Programa de Pós-Graduação Interunidades em Bioinformática
Universidade de São Paulo

# Machine Learning Prediction in Genomic Sequences of Prokaryotic Viruses from Metagenomic Datasets

Deyvid Emanuel Amgarten
Departamento de Bioquímica
Instituto de Química - USP

Date of deposit in SPG: 16/11/2021
São Paulo, Brazil, 2021

Deyvid Emanuel Amgarten

# Machine Learning Prediction in Genomic Sequences of Prokaryotic Viruses from Metagenomic Datasets

Original Version

Ph.D. Thesis presented to the Bioinformatics Graduate Program, Universidade de São Paulo as part of the requirements necessary to obtain the degree of Doctor of Science.
Concentration area: Bioinformatics

Advisor: Dr. João Carlos Setubal
Co-Advisor: Dr. Aline Maria da Silva

São Paulo, Brazil, 2021

To Molecular Biology, Computer Sciences, and Viruses.

GGGTCTCAAGCTTAAGTGGTATAAGCATCCCGCTCATAACGGGGAGATACAGGGTTCAAATCCCTGGAGACCCACCA

01010100 01101111 00100000 01000101 01110110 01101111 01101100 01110101 01110100 01101001 01101111 01101110

VHRLVASAFHENKDNLREVNHIDGNKLNNNACNLEWCTSSEN

# Acknowledgements

**To my advisor, Prof. Dr. João Carlos Setubal**, for the learning opportunity, guidance, patience, and autonomy during this work. Today I consider myself a fully formed bioinformatics professional, and that is why I learned from the best. Thank you for all the knowledge transmitted through these seven years.

**To my co-advisor, Prof. Dr. Aline Maria da Silva**, for all her unrestricted support. It has been a real pleasure to discuss and share knowledge with her in all these years. Thank you for sharing something else beyond knowledge, your passion for phage and microorganisms. It is said that you can only be great and work for a long time with something that you really love. Your career is a fine example of that. Hope one day I can be another.

**To Roberta Verciano, Layla Farage and Lucas Braga**, for their patience and crucial help for all the work built through this thesis. Mastering so different areas as wet-lab and bioinformatics is a real challenge and even though I had my difficulties in the wet-lab department, Roberta, Layla and Lucas were always there to provide guidance, help and neat explanations. This is a gentle reminder for future me that one can go much further having good friends and being a good collaborator.

**To Bruno Koshin Vázquez Iha**, for all the help in developing vHULK. Neural Networks are not a piece of candy, and it is impressive how you master them. Thank you for conducting this work as if it was your own, and for the amazing tool we built together.

**To Vinicius Flores**, for useful scripts and help provided during analyses. Glad to see the amazing bioformatician you have become.

**To Patricia Locosque Ramos and The Sao Paulo Zoo Park Foundation**, for kindly allowing us to collect samples and study their composting units in all these years of collaboration. More than that, for all the support and help.

**To Prof. Dr. Anna Helena Reali Costa, Prof. Dr. Christian Hoffmann and Prof. Dr. André Carlos Ponce de Leon Ferreira Carvalho**, for their feedback and guidance

during the development of this thesis. Thank you for making sure I was at the right path.

**To my father, Gilberto Leonel Amgarten**, for building the foundations that allowed me to be here.

**To all my good friends**, without whom I would never have ended this thesis. Thank you for being present when stress was sky-high and for the few (several?) times you had to convince me to keep going.

In special, I thank **Daniel Robles** for the Homeric task of putting up with me all these years. Your support helped me to get here.

This study was partially financed by the **Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES)** - Finance Code 001 through a PhD scholarship to Deyvid Amgarten.

Grant n. 2011/508706, from **São Paulo Research Foundation FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO (FAPESP)** funded reagents to sequence samples and computational infrastructure to perform bioinformatic analyses.

**To all the Brazilian people**, whose tax contributions allowed me to get here. Despite all efforts to hinder our development and undermine our science, researchers and scientists like me will keep going. We will keep creating solutions to save and improve lives of you all. We were there in the COVID-19 pandemics and we will still be there if needed more.

ABSTRACT

AMGARTEN, D. Machine Learning Prediction in Genomic Sequences of Prokaryotic Viruses from Metagenomic Datasets. 2021. 79 pages. Ph.D. Thesis. Bioinformatics Graduate Program, Universidade de São Paulo, São Paulo.

Environmental viruses are extremely diverse and abundant in the biosphere. Several studies have shown prokaryotic viruses (or simply phages) as major players in determining biogeochemical cycles in oceans as well as driving microbial diversification. Besides this ecological role, phages may also be used for clinical purposes since they can kill bacterial cells and terminate infections. A crucial step in this process is the isolation of new phages, which can target a specific bacterial pathogen. Thus, researchers employ screening techniques to find and isolate pathogen-specific phages from environmental samples, which are a rich source of new phages. However, this task remains mostly exploratory and laborious if the researcher has no detailed information about the sample and its potential viral diversity. Having this problem in mind, we propose the development of a bioinformatic workflow to identify genomic sequences belonging to phages in environmental datasets, as well as for host prediction of the identified phages based on their genomic sequences. To achieve this goal, we implemented a random forest classifier and created the tool named MARVEL (Metagenomic Analyses and Retrieval of Viral Elements), which is able to efficiently predict phage genomic sequences in bins generated from whole community metagenomic short reads. We also developed a toolkit, name vHULK (Viral Host Unveiling Kit), which can predict phage's host given only their genome as input. vHULK presents higher accuracy than available tools and it can predict both host species and genus in a multiclass prediction setting. Data generated by the application of both tools in public and private composting metagenomic datasets is used for recovery, annotation, and characterization of phage diversity in composting environments. Both tools are publicly available through a GitHub repository: https://github.com/LaboratorioBioinformatica/.

Keywords: Phages, prokaryotic viruses, host prediction, phage prediction, metagenomics, machine learning

RESUMO

AMGARTEN, D. Predição em sequências de vírus de procariotos através da aplicação de técnicas de aprendizado de máquina em dados metagenômicos. 2021. 79 páginas. Tese de Doutorado. Programa de Pós-Graduação Interunidades em Bioinformática, Universidade de São Paulo, São Paulo.

Vírus ambientais são extremamente diversos e abundantes na biosfera. Estudos têm demostrado que vírus que infectam procariotos (ou simplesmente fagos) são determinantes no direcionamento de ciclos biogeoquímicos em oceanos, além de influenciarem de forma significativa a diversificação de seus hospedeiros. Sem considerar esse papel ecológico, fagos também estão sendo utilizados para propósitos clínicos graças à habilidade de infectar bactérias e terminar infecções bacterianas. Um passo crucial para esta aplicação é o isolamento de fagos que tenham como alvo um determinado patógeno bacteriano de interesse. Para isso, pesquisadores geralmente recorrem a amostras ambientais num processo dispendioso de tentativa e erro de isolamento experimental. Ter informações importantes sobre a diversidade de fagos em uma amostra, assim como potenciais hospedeiros poderia ajudar neste processo. Sendo assim, nesta tese nós propomos o desenvolvimento de um pipeline de bioinformática para recuperação de genomas de fagos de amostras ambientais, assim como para predição de hospedeiros desses genomas. Para atingir esse objetivo, nós treinamos um classificador random forest para diferenciação de sequências de fagos e o implementamos na ferramenta chamada de MARVEL. Nós também desenvolvemos a ferramenta chamada vHULK, que é capaz de predizer hospedeiros bacterianos dada a sequência do genoma do fago. Ambas as ferramentas apresentam alta acurácia e performance quando comparadas com o estado da arte em cada problema de predição. Resultados gerados pela aplicação das ferramentas desenvolvidas nesta tese em datasets metagenômicos de compostagem e solo são apresentados como uma prova de conceito e estudo de caso. Ambas as ferramentas encontram-se disponíveis no repositório público: https://github.com/LaboratorioBioinformatica/.

Palavras-chave: Fagos, bacteriófagos, vírus de procariotos, metagenômica, predição de hospedeiro viral, aprendizado de máquina, vírus ambientais

# Table of Contents

# **Chapter 1**: Introduction

## 1.1. Background

### 1.1.1. Environmental viruses and genomic data

In the past few decades, our understanding of microbial life has profoundly changed owing to innovations in environmental sampling, high-throughput sequencing, and computational analyses. These changes have resulted in a microbial culture-independent field of study named metagenomics (Chen & Pachter, 2005; Handelsman et al., 1998). The uncultured majority of bacterial and archaeal diversity has been quickly unveiled since then, with an unknown universe of environmental viruses (DeLong, 2009; Handelsman, 2004; Rappé & Giovannoni, 2003; Solden et al., 2016). Viruses are the most abundant biological entities in the biosphere, outnumbering all bacteria and archaea in the oceans at least ten times over (Bergh et al., 1989). Most environmental viruses infect bacterial hosts and are known as bacteriophages (or **phages**); these are drivers of biogeochemical cycles on Earth and key players in directing and originating bacterial diversity (Braga et al., 2018; Brum et al., 2015; Falkowski et al., 2008).

A large amount of genomic data has now become available to the scientific community owing to technological improvements regarding the second (short reads) and third generations (long reads) of DNA sequencing machines (Rusk, 2009; Schuster, 2008), allowing access to genomes of whole microbial communities without the need for laboratory cultivation. Therefore, we face a point in microbiology where we have genome sequences of microbial organisms that have never been isolated or spotted under a microscope before.

Nevertheless, wet-lab isolation of environmental bacteria and phages is crucial for their complete biological characterization and exploration of their biotechnological and therapeutic applications (Hyman, 2019; Lagier et al., 2018). Major efforts have been made to isolate phages, catalog their diversity, and maintain repositories of

phage isolates (Abdelsattar et al., 2021; Abdelsattar et al., 2021; Lin et al., 2021). Among them, we may cite the pioneer *Actinobacteriophages Database* (http://phagesdb.org/), composed of more than 3,816 isolated phages with complete genome sequences (November 2, 2021). Database maintenance and expansion are carried out by a network of researchers and undergraduate students all over the United States, known as *Phage Hunters* (Jordan et al., 2014). Moreover, thousands of genomes belonging to isolated phages are currently available in the NCBI repository of biological sequences known as *GenBank* (https://www.ncbi.nlm.nih.gov/genbank/). In addition to biological sequences, huge repositories such as ActinoDB and Genbank store all types of metadata, presenting unique opportunities for data mining and the generation of learning models.

### 1.1.2. Machine learning approaches in genomics

Owing to the significant amount of genomic data available, it is unfeasible to perform experimental assays to characterize biological attributes in all newly discovered species whose genomes are readily available. The same is true for manual curation of information generated by *in silico* genome analysis, such as predicted proteins and metabolic pathways. Therefore, there is an imperative need for computational approaches that can handle this increasing quantity of genomic data and provide meaningful and reliable biological information.

Machine learning is a field of computer science concerned with developing and applying computer algorithms that can learn hidden patterns from data (James et al., 2000; Ruppert, 2004). In other words, an algorithm can be trained to recognize a specific attribute once a list of features as example data is provided. An attribute to be predicted is commonly referred to as a label or class in supervised learning; it can be any biological attribute or phenotype of interest in the genomic context. A supervised machine learning problem generally consists of assigning labels, given a new example and a list of features in which the algorithm is trained. Features in the genomic context are measurements at the sequence level, such as GC content, oligonucleotide frequency profiles, codon usage, and gene expression. An out-of-the-box way of seeing genomic features for machine learning are categorical variables, such as the presence or absence of specific genes or domains, as applied in the study by Yu et al., 2003).

Several machine learning approaches have been applied in genomic studies, including prediction of genes based on specific models trained with coding sequences of specific groups of organisms (e.g., phages, viruses, and intronless eukaryotes) (Besemer et al., 2001), finding the location of transcription start sites (Ohler et al., 2002), and predicting the optimum temperature to differentiate mesophilic and thermophilic proteins (Gromiha & Suresh, 2008). In viral ecology and metagenomics, Khot et al. reviewed tools and computational methods for identifying attributes such as viral sequences in metagenomics datasets (2020). Various approaches were surveyed (discussed in detail later in Chapter 2), including the MARVEL tool developed by our group in this thesis (Amgarten et al., 2018).

Interactions established between two organisms (e.g., host-phage susceptibility and virulence) could also be considered as attributes for prediction. In a review presented by Edwards et al. (2016), the predictive power of several *in silico* genomic signals was tested and assessed. The authors found that all reviewed signals significantly link phages to their hosts; more importantly, however, signals based on sequence similarity (using alignments) are the most effective at identifying known phage-host pairs. However, we noted that more sophisticated machine learning approaches, such as deep neural networks, were not assessed in this review. Recently, Clement Coclet and Simon Roux reviewed state-of-the-art methods for host prediction, including alignment-dependent, alignment-free, and integrative approaches (Coclet & Roux, 2021). For example, vHULK, a tool for host prediction developed by our group in this thesis, is cited as an alignment-free alternative that relies on deep neural networks (Amgarten et al., 2020). Approaches for host prediction and vHULK are discussed further in Chapter 3.

Bearing in mind the availability of genomic data and the advances provided by the machine learning techniques described above, we believe that there is a great opportunity for the development and improvement of machine learning tools aimed at predicting phage sequences and biological attributes such as host-phage interactions.

### 1.1.3. Phage therapy and clinical possibilities

In 2013, the *United States Center for Disease Control* (CDC) published a document outlining the biggest threats in terms of multidrug-resistant (MDR) bacterial

infections (https://www.cdc.gov/drugresistance/biggest_threats.html). Threats were classified into four hazard levels: Urgent, Serious, Concerning, and Watchlist. Common infections previously easily treated by antibiotics until the 2010s, such as diarrhea and gonorrhea, are now placed at the most dangerous level of threat owing to resistance to practically all known antibiotics. Moreover, the widespread bacteria *Pseudomonas aeruginosa*, a common cause of healthcare-associated infections (pneumonia, bloodstream infections, urinary tract infections, and surgical site infections) (Botelho et al., 2019), are placed at a serious hazard level. This is an alarming scenario, and new alternatives are imperative (McKenna, 2013).

Aside from the ecological role of phages (Braga et al., 2018), attention has recently been drawn to their possible application as antibacterial agents in human and animal infections, as some are the primordial enemies and predators of bacterial species (Brives & Pourraz, 2020; Brüssow, 2005; Kortright et al., 2019; Thiel, 2004). Furthermore, clinical trials have evaluated the phage efficiency in chronic otitis caused by *P. aeruginosa* (Wright et al., 2009). Safety tests were performed in human volunteers who received *Escherichia coli* T4 phage, with results showing that no fecal phage was detectable a week after a 2-day course of oral phage application (Bruttin & Brüssow, 2005). Lastly, Schooley et al. reported a method to produce a personalized bacteriophage-based therapeutic treatment for a 68-year-old diabetic patient with necrotizing pancreatitis complicated by an MDR *Acinetobacter baumannii* infection (Schooley et al., 2017). Despite multiple antibiotic courses, the patient deteriorated over a four-month period. Researchers and the medical team managed to receive approval from the *US Food and Drug Administration* (FDE) to try an experimental last resource technique based on personalized phages that specifically target the strain of *A. baumannii*. The phage cocktail was administered intravenously, and the patient returned to health a few weeks after the treatment, with clearance of the *A. baumannii* infection.

Thus, the field of phage therapy research is going through a renaissance. Phage therapy has been demonstrated to be effective, directed to specific pathogens (pathogen-targeted); therefore, this treatment might not disrupt the healthy human microbiome. Phage therapy has also been shown to be safe for human use. Nonetheless, it is clear that much needs to be pursued regarding basic phages and clinical research to address possible pitfalls and unclear questions (Podlacha et al.,

2021). This Ph.D. thesis is inserted in the context of basic phage and host interaction research.

## 1.2. Motivation

Metagenomics has emerged as a prominent field of research, bringing exciting opportunities to reveal and characterize the unknown diversity of environmental viruses. However, several challenges need to be addressed to access all of this pledged diversity. Metagenomics, in general, is associated with significant drawbacks in the assembly and recovery of genomes in a single continuous sequence. As this is not possible, recognizing viral sequences in an ocean of sequences belonging to other organisms and clustering them in bins is also a significant challenge. Moreover, viral genomes recovered from metagenomes usually do not include information about the host it infects. Without complete viral genomes and host information, an important portion of the knowledge associated with viral dark matter remains inaccessible. Therefore, addressing these challenges is the main motivation behind this thesis.

## 1.3. Thesis Organization

This thesis comprises an introduction chapter, three main result chapters, and a conclusion chapter. Chapter 1 provides a contextualization of the general theme of this thesis. It also provides the motivation and background of the work to be presented in the following chapters. Chapters 2 to 4 were thought to be independent articles for publication, containing the main results of this thesis. Chapter 2 presents the development of a tool for the prediction of phage genomes, named MARVEL. The original work was published in August 2018 at the journal Frontiers in Genetics. Chapter 3 presents the development of vHULK, a toolkit based on artificial neural networks for phage host prediction, deposited as a pre-print in bioRxiv in December 2020. Chapter 4 presents the data and results of MARVEL and vHULK usage to recover and characterize environmental viruses from SP Zoo Park composting and soil samples. Finally, Chapter 5 provides concluding remarks and future perspectives on the work presented in this thesis.

# References


Abdelsattar, A., Dawooud, A., Rezk, N., Makky, S., Safwat, A., Richards, P., & El-Shibiny, A. (2021). How to Train Your Phage: The Recent Efforts in Phage Training. *Biologics*, *1*(2), 70–88. https://doi.org/10.3390/biologics1020005

Abdelsattar, A. S., Dawoud, A., Makky, S., Nofal, R., Aziz, R. K., & El-Shibiny, A. (2021). Bacteriophages: from isolation to application. *Current Pharmaceutical Biotechnology*, *22*. https://doi.org/10.2174/1389201022666210426092002

Amgarten, D., Braga, L. P. P., da Silva, A. M., & Setubal, J. C. (2018). MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. *Frontiers in Genetics*, *0*(AUG), 304. https://doi.org/10.3389/FGENE.2018.00304

Amgarten, D., Iha, B. K. V., Piroupo, C. M., Silva, A. M. da, & Setubal, J. C. (2020). vHULK, a new tool for bacteriophage host prediction based on annotated genomic features and deep neural networks. *BioRxiv*, 2020.12.06.413476. https://doi.org/10.1101/2020.12.06.413476

Bergh, Ø., Børsheim, K. Y., Bratbak, G., & Heldal, M. (1989). High abundance of viruses found in aquatic environments. *Nature*, *340*(6233), 467–468. https://doi.org/10.1038/340467a0

Besemer, J., Lomsadze, A., & Borodovsky, M. (2001). GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research*, *29*(12), 2607–2618. https://doi.org/10.1093/nar/29.12.2607

Botelho, J., Grosso, F., & Peixe, L. (2019). Antibiotic resistance in Pseudomonas aeruginosa – Mechanisms, epidemiology and evolution. *Drug Resistance Updates*, *44*, 100640. https://doi.org/10.1016/J.DRUP.2019.07.002

Braga, L. P. P., Soucy, S. M., Amgarten, D. E., da Silva, A. M., & Setubal, J. C. (2018). Bacterial diversification in the light of the interactions with phages: The genetic symbionts and their role in ecological speciation. *Frontiers in Ecology and Evolution*, *6*(JAN). https://doi.org/10.3389/fevo.2018.00006

Brives, C., & Pourraz, J. (2020). Phage therapy as a potential solution in the fight against AMR: obstacles and possible futures. *Palgrave Communications 2020 6:1*, *6*(1), 1–11. https://doi.org/10.1057/s41599-020-0478-4

Brum, J. R., Ignacio-Espinoza, J. C., Roux, S., Doulcier, G., Acinas, S. G., Alberti, A., Chaffron, S., Cruaud, C., Vargas, C. de, Gasol, J. M., Gorsky, G., Gregory, A. C., Guidi, L., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Poulos, B. T., … Sullivan, M. B. (2015). Patterns and ecological drivers of ocean viral communities. *Science*, *348*(6237). https://doi.org/10.1126/SCIENCE.1261498

Brüssow, H. (2005). Phage therapy: The Escherichia coli experience. In *Microbiology* (Vol. 151, Issue 7, pp. 2133–2140). Microbiology Society. https://doi.org/10.1099/mic.0.27849-0

Bruttin, A., & Brüssow, H. (2005). Human volunteers receiving Escherichia coli phage T4 orally: A safety test of phage therapy. *Antimicrobial Agents and Chemotherapy*, *49*(7), 2874–2878. https://doi.org/10.1128/AAC.49.7.2874-2878.2005

Chen, K., & Pachter, L. (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. In *PLoS Computational Biology* (Vol. 1, Issue 2, pp. 0106–0112). Public Library of Science. https://doi.org/10.1371/journal.pcbi.0010024

Coclet, C., & Roux, S. (2021). Global overview and major challenges of host prediction methods for uncultivated phages. *Current Opinion in Virology*, *49*, 117–126. https://doi.org/10.1016/J.COVIRO.2021.05.003

DeLong, E. F. (2009). The microbial ocean from genomes to biomes. In *Nature* (Vol. 459, Issue 7244, pp. 200–206). Nature Publishing Group. https://doi.org/10.1038/nature08059

Edwards, R. A., McNair, K., Faust, K., Raes, J., & Dutilh, B. E. (2016). Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiology Reviews*, *40*(2), 258–272. https://doi.org/10.1093/femsre/fuv048

Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The microbial engines that drive earth's biogeochemical cycles. In *Science* (Vol. 320, Issue 5879, pp. 1034–1039). American Association for the Advancement of Science. https://doi.org/10.1126/science.1153213

Gromiha, M. M., & Suresh, M. X. (2008). Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. *Proteins: Structure, Function and Genetics*, *70*(4), 1274–1279. https://doi.org/10.1002/prot.21616

Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, *68*(4), 669–685. https://doi.org/10.1128/mmbr.68.4.669-685.2004

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry and Biology*, *5*(10). https://doi.org/10.1016/S1074-5521(98)90108-9

Hyman, P. (2019). Phages for phage therapy: Isolation, characterization, and host range breadth. In *Pharmaceuticals* (Vol. 12, Issue 1, p. 35). Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/ph12010035

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2000). An introduction to Statistical Learning. In *Current medicinal chemistry* (Vol. 7, Issue 10). https://doi.org/10.1007/978-1-4614-7138-7

