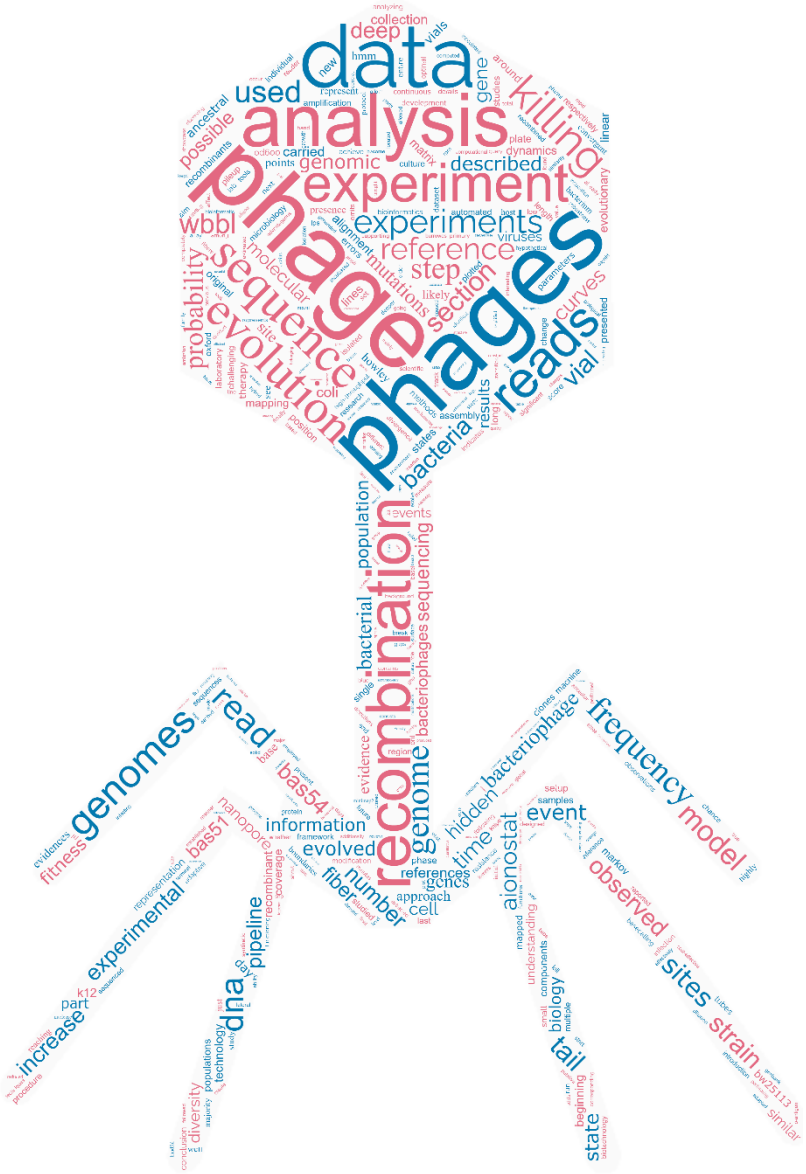
Bachelor's Degree in Genomics
University of Bologna

BIOZENTRUM
Universität Basel
The Center for
Molecular Life Sciences

Quantitative Tracking of Evolutionary and Recombination Dynamics in Bacteriophage Genomes



Supervision:
Prof. Dr. Richard Neher
Prof. Dr. Federico Manuel Giorgi

Defended by:
Giacomo Castagnetti

19 / 07 / 2024

Abstract

The rise of antimicrobial resistance poses a significant challenge to global health, necessitating alternative therapeutic strategies. Bacteriophages, the viruses that infect and kill bacteria, present a promising solution to this problem. However, the current research on bacteriophages is limited, focusing mainly on a handful of deeply studied phages or on broad metagenomics studies, where phages are rarely isolated for in-depth analysis. To thoughtfully develop synthetic phages, or evolve effective phages for therapy, a deeper and broader understanding of bacteriophage evolution is needed. To address this gap, this thesis introduces a framework for high-throughput bacteriophage evolution. This framework is based on the Aionostat, a machine capable of performing automatic, high-throughput, rapid, reproducible, and cost-effective bacteriophage adaptive laboratory evolution experiments. Alongside the machine, a deep sequencing bioinformatic analysis is presented to efficiently and precisely track genomic modification and recombination events that happen in the phage population, yielding significant insights into their adaptive mechanisms. This thesis describes a deep sequencing-based data analysis carried out on the data coming from two evolution experiments. In the first experiment, single phages were evolved in presence of a challenging bacterium, while in the second experiment, multiple phages were evolved together in the same environment. The data analysis of each experiment provides a comprehensive look over the evolutionary and recombination dynamics of the phage populations, revealing high diversity and convergent evolution between independent experiments. The presented framework has the potential not only to advance phage therapy by tailoring phages to combat resistant bacterial strains more effectively, but also to deepen our understanding of the molecular functions and evolutionary processes of bacteriophages.

Contents

Abstract **i**

1 Introduction **1**

2 Scientific background **5**

2.1 Bacteriophages 5

2.2 Aionostat 7

3 Analysis of deep-sequencing data from bacteriophage evolution experiments **11**

3.1 Experimental setup 12

3.2 Data analysis 12

3.3 Results 15

4 Inference of recombination break points in bacteriophage genomes from long read sequencing data **19**

4.1 Experimental setup 20

4.2 Data analysis 20

4.2.1 Pipeline 20

4.2.2 Hidden Markov model 22

4.3 Results 25

5 Materials and methods **31**

5.1 Phages and bacteria 31

5.2 Phage amplification 32

5.3 Phage DNA extraction 32

5.4 DNA sequencing 32

5.5 Plate reader killing curves 33

5.6 Genome assembly 33

5.7 Read mapping 34

6 Conclusion and future directions **35**

Acknowledgements **37**

Chapter 1

Introduction

Increasing antibiotic resistance rates in bacteria cause millions of deaths every year (Murray et al. (2022)). The rise in antimicrobial resistance is driven by the misuse and overuse of antibiotics. In addition, the reduced investments in antibiotic research and development are laying the basis for a future in which resistant bacteria cannot be treated. The antimicrobial resistance crisis has revived the research on alternatives to antibiotics, such as bacteriophages (Strathdee et al. (2023)). Bacteriophages are the viruses that infect and kill bacteria. The therapeutic use of bacteriophages, known as phage therapy, was discovered almost a century ago, but was left on the sidelines after the successful introduction of antibiotics. Now, phage therapy and phage research are experiencing a rebirth (Altamirano et al. (2019)). Beyond their therapeutic potential, phages are interesting because they have a central role in shaping ecosystems (e.g. gut microbiome), they are a source of molecular biology tools (e.g. restriction enzymes) and they can be used as a model for the evolution and spreading of human viruses.

To effectively create phages for therapy, two primary approaches are considered: synthetic phages and adapted natural phages. Synthetic phages can be engineered through genetic engineering tools, but this method requires extensive knowledge of the genetic background and molecular mechanisms of phages. Consequently, phage engineering can only be applied on a small percentage of the existing phages (Pires et al. (2016)). Natural phages are most frequently used in the few phage therapy clinical trials conducted so far. However, they often have issues such as a narrow host range and inefficient bacterial killing (Strathdee et al. (2023)). To increase their efficiency, phages are typically adapted to a specific host using directed laboratory evolution experiments, such as the Appelmans protocol (Burrowes et al. (2019)). Nonetheless, the precise experimental parameters and evolutionary mechanisms to

optimize phages are not yet fully understood, highlighting the need for further study of phage evolution.

Advancing in this field requires a deeper and broader understanding of phage evolution. The current state of phage evolution research is limited to a deep understanding of a handful of individual phages (Knipe and Howley (2013)), which cannot account for the enormous diversity of phages, and to broad environmental metagenomics studies, where phages are rarely isolated and poorly studied (Tisza and Buck (2021)).

To better understand and manipulate phage evolution this thesis presents a framework for automated, high-throughput, rapid, reproducible and cost-effective bacteriophage adaptive laboratory evolution experiments. This framework not only improves the procedure of evolving phages for phage therapy, but also allows to keep track of the genotype changes in a precise and efficient manner. The central component of the framework is the Aionostat (Druelle, Valentin (2024)) an automated machine for adaptive laboratory evolution experiments designed and built by Valentin Druelle. Supporting the machine, there is a data analysis procedure, which is the focus of this thesis, based on a deep sequencing approach that can efficiently and precisely keep track of genomic changes in phage genomes. This machine, used alongside large phage isolation studies such as the BASEL phage collection (Maffei et al. (2021)) or the Nahant phage collection (Kauffman et al. (2022)), can provide new insights into the evolution and diversity of phages and can help characterize new molecular details of phages.

The aim of the thesis is to introduce the Aionostat and showcase its capabilities, while giving details regarding the data analysis carried out on phage evolution data.

In chapter 2 (Scientific background), the basic knowledge of phage biology required to understand the following work is presented and the Aionostat is introduced. In chapter 3 (Bacteriophage linear evolution experiment), it is explained the automated evolution experiment that was carried out on single phages infecting a challenging bacterial strain, going into the details of the data analysis. The aim of this experiment is to quatitatively track all the modification that happen in the genomic landscape of the population and showcase the functioning of the Aionostat. Chapter 4 (Bacteriophage recombination experiment), contains a second experiment, in which multiple phages were evolved in presence of a challenging bacterium. Again, the details about the data analysis and about the employed model are presented. The objective of this experiment is to track all the recombination events happening between phages, in a quantitative and precise manner. In chapter

5 (Materials and methods), the wet lab techniques and the bioinformatic tools used in the analysis are described. Chapter 6 (Conclusion and future directions) gives a summary of the findings and the outlook for the future.

Chapter 2

Scientific background

2.1 Bacteriophages

Bacteriophages are the viruses that infect bacteria. They are the most diverse and numerous biological entities in the world. There are about 10^{31} individual tailed phage virions on the planet. Measurements of the persistence of phages in the environment suggest that the entire population of 10^{31} individual virions turns over every 4 to 5 days, and this leads to the estimate that it takes roughly 10^{24} productive infections per second to maintain the population (Suttle (2007)). Throughout history, phages allowed to make discoveries in the field of molecular biology, virology and epidemiology. There has also been interest in the application of phages and their components to practical problems, and this has continued to the present, including but certainly not limited to a renewed interest in phage therapy (Knipe and Howley (2013)).

