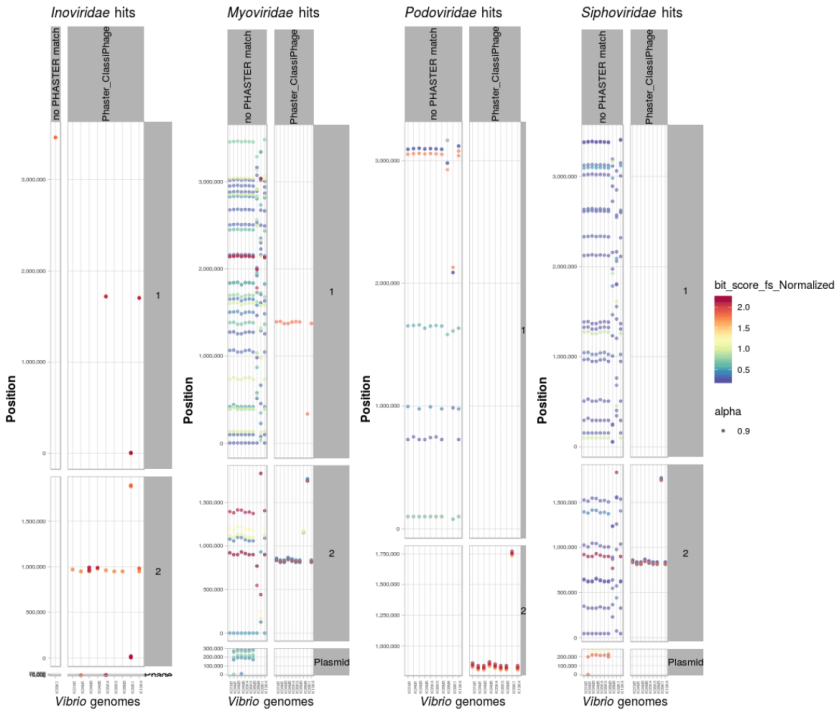


Figure 6. HMM search for prophages in *Vibrio* genomes with proven phage activities.Family specific HMMs constructed for *Ino-*, *Myo-*, *Podo-*, *andSiphoviridae* (grouped on *x*-axis) were used to scan all proteins derived from the genome of nine *V. alginolyticus* and one *V. typhli* genomes (*x*-axis per phage family grouping). In all of the *V. alginolyticus* genomes, regions encoding proteins matching to the profile HMMs were found (plotted per position and grouped per replicon on the *y*-axis). In cases where a region with consecutive HMM hits predicted as well by PHASTER was separately faceted.

3.9. PHASTER and ClassiPhage Scan of Published VibrioGenomes, Commonly and Additional Identified Phage Regions

PHASTER scan of 158 published closed *Vibrio* genomes resulted in the prediction of 458 prophages, out of which 143 were confirmed by the ClassiPhage scan (Table S15). Additionally, 64 regions where more than three consecutive HMM hits have been predicted by ClassiPhage that indicate protein genes of phage origin (Table S16). In addition to locus identification, ClassiPhage enabled us to taxonomically classify the prophages into *Ino-*, *Myo-*, *Podo-*, or *Siphoviridae* (Table S16, Figure S2). Most phages (>90%) could be classified into *Inoviridae* and some as *Podoviridae*. Our results further support the findings that *Inoviridae* are the most frequent phages infecting *Vibrio* species [21]. For *Myoviridae* and *Siphoviridae* HMM hits of one hypothetical protein it is not enough to classify.

On the other hand, the ClassiPhage method failed in identifying a set of 315 regions predicted by PHASTER. This set of genome regions encode proteins that match to proteins of phages infecting *Salmonella*, *E. coli*, *Bacilli* (Table S15, Figure S1) such as integrases, recombinase as well as proteins of

unknown functions. Note that phages share these kinds of proteins with other types of mobile genetic elements and that PHASTER characterized the vast majority of these loci as incomplete. However, in the case of a vibriophage reference in 61 cases, no HMM generation was possible due to the low number of proteins clustered during the HMM generation steps (vibriophage 12A4, vibriophage 12B12, *Vibrio* 8, *Vibrio* K139, *Vibrio* kappa, Vibrio N4, *Vibrio* pYD38, VfO3K6, Vf33, VfO4K68, VHML, VP4, VP882, VvAW1, X29).

Future developments, to overcome this limitation, would include using starting data sets not limited to vibriophages, rather using all available phages, generating HMMs and scanning diverse bacterial genomes. The possibility to generate more diverse and inclusive HMMs increases when more clusters generated out of closely related yet diverse phages are used, which reinforces the need to develop a method including more phage sequences, not limited to a host.

Additionally, the HMM scan resulted in hits that could not be assigned to a reference phage family. This might be evidence for vibriophages of so far unknown phage taxa or indicate false positive hits of the ClassiPhage method reflected by a low bit-score value, or due to HGT whenever the bit-score value was high. The scan of published *Vibrio* genomes generates much more hits than to a phage region, the reason why the combination of consecutive hits located in a certain region, size of the identified ORFs, annotation, E-value and bit scores are key to identify which hits belong to a phage and which do not. Hits not belonging to a prophage generally have low scores. In the case of a high score, proteins are annotated as "polymerases" or "flagellum" or "Transposon area", whereas phage related annotations are explained by being remnants of phages or by HGT.

The use of a combination of profile HMM hits for phage classification is a relatively new approach for the characterization of bacteriophages and thus further steps must be considered to better exploit the method [15].

4. Conclusions

In this work, we describe ClassiPhage, a method for phage classification independent of a shared molecular marker, based on combination of multiple profile HMM hits generated from a set of classified phage proteomes, and thus generating a Markov-based classification fitting the ICTV classification. We discussed the generation and refinement of profile HMMs, their validation across four different viral taxa and their application for viral taxonomic classification, focusing on vibriophages. Additionally, we used the generated HMMs to scan whole genomes and benchmarked the identified regions to PHASTER predicted prophage regions, to attempt viral identification prior to classification using the ClassiPhage method. We were able to show that the ClassiPhage method was able to reliably classify, by scanning the protein coding sequences of (i) a set of unclassified vibriophages; (ii) experimentally proven *Inoviridae*; and (iii) integrated phages in a set of closed and published *Vibrio* genomes, into one of the four phage families. We were also able to show that the method is not limited to vibriophages but the potential of the method extends towards phage subclassification, especially in the case of *Podoviridae*. This analysis supports the correlation of the generated HMMs per vibriophage family to the bacterial host. Lastly, we were able to show the potential of the method to be used as a phage identification and classification tool by scanning bacterial genomes using the refined HMMs and analyzing the protein sequence hits with regards to their consecutive location in the host genome. This method showed limitations for the case when scanned unclassified phages had one ambiguous hit to the refined HMMs and when phages identified by PHASTER which were missed by the ClassiPhage method. Phage identification must be coupled with sequence features for correct phage boundary identification. This limitation is a consequence of the quality and the constraints of the HMMs generation step, which makes it clear that fundamental steps must be considered to generate better and more comprehensive viral derived refined HMMs. We foresee that, with an ever-increasing amount of viral sequences and with the generation of robust and comprehensive viral HMMs, this method has the ability to classify phages into their taxonomic family in accordance with the ICTV scheme. The generated scans can subsequently be used in machine learning approaches to automatically classify viral sequences.

Supplementary Materials: The following are available online at http://www.mdpi.com/1999-4915/11/2/195/s1, Figure S1: PHASTER identified prophages and phage reference hit taxonomy, Figure S2: ClassiPhage scan of 158 Vibrio genomes hits and classification, Table S1: S1_Vibriophages_classifiedtitle, Table S2: S2_Phages_Test_dataset, Table S3: S3_VibrioGenomes_scanned, Table S4: S4_HMMs_calculation, Table S5: Proteins used to generate refined HMMs for Inoviridae vibriophages, Table S6: Proteins used to generate refined HMMs for Myoviridae vibriophages, Table S7: Proteins used to generate refined HMMs for Podoviridae vibriophages, Table S8: Proteins used to generate refined HMMs for Siphoviridae vibriophages, Table S9: Length of every generated HMM cluster per phage family, Table S10: Cross Scans of HMMs derived from the four different vibriophage families, Table S11: Refined profile HMMs scanning unclassified vibriophages, Table S12: Inoviridae vibriophages generated profile HMMs scanning all other Inoviridae phages, Table S13: Podoviridae vibriophages generated profile HMMs scanning all other Podoviridae phages, Table S14: All refined HMMs scanning the proteome of 10 Vibrio genomes with proven Inoviridae phage activity, Table S15: PHASTER predicted phages of 158 Vibrio genomes, Table S16: PHASTER predicted phages in common with IdentiPhage Vibrio scan.

Author Contributions: C.M.C. performed research, designed algorithm, wrote program, performed data analysis, and wrote the manuscript; A.F. generated Markov models, wrote the program, and performed data analysis; S.K. compared and visualized phage genomes; S.D. designed the algorithm; H.L. designed research, analyzed data, designed the algorithm, and wrote the manuscript.

Funding: KAAD for stipend, Department of Genomics and Applied Microbiology, Open access fund of DFG.

Acknowledgments: We thank Carolin Wendling and Olivia Roth for providing the *Vibrio alginolyticus* genomes as well as Tarek Morsi and Marc Dornieden for excellent IT-support. We acknowledge support by the German Research Foundation and the Open Access Fund of the Goettingen University.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chow, C.-E.T.; Suttle, C.A. Biogeography of Viruses in the Sea. *Annu. Rev. Virol.* **2015**, *2*, 41–66. [CrossRef] [PubMed]
2. Suttle, C.A. Marine viruses-Major players in the global ecosystem. *Nat. Rev. Microbiol.* **2007**, *5*, 801–812. [CrossRef] [PubMed]
3. Adams, M.J.; Lefkowitz, E.J.; King, A.M.Q.; Harrach, B.; Harrison, R.L.; Knowles, N.J.; Kropinski, A.M.; Krupovic, M.; Kuhn, J.H.; Mushegian, A.R.; et al. 50 years of the International Committee on Taxonomy of Viruses: Progress and prospects. *Arch. Virol.* **2017**, *162*, 1441–1446. [CrossRef] [PubMed]
4. Ackermann, H.W. Classification of Bacteriophages. In *The Bacteriophages*; Calendar, R., Ed.; Oxford University Press: New York, NY, USA, 2006; p. 746, ISBN 9780195148503.
5. Simmonds, P.; Adams, M.J.; Benk, M.; Breitbart, M.; Brister, J.R.; Carstens, E.B.; Davison, A.J.; Delwart, E.; Gorbalenya, A.E.; Harrach, B.; et al. Consensus statement: Virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.* **2017**, *15*, 161–168. [CrossRef] [PubMed]
6. Adriaenssens, E.M.; Rodney Brister, J. How to name and classify your phage: An informal guide. *Viruses* **2017**, *9*, 70. [CrossRef] [PubMed]
7. Roux, S.; Solonenko, N.E.; Dang, V.T.; Poulos, B.T.; Schwenck, S.M.; Goldsmith, D.B.; Coleman, M.L.; Breitbart, M.; Sullivan, M.B. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* **2016**, *4*, e2777. [CrossRef] [PubMed]
8. Roux, S.; Tournayre, J.; Mahul, A.; Debroas, D.; Enault, F. Metavir 2: New tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinform.* **2014**, *15*. [CrossRef] [PubMed]
9. Krupovic, M.; Dutilh, B.E.; Adriaenssens, E.M.; Wittmann, J.; Vogensen, F.K.; Sullivan, M.B.; Rumnieks, J.; Prangishvili, D.; Lavigne, R.; Kropinski, A.M.; et al. Taxonomy of prokaryotic viruses: update from the ICTV bacterial and archaeal viruses subcommittee. *Arch. Virol.* **2016**, *161*, 1095–1099. [CrossRef] [PubMed]
10. Lefkowitz, E.J.; Dempsey, D.M.; Hendrickson, R.C.; Orton, R.J.; Siddell, S.G.; Smith, D.B. Virus taxonomy: The database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.* **2018**, *46*, D708–D717. [CrossRef] [PubMed]
11. Shapiro, J.W.; Putonti, C. Gene Co-occurrence Networks Reflect Bacteriophage Ecology and Evolution. *MBio* **2018**, *9*, 1–14. [CrossRef] [PubMed]
12. Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P.; Peplies, J.; Glöckner, F.O. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **2013**, *41*, 590–596. [CrossRef] [PubMed]

13. Madslien, E.H.; Olsen, J.S.; Granum, P.E.; Blatny, J.M. Genotyping of B. licheniformis based on a novel multi-locus sequence typing (MLST) scheme. *BMC Microbiol.* **2012**, *12*, 230. [CrossRef] [PubMed]
14. Gonzalez-Escalona, N.; Jolley, K.A.; Reed, E.; Martinez-Urtaza, J. Defining a Core Genome Multilocus Sequence Typing Scheme for the Global Epidemiology of Vibrio parahaemolyticus. *J. Clin. Microbiol.* **2017**, *55*, 1682–1697. [CrossRef] [PubMed]
15. Reyes, A.; Alves, J.M.P.; Durham, A.M.; Gruber, A. Use of profile hidden Markov models in viral discovery: Current insights. *Adv. Genomics Genet.* **2017**, *7*, 29–45. [CrossRef]
16. Meier-Kolthoff, J.P.; Göker, M. VICTOR: Genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics* **2017**, *33*, 3396–3404. [CrossRef] [PubMed]
17. Roux, S.; Hallam, S.J.; Woyke, T.; Sullivan, M.B. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *Elife* **2015**, *4*, 1–20. [CrossRef] [PubMed]
18. Lima-Mendez, G.; Van Helden, J.; Toussaint, A.; Leplae, R. Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.* **2008**. [CrossRef] [PubMed]
19. Iranzo, J.; Krupovic, M.; Koonin, E.V. The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. *MBio* **2016**, *7*, 1–21. [CrossRef] [PubMed]
20. Deschavanne, P.; DuBow, M.S.; Regeard, C. The use of genomic signature distance between bacteriophages and their hosts displays evolutionary relationships and phage growth cycle determination. *Virol. J.* **2010**, *7*, 1–12. [CrossRef] [PubMed]
21. Castillo, D.; Kau, K.; Hussain, F.; Kalatzis, P.; Rørbo, N.; Polz, M.F.; Middelboe, M. Widespread distribution of prophage-encoded virulence factors in marine Vibrio communities. *Sci. Rep.* **2018**, *8*, 9973. [CrossRef] [PubMed]
22. Naser, I.B.; Hoque, M.M.; Abdullah, A.; Bari, S.M.N.; Ghosh, A.N.; Faruque, S.M. Environmental bacteriophages active on biofilms and planktonic forms of toxigenic Vibrio cholerae: Potential relevance in cholera epidemiology. *PLoS ONE* **2017**, *12*, e0180838. [CrossRef] [PubMed]
23. Bolduc, B.; Jang, H.B.; Doulcier, G.; You, Z.-Q.; Roux, S.; Sullivan, M.B. vConTACT: An iVirus tool to classify double-stranded DNA viruses that infect *Archaea* and *Bacteria*. *PeerJ* **2017**, *5*, e3243. [CrossRef] [PubMed]
24. Aiewsakun, P.; Adriaenssens, E.M.; Lavigne, R.; Kropinski, A.M.; Simmonds, P. Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: Steps towards a unified taxonomy. *J. Gen. Virol.* **2018**, *99*, 1331–1343. [CrossRef] [PubMed]
25. Kim, E.J.; Yu, H.J.; Lee, J.H.; Kim, J.-O.; Han, S.H.; Yun, C.-H.; Chun, J.; Nair, G.B.; Kim, D.W. Replication of *Vibrio cholerae* classical CTX phage. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 2343–2348. [CrossRef] [PubMed]
26. Fan, F.; Kan, B. Survival and proliferation of the lysogenic bacteriophage CTXΦ in Vibrio cholerae. *Virol. Sin.* **2015**, *30*, 19–25. [CrossRef] [PubMed]
27. Smeal, S.W.; Schmitt, M.A.; Pereira, R.R.; Prasad, A.; Fisk, J.D. Simulation of the M13 life cycle I: Assembly of a genetically-structured deterministic chemical kinetic simulation. *Virology* **2017**, *500*, 259–274. [CrossRef] [PubMed]
28. Mai-Prochnow, A.; Hui, J.G.K.; Kjelleberg, S.; Rakonjac, J.; McDougald, D.; Rice, S.A. Big things in small packages: The genetics of filamentous phage and effects on fitness of their host. *FEMS Microbiol. Rev.* **2015**, *39*, 465–487. [CrossRef] [PubMed]
29. Wendling, C.C.; Piecyk, A.; Refardt, D.; Chibani, C.; Hertel, R.; Liesegang, H.; Bunk, B.; Overmann, J.; Roth, O. Tripartite species interaction: Eukaryotic hosts suffer more from phage susceptible than from phage resistant bacteria. *BMC Evol. Biol.* **2017**, *17*, 98. [CrossRef] [PubMed]
30. Nelson, E.J.; Harris, J.B.; Morris, J.G., Jr.; Calderwood, S.B.; Camilli, A. Cholera transmission: The host, pathogen and bacteriophage dynamic. *Nat. Rev. Microbiol.* **2009**, *7*. [CrossRef] [PubMed]
31. Smeal, S.W.; Schmitt, M.A.; Pereira, R.R.; Prasad, A.; Fisk, J.D. Simulation of the M13 life cycle II: Investigation of the control mechanisms of M13 infection and establishment of the carrier state. *Virology* **2017**, *500*, 275–284. [CrossRef] [PubMed]
32. Senčilo, A.; Luhtanen, A.-M.; Saarijärvi, M.; Bamford, D.H.; Roine, E. Cold-active bacteriophages from the Baltic Sea ice have diverse genomes and virus-host interactions. *Environ. Microbiol.* **2014**. [CrossRef] [PubMed]
33. Doss, J.; Culbertson, K.; Hahn, D.; Camacho, J.; Barekzi, N. A review of phage therapy against bacterial pathogens of aquatic and terrestrial organisms. *Viruses* **2017**, *9*, 50. [CrossRef] [PubMed]

34. Tan, D.; Gram, L.; Middelboe, M. Vibriophages and their interactions with the fish pathogen Vibrio anguillarum. *Appl. Environ. Microbiol.* **2014**, *80*, 3128–3140. [CrossRef] [PubMed]
35. Alves, J.M.P.; De Oliveira, A.L.; Sandberg, T.O.M.; Moreno-Gallego, J.L.; De Toledo, M.A.F.; De Moura, E.M.M.; Oliveira, L.S.; Durham, A.M.; Mehnert, D.U.; De Zanotto, P.M.A.; et al. GenSeed-HMM: A tool for progressive assembly using profile HMMS as seeds and its application in Alpavirinae viral discovery from metagenomic data. *Front. Microbiol.* **2016**, *7*, 1–15. [CrossRef] [PubMed]
36. Enright, A.J.; Van Dongen, S.; Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **2002**, *30*, 1575–1584. [CrossRef] [PubMed]
37. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [CrossRef] [PubMed]
38. Eddy, S.R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **2011**, *7*. [CrossRef] [PubMed]
39. Lavigne, R.; Seto, D.; Mahadevan, P.; Ackermann, H.W.; Kropinski, A.M. Unifying classical and molecular taxonomic classification: Analysis of the Podoviridae using BLASTP-based tools. *Res. Microbiol.* **2008**, *159*, 406–414. [CrossRef] [PubMed]
40. Day, L.A. Inoviridae. *Virus Taxon.* **2012**, 375–383.
41. Veesler, D.; Cambillau, C. A Common Evolutionary Origin for Tailed-Bacteriophage Functional Modules and Bacterial Machineries. *Microbiol. Mol. Biol. Rev.* **2011**, *75*, 423–433. [CrossRef] [PubMed]
42. Lee, J.-H.; Shin, H.; Ryu, S. Characterization and comparative genomic analysis of bacteriophages infecting members of the Bacillus cereus group. *Arch. Virol.* **2014**, *159*, 871–884. [CrossRef] [PubMed]
43. Zahid, M.S.H.; Waise, T.M.Z.; Kamruzzaman, M.; Ghosh, A.N.; Nair, G.B.; Mekalanos, J.J.; Fanique, S.M. The cyclic AMP (cAMP)-cAMP receptor protein signaling system mediates resistance of Vibrio cholerae O1 strains to multiple environmental bacteriophages. *Appl. Environ. Microbiol.* **2010**, *76*, 4233–4240. [CrossRef] [PubMed]
44. Faruque, S.M.; Mekalanos, J.J. Phage-bacterial interactions in the evolution of toxigenic *Vibrio cholerae*. *Virulence* **2012**, *3*, 556–565. [CrossRef] [PubMed]
45. Aggarwal, G.; Ramaswamy, R. Ab initio gene identification: Prokaryote genome annotation with GeneScan and GLIMMER. *J. Biosci.* **2002**, *27*, 7–14. [CrossRef] [PubMed]
46. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **2014**, *30*, 2068–2069. [CrossRef] [PubMed]
47. Gao, F.; Zhang, C.-T. Ori-Finder: A web-based system for finding oriCs in unannotated bacterial genomes. *BMC Bioinform.* **2008**, *9*, 79. [CrossRef] [PubMed]
48. Linke, B.; McHardy, A.C.; Neuweger, H.; Krause, L.; Meyer, F. REGANOR: A gene prediction server for prokaryotic genomes and a database of high quality gene predictions for prokaryotes. *Appl. Bioinform.* **2006**, *5*, 193–198. [CrossRef] [PubMed]
49. Casjens, S. Prophages and bacterial genomics: What have we learned so far? *Mol. Microbiol.* **2003**, *49*, 277–300. [CrossRef] [PubMed]
50. Hertel, R.; Rodríguez, D.P.; Hollensteiner, J.; Dietrich, S.; Leimbach, A.; Hoppert, M.; Liesegang, H.; Volland, S. Genome-Based Identification of Active Prophage Regions by Next Generation Sequencing in Bacillus licheniformis DSM13. *PLoS ONE* **2015**, *10*, e0120759. [CrossRef] [PubMed]