Jordan, T. C., Burnett, S. H., Carson, S., Caruso, S. M., Clase, K., DeJong, R. J., Dennehy, J. J., Denver, D. R., Dunbar, D., Elgin, S. C. R., Findley, A. M., Gissendanner, C. R., Golebiewska, U. P., Guild, N., Hartzog, G. A., Grillo, W. H., Hollowell, G. P., Hughes, L. E., Johnson, A., … Hatfull, G. F. (2014). A broadly implementable research course in phage discovery and genomics for first-year undergraduate students. *MBio*, *5*(1). https://doi.org/10.1128/mBio.01051-13

Khot, V., Strous, M., & Hawley, A. K. (2020). Computational approaches in viral ecology. *Computational and Structural Biotechnology Journal*, *18*, 1605–1612. https://doi.org/10.1016/J.CSBJ.2020.06.019

Kortright, K. E., Chan, B. K., Koff, J. L., & Turner, P. E. (2019). Phage Therapy: A Renewed Approach to Combat Antibiotic-Resistant Bacteria. *Cell Host & Microbe*, *25*(2), 219–232. https://doi.org/10.1016/J.CHOM.2019.01.014

Lagier, J. C., Dubourg, G., Million, M., Cadoret, F., Bilen, M., Fenollar, F., Levasseur, A., Rolain, J. M., Fournier, P. E., & Raoult, D. (2018). Culturing the human microbiota and culturomics. In *Nature Reviews Microbiology* (Vol. 16, Issue 9, pp. 540–550). Nature Publishing Group. https://doi.org/10.1038/s41579-018-0041-0

Lin, R. C., Sacher, J. C., Ceyssens, P. J., Zheng, J., Khalid, A., & Iredell, J. R. (2021). Phage Biobank: Present Challenges and Future Perspectives. In *Current Opinion in Biotechnology* (Vol. 68, pp. 221–230). Elsevier Current Trends. https://doi.org/10.1016/j.copbio.2020.12.018

McKenna, M. (2013). Antibiotic resistance: The last resort. *Nature*, *499*(7459), 394–396. https://doi.org/10.1038/499394a

Ohler, U., Liao, G. chun, Niemann, H., & Rubin, G. M. (2002). Computational analysis of core promoters in the Drosophila genome. *Genome Biology*, *3*(12), 1–12. https://doi.org/10.1186/gb-2002-3-12-research0087

Podlacha, M., Grabowski, Ł., Kosznik-Kawśnicka, K., Zdrojewska, K., Stasiłojć, M., Węgrzyn, G., & Węgrzyn, A. (2021). Interactions of Bacteriophages with Animal and Human Organisms—

Safety Issues in the Light of Phage Therapy. *International Journal of Molecular Sciences 2021, Vol. 22, Page 8937*, *22*(16), 8937. https://doi.org/10.3390/IJMS22168937

Rappé, M. S., & Giovannoni, S. J. (2003). The Uncultured Microbial Majority. In *Annual Review of Microbiology* (Vol. 57, pp. 369–394). Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA. https://doi.org/10.1146/annurev.micro.57.030502.090759

Ruppert, D. (2004). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Journal of the American Statistical Association*, *99*(466), 567–567. https://doi.org/10.1198/jasa.2004.s339

Rusk, N. (2009). Cheap third-generation sequencing. *Nature Methods*, *6*(4), 244–245. https://doi.org/10.1038/nmeth0409-244a

Schooley, R. T., Biswas, B., Gill, J. J., Hernandez-Morales, A., Lancaster, J., Lessor, L., Barr, J. J., Reed, S. L., Rohwer, F., Benler, S., Segall, A. M., Taplitz, R., Smith, D. M., Kerr, K., Kumaraswamy, M., Nizet, V., Lin, L., McCauley, M. D., Strathdee, S. A., … Hamilton, T. (2017). Development and use of personalized bacteriophage-based therapeutic cocktails to treat a patient with a disseminated resistant Acinetobacter baumannii infection. *Antimicrobial Agents and Chemotherapy*, *61*(10). https://doi.org/10.1128/AAC.00954-17

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. In *Nature Methods* (Vol. 5, Issue 1, pp. 16–18). Nature Publishing Group. https://doi.org/10.1038/nmeth1156

Solden, L., Lloyd, K., & Wrighton, K. (2016). The bright side of microbial dark matter: Lessons learned from the uncultivated majority. In *Current Opinion in Microbiology* (Vol. 31, pp. 217–226). Elsevier Current Trends. https://doi.org/10.1016/j.mib.2016.04.020

Thiel, K. (2004). Old dogma, new tricks - 21st Century phage therapy. In *Nature Biotechnology* (Vol. 22, Issue 1, pp. 31–36). Nature Publishing Group. https://doi.org/10.1038/nbt0104-31

Wright, A., Hawkins, C. H., Änggård, E. E., & Harper, D. R. (2009). A controlled clinical trial of a therapeutic bacteriophage preparation in chronic otitis due to antibiotic-resistant Pseudomonas aeruginosa; A preliminary report of efficacy. *Clinical Otolaryngology*, *34*(4), 349–357. https://doi.org/10.1111/j.1749-4486.2009.01973.x

Yu, G. X., Ostrouchov, G., Geist, A., & Samatova, N. F. (2003). An SVM-based algorithm for identification of photosynthesis-specific genome features. *Proceedings of the 2003 IEEE Bioinformatics Conference, CSB 2003*, 235–243. https://doi.org/10.1109/CSB.2003.1227323

# **Chapter 2**: MARVEL, a tool for prediction of viral sequences from metagenomic bins

## 2.1. Contextualization

The following chapter describes the development and test of the tool named MARVEL. An article describing this work was published in August 2018 at the peer-reviewed journal Frontiers in Genetics (Amgarten et al., 2018). As of October 2021, this article was cited in 45 publications according to Web of Sciences database, including several review articles.

The article is reproduced here with few adaptations and with the writing consent of all authors. The authors contributions are as follows: Deyvid Amgarten conceived, coded, and implemented the tool. Joao Setubal, Aline Maria da Silva, and Lucas Braga discussed the tool's design and experimental set-up and results, providing feedback that led to improvements. Deyvid Amgarten and Joao Setubal wrote the manuscript. All authors read, revised, and approved the final draft.

## 2.2. Background

In the past few decades, our understanding of microbial life has been profoundly changed by techniques of environmental sampling and high-throughput sequencing (DeLong, 2009; Handelsman, 2004; Rappé & Giovannoni, 2003). The uncultured majority of Bacteria and Archaea has been unveiled and so has the unknown universe of their viruses (Yutin et al., 2018). They are the most abundant biological entities, outnumbering all bacteria and archaea in the oceans by a factor of 10-100x (Bergh et al., 1989). Most environmental viruses infect bacterial hosts and are so termed bacteriophages or phages. They have been shown to be important drivers of biogeochemical cycles on Earth (Roux et al., 2016), as well as key players in directing and originating bacterial diversity (Braga et al., 2018; Falkowski et al., 2008; Koskella & Brockhurst, 2014).

Isolation is the gold standard for characterizing and assessing phage diversity and many new phages are isolated every year from diverse environments such as oceans, composting and human sewage (Amgarten et al., 2017; Kumari et al., 2009; Sullivan et al., 2003). However, isolation of viruses is constrained with the isolation of the hosts. This hinders the prospection of the prokaryotic virosphere, given the current lack of possibilities to cultivate the majority of the microbes (Solden et al., 2016). In this context, tools for mining viral sequences from huge datasets of culture-independent metagenomic sequencing reads and contigs are crucial. They have the potential to provide information about never reported phage genomes and genes, which are awaiting to be identified in metagenomes from environmental samples yet to be sequenced, or in the many datasets already public in databases (Rosario & Breitbart, 2011).

Machine learning is a field of computer sciences concerned with the development and application of computer algorithms that improve with experience (James et al., 2000; Ruppert, 2004). In other words, an algorithm can be trained to recognize a specific biological attribute once a list of features as example data is provided. Attributes to be predicted are commonly referred as labels in supervised learning, and it can be any biological attribute or phenotype of interest in the molecular biology context. A machine learning problem generally consist of trying to assign labels, given a new example and a list of features in which the algorithm was trained. In this case, features are measurements at the sequence level as for example, GC content, oligonucleotide frequency profiles, codon usage, gene expression, etc. An out of the box way of seeing genomic features for machine learning are categorical variables, such as presence or absence of specific genes or domains as applied to predict photosynthesis-specific genomic features by (Yu et al., 2003).

Concerning viral sequences as the target attribute to be predicted, two main tools have been reported in the literature. VirSorter is a tool for prediction viral contigs in metagenomic datasets, which uses alignments and similarity to a database of known viruses to recognize viral sequences (Roux et al., 2015). In a different approach, VirFinder uses a machine learning classifier for the same purpose, but in this case, k-mer frequency profiles (frequency of nucleotide words) are extracted from contigs and given as input to a previously trained model (Ren et al., 2017). Both tools present remarkable results that are helping to shed light into the viral dark matter (Hurwitz et

al., 2018; Nigro et al., 2017). However, true positive rates have been discussed as an issue for these tools, since they are very precise in their prediction, but may be losing many true viral sequences probably belonging to the viral dark matter.

Here we present the MARVEL tool for prediction of dsDNA phage genomes in metagenomic bins. MARVEL uses a core algorithm based on machine learning and three simple genomic features extracted from contig sequences. Besides, MARVEL also takes advantage of more robust information contained in bins (in contrast with single contigs) to improve prediction and output phage draft genomes. Therefore, unlike previous tools which are focused on finding contigs from fragmented viral genomes, MARVEL is focused on the discovery of virus genomes from metagenomic bins.

## 2.3. Methods

### 2.3.1. Pipeline implementation

MARVEL was coded in Python 3 and uses Prokka (Seemann, 2014) and HMMscan (Finn et al., 2011) as important dependencies. As input, MARVEL requires a directory with metagenomic bins in FASTA format; it then generates a results directory containing bins predicted as phages. An auxiliary script was made available to generate bins from Illumina paired-end raw reads using standard tools and methods (Breitwieser et al., 2018). Source code, datasets and use instructions are available in MARVEL's repository page for easy reproducibility: https://github.com/laboratoriobioinformatica/MARVEL.

### 2.3.2. Training and testing datasets

To build and test MARVEL, the RefSeq microbial dataset was downloaded (January 2018) and only genomes belonging to the Bacteria domain (NCBI txid: 2) and dsDNA viruses from the *Caudovirales* order (NCBI txid:28883) were selected (this is the baseline dataset). Tailed phages were selected at this step as a representative group given that they constitute the majority of viruses present in most environmental samples (Ackermann, 2007; Ashelford et al., 2003; Filée et al., 2005). The baseline dataset was split into two subsets according to the GenBank deposit date: before January 2016; and January 2016 and thereafter. This time-based division is usually

applied in classifiers to simulate the use of the tool on newly isolated sequences (Ren et al., 2017; Roux et al., 2015).

Training dataset has 1247 phage genomes and 1029 bacterial complete genomes. Testing dataset has 335 bacterial genomes and 177 phage genomes. Training and testing datasets have no overlap and are available in MARVEL's repository page.

Training and testing datasets were further processed to generate mock datasets of contigs with specific lengths. For each fragment length analyzed in this study (2 kbp, 4 kbp, 8 kbp, 12 kbp and 16 kbp), complete genomes were randomly fragmented in 10 contigs of the specified length that may or may not have overlap. Next, contigs belonging to the same organism were clustered to form a simulated bin. This process was performed for both training and testing sets, and the resulting bins were used to train the machine learning algorithm, to assess MARVEL's performance, and to compare MARVEL against VirSorter (Roux et al., 2015) and VirFinder (Ren et al., 2017), which are tools for similar prediction of viral sequences.

### 2.3.3. Feature extraction and classifier development

As previous studies have shown, genomic features such as DNA k-mer profiles and GC content can be strong signals in linking or differentiating genome sequences from bacteria and viruses (Edwards et al., 2016; Ren et al., 2017). However, it is known that phages try to mimic host genome sequences in order to overcome their defenses (Bahir et al., 2009; Carbone, 2008). This causes classifiers based on k-mer frequencies to have poor performance in terms of overall accuracy and particularly recall. In other words, when one of these classifiers identifies a phage genome, it is almost always correct, but it is likely to miss many new phages present in environmental samples.

Seeking more robust features, we focused our efforts on characteristics related to genome structure and protein translational mechanisms of each organism. Such characteristics require a second layer of information, which may be added by utilization of results from gene prediction programs, such as Prodigal (Hyatt et al., 2010) and GeneMark (Besemer et al., 2001). Therefore, we evaluated phage and bacterial

genomes according to six of these genomic features extracted from the baseline dataset of RefSeq complete genomes.

These six features are: average gene length, average spacing between genes, density of genes, frequency of strand shifts between neighboring genes, ATG relative frequency, and fraction of genes with significant hits against the pVOGs database (Grazziotin et al., 2017). Average gene length was computed by adding up the length of all predicted CDSs in the genome or in the contigs in a bin (in bp) divided by the total number of predicted CDSs. Average spacing was calculated as the mean length in bp of regions between two CDSs. Density of genes was calculated as the total number of CDSs divided by genome length measured in kbp. Frequency of strand shifts was computed by adding up the number of strand shifts between neighboring genes, and dividing by the total number of CDSs in the genome. ATG relative frequency was computed by counting the number of ATG triplets in one of the strands, in all contigs in a bin or in the complete genome, divided by the total number of 3-mers in that sequence (one strand). Finally, each CDS in a genome was searched using HMMscan (Eddy, 2011) against the pVOGs database of viral HMM profiles (downloaded in January 2018); a significant hit was noted when the e-value was less than or equal to $10^{-10}$. The number of significant hits was divided by the total number of CDSs to generate the fraction of genes with significant hits against the pVOGs database. All values based on predicted CDSs were extracted from GenBank files as available for download in January 2018 (exploratory step) or predicted in simulated fragments by Prodigal as driven by Prokka.

Using Python Scikit Learn libraries, we tried different machine learning approaches based on the six features listed above. Specifically: Support Vector Machine (SVM), logistic regression, neural networks, and random forest. Classifiers were evaluated using the training set and k-fold cross-validation (k = 20), with the result that random forest was the best approach for our target prediction. Similar findings about suitability of random forest classifiers in bioinformatics have also been reported (Boulesteix et al., 2012; Zhang et al., 2017).

The ID3 implementation of the random forest technique performs an initial step of feature selection by measuring gain of information for each of the features tested, resulting in a clean model containing only the most relevant features for the target

prediction (Quinlan, 1986). In our case, the following features were selected as more informative: gene density, strand shifts, and fraction of genes with significant hits against pVOGs database (more details in the Preliminary Results section). We then extracted these three informative features from a complete training set of 8 kbp simulated bins, and a random forest classifier was trained to be MARVEL's prediction core. The random forest model was trained with 50 initial tree estimators and leaf pruning; other parameters were set to their default values.

### 2.3.4. Tests with simulated metagenomic bins

Simulated bins containing different contig sizes were generated for genomes of the testing set as previously described, to assess MARVEL's performance. Each test corresponding to a specific fragment length was performed in five randomly sampled replicates of 150 bins (75 bacteria and 75 dsDNA phages). Bins were submitted to MARVEL and predictions were evaluated for true positive rates, specificity, accuracy, and F1 scores.

### 2.3.5. Performance comparison of MARVEL, Virsorter, and Virfinder

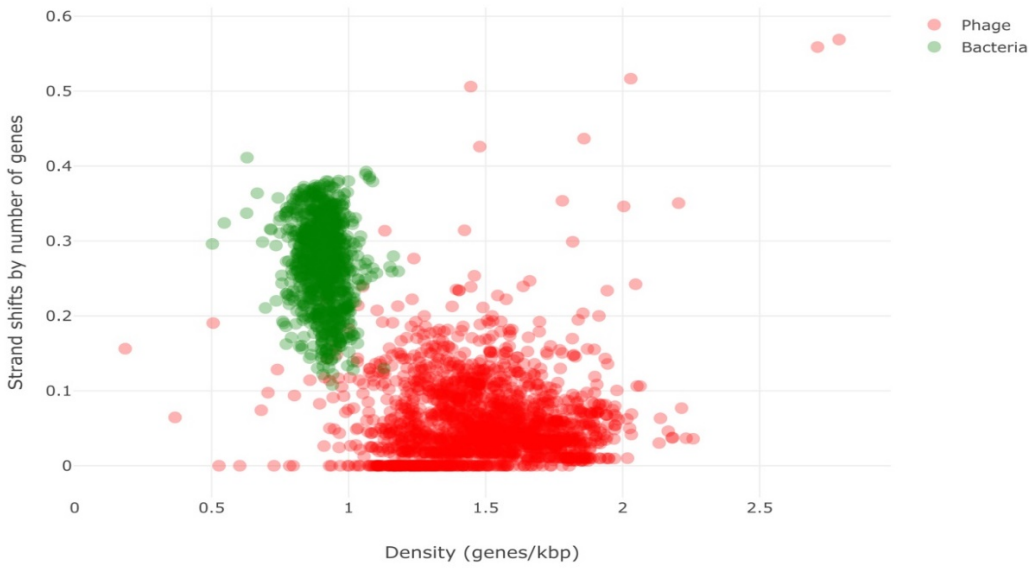
Each contig of a simulated bin (ten contigs in total) was individually given as input to VirSorter and VirFinder. For a given tool, an entire bin was considered to be a positive prediction in case at least one of its contigs were predicted as viral (note that in our experimental set-up, there are no bins with both bacterial and viral sequences). A contig was considered viral if predicted in categories I and II for VirSorter, and if the q-value was less than or equal to 0.01 for VirFinder. Tests were performed for different fragment lengths and in 30 randomly sampled replicates of 100 bins (50 bacteria and 50 dsDNA phages). Average values of true positive rate, specificity, and accuracy were compared using the Wilcoxon signed-rank test and were considered significant if the p-value was less than 0.001.

## 2.4. Results

### 2.4.1. Three simple genomic features yield accurate predictions of phage genomes

As mentioned in Methods section, we tested six different genomic features; the three best features for our target prediction were gene density, strand shifts, and fraction of significant pVOGs hits. Figure 2.1 shows the results for two of these features on the baseline dataset; numerical results for all three features are shown in Table 2.1.

**Figure 2.1:** Scatter plot of bacterial and phage genomes using two of the three features as axes: Strand shifts by total number of genes and density of genes. Green and red dots represent bacterial and phage genomes, respectively.




Note: Own work.

**Table 2.1**: Average and standard deviation values obtained by extracting features from the training dataset of dsDNA phage and bacterial genomes.

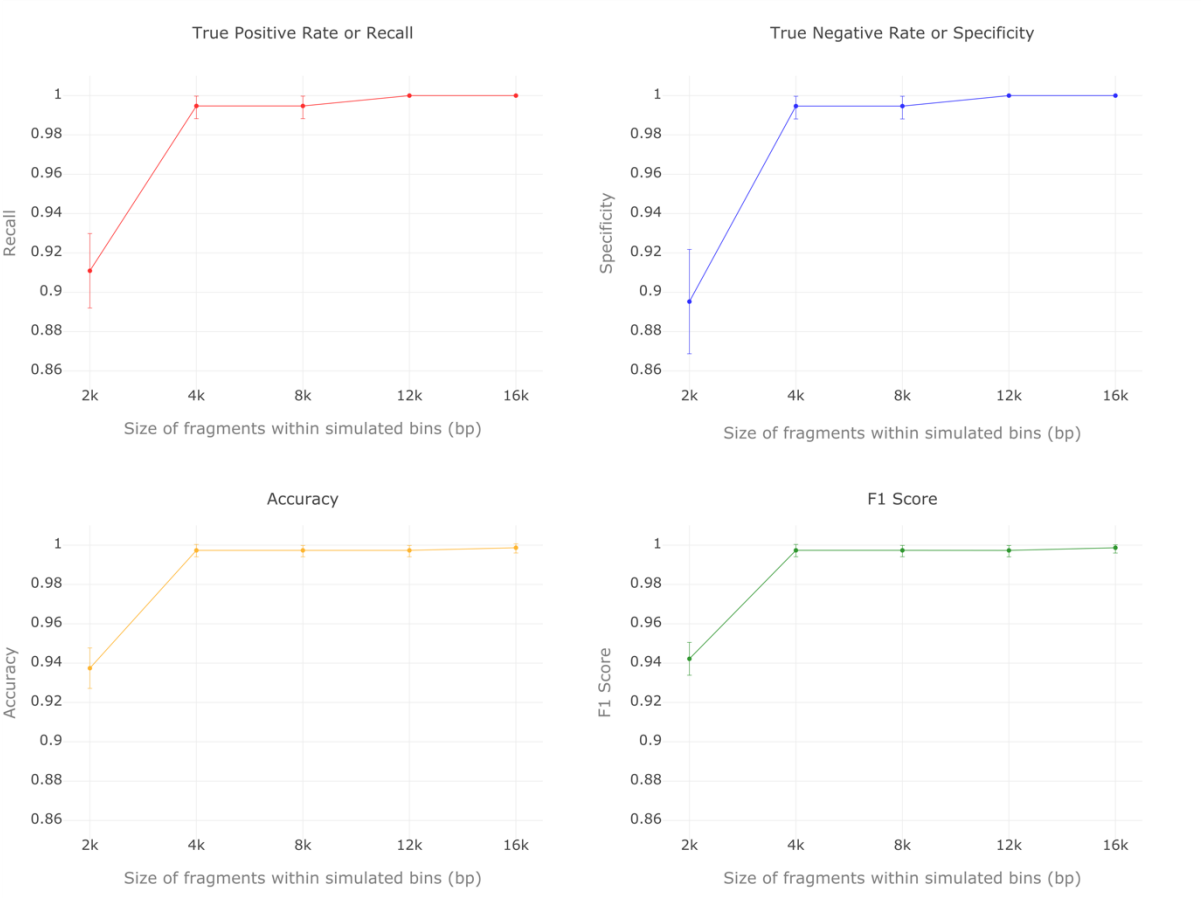
| | **Gene density (genes by kbp)** | **Strand shifts by total of genes** | **Fraction of pVOGs significant hits** |
|---|---|---|---|
| **Phage** | 1.44 (± 0.27) | 0.07 (± 0.05) | 0.68 (± 0.2) |
| **Bacteria** | 0.93 (±0.13) | 0.24 (± 0.08) | 0.1 (± 0.04) |

Note: Own work.