Phage research is concentrated on a small number of phages, with the result of having a very deep understanding of a handful of individual phages that do not account for the extreme diversity of phages. Bacteriophages taxonomy has not been standardized yet (Turner et al. (2023)), the commonly used classification is the one described by Dion et al. (2020). This classification is based on the genome type and morphology of phages. The vast majority of bacteriophage described to date are dsDNA bacteriophages. Among these, tailed bacteriophages, belonging to the caudovirales order, are by far the most studied phages (Knipe and Howley (2013)). They are extremely diverse gentically and their genome size goes from 19 kb to 500 kb. This thesis is focused on the Myoviridae family, belonging to the Caudovirales order. These phages have their genome highly packaged in an icosahedral protein capsid to which it is attached a contractile tail that injects the

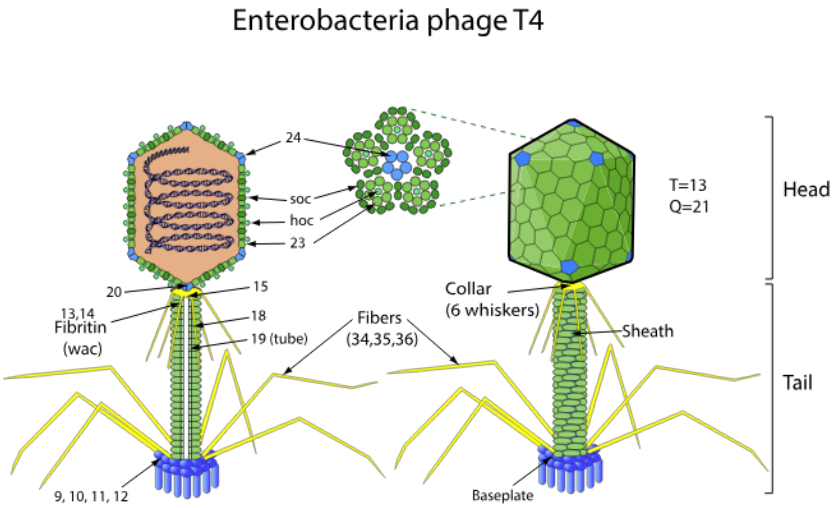


Figure 2.1: Scheme of T4 phage morphology, taken from https://viralzone.expasy.org/504

DNA into the cytoplasm of the host. The most studied phage belonging to this family is T4 phage. A representation of the structure of T4 phage is shown in figure 2.1.

The diversity of phages raises interests in their evolution mechanisms. Two ways have been employed to study phage evolution. The first consists of culturing phages in the laboratory, in a controlled environment, applying a selective pressure, and following them by genome sequencing. This approach allowed rigorous testing of a number of different aspects of evolutionary theory and made it possible to elucidate specific pathways of adaptation. This work has been valuable, but limited to a small number of phages and to a short evolution time (Knipe and Howley (2013)). This thesis presents a new version of this approach, aiming to make it high-throughput, fast, and cost-effective. The second approach is to isolate phages from the environment, sequencing them and comparing their genomes. This approach has the limitation that the sampling of the population is extremely sparse and biased, but it yielded a lot of information about the structures of the genomes produced by phage evolution, allowing to infer the mechanisms by which they got there (Knipe and Howley (2013)). An approach that can reduce sampling bias is metagenomics, which consists of sequencing all genetic material in an environment and then detecting the DNA pieces belonging to phages. The downside of this approach is that single phages are not isolated, therefore, it is impossible to study them in depth.

As mentioned before, tailed phages are the most studied, both at molecular

and evolutionary level. The genomes of tailed phages are genetic mosaics. The sequences of two genomes may be nearly identical over the length of a few genes and then abruptly change so that they match less well or not at all. These sites are evidences of non-homologous recombination in the ancestry of one of the phages. These characteristic of phage genomes lead to the conclusion that recombination is the major actor in phage evolution and diversification. Opportunities for recombination arise when a cell is infected simultaneously by two phages or when the infected cell carries one or more prophages. Phage recombination is thought to occur randomly across the entire genome without site specificity, but the vast majority of recombinants are eliminated by natural selection. The only recombinants that survive are those that do not disrupt the functions of genes. Therefore, it is mostly common to observe recombination boundaries in correspondence of gene or protein domain boundaries (Knipe and Howley (2013)).

2.2 Aionostat

The Aionostat (Druelle, Valentin (2024)) is a continuous culture device that performs adaptive laboratory evolution experiments designed and built by Valentin Druelle. The Aionostat operates similarly to a morbidostat (Toprak et al. (2013)), with some modifications that improve performance, ease of use, reliability, and make it phage compatible. It can be seen as an improvement of similar phage continuous culture devices thanks to its versatility (Holtzman et al. (2020)).

A picture of the Aionostat and its main components is shown in figure 2.2. The Aionostat has two main parts: the first part, which handles the liquids and the cultures, sits inside an incubator for temperature control; the second part acts as the brain and power source for the machine and can be programmed to perform a wide variety of experiments. Overall, the Aionostat was made from commercially available components and custom 3D printed parts (Druelle, Valentin (2024)).

The working principle of the Aionostat is shown in figure 2.3. The Aionostat is made of pairs of vials, each pair has a bacterial vial and a phage vial. In the bacterial vial the bacteria are kept in their exponential growth phase, without phages, in lysogeny broth medium (LB) at 37°C, at a constant optical density of 600 nm (OD600) of 0.5. The bacterial density is regulated by diluting the culture with raw LB and the excess liquid resulting from the dilution is directed to the phage vial where phages will receive new bacteria to infect. The volume in the phage vial is kept constant by discarding any surplus. This enables evolution of the phages over

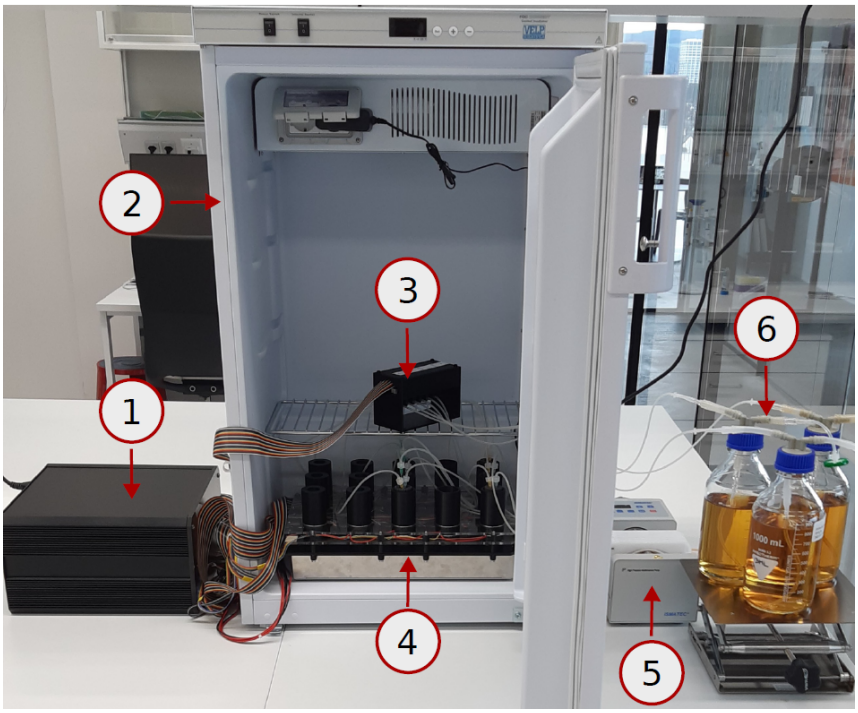


Figure 2.2: Picture of the Aionostat taken from Druelle, Valentin (2024). ① Electronic components' enclosure. Contains the central computer and custom circuits for the electric components. ② Incubator for temperature control. ③ Single channel piezoelectric pump array. ④ Array of experiment vials on stirrer plate. ⑤ Peristaltic exhaust pump. ⑥ Input media bottle. Waste bottle sits on the floor.

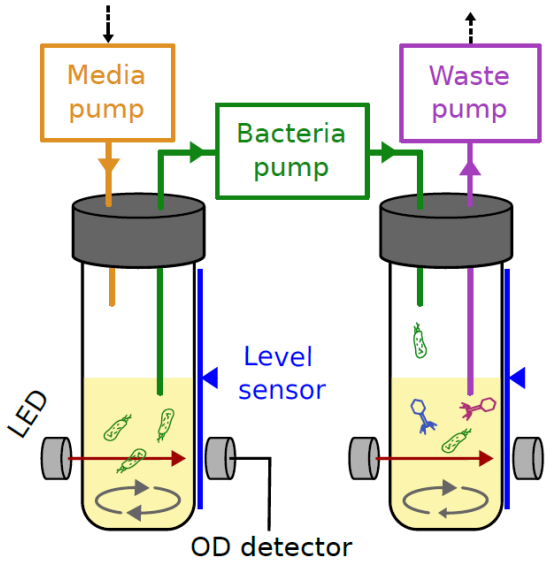


Figure 2.3: Schematic of the working principle of the Aionostat taken from Druelle, Valentin (2024). One vial is used to grow bacteria in exponential phase. These bacteria are then continuously transferred to the second vial and get infected by phages, which can evolve over time.

time. The constant dilution in the phage vial with new bacteria imposes a selective pressure on the phages. The bacteriophages that will by chance collect advantageous mutations will become more efficient and faster in killing the bacteria (fitter) and they will outcompete the other phages. At maximum capacity, the Aionostat can handle 15 vials simultaneously (Druelle, Valentin (2024)).

Chapter 3

Analysis of deep-sequencing data from bacteriophage evolution experiments

In this chapter, I present the design of a computational pipeline to analyze deep sequencing data obtained via nanopore sequencing of sequential samples from an evolution experiment performed by Valentin Druelle. The experimental setup, designed by Valentin Druelle, is presented first, followed by the description of the data analysis that I carried out and the results obtained from the analysis.