Supplementary information

Supplementary information for this manuscript can be found at the “Viruses” and “G2L” websites under the following addresses:

https://www.mdpi.com/1999-4915/11/2/195

http://appmibio.uni-goettingen.de/index.php?sec=sw

Additionally, supplementary figures and tables are provided along with the electronic version of this thesis (on DVD), under the following paths:

Additional Data:

Ino_refined_HMMs:
SupplementaryMaterial/ChapterII/ChapterII.3/HMMs/Ino_refined_HMMs/

Myo_refined_HMMs:
SupplementaryMaterial/ChapterII/ChapterII.3/HMMs/Myo_refined_HMMs/

Podo_refined_HMMs:
SupplementaryMaterial/ChapterII/ChapterII.3/HMMs/Podo_refined_HMMs/

Sipho_refined_HMMs:
SupplementaryMaterial/ChapterII/ChapterII.3/HMMs/Sipho_refined_HMMs/

Data:

SupplementaryMaterial/ChapterII/ChapterII.3/HMMs/Ign_hmm_scan.csv

Additional Figures:

Figure S1: SupplementaryMaterial/ChapterII/ChapterII.3/Figure S1.tiff

Figure S2: SupplementaryMaterial/ChapterII/ChapterII.3/Figure S2.tiff

Figure S3: SupplementaryMaterial/ChapterII/ChapterII.3/Figure S3.png

Additional Tables:

Table S1: SupplementaryMaterial/ChapterII/ChapterII.3/Table_S1.xlsx

Table S2: SupplementaryMaterial/ChapterII/ChapterII.3/Table_S2.xlsx

Table S3: SupplementaryMaterial/ChapterII/ChapterII.3/Table_S3.xlsx

Table S4: SupplementaryMaterial/ChapterII/ChapterII.3/Table_S4.xlsx

Table S5: SupplementaryMaterial/ChapterII/ChapterII.3/Table_S5.xlsx

Table S6: SupplementaryMaterial/ChapterII/ChapterII.3/Table_S6.xlsx

Table S7: SupplementaryMaterial/ChapterII/ChapterII.3/Table_S7.xlsx

Table S8: SupplementaryMaterial/ChapterII/ChapterII.3/Table_S8.xlsx

Table S9: SupplementaryMaterial/ChapterII/ChapterII.3/Table_S9.xlsx

Table S10: SupplementaryMaterial/ChapterII/ChapterII.3/Table_S10.xlsx

Table S11: SupplementaryMaterial/ChapterII/ChapterII.3/Table_S11.xlsx

Table S12: SupplementaryMaterial/ChapterII/ChapterII.3/Table_S12.xlsx

Table S13: SupplementaryMaterial/ChapterII/ChapterII.3/Table_S13.xlsx

Table S14: SupplementaryMaterial/ChapterII/ChapterII.3/Table_S14.xlsx

Table S15: SupplementaryMaterial/ChapterII/ChapterII.3/Table_S15.xlsx

Table S16: SupplementaryMaterial/ChapterII/ChapterII.3/Table_S16.xlsx

Table S17: SupplementaryMaterial/ChapterII/ChapterII.3/Table_S17.xlsx

Table S18: SupplementaryMaterial/ChapterII/ChapterII.3/Table_S18.xlsx

Scripts:

SupplementaryMaterial/ChapterII/ChapterII.3/Scripts/

SupplementaryMaterial/ChapterII/ChapterII.3/Scripts/ReadMe.txt

SupplementaryMaterial/ChapterII/ChapterII.3/Scripts/pipeline_overview.png

II.4 ClassiPhage 2.0: Sequence-based classification of phages using Artificial Neural Networks

ClassiPhage 2.0: Sequence-based classification of phages using Artificial Neural Networks

Cynthia Maria Chibani, Florentin Meinecke, Anton Farr, Sascha Dietrich, Heiko Liesegang

Chibani et al. (2019)**DOI** https://doi.org/10.1101/558171

Authors' contributions

CC performed research, designed algorithm,

FM designed algorithm, wrote program, performed data analysis,

AF wrote program to refine Hidden Markov Models,

SD designed algorithm,

HL designed research, analyzed data, corrected manuscript.

1 Title:

2 ClassiPhages 2.0: Sequence-based classification of phages using Artificial Neural Networks

3 Author's list:

4 Cynthia Maria Chibani, Florentin Meinecke, Anton Farr, Sascha Dietrich, Heiko Liesegang.

5 Author Information:

6 Institute for Microbiology and Genetics, Georg-August University Goettingen, Grisebachstr. 8,
7 37077, Goettingen, Germany

8 Corresponding author:

9 Heiko Liesegang, hlieseg@gwdg.de

10

Abstract:

Background/ Motivation:

In the era of affordable next generation sequencing technologies we are facing an exploding amount of new phage genome sequences. This requests high throughput phage classification tools that meet the standards of the International Committee on Taxonomy of Viruses (ICTV). However, an accurate prediction of phage taxonomic classification derived from phage sequences still poses a challenge due to the lack of performant taxonomic markers. Since machine learning methods have proved to be efficient for the classification of biological data we investigated how artificial neural networks perform on the task of phage taxonomy.

Results:

In this work, 5,920 constructed and refined profile Hidden Markov Models (HMMs), derived from 8,721 phage sequences classified into 12 well known phage families, were used to scan phage proteome datasets. The resulting Phage Family-proteome to Phage-derived-HMMs scoring matrix was used to develop and train an Artificial Neural Network (ANN) to find patterns for phage classification into one of the phage families. Results show that using the 100 fold cross-validation test, the proposed method achieved an overall accuracy of 84.18 %. The ANN was tested on a set of unclassified phages and resulted in a taxonomic prediction. The ANN prediction was benchmarked against the prediction resulting of multi-HMM hits, and showed that the ANN performance is dependent on the quality of the input matrix.

Conclusions:

We believe that, as long as some phage families on public databases are underrepresented, multi-HMM hits can be used as a classification method to populate those phage families, which in turn will improve the performance and accuracy of the ANN. We believe that the proposed method is an effective and promising method for phage classification. The good performance of the ANN and HMM based predictor indicates the efficiency of the method for phage classification, where we foresee its improvement with an increasing number of sequenced viral genomes.

Keywords:

Phage; Classification; HMM; Machine Learning; Artificial Neural Networks

Introduction:

Bacteriophages, bacterial viruses infecting bacteria, are of utmost importance due to the role they play in bacterial evolution (Roux et al. 2016). Virus classification is based on the idea of an evolutionary relationship between viruses and groups of viruses having more ability to exchange genetic material (Hans-W Ackermann 2011). Virus taxonomy is currently the responsibility of the International Committee on the Taxonomy of Viruses (ICTV). As of March 2017, there exist 4,404 approved Species, 735 Genera, 35 Subfamilies, 122 Families and 8 Orders (Lefkowitz et al. 2017).

The traditional method for the classification of phages is based on deciphering the type of nucleic acid and virion morphology using Transmission Electron Microscopy (TEM)(Rohwer & Edwards 2002). Experimental identification and classification of phages is based on physiological data and needs time to perform the experiments and expertise on the culture conditions of the corresponding host and phage system. However, within the explosive growth of phage sequences in the era of next generation sequencing technologies, there is an increasing amount of phage derived sequences that lack physiological data and knowledge on the host of the phages, especially in the case of metagenome data. This poses challenges to the successful implementation of a method which correctly classifies phages(Skewes-cox et al. 2014). Therefore, the development of a sequence based computational method, with the flexibility to integrate newly sequence derived phage descriptors, is necessary to allow rapid and accurate classification.

It is a known fact that phages do not have a ribosomal gene to place them on the tree of life (Rohwer & Edwards 2002).Phage classification based nucleotide pairwise comparison limits the process to similarities to phages found within reference databases (Bolduc et al. 2017). This poses a challenge to phage sequences identified from metagenomic datasets, where in one study by Paez-Espino et al (Paez-Espino et al. 2016), they identified over 125,000 contigs which revealed no sequence similarity to known viruses.

To that extent, taxonomic systems based on phage proteomes were suggested; however they come with their limitations (Meier-Kolthoff & Göker 2017). Clustering techniques optimized for viral classification were applied by Lima-Mendez et al. (Lima-Mendez et al. 2008)and Roux et al. (Roux et al. 2015), which showed the efficiency of the use of phage clustering as a basis of classification.

Profile HMMs proved to be a powerful method to model the sequence diversity of a set of orthologs, and thus are sensitive and more effective than pairwise alignment methods in detecting divergent viral sequences (Skewes-cox et al. 2014; Reyes et al. 2017). Additionally, Chibani et al. 2019 (accepted) showed that the use of a combination of phage derived profile HMM hits proved to be efficient to classify previously unclassified phage genomes into different phage families.

74 The emerging fields and use of machine learning and data mining in different biological fields are
75 proving to be instrumental in answering challenging questions by looking into millions of biological
76 data produced in the last decade. Because of their success with big data, ANNs and other machine
77 learning models have gained a considerable amount of interest as a promising framework for
78 biology. When combined with genomic information, novel machine learning and data mining
79 techniques can advance the extraction of critical information and predict future observations from
80 big data. Considerable progress has been made in the application of Support Vector Machines
81 (SVM) (Manavalan, Tae H. Shin, et al. 2018; Tan et al. 2018) and Naïve Bayes (Feng et al. 2013)
82 machine learning algorithms to identify phage virion proteins and in the application of ANN to
83 classify tailed phages (currently deprecated) (Lopes et al. 2014). However, the use of machine
84 learning for phage taxonomic classification has not been reported so far. Therefore, it is necessary
85 to apply meaningful feature extraction and selection methods to investigate the classification
86 method.

87 In order to address the limitations of current phage taxonomic classification software, we focused
88 on the question of how profile HMMs (Chibani et al 2019 (accepted)) perform within a machine
89 learning approach for the automated classification of phage genome sequences. We designed and
90 developed an ANN, a well known supervised Machine Learning (ML) algorithm, which has been
91 applied to several biological problems (Arango-Argoty et al. 2018; Seguritan et al. 2012). The ANN
92 takes protein hits scores to phage derived profile HMMs per phage family as input, by applying a
93 set of thresholds to select optimal features for a phage classification method. The performance of
94 supervised prediction algorithms depends on the quality of the training data set. We therefore
95 generated a training data set to train an ANN to classify new phage genomes and whether the public
96 available phage genomes are sufficient. To our knowledge, this is the first ever reported use of
97 ANN for the classification of phages into phage families with a trusted performance to accuracy
98 ratio for the predictions.

Materials and Methods:

A five-step guideline has increasingly been endorsed (Manavalan, Tae Hwan Shin, et al. 2018) in a series of recent publications, to develop a sequence-based predictor for a biological system that can easily be used, which goes as follow:

(i) generating a solid benchmarking dataset to train and test the prediction model; (ii) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) develop a powerful algorithm to generate a prediction; (iv) implement cross-validation tests to objectively evaluate the performance of the predictor; and finally, (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, we describe the achieved steps.

Data Collection

The raw phage dataset used in this research were retrieved from millardlab database (http://millardlab.org/bioinformatics/bacteriophage-genomes/).

As of 20 March 2018, the database contained in total 8,721 phage genomes (Table S1) belonging to 21 phage families summarized in **Table 1**.

Table 1: Summary table of the phage families and number of phages belonging to each phage family found in the millardlab database as of 20 March 2018

ds/ss	DNA/RNA	Phage Family	Number
Classified Phages			
ds	DNA	*Ampullaviridae*	6
ds	DNA	*Bicaudaviridae*	10
ds	DNA	*Myoviridae*	1,766
ds	DNA	*Podoviridae*	1,066
ds	DNA	*Siphoviridae*	3,466
ds	DNA	*Corticoviridae*	2
ds	RNA	*Cystoviridae*	15
ds	DNA	*Fuselloviridae*	22
ds	DNA	*Globuloviridae*	4
ds	DNA	*Guttaviridae*	1
ds	DNA	*Haloviruses*	30
ss	DNA	*Inoviridae*	119
ss	RNA	*Leviviridae*	40
ds	DNA	*Ligamenvirales (Lipothrixviridae* and *Rudiviridae)*	49
ss	DNA	*Microviridae*	734
ds	DNA	*Plasmaviridae*	2
ds/ss	unclassified	*Pleolipoviridae*	16
ds	DNA	*Salteproviridae*	2
ss	DNA	*Spiraviridae*	1
ds	DNA	*Tectiviridae*	19
ds	DNA	*Turriviridae*	4
Unclassified Phages			
-	-	Generally unclassified phages	1,175
ds	DNA	unclassified phages	105

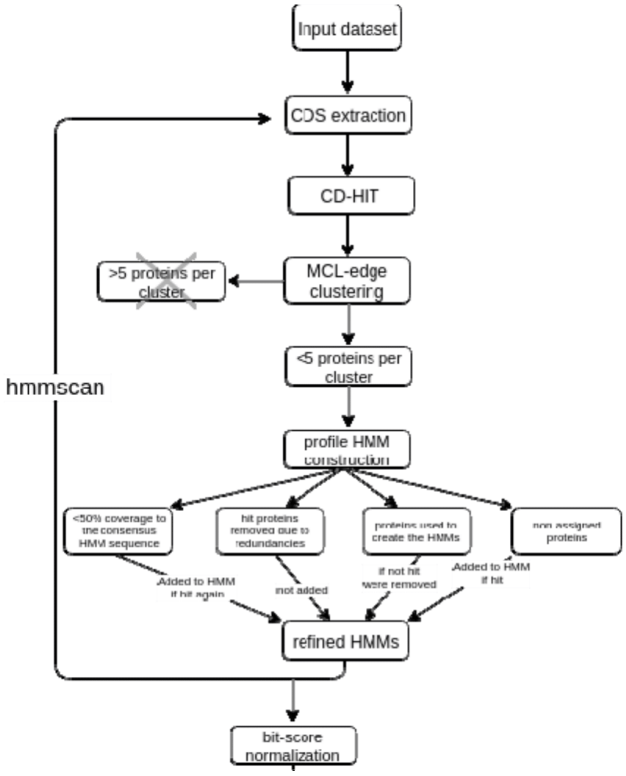
ds	DNA	*Caudovirales* unclassified phages	67

The first two columns represent the nucleic acid structure of the phage family. The third column represents the phage family and the fourth column represents the number of phages belonging to every phage family. ds: double stranded, ss: single-stranded, DNA: Deoxyribonucleic acid, RNA: Ribonucleic acid.

Data Construction

For the purpose of obtaining a reliable benchmark dataset, the following steps were considered. Phage families which had less than 15 phage genomes were excluded, in order to ensure diverse phages with diverse proteins for HMM generation. This step is crucial in order to differentiate between the highly biased number of *Siphoviridae* phages and least abundant ones. This resulted in 12 of the 21 phages families (*Cystoviridae*, *Fusselloviridae*, *Haloviruses*, *Inoviridae*, *Leviviridae*, *Ligamenvirales*, *Microviridae*, *Myoviridae*, *Pleolipoviridae*, *Podoviridae*, *Siphoviridae* and *Tectiviridae*) used for the benchmark dataset construction.

Figure 1: Overall framework of Phage_input_matrix construction.



Non-redundant CDS, extracted from classified phage gbk files, were used as input for the Markov Clustering algorithm (MCL-edge). Clusters including more than 5 proteins were used to generate profile HMMs. Profile HMMs were subjected to refinement steps after rescanning the input extracted CDS. Refinement included 1) proteins not reaching the coverage threshold of 50% of the HMM consensus sequence were removed, and if were hit again, added to the model; 2) proteins removed due to redundancies were not added to the model; 3) proteins used to create the HMMs themselves if were hit were kept, if not hit thus were removed from the model; 4) not yet assigned proteins were added to the model. Rescanning the input and refinement steps were repeated until no change was observed. Resulting HMM scan bit-scores were normalized, and a set of input features were extracted, using the generated HMMs scanning the input data set, resulting in a cross-scan matrix of HMM-Phage-Family correlation to Protein-Phage correlation, we call Phage_input_matrix.

HMM profiles from the 12 phage families were generated as described by Chibani et al. 2019 (accepted) (see **Figure 1** for an overview of the methodology). In summary, protein coding sequences were extracted from the phage Gbk files, and sequences containing non-standard amino acid residues were excluded, as their meanings are ambiguous. To avoid biases and over-fitting, redundant proteins defined by CD-HIT (v.4.5.4)(Li & Godzik 2006) program by applying a 100% sequence identity cut-off, were removed during HMM generation steps. It should be noted that redundant proteins were removed only from the dataset used for HMM construction and not for the testing dataset. MCL-edge (v12-068) (Enright 2002) was used to generate protein clusters out of a BLASTp scan of all-against-all input protein sequences. For the clusters which had more than 5 proteins, multi-sequence alignment (MSA) files were generated. Profile HMMs were generated, per MSA file, using “hmmbuild” from HMMER (v3.1b1) (Finn et al. 2011) with default parameters. Removed proteins were stored for later refinement.

The initially generated HMMs were then refined considering the following steps:

Firstly, the function "hmmemit" was used to create a consensus sequence from a generated profile HMM. This consensus sequence is closest in similarity to the majority of sequences used to create the respective HMM. Using "BLASTP" to align each protein of a cluster against the consensus sequence, proteins not reaching the coverage threshold of 50% were removed and stored for later refinement as well.

Secondly, the command "hmmpress" was used to create binary compressed data files (.h3m, .h3i, .h3f and .h3p) from a "profile HMM". With "hmmscan" the binary files were used to look for orthologous protein hits in the scanned dataset. Created profile HMMs were used to scan the input fasta files where protein hits could be mapped to a) proteins removed due to redundancies b) proteins used to create the HMMs themselves c) not yet assigned proteins.

Lastly, proteins which are hit and have not yet been assigned were added to the profile HMM. Proteins that were used to create the HMM and were not hit, were removed from the profile HMM. Proteins that are hit but were previously removed due to redundancies were not added. Whenever multiple HMMs hit the same sets of proteins as well as their inputs, they were merged. Refined HMMs were used to rescan the input fasta and, if needed, refinement steps of merging were repeated until no changes occur. Resulting HMM scan bit-scores were lastly normalized (see Data normalization section) for further analysis.

Feature extraction

The aim of this experiment was to train ANN Machine Learning (ML)-based model to accurately map input features generated from HMM scans, to predict the phage family a phage sequence belongs to, which is considered a multiclass classification problem. The key is to extract a set of informative features. We generated a set of input features for the ANN predictor, by scanning the proteomes of the 7,342 phages, of the remaining 12 phage families, using the generated 5,920 refined profile HMMs, which resulted in a cross-scan matrix of HMM-Phage-Family correlation to Protein-Phage correlation. The resulting bit-scores per HMM were extracted to generate input feature vectors for the training dataset with the phage family as the label.

For each individual phage of the phage family, one row is set up in the matrix, with the first two columns containing the bacteriophages name, which was later dropped, and phage family, which was used as the label. All other columns contain the bit-score value of the 5,920 HMM profiles scan of this phage protein sequences, or a default value of zero for no hit of that profile. We name our input matrix Phage_input_matrix.

Data normalization

The bit-score values were normalized by dividing the resulting HMM scan bit-score by the number of amino acids of the consensus sequence of every HMM cluster. Hits of insufficient quality were filtered (e-score value <1e-10,(Amgarten et al. 2018; Arango-Argoty et al. 2018)). Additionally, if the bias of a hit was larger than the bit-score it produced, or if the bit-score was below zero in the first place, the corresponding HMM profile hit was omitted. If negative bit-score values were allowed, this would increase the value of empty hit cells in the final input matrix to a value greater than zero, creating values of HMM profile hits in the training dataset where there are none in the input.

After the creation of the matrix is completed and prior to the training of the ANN, its values are normalized to range from of [0,1], by employing “Minmax” formula described in (Manavalan et al. 2014):

$$b = a - min\,(a)/max\,(a) - min(a) - min(a)$$

that can be used to reduce a k-dimensional array with any range to an array of the same shape covering a range from 0 to 1.