### 2.4.2. Assessing MARVEL prediction performance with simulated bins from known dsDNA phage and bacterial genomes

Simulated bins were randomly subsampled and given as input to MARVEL in five replicates. Predictions were performed for each simulated bin and results are shown in Figure 2.2. MARVEL has high F1 scores and accuracy for all sizes of bins analyzed, but especially for bins composed of contigs larger than 4 kbp (bins composed of 2 kbp contigs present a clear limitation in performance for reasons we explore in the Discussion section). True positive rates (recall) were particularly high for all fragment lengths. It is noteworthy that other tools struggle to obtain good recall values, as opposed to precision, for which the performance is usually very good (Ren et al., 2017; Roux et al., 2015). Altogether, these performance results in simulated data suggest that MARVEL is effectively able to predict dsDNA phage genomes in metagenomic bins.

**Figure 2.2:** MARVEL's performance in simulated bins obtained from the testing set of RefSeq genomes. Recall, specificity, accuracy and F1 score are shown for bins composed of different contig lengths.

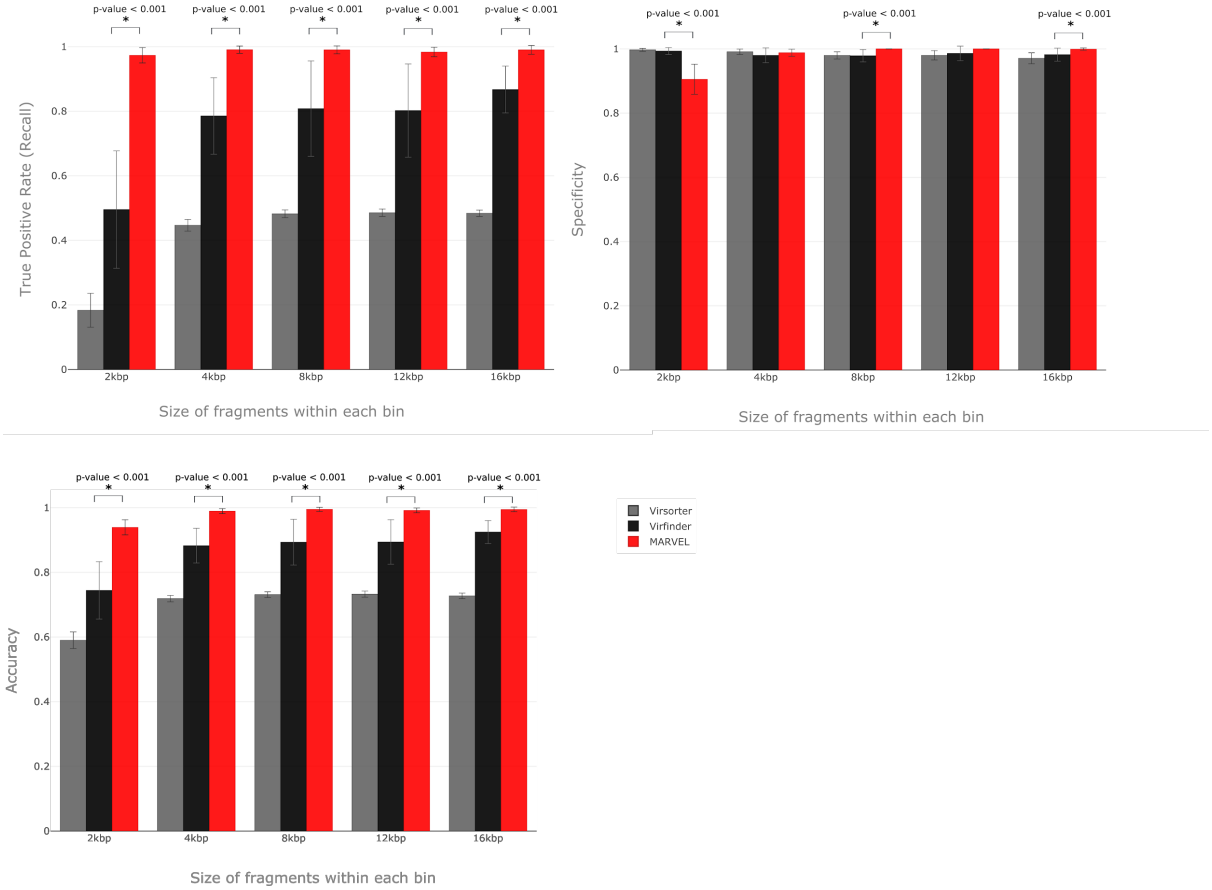



Note: Own work.

## 2.4.3. Performance Comparison of MARVEL, Virsorter, and Virfinder

All three tools have comparable results for specificity in most of the fragment lengths studied, except for 2 kbp fragments (Figure 2.3). Nevertheless, MARVEL's true positive rates were significantly higher in all cases (p-value < 0.001), which has been the main difficulty for viral sequence classifiers. MARVEL´s better true positive rates resulted in better overall accuracy compared to the two other tools in all scenarios.

**Figure 2.3:** Performance comparison of MARVEL, VirSorter and VirFinder. Means were compared using Wilcoxon signed-rank test. Standard deviation of 30 replicates is shown by error bars.




Note: Own work.

## 2.5. Discussion

To our knowledge, MARVEL is the first tool able to effectively separate metagenomic bins containing dsDNA phage sequences from those containing bacterial sequences. By doing this, it facilitates downstream metagenomic analyses aiming to characterize phage phylogenetic and functional diversity. VirSorter and VirFinder are two excellent tools optimized to analyze single contigs. It would be possible to use these tools in a pipeline to generate whole bin predictions, but in our estimation, this would perhaps require substantial additional work. Furthermore, we present results in simulated data showing significantly better true positive rates and accuracy for MARVEL's predictions. These improvements were achieved by the implementation and use of three specific genomic features, which we show as highly suitable for viral sequences prediction.

Results presented in this article suggest that higher gene density and lower rates of strand shift are important phage genomic hallmarks when compared with bacterial genomes (Figure 2.1). The length of phage genomes is physically constrained by the size of the capsid, which imposes a limited space for genes in the genome (Chirico et al., 2010), favoring selective pressure for increased gene density. Furthermore, dsDNA phage genomes present lower rates of strand shift than bacterial genomes. This could be translated as higher strand symmetry among codified genes, which may favor rapid and regulated transcription/translation of early, middle, and late genes of phage's infectious cycle (Miller et al., 2003; Mrázek & Karlin, 1998). Altogether, evidence supporting very compact phage genomes have also been reported by previously studies (Mahmoudabadi & Phillips, 2018; O'Connell, 2005; Roux et al., 2015).

Our results in simulated datasets suggest a performance limitation for bins composed of contigs with 2 kbp or less in length (Figure 2.2). This is a very clear limitation, since MARVEL uses CDS predictions as primary information in all three features on which our phage predictions are based. Sequences too short will contain very few or no CDSs, and at least two CDSs are required for calculating the features gene density and frequency of strand shifts. On the other hand, reports in the literature indicate that viral bins are often composed of large contigs, and in some cases contain almost complete viral genomes (Dutilh et al., 2014; Paez-Espino et al., 2017), suggesting that this limitation may not be a hindrance.

Upstream processing such as assembly and binning are two major factors that also influence MARVEL's performance. Chimeric contigs, as well as poorly binned bins may generate noisy data, which will certainly increase the number of erroneous predictions. Therefore, it is important to choose thresholds and parameter values to ensure quality of upstream processing (Mavromatis et al., 2007; Roux et al., 2017). There are several tools available for assembly and binning which generate good quality contigs and bins (Kang et al., 2015; Li et al., 2016; Nurk et al., 2017; Wu et al., 2016). We emphasize, however, that assessing quality of viral bins is not an easy task. CheckM is a tool for assessing marker genes, contamination, and completeness of metagenomic bins, but unfortunately only bacterial and archaeal datasets of marker genes are available (Parks et al., 2015).

In its present incarnation, as shown here, MARVEL is able to effectively predict tailed phages of the *Caudovirales* order only. Tailed phages are the majority of viruses present in most environmental samples, and we believe this reason justifies our choice (Ackermann, 2007; Ashelford et al., 2003; Filée et al., 2005). On the other hand, the features that we used for predictions in this work may not be as effective for viruses in general (Mahmoudabadi & Phillips, 2018). This may be one reason why recall rates in our tests were lower for VirSorter and VirFinder as compared to MARVEL, since those other tools are generic viral sequence finders.

We believe an effective generic viral model would be hard to achieve, given the heterogeneity of viral types and genome structures. Nevertheless, it is our intention to expand MARVEL's scope to predict other groups of viruses, by obtaining additional models specific to other major viral groups. Such models would be available to users as parameter choices in future versions of MARVEL. The program was designed with this objective in mind.

## Software availability:

The MARVEL tool, documentation, usage examples, and training and testing datasets are freely available in through an online repository: https://github.com/LaboratorioBioinformatica/MARVEL .

## Acknowledgments

We are grateful to Dr. Luiz Thiberio Rangel for providing helpful coding guidance and discussions to this project. We thank Dr. Melline Fontes Noronha for testing the tool and providing constant feedback to improvement. We also thank Carlos Morais for all technical support. DA was supported by a fellowship grant #2014/16450-8, São Paulo Research Foundation (FAPESP). DA and LB were supported by a fellowship from the Coordination for the Improvement of Higher Education Personnel (CAPES). JS and AS wish to acknowledge their respective research fellowships from CNPq. This work was supported in part by FAPESP grant 2011/50870-6.

# References


Ackermann, H. W. (2007). 5500 Phages examined in the electron microscope. In *Archives of Virology* (Vol. 152, Issue 2, pp. 227–243). Springer. https://doi.org/10.1007/s00705-006-0849-1

Amgarten, D., Braga, L. P. P., da Silva, A. M., & Setubal, J. C. (2018). MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. *Frontiers in Genetics*, *0*(AUG), 304. https://doi.org/10.3389/FGENE.2018.00304

Amgarten, D., Martins, L. F., Lombardi, K. C., Antunes, L. P., de Souza, A. P. S., Nicastro, G. G., Kitajima, E. W., Quaggio, R. B., Upton, C., Setubal, J. C., & da Silva, A. M. (2017). Three novel Pseudomonas phages isolated from composting provide insights into the evolution and diversity of tailed phages. *BMC Genomics*, *18*(1), 1–18. https://doi.org/10.1186/s12864-017-3729-z

Ashelford, K. E., Day, M. J., & Fry, J. C. (2003). Elevated abundance of bacteriophage infecting bacteria in soil. *Applied and Environmental Microbiology*, *69*(1), 285–289. https://doi.org/10.1128/AEM.69.1.285-289.2003

Bahir, I., Fromer, M., Prat, Y., & Linial, M. (2009). Viral adaptation to host: A proteome-based analysis of codon usage and amino acid preferences. *Molecular Systems Biology*, *5*(1), 311. https://doi.org/10.1038/msb.2009.71

Bergh, Ø., Børsheim, K. Y., Bratbak, G., & Heldal, M. (1989). High abundance of viruses found in aquatic environments. *Nature*, *340*(6233), 467–468. https://doi.org/10.1038/340467a0

Besemer, J., Lomsadze, A., & Borodovsky, M. (2001). GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research*, *29*(12), 2607–2618. https://doi.org/10.1093/nar/29.12.2607

Boulesteix, A. L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(6), 493–507. https://doi.org/10.1002/widm.1072

Braga, L. P. P., Soucy, S. M., Amgarten, D. E., da Silva, A. M., & Setubal, J. C. (2018). Bacterial diversification in the light of the interactions with phages: The genetic symbionts and their role in ecological speciation. *Frontiers in Ecology and Evolution*, *6*(JAN). https://doi.org/10.3389/fevo.2018.00006

Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2018). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, *20*(4), 1125–1139. https://doi.org/10.1093/bib/bbx120

Carbone, A. (2008). Codon bias is a major factor explaining phage evolution in translationally biased hosts. *Journal of Molecular Evolution*, *66*(3), 210–223. https://doi.org/10.1007/s00239-008-9068-6

Chirico, N., Vianelli, A., & Belshaw, R. (2010). Why genes overlap in viruses. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1701), 3809–3817. https://doi.org/10.1098/rspb.2010.1052

DeLong, E. F. (2009). The microbial ocean from genomes to biomes. In *Nature* (Vol. 459, Issue 7244, pp. 200–206). Nature Publishing Group. https://doi.org/10.1038/nature08059

Dutilh, B. E., Cassman, N., McNair, K., Sanchez, S. E., Silva, G. G. Z., Boling, L., Barr, J. J., Speth, D. R., Seguritan, V., Aziz, R. K., Felts, B., Dinsdale, E. A., Mokili, J. L., & Edwards, R. A. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human

faecal metagenomes. *Nature Communications*, *5*(1), 1–11. https://doi.org/10.1038/ncomms5498
Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195
Edwards, R. A., McNair, K., Faust, K., Raes, J., & Dutilh, B. E. (2016). Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiology Reviews*, *40*(2), 258–272. https://doi.org/10.1093/femsre/fuv048
Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The microbial engines that drive earth's biogeochemical cycles. In *Science* (Vol. 320, Issue 5879, pp. 1034–1039). American Association for the Advancement of Science. https://doi.org/10.1126/science.1153213
Filée, J., Tétart, F., Suttle, C. A., & Krisch, H. M. (2005). Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(35), 12471–12476. https://doi.org/10.1073/pnas.0503404102
Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *39*(SUPPL. 2). https://doi.org/10.1093/nar/gkr367
Grazziotin, A. L., Koonin, E. v., & Kristensen, D. M. (2017). Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Research*, *45*(D1), D491–D498. https://doi.org/10.1093/NAR/GKW975
Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, *68*(4), 669–685. https://doi.org/10.1128/mmbr.68.4.669-685.2004
Hurwitz, B. L., Ponsero, A., Thornton, J., & U'Ren, J. M. (2018). Phage hunters: Computational strategies for finding phages in large-scale 'omics datasets. In *Virus Research* (Vol. 244, pp. 110–115). Elsevier. https://doi.org/10.1016/j.virusres.2017.10.019
Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 1–11. https://doi.org/10.1186/1471-2105-11-119
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2000). An introduction to Statistical Learning. In *Current medicinal chemistry* (Vol. 7, Issue 10). https://doi.org/10.1007/978-1-4614-7138-7
Kang, D. D., Froula, J., Egan, R., & Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, *2015*(8), e1165. https://doi.org/10.7717/PEERJ.1165/SUPP-2
Koskella, B., & Brockhurst, M. A. (2014). Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiology Reviews*, *38*(5), 916–931. https://doi.org/10.1111/1574-6976.12072
Kumari, S., Harjai, K., & Chhibber, S. (2009). Characterization of Pseudomonas aeruginosa PAO Specific Bacteriophages Isolated from Sewage Samples. *American Journal of Biomedical Sciences*, *1*(2), 91–102. https://doi.org/10.5099/aj090200091
Li, D., Luo, R., Liu, C. M., Leung, C. M., Ting, H. F., Sadakane, K., Yamashita, H., & Lam, T. W. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, *102*, 3–11. https://doi.org/10.1016/J.YMETH.2016.02.020
Mahmoudabadi, G., & Phillips, R. (2018). A comprehensive and quantitative exploration of thousands of viral genomes. *ELife*, *7*. https://doi.org/10.7554/ELIFE.31955
Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A. C., Rigoutsos, I., Salamov, A., Korzeniewski, F., Land, M., Lapidus, A., Grigoriev, I., Richardson, P.,

Hugenholtz, P., & Kyrpides, N. C. (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods*, *4*(6), 495–500. https://doi.org/10.1038/nmeth1043

Miller, E. S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T., & Rüger, W. (2003). Bacteriophage T4 Genome. *Microbiology and Molecular Biology Reviews*, *67*(1), 86–156. https://doi.org/10.1128/mmbr.67.1.86-156.2003

Mrázek, J., & Karlin, S. (1998). Strand compositional asymmetry in bacterial and large viral genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(7), 3720–3725. https://doi.org/10.1073/pnas.95.7.3720

Nigro, O. D., Jungbluth, S. P., Lin, H. T., Hsieh, C. C., Miranda, J. A., Schvarcz, C. R., Rappé, M. S., & Steward, G. F. (2017). Viruses in the oceanic basement. *MBio*, *8*(2). https://doi.org/10.1128/mBio.02129-16

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. https://doi.org/10.1101/GR.213959.116

O'Connell, D. (2005). Small is beautiful. In *Nature Reviews Microbiology* (Vol. 3, Issue 7, p. 520). Nature Publishing Group. https://doi.org/10.1038/nrmicro1196

Paez-Espino, D., Chen, I. M. A., Palaniappan, K., Ratner, A., Chu, K., Szeto, E., Pillay, M., Huang, J., Markowitz, V. M., Nielsen, T., Huntemann, M., Reddy, T. B. K., Pavlopoulos, G. A., Sullivan, M. B., Campbell, B. J., Chen, F., McMahon, K., Hallam, S. J., Denef, V., … Kyrpides, N. C. (2017). IMG/VR: A database of cultured and uncultured DNA viruses and retroviruses. *Nucleic Acids Research*, *45*(D1), D457–D465. https://doi.org/10.1093/nar/gkw1030

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055. https://doi.org/10.1101/GR.186072.114

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/bf00116251

Rappé, M. S., & Giovannoni, S. J. (2003). The Uncultured Microbial Majority. In *Annual Review of Microbiology* (Vol. 57, pp. 369–394). Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA. https://doi.org/10.1146/annurev.micro.57.030502.090759

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, *5*(1), 69. https://doi.org/10.1186/s40168-017-0283-5

Rosario, K., & Breitbart, M. (2011). Exploring the viral world through metagenomics. In *Current Opinion in Virology* (Vol. 1, Issue 4, pp. 289–297). Elsevier. https://doi.org/10.1016/j.coviro.2011.06.004

Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., Poulos, B. T., Solonenko, N., Lara, E., Poulain, J., Pesant, S., Kandels-Lewis, S., Dimier, C., Picheral, M., Searson, S., Cruaud, C., Alberti, A., Duarte, C. M., Gasol, J. M., … Sullivan, M. B. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, *537*(7622), 689–693. https://doi.org/10.1038/nature19366

Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A., & Sullivan, M. B. (2017). Benchmarking viromics: An in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ*, *2017*(9), e3817. https://doi.org/10.7717/peerj.3817

Roux, S., Enault, F., Hurwitz, B. L., & Sullivan, M. B. (2015). VirSorter: Mining viral signal from microbial genomic data. *PeerJ*, *2015*(5), e985. https://doi.org/10.7717/peerj.985

Ruppert, D. (2004). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Journal of the American Statistical Association*, *99*(466), 567–567. https://doi.org/10.1198/jasa.2004.s339

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Solden, L., Lloyd, K., & Wrighton, K. (2016). The bright side of microbial dark matter: Lessons learned from the uncultivated majority. In *Current Opinion in Microbiology* (Vol. 31, pp. 217–226). Elsevier Current Trends. https://doi.org/10.1016/j.mib.2016.04.020

Sullivan, M. B., Waterbury, J. B., & Chisholm, S. W. (2003). Cyanophages infecting the oceanic cyanobacterium Prochlorococcus. *Nature*, *424*(6952), 1047–1051. https://doi.org/10.1038/nature01929

Wu, Y. W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, *32*(4), 605–607. https://doi.org/10.1093/BIOINFORMATICS/BTV638

Yu, G. X., Ostrouchov, G., Geist, A., & Samatova, N. F. (2003). An SVM-based algorithm for identification of photosynthesis-specific genome features. *Proceedings of the 2003 IEEE Bioinformatics Conference, CSB 2003*, 235–243. https://doi.org/10.1109/CSB.2003.1227323

Yutin, N., Bäckström, D., Ettema, T. J. G., Krupovic, M., & Koonin, E. v. (2018). Vast diversity of prokaryotic virus genomes encoding double jelly-roll major capsid proteins uncovered by genomic and metagenomic sequence analysis. *Virology Journal*, *15*(1), 1–17. https://doi.org/10.1186/s12985-018-0974-y

Zhang, M., Yang, L., Ren, J., Ahlgren, N. A., Fuhrman, J. A., & Sun, F. (2017). Prediction of virus-host infectious association by supervised learning methods. *BMC Bioinformatics*, *18*(3), 143–154. https://doi.org/10.1186/s12859-017-1473-7

# **Chapter 3**: vHULK, Accurate Identification of Hosts from Environmental Viruses Using Artificial Neural Networks and High-Level Features

## 3.1. Contextualization

The following chapter describes the development of the tool named vHULK. An article describing this work was published as pre-print in December of 2020 at bioRxiv (Amgarten et al., 2020). As of October 2021, this article was cited in 2 publications according to Google Scholar database, including a review article in Current Opinion in Virology (Coclet & Roux, 2021).

The article is reproduced here with some adaptations and with the writing consent of all authors. The authors contributions are as follows: Deyvid Amgarten conceived the tool, prepared datasets, and engineered features consumed by the model. Deyvid Amgarten and Bruno Iha implemented the Artificial Neural Networks (ANNs), trained and tested models and compared vHULK's results with existing tools. Carlos Morais Piroupo provided technical support and computational infrastructure for model training and dataset preparation. Joao Setubal and Aline Maria da Silva discussed the tool's design and experimental set-up and results, providing feedback that led to improvements. Deyvid Amgarten and Joao Setubal wrote the manuscript. All authors read, revised, and approved the final draft.

## 3.2. Background

Viruses are believed to be by far the most diverse and abundant biological entities in the biosphere (Breitbart & Rohwer, 2005). They are present in virtually any place or environment as long as a host is nearby. Acquisition of new knowledge about viruses has been restricted by our capacity of sampling the environment and isolating new strains through cultivation-based methods. Recent improvements in Next Generation Sequencing techniques and the rise of metagenomics as a prominent field of study have given us access to unprecedented genetic diversity. Researchers now have at

their disposal thousands of DNA sequence datasets from different types of environmental or host-associated samples, such as ocean, soil, and animal guts (Breitbart et al., 2018; Emerson, 2019; Reyes et al., 2012). These datasets are helping us to shed light on the so-called viral dark matter (Carroll et al., 2018; Martínez-García et al., 2014; Paez-Espino et al., 2016). Recently, there have been efforts to consolidate many of these datasets in general databases (Gregory et al., 2020; Roux et al., 2021), thereby facilitating investigation of the viral taxonomic and genomic diversity, as well as virus' roles in specific habitats (Brum et al., 2015).