In this experiment, a phage strain was evolved in the presence of a challenging bacterial strain that it did not infect well originally. Phages accumulate mutations during the infection, and the mutations that increase fitness the most, spread faster in the population. Therefore, the prevalence of a mutation in the population should be correlated with the increase in fitness caused by this mutation. The aim of this experiment is to quantitatively track every genomic modification (SNP, gap, or insertion) that appears in the phage population.

A deep sequencing approach is employed to monitor rare mutations and variants using solely the sequencing data of the population of evolved phages. Since the modern sequencing technologies allow to have great coverage, decent accuracy and long reads at a reduced cost, this approach results to be very efficient in comparison to sequence individually clones from the population.

The motivation behind this experiment is to showcase the functioning of the Aionostat machine in performing linear evolution of bacteriophages to enhance their fitness. This experiment aims to prove the robustness of the Aionostat and establish the foundation for high-throughput phage evolution experiments. Additionally, the

goal is to characterize the evolved phages both phenotypically and genotypically.

3.1 Experimental setup

The following experiment was designed and carried out by Valentin Druelle. Three vials were set up for this experiment, in each of them, phages from the BASEL collection (see section 5.1) were inoculated at the beginning of the experiment. The first vial contained bas60 phage, the second bas51 and the third bas54. During the experiment, the Aionostat (see section 2.2) took care of constantly filling the phage vials with *E. coli* K12 wbbl(+) (see section 5.1), coming from the bacterium vial.

The experiment lasted 5 days and daily samples were extracted from each phage vial. To prepare these phage samples for storage and later analysis, the bacteriophage samples were cleared of bacteria following the protocol described in section 5.3. The ancestor phages along with the phage population from day 1 3, 5 were amplified (see section 5.2) and sequenced using Oxford Nanopore Technology as described in section 5.4.

Additionally, turbidity-based killing curves of both *E. coli* K12 and *E. coli* K12 wbbl(+) were evaluated with the ancestral phages and the evolved phage populations following the experimental procedure described in section 5.5. This allowed to have a phenotypic (fitness) evaluation of the differences accumulated in the phage genotype.

3.2 Data analysis

The ONT sequencing data is handled through bioinformatics tools and python scripts. A conceptual representation of the analysis is shown in figure 3.1. The objective of this analysis is to identify the sites that accumulate mutations throughout the experiment and, for each site, quantitatively measure the portion of reads that carry the mutation. To achieve this, the reads of every time step are aligned to a reference sequence, and for each site the information given by the pool of reads is compared with the information of the reference. If, at any time point, the reads differ significantly from the reference because of a mutation, a gap or an insertion, the site is tracked throughout the whole experiment.

The first step of the analysis consists of building the initial consensus sequence of the phage genome. The reads of the ancestor phage are assembled in a genome through flye, as described in section 5.6, obtaining an average of 15 errors per assembly with respect to the sequences stored in GenBank. These errors are mainly

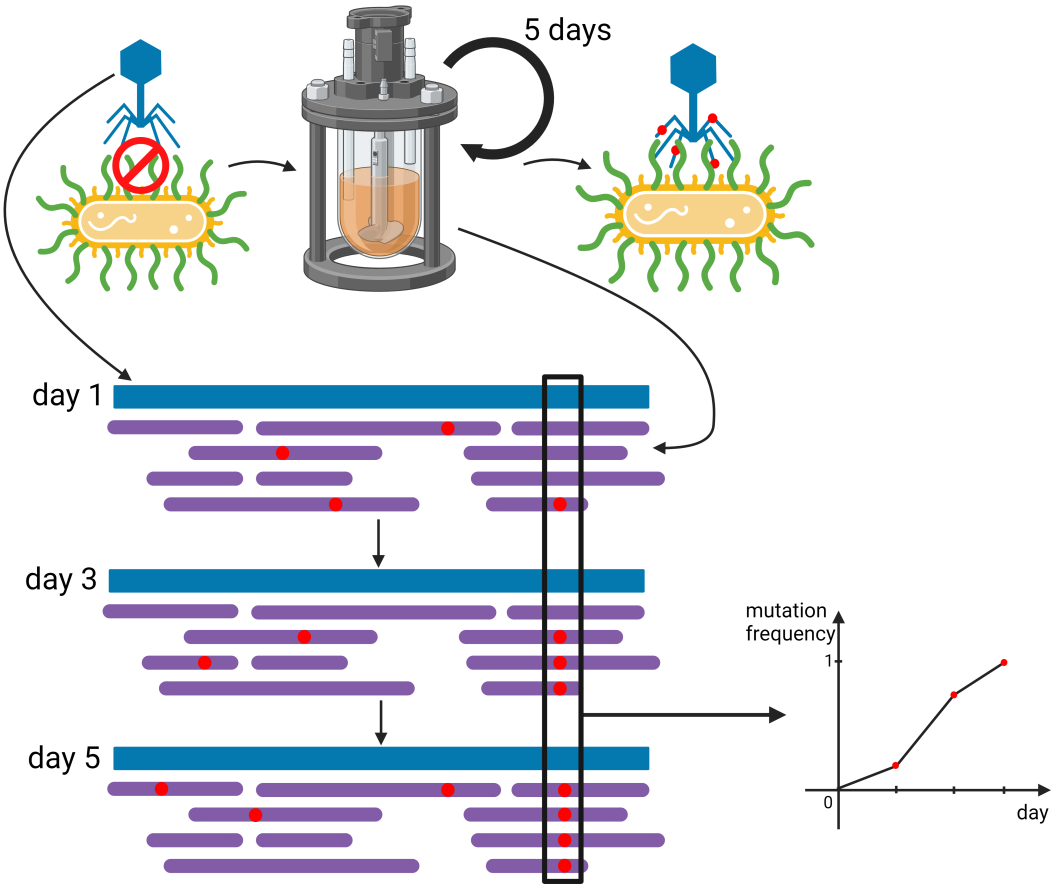


Figure 3.1: Schematic representation of linear evolution experiment setup and analysis. A phage strain was evolved in presence of a challenging bacterial strain. Throughout the experiment the genome of the phages was sequenced and compared with the initial information. During the infection the phages accumulated mutations and the fitter ones spread in the population. The sequencing data was analysed to find the sites that showed significant differences from the initial state, and these sites were tracked throughout the experiment.

SNPs located in correspondence of modified bases that ONT could not read properly. The reason of these modifications are methylases. The Dam methylase is an enzyme of *E. coli* which recognises "GATC" sequence and methylates (adds a methyl group) the adenine (this modification aids the repair system of the bacterial cell). Dam methylase can also modify the phage genome while it is infecting *E. coli*. The presence of the methyl group frequently causes the Nanopore to misread the base preceding the adenine (Delahaye and Nicolas (2021)).

Secondly, the reads of each of the successive time steps are mapped on the assembly using minimap2, as described in section 5.7. All the data of each alignment is summarized in pileups. A pileup is a data structure that stores and summarises the information of a bam file. For each position of the reference sequence a pileup contains: the coverage for each of the 4 bases, the number of gaps and the number of insertions. This data is stored separately for the reads mapped forward and reverse.

The next step is to extract from this data the sites that show significant differences compared to the reference. To do this it is established a divergence score for each measure of the pileup. The non consensus frequency score is the number of bases that differ from the reference sequence, at a specific site, over the total number of bases mapped on that position. The gap divergence score is the number of gaps mapping over a base, divided by the total number of reads that span that base. The insertion divergence score is the number of insertions mapping on a position divided by the number of reads mapping in that position.

The primary issue with this dataset is its significant noise levels. Nanopore sequencing is a technology that is prone to errors and biases and it is not possible to trust every piece of information that it provides as output. Given the sufficient coverage, it is possible to filter out the unreliable information to distinguish signal from noise. To achieve this, meaningful thresholds are set on read quality, site coverage, and forward-reverse concordance. Only nucleotides with a quality score greater than 20 are considered in the analysis. ONT uses the Phred quality score, a measure of confidence based on the estimated error rate, calculated as $-10 * log(Pe)$ where Pe is the probability of error. A Phred score of 20 corresponds to 0.01 error probability. The second threshold is set on the coverage, only the sites with more than 50x coverage are retained. Finally, sites where the information from forward-mapped reads significantly differs (more than 40%) from reverse-mapped reads are ignored. This control is needed because nanopores can read DNA fragments in both directions, generating reads that have the same information as the reference sequence and reads that are the reverse complementary of the reference sequence.

Sometimes, due to some characteristics of the DNA sequence, there can be problems in reading one of the two direction, resulting in basecalling errors. If the information of the forward reads is highly different from the information of the reverse reads it is impossible to know which of the two direction contains the true information and the site is discarded.

After extracting the divergence score for every site of the genome for each pileup measure at each time step, the data for a single pileup measure from all the time steps can be processed simultaneously. The expected outcome is that the vast majority of sites will have a low divergence score, with only a few sites showing a significant divergence from the consensus (single mutation, gap or insertion). As time progresses, more sites are expected to diverge. The final step involves identifying the most interesting sites, which are those showing the highest increase in divergence from the beginning to the end of the experiment. In the next section the significant sites that were found are presented.

3.3 Results

In this section, the genomic changes observed in the evolved phages will be presented first, followed by an analysis of the derived phenotype. Figures 3.2 and 3.3, show the frequency trajectories of single nucleotide polymorphisms (SNPs) and gaps throughout the experiment.

In the first vial, containing phage bas60, no significant changes were observed in the genomic composition of the phage. Therefore, the analysis will focus solely on bas51 (second vial) and bas54 (third vial).

In the second vial (figure 3.2), bas51 acquired two SNPs, at positions 36588 and 30397 (tail fiber genes), which took over in the population, reaching a frequency of 60%. The SNPs in position 31963 and 37242, which affect tail fiber genes, have a sudden increase at the beginning of the experiment, but they decrease already at day 3. While a SNP in position 7328, on a major capsid protein, shows a constant but slow increase in the population. Examining the gaps, a similar scenario is observed: gaps at positions 43327 and 28472, impacting a lateral tail fiber gene and a hypothetical protein, respectively, are highly prevalent in the population, whereas other gaps appear just at the beginning of the experiment.

