Evolutionary dynamics in the virosphere

From HIV-1 to Bacteriophage evolution

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
Vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Valentin Druelle

Basel, 2024

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Erstbetreuer: Prof. Dr. Richard Neher

Zweitbetreuer: Prof. Dr. Erik van Nimwegen

Externer Experte: Prof. Dr. Roland Regoes

Basel, 27.02.2024

Dekan,
Prof. Dr. Marcel Mayor

SUMMARY

Evolution is a fundamental force shaping all life on Earth. Viruses, the most numerous and diverse biological entities on the planet, excel in evolution and thrive in many hosts and environments. The study of their evolutionary dynamics, which are essential to their success, has significant implications for public health. Historical and recent pandemics have shown the considerable impact that viruses can have on society, and understanding their evolution is therefore essential to mitigate their effects, help control disease spread, design better vaccines and antiviral drugs, and create new innovative treatments.

Studies of HIV-1 biology and evolution enabled the creation of life-saving treatments for infected patients. Despite this considerable achievement, we lack a satisfactory explanation of how HIV-1's within-host evolution generates its global diversity. In the first part of this thesis, we sought to explain this discrepancy by investigating the evolutionary dynamics at play on both scales. We showed that between-host evolution can mostly be explained from within-host dynamics if one accounts for the changing immune pressure that the virus faces from one host to the next. The evolution of the virus, constrained by the immune response of the patient, leads to the emergence of many escape mutations that are relevant only in that specific host. When infecting a new host, the different immune pressure causes the reversion of previously acquired mutations to their original state. On longer time scales, we thus observe a slower evolution driven by adaptation to changing environments.

In the second part of this thesis, we study the evolution of another type of virus: the bacteriophages. These viruses infect bacteria and are much more numerous and diverse than human viruses. Bacteriophages hold great promise for a wide range of research fields such as ecology, healthcare and molecular biology. Their viral nature and diversity makes them great candidates to investigate viral evolutionary dynamics. However, phage research is currently limited to a handful of well-characterized bacteriophage models, or to broad metagenomics studies where the phages are rarely isolated and poorly characterized. The former limits the scope of the findings, while the latter cannot provide the detailed characterization that would require experimental intervention. This depth vs. breadth dichotomy hinders our ability to comprehensively study phage evolution, and we sought to bridge this gap in two ways. First by creating a collection of phages, the BASEL phage collection, that is representative of the natural diversity of *E.coli* phages but where individual phages are also well-characterized. This gives a detailed snapshot of the results of natural phage evolution,

which is informative of the evolutionary trade-offs that these phages face. Our second approach to address the dichotomy is to enable phage evolution experiments at scale. To achieve this, we created a high-throughput framework to perform bacteriophage evolution rapidly, reliably and at scale. The central piece of this framework is the continuous culture machine we crafted to perform the bacteriophage evolution experiment: the Aionostat. We present the machine and the results of two experiments to showcase its abilities. In these experiments, we evolved phages to increase their infectivity on a challenging bacterial strain, demonstrating that the Aionostat can drive the evolution of bacteriophages both vertically and through horizontal transfers. Both approaches complement each other and open new avenues for bacteriophage research.

ACKNOWLEDGMENTS

This PhD often felt like a 4 year adventure in the wild - it was full of surprises, slightly terrifying but ultimately rewarding. This dissertation, the output of years of research, learning, and more coffee, tea, pizza and cakes than I'd care to admit, wouldn't have been possible without the support of some incredible people that I would like to thank here.

First and foremost, I am tremendously grateful to my supervisor professor Richard Neher as well as my non-official supervisor professor Alexander Harms, who are both amazing scientists and mentors. Your guidance was like a lighthouse in the choppy seas of my PhD voyage. Through health challenges, the unexpected twists of a global pandemic, and some personal growth along the way, you were there to make sure the journey went well. Your endless support and teachings provided the push I needed to keep going, and I needed a lot! Coming from a physics background, every piece of bioinformatics or wet lab knowledge I know is thanks to you. In summary, thank you for seeing potential in me since the beginning. I hope this dissertation mirrors the dedication and perseverance you inspired in me.

