Project Specifications

Development of a Tool for the Analysis of the Co-localization of Anti-phage Defence Systems and Prophages in Bacterial Genomes

Auré Magloire Agbomakounzo
p2411922
Mariam El Khattabi
p2408949
Lorcan Brenders
p2104908

Master's Degree in Bioinformatics
Bioinformatic project 2
February 7, 2025

Contents

1 Project presentation

1.1 Context

The genus Pseudomonas is one of the most ecologically important bacterial groups, including commensal species of both animals and humans, such as *Pseudomonas aeruginosa* (Silby et al., 2011). It is an opportunistic pathogen and represents a particularly relevant target for phage therapy, as it is part of the ESKAPE group, a set of highly virulent and antibiotic-resistant bacterial pathogens. Those six bacteria are known for their ability to resist multiple antibiotics. They are grouped under that acronym and refer to *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter Baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter Species.*(Mortensen et al., 2021). These bacteria are notorious for their ability to "escape" the effects of commonly used antibiotics due to their increasing multidrug resistance (MDR).

As opportunistic pathogens are frequently associated with hospital environments, members of the ESKAPE group pose a significant threat to immunocompromised and critically ill patients, contributing to a high mortality rate. In particular, *P. aeruginosa* and *S. aureus* evolve among the most prevalent bacteria found in biofilms within healthcare settings. Those biofilms keep the bacteria in a protective and nourishing barrier that allows growth and adaptation to various environments, making infections even more difficult to treat (Gheorghita et al., 2023). Because of this, the World Health Organisation (WHO) has established a global priority pathogen list (PPL) to guide research and development of new antibiotics, classifying them into three levels: critical, high, and medium. Four of the ESKAPE pathogens fall under the critical priority category, while the remaining two are considered high priority, highlighting the global concern surrounding their resistance and the need for alternative therapeutic strategies such as phage therapy (*WHO Bacterial Priority Pathogens List 2024*, 2024). These bacteria have great metabolic diversity and can adapt well, allowing them to have a wide distribution in freshwater and soils (De Smet et al., 2017).

Bacteriophages, or phages, are viruses that infect bacteria. They are the most abundant entities on Earth, with 4.8×10^{31} phage particles. Phages can be found anywhere bacteria exist, from aquatic and soil environments to extreme ecosystems and microbiomes (Comeau et al., 2008; Suttle, 2005). The phages that infect *Pseudomonas* are also highly diverse, and although they have many differences, most are "global," meaning they are found in different parts of the world and are very successful, distributing globally. Phages are highly specific to a given bacterial species and even to specific strains of this species, thus supporting phage therapy with the advantage of specificity (De Smet et al., 2017).

Phages move by diffusion in the environment, without autonomous movement; they collide with bacteria randomly, and then the phage reversibly binds to the primary receptor on the host surface through noncovalent interactions. If the phage binds to secondary

receptors, irreversible adsorption starts. This binding is stronger, and the phage can no longer detach from the cell (Dennehy & Abedon, 2021).

Bacteria develop resistance in response to the strong selective pressure imposed by phages, bacteria have developed a wide range of protective mechanisms. These defences operate at different levels: externally, by modifying or hiding surface structures that serve as phage receptors, thus preventing adsorption; and internally, through various defence systems that can interfere with different stages of the phage replication cycle. They have mechanisms that prevent adsorption, such as reducing the synthesis of pili, where phages attach. However, phages respond by adapting their receptor-binding proteins, allowing them to continue infecting. Also, bacteria can modify their polysaccharide structures on the cell membrane, preventing phages from recognizing and attaching to them (Dennehy & Abedon, 2021). Finally, the conformational change that occurs after irreversible adsorption allows the phage genome to be expelled into the host. Following the expulsion, the phage can follow different replication cycles (Dennehy & Abedon, 2021):

In the lytic cycle, the phage takes control of the cell machinery to produce new viral particles. The phage lysins cause the cell death quickly and release the new viral particles, which then can infect other cells (Dennehy & Abedon, 2021).

In the lysogenic cycle, the phage integrates its genome into the host DNA, forming a prophage that can stay in a dormant state until a signal activates it. The prophages are also inherited by the descendant cells (Dennehy & Abedon, 2021).