Artificial Neural Network

We employed ANN as our algorithm, the objective of which is to learn to recognize patterns in a given dataset. Once it has been trained on samples of your data, it can make predictions by detecting similar patterns in future data(Schmidhuber 2015)). The "softMax" function (Manavalan, Tae H. Shin, et al. 2018), which is defined as $b = exp\,(ai)/\Sigma\, exp(zj)$(Andrew Skabar, Dennis Wollersheim 2006), with a being a k-dimensional array. The resulting array, b, of the same shape as a, holds values ranging from 0 to 1 where all values in b add up to 1. Softmax was implemented as the activation function of the ANN's output layers.

Based on the difference between the model's predictions and the correct values, an error rate is calculated and the weights in each layer of the network are adjusted to reduce the error of the prediction. This procedure is performed from the output layer through the entire network to the input layer, hence the term back-propagation. The extent to which weights are adjusted is controlled by a learning rate. While linear and exponential decay functions did result in an increase of accuracy, the decay had to be gradual for the model to reach good prediction accuracy. This was achieved with high numbers of training epochs. We adapted the cosine decay, as discussed by (Loshchilov & Hutter 2016), proved to be the most efficient approach to decay the learning rate in our tested ANN architecture. In this study, we used the TensorFlow 1.10 package.

Cross-Validation and Independent Testing

Usually, the benchmark dataset comprises a training dataset for training and a testing dataset for testing the model. Here, we performed 100-fold cross-validation on the training dataset and the trained model was tested on the independent dataset to confirm the generality of the developed method. For that, the benchmark dataset is split into 100 subsets, where $1/100^{th}$ of the initial data used for each of the testing subsets and the remainder used for training and cross-validation is performed using each of these 100 subsets as the testing dataset. The model trains for 100 individual sessions, once for each subset, as it must not have trained on any entry it later classifies in a testing set.

Here, all entries of the initial set are classified after the classification has ended, but the results can still vary due to the random distribution of entries in each training/testing subset. It should be noted

216 that we performed 5 independent 100-fold cross-validations to confirm the robustness of the ML
217 parameters.

Performance Evaluation Criteria

To provide a simple method to measure the prediction quality, the following three metrics, sensitivity (Sn), specificity (Sp) and accuracy (Acc) were used and expressed as:

(i) $Sn = TP/(TP + FN)$

$0 < Sn < 1$

(ii) $Sp = TN/(TP + FP)$

$0 < Sp < 1$

(iii) $Acc = (TP + TN)/(TP + FP + TN + FN)$

$0 < Acc < 1$

where TP is the number of phage correctly predicted to be of their corresponding phage families; TN is number of non-classified phages predicted to be not belonging to any phage family; FP in the number of is the number of non-classified phages predicted to belong to a phage family; and FN in the number of classified phages predicted not to belong to any phage family.

To further evaluate the performance of the ANN and determine suitable thresholds for the prediction values of the different families, we employed receiver operating characteristic (ROC) curves for the classification of each family. The ROC curve was plotted with the specificity as the x-axis and sensitivity as the y-axis by varying threshold. The area under the curve (AUC) was used for model evaluation, with higher AUC values corresponding to better performance of the classifier. The quality of the proposed method can be objectively evaluated by measuring the AUC.

Results

Data Construction

This method resulted in 5,920 refined profile HMMs, derived from 7,342 phages classified into 12 phage families (**Table 2**).

Table 2: Summary table of the number of refined HMMs resulting per phage family

Phage Family	Refined HMMs
Cystoviridae	2
Fuselloviridae	21
Haloviruses	48
Inoviridae	21
Leviviridae	4
Ligamenvirales	70
Microviridae	11
Myoviridae	2,851
Pleolipoviridae	3
Podoviridae	701
Siphoviridae	2,170
Tectiviridae	18

The first represents the phage family. The second column represents the number of refined HMMs generated per phage family.

The cross scan matrix resulting from the scan of HMMs derived from one phage family against the proteome of the 11 other phages families resulted in 60,560 protein hits by input HMM (Table S2).

Neural Network Training and Classifications

The accuracy of the model during training was monitored using a scatter plot, which records the models performance on the testing set at every 10^{th} epoch of model training. Further collected metrics, the accuracy of the classification of the training and the testing data, as well as the learning rate at the given training epoch, were collected and plotted when training was complete (**Figure 2**). An overall prediction accuracy of 84.18 % was achieved by adopting ANN with a 100-fold cross-validation method on all phages in the dataset.

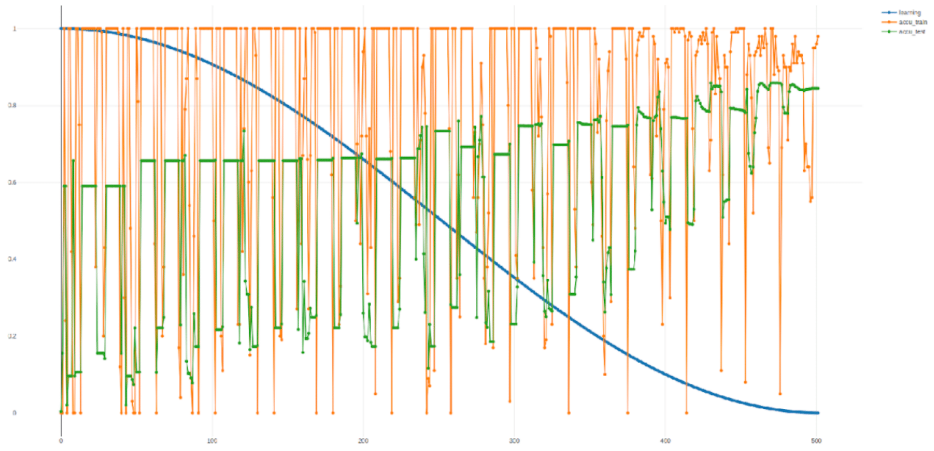


Figure 2: ANN performance on input matrix over training epochs.
The plot displays the trends of the learning rate, training set accuracy and testing set accuracy over 500 epochs. The high learning rate in early epochs shows the high fluctuation of accuracies between epochs, as the adjustment of the model's weights modifies it heavily. In the final epochs, the accuracy of the testing data classification reached 84.18%.

The scatter plot shows that the chosen batch size of 100 yielded the best result. We do not see information about possible issues with over- or under-fitting data. The model does not performs poorly on the testing set compared to the training set and thus did not result in over-fitting. Over-fitting results in a fluctuating training performance and low testing performance. Additionally, the

model did not result in a poorer performance on both the training and the testing set. Under-fitting of the model to the training set results in a training performance curve that is constantly higher than the testing curve. The learning rate displays a decrease with an increasing number of epochs, to reach 0, when the accuracy of the testing reaches its high of 84.18%. We conclude there is no reason to assume issues with an over- or under-fitting model.

Model performance and Metrics

The main output of the neural network is the label of the testing set and predictions of the model for each entry recorded at any training epoch. Using this information, the performance of the neural network can be accessed in detail for different stages of training. The labels of testing data are compared to the models assignments of the last recorded prediction by taking the maximum value of the models assignments.

As shown in **Table 3**, the TP, TN, FP, FN, Sp, Sn and Acc were calculated for the classification into the different phage families by using all 5,920 features.

Table 3: Predictive performance of the ANN per phage family

Phage Family	TP	TN	FP	FN	Sensitivity	Specificity	Accuracy
Cystoviridae	0	7,790	0	22	0	1	0.9971838
Fuselloviridae	0	7,782	0	15	0	1	0.9980762
Haloviruses	4	7,782	0	25	0.137931	1	0.9967994
Inoviridae	88	7,633	0	91	0.4916201	1	0.9883513
Leviviridae	25	7,776	0	11	0.6944444	1	0.9985919
Ligamenvirales	8	7,742	0	35	0.1860465	1	0.9955042
Microviridae	59	7,057	13	173	0.2543103	0.9981612	0.9745275
Myoviridae	577	6,548	40	647	0.4714052	0.9939284	0.9120584
Pleolipoviridae	0	7,796	0	16	0	1	0.9979519
Podoviridae	605	7,001	21	185	0.7658228	0.9970094	0.9736303
Siphoviridae	3,693	2,691	214	944	0.7964201	0.9325984	0.8517665
Tectiviridae	3	7,776	0	25	0.1071429	1	0.9967965

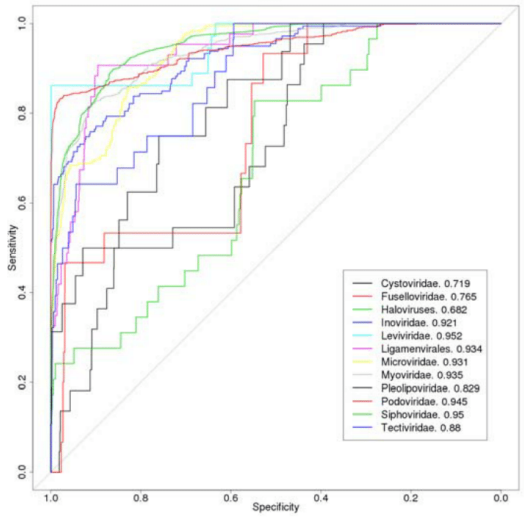
True or wrong phage classification prediction was assumed when the taxonomic prediction matched or did not match respectively the taxon that was given by the authors of the genome sequence. The number of correctly predicted phages (TP) of *Siphoviridae* (79.6%), *Podoviridae* (76.6%), *Leviviridae (*69.4 %), *Inoviridae* (49.1%), Myoviridae (45,5%), *Microviridae (*25.4%), *Haloviruses* (13.79%), *Ligamenvirales* (18.6%) and *Tectiviridae* (10.71%). Neither *Cystoviridae*, nor *Fuselloviridae*, or *Pleolipoviridae* were correctly predicted (TP = 0).

On the other hand, phage families where FP was predicted were *Microviridae*, *Myoviridae*, *Podoviridae* and *Siphoviridae*. All four phage families are known to infect bacterial hosts, however *Microviridae* are ss/DNA phages, whereas *Myo-*, *Podo-* and *Sipho-* are ds/DNA tailed phages belonging to the order of *Caudovirales*.

284 The clearest trend is the misclassification of entries to the *Siphoviridae* family. This occurs in
285 families that are closely related to *Siphoviridae* (*Myoviridae*, *Podoviridae*), but also in structurally
286 very distinct families such as *Fuselloviridae* and *Inoviridae*. This could indicate unexpected gene
287 flux between unrelated phage species (Shapiro & Putonti 2018).

288 ## ROC curves and thresholds

289 It is important to note that the confidence values in the final output of the model are not a
290 percentage of likelihood for the corresponding entry. For example, a value of 0.7 as the highest
291 value for an entry does not mean that the classification has a probability of 70% to be true.
292 However, it makes it possible to set a threshold value to distinguish between more and less
293 significant predictions. A higher threshold can improve the specificity of classification while a
294 lower threshold results in highly sensitive classification. One threshold may have different effects
295 on families, as the prediction scores are not calibrated between them. Thus, one score may be suited
296 to distinguish true positives from false positives in one family but inappropriate to do this in another
297 (Fawcett 2006). To determine suitable thresholds for the prediction values of different families,
298 ROC curves for the classification of each family were created and plotted using the R package



299 pROC (**Figure 3**).

300

Figure 3: ROC curve resulting from the ANN classification.

ROC curves out of the input matrix dataset prediction. The performance of the neural network ranges from near perfect prediction (AUC of 0.97 for the *Leviviridae* family) to almost random (AUC of 0.682 for the *Pleolipoviridae* family). The varying trends of the individual curves reflect that classifications of different families benefit from thresholds that are unique to them

From the ROC curves, AUC (Area Under the Curve) values were calculated, which provided insight into the prediction performance without a specific threshold. As the area in a ROC plot is always 1, the area under the curve can range from 0 to 1, with 0.5 representing no predictive power and 1 perfect prediction. It can be interpreted as an average performance metric for the classifier. All calculated AUCs for were displayed in the legend of the ROC curves (AUC of 0.719 for *Cystoviridae*, 0.765 for *Fuselloviridae*, 0.682 for *Haloviruses*, 0.921 for *Inoviridae*, 0.952 for *Leviviridae*, 0.934 for *Ligamenvirales*, 0.931 for *Microviridae*, 0.935 for *Myoviridae*, 0.829 for *Pleolipoviridae*, 0.945 for *Podoviridae*, 0.95 for *Siphoviridae* and 0.88 for *Tectiviridae*).

External dataset test

The proteomes of (~1,347) unclassified phages (Generally unclassified phages, ds/DNA unclassified phages and ds/DNA/*Caudovirales* unclassified phages) were scanned using the set of 5,920 refined profile HMMs. A matrix using the resulting bit-scores per HMM was generated, where the bit-scores were normalized as was described previously. We used the generated ANN to test the ability of the ClassiPhage 2.0 model to predict the phage family classification of the unclassified phages. Out of 1,175 generally unclassified phages, predicted phage families were *Inoviridae*, *Microviridae*, *Myoviridae*, *Pleolipoviridae*, *Podoviridae*, *Siphoviridae* and *Tectiviridae*. Out of 105 ds/DNA unclassified phages, predicted phage families were *Microviridae*, *Myoviridae*, *Podoviridae*, *Siphoviridae* and *Tectiviridae*. Finally, out of 67 ds/DNA/*Caudovirales* unclassified phages, predicted phage families were *Halovirus*, *Microviridae*, *Myoviridae*, *Podoviridae* and *Siphoviridae* (Table S8). *Haloviruses* and *Microviridae* can't be a classification for ds/DNA/*Caudovirales*, which shows that ClassiPhage 2.0 misclassifies phages where cross hits occur and enough family specific HMM hits.

We generate a heatmap of the prediction of the same set of unclassified vibriophages classified by Chibani et al 2019 (accepted) (**Figure 4**).

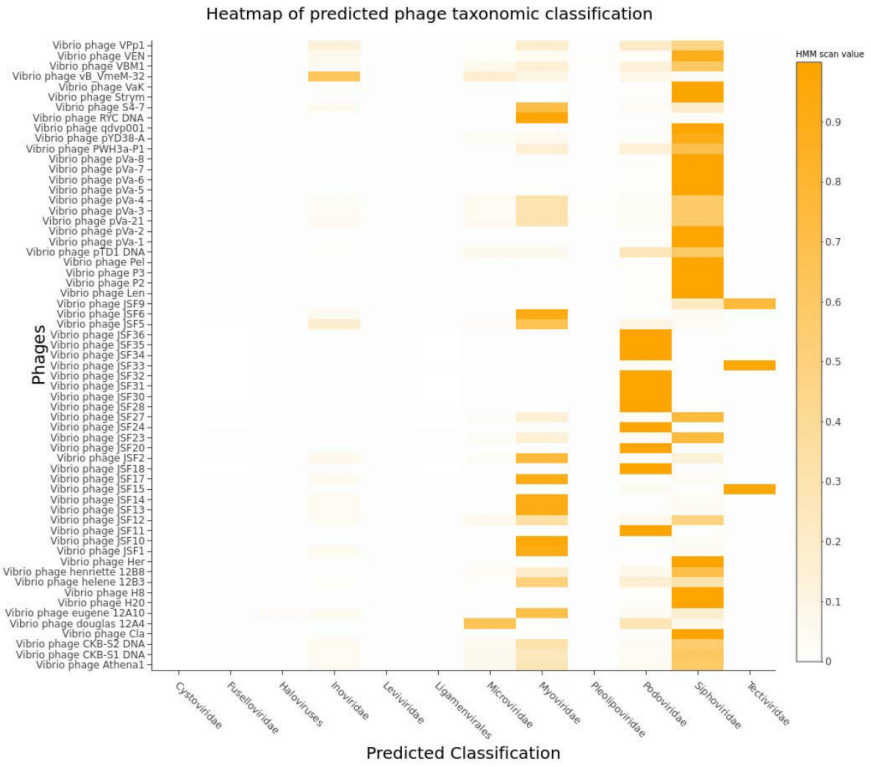


Figure 4: Heatmap of ClassiPhage 2.0 prediction of unclassified vibriophages.

A heatmap based on a phage family prediction of a set of unclassified vibriophages by the ClassiPhage 2.0 model, displaying the phage labels (y-axis) and phage family prediction (x-axis).

22 classified phages were consistent with the classification resulting in Chibani et al. 2019 (accepted). 23 phages which had an unclear classification were classified as *Siphoviridae* by ClassiPhage 2.0. Lastly, out of 17 phages which were not consistent between the two methods, the clearest trend was the misclassification of entries to the *Siphoviridae* phage family (Table S9).

Comparison to other methods

To the best of our knowledge, there exists no theoretical method for phage classification into phage families. Therefore, we cannot provide the comparison to analysis with published results to confirm that the model proposed here is superior to other methods. However, we generated a matrix out of the expected phage classification, as described in Chibani et al 2019 (accepted), to which we

compare the prediction of ClassiPhage 2.0 of the unclassified dataset. We display phage predictions resulting from ClassiPhage and ClassiPhage 2.0 (**Figure 5**).

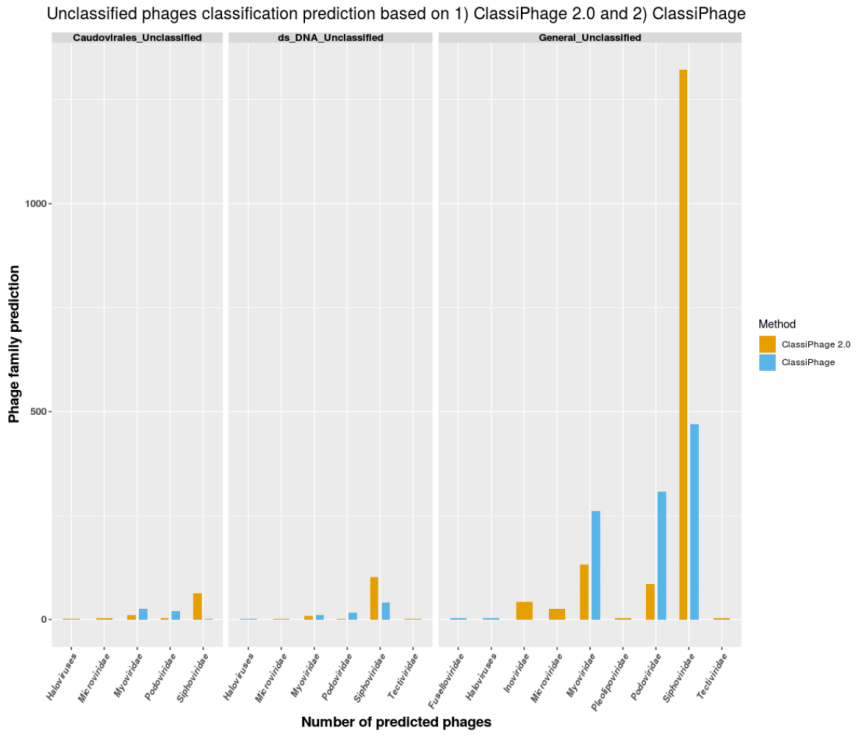


Figure 5: Barplot representing the classification of the unclassified phage dataset based on ClassiPhage 2.0 and ClassiPhage.

A bar plot summarizing phage classification prediction of 1) ds/DNA/*Caudovirales*, 2) ds/DNA unclassified phages and 3) generally unclassified phages based on ClassiPhage 2.0 (yellow bars) and ClassiPhage (blue bars). Displaying the count number (y-axis), and the grouped phage family prediction (x-axis).

HMM based phage classification, resulted in the classification of 835 out of 1,175 generally unclassified phages into 5 of the 12 phage families (3 *Fuselloviridae*, 3 *Haloviruses*, 261 *Myoviridae*, 307 *Podoviridae* and 261 *Siphoviridae*), and resulted in the classification of 67 out of 105 ds/DNA (1 *Halovirus*,10 *Myoviridae*,16 *Podoviridae* and 40 *Siphoviridae*) and 48 out of 67 ds/DNA/*Caudovirales* (26 *Myoviridae*, 20 *Podoviridae* and 2 *Siphoviridae*) (Tables S5 and S9).The performance of ClassiPhage 2.0 prediction in comparison to HMM based phage classification was

skewed towards *Siphoviridae* prediction, which is a consequence of the skewed input matrix of the ANN.

Discussion:

Phage classification based on phage sequencing data has long been a challenge, since phages have no conserved gene to place them on the tree of life (Rohwer & Edwards 2002). Although many pipelines exist for classification of prophages, these methods are based on the assumption that phages are monophyletic in origin and thus based on pairwise-alignment hits (Meier-kolthoff & Go 2018). This makes the classification of newly sequenced phages biased towards phage sequences available in the databases (Bolduc et al. 2017) and which is mostly skewed towards *Caudovirales (Skewes-cox et al. 2014)*. Therefore it is necessary to develop comprehensive computational methods for phage classification.