Bacteriophages (or simply phages) are a substantial part of the global virome, and as such their genomes are also being sampled at a rapid pace (Koonin et al., 2020; Paez-Espino et al., 2016). Phages have achieved prominence in the last few years thanks to phage therapy, which is seen by many as a viable resource against multidrug resistant bacteria (Kortright et al., 2019). A key step in phage therapy development is finding a phage that can infect and efficiently terminate with high specificity an infectious agent of interest. Establishing a link between phage and host is however far from trivial, since phages may present a variety of infecting molecular mechanisms, and a host range that varies from narrow (strain-specific) to very broad (hosts from a same genus, family, or higher taxonomic levels) (Chaturongakul & Ounjai, 2014; de Jonge et al., 2019; Samson et al., 2013). Furthermore, host range characterization mostly depends on laborious wet-lab experimental methods and may be very specific for species or even bacterial strains. As examples of such techniques, we may cite the use of fosmids, viral tagging with fluorescent die, and phageFISH (Allers et al., 2013; Chow et al., 2015; Deng et al., 2012). Because of the laboriousness and specificity of these methods, effective computational prediction of phage hosts should be of great help.

R. Edwards and colleagues have reviewed genomic signals used to link a phage to its host (Edwards et al., 2016). Oligonucleotide frequency profiles, sequence similarity, and CRISPR-spacers matches were evaluated. Sequence similarity between phage and host genome, as well as matches of CRISPR-spacers were found to be more effective in establishing phage-host pairs. However, only 40% accuracy was achieved with these features. Since then, a several other tools have become available. Tools such as PHASTER (Arndt et al., 2016), PhiSpy (Akhter et al., 2012), Phage_Finder (Fouts, 2006), and Prophinder (Lima-Mendez et al., 2008) focus on

finding phages integrated in a host genome (prophages). Despite establishing a clear phage-host relationship, these approaches are restricted to temperate phages that are integrated in their host genome at the moment of sampling.

More general approaches have been suggested, such as the use of Markov models based on k-mer frequencies implemented in the WIsH tool (Galiez et al., 2017). According to its authors, WIsH has 63% of mean accuracy when predicting host genus among 20 possible genera. Leite et al. (Leite et al., 2018) report predictive performance of approximately 90% for several metrics, using a machine learning approach. A more recent work proposed the use of neural networks based on an ensemble of distance measurements and other features, concepts that were implemented in the tool VirHostMatcher-net (VHM-net) (Wang et al., 2020). VHM-net achieved accuracies that varied from 43% to 59% at the host genus level only. VHM-net is theoretically able to predict thousands of different prokaryote hosts. The most recent program in this field is Random Forest Assignment of Hosts (RaFAH), which has been shown to outperform previous virus-host prediction methods at the phylum level (Coutinho et al., 2021).

In this work, we present a novel approach for phage host prediction based on a unique representation of annotated genomic features and neural networks. Features are extracted from phage genomes and searched against a protein database, which provides information about genes and their function. These features are converted to a pixel-like matrix and fed to multi-layer perceptron models (MLP) trained to predict 52 and 61 different host species and host genera, respectively. Trained models were implemented in a user-friendly tool named vHULK, which stands for Viral Host UnveiLing Kit. vHULK's predictions were evaluated on a dataset that simulates a real-life use case, and compared with CRISPR-spacers, VHM-net, and RaFAH. Using vHULK we obtained better performance than these other programs at both the genus and species levels.

## 3.3. Methods

### 3.2.1. Datasets

In March, 2019, we searched the Genbank nucleotide database for virus records with the following query: 'taxid:10239 AND Sequence_type:complete'. A total of 314,536

records in format .gb were downloaded and further processed for quality control. Records without features 'host', 'CDS', 'Translation', or containing any format error were removed. The field "host" was checked and records without Linnaeus binomial species name as a host were also filtered out (e.g., diarrhea cat, pig, swine, chicken). Only records whose host was in domains Bacteria or Archaea were selected at this point. A total of 8,616 records passed these quality control steps, resulting in what we call *baseline dataset*.

In order to remove redundancy in terms of presence of similar genomes in the baseline dataset, we calculated a pairwise matrix of AAI scores using the CompareM AAI_wf tool (https://github.com/dparks1134/CompareM). We parsed AAI scores matrices to generate three *AAI redundancy reduction datasets* for each prediction type (host species and host genus) as follows: Instances with AAI scores equal to or higher than 90/80/70% were removed and only one randomly chosen instance from each group was left in the AAI redundancy reduction dataset. The AAI redundancy reduction dataset was randomly split into training (for neural network fine-tuning purposes) and test sets in an 80/20 fashion, that is: 80% of the dataset was used to train the neural network and 20% to evaluate its predictions. Performance measurements were averaged for 10 random splits of the training/test datasets. Each random split test dataset had 1,414 phage genomes assessed for the genus model and 1,131 phage genomes assessed for the species model. We used these performance results to fine-tune the following neural network parameters: the number of layers, number of neurons in each layer, the activation function used in the hidden layers and in the regularizations of the neural network, and early stopping parameters to avoid overfitting.

With the neural network properly configured, we were able to generate two predictors from the baseline dataset: one for prediction of host species and another for prediction of host genus. Each of these models required its own training set; we call them the *species training set* and the *genus training set*. For adequate model training, for each host species or host genus to be predicted (i.e., to define a target), we needed minimum numbers of training instances. Thus, for a given species to be included as a target, we required the existence of at least 20 phage records with that host species; for genus, this requirement was at least 12 records. As a result, the dataset for training the genus model contained 61 targets and 7,993 instances (phage genomes), and the

species model contained 52 targets for prediction and 6,432 instances (Table 3.1). These datasets were used for final model training.

**Table 3.1.** Target hosts and number of instances in the final training set for each model type (species and genus).

| *Identifier* | *Target species* | *Number of instances* | *Target genus* | *Number of instances* |
|---|---|---|---|---|
| *1* | *Mycobacterium smegmatis* | 1553 | *Mycobacterium* | 1612 |
| *2* | *Escherichia coli* | 992 | *Escherichia* | 1004 |
| *3* | *Lactococcus lactis* | 294 | *Pseudomonas* | 392 |
| *4* | *Pseudomonas aeruginosa* | 289 | *Salmonella* | 345 |
| *5* | *Gordonia terrae* | 249 | *Lactococcus* | 309 |
| *6* | *Salmonella enterica* | 232 | *Streptococcus* | 294 |
| *7* | *Staphylococcus aureus* | 215 | *Bacillus* | 284 |
| *8* | *Streptococcus thermophilus* | 166 | *Gordonia* | 283 |
| *9* | *Propionibacterium acnes* | 147 | *Staphylococcus* | 277 |
| *10* | *Klebsiella pneumoniae* | 125 | *Vibrio* | 220 |
| *11* | *Vibrio cholerae* | 109 | *Arthrobacter* | 209 |
| *12* | *Bacillus thuringiensis* | 95 | *Synechococcus* | 192 |
| *13* | *Microbacterium foliorum* | 88 | *Streptomyces* | 179 |
| *14* | *Streptococcus pneumoniae* | 79 | *Propionibacterium* | 164 |
| *15* | *Acinetobacter baumannii* | 71 | *Klebsiella* | 136 |
| *16* | *Bacillus cereus* | 69 | *Microbacterium* | 107 |
| *17* | *Streptomyces griseus* | 67 | *Lactobacillus* | 90 |
| *18* | *Enterococcus faecalis* | 65 | *Acinetobacter* | 87 |
| *19* | *Campylobacter jejuni* | 62 | *Enterococcus* | 78 |
| *20* | *Erwinia amylovora* | 61 | *Shigella* | 76 |
| *21* | *Paenibacillus larvae* | 50 | *Clostridium* | 70 |
| *22* | *Listeria monocytogenes* | 48 | *Erwinia* | 68 |
| *23* | *Ralstonia solanacearum* | 47 | *Campylobacter* | 68 |
| *24* | *Sulfolobus islandicus* | 43 | *Sulfolobus* | 64 |
| *25* | *Rhodococcus erythropolis* | 42 | *Rhodococcus* | 63 |
| *26* | *Cellulophaga baltica* | 41 | *Pectobacterium* | 61 |
| *27* | *Vibrio parahaemolyticus* | 40 | *Aeromonas* | 61 |

| | | | | |
|---|---|---|---|---|
| 28 | *Pectobacterium atrosepticum* | 38 | *Burkholderia* | 55 |
| 29 | *Pseudomonas syringae* | 37 | *Listeria* | 53 |
| 30 | *Aeromonas salmonicida* | 35 | *Paenibacillus* | 50 |
| 31 | *Shigella flexneri* | 35 | *Ralstonia* | 49 |
| 32 | *Clostridium difficile* | 34 | *Citrobacter* | 41 |
| 33 | *Cronobacter sakazakii* | 34 | *Cellulophaga* | 41 |
| 34 | *Moraxella catarrhalis* | 32 | *Cronobacter* | 38 |
| 35 | *Arthrobacter globiformis* | 31 | *Yersinia* | 34 |
| 36 | *Helicobacter pylori* | 31 | *Caulobacter* | 33 |
| 37 | *Bacillus megaterium* | 31 | *Moraxella* | 32 |
| 38 | *Pseudomonas fluorescens* | 30 | *Brucella* | 31 |
| 39 | *Citrobacter freundii* | 30 | *Helicobacter* | 31 |
| 40 | *Bacillus subtilis* | 30 | *Pseudoalteromonas* | 31 |
| 41 | *Clostridium perfringens* | 29 | *Leuconostoc* | 30 |
| 42 | *Caulobacter crescentus* | 26 | *Xanthomonas* | 29 |
| 43 | *Shigella sonnei* | 26 | *Flavobacterium* | 28 |
| 44 | *Staphylococcus epidermidis* | 25 | *Prochlorococcus* | 27 |
| 45 | *Salmonella typhimurium* | 25 | *Enterobacter* | 25 |
| 46 | *Streptomyces lividans* | 24 | *Corynebacterium* | 24 |
| 47 | *Bacillus pumilus* | 24 | *Dickeya* | 24 |
| 48 | *Vibrio anguillarum* | 23 | *Proteus* | 21 |
| 49 | *Brucella abortus* | 22 | *Stenotrophomonas* | 19 |
| 50 | *Yersinia enterocolitica* | 21 | *Ruegeria* | 18 |
| 51 | *Proteus mirabilis* | 20 | *Mannheimia* | 18 |
| 52 | *Streptomyces venezuelae* | 20 | *Rhizobium* | 17 |
| 53 | | | *Acidianus* | 16 |
| 54 | | | *Serratia* | 16 |
| 55 | | | *Brevibacillus* | 16 |
| 56 | | | *Sinorhizobium* | 15 |
| 57 | | | *Achromobacter* | 15 |
| 58 | | | *Haloarcula* | 15 |
| 59 | | | *Thermus* | 14 |
| 60 | | | *Halorubrum* | 12 |
| 61 | | | *Xylella* | 12 |

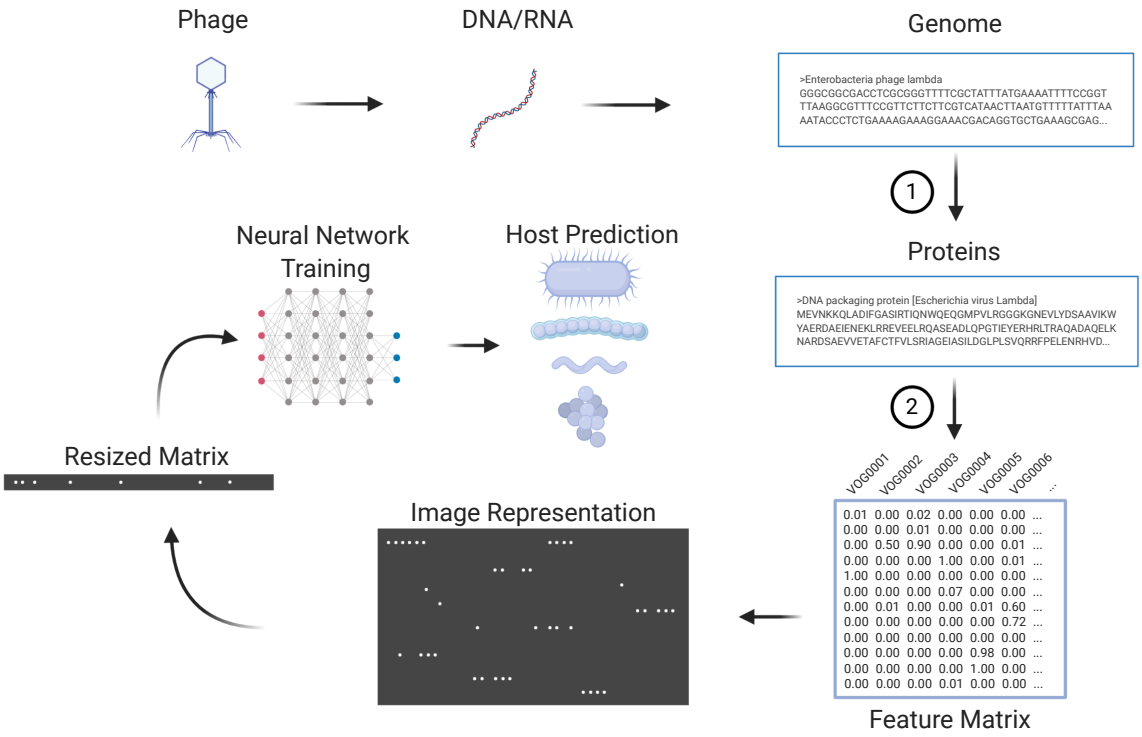
Note: Own work.

As an approximation of a real-life use case, we created the *newly-deposited-genomes (NDG) dataset*. This dataset included GenBank records released between April 1st, 2019 and March 31st, 2020, which ensured that there was no overlap with training sets. The QC filters applied to generate the baseline dataset were applied to this dataset, and complete genomes from viruses infecting any Bacteria and Archaea hosts were considered (including host targets which were not present in the training sets). A total of 2,178 genomes met these criteria.

### 3.2.2. Feature Engineering and Representation

Genomes in the training species/genus datasets were individually submitted to Prokka (Seemann, 2014) for automatic gene prediction and annotation. Predicted protein sequences (FASTA amino acid format) were searched against the pVOGs database of phage protein families (Grazziotin et al., 2017) using HMMscan (Eddy, 1998). Search results were parsed to populate a matrix of pVOG identifiers in columns and phage genes in rows. The matrix cell values were based on the e-value of the best alignment of each gene against the pVOGs database. We call each matrix cell value a Measure of Hit Significance (MHS), defined as follows: if e-value < 1 then MHS = 1 – e-value; if e-value ≥ 1 then MHS = 0. Values close to one mean that significant hits regarding specific phage protein families are present among phage genes for that genome. With this approach, each genome is represented by a matrix of numbers between zero and one. For efficiency reasons, matrices were then transformed into vectors by adding up values in a same column; these vectors were then used as input to the neural network (Figure 3.1). The qualitative difference between original matrices and their vectors is that matrices also carry synteny information (relative position of a gene in the genome). We chose to use vectors instead of matrices because in preliminary testing we obtained similar performance results from both, with matrices being much more computationally intense to train. Feature statistics (those shown in Supplementary Table S1) and multidimensional scaling were weighted for class representation (which helps in minimizing bias because of overrepresented classes) and were calculated using Orange Data Mining Python 3 libraries.

**Figure 3.1**. Pipeline for phage prediction. Schematic representation of how high-level features are engineered to feed the neural network and to perform host prediction. 1 - Gene prediction and annotation by Prokka; 2 - HMMscan search against pVOGs database of phage protein families and generation of a matrix with MHS values.




Note: Created using Biorender (https://biorender.com/). Own work.

### 3.2.3. Machine learning model architecture, configuration, and training

As mentioned above, two separate models were generated for species and genus host predictions. Models are Multi-layer Perceptron (MLP) (Hornik et al., 1989) constituted of one input layer, two hidden layers, and one output layer. The hidden layers (2nd and 3rd) have respectively 1000 and 500 neuron units with a leaky ReLU activation function. Output layers have the exact number of neurons as the number of output classes expected for each model. The softmax activation function was used for neurons at the output layer. Optimization in training was achieved with the ADAM algorithm (Kingma & Ba, 2015). Categorical cross-entropy was used as loss function with weight balancing for each host class because we had an imbalanced dataset. Regularization L2 was implemented in the first neural layer. (We also tested regularization L1, but L2 yielded better results.) Early stopping was implemented during training to avoid overfitting; this allows returning the model to the last step that

showed considerable decrease in the value of the loss function. The final prediction output is a vector composed of 52 prediction scores for species or 61 prediction scores for genus.

To provide the user with something more definite and easier to understand than just a list of scores, a simple heuristic decision is made based on the score of the output layer. An empirically derived threshold of 0.5 is used to define a host prediction. If both species and genus models output scores higher than or equal to the threshold, then the output prediction is that for species. Note that a possible outcome of this heuristic decision is the prediction 'none' (both scores are less than 0.5), which means the prediction is 'none of the hosts is among the available targets'. (Note that this prediction could be a false negative, in the sense that the actual host is, in fact, among the available targets.) Even when the final prediction is 'none', users can still check the list of scores. We emphasize that a heuristic decision was implemented only to help the user in the final decision. Additionally, each score has an entropy value associated with it (more on this below), which may be used to gauge the uncertainty of each score.

Final models were trained with the entire species/genus-datasets and saved with the format TensorFlow SavedModel. Training code was written in Python 3.8.5 using TensorFlow 2.4.1 and Keras 2.4 libraries. Training was performed in a server equipped with a Nvidia GPU Quadros P6000 processor.

### 3.2.4. Tool implementation

Models for prediction of host species and genus were embedded in a standalone toolkit developed in Python 3, named vHULK, for Viral Host UnveiLing Kit. The tool receives phage genomes as FASTA format and outputs top prediction scores for each internal model. To complete the toolkit, vHULK also outputs an entropy value for the vector of prediction scores of the genus model, which reflects how evenly distributed are the scores for each of the target hosts. In other words, entropy will provide a measure for prediction confidence (see additional comments in the Results section). High values of entropy suggest that the model is unsure about a prediction or that a phage may infect different hosts (broad host range). Scipy implementation was used to calculate entropy from a prediction vector.

### 3.2.5. Comparison of vHULK with VirHostMatcher-net, CRISPR spacers matches, and RaFAH

Phage genomes of the NDG dataset were submitted for host prediction using four different approaches: vHULK, neural networks with distance measurements and other features used by VirHostMatcher-net (VHM-net) (Wang et al., 2020), RaFAH (Coutinho et al., 2021), and CRISPR-spacers as links between host and phage genomes. For vHULK, phage genomes were individually submitted to the tool using default parameters. For VHM-net, individual genomes were submitted for prediction using default parameters. Default datasets for possible hosts were used in VHM-net. For RaFAH, default parameters were used. For the CRISPR-spacers approach, individual genomes were searched against a database of CRISPR spacers of the CRISPRCasDB (Pourcel et al., 2020) using the NCBI blastn tool (Altschul et al., 1990). To adapt our search to CRISPR spacers characteristics, word_size was set to 9 nt, and thresholds of e-value $\leq$ 0.001 and number of mismatches $\leq$ 2 were applied to select significant hits. Only the best scoring hit was considered in each approach. In all four cases, predicted hosts were compared to known phage hosts as specified by the tag ‘host’ or ‘lab_host’ in the GenBank record. The evaluation metrics were precision, recall, F1, and their respective macro versions (macro-precision, macro-recall and macro-F1). The macro version of a metric is the unweighted mean of all values obtained for that metric. We also use accuracy, defined as (number of correct predictions) / (total number of instances).

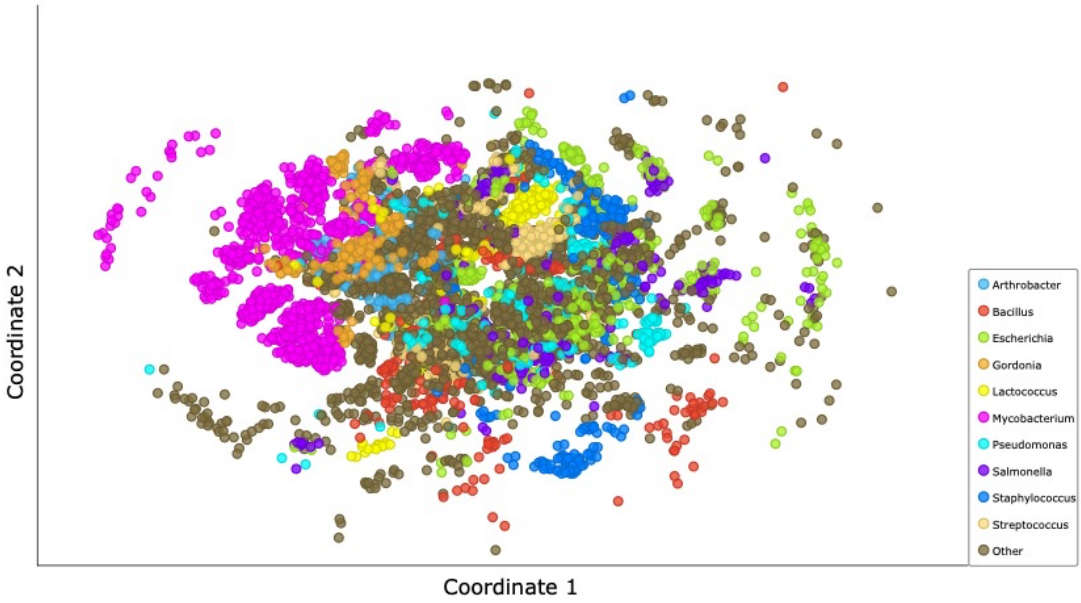
## 3.4. Results

### 3.4.1. Phage-host predictions features

A total of 9,504 features were generated to train the vHULK neural network models in predicting phage hosts. Features are the MHS (Measure of Hit Significance) values obtained from similarity searches to a corresponding phage orthologous group (pVOG) (Grazziotin et al., 2017), as shown in Figure 3.1. Feature selection was not necessary, as neural networks with regularization perform their own feature selection by adjusting edge weights to zero (when a feature is not important for the target prediction) or to one (when it is a very informative feature). Multidimensional Scaling (MDS) representation for the entire set of features considering genus targets is shown in

Figure 3.2. In this two-dimensional projection, some hosts appear better clustered than others. For example, the points that represent Mycobacterium, Gordonia, Staphylococcus, and Streptococcus are better clustered than those that represent Salmonella, Escherichia, and Pseudomonas. This suggests that certain hosts (those that do not cluster well) may be harder to predict with this set of features than others.