I would like to thank professor Roland Regoes and professor Erik van Nimwegen for kindly accepting to be part of the committee alongside professor Richard Neher and professor Alexander Harms. Thank you for your feedback over the year. Extra thanks to professors Roland Regoes and Richard Neher for their patience and time to evaluate this dissertation.

A massive shoutout to my colleagues, both past and present members of the Neher Lab and the Phage Hunter team. Whether it was by sharing knowledge, helping troubleshoot experiments, or simply keeping spirits high with your jokes, you have all contributed to making this a great journey. In particular, I thank Marco and Pierre for all the input and fun they provided over the years, along with the endless supply of sweets to keep the morale high. Janos, thank you for being a great officemate and inadvertently training my immunity. Emma, thank you for being so cheerful and bringing some British humour in the lab. Giacomo, although your stay with us was short, it was super nice and helpful, thanks a lot ! Huge thanks as well to all my wet lab colleagues that helped show me the ropes of this scary world. Reto, thanks a lot for your help in building and running the lab and nanopore, as well as for our daily interactions! The same thanks goes to Enea, Aisylu and Yannik, without whom I would not be able to hold a pipette correctly or perform experiments on phages. Finally, a huge thank you to my colleagues outside of the Neher lab and phage

hunter team. In particular, many thanks to Leonardo and Michelle for their cheerful attitude, help and tips that prevented many phases of despair, I am truly grateful for having such great friends. Nico, Beni and Pablo, many thanks as well for the great help, protocols and insights.

I also need to dedicate their own paragraph to the people who helped me with some technical aspects of my project. Massive round of applause for the people from the Research Instrumentation Facility, electronic workshop and mechanical workshop of the Biozentrum. I especially need to thank Lajko and Richard from the RIF for teaching me how to use all the great machines in their facility as well as providing help and feedback on my projects. You made me a 3D printing and DIY addict, and I hope you are proud of yourselves because I am having a blast! Big shoutout as well to Simon and Christian. The Aionostat would be dangerous, ugly and far worse without your help, and that would be a shame! Patrick, thank you as well for your help regarding everything mechanical.

To all my friends, whether I crossed paths with you during my PhD in Basel, my undergraduate years in Lausanne, or even earlier, each of you has played a unique role in shaping my life as it is today. Thank you for enduring the endless "PhD talk" and thank you for adding richness to my life and making it enjoyable every day.

Finally I want to thank my family, who contributed to making me the person I am today. To my grandparents, who are the best I could hope for. To my parents, whom I thank for their guidance, love and unwavering support, along with the many nice dishes and "cake au saumon" that fueled my brain throughout these years. The same thanks goes to my brother, who I know is also supportive and proud of what I do despite the distance. Last but not least, thank you to Jocelyne for sharing my life and making my PhD time as enjoyable as it is.

PUBLICATIONS

- Druelle, Valentin and Neher, Richard A. "Reversions to consensus are positively selected in HIV-1 and bias substitution rate estimates." In: *Virus Evolution*, (2023), veac118.
- Maffei, Enea and Shaidullina, Aisylu and Burkolter, Marco and Heyer, Yannik and Estermann, Fabienne and Druelle, Valentin and Sauer, Patrick and Willi, Luc and Michaelis, Sarah and Hilbi, Hubert and others. "Systematic exploration of Escherichia coli phage–host interactions with the BASEL phage collection." In: *PLoS Biology*, 19.11 (2021), e3001424.
- Neher, Richard A and Dyrdak, Robert and Druelle, Valentin and Hodcroft, Emma B and Albert, Jan. "Potential impact of seasonal forcing on a SARS-CoV-2 pandemic." In: *Swiss medical weekly*, (2020), w20224.
- Noll, Nicholas B and Aksamentov, Ivan and Druelle, Valentin and Badenhorst, Abrie and Ronzani, Bruno and Jefferies, Gavin and Albert, Jan and Neher, Richard A. "COVID-19 Scenarios: an interactive tool to explore the spread and associated morbidity and mortality of SARS-CoV-2." In: *MedRxiv*, (2020), pp. 2020–05.
- Wenner, Nicolas and Bertola, Anouk and Larsson, Louise and Rocker, Andrea and Bekele, Nahimi Amare and Sauerbeck, Chris and Rocha, Leonardo F Lemos and Druelle, Valentin and Harms, Alexander and Diard, Mederic "Phenotypic heterogeneity drives phage-bacteria coevolution in the intestinal tract." In: *bioRxiv*, (2023), pp. 2023–11.