As stated by (Reyes & Gruber 2016), profile HMMs have an advantage over pairwise alignment in detecting remote homologs that are not part of the original MSA file used for the model's generation. Thus profiles HMMs are more sensitive when dealing with the highly complex and diverse phages and have the potential to increase the spectrum of detectable entities. On the other hand, since HMMs rely, to some degree, on the similarity to already known sequences available in the database, and since they represent a few sequences for a few over represented viral families, means that characterizing a greater number of viral sequences and regularly updating sequence databases are crucial for this method to be effective in the future (Skewes-cox et al. 2014; Reyes et al. 2017; Reyes & Gruber 2016).Although no HMMs exist for all phage proteins, the high scoring hits to a number of HMMs derived from a phage family were enough to classify a phage based on sequence information (Chibani et al. 2019, accepted). This means that combining multiple HMM hits is crucial since no single profile HMM can assess the true viral diversity of any sequenced dataset.

To this end, we developed and applied a novel ML approach called ClassiPhage 2.0, which allows the classification of phages based on their hits into one of 12 phage families. We demonstrate that by using multiple profiles HMM as input features, derived from phage proteins out of 12 phage families, we were able to predict the phage's taxonomic classification. Overall, we found that the method proved to be quite robust, within a range of reasonable parameter values, for the classification of the testing phage dataset, and for the assignment of a taxonomic classification of the unclassified phage dataset. However, supervised learning algorithms highly depend on the amount and quality of input data (Schmidhuber 2015). As it has been shown, phage information available in public databases is heavily biased with sequenced *Caudovirales (Skewes-cox et al. 2014; Reyes et al. 2017; Grazziotin et al. 2017)* and a large proportion of phage families are

underrepresented. This further emphasizes the importance of better and more comprehensive viral databases, enriching sequence representation of each of the viral taxa, which in turn will lead to robust models constructions and thus more sensitive and comprehensive input for ML classifiers (Manavalan, Tae H. Shin, et al. 2018; Arango-Argoty et al. 2018; Amgarten et al. 2018). A misclassification resulting from this approach is due to the random split nature of k-fold cross-validation. This creates the risk for the model to predict an entry of a family that was entirely absent from its training data, due to the presence of phage families with low number of HMMs associated. As our method's accuracy is highly dependent on the quality and accuracy of the input data, the better and more diverse the HMM models are, the better the neural network performs. That is to say that 1) whenever HMM hits are generally shared between multiple phage families such as "polymerases" or 2) if no HMM score was generated when scanning a phage proteome with the profile HMM models, then predictions are ambiguous in the first or cannot be made in the latter case. When scan outputs are not generated, the cause is that the phage belongs to a new phage family or is distant from the known phages (Roux et al. 2015). Finally, we expect the population of phage families with low abundant phages, from viral metagenomic datasets analysis. Since ANNs are known to perform better with an increasing size of a benchmark dataset (Morota et al. 2018), we foresee the improvement of ClassiPhage 2.0.

Conclusion:

In this study, we introduced a novel method which we call ClassiPhage 2.0. The method predicts a taxonomic phage family classification, resulting from multi-HMM hits of phages proteomes. We constructed ClassiPhage 2.0 using 5,920 refined profile HMMs as input features, derived from 7,342 phages classified into 12 phage families.

The results indicated that ClassiPhage 2.0 can be applied to predict a phage taxonomic classification at the family level with high accuracy. While these results are promising when observing the classification performance of one family on its own, it has proven challenging to accurately represent them in the context of all investigated families. To further elevate the performance of the neural network, as more phage data becomes available, more specific profile HMMs could be generated, improving the input datasets. In addition, the model could also be extended to include more features than HMM profile hits. This method can be further applied, for the prediction of well-delimited taxonomic groups such as subfamilies or families when profiles HMMs per subfamilies become well defined. Furthermore, the spectrum of potential applications of this approach is a general one and doesn't have to be limited to viral classification, rather could be applied to many other classification problems in bioinformatics.

This is a tool under active development to be made available as a publicly accessible easy-to-use web service, and we envisage its growing application on a variety of forthcoming projects.

Supplementary Data:

Supplemental Figure 1:

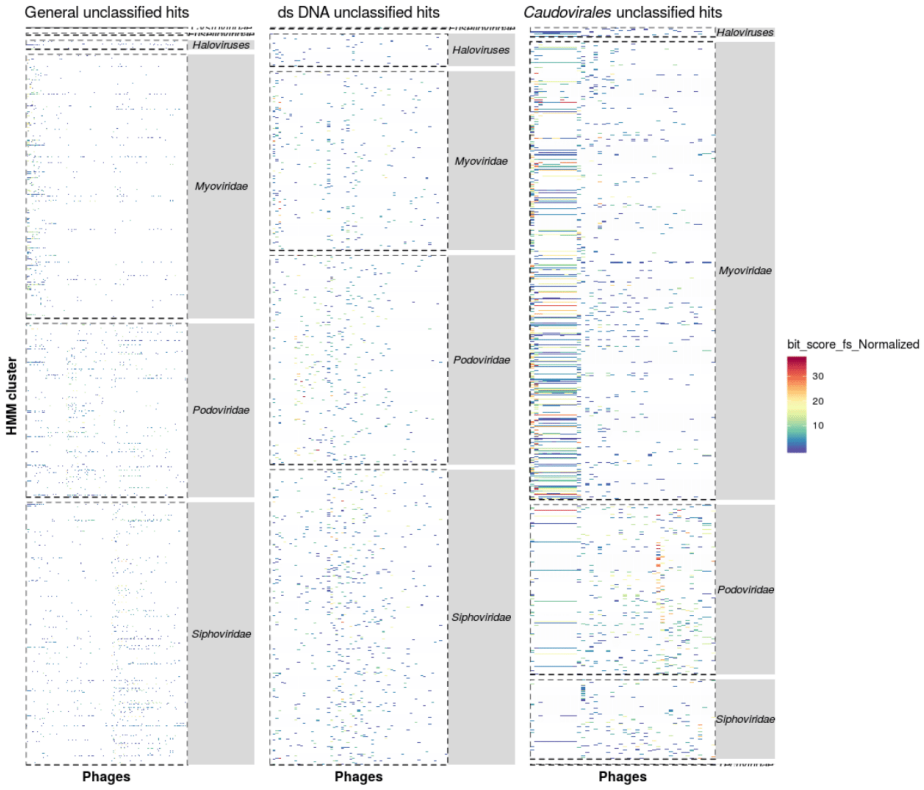


Figure S 1: Heatmap of phage family prediction of *Caudovirales* unclassified phages depending on combination of HMM hits.

The scan of the protein sequences derived from unclassified phages, was conducted by the profile HMMs of 12 phage families. The heatmap is split into 3 subplots (Generally unclassified phages, ds/DNA unclassified phages and ds/DNA/*Caudovirales*) where the phage family prediction is presented on the y-axis. The bit-score of the HMM matches was normalized by the size (in bp) of the HMM's consensus sequence (data see Table S5). The results are color-coded from blue (low-score) to red (high-score).

Supplemental Table S1: All phage dataset information

436 Phages test dataset downloaded from the millardlab database. The table contains information for the
437 phage, its classification and subclassification, size and accession number.
438 Supplemental Table S2: InputFamily generated HMMs scanning TargetFamily CDS
439 Refined HMMs derived from classified phages scanning all downloaded classified phage
440 proteomes. This table contains information for the cluster and its length, protein hit information,
441 which phage the protein is extracted from, the phages host, the input phages classification, the
442 scanned CDS phage classification and hmmscan information.
443 Supplemental Table S3: ClassiPhage 2.0 input matrix
444 Input matrix generated used as input to train and test ClassiPhage 2.0. This table contains
445 information of the phage, its classification and bit-score values resulting from refined HMMs scan
446 of the phage derived CDS.
447 Supplemental S4: Prediction layout of the ANN performed on the input matrix
448 ClassiPhage 2.0 predicted classification of classified phages. This table contains information about
449 the phage, it's published classification and ClassiPhage's 2.0 classification value ranging from [0,1].
450 An output close to 1 is ClassiPhage's 2.0 best predicted taxonomic classification.
451 Supplemental Table S5: InputFamily generated HMMs scanning unclassified phage CDS
452 Refined HMMs derived from classified phages scanning all downloaded classified phage
453 proteomes. This table contains information for the cluster and its length, protein hit information,
454 which phage the protein is extracted from, the phages host, the input phages classification and
455 hmmscan information.
456 Supplemental Table S6: Unclassified phage dataset matrix input for ClassiPhage 2.0
457 Input matrix generated used as an external dataset for classification using ClassiPhage 2.0 model.
458 This table contains information of the phage, unknown classification tag classification and bit-score
459 values resulting from refined HMMs scan of the phage derived CDS.
460 Supplemental Table S7: Prediction layout of the ANN for the unclassified phages dataset
461 ClassiPhage 2.0 predicted classification of unclassified phages. This table contains information
462 about the phage, 0 values for published classification and ClassiPhage's 2.0 classification values
463 ranging from [0,1]. An output close to 1 is ClassiPhage's best predicted taxonomic classification.

464 Supplemental Table S8: Unclassified phage dataset predicted taxonomic classification via

465 ClassiPhage 2.0 and ClassiPhages methods.

466 Supplemental Table S9: ANN prediction of unclassified Vibriophage dataset classified in Chibani et

467 al. 2019(accepted).

468 Excerpt out of Table S7, which contains information about ClassiPhage 2.0 output of the same set

469 of unclassified vibriophages classified by Chibani et al. 2019(accepted).

470

References:

Amgarten, D. et al., 2018. MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. *Frontiers in Genetics*.

Andrew Skabar, Dennis Wollersheim, T.W., 2006. Multi-label Classification of Gene Function using MLPs. In *International Joint Conference on Neural Networks*.

Arango-Argoty, G. et al., 2018. DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*.

Bolduc, B. et al., 2017. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect *Archaea* and *Bacteria*. *PeerJ*.

Enright, A.J., 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7), pp.1575–1584.

Fawcett, T., 2006. An introduction to ROC analysis Tom. *Pattern Recognition Letters*, (27), pp.861–874.

Feng, P.M. et al., 2013. Naïve bayes classifier with feature selection to identify phage virion proteins. *Computational and Mathematical Methods in Medicine*.

Finn, R.D., Clements, J. & Eddy, S.R., 2011. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, 39(SUPPL. 2), pp.29–37.

Grazziotin, A.L., Koonin, E. V & Kristensen, D.M., 2017. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. , 45(October 2016), pp.491–498.

Hans-W Ackermann, 2011. Bacteriophage Taxonomy. *Microbiology Australia*, 32(2), pp.90–94.

Lefkowitz, E.J. et al., 2017. *Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2017)*,

Li, W. & Godzik, A., 2006. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13), pp.1658–1659.

Lima-Mendez, G. et al., 2008. Reticulate representation of evolutionary and functional relationships between phage genomes. *Molecular Biology and Evolution*.

Lopes, A. et al., 2014. Automated classification of tailed bacteriophages according to their neck organization. *BMC Genomics*, 15(1), pp.1–17.

Loshchilov, I. & Hutter, F., 2016. SGDR: Stochastic Gradient Descent with Warm Restarts.

Manavalan, B., Lee, J. & Lee, J., 2014. Random forest-based protein model quality assessment (RFMQA) using structural features and potential energy terms. *PLoS ONE*.

Manavalan, B., Shin, T.H. & Lee, G., 2018. DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget*.

Manavalan, B., Shin, T.H. & Lee, G., 2018. PVP-SVM: Sequence-based prediction of phage virion proteins using a support vector machine. *Frontiers in Microbiology*.

Meier-kolthoff, J.P. & Go, M., 2018. Phylogenetics VICTOR□: genome-based phylogeny and classification of prokaryotic viruses. , 33(July 2017), pp.3396–3404.

Meier-Kolthoff, J.P. & Göker, M., 2017. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics (Oxford, England)*, 33(21), pp.3396–3404.

Morota, G. et al., 2018. BIG DATA ANALYTICS AND PRECISION ANIMAL AGRICULTURE SYMPOSIUM: Machine learning and data mining advance predictive big data analysis in precision animal agriculture1. *Journal of Animal Science*, 96(4), pp.1540–1550. Available at: https://academic.oup.com/jas/article/96/4/1540/4828311.

Paez-Espino, D. et al., 2016. Uncovering Earth's virome. *Nature*.

Reyes, A. et al., 2017. Use of profile hidden Markov models in viral discovery: current insights. *Advances in Genomics and Genetics*, Volume 7(July), pp.29–45. Available at: https://www.dovepress.com/use-of-profile-hidden-markov-models-in-viral-discovery-current-insight-peer-reviewed-article-AGG.

Reyes, A. & Gruber, A., 2016. GenSeed-HMM□: A Tool for Progressive Assembly Using Profile HMMs as Seeds and its Application in Alpavirinae Viral Discovery from Metagenomic Data. , 7(March), pp.1–15.

Rohwer, F. & Edwards, R., 2002. The phage proteomic tree: A genome-based taxonomy for phage. *Journal of Bacteriology*, 184(16), pp.4529–4535.

Roux, S. et al., 2016. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, 537(7622), pp.689–693. Available at: http://dx.doi.org/10.1038/nature19366.

Roux, S. et al., 2015. VirSorter: mining viral signal from microbial genomic data. *PeerJ*.

Schmidhuber, J., 2015. Deep learning – An overview. *International Journal of Applied Engineering Research*.

Seguritan, V. et al., 2012. Artificial Neural Networks Trained to Detect Viral and Phage Structural Proteins. *PLoS Computational Biology*.

Shapiro, J.W. & Putonti, C., 2018. Gene co-occurrence networks reflect bacteriophage ecology and evolution. *mBio*.

Skewes-cox, P. et al., 2014. Profile Hidden Markov Models for the Detection of Viruses within Metagenomic Sequence Data. , 9(8).

Tan, J.X. et al., 2018. Identifying phage virion proteins by using two-step feature selection methods. *Molecules*, 23(8), pp.1–13.

Funding:

KAAD for stipend, Department of Genomics and Applied Microbiology, Open access fund of DFG.

Availability of data and materials:

HMMs download available on http://appmibio.uni-goettingen.de/index.php?sec=sw

(To be made public once manuscript is accepted)

Competing interests

The authors declare that they have no competing interests.

Author's contributions

CC performed research, designed algorithm, performed data analysis, wrote manuscript, FM designed algorithm, wrote program, performed data analysis, AF wrote program to refine Markov Models, SD designed algorithm, HL designed research, analyzed data, wrote manuscript.

Acknowledgements

We thank Tarek Morsi and Marc Dornieden for excellent IT-support. We thank the Goettinge Genomics Laboratory G2L for hosting. We acknowledge the support by the German research Foundation and the Open Access Fund of the Goettingen University.

Consent for publication

Not applicable.

Supplementary information

Supplementary information for this manuscript can be found at the “bioRxiv” website under the following address:

https://www.biorxiv.org/content/10.1101/558171v1.supplementary-material

Additionally, supplementary figures and tables are provided along with the electronic version of this thesis (on DVD), under the following paths:

Additional Data:

SupplementaryMaterial/ChapterII/ChapterII.4/All_refined_HMMs.zip

Cytoviridae HMMs: Cystoviridae_refined_HMMs/

Fuselloviridae HMMs: Fuselloviridae_refined_HMMs/

Haloviruses HMMs: Haloviruses_refined_HMMs/

Inoviridae HMMs: Inoviridae_refined_HMMs/

Leviviridae HMMs: Leviviridae_refined_HMMs/

Ligamenvirales HMMs: Ligamenvirales_refined_HMMs/

Microviridae HMMs: Microviridae_refined_HMMs/

Myoviridae HMMs: HMMs/Myoviridae_refined_HMMs/

Pleolipoviridae HMMs: Pleolipoviridae_refined_HMMs/

Podoviridae HMMs: Podoviridae_refined_HMMs/

Siphoviridae HMMs: Siphoviridae_refined_HMMs/

Tectiviridae HMMs: Tectiviridae_refined_HMMs/

Additional Figures:

Figure S1: SupplementaryMaterial/ChapterII/ChapterII.4/Figure S1.tiff

Additional Tables:

Table S1: SupplementaryMaterial/ChapterII/ChapterII.4/TableS1.xlsx

Table S2: SupplementaryMaterial/ChapterII/ChapterII.4/TableS2.xlsx

Table S3: SupplementaryMaterial/ChapterII/ChapterII.4/TableS3.csv

Table S4: SupplementaryMaterial/ChapterII/ChapterII.4/TableS4.xlsx

Table S5: SupplementaryMaterial/ChapterII/ChapterII.4/TableS5.xlsx

Table S6: SupplementaryMaterial/ChapterII/ChapterII.4/TableS6.csv

Table S7: SupplementaryMaterial/ChapterII/ChapterII.4/TableS7.xlsx

Table S8: SupplementaryMaterial/ChapterII/ChapterII.4/TableS8.xlsx

Table S9: SupplementaryMaterial/ChapterII/ChapterII.4/TableS9.xlsx

Scripts:

SupplementaryMaterial/ChapterII/ChapterII.4/Scripts/

SupplementaryMaterial/ChapterII/ChapterII.4/Scripts/ReadMe.txt

SupplementaryMaterial/ChapterII/ChapterII.4/Scripts/pipeline_overview.png

II.5 IdentiPhage: Integrated Phages Identification using DNN

IdentiPhage: Integrated Phage Identification using DNN

Cynthia Maria Chibani, David Pawlowicz, Sascha Dietrich, Heiko Liesegang

Authors' contributions

CC performed research, designed algorithm, performed data analysis, wrote the manuscript,

DP, designed algorithm, wrote the program, performed data analysis,

SD designed algorithm,

HL designed research, analyzed data, designed algorithm, wrote the manuscript.

IdentiPhage: Integrated Phage Identification using DNN

Cynthia Maria Chibani, David Pawlowicz, Sascha Dietrich, Heiko Liesegang

Institute for Microbiology and Genetics, Georg-August University Goettingen, Grisebachstr. 8, 37077, Goettingen, Germany

*** Correspondence:**

Corresponding Author

hlieseg@gwdg.de

Abstract

Background:

Prophages are known to have a tremendous impact on their bacterial host. However, accurately identifying integrated prophage regions within bacterial genomes remains a problem. The majority of existing tools rely on hits to known phage sequences, which limits the identification of distantly related prophage regions.

Results:

In this study, we present IdentiPhage, a method for the prediction of integrated bacteriophage sequences within bacterial genomes. IdentiPhage uses a deep neural network machine learning approach. We trained IdentiPhage on a set of genomic features generated from a dataset of 11,373 bacterial and 8,721 phage genomes. We show that features such as GC%, GC content deviation, dinucleotide skew, number of CDS per window, overlapping CDS per window and an average gene size were sufficient to locate integrated prophages. These features canidentify prophages without any sequence similarities to known phages. Our positive phage label, for the supervised machine learning approach, was a BLAST hit of over 1 kb of the phage sequence to the bacterial sequence database. IdentiPhage achieved a specificity of 80.14 % during hold-out cross-validation. We compared the performance of IdentiPhage to PHASTER and phiSpy which are popular tools used for phage prophage identification.

Conclusions:

Our results show that IdentiPhage can be used as a complementary tool to existing tools. In a simple test with real data, where prophages in 9 *Vibrio alginolyticus* genomes were experimentally confirmed, IdentiPhage identified all known prophages. IdentiPhageis a tool under active development, to be made available as a publicly accessible easy-to-use web service.

Keywords:

Prophage, integrated phages, genomic features, machine learning, artificial neural networks

1. Introduction

Bacteriophages, viruses infecting bacteria, are estimated to be one of the most biological entities on earth (Amgarten et al. 2018; Arndt et al. 2017; Jurtz et al. 2016; Roux et al. 2014). They are recognized as the major driving forces of i) virulence of facultative pathogens (Roux et al. 2016; Roux et al. 2015; Busby et al. 2013), ii) microbial evolution and adaptation to new ecological niches (Arndt et al. 2017; Howard-Varona et al. 2017), and iii) marine carbon and nutrient cycling, such as nitrogen, phosphate and sulfur (Howard-Varona et al. 2017; Jurtz et al. 2016; Roux et al. 2016; Roux et al. 2015).

Bacteriophages are known to, either use the replication machinery of the host for replication and lyse the host, and thus have a lytic life cycle, or can integrate into the host genomes and replicate with the replicating host and therefore has a temperate lifestyle (Howard-Varona et al. 2017; Wendling et al. 2017; Akhter et al. 2012;Zhou et al. 2011). The latter, termed prophages, were identified in over 50% of bacterial genomes (Touchon et al. 2016), and moreover, bacterial genomes can contain over 20% of prophages and cryptic prophages (Arndt et al. 2017; Casjens 2003). The computational identification of prophages still poses a challenge due to the extensive genetic exchange between phages and their hosts, which increases the complexity of phage identification (Hurwitz et al. 2018). Multiple tools exist for the identification of integrated prophages within bacterial genomes. Tools such as PHAST (Zhou et al. 2011), PHASTER (Arndt et al. 2016), PHASTEST (Arndt et al. 2017), Phage_Finder(Fouts 2006) and Prophinder (Lima-Mendez et al. 2008)are based on annotated

genes and codingsequencesorsimilaritiestoknownreferencephagegenomes. The downside is that identification by sequence comparison to a database of known phages limits the possibility of identifying new phages to those similar to the phages within those databases (Zhao et al. 2017).