Upon successful adsorption, phages must face bacterial intracellular immunity, which consists of highly variable defence systems (Hussain et al., 2021). Thus, phages infect more hosts than they kill. Until recently, only a few intracellular antiphage mechanisms were known, such as systems that detect and degrade invading nucleic acids (Restriction-Modification, CRISPR-Cas, BREX) and abortive infection ones. However, advances in high-throughput bioinformatics have expanded our knowledge of bacterial antiphage tools, revealing over a hundred diverse defence systems and mechanisms, often clustered in so-called *defence islands*. Those recent discoveries have shown unexpectedly complex bacterial immune strategies, which include small antiphage molecules production, intracellular signal transduction through the production of signalling molecules, the recognition of conserved structural viral protein patterns to trigger immune responses (Georjon & Bernheim, 2023). The presence and composition of these defense systems vary among individual strains.

Through mutations and adaptations, phages can overcome bacterial defence barriers and continue their infection cycle. Once inside, prophages integrate into the mobile part of the bacterial genome, potentially providing the host with new functions. Some phages can prevent the host cell from forming pili, structures that phages normally use to attach and infect the cell. For example, the temperate *Pseudomonas aeruginosa* phage D3112 encodes a protein called Tip, which binds to PilB, blocking the formation of type IV pili. By preventing pili formation, phage D3112 stops other phages, which rely on pili, from

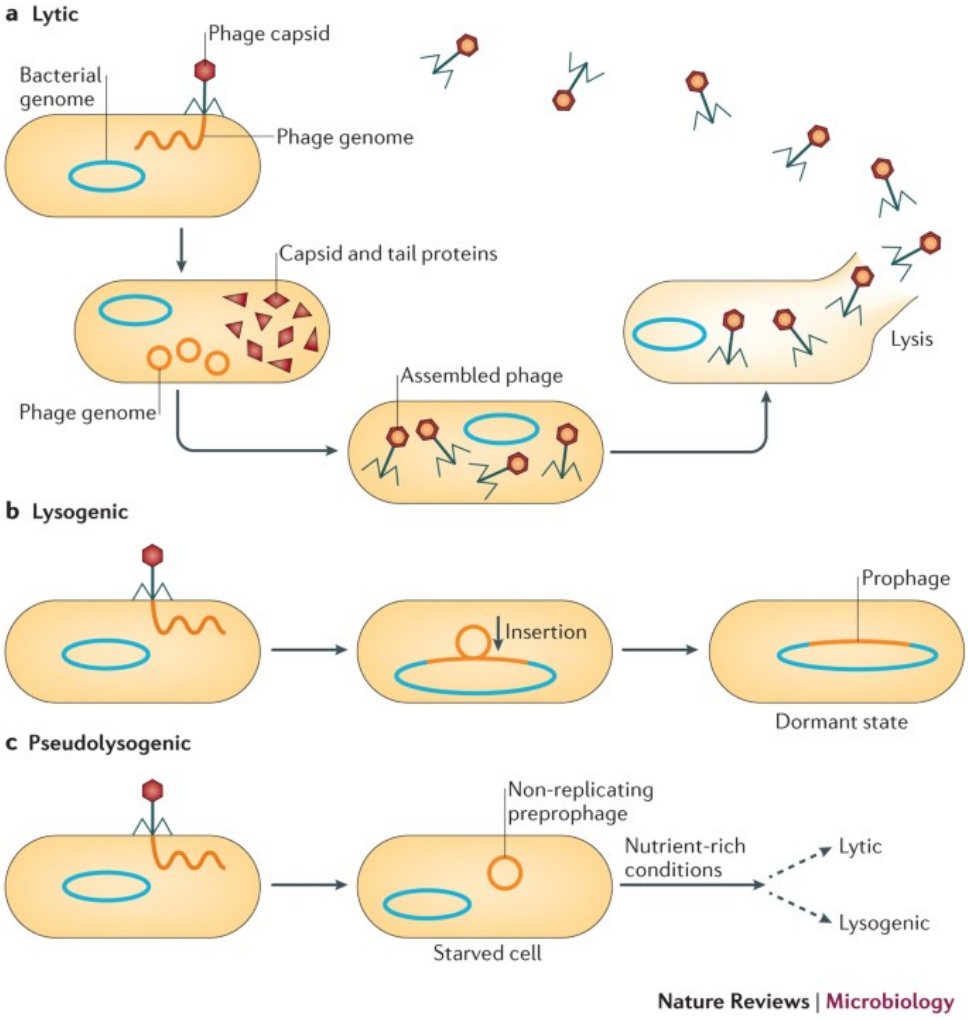


Figure 1: The phage replication cycles (Feiner et al., 2015)

infecting the host cell (Chung et al., 2014).