In the third vial (figure 3.3), bas54 presented only one SNP that reached a high frequency (around 80%) located in a lateral tail fiber gene (36619) and two SNPs (7695 and 7696) that slowly increased during the experiment involving a major

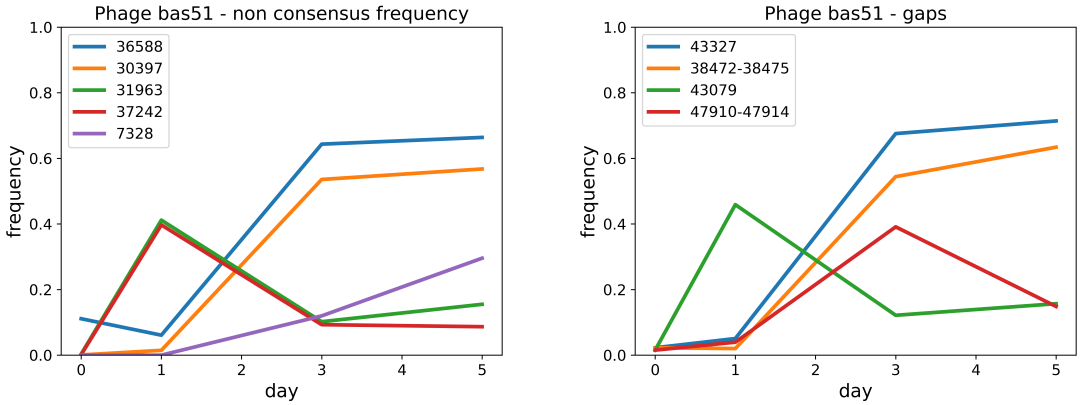


Figure 3.2: Results of the linear evolution experiment of phage Bas51. The frequency of each modification (i.e., its prevalence in the population) is plotted over time. The majority of the reported modification affect tail fiber genes.

capsid protein gene. In bas54 no gap reaches a high frequency but all of them are located on tail fiber genes.

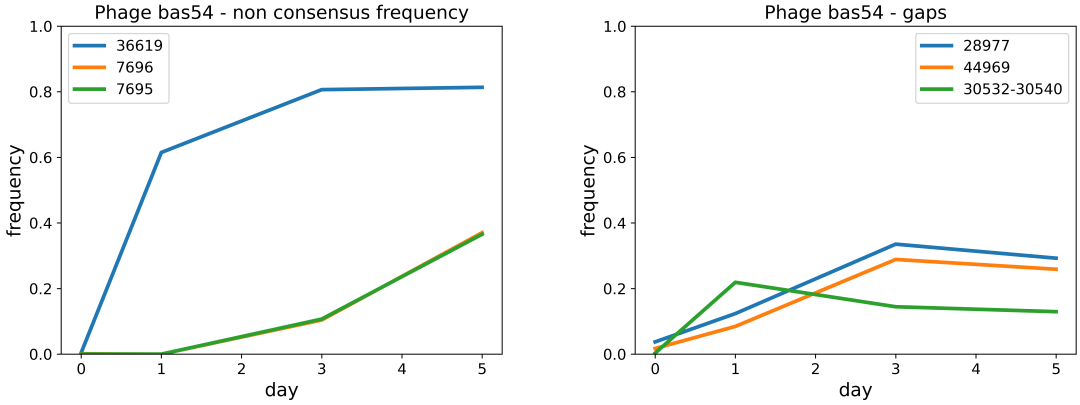


Figure 3.3: Results of the linear evolution experiment of phage Bas54. The frequency of each modification (i.e., its prevalence in the population) is plotted over time. The majority of the reported modification affect tail fiber genes.

Additionally, in both phages, a subpopulation characterized by a large deletion in their genomes establishes at a low frequency. For bas51, the deletion spans from position 43300 to 51800, and for bas54, it extends from position 46680 to 50600. These deletions mainly involve a lateral tail fiber gene. No other major rearrangements have been observed in the phage genomes.

The pattern that emerges from this results is that tail fiber genes are by far the genes that are mostly affected by modifications. This comes from their crucial role in bacterial infection. Additionally, advantageous mutations take time to appear

and become established in the population. During this process other mutations, presumably less beneficial, rise and then decrease in frequency. Interestingly, newly acquired mutations do not appear to completely fix in the population, maintaining some level of diversity. Most importantly, convergent evolution is observed among phages independently evolved in separate vials. This underscores the robustness of the Aionostat, demonstrating its ability to consistently produce reliable results.

To test whether the adaptive laboratory experiment actually resulted in an increase in fitness in the phages, Valentin Druelle performed killing assays with evolved and ancestor phages on wild-type (WT) and wbbl(+) bacteria. Figure 3.4 displays the resulting killing curves. By looking at the killing assay carried out on the original bacterial *E. coli* strain (dashed lines), it is possible to observe how both the ancestor phage (blue dashed line) and the evolved phage (orange dashed line) cause a decline in the OD600 values after around 3 hours of incubation going down to the detection limit. This suggests that both ancestor and evolved phages kill the original bacteria at comparable rates.

By observing the killing curves of the wbbl(+) strain (solid lines), the situation appears to be really different; the ancestor phage (blue solid line) after around 3 hours of incubation starts to cause a slight decrease in the bacterial density, but soon after, the bacteria start to grow again and reach a plateau. On the other hand, the evolved phage (orange solid line) is able to completely clear the bacterial population. This indicates that the evolved versions of both bas51 and bas54 are more efficient in killing wbbl(+) bacteria than their predecessors.

Linking the observations on the killing curves with the genotypic analysis, it can be inferred that the genomic changes played a pivotal role in the phages' enhanced capability to kill the wbbl(+) strain. It is hypothesized that the modifications on the tail fiber proteins likely impact the binding efficiency of the phages on the bacteria. This is supported by the fact that the bacterial strain used in the experiment is genetically identical to the isolation strain of the phages, with the exception of the restored O-antigen (wbbl+) that adds long chains on the lipopolysaccharide (LPS) of the bacterial wall. The phages use tail fibers to bind the bacteria LPS and initiate infection; therefore, a change in the LPS produces a coevolutionary change in the phage tail fibers. Deeper molecular analysis of the detected mutations could provide extra knowledge on the function of every protein.

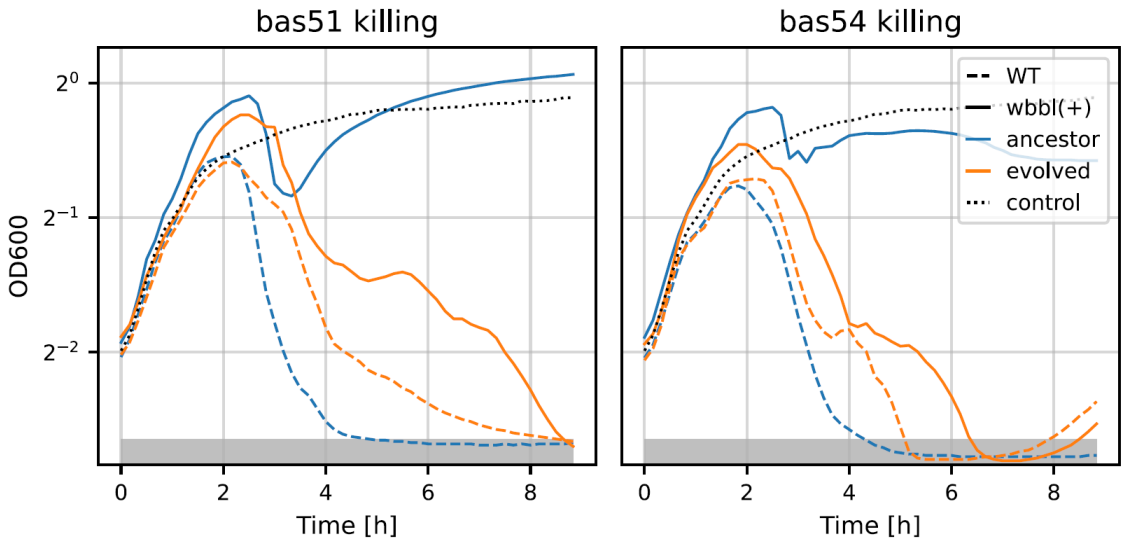


Figure 3.4: Killing curves of linear evolution experiment taken from Druelle, Valentin (2024). The killing curves of evolved bas51 and bas54 are represented on the left and right respectively. The dashed blue line represents the ancestral phage killing the WT bacteria (*E. coli* K12 BW25113) and the solid blue line represents the ancestral phage not killing the wbbl(+) (challenging) bacterium. The dashed and solid orange lines represent respectively the evolved phage killing the WT and wbbl(+) bacteria. The evolution experiment caused an increase in the fitness of the phages.

Chapter 4

Inference of recombination break points in bacteriophage genomes from long read sequencing data

In this chapter, I present the design of a computational pipeline to analyze deep sequencing data obtained via nanopore sequencing of sequential samples from a recombination-driven evolution experiment performed by Valentin Druelle. The experimental setup, designed by Valentin Druelle, is presented first, followed by the description of the data analysis that I carried out and the results obtained from the analysis.

It has been widely demonstrated that tailed phages evolve largely by horizontal gene transfer (Hendrix (2008) and Kauffman et al. (2022)). Recombination events can occur between lytic phages when they co-infect the same bacterial cell simultaneously (Kunisaki and Tanji (2009)). In this experiment, a population of diverse phages is infecting a challenging bacteria strain.

The objective of this experiment is to have a complete view of the horizontal gene transfer dynamics during phage infection. The aim is to develop a bioinformatics pipeline to quantitatively track the abundance of each phage genome in the population and precisely localize the position and size of the recombination events. To achieve this, a deep sequencing approach is employed, which uses only sequencing data of the entire phage population. Additionally, the experiment seeks to determine the impact of these rearrangements on the phage phenotype by comparing the killing potential of the chimeric phages with that of the ancestral ones.

The motivation behind this experiment is to lay the foundations for high-throughput experimental horizontal evolution of phages. The Aionostat machine aims to make

adaptive laboratory evolution experiments of phages fast and effortless. Since phage diversity and evolution is highly dependent on horizontal gene transfer, leveraging this mechanism for evolving phages in the lab has many possible applications in fields like phage therapy. Scientists developed experimental methods that exploit recombination, such as the Appelsman protocol (Burrowes et al. (2019)), that could be automated thanks to the Aionostat.