Moreover, prophage identification relying on phage specific annotations can be biased, depending on the software used to generate those annotations due to the high number of poorly or incorrectly annotated proteins (de Crécy-Lagard 2016), for instance, due to the fact that most automated annotation pipelines are not refined for the detection of small phage ORFs (Linial 2003). PhiSpy, a tool introduced in 2012, combines multiple phage sequence characteristics, including some non-similarity based features, thus increased the accuracy of prophage predictions (Hurwitz et al. 2018). Yet, it additionally relies on identifying viral genes based on homology to known viral genes that represent only a small portion of viral diversity. To this end, a plethora of tools are available and additionally being developed for mining viral sequences in large metagenomic datasets but are not suitable to identify prophages integrated within bacterial genomes (Hurwitz et al. 2018)

Here, we present IdentiPhage a tool for the prediction of integrated prophages within bacterial genomes. IdentiPhage applies a machine learning approach that evaluates 12 non-similarity based genomic features by applying a set of thresholds to select optimal parameters for the prediction. The use of machine learning algorithms have been successfully applied to several biological problems, such as the prediction of antibiotic resistance genes from metagenomic data (Arango-Argoty et al. 2018), the prediction of prokaryotic hosts from metagenomic phage contigs (Galiez et al. 2017), the taxonomic classification of phages (Chibani et al. 2019), the prediction of phage sequences in metagenomic bins (Amgarten et al. 2018), and the prediction of phage proteins (Manavalan, Tae H. Shin, et al. 2018; Ding et al. 2014; Feng et al. 2013). The resulting predicted regions were benchmarked against the prediction of popular prophage prediction tools phiSpy and PHASTER.

2. Material and Methods

We follow the 5 step guideline endorsed in a series of publications, for the development of a sequence-based predictor for a biological system (Manavalan, Tae H. Shin, et al. 2018 (a); Manavalan, Tae Hwan Shin, et al. 2018 (b); Manavalan et al. 2017).

Training and Testing Datasets

To build and test IdentiPhage, bacterial GenBank files were downloaded from NCBI on 04 June 2018. The accession numbers of the phage dataset used in this research were retrieved from the millardlab database (http://millardlab.org/bioinformatics/bacteriophage-genomes/) and downloaded from NCBI. As of 20 March 2018, the database contained 8,721 number of phage genomes (Chibani et al. 2019). A blastn search was done by aligning phage nucleotide sequences against nucleotide sequences of the downloaded bacterial genomes. The BLASTn results were used to extract positive sample data from the bacterial genomes. Every hit with a size of ≥1 Kbpwas used as the resulting range of positive phage samples.

Feature Extraction

Seeking robust phage descriptive features, we computed 12 prophage descriptivesequence-derived features. Feature data were generated per window, in a sliding window approach, for a specific window size on the bacterial host replicons. Window size was arbitrarily set to 500 bp (which is approximately half of the size of an average gene). Shifts per nucleotide usage, reflected in local changes or distortion in the cumulative skew distribution could be a result of the integration of foreign DNA (Akhter et al. 2012). Thus we considered GC%, GC skews (i), AT skews (ii) and GC content deviation (iii) which were calculated as follow:

$$\text{(i) } GCskew = (G-C)/(G+C)$$

$$\text{(ii) } ATskew = (A-T)/(A+T)$$

$$\text{(iii) } GCdeviation = GCwindow/GCreplicon$$

$$\text{(iv) } Slope = \Delta value/\Delta position$$

Additionally, we computed slope values (iv) for dinucleotide skews and the GC% deviation and GC-deviation switch to indicate a sudden deviation of the GC% deviation from below 1 to above 1 or vice versa. This switch was used to mark hotspots by searching for a steep GC%-deviation slope (the highest 3% slopes) within 300 bp.

Furthermore, phages are known to encode shorter genes (Akhter et al. 2012), thus we considered features such gene density (Amgarten et al. 2018), the average gene length per window (Akhter et al. 2012; Amgarten et al. 2018), and overlapping CDSs (Brandes & Linial 2016). Gene density per window was calculated as the total number of CDS per window divided by window length measured in bp. The average gene length per window was calculated by adding up the length of all predicted CDSs in a window divided by the total number predicted CDS per window. All estimated values were generated based on the predicted CDSs which were extracted from the downloaded bacterial GenBank files and were calculated using a window size of 500 bp, and an overlap of 400 bp. The generated feature data files were processed, adding positive phage information as a label where 0 stands for no phage hit per window or 1 for phage hit per window. Note that all the above-mentioned features were normalized to the range of [0,1] as input for the DNN.

Data Normalization

To be able to use the data set in machine learning algorithms, its data has been normalized using the StandardScaler from sklearn.preprocessing, normalizing all data to values between -1 and 1. The normalized data set was, as well as the StandardScaler -object, dumped to a binary file for faster access using the pickle module, calling pickle.dump. Due to the vast imbalance of positive and negative samples (5,501,976 of 349,024,443 samples were positive), a smaller subset was created using a 50/50-ratio for positive and negative samples, collected randomly from the entire data set. This dataset contained ~5 million positive and 5 million negative samples.

Classifier Development

Using Python Scikit Learn Libraries and the Keras module configured to use TensorFlow as its back-end; we developed and trained a Deep Neural Network (DNN). We experimented with various DNN architectures as follow: a first hidden dense layer with twice the nodes than the input dimension as an inputlayer and three additional hidden dense layers with triple the nodes than the input dimension. After every hidden layer, starting with the second one, a "Dropout layer" was used, set to 25% dropout rate which is capable of better generalization of the model and avoiding overfitting (Srivastava 2014). Lastly, the output layer of the deep neural network consists of 2 units that correspond to whether a window belongs to a phage region or not. The DNN uses a rectified linear unit (relu) activation function (Arora et al. 2016) that computes the probability of the input window sequence against one of the two possible outcomes; window belonging to phage or not. The training data was split into two subsets at an 80-20% division where we refer to 80% of the data as the training dataset and the 20% of the data as the validation dataset. The training dataset was used to train and generate a model for the prediction of phage windows while the testing dataset was used to provide an evaluation of the final fit IdentiPhage model what is referred to as the hold-out cross-validation method. Heavy computation is required only once to obtain the deep learning model, and the prediction routines do not need such computational resources.

Metrics Measurement

To assess IdentiPhage's performance and robustness, we repeat the process of random selection of the training and testing datasets, model-building and model-evaluating using 3 parameters: overall prediction v) sensitivity (Sn), vi) specificity (Sp) and vii) accuracy (Acc). These measured metrics would help us determine how the model would perform on new datasets(Pan et al. 2018). The parameters are defined as follows:

$$\text{(v) } Sn = TP/(TP+FN)$$

$$0<Sn<1$$

$$\text{(vi) } Sp = TN/(TN+FP)$$

$$0<Sp<1$$

$$\text{(vii) } Acc = (TP+TN)/(TP+FP+TN+FN)$$

$$0<Acc<1$$

Where TP (true positives) is the number of predicted phage windows; TN (true negatives) is the number of nonphage predicted windows; FP (false positives) is the number of bacterial window sequences predicted as phage windows, and FN (false negatives) is the number of phage windows predicted as bacterial window sequences. In our experiment, the Sn is the proportion of bacteriophage sequences that were correctly identified. The Sp measures the proportion of non-bacteriophage sequences that were correctly identified. The Acc is the proportion of true results (the percentage of correctly identified bacteriophage sequences and non-bacteriophage sequences) among the total number of samples.

To further evaluate the performance of IdentiPhage and to determine suitable thresholds for the prediction values of the different windows, we generated receiver operating curves (ROC), where we plotted the FP rate as the x-axis and the TP rate as the y-axis. ROC curves depict the tradeoff between sensitivity and specificity (any increase in sensitivity is coupled with a decrease in specificity) (Pan et al. 2018). The area under the curve (AUC), which is a measure of discrimination, was used for IdentiPhages evaluation, with higher AUC values corresponding to the better performance of the model. The value of AUC score ranges from 0 to 1, with a score 0.5 corresponding to a random guess and a score of 1.0 indicating a perfect separation. The AUC is a measure of the ability of the model to correctly classify a sequence window into belonging to a bacteriophage or not. Based on determining a model's final predictions a confusion matrix of the true and predicted phage labels with the number of TP was created and plotted.

Pipeline Implementation

IdentiPhage was coded in the Python 3 programming language in version 3.5.3, using the scientific Python distribution Anaconda. As input, IdentiPhage requires a directory with GenBank or fasta files. It generates a result directory with an extracted multi-CDS file per predicted phage region.

Prophage prediction

Once sequences are classified, the maximum distance between two positively predicted positions is calculated to evaluate the size of the predicted region. The minimum distance is measured as well to validate the range, and if it is smaller than the smallest considered phage genome, then it is dropped (*Campylobacter* phage C10, GenBank accession number MG065651.1, size 1.4kb).

Test with *Vibrio alginolyticus* independent dataset

An external dataset of 9 sequenced *Vibrio alginolyticus* genomes, where the location of active integrated prophages was experimentally proven (Wendling et al. 2017), was used to test IdentiPhage's performance. The feature matrix of the *V. alginolyticus* genomes was generated and normalized as previously described. The matrixwas used as an independent dataset input for the model to classify sequences whether they are of a prophage or not. The predicted phage regions were compared to the positions of the known integrated prophages.

Performance Comparison of IdentiPhage to other tools

The same set of *Vibrio alginolyticus* genomes, used as an independent dataset for IdentiPhage, was used as input for phiSpy and PHASTER. PhiSpy was benchmarked against phage_finder and prophinder, and the authors proved that phiSpy outperformed the mentioned tools (Akhter et al. 2012) and thus were not considered in this analysis. The average values of the true positive rate of the tools were compared based on the tools abilities to predict the known phages correctly.

3. Results

Framework of the Proposed Predictor

Firstly, we constructed the benchmark dataset; then we extracted the various described features from the primary sequences, including GC%, GC%-deviation, dinucleotide skew, gene density, overlapping CDS and the average number of CDS per the size of a specified window in bp. These different features were used as an input for a DNN to develop a prediction model. The described metrics were evaluated for the generated model. Finally, IdentiPhage was tested with

an external dataset, and its performance was benchmarked against two popular phage prediction tools.

Data Construction and Input Features Variance

To show the percentage of variances explained by each principal component, the eigenvalues were computed and ordered from the largest to smallest to generate a scree plot (Figure 1). Eigenvalues (Table S1) are used to determine the number of principal components, which show an interesting pattern in the data, to keep after PCA (principal component analysis) (Figure S1). In our study, the first three principal components explain 38 % of the variation which is a largely acceptable percentage.

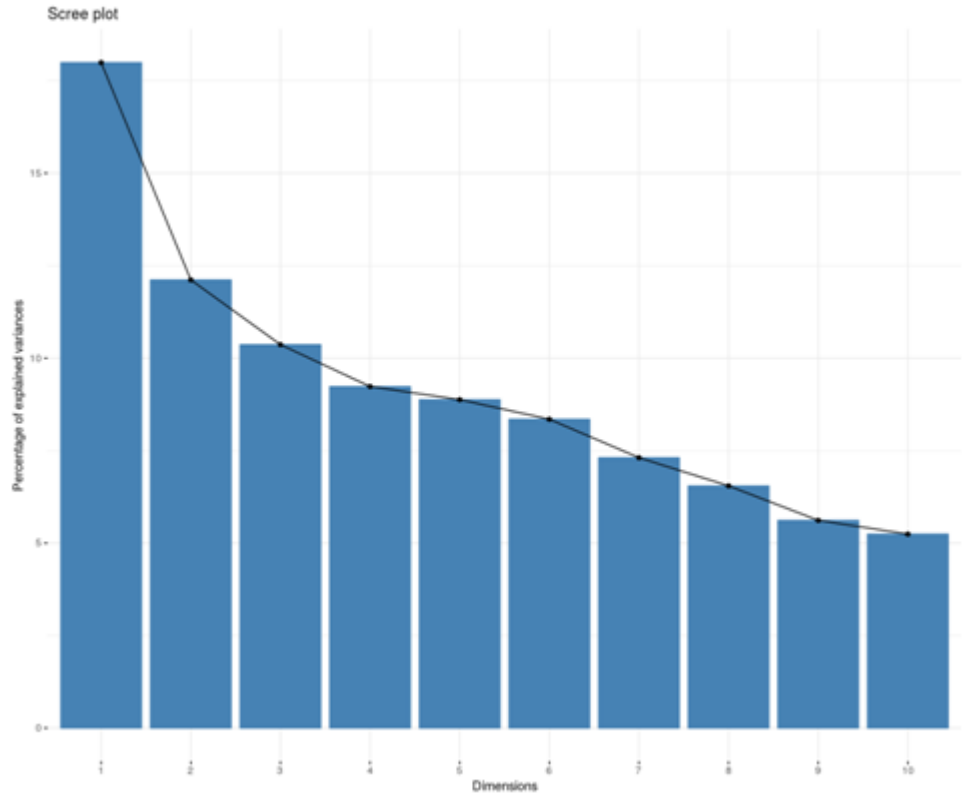


Figure 1: Scree plot of the total variance associated with each input factor.

Scree plot of the normalized input dataset.The x-axis shows the number of principal components. The scree plot showed that the first three principal components explained the maximum variation (38 %) in the dataset.

Subsequently, Principal Component Analysis (PCA) and correlation plots were performed to highlight the most contributing variables for each dimension and to determine the most informative features in the generated dataset. The analysiswas first created for all input features (Figure S2, Figure S4), and then for the 3 most descriptiveelements (Figure S1, Figure S3, and Figure S5).

Model performance and Metrics

The DNN was tested by using either 3 or 4 hidden layers, either 12 or 24 number of nodes in the first layer and either 36 or 48 as the number of nodes in the hidden layers. The main output of the DNN is the label of the testing set and predictions of the model for each entry recorded at any training epoch. Using this information, the performance of IdentiPhage can be assessed in detail for different stages of training. The labels of testing data are compared to the model's assignments of the last recorded prediction, by taking the maximum value of the model'sassignments. Here we present the metrics measured for IdentiPhage (Table 1), where the Acc of the selected model was 72.88 %.

Table 1: Sensitivity and specificity of the different architectures of the DNNs

Hidden Layers	1st node	2nd node	Sensitivity	Specificity
3	12	36	94.52	37.71
3	12	48	71.16	74.19
3	24	36	71.97	75.49
3	**24**	**48**	**80.14**	**71.05**
4	12	36	94.17	39.12

4	12	48	70.57	73.43
4	24	36	68.52	74.8
4	24	48	70.95	75.97

*The values marked in bold were further represented in a confusion matrix and ROC curve.

The true positive and false positive rate on the test data at different thresholds for the classifiers using the top 12 features are displayed as a confusion matrix and as a ROC curve (Figure 2).

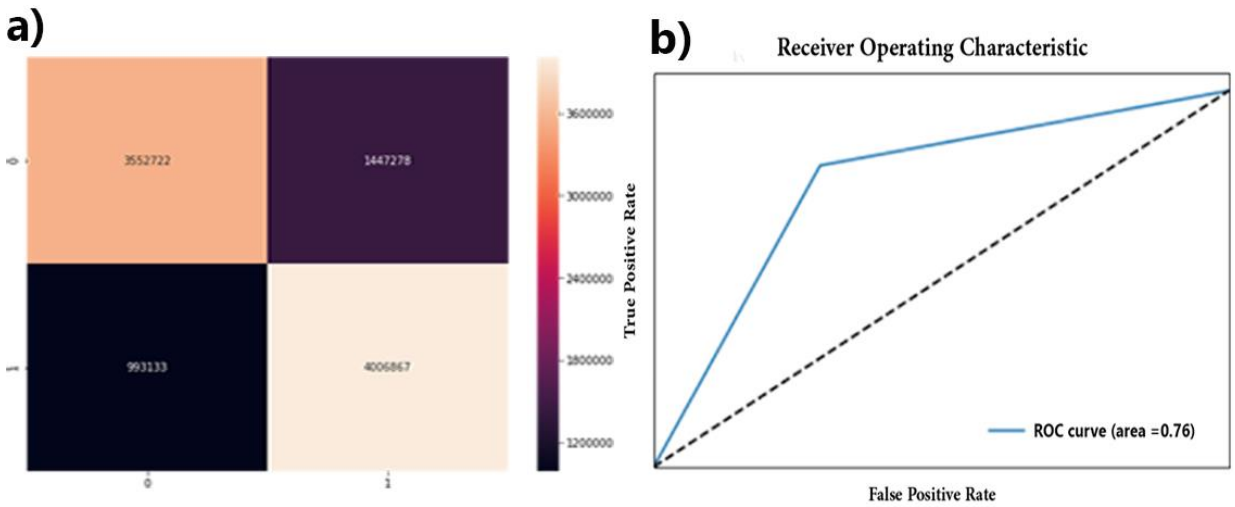


Figure 2: Confusion matrix a) and ROC curve (b) for predicted prophage regions.

The confusion matrix (a) shows the true value on the y-axis, meaning the samples which were 0 are in the upper, and the ones which were 1 are in the lower half of the matrix. Resulting values are from left to right, top to bottom: true negatives (3,552,722), false positives (1,447,278), false negatives (993,133) and true positives (4,006,867).

The ROC curve (b) for the prediction of phage region. The diagonal dot line denotes a random guess with the auROC of 0.5. An AUC of 0.76 was obtained in a hold-out cross-validation test.

Independent dataset testing

We evaluated the performance of IdentiPhage using an independent dataset. 15 out of 16 prophage regions were hit at least partially (Table 2). Generally, all hits had a delayed start and always hit the end of the phage regions with a precision within 1Kbp.

Table 2: Coverage of phage hits in *V. alginolyticus*

V. alginolyticus	Phage Region		Coverage* [bp]			Relative Coverage [%]	
	Start	Stop	Fragment	Joined	Size	Fragment	Joined
K01M1_2	965,930	975,241	2,500	7,200	9,311	26.85	77.33
K04M1_2	945,046	953,135	1,035	6,235	8,089	12.80	77.08
K04M3_2	945,065	969,721	5,400	23,300	24,656	21.90	94.50
	979,135	994,596	7,000	11,600	15,461	45.28	75.03
K04M5_2	975,221	991,351	4,851	14,521	16,130	30.07	88.35
K05K4_2	957,739	959,718	0	0	1,979	0.00	0.00
	964,327	965,822	222	222	1,495	14.85	14.85
K06K5_2	945,036	953,638	1,532	6,738	8,602	17.88	78.33
K08M3_2	945,048	953,650	1,550	6,750	8,602	18.02	78.47

K09K1_1	1	22,357	6,700	19,000	22,356	29.97	84.99
	1,876,616	1,897,209	9,100	17,300	20,593	44.19	84.01
K09K1_2	1	11,467	5,100	9,800	11,466	44.48	85.47
	3,460,762	3,465,618	700	100	4,856	14.41	14.41
K10K4_1	1,701,803	1,709,837	1,937	5,337	8,034	24.11	66.43
K10K4_2	945,030	953,121	1,521	6,211	8,091	18.81	76.89
	977,741	985,316	2,200	2,200	7,575	29.04	29.04

*Coverage is defined as the number of phage bp identified out of the known phage regions.

Shown in this table are the *V. alginolyticus* strains, starting and stopping position of known phage regions, and the coverage predicted for phage regions. For the values in the fragmented columns, only the regions which were explicitly predicted were used, whereas, for the ’Joined‘ values, the entire region spans were considered. The joined coverage averages to 64.07%.

Comparison with other methods

The comparison was made by comparing the prediction power of the different considered tools, PHASTER, and phiSpy, of active *Inoviridae* prophages, integrated into 9 *V. alginolyticus* genomes. The inactive phages were not considered for this analysis since the number, and the correct boundaries of those phages can’t be quantified.

PHASTER prediction using the independent dataset

PHASTER identified 15 of 16 (93.75) known prophages. Generally, PHASTER determined boundaries were either greater or smaller than the curated prophage boundaries. PHASTER

was able to predict the prophage IdentiPhage missed, while on the other hand one prophage was missed entirely (Table 3).

Table 3: PHASTER prophage predictions of the independent dataset

V. alginolyticus	Phage Region		PHASTER		Relative coverage	Additional Predicted Phages
	Start [bp]	Stop [bp]	Start [bp]	Stop [bp]	[%]	
K01M1_2	965,930	975,241	954,239	974,702	219.77	1
K04M1_2	945,046	953,135	933,351	953,814	252.97	1
K04M3_2	945,065	969,721	933,352	969,913	148.28	1
	979,135	994,596	982,291	994,621	79.75	
K04M5_2	975,221	991,351	963,526	991,566	173.84	1
K05K4_2	957,739	959,718	946,048	966,511	1,034.01	1
	964,327	965,822	-	-		
K06K5_2	945,036	953,638	933,345	953,808	237.89	1
K08M3_2	945,048	953,650	658	11,534	126.44	1
K09K1_1	1	22,357	1,701,790	1,709,314	33.66	1
	1,876,616	1,897,209	933,339	953,802	99.37	

K09K1_2	1	11,467	977,819	985,329	65.50	1
	3,460,762	3,465,618	3,461,560	3,465,606	83.32	
K10K4_1	1,701,803	1,709,837	1,701,790	1,709,314	93.65	1
K10K4_2	945,030	953,121	933,339	953,802	252.91	1
	977,741	985,316	977,819	985,329	99.14	

* The Last column shows the number of phage regions predicted by PHASTER in addition to the ones which were experimentally proven.