Because *Pseudomonas aeruginosa* is resistant to antibiotics, bacteriophages have been studied as a possible treatment. These viruses can attack and destroy bacteria without harming the human microbiota (Paz-Zarza et al., 2019). To ensure an efficient phage therapy, phages must be lytic. If a phage enters the lysogenic cycle and inserts its DNA sequence into the bacterial genome, it can confer protection upon the infected bacterium against other phages. Assuming that we aim to kill the bacterium, this stage is not adapted for our goal. (Petrovic Fabijan et al., 2023).

Understanding the role of prophages in bacterial defense against phages is important for several reasons:

1. If phage defence systems are granted by prophages, they can be horizontally transferred through lysogeny, thus accelerating the evolution of bacterial defences.

2. Some prophages provide immunity against other phages, helping lysogenic bacteria survival against viral infections. Therefore, understanding this phenomenon sheds light on how bacteria balance the advantages of lysogeny with the risk of phage attacks.

3. Due to antimicrobial limitations, phages are being explored as alternatives to antibiotics. If prophages play an important role in bacterial resistance to phages, this could affect the effectiveness of these therapies.

4. Although phage therapy is more precise than conventional antibiotics, this does not mean that it does not affect the bacterial population.

5. Identifying whether defense systems are mainly located in prophages or in the bacterial chromosome provides information on how these mechanisms evolve, spread, and interact with other protective systems such as CRISPR-Cas or phage inhibitors.

In summary, studying the role of prophages in defense against phages also helps us better understand bacterial evolution, microbial ecology, and potential biomedical and biotechnological applications.

1.2 Objectives

The objective of this project is to analyze the colocalization of prophages and defense systems in bacterial genomes to better understand the defence mechanisms that bacteria develop against bacteriophages. Given that some bacteria can become immune to a series of phages through the acquisition of specific defence systems, it is fundamental to identify these processes to prevent them from interfering with the effectiveness of phage therapy.

In particular, we seek to identify which phages can confer immunity to *Pseudomonas aeruginosa* against other phages, which will allow us to exclude them from therapeutic use, so we want to select phages capable of entering a lytic cycle, specifically destroying bacterial cells without affecting the human microbiota. This will allow designing more effective, precise, and adapted phage therapy strategies for each type of bacteria.

1.3 Status quo

Currently, antibiotic resistance is seriously increasing and therefore impacting the antibiotic prescriptions (Bottery et al., 2021). This explains why we need to develop alternatives or improve already-existing ones, such as phage therapy. To improve it, it is essential to understand the interaction of mobile elements and bacterial defence systems, because this helps us to identify more effective phages and develop precise strategies.

Recently, tools have been developed to identify the bacterial defense systems and to detect prophages. However, no tool integrates both analyses to allow the verification of the co-localization of defense systems and prophages. We will therefore try to solve this issue in this project.

2 Requirements specification

2.1 Functional requirements

To fulfil this project, we will need to:

1. Create a script that runs defenseFinder to detect anti-phage defence systems in bacterial genomes. The script needs to include the preparation of the input files, the tool run, and the results outputting.
2. Create a script that runs tools like Genomad or PHASTEST to detect prophages in bacterial genomes.
3. Create a script to merge the two results by combining the information from the two previous analyses (defence systems and prophages) into one single file or structured format for joint analysis.
4. Create a script for visualising the results using Plotly or any other Python library to show the colocalization between anti-phage defence systems and prophages in bacterial genomes.

2.2 Non-functional requirements

The project aims to provide the project owner with a usable and easy-to-install tool. Fully explained documentation is required to ensure the tests' reproducibility. It should specify the dependencies versions used in our scripts to guarantee tool compatibility.