CONTENTS

1 INTRODUCTION

In this chapter we present an overview of the general biology and concepts necessary to understand the thesis. We will introduce the topics treated in the following chapters in detail when it becomes relevant. Section 1.1 provides a general motivation to the field of viral evolution. Section 1.2 is an elementary introduction to the biology traits shared by viruses, both in terms of structure and lifecycle. Section 1.3 introduces viral evolution, the fundamental process at play and the interesting evolutionary dynamics that emerge from these processes. Section 1.4 provides some background to viral evolution research and how it is performed, which introduces some of the approaches used in this thesis. Finally section 1.5 defines the aims of this thesis.

1.1 MOTIVATION

Evolution is a fundamental force that drives changes in all forms of life, from the simplest organism to the most complex. This relentless process of change and adaptation is particularly evident in the world of viruses. The incredible diversity of viruses out there is both a testament to and source of many evolutionary changes, making them an ideal subject for studying the principles of evolution. Understanding how viruses evolve not only sheds light on these tiny yet impactful entities, but also provides broader insights into evolutionary mechanisms that affect all life forms.

In public health, the significance of viral evolution is evident. Historical outbreaks like the 1918 Influenza pandemic (often referred to as Spanish flu) and the 1980s HIV/AIDS crisis highlight the devastating impact of evolving viruses on human populations [1, 2]. More recent outbreaks, such as those caused by SARS-CoV-2 and Ebola, reinforce this point. These events have not only led to widespread health consequences and morbidity, but also profound socio-economic disruptions. The COVID-19 pandemic, in particular, has demonstrated the sheer impact a virus can have on society, necessitating widespread confinement measures and reshaping daily life [3, 4].

The control of viral spread and impact is intertwined with the understanding of their evolution. Accurate predictions of viral evolution are crucial for designing effective vaccines, as seen in the flu and COVID-19 responses [5]. Similarly, the development of antiviral drugs relies on anticipating and circumventing viral resistance mechanisms that could appear through viral evolution. Interestingly, the study of viral evolution is not only about combating harmful viruses; it can also help

us tackle other menacing health issues. For instance, bacterial viruses can be used to fight infections from harmful bacteria, a treatment known as phage therapy [6, 7]. In this area, a deep understanding of bacteriophage evolution is also key to the success of these therapeutic strategies.

Beyond human health, viral evolution significantly influences ecological dynamics. The interplay between viruses and their hosts profoundly influences ecosystems across the globe, from the depths of the oceans to the expanse of forests. The predation of viruses often drives processes that help maintain the balance of ecosystems, a prime example being the role of bacteriophages in the ocean carbon shunt [8]. Their ability to transfer genetic material between different species also contributes to biodiversity and the evolutionary trajectories of countless organisms.

In summary, the study of viral evolution is not just a matter of addressing immediate health concerns. It is also about deepening our understanding of evolution as a fundamental biological process on multiple scales.

1.2 Fundamentals of virus biology

Viruses exhibit a remarkable diversity in structure and function, yet they all share a defining feature: they need to infect host cells to replicate and produce more copies of themselves. This common objective gives rise to shared fundamental traits across different viral species. In this section we present these shared features, starting with their structure and then following with their life cycle. We purposefully ignore traits that are not shared widely by viruses. Additional details relevant for understanding HIV-1 and bacteriophages biology will be introduced in chapter 2 and 3.