4.1 Experimental setup

The following experiment was designed and carried out by Valentin Druelle. In this experiment, 3 vials were set up, two of them having three phage strains each (bas51, bas54, and bas60) and one without phages, as a negative control. The phages were inoculated only one time, at the beginning of the experiment. The rest of the Aionostat setup remained unchanged from the first experiment.

During the experiment, samples were taken from each vial on days 1,3,5 and 7, the phage population was isolated and sequenced with Oxford Nanopore Technology. Furthermore, at the end of the experiment, the phage populations were plated, and four individual clones were selected, amplified, and sequenced.

4.2 Data analysis

An overview of the pipeline developed to process the data produced by the recombination experiment is presented in figure 4.1. In the subsequent sections, explanations of the pipeline components and their functioning will be provided. The pipeline was implemented in Snakemake (Molder et al. (2021)) and run on sciCORE (http://scicore.unibas.ch/), the scientific computing center at University of Basel. The code can be found at https://github.com/kcajj/recombinant_population_analysis.

4.2.1 Pipeline

The pipeline takes 3 inputs: the sequencing reads of the recombinant populations, the reference sequences of the original phages used in the experiment and the parameters for the Hidden Markov model (see below).

The mechanism of the pipeline consists in aligning the reads to the references and then, for each read, predicting whether it belongs to one or the other reference based on its similarities. To quickly produce a precise alignment of each read with

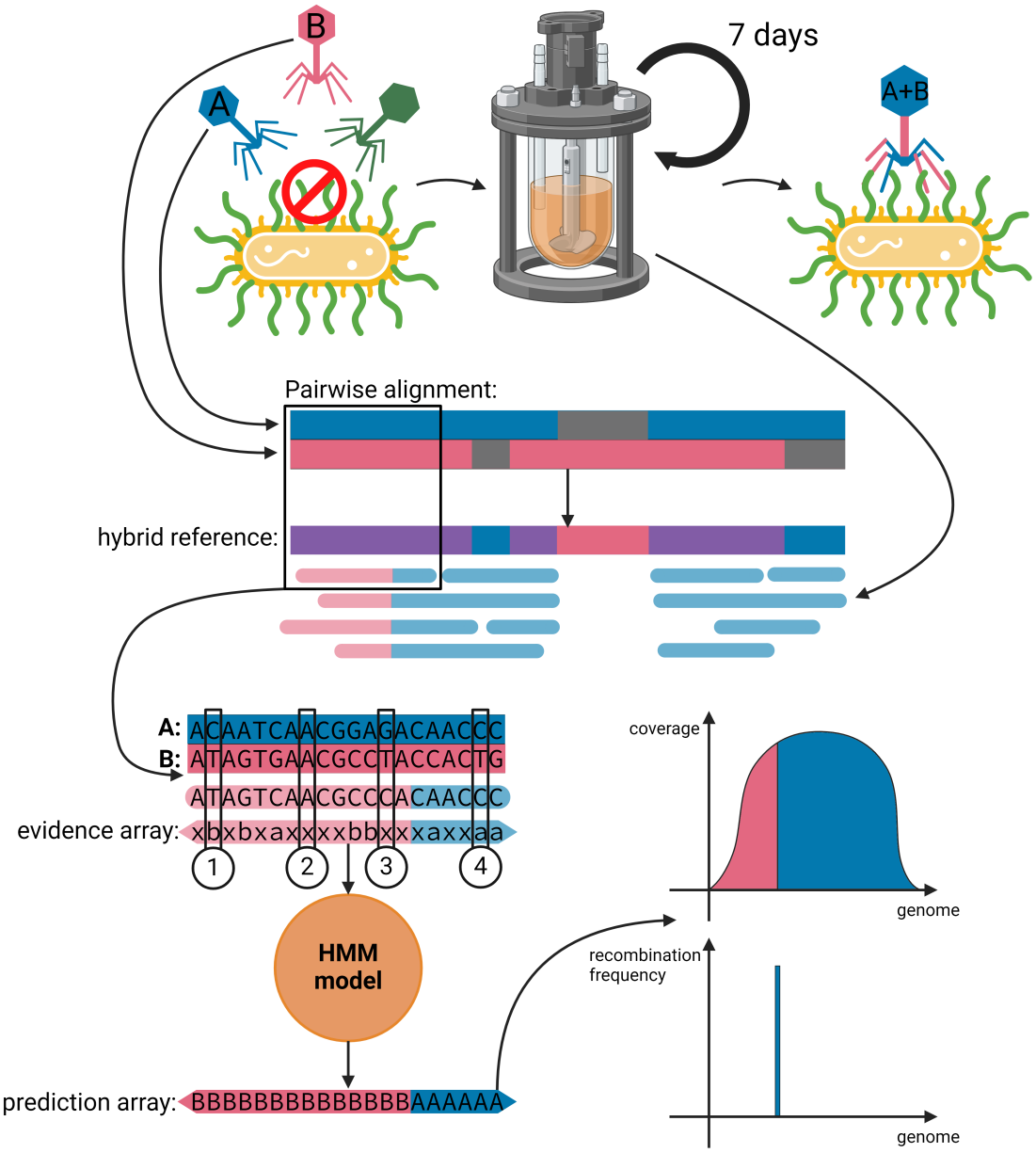


Figure 4.1: Schematic representation of the recombination experiment's setup and analysis. Three phages were evolved in presence of a challenging bacteria. During co-infection, two phages recombined and one got extinct. To detect recombination events, a hybrid reference is constructed from pairwise alignments of two references. Recombinant reads are mapped onto the hybrid reference to create a multiple sequence alignment with the two references. At each position of each read, evidence of similarity to phage A (④), B (♠) or no evidence (② and ③) is collected. The evidences are fed into an HMM that infers the most likely recombination border. The coverage of each phage and the positions of each recombination event are plotted.

both references a hybrid reference approach is used. The hybrid reference is created starting from the pairwise alignment of the two references. This sequence is constructed by taking randomly the information from one sequence or the other in the regions where they match and taking the information from the only sequence possible in the case of an indel. The hybrid sequence can serve as a reference for mapping the recombinant reads. In this way, a heuristic read aligner, such as minimap2, can quickly and precisely map reads coming from either reference to a single sequence. The resulting alignment can be approximated as the MSA of the read and the two references. From the MSA of the three sequences, evidence supporting the origin of the read from one reference over the other can be gathered. In this experiment, the references are highly similar, differing in only 3% of sites, which are the informative ones. The query information at these sites gives information about whether the region is likely to be inherited from one reference or the other. Therefore, each base of the read presents three possible scenarios:

1. The base gives evidence of the read being similar to reference A and not to reference B (see ④ in figure 4.1).

2. The base gives evidence of the read being similar to reference B and not to reference A (see ♠ in figure 4.1).

3. The base does not give any information on whether the read is similar to reference A or B: this happens if the two references have the same information (see ② in figure 4.1) or if the base of the read is different from both references (see ③ in figure 4.1).

The sequence of evidences extracted from a read can be interpreted by a Hidden Markov model (see below) that can predict whether the whole read or just part of it belongs to a reference or the other. The model infers the most likely position of the reference switch. From the analysis of each read, the array of its hidden states is obtained. This information can be summarized to represent the coverage of each phage and the hotspots where the recombination events (switches between hidden states) were detected (see figures 4.5 and 4.4). The data coming from each time step can be analysed to track the recombination dynamics throughout the experiment.

4.2.2 Hidden Markov model

At the core of the pipeline there is a Hidden Markov model (HMM) that can infer whether a read or just part of is inherited from a reference sequence or the other.

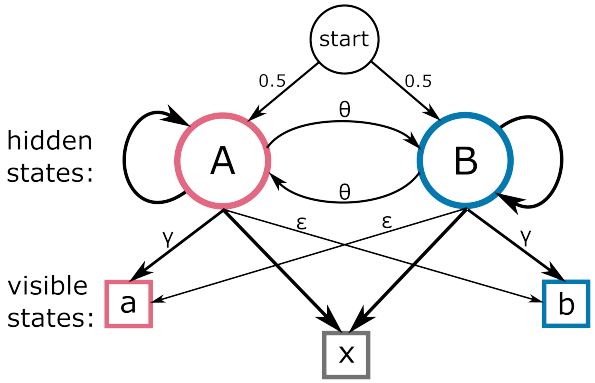


Figure 4.2: Hidden Markov Model (HMM) for the inference of recombination. The two hidden states (A and B) correspond to the two reference phage genomes from which each read has been generated. The visible states (a, b and x) correspond to the evidences that are observed when the read is aligned to the two references.

To do this, the HMM takes in input the sequence of evidences of similarity of the read with respect to the two references and returns the most likely path of the read between the two sequences. Figure 4.2 shows a schematic representation of the model. The HMM has two hidden states, corresponding to the two references (A and B), and three visible states, corresponding to the evidences of similarity described in the previous section (a is evidence for reference A, b is evidence for reference B, and x is no evidence). The probability of jumping from one hidden state to the next depends on a transition matrix (table 4.1), where θ is the transition probability, which is a hyperparameter (optimised below) related to the recombination frequency of the dataset. The probability of emitting one of the three visible states depends on the value of the hidden state. In the emission matrix (table 4.2), γ is the probability for a hidden state, to emit a correct evidence (A emits a or B emits b), while ϵ is the (small) probability of emitting a misleading evidence corresponding to the other hidden state (A emits b or B emits a). Since the two references are identical at most sites, there will often be no discernible evidence (x). Lack of evidence will also occur when ONT makes an error. The probability of having the correct evidence (γ) represents the fraction of genome that is different between the two references minus the error rate of the sequencing technology. The probability of emitting a misleading evidence (ϵ) corresponds to the random sequencing errors that end up suggesting the opposite base. The model starts with the same probability of being in state A and B.