We will also make the scripts runnable through a single Unix-based terminal command. This command will need to enable an input of a directory or a single file.

Besides, throughout the different installations required for the project, many dependencies will be associated with the tools. Creating a `conda` environment will allow us to capture every version of the libraries on which the project will rely. This step aims to increase tool compatibility.

3 Constraints

3.1 Functional constraints

The scripts will need to adapt during different stages of development. First, they will be tested on a whole reference genome, which will need to accept contig-based inputs, as this is the type of data produced by the project owner's laboratory. Besides, since many different tools exist for phage detection, two will be tested: PHASTEST and GeNomad, as suggested by the project owner. The latter will guide us in choosing one tool over the other.

The installation process, which will be described in the documentation, will rely on the main methods for tool installation: `git clone` command or `pip`-installable Python package.

3.2 Technical constraints

Several constraints will also need to be taken into consideration by the team members.

For instance, GeNomad only accepts nucleotide sequences in input fasta files, whereas DefenseFinder accepts both protein and nucleotide ones. It implies the need to add verification processes within the script to guarantee the performance of our tools. In addition, a common tool for gene prediction can and must be used, as the geNomad prediction may differ from the one held by DefenseFinder. Given that GeNomad and DefenseFinder rely on `Prodigal` to annotate the given sequence genes, the same tool will be implemented in our scripts to make sure that PHASTEST is also set with an identical proteome. Thus, output processing will be easier to perform as all software is expected to describe the same amount of proteins.

Furthermore, while GeNomad and DefenseFinder are locally installed and return `.tsv` files, PHASTEST can be called through an API and returns a URL for a `.zip` file. Thus, some additional processing will be required if this tool is used.

Besides, as it is recommended by both DefenseFinder and GeNomad documentation, those tools will be installed through a `conda` environment. Our tool will therefore need to be compatible with such an environment.

3.3 Time constraints

Depending on the size of the bacteria's genome, the performance may differ from one run to the next by affecting the tools' efficiency. In case of RAM-related errors, or if the processing time is too long, we are likely to use the Master's degree pedagogical resources (`Pedago NGS` server).