1.2.1 *Viral structure*

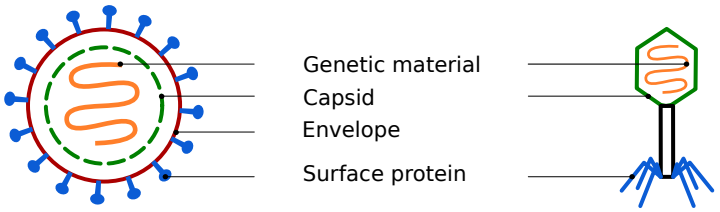


Figure 1.1: General structure of a virus. The left schematic typically resembles a human virus, while the right schematic typically resembles a bacteriophage. Viral morphologies are extremely diverse in reality.

GENETIC MATERIAL: At the heart of every virus lies its genetic material, which stores the information necessary for viral replication.

This genetic code is the blueprint for producing more viruses. The material on which this code is stored can vary, it might be RNA or DNA, single-stranded or double-stranded. The Baltimore classification, one of the most commonly used systems to classify viruses, is based on the genomic material of viruses and its characteristics, which are used to separate them into 7 distinct groups.

CAPSID AND ENVELOPE: Encasing the genetic material is usually a protective protein shell called a capsid. Its main function is to protect the virus genetic material from the outside. Some viruses, like SARS-CoV-2 or HIV, go a step further by having an envelope [9, 10]. This envelope, derived from the host's own cell membrane, adds an extra layer of protection and aids in the process of infecting new host cells. Such viruses are often more fragile than their non-enveloped counterparts [11, 12].

SURFACE PROTEINS: In proximity to the capsid (or in the envelope when relevant) are surface proteins, which role is to mediate the interaction with the host cells. They are called receptor binding proteins and are located on other structures such as tail fibers in the context of bacteriophages [13]. These proteins are key to the virus's entry into the host cell. An exemplary case is the spike protein of SARS-CoV-2, which binds to the ACE2 receptor on human cells [14]. This binding is the first critical step in the virus's infectious cycle, providing host recognition and facilitating entry into the cell. Since these proteins are exposed at the surface of the virus, they are often targeted by antibodies and are therefore under strong evolutionary pressure.

1.2.2 *Virus infection cycle*

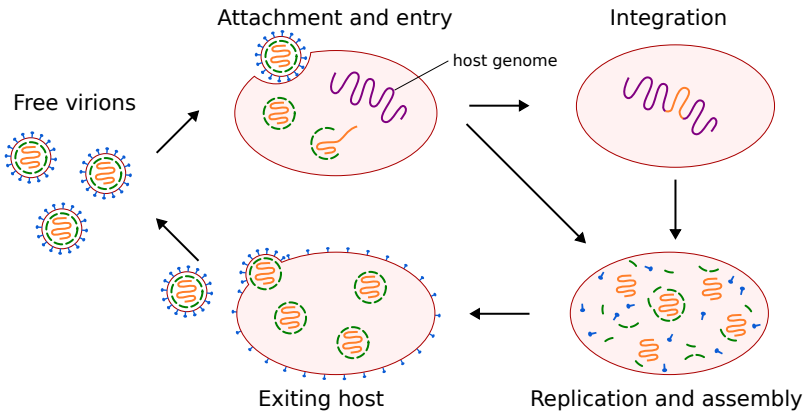


Figure 1.2: Generalized life cycle of a virus.

ATTACHMENT AND ENTRY: The infection cycle begins with the virus attaching itself to a host cell. This attachment is typically mediated by surface proteins on the virus, which recognize and bind to receptors on the host cell's surface. After entry, the viral genome is released into the host cell.

SPECIAL CASE - INTEGRATION: In some instances, such as with retroviruses or lysogenic viruses, the viral genome integrates into the host genomic material. This integration represents a unique strategy, allowing the viral genome to stay within the host cell for extended periods before resuming its replication cycle.

REPLICATION AND ASSEMBLY: Once inside the host cell (and possibly after a period of integration), the virus starts replicating its genome and producing the building blocks for new viral particles using the host cell resources. The newly made genetic material and structural components then assemble into complete, functional new viruses.