PhiSpy prediction using the independent dataset

PhiSpy identified 14 of 16 (87.5%) known prophages. PhiSpy predicted a surplus of ~ 20 kps upstream of every phage region, as well as additional ~ 2–30 Kbp downstream. It was able to identify the ≤2Kbp region that was missed by IdentiPhage. On the other hand, two areas were missed completely (Table 4).

Table 4: PhiSpy prophage predictions of the independent dataset

V. alginolyticus	Phage Region		phiSpy		Relative coverage	Additional Predicted Phages
	Start [bp]	Stop [bp]	Start [bp]	Stop [bp]	[%]	
K01M1_2	965,930	975,241	946,652	976,544	321.0	2
K04M1_2	945,046	953,135	926,071	955,131	359.3	2
K04M3_2	945,065	969,721	931,711	995,530	258.8	1

	979,135	994,596	931,711	995,530	412.8	
K04M5_2	975,221	991,351	955,939	986,758	191.1	2
K05K4_2	957,739	959,718	938,768	967,629	1458.4	1
	964,327	965,822	938,768	967,629	1930.5	
K06K5_2	945,036	953,638	925,758	955,125	341.4	2
K08M3_2	945,048	953,650	937,804	954,938	199.2	3
K09K1_1	1	22,357	1	24,369	109.0	2
	1,876,616	1,897,209	1,857,653	1,897,209	192.1	
K09K1_2	1	11,467	1	18,754	163.6	9
	3,460,762	3,465,618	-	-		
K10K4_1	1,701,803	1,709,837	-	-		7
K10K4_2	945,030	953,121	926,859	982,436	686.9	1
	977,741	985,316	926,859	982,436	733.7	

* The Last column shows the number of phage regions predicted by PHASTER in addition to the ones which were experimentally proven.

These results indicate that prophage prediction is far from adequate and no tool can precisely prophages. Additionally, we can state that IdentiPhage can play a complementary role to existing tools for prophage prediction.

4. Discussion

In this project, we evaluated 12 genomic features concerning their usability for prophage detection within host genomes. In addition to the features used in phiSpy (GC%, GC skews, AT skews, GC content deviation) (Akhter et al. 2012) we used gene density, average gene length, and overlapping CDSs, which proved to be extremely informative for prophage prediction. To investigate which combinations of individually weighted features perform best we designed and applied a deep neural network (DNN).

For IdentiPhage, we considered a much more elaborate benchmark dataset compared to phiSpy. The features used, were the GC%, GC skews, AT skews, GC content deviation and additional dependent features, which were shown to work better for related organisms and organisms with extreme AT and GC deviations (Akhter et al. 2012). Additionally, we used gene density, average gene length, and overlapping CDSs, which proved to be extremely informative for prophage prediction. We considered a window size of 500 bp which might have affected feature calculation, while in phiSpy a window of 40 genes was considered. We computed gene density and average gene length per window size, contrary to phiSpy where those features were calculated by replicon size. The PCA analysis showed that the three most important featureswere i) slopes, ii) hotspots size and iii) GCD switch. Thus for the future development of IdentiPhage, we will consider replacing the insignificant features by descriptive features such as the transcription strand directionality, the average spacing between genes, the median of all protein lengths. Using an input feature such as HMM hits, could limit the identification of new phages; thus we will include HMM hits in a secondary step, after initial identification of genomic hallmarks. We additionally will consider adjusting the window size and the computation of the gene density per replicon.

We briefly mention PCA, however, we proceeded by using the whole dataset as input for the DNN since PCA is known to alter the original representation of the variables (Pan et al. 2018).

Thus for the future development of IdentiPhage, we will consider a less expensive approach to dimensionality reduction for feature selection. Feature selection processes proved effective in reducing the dimensionality of the data and improving the performance of the predictors (Manavalan, Tae H. Shin, et al. 2018 (a); Manavalan, Tae Hwan Shin, et al. 2018 (b); Pan et al. 2018; Manavalan et al. 2017; Feng et al. 2013). This is an important step to exclude redundant, irrelevant and noisy information found in high dimensional features, and thus to find a minimum set of features that achieve maximum classification performance.

Random Forest algorithms were used in MARVEL, where they selected three features as the most informative: gene density, strand shifts, and significant matches to the pVOG database. In other methods, Analysis of variance (ANOVA) was used as a feature selection process (Pan et al. 2018; Ding et al. 2014), while Tan et al. (2018) used Minimal-Redundancy-Maximal-Relevance (mRMR) in addition to ANOVA as the second step in their process. ANOVA method calculates the variance among groups and thus gives a clear understanding of each feature capabilities for the model; while mRMR filters out the most informative features to minimize information redundancy and gather the most concise feature subset with no loss of useful information(Tan et al. 2018). Feature redundancy may be an issue since some of the considered descriptors may be derived from each other, for instance, GC% and GC% deviation and the corresponding slopes. Thus, we will consider exploring different feature selection methods to achieve a maximum variance with minimal redundancy in our input dataset.

Generally, it is crucial to explore various ML-methods on the same dataset, and then to select the best method, since ML-based predictors are problem-specific (Amgarten et al. 2018). Moreover, IdentiPhage uses the hold-out cross-validation method to evaluate the predictive ability of our predictor. However, the K-fold cross-validation method and the jackknife test are more rigorous (Amgarten et al. 2018; Tan et al. 2018). The five-fold cross-validation is widely used by scientiststo save computation time (Tan et al. 2018). Thus for the future development of IdentiPhage, we will consider exploring Support Vector Machines (SVM) and RF algorithms, as well as the different cross-validation tests.

Lastly, the prediction model tends to over-optimize to attain higher accuracy. Therefore, it is always necessary to evaluate the prediction model using an independent dataset, to evaluate the generalizability and the transferability of the method (Manavalan, Tae Hwan Shin, et al. 2018; Manavalan et al. 2017). Hence, we evaluated our prediction model on an independent dataset, which harbors a manually curated set of prophages (Wendling et al. 2017), and benchmarked the prediction of prophage prediction popular tools against this dataset. Our study demonstrated that IdentiPhage can play a complementary role for the prediction of prophages overseen by the existing tools.

To support the scientific community, we are working on a user-friendly web interface is to be made available to allow researchers access to the prediction method on our servers. The IdentiPhage method represents a powerful and cost-effective approach for prophage prediction suitable for high throughput analysis of genomic data. Therefore, IdentiPhage might be useful for prophage prediction, facilitating hypothesis-driven experimental design.

5. Conclusion

In this study, we introduced a novel method which we call IdentiPhage. The method predicts prophage regions in bacterial genomes using a combined set of 12 sequence-derived features. The results indicated that IdentiPhage could be applied to predict prophages on high quality closed genomes. While these results are promising on well-characterized prophage classes, it has proven challenging to choose the best features for accurate prediction. It is shown that an effective generic viral prediction pipeline using the 12 investigated features in this study can be hard to achieve. However, given the heterogeneity of viral types and genome structures, we believe that we can elevate the performance of the method on so far unknown phages by integrating additional descriptive features. We intend to expand IdentiPhage's scope to include additional phage specific features, feature selection protocol and to test additional ML algorithms; the program was designed with this objective in mind. Furthermore, the spectrum of potential applications of this approach is a general one and doesn't have to be limited for prophage identification, rather could be applied to many other classification problems in bioinformatics. This is a tool under active development to be made available as a publicly

accessible easy-to-use we service, and we envisage its growing application on a variety of forthcoming projects.

6. References

Akhter, S., Aziz, R.K. & Edwards, R.A., 2012. PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. Nucleic Acids Research, 40(16), pp.1–13.

Amgarten, D. et al., 2018. MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. Frontiers in Genetics.

Arango-Argoty, G. et al., 2018. DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome.

Arndt, D. et al., 2017. PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes. Briefings in Bioinformatics, (May), pp.1–8. Available at: http://academic.oup.com/bib/article/doi/10.1093/bib/bbx121/4222653/PHAST-PHASTER-and-PHASTEST-Tools-for-finding.

Arndt, D. et al., 2016. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic acids research, 44(W1), pp.W16–W21.

Arora, R. et al., 2016. Understanding Deep Neural Networks with Rectified Linear Units. , pp.1–17. Available at: http://arxiv.org/abs/1611.01491.

Brandes, N. & Linial, M., 2016. Gene overlapping and size constraints in the viral world. Biology Direct, 11(1), pp.1–15. Available at: http://dx.doi.org/10.1186/s13062-016-0128-3.

Busby, B., Kristensen, D.M. & Koonin, E. V., 2013. Contribution of phage-derived genomic islands to the virulence of facultative bacterial pathogens. Environmental Microbiology, 15(2), pp.307–312.

Casjens, S., 2003. MicroReview Prophages and bacterial genomics : what have we learned so far ? , 49, pp.277–300.

Chibani, C.M. et al., 2019. ClassiPhages 2.0: Sequence-based classification of phages using Artificial Neural Networks. Available at: http://dx.doi.org/10.1101/558171.

de Crécy-Lagard, V., 2016. Quality Annotations, a Key Frontier in the Microbial Sciences. Microbe Magazine, 11(7), pp.303–310.

Ding, H. et al., 2014. Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. Molecular BioSystems, 10(8), pp.2229–2235.

Feng, P.M. et al., 2013. Naïve bayes classifier with feature selection to identify phage virion proteins. Computational and Mathematical Methods in Medicine.

Fouts, D.E., 2006. Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Research, 34(20), pp.5839–5851.

Galiez, C. et al., 2017. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs.

Bioinformatics, 33(19), pp.3113–3114.

Howard-varona, C. et al., 2017. Lysogeny in nature : mechanisms , impact and ecology of temperate phages. Nature Publishing Group, 11(7), pp.1511–1520. Available at: http://dx.doi.org/10.1038/ismej.2017.16.

Hurwitz, B.L., Ren, J.M.U. & Youens-clark, K., 2018. Computational prospecting the great viral unknown. , (January 2016), pp.1–12.

Jurtz, V.I. et al., 2016. MetaPhinder — Identifying Bacteriophage Sequences in Metagenomic Data Sets. , pp.1–14.

Lima-Mendez, G. et al., 2008. Prophinder: A computational tool for prophage prediction in prokaryotic genomes. Bioinformatics.

Linial, M., 2003. How incorrect annotations evolve - The case of short ORFs. Trends in Biotechnology, 21(7), pp.298–300.

Manavalan, B. et al., 2017. MLACP: machine-learning-based prediction of anticancer peptides, Available at: www.impactjournals.com/oncotarget.

Manavalan, B., Shin, T.H. & Lee, G., 2018. DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. Oncotarget.

Manavalan, B., Shin, T.H. & Lee, G., 2018. PVP-SVM: Sequence-based prediction of phage virion proteins using a support vector machine. Frontiers in Microbiology.

Pan, Y. et al., 2018. Identification of bacteriophage virion proteins using multinomial Naïve bayes with g-gap feature tree. International Journal of Molecular Sciences.

Roux, S. et al., 2016. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature, 537(7622), pp.689–693. Available at: http://dx.doi.org/10.1038/nature19366.

Roux, S. et al., 2014. Metavir 2: New tools for viral metagenome comparison and assembled virome analysis. BMC Bioinformatics, 15(1).

Roux, S. et al., 2015. VirSorter : mining viral signal from microbial genomic data. , pp.1–20.

Srivastava, N.G.H.A.K.I.S.S., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 15.1, pp.1929–1958.

Tan, J.X. et al., 2018. Identifying phage virion proteins by using two-step feature selection methods. Molecules, 23(8), pp.1–13.

Touchon, M., Bernheim, A. & Rocha, E.P.C., 2016. Genetic and life-history traits associated with the distribution of prophages in bacteria. ISME Journal, 10(11), pp.2744–2754. Available at: http://dx.doi.org/10.1038/ismej.2016.47.

Wendling, C.C. et al., 2017. Tripartite species interaction: eukaryotic hosts suffer more from phage susceptible than from phage resistant bacteria. BMC Evolutionary Biology, 17(1).

Zhao, G. et al., 2017. VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. Virology, 503(January), pp.21–30. Available at: http://dx.doi.org/10.1016/j.virol.2017.01.005.

Zhou, Y. et al., 2011. PHAST: A Fast Phage Search Tool. Nucleic Acids Research, 39(SUPPL. 2), pp.347–352.

Supplementary Material

Supplemental Figure S1: Scree plot of the total variance associated with the most relevant input factors.

Supplemental Figure S2: PCA analysis plot of the different input variables, based on their contribution. These results are color coded from blue (low-score) to orange (high-score).

Figure S3: PCA analysis plot of the three most relevant input variables, based on their contribution. These results are color coded from blue (low-score) to orange (high-score).

Supplemental Figure S4: Correlation plot highlighting the most contributing variables for each dimension.

Supplemental Figure S5: Correlation plot highlighting the three most contributing variables for each dimension.

Supplemental Table S1: Eigenvalues measuring the amount of variance retained by each principal component.

Supplemental Table S2: PHASTER all prophages prediction in an independent dataset.

Supplemental Table S3: phiSpy all prophages prediction in an independent dataset.

Supplementary information

Additionally, supplementary figures and tables are provided along with the electronic version of this thesis (on DVD), under the following paths:

Additional Figures:

Figure S1: SupplementaryMaterial/ChapterII/ChapterII.5/Figure S1.png

Figure S2: SupplementaryMaterial/ChapterII/ChapterII.5/Figure S2.png

Figure S3: SupplementaryMaterial/ChapterII/ChapterII.5/Figure S3.png

Figure S4: SupplementaryMaterial/ChapterII/ChapterII.5/Figure S4.png

Figure S5: SupplementaryMaterial/ChapterII/ChapterII.5/Figure S5.png

Additional Tables:

Table S1: SupplementaryMaterial/ChapterII/ChapterII.5/Table_S1.xlsx

Table S2: SupplementaryMaterial/ChapterII/ChapterII.5/Table_S2.xlsx

Table S3: SupplementaryMaterial/ChapterII/ChapterII.5/Table_S3.xlsx

Scripts:

SupplementaryMaterial/ChapterII/ChapterII.5/Scripts/

SupplementaryMaterial/ChapterII/ChapterII.5/Scripts/ReadME.txt

CHAPTER III: Discussion

III.1 General Discussion

The access to viral genomes at unprecedented rates, using continuously improving Next-Generation Sequencing (NGS) techniques, aggravates the existing gap of viral classification based on sequence information (Bolduc et al. 2017). This gap remains in consequence of the lengthy procedure required by the International Committee on Taxonomy of Viruses (ICTV) for the deposition of new viral genomes into their maintained database (Fauquet & Fargette 2005). Nonetheless, the advancements in NGS technologies permitted scientists to investigate the enormous number of uncharacterized viral sequences termed "Viral Dark Matter"(Reyes et al. 2017; Roux, Hallam, et al. 2015; Youle et al. 2012). The "Viral Dark Matter" extends beyond the three most common viral families available in public databases (Roux, Enault, et al. 2015) and often shares no similarities to known viral sequences (Roux, Hallam, et al. 2015). As an example, the *Alpavirinae* phage family was identified from the analysis of viral metagenomic datasets and was otherwise unidentifiable through classical phage culturing techniques (Alves et al. 2016).

For the modeling of viral sequence diversity, one proposed method of analysis is using Hidden Markov Models (HMM) of shared proteins of phage genomes. This proved to be successful in reconstructing distant homologs of the *Alpavirinae* phage family (Reyes et al. 2017; Aiewsakun et al. 2018). Thus HMMs can be used, to some extent, for the characterization of viral sequences within the "Viral Dark Matter" (Bolduc et al. 2017). However, this remains an iterative process, as the characterization of more diverse viral sequences remains instrumental in improving the sensitivity of HMMs (Reyes et al. 2017; Grazziotin et al. 2017).

An additional source of viral sequences can be retrieved by exploring microbial genomes, where it was estimated that over 62% harbor at least one prophage (Casjens 2003). Moreover, Roux et al. (Roux, Hallam, et al. 2015) used VirSorter(Roux, Enault, et al. 2015) to identify 12,498 viral sequences from 14,977 microbial and archaeal genomes. Prophages were predicted in novel bacterial hosts where further experiments are needed to confirm the identification of an entirely new viral order (Roux, Hallam, et al. 2015). Thus, Roux et al. (2015) further endorse the importance of exploring prophage diversity within sequenced bacterial genomes deposited in public databases.

The first aim of this Ph.D. thesis was to generate a scalable method for the taxonomic classification of phages based on their sequence information into the existing phage families defined by the ICTV. An in-depth analysis was performed for the classification of phages by creating robust protein profile HMMs for homology searches as seed (Chapter II.3). To address the limitation of current best-hit approaches, we evaluated the use of artificial neural networks (ANN) for phage multiclass classification based on various HMM hits combination (Chapter II.4).

The second aim of this thesis was to identify lysogens within bacterial host genomes. Therefore, we computed and rigorously assessed a set of descriptive sequence properties combined as features in a deep learning approach to allow an accurate prophage prediction (Chapter II.5).

Lastly, the valuable knowledge of prophage identification and classification using traditional microbiology and molecular biology methods compared to computational methods can't be disregarded. On that account, a set of phages infecting different *Vibrio alginolyticus* strains were experimentally characterized,and their integration sites were empirically verified (Chapter II.1andChapter II.2). These experiments provided valuable information for interpreting the resulting predictions out of the IdentiPhage and ClassiPhage approaches.

III.1.1 Experimentally verified *Inoviridae*

In this thesis, ten bacterial isolates, including nine strains of *Vibrio alginolyticus* as well as one strain of a new *Vibrio* species arbitrarily names *Vibrio typhli* were sequenced, assembled and analyzed with a specific focus on prophage content. The isolated strains resulted from a study initiated by Dr. Carolin Wendling and Dr. Olivia Roth (Wendling et al. 2017). The study is based on a tripartite interaction system where a phage infects a *Vibrio* host which in turn infects the pipefish *Syngnathus typhle* (Chapter II.1). To understand the dynamic relationship between phages and their hosts, prophages were induced from 75 *Vibrio* isolates, and cross-infection experiments were carried out. Thus a 75 x 75 phage-bacteria infection matrix was generated where bacteria were grouped according to their resistance to phage containing supernatant of the 75 isolates. Out of the three different phage resistant groups, 10 strains were chosen for downstream in-depth analysis. The induced phages out of the 10isolates were

sequenced and classified as *Inoviridae* based on phage morphology (TEM) and sequence similarity. In total, the ten isolates comprised a set of nineteen *Inoviridae* prophages which were used for an extensive analysis. Therefore, the prophages were compared to published *Inoviridae* vibriophages which revealed a highly specific genome organization of the phage family as is corroborated by Mai-Prochnow et al. (Mai-Prochnow et al. 2015) (Chapter II.2). Generally, *Inoviridae* are filamentous phages with a circular ssDNA genome. Their sizes usually range between 4 to 12 Kbp. Their genomes encode 10 core proteins which are grouped by functional units into a "Replication" unit, a "Structural" unit and an "Assembly and Secretion" unit (Mai-Prochnow et al. 2015). The nineteen identified *Inoviridae* phages unveiled suitable genome sizes and functional units (Chapter II.2). The advantages given by this study made it feasible to generate a highly reliable external dataset for evaluating the methods reported (Chapter II.3, Chapter II.4,andChapter II.5). On the one hand, the benefit of having mappable induced prophages (Hertel et al. 2015) onto reference *V. alginolyticus* replicons gave us the unique advantage of evaluating the identification prophages using i) HMMs as seed in the "ClassiPhage" method and ii) sequence derived characteristics in the "IdentiPhage" method. On the other hand, the benefit of having Transmission Electron Microscopy (TEM) images of the induced phages gave us the unique advantage to investigate the morphology of phages and classify them according to the ICTV scheme. Moreover, those results were used for evaluating the phage classification methods "ClassiPhage" and "ClassiPhage 2.0".

III.1.2 ClassiPhage and ClassiPhage 2.0

Phage classification is instrumental for inferring ecological and evolutionary relationships (Roux, Hallam, et al. 2015). As new virus genomes are expected to be sequenced, new challenges for taxonomy are expected to arise (Bolduc et al. 2017). Considerable efforts are being made to shift towards a comprehensive automated viral taxonomy (Bolduc et al. 2017; Roux, Enault, et al. 2015; Aiewsakun et al. 2018). Recently, the ICTV issued a consensus statement endorsing this shift which is a critical step given the growing number of metagenome-derived viral sequences (Simmonds et al. 2017).