The model runs over the sequence of evidences using the Viterbi algorithm, a dynamic programming approach that finds the optimal sequence of hidden states

	A	B
A	1-θ	θ
B	θ	1-θ

Table 4.1: Transition matrix of the HMM. θ is the transition probability.

	a	b	x
A	γ	ϵ	1-γ-ϵ
B	ϵ	γ	1-γ-ϵ

Table 4.2: Emission matrix of the HMM. γ is the probability of observing a true evidence and ϵ is the probability of observing a misleading evidence.

starting from a sequence of emissions. Starting from the first element, the algorithm fills a dynamic programming matrix with two rows (hidden states) and n columns (length of the genome). For each cell, it computes the likelihood of reaching it from the two previous hidden states, keeping only the optimal value. The likelihood is computed by multiplying the probability computed up to the previous cell (0.5 at the beginning) with the appropriate probability from the transition matrix (1-θ if the hidden state is the same and θ if the hidden state switches) and the respective emission probability. When every cell of the matrix is filled, backtracking can start. From the final cell with highest likelihood, the sequence of optimal steps is reconstructed going backward, taking each time the optimal step that was computed in the initial phase.

The emission parameters of the exeriment were estimated using a dataset without recombination, where the hidden state was known. The initial phages (bas51 and bas54) were sequenced alone, and the sequencing data was mapped to both references to measure the frequency of the visible states. The conditional emission probabilities for hidden state A were approximated based on the empirical frequency of a,b, and x states observed in reads from the sequencing run of reference A. These parameters amount to 0.967 for x, $\gamma = 0.03$ for the true evidence and $\epsilon = 0.003$ for the misleading evidence. The transition parameter θ was estimated on a subset of the longest reads coming from the sequencing run of the recombinant population. The model was tested with various values θ, and the final log-likelihood of each read prediction was summed. A peak in the total likelihood was observed at approximately $\theta = 10^{-5}$. This parameter correlates with the recombination frequency in the dataset, suggesting an expected recombination event roughly once every 10^5 bases.

By testing the model on artificially generated sequences, an accuracy of 0.9996

was achieved in correctly assigning hidden states.

4.3 Results

The results will be presented as follows: first, the genome of the isolated phage clones will be shown; next, additional insights will be derived from analyzing the population data; and finally, the phenotypic effects of all genomic changes will be investigated by analyzing the killing curves of the recombinant phages.

The isolated clones from vial 1 were mainly composed of bas51 genome with some portions of bas54 genome that were exchanged, as shown in figure 4.3. All clones had a big recombination event (labeled recombination event 1) that involved 28 kb of bas54 genome that were inserted into bas51 genome. The recombined region starts from the repeated terminal region, the physical breakpoint of the genome, and finishes 28 kb later. The recombined region contains mainly putative genes, so it is not easy to hypothesize its function but it must have produced enough fitness increase to outcompete the two ancestral phages. The second recombination event (labeled as recombination event 2) was present in only 3 clones out of 4 and involved a single gene of a putative homing nuclease that probably jumped from bas54 to bas51 during coinfection. Homing endonucleases are enzymes that can catalyze the transfer of their onwn genes and flanking sequences by cleaving the recipient DNA in a site-specific manner (barzel 2011). It's therefore plausible that a bas54 endonuclease recognised the target site in bas51 and catalysed the insertion of its sequence. All vial 1 clones also presented a large deletion close to a tail fiber gene, similar to the one observed at the end of the linear evolution experiment. The clones isolated from vial 2 were just bas51 phages with some point mutations and 2 out of 4 reported a large deletion close to a tail fiber gene.

By analyzing the sequencing data of the entire population of each vial as explained in section 4.2, it is possible to obtain more quantitative information on the genome rearrangement that occurred between the two phages.

When examining the population composition at the last time step, at the bottom of figure 4.4, one can observe precisely which parts of each phage's genome are present in the population at the end of the experiment and in which quantity. As expected, most of the genome is composed of bas51, with some portions substituted with bas54 genome. Recombination event 1 is clearly observable, and bas54 genome reaches almost the maximum frequency in this region. However, a low level (5%) of bas51 genome was still detected. The second recombination event is also observable

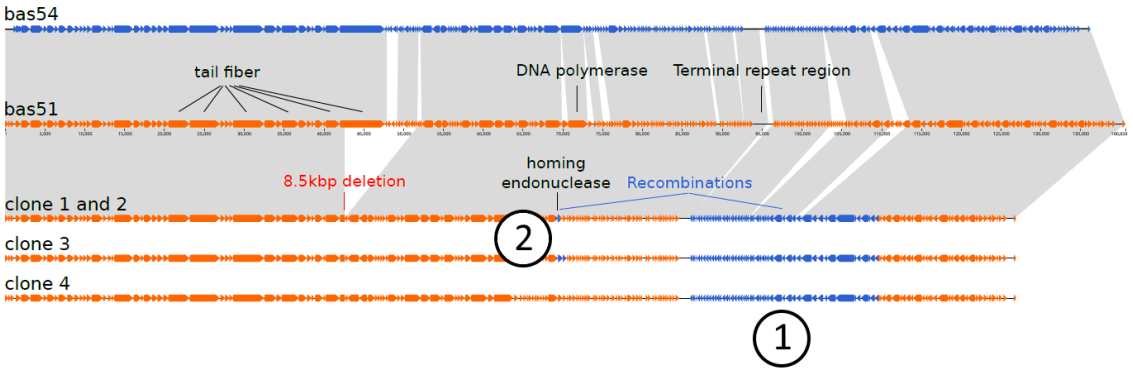


Figure 4.3: Figure taken from Druelle, Valentin (2024). bas51 and bas54 genome in comparison with recombined phage clones. ① indicates recombination event 1. ② indicates recombination event 2

and is present in 50% of the phage population. This explains why it was possible to find, by chance, a phage clone that lacks the bas54 homing nuclease.

In figure 4.5 the positions of the boundaries of recombination events predicted by the model are illustrated. By looking at the last time step, observations can be made regarding the recombination boundaries of recombination events 1 and 2 at the end of the experiment. For the first recombination event, it is observed that the recombination starts consistently at 95262 (in correspondence of the terminal repeated region) across all recombinants. Generally, it ends at base 123100, which falls within a putative P-loop NTPase protein. However, it is possible to observe a small fraction of recombinants (2%) that carry bas54 genome up to base 125434. Regarding the second recombination event, it always starts at 78670 but there is no univocal end. This heterogenicity could be due to a non-robust prediction of the model or to multiple independent transfers of bas54 homing nuclease that happened with slightly different boundaries.

Considering all the time steps represented in the two figures (4.4 and 4.5), observations can be made regarding the dynamics of the recombinant genomes throughout the experiment. It is evident that the largest recombination events start at a frequency of around 20% on day 1 and increase linearly up to day 7, reaching a frequency of 80%. On the other hand, the second recombination event appears in the population only on day 5, reaching 50% at day 7. Other interesting observations can be drawn by searching for recombination events that did not manage to spread in the population. At the first time step of figure 4.4, it is possible to observe a disproportion in the relative abundance of genomes between 70 and 80 kb; consequently, many recombination events are reported in that region in figure 4.5. These signals suggest that bas54 acquired 10 kb of bas51 at the beginning of

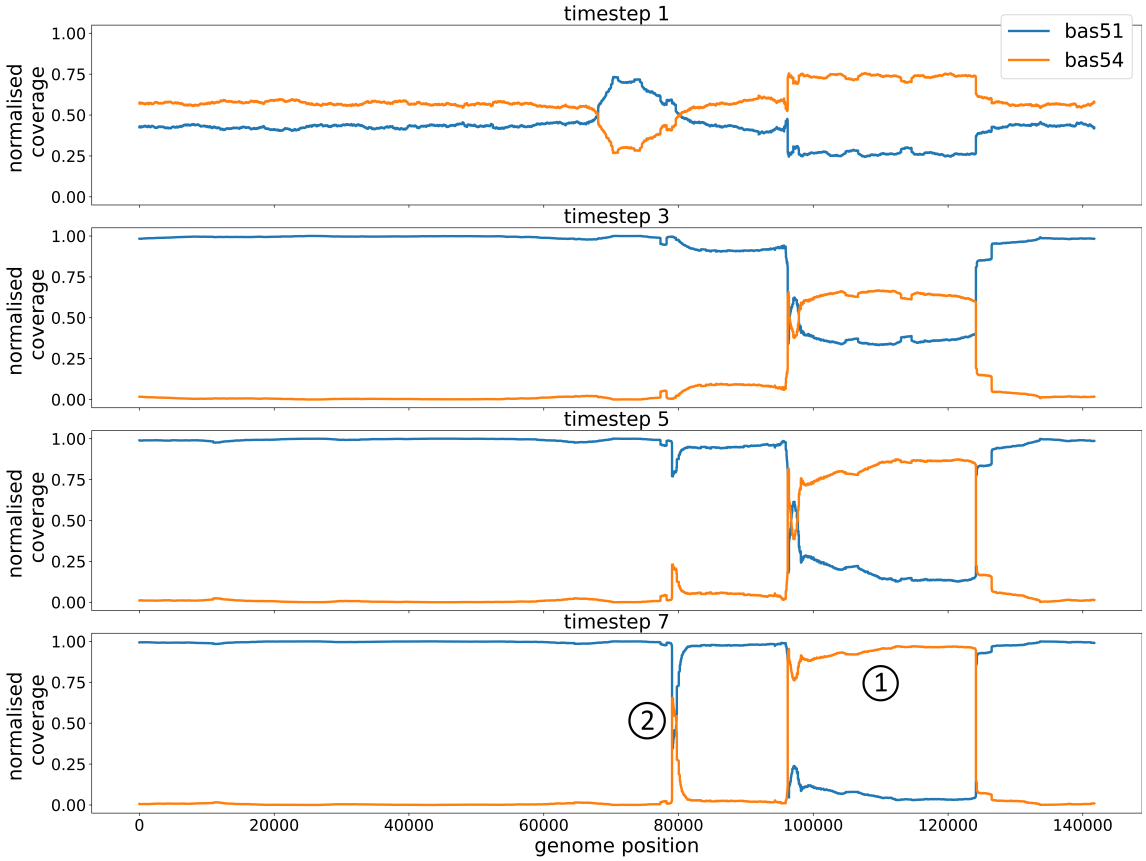


Figure 4.4: Plot of the normalised coverage for each time step of the recombination experiment, first vial. For each base, the fraction of reads belonging to each phage is plotted. ① indicates recombination event 1, which reaches around 90% of frequency. ② indicates recombination event 2, which reaches only 50% of frequency, this means that only half of the phage population will harbour such recombination event

the experiment. However, these recombinants disappear already at time step three, suggesting that the increase in fitness was not sufficient to allow their spread in the population.