**Figure 3.2**. Multidimensional scaling 2D projection based on Euclidean distances of the 9,504 features used in this study. Each dot indicates a phage genome of the baseline dataset. Color indicates host for the most frequent genera.




Note: Own work.

We assessed feature relevance by using a predominant correlation approach as implemented in the Fast Correlation-Based Filter algorithm (FCBF). In this way we identified a set of features that seem to be strongly correlated with host prediction in specific groups of phages. Among these, features that represent the orthologous groups VOG4545, VOG4558 and VOG4544 had the highest correlation values. VOG4545 is an orthologous group ubiquitous among the three tailed phages families (*Myoviridae*, *Siphoviridae* and *Podoviridae*), composed of 1,747 known orthologous proteins. Most of the VOG4545 member proteins are annotated as a tape measure protein, which is thought to vary in length according to phage tail size (Belcaid et al., 2011). Proteins of the orthologous group VOG4558 are present in phages of the *Siphoviridae* family, mainly infecting hosts of the genera *Mycobacterium* and *Gordonia*.

They are annotated as a structural minor tail protein. The orthologous group VOG4544 is composed of proteins annotated as Terminase Large Subunit (TerL). This VOG can be found in 2281 genomes, from 68 genera found in three virus families (information from the pVOGs database: http://dmk-brain.ecn.uiowa.edu/pVOGs/), being therefore rather widespread among phages. It is worth noting that pVOG annotations of most of the top correlated features indicate proteins directly or indirectly related to the infection mechanism. Unweighted and weighted information theory metrics for all features used in this study are available as Supplementary Table S1.

### 3.4.2. Neural network testing on the AAI redundancy reduction datasets

The results of the vHULK tool in the AAI redundancy reduction datasets is reported in Table 3.2. F1-scores ranged from 74% to 83% for genus predictions, while species predictions resulted in F1-scores ranging from 69% to 79%. Macro versions of the F1-score resulted in lower values (60-69% at genus level and 57-64% at the species level).

**Table 3.2.** Performance results (in %) on the AAI redundancy reduction datasets using different levels of redundancy removal. Values represent the average of ten random splits, and standard deviations are shown in parentheses.

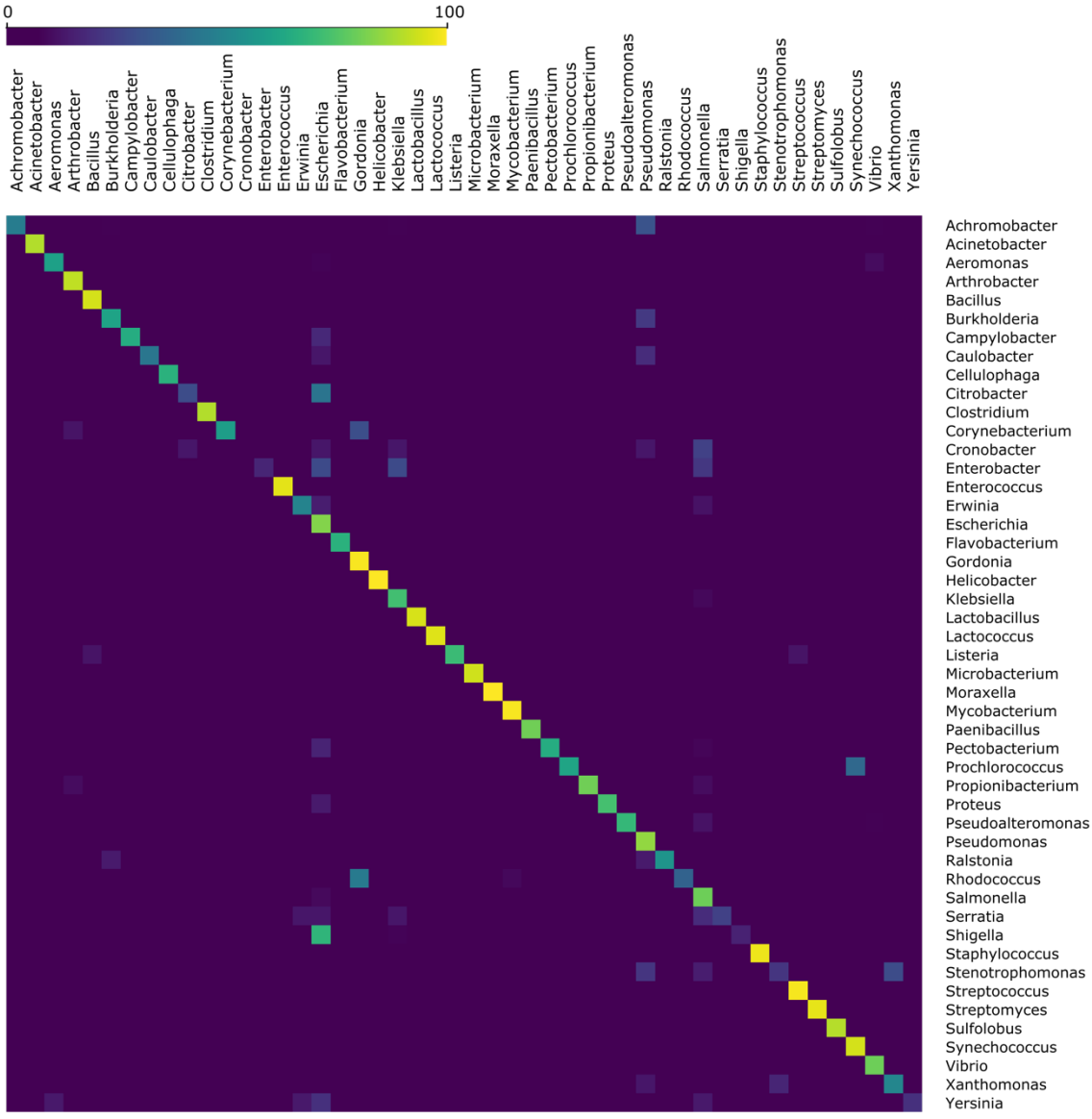
| | Genus 70 | Genus 80 | Genus 90 | Species 70 | Species 80 | Species 90 |
|---|---|---|---|---|---|---|
| **Precision** | 76.2 (2.0) | 81.0 (2.0) | 85.8 (0.6) | 73.7 (2.2) | 76.9 (2.8) | 81.0 (1.7) |
| **Macro-Precision** | 63.0 (2.2) | 64.1 (2.6) | 70.8 (1.9) | 61.8 (2.6) | 63.8 (3.1) | 65.8 (1.5) |
| **Recall** | 73.9 (2.1) | 79.5 (1.8) | 82.8 (1.5) | 69.1 (1.9) | 73.3 (2.1) | 78.9 (1.6) |
| **Macro-Recall** | 61.6 (3.3) | 62.3 (1.8) | 70.7 (2.8) | 58.2 (5.1) | 61.8 (2.3) | 66.1 (2.7) |
| **F1** | 73.9 (2.3) | 79.4 (1.6) | 83.4 (1.2) | 69.3 (2.0) | 73.2 (2.4) | 78.8 (1.9) |
| **Macro-F1** | 59.7 (1.8) | 61.2 (1.5) | 69.0 (1.9) | 57.0 (3.7) | 59.6 (2.8) | 63.7 (1.8) |

Note: Own work.

Confusion matrix for predictions at the genus level using the 90% AAI redundancy reduction dataset is shown in Figure 3.3. As expected, most of the predictions are centered around the diagonal line. Most of the confusion happened with hosts from the same taxonomic family. These results are expected and may be

biologically meaningful, since wide host-range phages have been reported to infect species belonging to different genera from the same family (37,38). Detailed confusion matrices showing average percentage of actual classes for five random split train/test datasets are presented in Supplementary Table S2.

**Figure 3.3**. Confusion matrix of host predictions. Confusion matrix for predictions at the genus level using the 90% AAI redundancy reduction dataset. Color scale corresponds to average percentage values of five random test splits. Corrected predictions are centered on the diagonal line.




Note: Own work.

### 3.4.3. Performance results on the Newly Deposited Genomes dataset

We used the NDG dataset for as an approximation of a real-life use case (see Methods). Performance on the NDG dataset varied depending on the host and type of prediction (genus or species level). Table 3.3 shows performance results for hosts of the ESKAPE group, which are bacteria known to be multi-drug resistant bacteria. A complete description of vHULK output results as well as performance metrics for all phages of the NDG dataset are available as Supplementary Tables S3, S4 and S5.

**Table 3.3**: Precision/Recall details for six bacterial hosts of the ESKAPE group. Total counts of genomes in the NDG dataset for each host is shown, as well as precision/recall percentages. Enterococcus faecium is not one of the possible hosts in vHULK's predictions; therefore, predictions are shown at the genus level only.

| | *Genus count* | *Species count* | *Genus Recall Level (%)* | *Genus Precision Level (%)* | *Species Recall Level (%)* | *Species Precision Level (%)* |
|---|---|---|---|---|---|---|
| *Enterococcus faecium* | 40 | -- | 100 | 100 | -- | -- |
| *Staphylococcus aureus* | 18 | 18 | 98 | 100 | 100 | 45 |
| *Klebsiella pneumoniae* | 86 | 86 | 55 | 82 | 62 | 65 |
| *Acinetobacter baumannii* | 38 | 38 | 95 | 93 | 97 | 93 |
| *Pseudomonas aeruginosa* | 103 | 103 | 84 | 95 | 89 | 98 |
| *Enterobacter spp.* | 10 | -- | 10 | 25 | -- | -- |

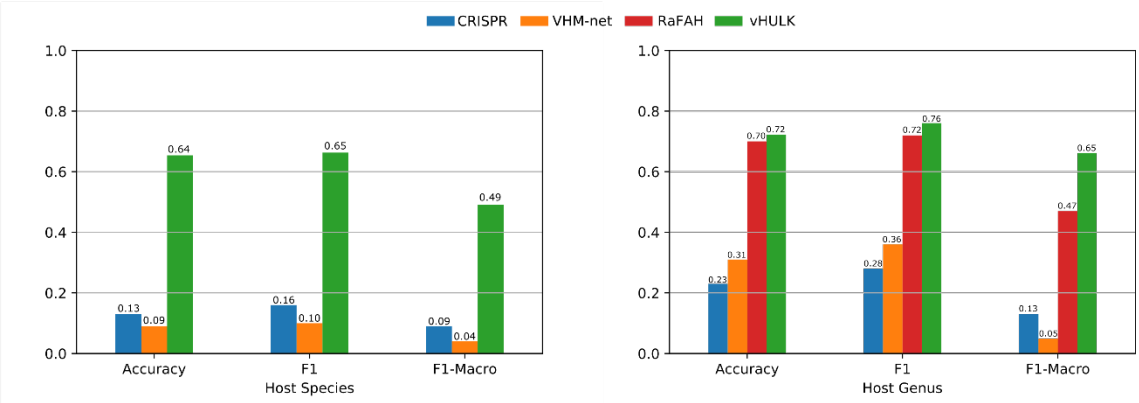
Note: Own work.

### 3.4.4. Comparing phage host prediction programs

We have compared vHULK host predictions to those provided by VHM-net, RaFAH, and CRISPR-spacers matches. The NDG dataset was used for this purpose. As shown in Figure 3.4, F1-score for vHULK predictions was 4 times higher than the second-best approach when predictions at the species level are considered (65% against 16% of CRISPR-spacers). When predictions at the genus level are considered, vHULK obtained an F1-score of 76% across 61 possible host targets, while RaFAH (the second best) obtained 72%. Despite very close F1-scores at the genus level, vHULK presented a larger difference when the macro version of F1-scores was considered.

These results could mean that RaFAH performs very well in specific group of well-represented hosts and not as well in less represented hosts, which would be less of an issue for vHULK's predictions. Detailed information about predictions in the NDG dataset for the four methods is available in Supplementary Table S6.

**Figure 3.4**. Perfomance metrics of vHULK, VirHostMatcher-net RaFAH tools, as well as the CRISPR-spacers matches approach, on the NDG dataset. Predictions were performed for 2,178 phage genomes of the NDG dataset, both at the species and genus levels.




Note: Own work.

All four programs ran on a server with Intel Xeon 2 GHz CPUs, 22 threads and 32 GB of RAM. vHULK took approximately 18 h of wall time, VHM-net took approximately 32 h, CRISPR-spacers took approximately 50 minutes, and RaFAH took approximately 15 h. Genomes were submitted sequentially, and no additional parallelization was used in any of the tools.

## 3.5. Discussion

We have presented vHULK, a new program to predict phage hosts based on phage genome sequences. A comparison between vHULK's and three other tools/methods for host prediction showed that our program had superior performance at both the species and genus levels, and in all three main metrics considered (Accuracy, F1, and F1-Macro).

Testing of vHULK on datasets with decreased protein sequence redundancy levels showed F1-scores that varied from 79% to 69% at the species level and 83% to 74% at the genus level. These results show clearly that the less redundant a dataset

is in terms of protein similarity, the lower the scores achieved by vHULK. This points to the dependency that our tool has on protein sequence similarities between training data and test data, as one would expect, given the tool's design. Nevertheless, the small decrease in performance with decreased redundancy suggests that the tool is relatively robust with respect to lack of redundancy.

vHULK was trained to predict 52 different species and 61 different genera. These numbers of possible hosts were limited only by the number of instances available to train the model and by our requirement that there had to be at least 20 records for any given species and at least 12 records for any given genus. One can ask whether these 52 species and 61 genera are representative of currently available data. We can answer this question in part by noting that the 52 species account for 70% of all host species present in the NDG dataset, and the 61 genera account for 97% of all host genera present in the NDG dataset. On the other hand, 52 and 61 are tiny numbers when compared with the presumed richness of phage hosts that exist in the entire biosphere. We expect however that the number of species and genus hosts in the vHULK repertoire will grow with future releases of the software.

It is known that many machine learning prediction models may struggle when facing instances for which there was no target in training (Galiez et al., 2017; Villarroel et al., 2016; Wang et al., 2020). We have tried to mitigate this problem by providing users with entropy values associated with scores; we have shown that these entropy values can be used as proxies for prediction confidence.

Despite the relatively small host repertoire, we envision that vHULK will have wide application. One use could be the mining of metagenomic datasets for new phages that may have as host one of the six multidrug resistant bacterial pathogens that compose the ESKAPE group (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Enterobacter* spp.) (Pendleton et al., 2013). All of them are part of the vHULK host repertoire. We note also that methods for recovery of phage genomes from metagenomics datasets have been improving, and that it has become relatively common that one can restrict recovered phage genomes to 90% completeness (using, for example, checkV (Nayfach et al., 2021)), and still obtain hundreds of genomes from a given sample (depending on ecological niche and sequence coverage). This means

that vHULK's requirement that query genomes be complete or nearly complete should not be a problem in most applications.

Regarding the design of vHULK, we would like to point out that the program embodies a novel approach for extracting and representing genomic features for use in neural networks. The main novelty is the use of alignment significance scores (MHS) when evaluating presence/absence of protein families in phage genomes, and their number when they are present, as features in the neural network. Our results demonstrate that the use of these features was a determining factor in achieving good accuracy. We conjecture that this same approach can be extended to the prediction of other viral biological attributes, such as viral life cycle and virulence.

F. Young and colleagues have recently evaluated different levels of genomic features for phage host prediction (Young et al., 2020). These authors defined four levels of features that could be extracted from genomes by applying different annotation processes: Oligonucleotide frequencies, amino acid frequencies, protein physic-chemical properties, and protein domains. They conclude that all feature levels yield significant signals for host prediction, probably being complementary to one another. The protein features in vHULK are somewhat related to the protein domains of Young et al. (Young et al., 2020). However, whereas they proposed a count of domains, we use alignment significance scores. To our knowledge, this type of machine learning feature has never been used in this application.

In sum, our results show that vHULK achieves the best performance among existing programs for phage host prediction. We hope that its availability will be found useful by the research community interested in such predictions.

## Software and Datasets Availability

The tool, the models, and the testing examples and documentation are publicly available in GitHub: https://github.com/LaboratorioBioinformatica/vHULK. We also provide GenBank records used to train vHULK's models and accessions numbers for the NDG dataset in: http://projetos.lbi.iq.usp.br/phaghost/vHULK/datasets.

## Supplementary Material availability

Supplementary tables and files cited in this chapter are available in: https://github.com/LaboratorioBioinformatica/vHULK.

## 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), e126–e126. https://doi.org/10.1093/nar/gks406

Allers, E., Moraru, C., Duhaime, M. B., Beneze, E., Solonenko, N., Barrero-Canosa, J., Amann, R., & Sullivan, M. B. (2013). Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses. *Environmental Microbiology*, *15*(8), 2306–2318. https://doi.org/10.1111/1462-2920.12100

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amgarten, D., Iha, B. K. V., Piroupo, C. M., Silva, A. M. da, & Setubal, J. C. (2020). vHULK, a new tool for bacteriophage host prediction based on annotated genomic features and deep neural networks. *BioRxiv*, 2020.12.06.413476. https://doi.org/10.1101/2020.12.06.413476

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. https://doi.org/10.1093/nar/gkw387

Belcaid, M., Bergeron, A., & Poisson, G. (2011). The evolution of the tape measure protein: Units, duplications and losses. *BMC Bioinformatics*, *12*(SUPPL. 9), S10. https://doi.org/10.1186/1471-2105-12-S9-S10

Breitbart, M., Bonnain, C., Malki, K., & Sawaya, N. A. (2018). Phage puppet masters of the marine microbial realm. In *Nature Microbiology* (Vol. 3, Issue 7, pp. 754–766). Nature Publishing Group. https://doi.org/10.1038/s41564-018-0166-y

Breitbart, M., & Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? In *Trends in Microbiology* (Vol. 13, Issue 6, pp. 278–284). Elsevier Current Trends. https://doi.org/10.1016/j.tim.2005.04.003

Brum, J. R., Cesar Ignacio-Espinoza, J., Roux, S., Doulcier, G., Acinas, S. G., Alberti, A., Chaffron, S., Cruaud, C., de Vargas, C., Gasol, J. M., Gorsky, G., Gregory, A. C., Guidi, L., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Poulos, B. T., … Weissenbach, J. (2015). Patterns and ecological drivers of ocean viral communities. *Science*, *348*(6237). https://doi.org/10.1126/science.1261498

Carroll, D., Daszak, P., Wolfe, N. D., Gao, G. F., Morel, C. M., Morzaria, S., Pablos-Méndez, A., Tomori, O., & Mazet, J. A. K. (2018). The Global Virome Project. In *Science* (Vol. 359, Issue 6378, pp. 872–874). https://doi.org/10.1126/science.aap7463

Chaturongakul, S., & Ounjai, P. (2014). Phage-host interplay: Examples from tailed phages and Gram-negative bacterial pathogens. In *Frontiers in Microbiology* (Vol. 5, Issue AUG, p. 442). Frontiers Media S.A. https://doi.org/10.3389/fmicb.2014.00442

Chow, C.-E. T., Winget, D. M., White, R. A., Hallam, S. J., & Suttle, C. A. (2015). Combining genomic sequencing methods to explore viral diversity and reveal potential virus-host

interactions. *Frontiers in Microbiology*, *6*(APR), 265. https://doi.org/10.3389/fmicb.2015.00265

Coclet, C., & Roux, S. (2021). Global overview and major challenges of host prediction methods for uncultivated phages. In *Current Opinion in Virology* (Vol. 49, pp. 117–126). Elsevier. https://doi.org/10.1016/j.coviro.2021.05.003

Coutinho, F. H., Zaragoza-Solas, A., López-Pérez, M., Barylski, J., Zielezinski, A., Dutilh, B. E., Edwards, R., & Rodriguez-Valera, F. (2021). RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. *Patterns*, *2*(7), 100274. https://doi.org/10.1016/J.PATTER.2021.100274

de Jonge, P. A., Nobrega, F. L., Brouns, S. J. J., & Dutilh, B. E. (2019). Molecular and Evolutionary Determinants of Bacteriophage Host Range. In *Trends in Microbiology* (Vol. 27, Issue 1, pp. 51–63). Elsevier Ltd. https://doi.org/10.1016/j.tim.2018.08.006

Deng, L., Gregory, A., Yilmaz, S., Poulos, B. T., Hugenholtz, P., & Sullivan, M. B. (2012). Contrasting life strategies of viruses that infect photo- and heterotrophic bacteria, as revealed by viral tagging. *MBio*, *3*(6). https://doi.org/10.1128/mBio.00373-12

Eddy, S. R. (1998). Profile hidden Markov models. In *Bioinformatics* (Vol. 14, Issue 9, pp. 755–763). Oxford University Press. https://doi.org/10.1093/bioinformatics/14.9.755

Edwards, R. A., McNair, K., Faust, K., Raes, J., & Dutilh, B. E. (2016). Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiology Reviews*, *40*(2), 258–272. https://doi.org/10.1093/femsre/fuv048

Emerson, J. B. (2019). Soil Viruses: A New Hope. *MSystems*, *4*(3). https://doi.org/10.1128/msystems.00120-19

Fouts, D. E. (2006). Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Research*, *34*(20), 5839–5851. https://doi.org/10.1093/NAR/GKL732

Galiez, C., Siebert, M., Enault, F., Vincent, J., & Söding, J. (2017). WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics*, *33*(19), 3113–3114. https://doi.org/10.1093/bioinformatics/btx383

Grazziotin, A. L., Koonin, E. v., & Kristensen, D. M. (2017). Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Research*, *45*(D1), D491–D498. https://doi.org/10.1093/NAR/GKW975

Gregory, A. C., Zablocki, O., Zayed, A. A., Howell, A., Bolduc, B., & Sullivan, M. B. (2020). The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host & Microbe*. https://doi.org/10.1016/j.chom.2020.08.003

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Kingma, D. P., & Ba, J. L. (2015, December 22). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. https://arxiv.org/abs/1412.6980v9