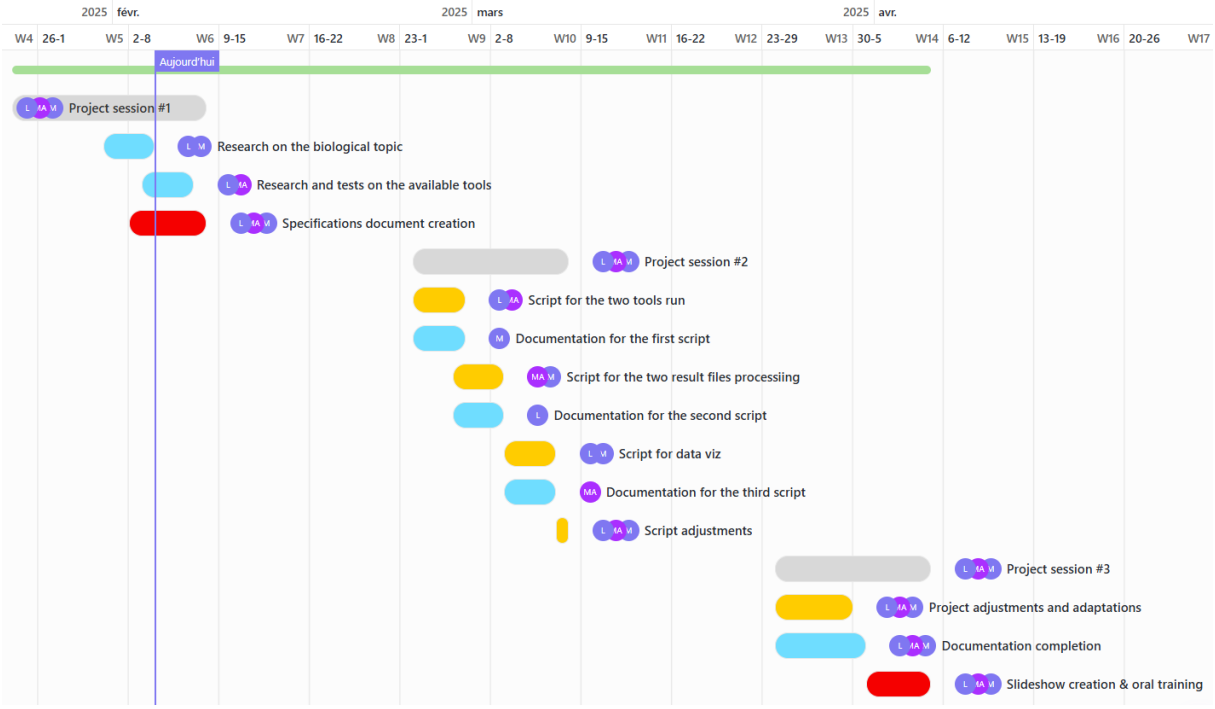
4 Project phases

The project is composed of three stages, as follows:

1. **Project conception and specifications writing**, from Monday, 27 January, to Friday, 7 February
2. **Tool development and debugging**, from Monday, 24 February, to Friday, 7 March
3. **Project finalisation and *viva voce* preparation**, from Monday, 24 March, to Friday, 4 April

4.1 Project timeline

Throughout the different phases, the following timeline will be respected:



5 Diverse specifications

5.1 Documentation

The project will contain a `readme.md` file, which will also serve as documentation for the newly created tool. Divided into several parts, it will include the following information:

- Instructions on how to install the tool
- Required dependencies that need to be installed previously or that will be installed automatically
- Quick run instructions to launch the tool
- Specifications regarding the type of data that needs to be imported
- Description of the result files

This documentation will be added at the repository root, in the `main` branch.

5.2 Methods and tools used

Several tools may be used throughout the project steps:

5.2.1 DefenseFinder

This tool relies on Python and returns the defence systems included within the input sequences for a given fasta file.

The search is done by comparing the genome with Hidden Markow Models from a defence system database, implying regular updates are needed. This Python package is related to the following dependencies:

— macsyfinder

— colorlog

— pyyaml

— packaging

— networkx

Though all those dependencies are installed with the `pip` command.

The results are saved in a `.tsv` file containing, among others, the beginning and ending positions of the defence system. (Couvin et al., 2018; Néron et al., 2023; Tesson et al., 2022).

DefenseFinder is a powerful tool for detecting bacterial antiviral systems, with broad coverage, great accessibility, and availability, as both a command-line version and a web service can be used. Additionally, it can analyse up to 5,000 genomes in less than a day using a standard laptop, making it accessible to many users. It also uses the MacSyFinder architecture, allowing the tool to be regularly updated to match scientific literature advances on antiviral systems. However, it has several limitations, such as the lack of understanding of some antiviral systems that are not yet well understood. It also strongly relies on computational inferences since these systems have not been fully experimentally validated. This can lead to interpretation errors and reduced precision because no experimental reference helps to confirm their accuracy. In conclusion, despite those limitations, DefenseFinder provides a valuable quantitative view of antiviral systems in bacteria and is often described as a reference for defence system search. (Tesson et al., 2022).

5.2.2 Prodigal

This tool will ensure that DefenseFinder, GeNomad, and PHASTEST have the same protein data by translating the nucleotide fasta into protein fasta. It will allow fewer processing and filtering steps afterwards (Hyatt et al., 2010).

5.2.3 GeNomad

GeNomad is a pipeline for mobile element detection that combines two approaches: those based on alignment methods and alignment-free models with neural networks to detect those elements. This overcomes the limitations of previous tools, which were based on only one of the two approaches, meaning that alignment-based tools could not detect viruses that were not present in the database, and neural network-based tools were not precise enough due to false positives.

Therefore, GeNomad allows us greater precision with superior classification across all sequence lengths, particularly in short sequences ($< 6kb$), where other tools stall. It also has improved sensitivity in RNA virus detection and giant viruses (Camargo et al., 2024).