Other 2 smaller recombination events can be detected thanks to the sensitivity of the analysis. The first is a low-frequency recombination event that stretches from base 76956 to base 77864 (just before the second recombination event). This region, which contains one hypothetical protein, is transferred from bas54 to bas51 already at the first time step, but it never reaches a high frequency (staying between 1 and 4%). Finally, right after the start of the first recombination event, a disproportion in relative genome coverage is observed. In this case, it seems that part of the terminal repeat region and some hypothetical proteins of bas51 are being transferred to the portion of bas54 that is retained in the recombinants.

In the second vial, bas51 phage took over on day 1, constituting more than 99%

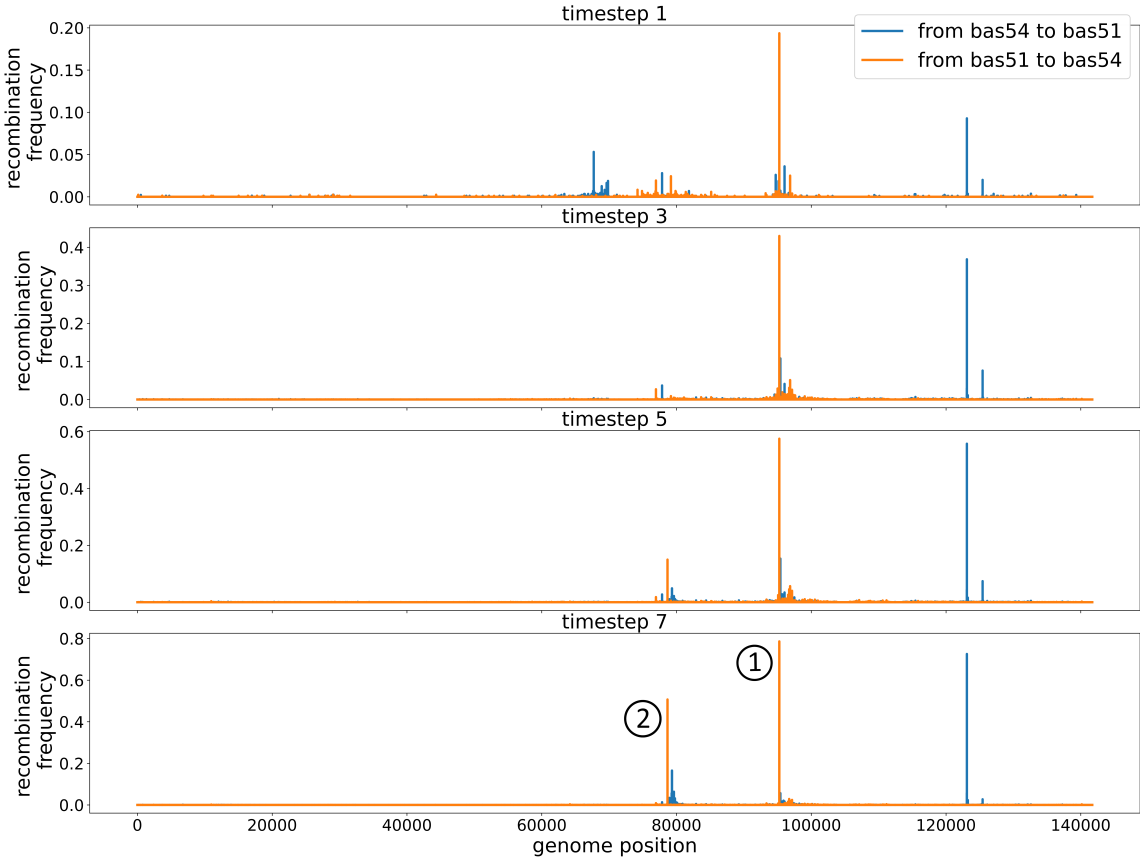


Figure 4.5: Plot of the recombination frequency for each time step of the recombination experiment, first vial. For each base, the fraction of reads that have a switch between the references in that position is plotted. ① indicates recombination event 1. ② indicates recombination event 2.

of the phage population, eliminating every chance of seeing any random recombination events. If a great imbalance between the two phages is established, the chances of a coinfection drop significantly. However, a clear increase in the bas54 genome in correspondence of the region involved in the first recombination event of vial 1 is detectable. This part of the bas54 genome reaches a quantity of 10% in the population at the end of the experiment. Figure 4.6 shows the normalized coverage of the last time step of the second vial. This is a sign of convergent evolution, supporting the idea that the observed recombination event is beneficial for the phages.

To determine whether genome recombination was actually advantageous for the phages, Valentin Druelle performed killing assays with the ancestral and evolved phages on *E. coli* K12 BW25113 wbbl(+) strain. In figure 4.7 the resulting killing curves are displayed. The blue and orange lines represent the killing ability of the original phage strains on wbbl(+) bacteria. the exponential growth of bacteria stops after around 3 hours but the OD600 value never drops significantly. On the other

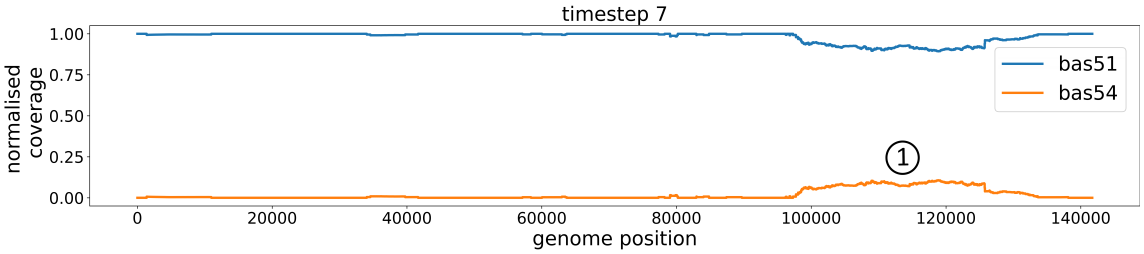


Figure 4.6: Plot of the normalised coverage in the second vial at the last time step. ① indicates a recombination event resembling recombination event 1 in the first vial. This is an evidence of convergent evolution.

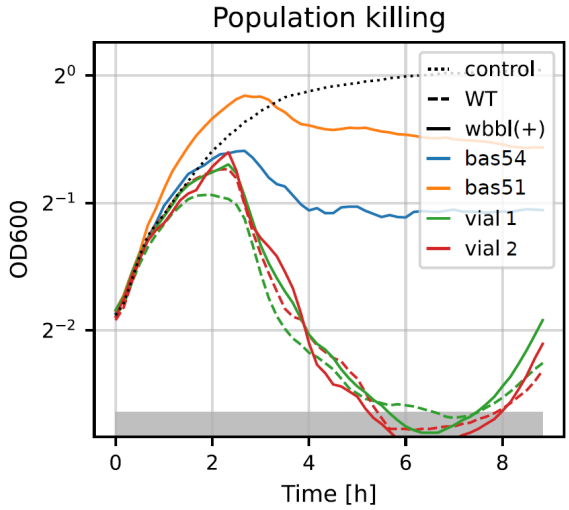


Figure 4.7: Killing curves of recombination experiment. The blue and orange lines represent the ancestral phage killing the wbbl(+) (challenging) bacterium. The green and red lines represent the evolved (recombinant) phages, from vial 1 and 2 respectively, killing the WT (dashed) and wbbl(+) (solid) bacteria. The evolution experiment caused an increase in the fitness of the phages.

hand, the green and red lines, representing respectively the phage populations from vial 1 and 2, cause a greater decrease in the bacterial density, suggesting that the evolved phages have a fitness advantage over their predecessors.

Chapter 5

Materials and methods

5.1 Phages and bacteria

The bacteriophages used in the experiments described in this thesis come from the BASEL collection (Maffei et al. (2021)), these phages belong to the Myoviridae family and Vequintavirinae subfamily. The phages are referred to as: phage WalterGehring (bas51, NCBI GenBank accession MZ501111.1), phage MaxBurger (bas54, accession MZ501093.1), and phage PaulScherrer (bas60, accession MZ501100.1). Phages from the BASEL collection are well genotypically characterized, and they are characterized for some important functions such as the target bacterial surface receptor, the interaction with host surface glycans, and resistance to bacterial immunity systems. These phages have genomes of 131 to 140 kb size that encode 3 sets of tail fibers that are coexpressed. This is a peculiar but shared feature among this group of phages, which likely enables them to attach to a variety of surface motifs on the bacterial surface.

These bacteriophages are well adapted to *E. coli* K12 BW25113, the partental strain from the Keio collection (Baba et al. (2006)). During the adaptive laboratory evolution experiments, a more challenging strain was used to allow the phages to mutate and adapt to it. The employed strain is *E. coli* K12 BW25112 wbbl (+), which has a restored O-16 type O-antigen that adds long chains on the lipopolysaccharide (LPS) and other expolysaccharides that effectively shield the cell surface, unless they can be degraded or serve as the phage's primary receptor (Maffei et al. (2021)). The phages used in this study are imparied by the O-antigen, but infection is not completely inhibited. This is likely due to their ability to bind another surface glycan. (Sellner et al. (2021))

5.2 Phage amplification

The samples taken from the evolution experiment typically ranged between 10^6 and 10^9 PFU/mL. For applications requiring higher quantities of phages, such as DNA sequencing, the phage samples were amplified in liquid cultures. To reduce bias from the original sample, the amplification process was designed with a high initial phage concentration and a short incubation period to limit the number of replication cycles.

To prepare each amplified phage stock, tubes were first inoculated with 1 mL of LB and 300 μL of *E. coli* BW25113 wbbl(+) from an overnight culture. These tubes were incubated at 37°C, shaking at 600 RPM for 20 minutes to reactivate bacterial growth. Following this, 100 μL of the phage sample intended for amplification was added to the tubes, which were then incubated for 3 hours at 37°C, shaking at 600 RPM. Finally, the tubes were cleared of bacteria by adding 1% chloroform, vortexing and centrifuging the tubes for 10 minutes at 8000g. The supernatant was then collected and titered, generally reaching concentrations between 10^{10} and 10^{12} PFU/mL. These amplified samples were stored at 4°C until needed for successive experiments (e.g. sequencing).