Koonin, E. v., Dolja, V. v., Krupovic, M., Varsani, A., Wolf, Y. I., Yutin, N., Zerbini, F. M., & Kuhn, J. H. (2020). Global Organization and Proposed Megataxonomy of the Virus World. *Microbiology and Molecular Biology Reviews*, *84*(2). https://doi.org/10.1128/MMBR.00061-19/ASSET/A150F1AA-592D-4605-9D35-99DF755DC6A0/ASSETS/GRAPHIC/MMBR.00061-19-F0012.JPEG

Kortright, K. E., Chan, B. K., Koff, J. L., & Turner, P. E. (2019). Phage Therapy: A Renewed Approach to Combat Antibiotic-Resistant Bacteria. *Cell Host & Microbe*, *25*(2), 219–232. https://doi.org/10.1016/J.CHOM.2019.01.014

Leite, D. M. C., Brochet, X., Resch, G., Que, Y. A., Neves, A., & Peña-Reyes, C. (2018). Computational prediction of inter-species relationships through omics data analysis and machine learning. *BMC Bioinformatics*, *19*(S14), 420. https://doi.org/10.1186/s12859-018-2388-7

Lima-Mendez, G., van Helden, J., Toussaint, A., & Leplae, R. (2008). Prophinder: A computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics*, *24*(6), 863–865. https://doi.org/10.1093/bioinformatics/btn043

Martínez-García, M., Santos, F., Moreno-Paz, M., Parro, V., & Antón, J. (2014). Unveiling viral-host interactions within the “microbial dark matter.” *Nature Communications*, *5*(1), 1–8. https://doi.org/10.1038/ncomms5542

Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., & Kyrpides, N. C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology*, *39*(5), 578–585. https://doi.org/10.1038/s41587-020-00774-7

Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N. N., & Kyrpides, N. C. (2016). Uncovering Earth’s virome. *Nature*, *536*(7617), 425–430. https://doi.org/10.1038/nature19094

Pendleton, J. N., Gorman, S. P., & Gilmore, B. F. (2013). Clinical relevance of the ESKAPE pathogens. In *Expert Review of Anti-Infective Therapy* (Vol. 11, Issue 3, pp. 297–308). Taylor & Francis. https://doi.org/10.1586/eri.13.12

Pourcel, C., Touchon, M., Villeriot, N., Vernadet, J. P., Couvin, D., Toffano-Nioche, C., & Vergnaud, G. (2020). CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. *Nucleic Acids Research*, *48*(D1), D535–D544. https://doi.org/10.1093/nar/gkz915

Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F., & Gordon, J. I. (2012). Going viral: Next-generation sequencing applied to phage populations in the human gut. In *Nature Reviews Microbiology* (Vol. 10, Issue 9, pp. 607–617). Nature Publishing Group. https://doi.org/10.1038/nrmicro2853

Roux, S., Páez-Espino, D., Chen, I. M. A., Palaniappan, K., Ratner, A., Chu, K., Reddy, T., Nayfach, S., Schulz, F., Call, L., Neches, R. Y., Woyke, T., Ivanova, N. N., Eloe-Fadrosh, E. A., & Kyrpides, N. C. (2021). IMG/VR v3: An integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Research*, *49*(D1), D764–D775. https://doi.org/10.1093/nar/gkaa946

Samson, J. E., Magadán, A. H., Sabri, M., & Moineau, S. (2013). Revenge of the phages: Defeating bacterial defences. In *Nature Reviews Microbiology* (Vol. 11, Issue 10, pp. 675–687). Nature Publishing Group. https://doi.org/10.1038/nrmicro3096

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Villarroel, J., Kleinheinz, K. A., Jurtz, V. I., Zschach, H., Lund, O., Nielsen, M., & Larsen, M. V. (2016). HostPhinder: A Phage Host Prediction Tool. *Viruses 2016, Vol. 8, Page 116*, *8*(5), 116. https://doi.org/10.3390/V8050116

Wang, W., Ren, J., Tang, K., Dart, E., Ignacio-Espinoza, J. C., Fuhrman, J. A., Braun, J., Sun, F., & Ahlgren, N. A. (2020). A network-based integrated framework for predicting virus–prokaryote interactions. *NAR Genomics and Bioinformatics*, *2*(2). https://doi.org/10.1093/NARGAB/LQAA044

Young, F., Rogers, S., & Robertson, D. L. (2020). Predicting host taxonomic information from viral genomes: A comparison of feature representations. *PLoS Computational Biology*, *16*(5), 1–24. https://doi.org/10.1371/journal.pcbi.1007894

# **Chapter 4**: Genome Recovery and Computational Characterization of Environmental Viruses in Composting and Soil Samples from the Sao Paulo Zoo Park

## 4.1. Contextualization

The work described in this chapter was conceived as a proof of concept to validate the methodology described in the two previous chapters of this thesis. Here, we implemented an end-to-end workflow, which involves sample collection, sequencing data generation, diversity analyses, assembly, binning, prediction and recovery of putative phage genomes, host prediction, and further characterization. The results presented here are original research and have not been published elsewhere.

## 4.2. Background

Environmental samples are a major source of microbial diversity (Danko et al., 2021; Gilbert et al., 2014; Moran, 2015). Many new metagenome-assembled genomes (MAGs) are being recovered every year, helping researchers to catalog the unknown genomic and proteomic diversity held by microorganisms (Chen et al., 2019; Nayfach, Roux, et al., 2021; Parks et al., 2017). Viruses are likely the most abundant and diverse organisms in the biosphere; however, only a small portion of their diversity is well represented in public databases (Krishnamurthy & Wang, 2017; Roux et al., 2015). Consequently, studies and protocols to recover viral MAGs (vMAGs) are important in this area (Benler et al., 2021; Roux et al., 2018).

Among the many environments investigated in these studies, soil and marine samples have been referred to as hotspots of microbial diversity on Earth (Daniel, 2005; Sunagawa et al., 2020). In addition, the human gut is one of the most investigated samples, with large, publicly available datasets. These datasets are constantly mined, and, in some cases, thousands of vMAGs have been recovered (Benler et al., 2021; Camarillo-Guerrero et al., 2021; Nayfach, Páez-Espino, et al.,

2021; Roux et al., 2019; Tisza & Buck, 2021). Our research group has extensively studied composting samples, finding that they harbor an unexplored diversity of bacterial and viral organisms (Amgarten et al., 2017; Antunes et al., 2016; Braga et al., 2021; Lemos et al., 2017; Martins et al., 2013). Altogether, this indicates the substantial potential for recovering new virus genomes from the aforementioned environments, or virtually any environmental sample.

Approaches to recognizing vMAGs among various bacterial sequences may vary and are mostly dependent on similarity searches against known viral genomes or protein clusters (Amgarten et al., 2018; Guo et al., 2021). After vMAGs are assembled and recovered, a quality assessment is required to validate the findings and ensure that they have minimal contamination (Roux et al., 2018); tools such as CheckV (Nayfach, Camargo, et al., 2021), vCONTACT, (Bolduc et al., 2017), and PhaGCN (Shang et al., 2021) are typically used for this purpose. In addition, new vMAGs also need to be characterized in terms of important biological features such as host and taxonomic affiliations. Host prediction may be performed using approaches such as similarity searches of known phage genomes or host CRISPR spacers (Coutinho et al., 2021; Villarroel et al., 2016). Additionally, more sensitive approaches have been proposed, using elaborate machine-learning techniques and artificial neural networks (Amgarten et al., 2020; Wang et al., 2020). Generating a set of minimal information about vMAGs is crucial for cataloging viral diversity and enabling further experimental investigation.

In this chapter, we present an end-to-end workflow to recover vMAGs from composting and soil metagenomic datasets—hotspots of unexplored viral diversity. Two of the tools used in this workflow have been described in the previous chapters of this thesis. Therefore, the results presented here are intended as proof of concept for such approaches.

## 4.3. Methods

### 4.3.1. Sample collection

Samples were collected from Sao Paulo Zoo Park between September 18, 2017 and April 26, 2018. On these occasions, we collected samples from three different semi-

industrial thermophilic composting units, which were on days 1 (using layers from the moment of assembly), 24, and 72 of a 90-day process. Samples from composting on day 1 were termed C1; a pool of samples from composting on days 24 and 72 was termed C2. Composting was performed as described previously (Antunes et al., 2016). Briefly, soil samples were collected from a fragment of Atlantic rain forest within the Sao Paulo Zoo Park at 23°38'55.2"S 46°36'58.4", and the sampling procedure was as follows: sterilized 50-mL tubes were used to collect the 0–20-cm top layer of the soil. Two different collection areas were chosen inside the forest (termed S1 and S2), at least 20 m from the border region and 1 m from the tree roots. At the central spot of each area, five samples were collected with a minimum distance of 50 cm between them, and the samples were pooled to obtain one composite sample per point. Composting and soil samples were stored at -80 °C until further experiments. All samples were collected as part of the Metazoo FAPESP thematic project (2011/508706) and under the license 02001.000693/2013-89 issued by the Brazilian Institute of the Environment and Renewable Natural Resources (IBAMA).

### 4.3.2. Sample processing and next generation sequencing

Each sample (C1, C2, S1, and S2) was split into two aliquots. The first aliquot was subjected to purification and total DNA extraction to generate metagenomic sequencing libraries according to procedures established in our laboratory and reported previously (Antunes et al., 2016). The second aliquot was used for viral particle enrichment (virome), following the adapted protocol described by Thurber et al. (2009). Briefly, composting or soil (2.5 g) was added to 15 mL of Sorensen's phosphate buffer (0.133 M, pH 7.2). Samples were vortexed for 10 min and centrifuged for 5 min at 2,500 rpm. Supernatant was filtered using a sterile syringe filter (Millex-HP Syringe Filter Unit, polyethersulfone membrane, 0.45-µm pore size, 33-mm diameter) and then concentrated using Amicon Ultra 0.5-mL centrifugal filters of 10 kDa. A 200-µL aliquot of the concentrated sample was subjected to DNA extraction using the MoBio DNA Power Soil kit. DNA purity and concentration were evaluated using an ND-1000 spectrophotometer (Nano Drop Technologies, USA) at 260 nm, 280 nm, and 230 nm. Further quantification was performed using a Quant-iT Picogreen dsDNA assay kit (Life Technologies, USA).

Fragment libraries of C1, C2, S1, and S2 total and virome DNA were prepared using the Illumina Nextera DNA library preparation kit (Illumina, Inc., USA) with a DNA input of 20–35 ng. The resulting libraries were cleaned using Agencourt AMPure XP beads (Beckman Coulter, Inc., USA), and fragment size within the range of 400–700 bp was verified by running on a 2100 Bioanalyzer using an Agilent High Sensitivity DNA chip. Sequencing was performed using the Illumina MiSeq equipment with Illumina paired-end V2 sequencing kits for 500 cycles. All libraries were successfully sequenced, except for S1-total, which did not yield any reads. We performed a second attempt at sequencing this sample with the same results. We have no explanation for this failure and did not have an additional S1 sample to perform a new round of DNA isolation and library preparation. The details and yields of the sequencing are shown in Table 4.1.

**Table 4.1:** Description of the four metagenomic samples used in this work to validate the proposed workflow for recovery and characterization of new phages from environmental samples. S1-total sample did not yield any reads and was omitted.

| *ID* | Sample type | Experimental Procedure | Total reads |
|---|---|---|---|
| *C1-total* | Composting | Total DNA Metagenomics | 2,809,521 |
| *C1-viral* | Composting | Virome | 2,291,758 |
| *C2-total* | Composting | Total DNA Metagenomics | 7,360,600 |
| *C2-viral* | Composting | Virome | 5,302,793 |
| *S1-viral* | Soil | Virome | 4,994,576 |
| *S2-total* | Soil | Total DNA Metagenomics | 1,180,370 |
| *S2-viral* | Soil | Virome | 4,429,989 |

Note: Own work.

### 4.3.2. Bioinformatics analyses

Raw reads were subjected to quality assessment using FASTQC (https://www.bioinformatics.babraham.ac.uk). Sequences shorter than 50 bp were removed, and trimming was performed (5 bp at the 5′ end and 8 bp at the 3′ end) to remove poor quality ends using Cutadapt (Martin, 2011). Passed-QC reads were assembled individually using Spades v.1.15 (Prjibelski et al., 2020), with the '--meta' tag set to 'True.' Contigs for each sample (e.g., contigs generated from C1-total plus

C1-viral) were merged in a single FASTA and deduplicated using Dedupe from BBMap (https://sourceforge.net/projects/bbmap). Deduplicated contigs were used as a reference to map the original sample reads using the BWA tool (Li & Durbin, 2009). The BAM files and contigs for each sample were binned using metaBAT2 (Kang et al., 2019). Contigs in each individual bin were re-assembled using Phrap (Bastide & McCombie, 2007), which looks for overlaps at the ends of contigs to generate continuous scaffolds. A minimum overlap of 20 bp and a maximum of 1 mismatch were set to allow merging of overlapping contigs.

Metagenomic bins were submitted for phage prediction using MARVEL (Amgarten et al., 2018) with default parameters. Predicted bins were filtered by length (≤ 10 kbp), and a pairwise ANI comparison matrix was calculated using the OrthoANIu algorithm (Yoon et al., 2017). Redundant bins with more than 95% identity covering more than 70% of the smaller bin were clustered, and only the larger bin was retained. Putative phage bins were further assessed for quality, completeness, and contamination using CheckV (Nayfach, Camargo, et al., 2021) and CheckM (Parks et al., 2015).

Host prediction for each bin was performed using vHULK (Amgarten et al., 2020) and CRISPR spacer matches. For the first approach, default parameters and score thresholds were used. For the CRISPR-spacer approach, individual genomes were searched against CRISPRCasdb, a database of CRISPR spacers (Pourcel et al., 2020), using the NCBI blastn tool (Altschul et al., 1990). To adapt our search to CRISPR spacer characteristics, *word_size* was set to 9 nt, E-value threshold was set to ≤ 0.001, and the number of mismatches was ≤ 2 to select significant hits; only the best-scoring hit was considered.

Taxonomic identification of reads and contigs was performed using Kraken2 with the Standard plus protozoa and fungus database (May 17, 2021) and CCMetagen using the ncbi_nt_no_env_11jun2019 database (Marcelino et al., 2020). The Kraken2 counts were further normalized using Bayesian statistics, as implemented by Bracken (Wood et al., 2019). Taxonomic assignment of the putative phage bins to the family level was performed using PhaGCN (Shang et al., 2021) with '--len 4000' and further default parameters. Predictions were performed by individual contigs of a bin, and in the case of divergent predictions, the most frequent assignment was selected.

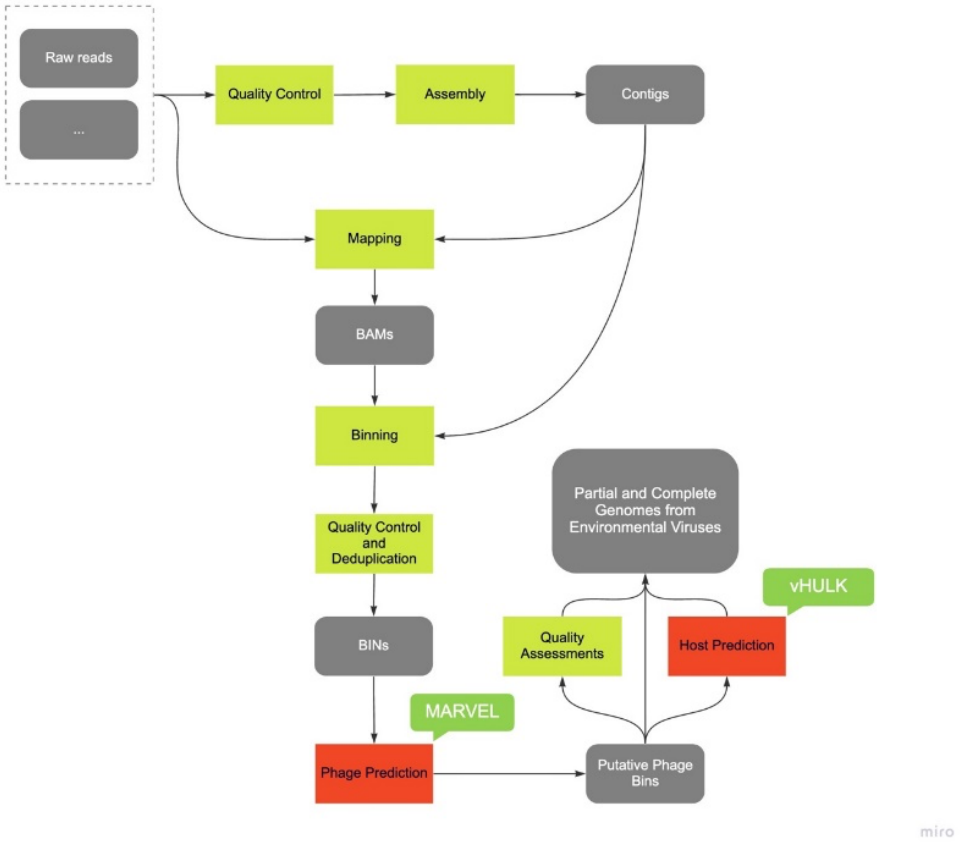
Bins indicated as complete genomes by CheckV were further annotated using the PATRIC v3.6.12 online platform (Davis et al., 2020) and RAST tool kit (Brettin et al., 2015). Similarity searches were performed using the NCBI blastn tool (Altschul et al., 1990) against the complete non-redundant NT database (October 12, 2021) and against the IMG/VR viral sequence database (Roux et al., 2021) (October 12, 2021). For nt searchers, a hit was considered significant if the E-value was ≤ 10e-5 and *query_cover* ≥ 30%. For IMG/VR searchers, only a threshold of E-value ≤ 10e-5 was defined, and the best-scoring hit was considered.

## 4.4. Results

### 4.4.1. Complete workflow for recovery of phage genomes from environmental samples

We primarily developed a bioinformatics pipeline to recover partial and complete genomes from environmental viruses. Our pipeline includes a final step of characterization with quality assessments and host prediction. Several public tools are used in this pipeline, including tools that were specifically developed in this thesis to fill gaps in the area of research. Figure 4.1 shows a flowchart of the pipeline with an emphasis on the MARVEL and vHULK tools developed by our group. Bash scripts for the automated steps of the pipeline are publicly available at https://github.com/LaboratorioBioinformatica/vMAGS.

**Figure 4.1**: Flowchart of the pipeline developed to recover viral metagenome-assembled genomes (vMAGs) from metagenomic datasets.



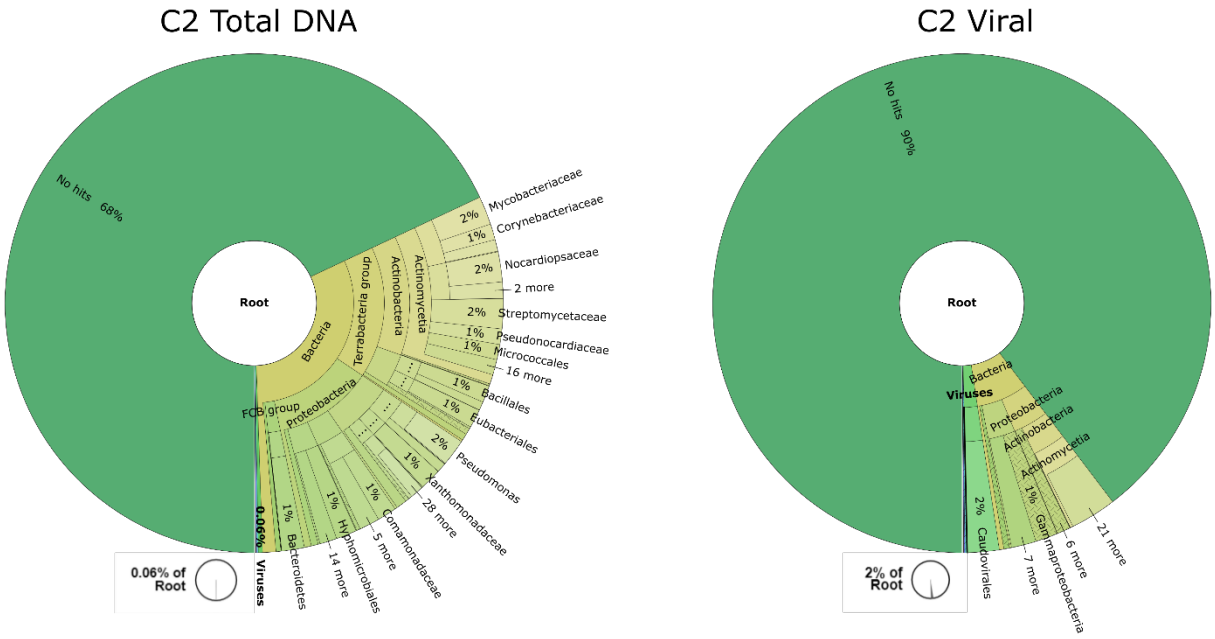

Note: Flowchart created using Miro (https://miro.com/). Own work.

### 4.4.2. Overall diversity and abundance

Two datasets were generated for each sample in this study: total DNA metagenomics and virome metagenomics, which involves the sequencing of a sample enriched for the viral fraction in the composting and soil microbiome. Total DNA datasets may provide essential information about bacterial hosts in the samples, while virome datasets improve the recovery of high-quality and complete phage genomes. Virome datasets presented about two times more identifiable viral sequences than their respective total DNA-related samples (see the example for sample C2 in Figure 4.2). Interestingly, the proportion of sequences with no hits against the Kraken 2 standard database was higher in the virome datasets compared to their total DNA pairs. As environmental prokaryotic viruses are fairly diverse and not well represented in public datasets, we speculate that sequences with no hits may be unknown phages. Krona

diversity plots for all samples are available as Supplementary Files at https://github.com/LaboratorioBioinformatica/vMAGS.

**Figure 4.2**: Diversity plot for Kraken2 taxonomic identification in datasets of the C2 composting sample. The total DNA dataset was generated from total DNA extraction and sequencing, whereas the viral dataset was generated using a protocol for viral enrichment.




Note: Images were generated using Krona tools (Ondov et al., 2011).

Notably, bacterial species typically reported as those present in thermophilic composting samples (Antunes et al., 2016; Braga et al., 2021) were also among the top identified species in our composting samples. For instance, *Thermobifida fusca, Thermobispora bispora, and Rhodothermus marinus* were among the 10 most frequent hits. For soil samples, *Streptomyces lividans, Bradyrhizobium diazoefficiens, and Sorangium cellulosum* were among the 10 most frequent hits (Table 4.2). Detailed count and abundance estimates for all identified species in all samples studied in this thesis are provided in Supplementary Table S1.