Studies based on genome pairwise comparison (Rohwer & Edwards 2002) were extremely valuable, however, became widely unpopular since they failed to capture the diversity

represented by viral metagenomic datasets (Simmonds et al. 2017). VICTOR (Meier-Kolthoff & Göker 2017), a recently developed tool, fails in classifying environmental viruses that do not share any similar gene to known reference genomes (Jang et al. 2019). Due to growing evidence of high mosaicism in viral genomes, gene sharing networks were first introduced by Lima-Mendez et al. (2008) and later widely adopted. Gene sharing networks permitted phage classification without prior knowledge and were largely consistent with ICTV proposed taxa (Lima-Mendez et al. 2008; Iranzo, Krupovic, et al. 2016; Iranzo, Koonin, et al. 2016; Shapiro & Putonti 2018; Bolduc et al. 2017). Prophinder was generated using a monopartite-network based on 306 phages known at that time and showed a high accuracy of ~92% (Lima-Mendez et al. 2008). Thereafter, a bipartite-network approach was used to analyze the dsDNA virosphere and addressed viral subfamilies (Iranzo, Krupovic, et al. 2016) and further extended to analyze archaeal viruses (Iranzo, Koonin, et al. 2016). Both approaches allow the investigation of gene sharing across viral genomes. However bipartite-networks can be more accurate in comparison to monopartite-networks due to the additional knowledge from the representation of gene families and genomes (Iranzo, Krupovic, et al. 2016).

Bolduc et al. (2017) were able to generate viral clusters that are 75% consistent with ICTV taxa,however; the monopartite gene-sharing network-based method creates artifact clusters for undersampled genomes and for highly overlapping genomes (Bolduc et al. 2017). Thus an accurate approach that is scalable with the growing amount of data appears to be still missing.

II.1.2.1 HMM-based classification

For this thesis, we explored the possibility of the use of HMMs derived from classified phages for accurately classifying sets of unclassified phages (Chapter II.3 and Chapter II.4). The use of HMMs for classification has been reported in multiple studies (Grazziotin et al. 2017; Fouts 2006; Aiewsakun et al. 2018; Skewes-cox et al. 2014). HMMs were the method of choice for the comparison of protein families since they are powerful for the efficient representation of variation and have the potential to detect three times more remote homologs than conventional pairwise methods (Barrett et al. 1998). Even though profile HMMs are a powerful tool, pitfalls and challenges exist that need to be considered to generate the best possible model. The quality of the Multiple Sequence Alignment (MSA) will be reflected in accuracy and the detection potential of the HMM. An MSA should contain a correct balance of sequences to represent the

diversity of an orthologous group while avoiding oversampling biases (Reyes et al. 2017). Multidomain proteins are negligible in phages; thus full-length protein MSA is used for HMM generation (Grazziotin et al. 2017). For protein profile HMM generation for ClassiPhage and ClassiPhage 2.0, i) excluding redundant proteins from the training dataset to avoid over-fitting (Manavalan et al. 2017), ii) the use of a Markov clustering algorithm MCL (Enright A.J., Van Dongen, S. and Ouzounis 2002), and finally ii) an iterative process of HMM refinement resulted in diversity representation per model.

It should be noted that similar peptides were removed only from the training dataset and not from the benchmarking dataset (Manavalan et al. 2017). For ClassiPhage, we additionally investigated missing ORFs for phage nucleotide sequences and included them in the input MSA for HMM generation and showed that the initial iterative process captured MSA diversity. Thiswas reflected in the negligible improvements resulting from an hmmscan with the additionally refined HMMs.

pVOG is a maintained online database with phage specific HMMs readily available for download and use (Grazziotin et al. 2017). However pVOG HMMs were generated by pooling all phage CDS together without distinguishing which phages belong to which phage family and thus an hmm scan displays the same bit-score value across different phage family for a protein hit. Contrary to pVOG, for the ClassiPhage approach, phages are initially grouped per phage family priorto the protein cluster and HMM generation. The hmmscan displays a variable bit-score value discriminating between different phage families. We additionally demonstrated that HMMs couldbe used for phage classification of phages on the genera, family and subfamily levels (Chapter II.3).

Lastly, an important aspect to consider is that HMM-based methods rely to some extent on similarities to already known viruses. Consequently, it is essential to regularly update sequence databases for the future effectiveness of such methods (Skewes-cox et al. 2014).

II.1.2.2 HMMs as input for an ANN classifier

For ClassiPhage 2.0 (Chapter II.4), we explored the use of phage family specific constructed HMMs scanning phage proteomes as an input matrix for an artificial neural network (ANN) classifier to group phages into the existing phage families. One of the significant advantages of

ANN is their ability to discern relationships between the relevant features with no human interference (Min et al. 2017; Arango-Argoty et al. 2018). Additionally, ANN can be used for multi-label classification problems contrary to other machine learning algorithms (Boutell et al. 2004).

The similarity indicator selected for the chosen classifier was the bit-score. Unlike E-values, bit-scores take into account the degree of identity between sequences and is independent of the database size (Pearson 2013; Arango-Argoty et al. 2018). ClassiPhage 2.0 reached an accuracy of 84.18% and when tested on a benchmark dataset, showed the potential application of the method for accurate classification of phage consistently with ICTV classification. However, ClassiPhage 2.0 suffers from over-representation of *Caudovirales* derived HMMs and thus affecting the ANN input nodes weights. This is an inherent problem since *Caudovirales* represent over 86% of phage sequences in public databases (Bolduc et al. 2017). We expect an improved accuracy of ANN prediction as currently underrepresented taxa get populated. Thus additional sources of viral sequences, such bacterial genomes explored for the prophage content and viral metagenomic datasets, must be examined for enriching low abundant phage families (Bolduc et al. 2017; Roux, Enault, et al. 2015; Simmonds et al. 2017).

Lastly, an important aspect to consider is the need for validating the benchmarking dataset against the ICTV master species list (https://talk.ictvonline.org/files/master-species-lists/). The classification assigned in GenBank files is not yet confirmed. Therefore the examination of the designated classification in GenBank files against the ICTV's golden standard classified phages is to be verified in an initial step. At a second step classification of the unclassified phage datasets can be considered.

III.1.3 IdentiPhage

Prophages identification is instrumental for the understanding of the dynamic relationship between phages and their host in addition to the understanding of the ecology and evolution of bacteria (Hans-W Ackermann 2011). On the one hand, multiple tools are available for prophage prediction from their sequence information. Existing tools such as PHASTER, Phage_finder, and profinder are based on sequence comparison to a phage database.Thereforeprophage identification is highly dependent on similarities to already known

phages, existing in those databases (Arndt et al. 2016; Fouts 2006; Lima-Mendez et al. 2008). On the other hand, the identification of prophages based on a sequence derived features, is key to identify prophages without any sequence similarities to known prophages (Akhter et al. 2012). To date, phiSpy is the only existing tool, which was developed based on 7 different prophage characteristics (Akhter et al. 2012). Thus, phiSpy proved to predict much more unknown phages compared to those tools.

Additionally, distinctive prophage features are widely used for the identification of phage bins from metagenomic datasets (Amgarten et al. 2018). VirSorter was developed for prophage identification but performs better for viral sequence identification from metagenomic bins (Roux, Enault, et al. 2015). Recently, MARVEL was developed for the identification of viral sequences from metagenomic bins as well, using three phage characteristics as an input for a Random Forest (RF) classifier. MARVEL showed higher sensitivity in comparison to VirSorter(Roux, Enault, et al. 2015).

Aforementioned distinctive and prophage features include the gene density and strand shifts which are considered in multiple studies (Amgarten et al. 2018; Akhter et al. 2012). Increased gene density has been suspectedof being a direct outcome of the limited phage capsid size (Chirico et al. 2010; Roux, Enault, et al. 2015; Mahmoudabadi & Phillips 2018; Amgarten et al. 2018). Lower strand-shift rates can be a result of the co-regulated transcriptional and translational unit to ensure competitive superiority (Amgarten et al. 2018; Akhter et al. 2012).

Additional explored features include K-mer(Pan et al. 2018) and GC content which tend to have weak performance since it is known that phages try to adapt to their host (Amgarten et al. 2018). AT and GC skews, as well as the abundance of phage words based on their oligonucleotide composition was shown to perform better for closely related genomes (Akhter et al. 2012). Features such as protein lengths (Amgarten et al. 2018; Akhter et al. 2012), the average spacing between genes (Amgarten et al. 2018), gene density (Amgarten et al. 2018), transcription strand directionality (Amgarten et al. 2018; Akhter et al. 2012) and ATG relative frequency (Amgarten et al. 2018) have the ability to locate prophages without any sequence similarities to known phages. However, Akhter et al (2012) showed that the median of all protein lengths displays a sharp change at the beginning of a phage region, contrary to the

average protein length used by Amgarten et al. (2018), which revealed a gradual change. All these features are a consequence of groups of small peptides encoded by closely collocated ORFs in phage genomes.

Additionally, features such as the insertion points of phages (Arndt et al. 2016;Arndt et al. 2017; Akhter et al. 2012; Zhou et al. 2011) and the phage proteins homology search (Amgarten et al. 2018; Akhter et al. 2012)are dependent on similarities to already known phages. However, contrary to phiSpy, MARVEL uses hits to known phage HMMs from the pVOG database, which are known to locate distantly related homologous protein and thus, the combination of matches lead to the identification of distantly related phages (Amgarten et al. 2018).

Even though those tools exist, their performance can still be enhanced. Amgarten et al. (Amgarten et al. 2018) stated that even with the additional considered features, the developed MARVEL tool could only effectively predict *Caudovirales* phages. Thus, we utilized the available bacterial GenBank files and available phage sequences to develop a novel computational method we call IdentiPhage(Chapter II.5). We use a combination of sequence derived features as an input for a Deep Neural Network (DNN) classifier to predict prophage regions. Also, we used genes overlap as a feature which was not considered in any of the available tools although it has been long known as a phage specific feature (Chirico et al. 2010; Brandes & Linial 2016). IdentiPhage was able to identify numerous genomic hallmarks in bacterial genomes. When compared to existing tools using the *V. alginolyticus* benchmarking dataset (Chapter II.2), IdentiPhage was able to identify one additional phage missed by PHASTER and PhiSpy, however, missed one region predicted by both tools — thus concluding that with its current status, the method can be used to complement the existing tools for phage prediction.

A DNN can use tens of thousands of parameters. As such, it can overfit easily with a small sample set and often requires convenient regularization, such as including a dropout layer, for successful performance (Srivastava 2014). However, other Machine Learning (ML) algorithms, such as recurrent neural networks (Morota et al. 2018) or support vector machines (Manavalan, Tae H. Shin, et al. 2018), should be explored in combination with rigorous cross-validation

methods and their performance should be compared on the same dataset for selecting the best predictor (Amgarten et al. 2018; Tan et al. 2018). To further elevate the performance of the classifier, a feature selection process such as ANOVA or RF can be employed to select the most informative input features from noisy datasets (Pan et al. 2018; Tan et al. 2018; Ding et al. 2014).

III.2 How does everything come together

Given that more microbial and viral genomes are expected to be discovered(Roux, Hallam, et al. 2015), IdentiPhage and ClassiPhage approaches were flexibly designed to adapt to the anticipated changes. For both methods, the described five-steps guidelines to develop prediction models were followed (Manavalan, Tae Hwan Shin, et al. 2018; Manavalan et al. 2017; Feng et al. 2013).ClassiPhage can be applied to taxonomically classify prophage regions predicted via IdentiPhage(Figure 3). Future considerations for the development of IdentiPhage include i) the use of phage features such as strand shifts (Amgarten et al. 2018), oligomer bias (Jurtz et al. 2016) and integration repeats (Arndt et al. 2016) and ii) the use of a feature selection process for the evaluation of the sequence derived feature. Future considerations for IdentiPhage include i) the consideration of an iterative process for the proteins not used for HMM generation using BLASTp to eventually generate a set of diverse homologues adequate for a HMM and ii) the use of input features such as the order of identified hits and the average distance between the identified hits (Amgarten et al. 2018; Akhter et al. 2012) as additional input features for the ANN. Lastly, to overcome biases introduced when different qualities of annotations are combined (de Crécy-Lagard 2016), open reading frame (ORF) prediction for phage nucleotide sequences is to be performed using the generated refined HMMs as seed.

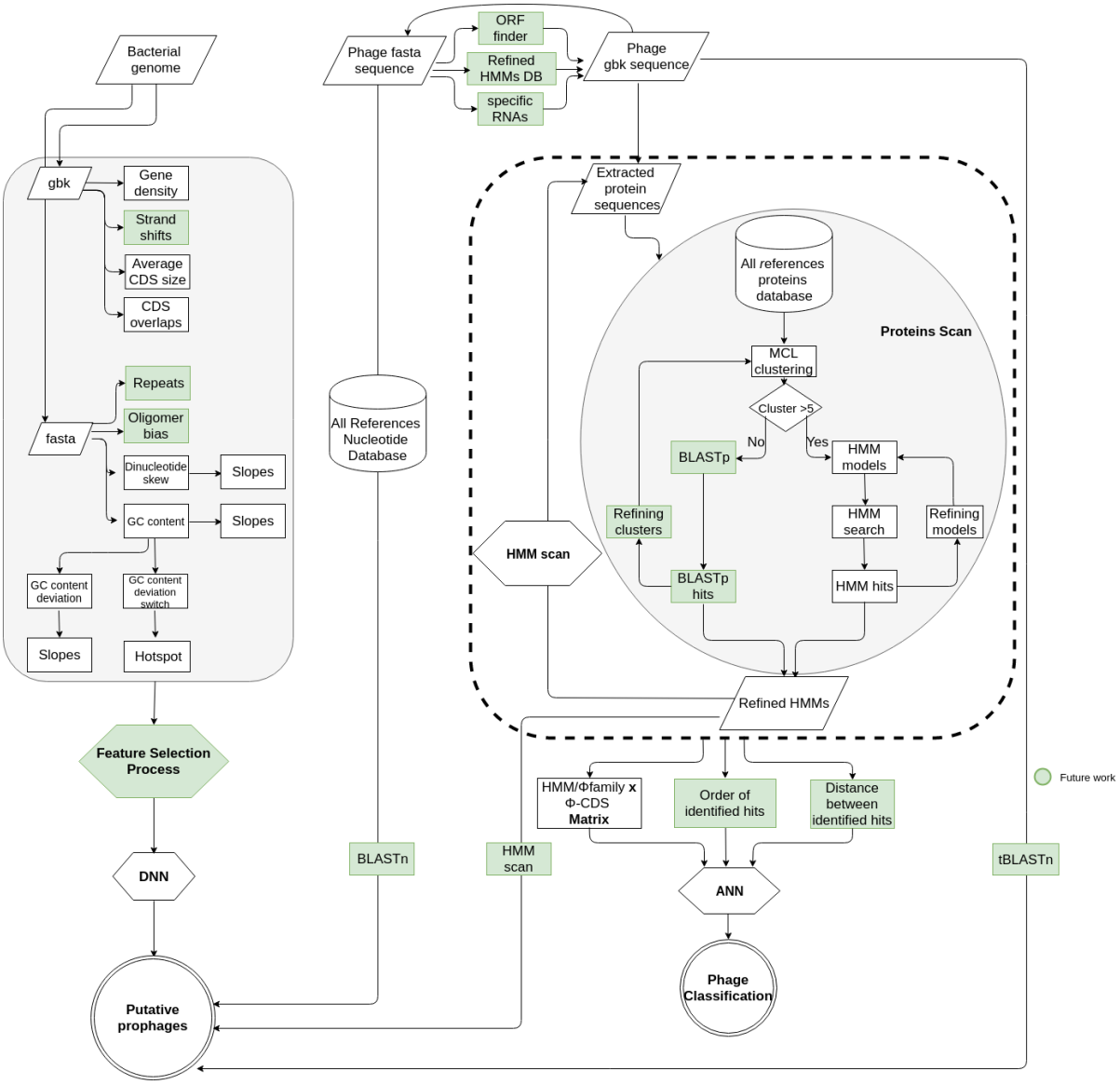


Figure 3: Comprehensive workflow summary of the investigated projects for prophage identification and classification.

The workflow consists of two major pipelines. The workflow described on the right is outlining ClassiPhage and ClassiPhage 2.0. Phage CDS sequences were extracted and clustered for HMM generation and refinement. The refined protein profile HMMs are used to scan the initial phage CDS results. The resulting matrix, with additional features, is used as an input for an ANN for phage sequence taxonomic classification. The workflow described on the left is describing IdentiPhage. It outlines sequence derived features computed out of GenBank and fasta files, and additional ones to be considered in future work. The resulting matrix is used as an input for a DNN for the identification of putative prophage regions. The two pipelines are to be linked, by classifying prophage regions predicted by the IdentiPhage pipeline. To overcome different qualities of annotation, improving phage CDS prediction is outlined. Future work plans are colored in green.

CHAPTER IV: Summary, Conclusion,and Outlook

IV.1 Summary

This work reports the potential of a fully automated genome based phage prediction and classification method with ever-increasing amounts of sequencing data. Firstly, we present an approach we call ClassiPhage and ClassiPhages 2.0 which was established describing phage taxonomical classification. ClassiPhagewas generated as a proof of principle on a defined set of phage families infecting *Vibrio* species while ClassiPhage 2.0 was broadly applied to include all phage families available. The method is based on generating and refining protein profile Hidden Markov Models (HMM) for every group of 12 phage families in total. To test sensitivity and specificity, 5,920 HMMs were used to scan the initial phage protein-coding sequences from 8,721 phages. Thus a cross-scan scoring matrix was generated. We profited from machine learning techniques which are proving to be valuable for extracting critical information and outcome prediction from big data. Thus the cross-scan matrix was used as an input for an artificial neural network (ANN) for phage classification. The accuracy of the ANN reached 84.18 % indicating the efficiency of the method. The method was tested on a set of vibriophages classified via multiple HMM hits results. Our results emphasize the need for more comprehensive and representative phage sequencing data in public databases.

Secondly, a method we call IdentiPhagewas established describing the prediction of integrated prophages in bacterial genome hosts. The method uses a set of 12 sequence derived features generated from a dataset of 11,373 bacterial using a sliding window approach. To assign a positive phage label to the matrix, we employed 8,721 phage genomes as a reference database for a BLASTn approach. The generated matrix was used as an input for a Deep Neural Network (DNN) for the prediction of potential prophage regions and achieved a specificity of 80.14%. We show that IdentiPhagecan locate prophages without any sequence similarities to known phages by testing the method on a set of experimentally identified *Inoviridae* phages infecting various *Vibrio alginolyticus* species. Our results indicate that IdentiPhage plays a complementary role to existing tools. However it would benefit from a feature selection process to select the most informative sequence features for future developments.

IV.1 Conclusion

In conclusion, an ever-increasing amount of phage genome sequence data is being generated and deposited into existing databases with no taxonomic assignment. Even though multiple computational methods exist which show encouraging results, a broad phage classification method is far from complete as long as there exist under-sampled phage families. The numbers demonstrate how distant we are from an accurate representation of viral diversity in public databases. To make such databases more comprehensive and useful, it is of paramount importance to characterize a more significant amount of viral sequences from a broader taxonomic range.

For the first research topic, we designed a flexible method to accommodate advances and changes over time.

- We generated and refined phage family specific protein profile HMMs.
- We demonstrated the potential of combined protein profile HMMs for phage taxonomic classification.
- We classified a set of published but preliminary unclassified vibriophages.
- We demonstrated that our classification is in accordance with experimentally characterized phages proved to belong to the *Inoviridae* phage family.

Automation is necessary to routinely classify sequenced phages using features ensuring accuracy.

- We demonstrated the potential of the use of artificial neural networks for phage characterization based on the combination of HMM hits.

The comparatively low cost and minimal time required for the computational identification of prophages in comparison to the labor-intensive and expensive experimental approaches make these tools indispensable among scientists.

- For the second research topic, we demonstrated the potential of identifying prophages based on sequence information using a Deep Neural Network.

- We computed 12 sequence-derived features singling out genomic hallmarks in bacterial hosts.
- We demonstrated the potential of the application of a DNN together with the sequence-derived features for prophage identification.

IV.3 Outlook

In this work we presented two methods based on sequence information, one to identify phages we call “IdentiPhage” and one to classify phages we call “ClassiPhage”. We show that both methods achieved the intended purposes but would greatly benefit from an increasing number of low populated phage families in public databases.

For future considerations, the ICTV would need to have a decisive framework for the integration of sequenced phages into their current taxonomic scheme. Scientists would need to combine their efforts in populating the under-representing viral families by exploring various metagenomic datasets. Subsequently, HMMs generation for under-represented phage families would be achievable, andClassiPhage’s performance would improve.

The further development of IdentiPhage together with informative sequence-derived features can effectively identify and characterize putative boundaries todetermine true phages. The putative prophages would then be subjected to taxonomic classification using the generated HMMs and the ClassiPhage 2.0 model.

In the future, it will be of great value to create a publicly accessible web server for prophage identification and classification from sequence data based on the described methods.

CHAPTER V: General References

V.1 Introduction References

Abedon, S.T. et al., 2011. Phage treatment of human infections. *Bacteriophage*, 1(2), pp.66–85.