5.3 Phage DNA extraction

DNA of bacteriophages was prepared from high-titer stocks produced as explained in section 5.2. The DNA was extracted using the Norgen Biotek Phage DNA Isolation Kit according to the manufacturer guidelines. When the DNA amount was to low for subsequent sequencing, the samples were concentrated using a SpeedVac vacuum concentrator. Quality of the DNA was controlled using a Nanodrop device and was sequenced as explained in section 5.4.

5.4 DNA sequencing

The isolated phage genomes (see section 5.3) were sequenced in-house, using Oxford Nanopore sequencing Techonology. The MinION Mk1B device was used for sequencing, with V14 chemistry coupled with R10.4.1 pores and FLO-MIN144 flowcells. The multiple samples coming from the time steps were sequenced all at once using the rapid barcoding sequencing kit 24, specifically the kit SQK-RBK114.24. For the basecalling process, Dorado version 0.4.1+6c4c636 was employed, using the

basecalling model dna_r10.4.1_e8.2_400bps_sup version 4.2.0, on a cluster node with 32 cores, 100 gb of memory and 4 A100 GPUs. The basecalling pipeline used is available here: https://github.com/vdruelle/nanopore_basecalling.

5.5 Plate reader killing curves

The killing curves shown in Figures 3.4 and 4.7 were generated using an Epoch2 plate reader in absorbance mode (OD600). Phages were tested on *E. coli* BW25113 and *E. coli* BW25113 wbbl(+) at a multiplicity of infection (MOI) of 1 to 1000. Each well contained 180 μL of a bacterial dilution in LB at 5×10^8 CFU/mL and 20 μL of phages at 5×10^6 PFU/mL. Phage stocks were titrated and diluted in PBS to achieve the target MOI of 1 to 1000. The plate was incubated in the Epoch2 plate reader for 15 hours at 37°C with 450 RPM double orbital rotation, and OD600 readings were taken every 10 minutes.

5.6 Genome assembly

Genome assembly refers to the process of taking a large number of short DNA sequences and putting them back together to create a representation of the original chromosomes from which the DNA originated (Li et al. (2021)). In this thesis, de novo assembly of bacteriophage genomes was performed.

To perform denovo genome assembly, the Flye assembler (Kolmogorov et al. (2019)) was used on a cluster node with 4 cores and 16 gb of memory. Flye's algorithms use a repeat graph as core data structure. The edges of the repeat graph represent genomic sequence, and nodes define the overlapping segments found by approximate sequence matching. The genome traverses the graph in an unknown way, so as each unique edge appears exactly once in this traversal. Repeat graphs are useful for repeat analysis and resolution - which are one of the key genome assembly challenges (Kolmogorov et al. (2019)).

The parameters used for the phage genome assembly are: –nano-hq, to handle Oxford Nanopore Technology reads basecalled with Guppy5+; –genome-size with the length of the genbank references. –asm-coverage specifies the target coverage for the initial disjoint assembly, set to 40X.

5.7 Read mapping

Read mapping is the procedure of mapping reads against a reference genome (Schbath et al. (2012)). Minimap2 aligner (Li 2018) was used for all read mapping procedures of this thesis. Minimap2 is a versatile sequence alignment program that aligns DNA or mRNA sequences against a large reference database. Minimap2 uses a heuristic approach to solve the read mapping problem in a fast and accurate way, employing the seed-chain-extend algorithm. This algorithm consists of four steps. First of all the reference sequence is indexed, meaning that k-mers are created and minimised. The sequence is divided in overlapping fragments of length k and then only the most representative k-mers are kept. Successively, in the seeding step, the aligner searches for all the k-mer matches between the query and the reference. Then, in the chaining phase, the collinear seeds (k-mers that appear in the same order in both the reference and the query sequence) are grouped in a chain. Finally in the extension phase, each seed part of a chain is extended into a full alignment using dynamic programming techniques. In every mapping task performed in this thesis the -ax map-ont option of minimap2 was used, which works best with Oxford Nanopore Technology reads.

Chapter 6

Conclusion and future directions

This thesis presents a high-throughput framework for performing automatic experimental bacteriophage evolution. At the core of this framework is the Aionostat machine, supported by a comprehensive data analysis procedure for deep sequencing data. The described bioinformatic pipelines enable the quantitative tracking every genomic change happening in the phage genome, including recombination events, using a single deep sequencing run of a phage population.

The key result of this thesis is the convergent evolution observed in independently evolved phage populations. Similar mutations observed in different phages demonstrate the biological relevance and robustness of this approach. Moreover, the deep sequencing approach reliably detects recombination events from population data, opening the door to high-throughput studies of accurate recombination dynamics between phages. This method is more informative and effective compared to analyzing single clones from the population. This approach has become feasible only recently, thanks to the reduced costs and increased accuracy and read length of the sequencing technology.

This framework, together with functional analyses, can provide advancements in the understanding of phage molecular biology. Future research directions could also include exploring its use in the study of complex microbial communities, resistance evolution in bacteria, and the development of phage therapies. Nonetheless, this machine has limitations, such as the difficulty of construction and in the impossibility to replicate environmental complexity.

In conclusion, automatic evolution experiments together with deep sequencing analyses constitute an important step towards achieving high-throughput evolution studies. This approach can improve the understanding of phage biology and speed up the development of phage therapies.

Acknowledgements

I am deeply grateful to Professor Richard Neher for welcoming me into his lab and making me feel like part of a close-knit, supportive group. His supervision and ideas have been indispensible to my research experience. I extend my heartfelt gratitude to my supervisors, Valentin Druelle and Marco Molari, for their extensive guidance from wet lab techniques to bioinformatics. I appreciated a lot their patience and dedication throughout my learning journey. I am thankful to Professor Federico Manuel Giorgi for his supervision and guidance in the writing of this thesis. Finally, I am eternally greateful to the Biozentrum research summer program and its administrative organisers for providing me with this incredible opportunity. This experience not only led to the development of this thesis but also shaped my future and gave me the chance to make wonderful new friends.

Bibliography

Altamirano, G., L., F., and Barr, J. J. (2019). Phage therapy in the postantibiotic era. *Clinical Microbiology Reviews*, 32(2).

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L., and Mori, H. (2006). Construction of escherichia coli kâ12 inâframe, singleâgene knockout mutants: the keio collection. *Molecular Systems Biology*, 2(1).

Burrowes, B., Molineux, I., and Fralick, J. (2019). Directed in vitro evolution of therapeutic bacteriophages: The appelmans protocol. *Viruses*, 11(3):241.

Delahaye, C. and Nicolas, J. (2021). Sequencing dna with nanopores: Troubles and biases. *PLOS ONE*, 16(10):e0257521.

Dion, M. B., Oechslin, F., and Moineau, S. (2020). Phage diversity, genomics and phylogeny. *Nature Reviews Microbiology*, 18(3):125–138.

Druelle, Valentin (2024). Evolutionary dynamics in the virosphere: from hiv-1 to bacteriophage evolution.

Hendrix, R. W. (2008). *Evolution of dsDNA Tailed Phages*, pages 219–227. Elsevier.

Holtzman, T., Globus, R., Molshanski-Mor, S., Ben-Shem, A., Yosef, I., and Qimron, U. (2020). A continuous evolution system for contracting the host range of bacteriophage t7. *Scientific Reports*, 10(1).

Kauffman, K. M., Chang, W. K., Brown, J. M., Hussain, F. A., Yang, J., Polz, M. F., and Kelly, L. (2022). Resolving the structure of phageâbacteria interactions in the context of natural diversity. *Nature Communications*, 13(1).

Knipe, D. M. and Howley, P. (2013). *Fields virology*. Wolters Kluwer, Philadelphia, PA, 6th edition.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37(5):540–546.

Kunisaki, H. and Tanji, Y. (2009). Intercrossing of phage genomes in a phage cocktail and stable coexistence with escherichia coli o157:h7 in anaerobic continuous culture. *Applied Microbiology and Biotechnology*, 85(5):1533–1540.

Li, J., de Vries, R. P., and Peng, M. (2021). *Bioinformatics Approaches for Fungal Biotechnology*, pages 536–554. Elsevier.

Maffei, E., Shaidullina, A., Burkolter, M., Heyer, Y., Estermann, F., Druelle, V., Sauer, P., Willi, L., Michaelis, S., Hilbi, H., Thaler, D. S., and Harms, A. (2021). Systematic exploration of escherichia coli phageâhost interactions with the basel phage collection. *PLOS Biology*, 19(11):e3001424.

Molder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., and Koster, J. (2021). Sustainable data analysis with snakemake. *F1000Research*, 10:33.

Murray, C. J. L., Ikuta, K. S., Sharara, F., et al. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, 399(10325):629–655.

Pires, D. P., Cleto, S., Sillankorva, S., Azeredo, J., and Lu, T. K. (2016). Genetically engineered phages: a review of advances over the last decade. *Microbiology and Molecular Biology Reviews*, 80(3):523–543.

Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V., and Gibrat, J.-F. (2012). Mapping reads on a genomic sequence: An algorithmic overview and a practical comparative analysis. *Journal of Computational Biology*, 19(6):796–813.

Sellner, B., Prakapaite, R., van Berkum, M., Heinemann, M., Harms, A., and Jenal, U. (2021). A new sugar for an old phage: a c-di-gmp-dependent polysaccharide pathway sensitizes escherichia coli for bacteriophage infection. *mBio*, 12(6).

Strathdee, S. A., Hatfull, G. F., Mutalik, V. K., and Schooley, R. T. (2023). Phage therapy: From biological mechanisms to future directions. *Cell*, 186(1):17–31.

Suttle, C. A. (2007). Marine viruses â major players in the global ecosystem. *Nature Reviews Microbiology*, 5(10):801–812.

Tisza, M. J. and 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).

Toprak, E., Veres, A., Yildiz, S., Pedraza, J. M., Chait, R., Paulsson, J., and Kishony, R. (2013). Building a morbidostat: an automated continuous-culture device for studying bacterial drug resistance under dynamically sustained drug inhibition. *Nature Protocols*, 8(3):555–567.

Turner, D., Shkoporov, A. N., Lood, C., et al. (2023). Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ictv bacterial viruses subcommittee. *Archives of Virology*, 168(2).