**Table 4.2**: Summary of the five bacterial species with most reads assigned by Kraken 2 after Bracken normalization.

| ***C1 composting total DNA metagenomics*** | | |
|---|---|---|
| *Species* | Estimated number of reads | Fraction of the total (%) |
| *Escherichia coli* | 161,842 | 13.936 |
| *Escherichia fergusonii* | 77,760 | 6.696 |
| *Pseudomonas mendocina* | 73,542 | 6.332 |
| *Comamonas kerstersii* | 73,128 | 6.297 |
| *Caldibacillus thermoamylovorans* | 29,674 | 2.555 |
| ***C2 composting total DNA metagenomics*** | | |
| *Species* | Estimated number of reads | Fraction of the total (%) |
| *Thermobifida fusca* | 124,706 | 5.979 |
| *Mycolicibacterium hassiacum* | 51,546 | 2.471 |
| *Comamonas kerstersii* | 49,003 | 2.349 |
| *Rhodothermus marinus* | 37,843 | 1.814 |
| *Klebsiella pneumoniae* | 33,910 | 1.626 |
| ***S2 soil total DNA metagenomics*** | | |
| *Species* | Estimated number of reads | Fraction of the total (%) |
| *Streptomyces lividans* | 3,235 | 1.734 |
| *Bradyrhizobium diazoefficiens* | 2,149 | 1.152 |
| *Delftia acidovorans* | 1,871 | 1.003 |
| *Sorangium cellulosum* | 1,598 | 0.856 |
| *Rhodoplanes sp. Z2-YC6860* | 1,587 | 0.85 |

Note: Own work.

Table 4.3 shows known phage species, with the highest number of sequencing reads assigned by Kraken2 after normalization using Bracken. The higher counts of reads for composting samples (C1 and C2) when compared to soil samples may indicate a higher abundance and that composting phages may be more similar to the NCBI Refseq database of known phages than soil phages.

**Table 4.3**: Summary of the five phage species with most reads assigned by Kraken 2 after Bracken normalization.

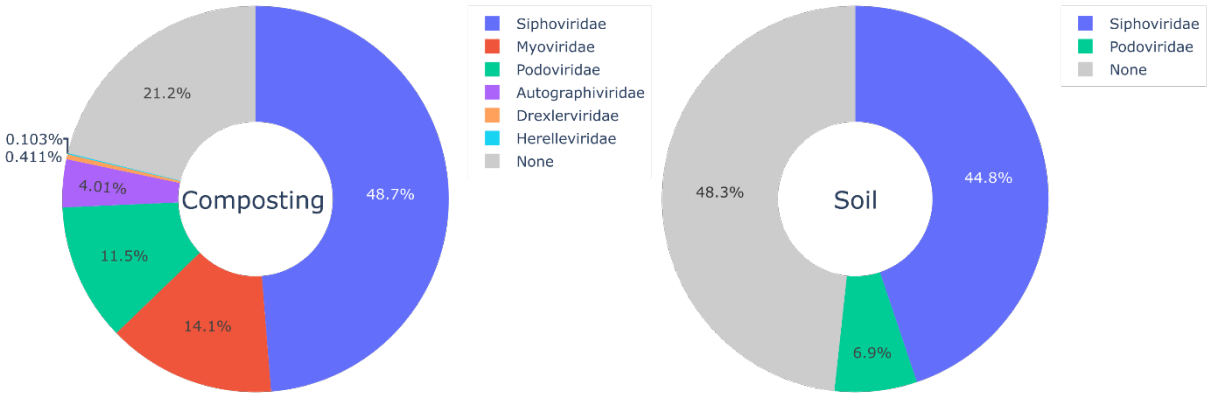
| ***C1 composting virome metagenomics*** | | |
|---|---|---|
| *Species* | Estimated number of reads | Fraction of the total (%) |
| *Lactococcus phage M5938* | 738 | 0.381 |
| *Lactococcus phage D4412* | 346 | 0.179 |
| *Aeromonas phage LAh_7* | 230 | 0.119 |
| *Paenibacillus phage PG1* | 153 | 0.079 |
| *Erwinia phage pEp_SNUABM_08* | 152 | 0.078 |

| *C2 composting virome metagenomics* | | |
|---|---|---|
| *Species* | Estimated number of reads | Fraction of the total (%) |
| *Streptomyces phage Manuel* | 6,669 | 1.38 |
| *Mycobacterium phage Wile* | 6,463 | 1.338 |
| *Salmonella phage 64795_sal3* | 2,638 | 0.546 |
| *Salmonella phage IME207* | 2,251 | 0.466 |
| *Mycobacterium phage Sheen* | 1,722 | 0.356 |
| ***S1 soil virome metagenomics*** | | |
| *Species* | Estimated number of reads | Fraction of the total (%) |
| *Mycobacterium phage Tonenili* | 4 | 0.001 |
| *Enterobacterial phage mEp390* | 6 | 0.001 |
| *Arthrobacter phage DrYang* | 2 | 0.0003 |
| *Gordonia phage Kenosha* | 3 | 0.0005 |
| *Mycobacterium phage Aminay* | 2 | 0.0003 |
| ***S2 soil virome metagenomics*** | | |
| *Species* | Estimated number of reads | Fraction of the total (%) |
| *Streptomyces phage Wofford* | 9 | 0.002 |
| *Treponema phage denis* | 4 | 0.001 |
| *Mycobacterium phage Serendipitous* | 3 | 0.001 |
| *Microbacterium phage Phedro* | 3 | 0.001 |
| *Microbacterium phage Microbacterium phage KaiHaiDragon* | 5 | 0.001 |

Note: Own work.

Putative phage genomes were classified in one of the eight families of the order *Caudovirales* using PhaGCN; the results are shown in Figure 4.3. The majority of putative phage genomes were predicted to belong to the three main families in *Caudovirales*: *Siphoviridae, Myoviridae,* and *Podoviridae* (738 of 1,002). Interestingly, 44 genomes from composting samples were assigned to recently created families (https://talk.ictvonline.org/), namely *Autographiviridae* (39), *Drexlerviridae* (4), and *Herelleviridae* (1). Approximately 21% of the composting putative phage genomes were not assigned to any family. This fraction was higher for the putative phage genomes recovered from the soil samples (48%). A detailed table with individual predictions for each bin is available in Supplementary Table S2.

**Figure 4.3**: Taxonomic affiliation at the family level of the 1,002 putative phage bins according to PhaGCN predictions.




Note: Figures generated using the Plotly library (https://plotly.com/) from Python3. Own work.

### 4.4.3. Recovered phage genomes and quality assessments

A total of 1,647 bins were generated from the two composts and two soil samples from the Sao Paulo Zoo Park. These bins were submitted to phage prediction using MARVEL, indicating 1,052 bins belonging to phage genomes. Bins were filtered by size (≤ 10 kbp), and redundancy among samples was removed using ANI clustering to yield 1,002 non-redundant putative phage bins. The details are listed in Table 4.4.

**Table 4.4:** Summary of bins recovered and remaining after each step of the analyses.

| Sample | Total bins | Predicted as phages using MARVEL | Larger than 10 kbp | Deduplicated by ANI clustering |
|---|---|---|---|---|
| **C1** | 404 | 258 | 244 | 244 |
| **C2** | 1094 | 764 | 730 | 729 |
| **S1** | 60 | 15 | 14 | 14 |
| **S2** | 89 | 15 | 15 | 15 |

Note: Own work.

In terms of general metrics, bins ranged in length from 10,011 bp to 150,302 bp. The average size was 27,680 bp (±19,644 bp), and the median was 21,314 bp. The number of contigs ranged from 1 to 33, with a median of three contigs. The largest single contig recovered in a bin was 64,814 bp, and the smallest was 2,028 bp (Table 4.5).

**Table 4.5**: Descriptive metrics of the putative phage bins recovered by MARVEL.

| | Genome size (bp) | Number of contigs | N50 (contigs) | Longest contig (bp) | GC (%) |
|---|---|---|---|---|---|
| *Mean* | 27,681 | 4.6 | 12,834 | 14,353 | 54.4 |
| *std* | 19,644 | 4.4 | 12,211 | 11,960 | 10.8 |
| *Min* | 10,011 | 1 | 1,599 | 2,028 | 27.6 |
| *Q1* | 13,981 | 2 | 4,179 | 6,268 | 45.4 |
| *Median* | 21,314 | 3 | 8,512 | 10,561 | 56.8 |
| *Q3* | 35,123 | 6 | 16,020 | 17,451 | 63.6 |
| *Max* | 150,302 | 33 | 64,814 | 64,814 | 72.4 |

Note: Own work.

Putative phage bins were also assessed using several quality metrics provided by CheckV and CheckM. For instance, bins were presented from 7 to 207 genes, with a median of 30 genes. The mean fraction of genes that CheckV recognized as being of viral origin was 39.9% (± 20.6%). Completeness ranged from 2.8% to 100%, with a mean of 50.4% and a median of 42.9% (Table 4.6). Bins presented an average of 0.7% contamination, with 17 bins predicted as a provirus. As CheckV finds delimitations for the provirus, the average size of the viral portion was 19,959 bp. Additionally, CheckM searched for bacterial marker genes in the putative phage bins, and only three bins returned one marker gene each; one (C2.bin.317) was also predicted to be a provirus by CheckV.

**Table 4.6**: Descriptive metrics of numeric variables for assessing quality of putative phage bins obtained by CheckV.

| | Gene count | Number of viral genes | Viral genes (%) | Number of host genes | Host Genes (%) | Completeness (%) | Contamination (%) |
|---|---|---|---|---|---|---|---|
| *Mean* | 40.0 | 14.9 | 39.9 | 0.7 | 2.5 | 50.4 | 0.7 |
| *Std* | 30.1 | 13.0 | 20.6 | 2.6 | 9.3 | 27.8 | 6.3 |
| *Min* | 7.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.8 | 0.0 |
| *Q1* | 19.0 | 7.0 | 25.0 | 0.0 | 0.0 | 27.0 | 0.0 |
| *Median* | 30.0 | 12.0 | 36.5 | 0.0 | 0.0 | 42.9 | 0.0 |
| *Q3* | 51.0 | 19.0 | 53.3 | 1.0 | 1.1 | 74.3 | 0.0 |
| *Max* | 207.0 | 186.0 | 100.0 | 44.0 | 92.9 | 100.0 | 82.5 |

Note: Own work.

CheckV automatically classifies putative phage bins into one of four categories: low-quality, medium-quality, high-quality, and complete. The fifth “Not Determined” category is used for bins in which CheckV finds little if any evidence of completeness, contamination, or similarity to known viral genomes. According to CheckV set of criteria, 15 (1.5%) bins were defined as “Complete,” 118 (11.7%) as “High-quality,” 293 (29.2%) as “Medium-quality,” 559 (55.8%) as “Low-quality,” and 17 (1.7%) as “Not determined.” Moreover, CheckV also uses criteria defined by the Minimum Information about an Uncultivated Virus Genome (MIUViG) (Roux et al., 2018) to classify putative phage bins in two categories: “High-Quality” and “Genome-fragment.” CheckV classified 133 (13.8%) putative phage as “High-Quality” and 869 (86.7%) as “Genome-fragment.” Putative phage genomes placed in “Complete” and “High-quality” CheckV categories (133) fully agreed with the “High-quality” category from MIUViG standards (Figure 4.4). A complete and detailed list of bins with all the evidence gathered by CheckV and CheckM approaches is available in Supplementary Table S3.

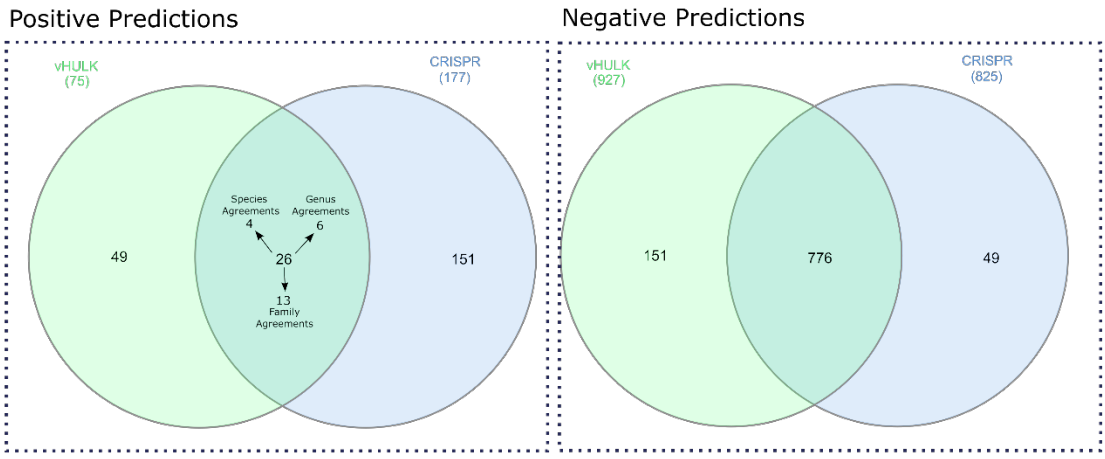
**Figure 4.4**: Total of 1,002 putative phage bins classified according to the CheckV and MIUViG quality criterion sets.


Note: Images generated using the PyWaffle library (https://github.com/gyli/PyWaffle) from Python 3. Own work.

### 4.4.4. Phage host predictions

Furthermore, vHULK and matches to CRISPR spacers were used to predict the putative host to the 1,002 phage bins. A total of 226 (22.5%) bins yielded a positive prediction using either vHULK or CRISPR and 26 (2.6%) using both vHULK and CRISPR approaches. Among the 26 positive predictions, there were 23 bins (88.4%) of agreement at the minimum family level. vHULK and CRISPR produced rather disjointed predictions (88.4% of the total positive predictions), possibly indicating that the two approaches are suitable for use together. If negative predictions are considered, the vHULK and CRISPR agreements are 80% (Figure 4.5).

**Figure 4.5**: Venn diagrams showing intersections between positive and negative host predictions using the vHULK and CRISPR spacer match approaches.

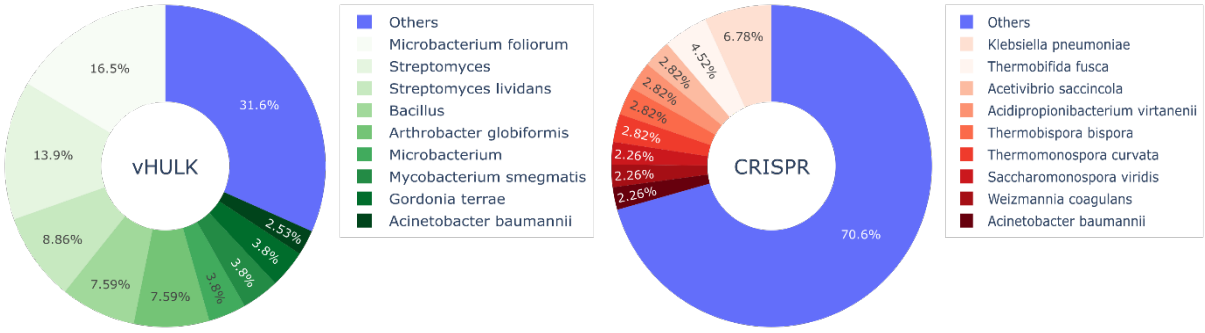



Note: Vector images generated using Inkscape (https://inkscape.org/). Own work.

When predictions by environment type were considered (composting or soil), none of the 29 putative phage genomes recovered from soil yielded any kind of prediction by both vHULK and CRISPR spacer matches. vHULK's most frequent hosts for composting putative phage genomes were *Microbacterium foliorum, Streptomyces* spp*., Streptomyces lividans, Bacillus* spp*.,* and *Arthobacter globiformis*. vHULK, in contrast to the CRISPR approach, may predict a genus or a species as a host. This depends on the probability scores provided by each vHULK internal model (softmax species and softmax genus). By contrast, the most frequent host predictions using CRISPR were *Klebsiella pneumoniae, Thermobifida fusca, Acetivibrio saccincola, Thermobispora bispora,* and *Thermobispora curvata* (Figure 4.6). The most notable

composting bacteria were among the most frequent predictions of hosts using CRISPR spacer matches, which is a coherent result. These results agree with the diversity and abundance results presented in Section 4.4.2. Unfortunately, these thermophilic composting bacteria are not among the hosts vHULK was trained to predict. A detailed list of host predictions with CRISPR and vHULK evidence is available as Supplementary Table S4.

**Figure 4.6**: Host predictions using vHULK and CRISPR spacer matches in the two composting samples.




Note: Images generated using the Plotly library (https://plotly.com/) from Python 3. Own work.

## 4.4.5. Case study with complete phage genomes

Fifteen genomes were recovered in a single contig and had direct terminal repeats (DTRs), providing compelling evidence for a complete phage genome. Among these genomes, the fraction of unknown genes, that is, genes with no similarity to any of the viral HMM databases used by CheckV (Nayfach, Camargo, et al., 2021), ranged from 17% to 82%. Additionally, we searched the IMG/VR environmental viruses sequence database for similar genomes, and the results showed hits satisfying E-value criterion (≤ 10e-5) but low query coverage (below 30%). However, the best-scoring hits were investigated regarding the environmental source of their metagenomes; the results are shown in Table 4.7. Interestingly, the best-scoring hits belonged to environments remarkably similar to composting. For instance, a similar phage genome was recovered from the western lowland gorilla feces of a Zoo Park in Wisconsin, USA. Primate feces are a component of São Paulo Zoo composting. In addition, the four phage genomes recovered in our study presented the best-scoring hits to phages from

steer compost in Utah, USA. PhaGCN also identified phages from the *Siphoviridae*, *Myoviridae*, and *Autographiviridae* families. *Autographiviridae* is also known as the T7 supergroup, previously placed in the *Podoviridae* family. If vHULK's best-scoring host predictions were considered for the complete phage genomes, a total of nine different hosts would be predicted (Table 4.7). Detailed information about the 15 complete phage genomes is available in Supplementary Table S5.

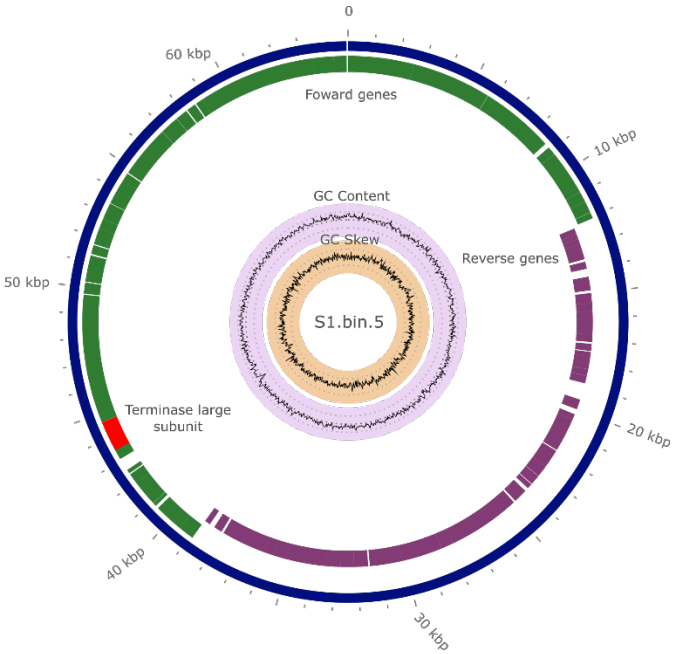
**Table 4.7**: Description of the 15 phage genomes classified as complete by CheckV owing to the presence of DTRs. Genomes were searched against the IMG/VR; the environment and location of the best-scoring hit are shown.

| Genome | Length (bp) | Novelty genes | vHULK Host | vHULK score | Family by PhaGCN | IMG/VR Viral | IMG/VR Location |
|---|---|---|---|---|---|---|---|
| **C1.bin.109** | 58946 | 67.9% | *Microbacterium* | 0.536 | *Siphoviridae* | Leaf-cutter ant dump | WS/USA |
| **C1.bin.204** | 42636 | 47.9% | *Acinetobacter* | 0.792 | *Myoviridae* | Fresh water | Congo |
| **C1.bin.211** | 12428 | 62.5% | *Pseudomonas* | 0.071 | *Myoviridae* | Steer compost | UT/USA |
| **C1.bin.213** | 12401 | 68.7% | *Pseudomonas* | 0.118 | *Myoviridae* | Human skin | PA/USA |
| **C1.bin.342** | 42603 | 17.1% | *Salmonella enterica* | 0.681 | *Siphoviridae* | Western lowland gorilla feces | WS/USA |
| **C1.bin.343** | 40014 | 55.6% | *Ralstonia solanacearum* | 0.583 | *Autographiviridae* | Human feces | China |
| **C1.bin.401** | 46313 | 45.5% | *Pseudomonas aeruginosa* | 0.391 | *Myoviridae* | Silver grass rhizosphere | MI/USA |
| **C1.bin.86** | 41992 | 69.5% | *Pseudomonas* | 0.231 | *Autographiviridae* | -- | -- |
| **C2.bin.27** | 31183 | 74.0% | *Bacillus* | 0.151 | *Siphoviridae* | Feces from healthy child | CA/USA |
| **C2.bin.307** | 60752 | 59.3% | *Rhodococcus* | 0.394 | *Siphoviridae* | Steer compost | UT/USA |
| **C2.bin.477** | 58605 | 73.9% | *Microbacterium* | 0.368 | *Siphoviridae* | Steer compost | UT/USA |
| **C2.bin.547** | 49049 | 72.3% | *Microbacterium* | 0.501 | *Siphoviridae* | Leaf-cutter ant dump | WS/USA |
| **C2.bin.684** | 41881 | 31.1% | *Ralstonia solanacearum* | 0.480 | *Autographiviridae* | Wastewater from oil extraction | Canada |
| **C2.bin.742** | 64690 | 56.9% | *Gordonia* | 0.470 | *Siphoviridae* | Steer compost | UT/USA |
| **S1.bin.5** | 64814 | 82.8% | *Sinorhizobium* | 0.093 | *Siphoviridae* | Soil from agricultural site | NY/USA |

Note: Own work.