5.2.4 PHASTEST

PHASTEST is the third version of a tool first called PHAST, then PHASTER. It relies on a web service that provides complete genome annotations and enhanced genome visualisation capabilities with a modern and user-friendly interface. Also usable through an API, one main challenge with this tool would be the minimum length and maximum size of the POST request that can only include nucleotide sequences with at least 1,500 base pairs, and less than 15 Mo of data. Thus, the analysis could take longer because of the single sequence request recommendation. Despite still facing false positives and undetected prophages minor concerns, mostly due to the complexity of prophage sequences and the limitations of computational methods, PHASTEST remains highly effective and is often cited as a reference in its field. (Arndt et al., 2016; Wishart et al., 2023; Zhou et al., 2011).

5.3 Reproductibility

To ensure scientific reproducibility and allow the project marking by the project owner and supervisor, a git repository will be created. It will include the task history and results

in the `issues` section, and the development files and resources will be in a separate git branch entitled `project`.

6 Responsibilities

6.1 Project owner

The project owner is Clarisse PLANTADY, PhD student at CNRS-IBCP-MMSB-UMR5086, Molecular Ecology and Evolution of Phages' team (MEEP)/Phaxiam Therapeutics/Paris-Cité University, and her PhD director Anne CHEVALLEREAU, team leader of MEEP at CNRS-IBCP-MMSB/Claude Bernard University.

6.2 Project supervisor

The project supervisor is Charles COLUZZI, lecturer and researcher at Lyon 1-Claude Bernard University, Laboratory of Biometry and Evolutionary Biology.

6.3 Team members

The project will be undertaken by the following students who are enrolled in the Lyon 1 Master's degree in bioinformatics:

— Magloire Auré AGBOMAKOUNZO

— Mariam EL KHATTABI

— Lorcan BRENDERS

References

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., & Wishart, D. S. (2016). PHASTER: A better, faster version of the PHAST phage search tool. *Nucleic Acids Research*, *44*(W1), W16–21. https://doi.org/10.1093/nar/gkw387

Bottery, M. J., Pitchford, J. W., & Friman, V.-P. (2021). Ecology and evolution of antimicrobial resistance in bacterial communities. *The ISME Journal*, *15*(4), 939–948. https://doi.org/10.1038/s41396-020-00832-7

Camargo, A. P., Roux, S., Schulz, F., Babinski, M., Xu, Y., Hu, B., Chain, P. S. G., Nayfach, S., & Kyrpides, N. C. (2024). Identification of mobile genetic elements with geNomad. *Nature Biotechnology*, *42*(8), 1303–1312. https://doi.org/10.1038/s41587-023-01953-y

Chung, I.-Y., Jang, H.-J., Bae, H.-W., & Cho, Y.-H. (2014). A phage protein that inhibits the bacterial ATPase required for type IV pilus assembly [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *111*(31), 11503–11508. https://doi.org/10.1073/pnas.1403537111

Comeau, A. M., Hatfull, G. F., Krisch, H. M., Lindell, D., Mann, N. H., & Prangishvili, D. (2008). Exploring the prokaryotic virosphere. *Research in Microbiology*, *159*(5), 306–313. https://doi.org/10.1016/j.resmic.2008.05.001

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E. P. C., Vergnaud, G., Gautheret, D., & Pourcel, C. (2018). CRISPRCas-Finder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research*, *46*(W1), W246–W251. https://doi.org/10.1093/nar/gky425

De Smet, J., Hendrix, H., Blasdel, B. G., Danis-Wlodarczyk, K., & Lavigne, R. (2017). Pseudomonas predators: Understanding and exploiting phage–host interactions [Publisher: Nature Publishing Group]. *Nature Reviews Microbiology*, *15*(9), 517–530. https://doi.org/10.1038/nrmicro.2017.61

Dennehy, J. J., & Abedon, S. T. (2021). Adsorption: Phage Acquisition of Bacteria. In D. R. Harper, S. T. Abedon, B. H. Burrowes, & M. L. McConville (Eds.), *Bacteriophages* (pp. 93–117). Springer International Publishing. https://doi.org/10.1007/978-3-319-41986-2_2

Feiner, R., Argov, T., Rabinovich, L., Sigal, N., Borovok, I., & Herskovits, A. A. (2015). A new perspective on lysogeny: Prophages as active regulatory switches of bacteria [Publisher: Nature Publishing Group]. *Nature Reviews Microbiology*, *13*(10), 641–650. https://doi.org/10.1038/nrmicro3527