EXITING HOST TO REPEAT CYCLE: The final stage in the virus's life cycle is to release the newly formed virion from the cell. Different viruses have different exit strategies - some bud off from the host cell, enveloping themselves in a piece of the cell membrane, while others cause the host cell to burst (lysis), releasing the new viruses. Once outside the host cell, these new viruses can in turn infect new cells and reproduce the same cycle. The duration of a full infection cycle varies among viruses and is influenced by other factors such as the host's state. This is typically around 10 to 20 hours for human viruses, while it can be as short as 20-25 minutes for bacteriophages [15, 16].

1.3 VIRAL EVOLUTION

Each replication cycle of a virus presents an opportunity for modification. As viruses replicate, changes can occur in their genetic material. These modifications, when combined with natural selection, drive the evolution of viruses over time, leading to various evolutionary dynamics that make viruses so successful.

1.3.1 *Fundamental evolutionary processes*

VERTICAL EVOLUTION: During the replication of a virus's genome, random errors can occur. These mutations are often insignificant or detrimental, but a subset of them may benefit the newly formed virus. Recent methods such as mutational scans allow us to generate many mutated viruses to study the mutational fitness effects [17–20]. These fitness effects can also be measured in-vivo by studying

mutations trajectories over time [21]. Overall it seems that among these random mutations, the majority are either lethal or detrimental, with a small fraction being neutral or having a positive impact on the fitness of the virus. This picture varies drastically between synonymous and non-synonymous mutations though. These mutations can be transmitted from parent to offspring virus, and as time progresses, natural selection can amplify the prevalence of beneficial mutations and remove the detrimental ones.

HORIZONTAL EVOLUTION: When two distinct viruses infect the same host cell simultaneously, it creates an opportunity for their genetic material to mix, which offers additional pathways for viral evolution. This process can occur in two ways: reassortment and recombination. The former is relevant for viruses with segmented genomes, such as Influenza. In this scenario, offspring viruses can inherit distinct genomic fragments from both parent viruses [22, 23]. Recombination, on the other hand, involves the exchange of genetic material within a single genome segment and is particularly relevant for viruses such as HIV-1 [24]. In both cases, the offspring viruses inherit genetic material from both parental viruses, a process which enables exchange of large sequences of DNA between viruses and promotes genetic diversity. Without horizontal transfers viruses would evolve like an asexual population. Horizontal evolution introduces some gene shuffling, which helps to maintain diversity and explore a broader range of genetic possibilities. This typically improves the ability to adapt to changing conditions, a trait which is likely beneficial to viruses. Overall, horizontal transfers are widely recognized for their importance in viral evolution, but studying their dynamics and impact remains challenging and is therefore and active field of research [25, 26].

SELECTION: Selection acts on the viral variants created through the processes mentioned above, selecting for variants that are better in the context of this selective pressure. In viruses, selection often favors traits that enhance survival, replication efficiency, or transmission capabilities. Factors such as the environment, host immunity, and therapeutic interventions all contribute to the selective pressure on viruses. These selective pressures, coupled with the creation of new variants, are the driving forces behind viral evolution. It enables viruses to adapt to changing conditions, which can lead to interesting evolutionary dynamics.

1.3.2 *Evolutionary dynamics*

Different viral lifestyles and the pressures they encounter lead to a variety of evolutionary dynamics. These dynamics often overlap and

interconnect, reflecting the complex nature of viral adaptation and survival strategies. Here we introduce several evolutionary dynamics that are relevant for the work presented in chapter 2 and 3. The extent to which they influence viral behavior depends greatly on the virus's lifestyle. Factors like the type of genetic material, replication strategies, and interactions with hosts play significant roles in determining the evolutionary dynamics observed.