Complete genomes were also annotated using PATRIC, and reports are available as Supplementary Files at https://github.com/LaboratorioBioinformatica/vMAGS. The only complete phage genome recovered from soil is presented here as a case study. A total of 89 ORFs were found, with only one protein presenting a known function: a terminase large subunit of 401 amino acids. The remaining 88 proteins were annotated as hypothetical proteins or with the generic name of "phage protein." A circular genome plot of the genome S1.bin.5 is shown in Figure 4.7.

**Figure 4.7**: Circular genome plot of the single complete genome recovered from soil (S1.bin.5). The tracks are displayed as concentric rings, from outermost to innermost: Position, Contig, CDS-forward, CDS-reverse, GC Content, and GC Skew. Highlighted in red is the only gene whose function could be annotated.




Note: Image generated using the PATRIC online platform and modified using Inkscape. Own work.

## 4.5. Discussion

In this chapter, we present a compendium of 1,002 non-redundant phage genomes recovered from two composting and two soil samples (973 and 29, respectively). From these, 133 presented compelling evidence for being complete or of high quality according to the MIUViG standards. To achieve these results, a complete workflow was developed to address important challenges in the area. First, we collected samples and tested a viral enrichment protocol to improve the quantity and quality of sequences generated through metagenomic sequencing. This was a significant challenge, for which the scientific literature provides many alternatives (Thurber et al., 2009). It is important to note that each sample type (e.g., soil and feces) presents its own challenges. In our case (composting and soil), filtration with 0.4-µM low-binding syringe filters and concentration with 10-kDa centrifuge filters presented optimized results for enriching viral sequences.

Once the sequencing datasets were efficiently generated, steps of quality control, assembly, and binning needed to be piped and the parameters adjusted accordingly. We considered several protocols in the literature (Nayfach, Camargo, et al., 2021; Overholt et al., 2020; Roux et al., 2018) that we considered to develop our own pipeline. As an important result of this thesis, scripts for the automated steps are publicly available at: https://github.com/LaboratorioBioinformatica/vMAGS.

Predicting whether a bin or contig belongs to a viral genome was a difficult task when we began this work. Nonetheless, multiple tools were developed to address this issue; Virsorter2 and VirFinder (Guo et al., 2021; Ren et al., 2020) are suitable examples of such tools. MARVEL (Amgarten et al., 2018) was developed and used in the pipeline. Most bins predicted by MARVEL as phage genomes had overwhelming authenticity evidence, passing through quality standards as defined by CheckV and MIUViG.

With putative phage genomes, it is particularly important to annotate and characterize these sequences to perform further studies. Additionally, vHULK was developed to predict perhaps the most important feature of a virus: the host it infects. Currently, there are several tools and approaches for predicting a virus host, such as RaFAH (Coutinho et al., 2021), VirHostMatcher-Net, (Wang et al., 2020), and HostPhinder (Villarroel et al., 2016). As some of the cited approaches were already compared to vHULK in Chapter 3, we decided to only use vHULK and CRISPR spacer matches to predict the host in putative phage genomes from our dataset. The results show little overlap between predictions provided by vHULK and CRISPR-spacer matches, suggesting their use as complementary approaches. However, vHULK provides predictions for a small portion of the bins (75 out of 1,002). Sensitivity is a major problem in any methodology proposed to address the host prediction problem. The vHULK score thresholds were set to be very conservative by default (≥ 0.5, in a multiclass setting of 61 possible genera and 52 possible species) and may be changed by the user to increase the number of predictions. By contrast, a random classifier would generate prediction scores of approximately 0.016 for predictions at the genus level and 0.019 at the species level. We are also studying ways to expand the number of hosts presented in the training datasets of vHULK. It is possible to use datasets of environmental phage genomes, whose strong host evidence is present (e.g., CRISPR-spacer matches).

To the best of our knowledge, this is the first compendium of vMAGs from composting reported at the time of writing. Soil samples are a diversity hotspot, with few studies reporting vMAGs in this environment. Han et al. investigated 19 soil samples across China, focusing on the way by which viral elements influence the phosphorus geochemical cycle (Han & Rohwer, 2021). To achieve this, they only performed an assembly step (without binning) and submitted contigs larger than 1,000 bp for CheckV quality assessment (without using a predictor of phage sequences, such as Virsorter2 or MARVEL). The authors claimed that 64,579 viral contigs were obtained within an astonishing diversity of 27 viral families. A second study by Schulz et al. reported an interesting hidden diversity of giant viruses in soil samples from the Harvard Forest (Schulz et al., 2018). Sixteen novel giant viruses were reported, significantly expanding the known phylogenetic diversity of the *Mimiviridae* family. It is worth emphasizing that both studies had different targets of investigation regarding viral diversity. The former focused on the viral diversity of ssDNA and dsDNA viruses, whereas the latter focused on giant *Mimivirus*. As an important limitation of MARVEL stated in Chapter 2, it was only trained to recognize sequences from tailed phages (*Caudovirales* order). Consequently, our study targets only the viral diversity of tailed phages.

If host-associated samples such as the human gut are considered, 3,738 (Benler et al., 2021), 8,000, (Parks et al., 2017), and 189,680 (Nayfach, Páez-Espino, et al., 2021) vMAGs have been recovered; impressive numbers obtained from excellent datasets of samples (up to 11,810 human stool metagenomes mined in Nayfach et al.). The Metasub project generated a compendium of 10,928 recovered environmental viruses from a total of 4,728 samples and 8 trillion base pairs sequenced (Danko et al., 2021). Here, we show a modest number of recovered genomes compared to these studies. However, it should be noted that there were four samples analyzed in our study. Only two composting samples generated a total of 973 putative phage genomes (245 in C1 and 728 in C2). Counts of recovered genomes for soil samples were lower (14 in S1 and 15 in S2); therefore, we hypothesize that the high diversity of the microbial community is linked with these results. More sequencing data are needed to efficiently access vMAGs in highly diverse samples.

Altogether, the methodology and tools developed in this work successfully address relevant challenges in the area, contributing to the recovery and

characterization of thousands of new environmental virus genomes of the so-called viral dark matter.

## Supplementary Material Availability

Supplementary tables and files cited in this chapter are publicly available in: https://github.com/LaboratorioBioinformatica/vMAGS.

## References


Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amgarten, D., Braga, L. P. P., da Silva, A. M., & Setubal, J. C. (2018). MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. *Frontiers in Genetics*, *0*(AUG), 304. https://doi.org/10.3389/FGENE.2018.00304

Amgarten, D., Iha, B. K. V., Piroupo, C. M., Silva, A. M. da, & Setubal, J. C. (2020). vHULK, a new tool for bacteriophage host prediction based on annotated genomic features and deep neural networks. *BioRxiv*, 2020.12.06.413476. https://doi.org/10.1101/2020.12.06.413476

Amgarten, D., Martins, L. F., Lombardi, K. C., Antunes, L. P., de Souza, A. P. S., Nicastro, G. G., Kitajima, E. W., Quaggio, R. B., Upton, C., Setubal, J. C., & da Silva, A. M. (2017). Three novel Pseudomonas phages isolated from composting provide insights into the evolution and diversity of tailed phages. *BMC Genomics*, *18*(1). https://doi.org/10.1186/s12864-017-3729-z

Antunes, L. P., Martins, L. F., Pereira, R. V., Thomas, A. M., Barbosa, D., Lemos, L. N., Silva, G. M. M., Moura, L. M. S., Epamino, G. W. C., Digiampietri, L. A., Lombardi, K. C., Ramos, P. L., Quaggio, R. B., de Oliveira, J. C. F., Pascon, R. C., Cruz, J. B. da, da Silva, A. M., & Setubal, J. C. (2016). Microbial community structure and dynamics in thermophilic composting viewed through metagenomics and metatranscriptomics. *Scientific Reports 2016 6:1*, *6*(1), 1–13. https://doi.org/10.1038/srep38915

Bastide, M., & McCombie, W. R. (2007). Assembling Genomic DNA Sequences with PHRAP. *Current Protocols in Bioinformatics*, *17*(1). https://doi.org/10.1002/0471250953.bi1104s17

Benler, S., Yutin, N., Antipov, D., Rayko, M., Shmakov, S., Gussow, A. B., Pevzner, P., & Koonin, E. v. (2021). Thousands of previously unknown phages discovered in whole-community human gut metagenomes. *Microbiome 2021 9:1*, *9*(1), 1–17. https://doi.org/10.1186/S40168-021-01017-W

Bolduc, B., Jang, H. bin, Doulcier, G., You, Z.-Q., Roux, S., & Sullivan, M. B. (2017). vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ*, *5*(5), e3243. https://doi.org/10.7717/PEERJ.3243

Braga, L. P. P., Pereira, R. V., Martins, L. F., Moura, L. M. S., Sanchez, F. B., Patané, J. S. L., da Silva, A. M., & Setubal, J. C. (2021). Genome-resolved metagenome and metatranscriptome analyses of thermophilic composting reveal key bacterial players and their metabolic interactions. *BMC Genomics 2021 22:1*, *22*(1), 1–19. https://doi.org/10.1186/S12864-021-07957-9

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., Olson, R., Overbeek, R., Parrello, B., Pusch, G. D., Shukla, M., Thomason, J. A., Stevens, R., Vonstein, V.,

Wattam, A. R., & Xia, F. (2015). RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports*, *5*. https://doi.org/10.1038/srep08365

Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D., & Lawley, T. D. (2021). Massive expansion of human gut bacteriophage diversity. *Cell*, *184*(4), 1098-1109.e9. https://doi.org/10.1016/J.CELL.2021.01.029

Chen, I.-M. A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Huntemann, M., Varghese, N., White, J. R., Seshadri, R., Smirnova, T., Kirton, E., Jungbluth, S. P., Woyke, T., Eloe-Fadrosh, E. A., Ivanova, N. N., & Kyrpides, N. C. (2019). IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Research*, *47*(D1), D666–D677. https://doi.org/10.1093/NAR/GKY901

Coutinho, F. H., Zaragoza-Solas, A., López-Pérez, M., Barylski, J., Zielezinski, A., Dutilh, B. E., Edwards, R., & Rodriguez-Valera, F. (2021). RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. *Patterns*, *2*(7), 100274. https://doi.org/10.1016/J.PATTER.2021.100274

Daniel, R. (2005). The metagenomics of soil. *Nature Reviews Microbiology 2005 3:6*, *3*(6), 470–478. https://doi.org/10.1038/nrmicro1160

Danko, D., Bezdan, D., Afshin, E. E., Ahsanuddin, S., Bhattacharya, C., Butler, D. J., Chng, K. R., Donnellan, D., Hecht, J., Jackson, K., Kuchin, K., Karasikov, M., Lyons, A., Mak, L., Meleshko, D., Mustafa, H., Mutai, B., Neches, R. Y., Ng, A., … Zubenko, S. (2021). A global metagenomic map of urban microbiomes and antimicrobial resistance. *Cell*, *184*(13), 3376-3393.e17. https://doi.org/10.1016/J.CELL.2021.05.002

Davis, J. J., Wattam, A. R., Aziz, R. K., Brettin, T., Butler, R., Butler, R. M., Chlenski, P., Conrad, N., Dickerman, A., Dietrich, E. M., Gabbard, J. L., Gerdes, S., Guard, A., Kenyon, R. W., MacHi, D., Mao, C., Murphy-Olson, D., Nguyen, M., Nordberg, E. K., … Stevens, R. (2020). The PATRIC Bioinformatics Resource Center: Expanding data and analysis capabilities. *Nucleic Acids Research*, *48*(D1), D606–D612. https://doi.org/10.1093/nar/gkz943

Gilbert, J. A., Jansson, J. K., & Knight, R. (2014). The Earth Microbiome project: successes and aspirations. *BMC Biology 2014 12:1*, *12*(1), 1–4. https://doi.org/10.1186/S12915-014-0069-1

Guo, J., Bolduc, B., Zayed, A. A., Varsani, A., Dominguez-Huerta, G., Delmont, T. O., Pratama, A. A., Gazitúa, M. C., Vik, D., Sullivan, M. B., & Roux, S. (2021). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome 2021 9:1*, *9*(1), 1–13. https://doi.org/10.1186/S40168-020-00990-Y

Han, L.-L., & Rohwer, F. (2021). Distribution of soil viruses across China and their potential role in phosphorous metabolism. *Research Square*, 1–21. https://doi.org/10.21203/rs.3.rs-361706/v1

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, *7*(7), e7359. https://doi.org/10.7717/PEERJ.7359

Krishnamurthy, S. R., & Wang, D. (2017). Origins and challenges of viral dark matter. *Virus Research*, *239*, 136–142. https://doi.org/10.1016/J.VIRUSRES.2017.02.002

Lemos, L. N., Pereira, R. v., Quaggio, R. B., Martins, L. F., Moura, L. M. S., da Silva, A. R., Antunes, L. P., da Silva, A. M., & Setubal, J. C. (2017). Genome-Centric Analysis of a Thermophilic and Cellulolytic Bacterial Consortium Derived from Composting. *Frontiers in Microbiology*, *0*(APR), 644. https://doi.org/10.3389/FMICB.2017.00644

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Marcelino, V. R., Clausen, P. T. L. C., Buchmann, J. P., Wille, M., Iredell, J. R., Meyer, W., Lund, O., Sorrell, T. C., & Holmes, E. C. (2020). CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. *Genome Biology 2020 21:1*, *21*(1), 1–15. https://doi.org/10.1186/S13059-020-02014-2

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12. https://doi.org/10.14806/EJ.17.1.200

Martins, L. F., Antunes, L. P., Pascon, R. C., Oliveira, J. C. F. de, Digiampietri, L. A., Barbosa, D., Peixoto, B. M., Vallim, M. A., Viana-Niero, C., Ostroski, E. H., Telles, G. P., Dias, Z., Cruz, J. B. da, Juliano, L., Verjovski-Almeida, S., Silva, A. M. da, & Setubal, J. C. (2013). Metagenomic Analysis of a Tropical Composting Operation at the São Paulo Zoo Park Reveals Diversity of Biomass Degradation Functions and Organisms. *PLOS ONE*, *8*(4), e61928. https://doi.org/10.1371/JOURNAL.PONE.0061928

Moran, M. A. (2015). The global ocean microbiome. *Science*, *350*(6266). https://doi.org/10.1126/SCIENCE.AAC8455

Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., & Kyrpides, N. C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology*, *39*(5), 578–585. https://doi.org/10.1038/s41587-020-00774-7

Nayfach, S., Páez-Espino, D., Call, L., Low, S. J., Sberro, H., Ivanova, N. N., Proal, A. D., Fischbach, M. A., Bhatt, A. S., Hugenholtz, P., & Kyrpides, N. C. (2021). Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nature Microbiology 2021 6:7*, *6*(7), 960–970. https://doi.org/10.1038/s41564-021-00928-6

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I. M., Huntemann, M., Palaniappan, K., Ladau, J., Mukherjee, S., Reddy, T. B. K., Nielsen, T., Kirton, E., Faria, J. P., Edirisinghe, J. N., Henry, C. S., … Eloe-Fadrosh, E. A. (2021). A genomic catalog of Earth's microbiomes. *Nature Biotechnology*, *39*(4), 499–509. https://doi.org/10.1038/s41587-020-0718-6

Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, *12*(1), 1–10. https://doi.org/10.1186/1471-2105-12-385

Overholt, W. A., Hölzer, M., Geesink, P., Diezel, C., Marz, M., & Küsel, K. (2020). Inclusion of Oxford Nanopore long reads improves all microbial and viral metagenome-assembled genomes from a complex aquifer system. *Environmental Microbiology*, *22*(9), 4000–4013. https://doi.org/10.1111/1462-2920.15186

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055. https://doi.org/10.1101/GR.186072.114

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P., & Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology 2017 2:11*, *2*(11), 1533–1542. https://doi.org/10.1038/s41564-017-0012-7

Pourcel, C., Touchon, M., Villeriot, N., Vernadet, J.-P., Couvin, D., Toffano-Nioche, C., & Vergnaud, G. (2020). CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. *Nucleic Acids Research*, *48*(D1), D535–D544. https://doi.org/10.1093/NAR/GKZ915

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics*, *70*(1), e102. https://doi.org/10.1002/cpbi.102

Ren, J., Song, K., Deng, C., Ahlgren, N. A., Fuhrman, J. A., Li, Y., Xie, X., Poplin, R., & Sun, F. (2020). Identifying viruses from metagenomic data using deep learning. *Quantitative Biology*, *8*(1), 64–77. https://doi.org/10.1007/s40484-019-0187-4

Roux, S., Adriaenssens, E. M., Dutilh, B. E., Koonin, E. v, Kropinski, A. M., Krupovic, M., Kuhn, J. H., Lavigne, R., Brister, J. R., Varsani, A., Amid, C., Aziz, R. K., Bordenstein, S. R., Bork, P., Breitbart, M., Cochrane, G. R., Daly, R. A., Desnues, C., Duhaime, M. B., … Eloe-Fadrosh, E. A. (2018). Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nature Biotechnology 2018 37:1*, *37*(1), 29–37. https://doi.org/10.1038/nbt.4306

Roux, S., Hallam, S. J., Woyke, T., & Sullivan, M. B. (2015). Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *ELife*, *4*. https://doi.org/10.7554/elife.08490

Roux, S., Krupovic, M., Daly, R. A., Borges, A. L., Nayfach, S., Schulz, F., Sharrar, A., Matheus Carnevali, P. B., Cheng, J.-F., Ivanova, N. N., Bondy-Denomy, J., Wrighton, K. C., Woyke, T., Visel, A., Kyrpides, N. C., & Eloe-Fadrosh, E. A. (2019). Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nature Microbiology 2019 4:11*, *4*(11), 1895–1906. https://doi.org/10.1038/s41564-019-0510-x

Roux, S., Páez-Espino, D., Chen, I.-M. A., Palaniappan, K., Ratner, A., Chu, K., Reddy, T. B. K., Nayfach, S., Schulz, F., Call, L., Neches, R. Y., Woyke, T., Ivanova, N. N., Eloe-Fadrosh, E. A., & Kyrpides, N. C. (2021). IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Research*, *49*(D1), D764–D775. https://doi.org/10.1093/NAR/GKAA946

Schulz, F., Alteio, L., Goudeau, D., Ryan, E. M., Yu, F. B., Malmstrom, R. R., Blanchard, J., & Woyke, T. (2018). Hidden diversity of soil giant viruses. *Nature Communications*, *9*(1), 1–9. https://doi.org/10.1038/s41467-018-07335-2

Shang, J., Jiang, J., & Sun, Y. (2021). Bacteriophage classification for assembled contigs using graph convolutional network. *Bioinformatics*, *37*(Supplement_1), i25–i33. https://doi.org/10.1093/BIOINFORMATICS/BTAB293

Sunagawa, S., Acinas, S. G., Bork, P., Bowler, C., Eveillard, D., Gorsky, G., Guidi, L., Iudicone, D., Karsenti, E., Lombard, F., Ogata, H., Pesant, S., Sullivan, M. B., Wincker, P., & de Vargas, C. (2020). Tara Oceans: towards global ocean ecosystems biology. *Nature Reviews Microbiology 2020 18:8*, *18*(8), 428–445. https://doi.org/10.1038/s41579-020-0364-5

Thurber, R. v, Haynes, M., Breitbart, M., Wegley, L., & Rohwer, F. (2009). Laboratory procedures to generate viral metagenomes. *Nature Protocols 2009 4:4*, *4*(4), 470–483. https://doi.org/10.1038/nprot.2009.10

Tisza, M. J., & Buck, C. B. (2021). A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proceedings of the National Academy of Sciences*, *118*(23). https://doi.org/10.1073/PNAS.2023202118

Villarroel, J., Kleinheinz, K. A., Jurtz, V. I., Zschach, H., Lund, O., Nielsen, M., & Larsen, M. V. (2016). HostPhinder: A Phage Host Prediction Tool. *Viruses 2016, Vol. 8, Page 116*, *8*(5), 116. https://doi.org/10.3390/V8050116

Wang, W., Ren, J., Tang, K., Dart, E., Ignacio-Espinoza, J. C., Fuhrman, J. A., Braun, J., Sun, F., & Ahlgren, N. A. (2020). A network-based integrated framework for predicting virus–prokaryote interactions. *NAR Genomics and Bioinformatics*, *2*(2). https://doi.org/10.1093/NARGAB/LQAA044

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology 2019 20:1*, *20*(1), 1–13. https://doi.org/10.1186/S13059-019-1891-0

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. https://doi.org/10.1007/s10482-017-0844-4

# **Chapter 5**: Conclusion

To conclude this thesis, I would like to remember the main motivation which has driven all the work developed until this point. Metagenomics has emerged as a prominent field of research and has brought exciting opportunities to reveal and characterize the unknown diversity held by environmental viruses. However, it was clear at the time that several challenges needed to be addressed in order to access all this promising diversity. Metagenomics in general is attached to significant drawbacks regarding the assembly and recovery of genomes in a single continuous sequence. Moreover, viral metagenome associated genomes usually do not come along with information of the host it infects. Without complete genomes and the information of host, an important portion of the knowledge associated with viral dark matter still remains locked. Bearing this in mind, our proposal was to develop tools and protocols that would help addressing these challenges. Looking at the results presented until here, we list as main achievements of this thesis:

1. Feature engineering and model training methods for accurate prediction of viral sequences implemented in a tool, called MARVEL, for prediction of phage bins in a universe of metagenomic sequences.
2. Feature engineering and model training methods for accurate prediction of viral hosts implemented in a tool, called vHULK, for prediction of host in phage genomes recovered from metagenomics datasets.
3. Development of a pipeline to go from raw metagenomic sequencing data to complete or high-quality phage genomes.
4. Recovery of a compendium of 1,002 phage genomes from composting and soil samples, along with basic characterization and host predictions.

It is evident that the compendium of 1,002 newly recovered phages from composting and soil need to be further characterized as for basic microbiological characteristics. As perspectives for future work in our group, we list:

1. Complete functional annotation of the 1,002 phage genomes and their proteins.
2. Deposit of the curated and high-quality genomes in public databases.

3. Wet-lab isolation of new phages in composting and soil samples from the Sao Paulo Zoo Park using bacterial hosts of special interest and levering information gathered from the compendium of 1,002 phage genomes.
4. Create a biobank of isolated phages along with genomes for further characterization aiming phage therapy application.

Tools and pipelines developed here are all publicly available to the scientific community in the hope they will help researchers with similar needs and challenges to overcome.