Georjon, H., & Bernheim, A. (2023). The highly diverse antiphage defence systems of bacteria. *Nature Reviews Microbiology*, *21*(10), 686–700. https://doi.org/10.1038/s41579-023-00934-x

Gheorghita, A. A., Wozniak, D. J., Parsek, M. R., & Howell, P. L. (2023). Pseudomonas aeruginosa biofilm exopolysaccharides: Assembly, function, and degradation. *FEMS Microbiology Reviews*, *47*(6), fuad060. https://doi.org/10.1093/femsre/fuad060

Hussain, F. A., Dubert, J., Elsherbini, J., Murphy, M., VanInsberghe, D., Arevalo, P., Kauffman, K., Rodino-Janeiro, B. K., Gavin, H., Gomez, A., Lopatina, A., Le Roux, F., & Polz, M. F. (2021). Rapid evolutionary turnover of mobile genetic

elements drives bacterial resistance to phages. *Science*, *374*(6566), 488–492. https://doi.org/10.1126/science.abb1083

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), 119. https://doi.org/10.1186/1471-2105-11-119

Mortensen, K., Lam, T. J., & Ye, Y. (2021). Comparison of CRISPR–Cas Immune Systems in Healthcare-Related Pathogens. *Frontiers in Microbiology*, *12*, 758782. https://doi.org/10.3389/fmicb.2021.758782

Néron, B., Denise, R., Coluzzi, C., Touchon, M., Rocha, E. P. C., & Abby, S. S. (2023). MacSyFinder v2: Improved modelling and search engine to identify molecular systems in genomes. *Peer Community Journal*, *3*. https://doi.org/10.24072/pcjournal.250

Paz-Zarza, V. M., Mangwani-Mordani, S., Martínez-Maldonado, A., Álvarez-Hernández, D., Solano-Gálvez, S. G., Vázquez-López, R., Paz-Zarza, V. M., Mangwani-Mordani, S., Martínez-Maldonado, A., Álvarez-Hernández, D., Solano-Gálvez, S. G., & Vázquez-López, R. (2019). Pseudomonas aeruginosa: Patogenicidad y resistencia antimicrobiana en la infección urinaria [Publisher: Sociedad Chilena de Infectología]. *Revista chilena de infectología*, *36*(2), 180–189. https://doi.org/10.4067/S0716-10182019000200180

Petrovic Fabijan, A., Iredell, J., Danis-Wlodarczyk, K., Kebriaei, R., & Abedon, S. T. (2023). Translating phage therapy into the clinic: Recent accomplishments but continuing challenges. *PLOS Biology*, *21*(5), e3002119. https://doi.org/10.1371/journal.pbio.3002119

Silby, M. W., Winstanley, C., Godfrey, S. A. C., Levy, S. B., & Jackson, R. W. (2011). Pseudomonas genomes: Diverse and adaptable. *FEMS microbiology reviews*, *35*(4), 652–680. https://doi.org/10.1111/j.1574-6976.2011.00269.x

Suttle, C. A. (2005). Viruses in the sea. *Nature*, *437*(7057), 356–361. https://doi.org/10.1038/nature04160

Tesson, F., Hervé, A., Mordret, E., Touchon, M., d'Humières, C., Cury, J., & Bernheim, A. (2022). Systematic and quantitative view of the antiviral arsenal of prokaryotes [Publisher: Nature Publishing Group]. *Nature Communications*, *13*(1), 2561. https://doi.org/10.1038/s41467-022-30269-9

WHO Bacterial Priority Pathogens List 2024: Bacterial Pathogens of Public Health Importance, to Guide Research, Development, and Strategies to Prevent and Control Antimicrobial Resistance (1st ed). (2024). World Health Organization.
Description based on publisher supplied metadata and other sources.

Wishart, D. S., Han, S., Saha, S., Oler, E., Peters, H., Grant, J. R., Stothard, P., & Gautam, V. (2023). PHASTEST: Faster than PHASTER, better than PHAST. *Nucleic Acids Research*, *51*(W1), W443–W450. https://doi.org/10.1093/nar/gkad382

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., & Wishart, D. S. (2011). PHAST: A fast phage search tool. *Nucleic Acids Research*, *39*(Web Server issue), W347–352. https://doi.org/10.1093/nar/gkr485