VIRAL EVOLUTION RATE AND DIVERSIFICATION: When selection does not change too drastically, the rate at which viruses evolve is relatively constant on the timescale of years [27]. It can be measured by counting the number of mutations accumulated over time. This rate is often called the molecular clock of a virus, and it can provide insights into how quickly a virus can adapt to new conditions, which varies between viruses. High mutation rates, common in RNA viruses like HIV-1 or SARS-CoV-2, lead to rapid diversification and adaptation to changing immune landscape. This rapid evolution plays a critical role in the virus's ability to evade immune responses and develop resistance to antiviral therapies. Understanding this evolution rate is essential for public health interventions, as it influences the development of effective vaccines, the prediction of virus spread, and the strategic deployment of antiviral drugs.

SPILLOVERS AND HOST ADAPTATION: Viral spillovers, where a virus jumps from one species to another, are pivotal events in viral evolution. Successful adaptation to a new host requires a virus to overcome numerous new biological barriers, which is a strong selective force [28]. These barriers can be at the attachment stage, where the different cell receptors of a new host are not well recognized by the viral surface proteins, at the replication stage, where the changes in host cell machinery can impair the creation of new virions, or even regarding the transmission from one host to the next. Overcoming these barriers is often promoted by recombination events that can provide new traits to the virus and is facilitated by the virus's inherent ability to mutate to subsequently adapt to this host. Such spillovers and adaptations are very relevant for public health as they seem to be at the source of many recent pandemics, such as the ones caused by Influenza H1N1 (2009) and potentially SARS-CoV-2 (2019), highlighting the need for research in this area [29, 30]. We study adaptation to a new host in the experiments presented in chapter 3.

CHANGING SELECTION AND ARMS RACE: Viruses and their hosts are engaged in a continuous evolutionary arms race. Host organisms develop mechanisms to detect and eliminate viruses, while viruses evolve strategies to evade host defenses. For instance, hosts can mutate their receptor proteins to prevent viral attachment, while viruses can

evolve to adapt their surface proteins to this modified target. This ever-changing selection due to host defenses can lead to significant changes in the virus, sometimes leading to co-evolution of the host and the virus. Additionally, the selective pressure from different hosts is sometimes sufficiently divergent that viruses face trade-offs between short-term benefits in the current host(s), and long-term survival strategies, such as transmissibility to new hosts. This concept is particularly evident in HIV, which we will cover in more detail in chapter 2. Due to their fast evolution and generation time, viruses are a great system to study host-pathogen arms race. Such dynamics are not restricted to viruses and their host, they are also extensively studied in other systems such as parasites and bacteria [31–34].

1.4 Approaches in viral evolution research

We have seen in the previous section that viruses have the means to evolve over time. This section presents an overview of the methodologies employed in studying this viral evolution, emphasizing the importance of several fields such as bioinformatics, epidemiology and molecular biology. Key examples from past research will be highlighted to illustrate how these methodologies have helped develop our understanding of viral evolution. Additional details will be provided on approaches that are particularly relevant to understanding the work presented in chapter 2 and 3.

1.4.1 *Sequencing data*

Studying viral evolution is all about tracking and understanding the genomic changes that happen over time in viral populations. Sequencing technologies provide a window to access the genomic sequences of viruses and are thus a central tool in studying viral evolution. The recent improvements in our understanding of viral evolution are closely linked to the increase in sequencing data available for such viruses.

Technologies

The increase in sequencing data available over the years is largely attributable to improvements in sequencing technologies coupled with a substantial reduction in sequencing costs. Although some technologies are capable of sequencing RNA directly, our focus will be on DNA sequencing methods, as they are more commonly used. When working with RNA viruses, the genomic material can be converted to DNA using methods such as reverse transcription prior to sequencing.