Adams, M.J. et al., 2017. 50 years of the International Committee on Taxonomy of Viruses: progress and prospects. *Archives of Virology*, 162(5), pp.1441–1446.

Aiewsakun, P. et al., 2018. Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: Steps towards a unified taxonomy.*Journal of General Virology*.

Aiewsakun, P. & Simmonds, P., 2018. The genomic underpinnings of eukaryotic virus taxonomy: Creating a sequence-based framework for family-level virus classification. *Microbiome*, 6(1), pp.1–24.

Akhter, S., Aziz, R.K. & Edwards, R.A., 2012. PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. *Nucleic Acids Research*, 40(16), pp.1–13.

Alves, J.M.P. et al., 2016. GenSeed-HMM: A tool for progressive assembly using profile HMMS as seeds and its application in Alpavirinae viral discovery from metagenomic data.*Frontiers in Microbiology*.

Amgarten, D. et al., 2018. MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. *Frontiers in Genetics*.

Arndt, D. et al., 2017. PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes. *Briefings in Bioinformatics*, (May), pp.1–8. Available at: http://academic.oup.com/bib/article/doi/10.1093/bib/bbx121/4222653/PHAST-PHASTER-and-PHASTEST-Tools-for-finding.

Arndt, D. et al., 2016. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic acids research*, 44(W1), pp.W16–W21.

Asare, P.T. et al., 2015. Putative type 1 thymidylate synthase and dihydrofolate reductase as signature genes of a novel bastille-like group of phages in the subfamily Spounavirinae. *BMC Genomics*, 16(1). Available at: http://dx.doi.org/10.1186/s12864-015-1757-0.

Barrett, C. et al., 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods.*Journal of Molecular Biology*, 284(4), pp.1201–10. Available at: http://www.ncbi.nlm.nih.gov/sites/entrez.

Bolduc, B., Jang, H. Bin, Doulcier, G., You, Z.-Q., et al., 2017. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect *Archaea* and *Bacteria*. *PeerJ*.

Brüssow, H. & Hendrix, R.W., 2002. Phage Genomics: Small is beautiful. *Cell*.

Chow, C.-E.T. & Suttle, C.A., 2015. Biogeography of Viruses in the Sea. *Annual Review of Virology*, 2(1), pp.41–66.

Corel, E. et al., 2016. Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. *Trends in Microbiology*, 24(3), pp.224–237. Available at: http://dx.doi.org/10.1016/j.tim.2015.12.003.

Ding, H. et al., 2014. Identification of bacteriophage virion proteins by the ANOVA feature selection and

analysis.*Molecular BioSystems*, 10(8), pp.2229–2235.

Feng, P.M., Ding, H., et al., 2013. Naïve bayes classifier with feature selection to identify phage virion proteins. *Computational and Mathematical Methods in Medicine*.

Feng, P.M., Lin, H. & Chen, W., 2013. Identification of antioxidants from sequence information using naïve bayes. *Computational and Mathematical Methods in Medicine*.

Fidelma Boyd, E. & Brussow, H., 2002. Common themes among bacteriophage-encoded virulense factors and diversity among the bacteriophages involved. *TREND in Microbiology*, 10(No. 11).

Fortier, L.C. &Sekulovic, O., 2013. Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence*, 4(5), pp.354–365.

Fouts, D.E., 2006. Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Research*, 34(20), pp.5839–5851.

Frazer, K.A. et al., 2004. VISTA: Computational tools for comparative genomics. *Nucleic Acids Research*, 32(WEB SERVER ISS.), pp.273–279.

Frederick, B.Y. & Twort, W., 1931. Ultra-Microscopic Viruses and Their Cultivation. , pp.204–235.

Grazziotin, A.L., Koonin, E. V &Kristensen, D.M., 2017. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. , 45(October 2016), pp.491–498.

Hans-W Ackermann, 2011. Bacteriophage Taxonomy. *Microbiology Australia*, 32(2), pp.90–94.

Housby, J.N. & Mann, N.H., 2009. Phage therapy. *Drug Discovery Today*, 14.

Iranzo, J., Koonin, E. V., et al., 2016. Bipartite Network Analysis of the Archaeal Virosphere: Evolutionary Connections between Viruses and Capsidless Mobile Elements.*Journal of Virology*, 90(24), pp.11043–11055.

Iranzo, J., Krupovic, M. & Koonin, E. V., 2016. The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. *mBio*, 7(4), pp.1–21.

Krupovic, M., Dutilh, B.E. & Adriaenssens, E.M., 2016. Taxonomy of prokaryotic viruses : update from the ICTV bacterial and archaeal viruses subcommittee.*Archives of Virology*, 161(4), pp.1095–1099.

Lavigne, R. et al., 2008. Unifying classical and molecular taxonomic classification: analysis of the Podoviridae using BLASTP-based tools. *Research in Microbiology*, 159(5), pp.406–414.

Lawrence, J.G., Hatfull, G.F. & Hendrix, R.W., 2002. Imbroglios of viral taxonomy: Genetic exchange and failings of phenetic approaches. *Journal of Bacteriology*, 184(17), pp.4891–4905.

Lefkowitz, E.J. et al., 2017. *Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2017)*,

Lefkowitz, E.J. et al., 2018. Virus taxonomy: The database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Research*, 46(D1), pp.D708–D717.

Lima-Mendez, G. et al., 2008. Prophinder: A computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics*.

Lopes, A. et al., 2014. Automated classification of tailed bacteriophages according to their neck organization.

BMC Genomics, 15(1), pp.1–17.

Manavalan, B. & Lee, J., 2017. SVMQA: support-vector-machine-based protein single-model quality assessment. *Bioinformatics (Oxford, England)*.

Manavalan, B., Lee, J. & Lee, J., 2014. Random forest-based protein model quality assessment (RFMQA) using structural features and potential energy terms. *PLoS ONE*.

Manavalan, B., Shin, T.H. & Lee, G., 2018. PVP-SVM : Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine. , 9(March), pp.1–10.

McNair, K., Bailey, B.A. & Edwards, R.A., 2012. PHACTS, a computational approach to classifying the lifestyle of phages. *Bioinformatics*, 28(5), pp.614–618.

Meier-Kolthoff, J.P. & Göker, M., 2017. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics (Oxford, England)*, 33(21), pp.3396–3404.

Merabishvili, M. et al., 2011. Phenotypic and genotypic variations within a single bacteriophage species. *Virology Journal*, 8(1), p.134.

Morota, G. et al., 2018. BIG DATA ANALYTICS AND PRECISION ANIMAL AGRICULTURE SYMPOSIUM: Machine learning and data mining advance predictive big data analysis in precision animal agriculture1. *Journal of Animal Science*, 96(4), pp.1540–1550. Available at: https://academic.oup.com/jas/article/96/4/1540/4828311.

Novik, G., Ladutska, A. & Rakhuba, D., 2017. Bacteriophage taxonomy and classification. *Antimicrobial Research: Novel bioknowledge and educational programs*, pp.251–259. Available at: https://pdfs.semanticscholar.org/7403/5f7ed80d1dd17cd4521f2c461b81f240dd43.pdf.

Ofir, G. & Sorek, R., 2018. Review Contemporary Phage Biology : From Classic Models to New Insights. *Cell*, 172(6), pp.1260–1270. Available at: https://doi.org/10.1016/j.cell.2017.10.045.

Pan, Y. et al., 2018. Identification of bacteriophage virion proteins using multinomial Naïve bayes with g-gap feature tree.*International Journal of Molecular Sciences*.

Ren, J. et al., 2017. VirFinder : a novel k -mer based tool for identifying viral sequences from assembled metagenomic data. , pp.1–20.

Reyes, A. et al., 2017. Use of profile hidden Markov models in viral discovery: current insights.*Advances in Genomics and Genetics*, Volume 7(July), pp.29–45. Available at: https://www.dovepress.com/use-of-profile-hidden-markov-models-in-viral-discovery-current-insight-peer-reviewed-article-AGG.

Reyes, A. & Gruber, A., 2017. Use of profile hidden Markov models in viral discovery : current insights. , (July).

Roberts, R.J., 2005. How restriction enzymes became the workhorses of molecular biology.*Proceedings of the National Academy of Sciences*, 102(17), pp.5905–5908.

Rohwer, F. &Edwards, R., 2002. The phage proteomic tree: A genome-based taxonomy for phage.*Journal of Bacteriology*, 184(16), pp.4529–4535.

Roux, S. et al., 2016. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, 537(7622), pp.689–693. Available at: http://dx.doi.org/10.1038/nature19366.

Roux, S. et al., 2019. Minimum information about an uncultivated virus genome (MIUVIG). *Nature Biotechnology*, 37(1), pp.29–37.

Roux, S., Hallam, S.J., et al., 2015. Viral dark matter and virus – host interactions resolved from publicly available microbial genomes. , (January), pp.1–20.

Roux, S., Enault, F., et al., 2015. VirSorter : mining viral signal from microbial genomic data. , pp.1–20.

Salmond, G.P.C. & Fineran, P.C., 2015. andfuture. *Nature Publishing Group*, 13(12), pp.777–786. Available at: http://dx.doi.org/10.1038/nrmicro3564.

Seguritan, V. et al., 2012. Artificial Neural Networks Trained to Detect Viral and Phage Structural Proteins. *PLoS Computational Biology*, 8(8).

Simmonds, P., Adams, M.J. & Benko, M., 2017. Virus taxonomy in the age of metagenomics. *Nature Reviews*, 15.

Skewes-cox, P. et al., 2014. Profile Hidden Markov Models for the Detection of Viruses within Metagenomic Sequence Data. , 9(8).

Sorek, R. et al., 2017. Communication between viruses guides lysis–lysogeny decisions. *Nature*, 541(7638), pp.488–493. Available at: http://dx.doi.org/10.1038/nature21049.

Sorek, R., Lawrence, C.M. & Wiedenheft, B., 2013. CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea. *Annual Review of Biochemistry*, 82(1), pp.237–266.

Summers, W.C., 2017. Félix Hubert d'Herelle (1873–1949): History of a scientific mind. *Bacteriophage*, 6(4), p.e1270090.

Sundstro, A., 2012. Gegenees : Fragmented Alignment of Multiple Genomes for Determining Phylogenomic Distances and Genetic Signatures Unique for Specified Target Groups. , 7(6).

Tan, J.X. et al., 2018. Identifying phage virion proteins by using two-step feature selection methods. *Molecules*, 23(8), pp.1–13.

Whitman, W.B., Coleman, D.C. & Wiebe, W.J., 1998. Prokaryotes: The unseen majority. *The National Academy of Sciences*.

Zhou, Y. et al., 2011. PHAST: A Fast Phage Search Tool. *Nucleic Acids Research*, 39(SUPPL. 2), pp.347–352.

V.2 Discussion References

Aiewsakun, P. et al., 2018. Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: Steps towards a unified taxonomy. *Journal of General Virology*.

Akhter, S., Aziz, R.K. & Edwards, R.A., 2012. PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. *Nucleic Acids Research*, 40(16), pp.1–13.

Alves, J.M.P. et al., 2016. GenSeed-HMM: A tool for progressive assembly using profile HMMS as seeds and its application in Alpavirinae viral discovery from metagenomic data. *Frontiers in Microbiology*.

Amgarten, D. et al., 2018. MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. *Frontiers in Genetics*.

Arango-Argoty, G. et al., 2018. DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*.

Arndt, D. et al., 2017. PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes. *Briefings in Bioinformatics*, (May), pp.1–8. Available at: http://academic.oup.com/bib/article/doi/10.1093/bib/bbx121/4222653/PHAST-PHASTER-and-PHASTEST-Tools-for-finding.

Arndt, D. et al., 2016. PHASTER : a better , faster version of the PHAST phage search tool. , 44(August), pp.16–21.

Barrett, C. et al., 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *Journal of Molecular Biology*, 284(4), pp.1201–10. Available at: http://www.ncbi.nlm.nih.gov/sites/entrez.

Bolduc, B. et al., 2017. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect *Archaea* and *Bacteria*. *PeerJ*.

Boutell, M.R. et al., 2004. Learning multi-label scene classification. *Pattern Recognition*, 37(9), pp.1757–1771.

Brandes, N. & Linial, M., 2016. Gene overlapping and size constraints in the viral world. *Biology Direct*, 11(1), pp.1–15. Available at: http://dx.doi.org/10.1186/s13062-016-0128-3.

Casjens, S., 2003. MicroReview Prophages and bacterial genomics : what have we learned so far ? , 49, pp.277–300.

Chirico, N., Vianelli, A. & Belshaw, R., 2010. Why genes overlap in viruses. *Proceedings of the Royal Society B: Biological Sciences*, 277(1701), pp.3809–3817.

de Crécy-Lagard, V., 2016. Quality Annotations, a Key Frontier in the Microbial Sciences. *Microbe Magazine*, 11(7), pp.303–310.

Ding, H. et al., 2014. Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. *Molecular BioSystems*, 10(8), pp.2229–2235.

Enright A.J., Van Dongen, S. and Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein

families. *Nucleic Acids Research*, 30(7), pp.1575–1584.

Fauquet, C.M. & Fargette, D., 2005. International Committee on Taxonomy of Viruses and the 3,142 unassigned species. *Virology Journal*, 2, pp.1–10.

Feng, P.M. et al., 2013. Naïve bayes classifier with feature selection to identify phage virion proteins. *Computational and Mathematical Methods in Medicine*.

Fouts, D.E., 2006. Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Research*, 34(20), pp.5839–5851.

Grazziotin, A.L., Koonin, E. V & Kristensen, D.M., 2017. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. , 45(October 2016), pp.491–498.

Hans-W Ackermann, 2011. Bacteriophage Taxonomy. *Microbiology Australia*, 32(2), pp.90–94.

Hertel, R. et al., 2015. Genome-based identification of active prophage regions by next generation sequencing in Bacillus licheniformis DSM13. *PLoS ONE*, 10(3), pp.1–18.

Iranzo, J., Koonin, E. V., et al., 2016. Bipartite Network Analysis of the Archaeal Virosphere: Evolutionary Connections between Viruses and Capsidless Mobile Elements. *Journal of Virology*, 90(24), pp.11043–11055.

Iranzo, J., Krupovic, M. & Koonin, E. V., 2016. The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. *mBio*, 7(4), pp.1–21.

Jang, H.B. et al., 2019. Gene sharing networks to automate genome-based prokaryotic viral taxonomy.

Jurtz, V.I. et al., 2016. MetaPhinder — Identifying Bacteriophage Sequences in Metagenomic Data Sets. , pp.1–14.

Lima-Mendez, G. et al., 2008. Prophinder: A computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics*.

Mahmoudabadi, G. & Phillips, R., 2018. A comprehensive and quantitative exploration of thousands of viral genomes. *eLife*, 7, pp.1–26.

Mai-Prochnow, A. et al., 2015. "Big things in small packages: The genetics of filamentous phage and effects on fitness of their host." *FEMS Microbiology Reviews*, 39(4), pp.465–487.

Manavalan, B. et al., 2017. *MLACP: machine-learning-based prediction of anticancer peptides*, Available at: www.impactjournals.com/oncotarget.

Manavalan, B., Shin, T.H. & Lee, G., 2018. DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget*.

Manavalan, B., Shin, T.H. & Lee, G., 2018. PVP-SVM: Sequence-based prediction of phage virion proteins using a support vector machine. *Frontiers in Microbiology*.

Meier-Kolthoff, J.P. & Göker, M., 2017. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics (Oxford, England)*, 33(21), pp.3396–3404.

Min, S., Lee, B. & Yoon, S., 2017. Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5), pp.851–869.

Morota, G. et al., 2018. BIG DATA ANALYTICS AND PRECISION ANIMAL AGRICULTURE SYMPOSIUM: Machine learning and data mining advance predictive big data analysis in precision animal agriculture1. *Journal of Animal Science*, 96(4), pp.1540–1550. Available at: https://academic.oup.com/jas/article/96/4/1540/4828311.

Pan, Y. et al., 2018. Identification of bacteriophage virion proteins using multinomial Naïve bayes with g-gap feature tree. *International Journal of Molecular Sciences*.

Pearson, W.R., 2013. An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics*, (SUPPL.42), pp.1–8.

Reyes, A. et al., 2017. Use of profile hidden Markov models in viral discovery: current insights. *Advances in Genomics and Genetics*, Volume 7(July), pp.29–45.

Rohwer, F. & Edwards, R., 2002. The phage proteomic tree: A genome-based taxonomy for phage. *Journal of Bacteriology*, 184(16), pp.4529–4535.

Roux, S., Hallam, S.J., et al., 2015. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife*.

Roux, S., Enault, F., et al., 2015. VirSorter : mining viral signal from microbial genomic data. , pp.1–20.

Shapiro, J.W. & Putonti, C., 2018. Gene co-occurrence networks reflect bacteriophage ecology and evolution. *mBio*.

Simmonds, P., Adams, M.J. & Benko, M., 2017. Virus taxonomy in the age of metagenomics. *Nature Reviews*, 15.

Skewes-cox, P. et al., 2014. Profile Hidden Markov Models for the Detection of Viruses within Metagenomic Sequence Data. , 9(8).

Srivastava, N.G.H.A.K.I.S.S., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15.1, pp.1929–1958.

Tan, J.X. et al., 2018. Identifying phage virion proteins by using two-step feature selection methods. *Molecules*, 23(8), pp.1–13.

Wendling, C.C. et al., 2017. Tripartite species interaction: eukaryotic hosts suffer more from phage susceptible than from phage resistant bacteria. *BMC Evolutionary Biology*, 17(1).

Youle, M.; Haynes, M.. R., 2012. Scratching the Durface of Biology's Dark Matter. *Springer, Dordrecht*, pp.61–81.

Zhou, Y. et al., 2011. PHAST: A Fast Phage Search Tool. *Nucleic Acids Research*, 39(SUPPL. 2), pp.347–352.

CHAPTER VI: Appendix

VI.1 ACKNOWLEDGEMENT

First, I would like to thank my advisor Heiko Liesegang,and for his constant advice for the last years, I will forever be indebted. I joined the G2L team in 2015 during my Master studies,and Heiko introduced me to the phage topic for my Master thesis. It is a great project,and I appreciate the opportunity to work on it. I am also thankful for the freedom, flexibility and the trust he has given me.

I am also grateful to have Rolf and Burkhard as members of my thesis committee. Their suggestions were valuable to me and this project.

I want to say thank you to Sascha. He was the person who taught me a lot, endured a lot with me and tortured me into learning new things out of my scope. It was a great experience working with him.

I want to thank all the former and current members of G2L. I am grateful to Jacqueline Hollensteiner, Inka Willms, Robert Hertel, Anja Poehlein, Stefanie Diaz and the IT team for their help and support. They were my support system at the department during the time I spent here. I want to thank everyone for the friendly atmosphere in this department.

I need to thank my students as well; Anton, Florentin, and David, without whose brilliant and hard work, this work would never have been possible.

I need to thank all my dear friends in Göttingen, Ghida, Rayan, Moazz, Anette, Blanca,and Nacho. I can’t put into words the amount of love and support I have received from these people. They have been my support system and my family here, in Goettingen, throughout the years.

I want to thank my Mom and Brother for their love and support during it all; I would never be the person I am without them in my life.

And lastly, I owe it all the KAAD team, without them believing in me and granting me the scholarship for the financial support for the last three years, this Ph.D. project wouldn’t have been possible.

I am sincerely grateful.

VI.2 Thesis Declaration

Declaration of plagiarism

I hereby confirm that I have written the doctoral thesis entitled “Functional Phage Genomics of selected Taxa” independently. I have not used other sources or facilities others than the ones mentioned in the chapters. The contributions of the authors are given preceding the respective manuscripts. Moreover, I have not used unauthorized assistance and have not submitted this thesis previously in any form for another degree at any institution or university.

_____

City, date, name Cynthia Maria Chibani, Göttingen

VI.3 Additional publications

1. Groß, U., Brzuszkiewicz, E., Gunka, K., Starke, J., Riedel, T., Bunk, B., Spröer, C., Wetzel, D., Poehlein, A., **Chibani, C.** and Bohne, W., 2018. Comparative genome and phenotypic analysis of three Clostridioides difficile strains isolated from a single patient provide insight into multiple infection of C. difficile. *BMC genomics*,*19*(1), p.1.

2. Poehlein, A., Alghaithi, H.S., Chandran, L., **Chibani, C.M.**, Davydova, E., Dhamotharan, K., Ge, W., Gutierrez-Gutierrez, D.A., Jagirdar, A., Khonsari, B. and Nair, K.P.P., 2014. First insights into the genome of the amino acid-metabolizing bacterium Clostridium litorale DSM 5388.*Genome announcements*, *2*(4), pp.e00754-14.

3. Dannheim, H., Riedel, T., Neumann-Schaal, M., Bunk, B., Schober, I., Spröer, C.,**Chibani, C.M.,**Gronow, S., Liesegang, H., Overmann, J. and Schomburg, D., 2017. **Manual curation and reannotation of the genomes of Clostridium difficile 630Δerm and C. difficile 630**.*Journal of medical microbiology*, *66*(3), pp.286-293.

VI.4 Curriculum Vitae

VI.5 DVD