The improvement in sequencing capabilities began with Sanger sequencing, a method developed in 1977 that enabled the reading

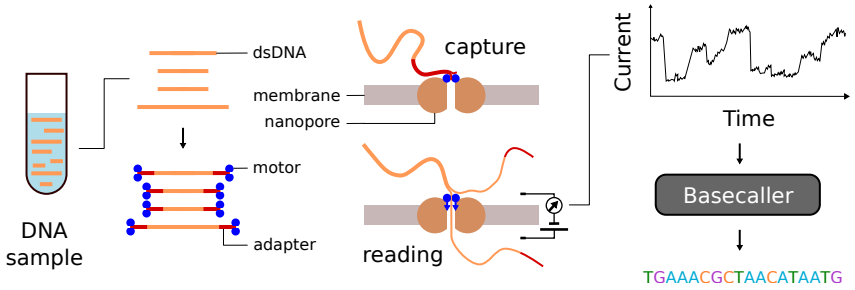


Figure 1.3: Schematic of the Nanopore sequencing technology. The DNA sample is first prepared and then spotted on the nanopore flowcell, where nanopores capture the DNA strands and motor proteins push a single strand through the pore. A potential difference is applied between both sides of the pore creating a current through the pore. The changes in this current are measured over time, and the basecaller translates this changes in current to nucleotide sequence.

of DNA sequences up to a few hundred base pairs in length [35]. It became the most widely used sequencing method for approximately 40 years, and is still used up to this day. It operates by using DNA polymerase to synthesize new DNA strands, incorporating dideoxynucleotides that cause chain termination at various points. This results in fragments of different lengths with distinct terminal nucleotide. Fragments are then separated by length through methods such as electrophoresis and the last nucleotide is identified using its radioactive or fluorescent labeling which enables reconstruction of the full sequence. Although it is highly accurate, it is limited in scalability and length of sequences it can read.

To overcome these limitations, next-generation sequencing (NGS) technologies, particularly Illumina sequencing, were developed in the mid-2000s. With this technology, many short DNA fragments from the sample are sequenced in parallel by measuring the signal of fluorescent labeled nucleotide as the DNA fragments are copied [36]. Illumina sequencing revolutionized the field by greatly increasing the throughput, allowing for the parallel sequencing of many small DNA fragments and consequently reducing costs. This technology is still widely used, but it is limited by its ability to sequence short fragments only.

New sequencing technologies such as Oxford Nanopore sequencing and PacBio sequencing appeared in the years 2010s, further improving our sequencing capabilities. We focus on Oxford Nanopore sequencing since this is the method we used for the work in chapter 3. Nanopore sequencing involves moving single DNA molecules through tiny protein pores, known as nanopores, embedded in a synthetic membrane [37]. As each nucleotide of the DNA strand passes through the nanopore, it causes a disruption of the electric current that

flows through the pore. This disruption is measured and recorded over time. A subsequent step called basecalling converts the electric signals into the corresponding nucleotide sequence. The working principle of Nanopore sequencing is illustrated in figure 1.3.

One of the remarkable features of Nanopore sequencing is its ability to read very long native DNA fragments, up to hundreds of thousands of bases. This contrasts with the shorter reads of Illumina sequencing and offers unique insights about difficult to sequence regions such as repeat regions. Additionally Nanopore sequencing is unique in its portability, real-time data generation, ease of use and has recently become relatively cheap to use. Nanopore's main limitation is the higher amount of miscalled nucleotides compared to methods such as illumina, but recent advancements have drastically improved the quality. For these reasons, we decided to do Nanopore only sequencing for the results presented in chapter 3, a service that we have also made available for the whole Biozentrum.

Sequencing approaches

PARTIAL / FULL GENOME SEQUENCING: The choice between partial and full genome sequencing depends on the research objectives. Partial genome sequencing focuses on specific regions of the viral genome and is generally more straight-forward and cost effective. It is often used for studying regions known to be highly variable or significant, such as receptor-binding domains in viruses. For example in influenza, the sequencing is often focused on the hemagglutinin (HA) and neuraminidase (NA) regions. In contrast, full-genome sequencing provides the full picture of the viral genome, which is essential for understanding the virus as a whole. Full-genome sequencing is usually more challenging and costly due to the increased amount of genetic material that needs to be sequenced.

CONSENSUS SEQUENCING: This involves sequencing several copies of the viral genome to produce a single, high-quality representative sequence. It can be done both with partial and full genome sequencing. This method smooths out individual variations between virions, or noise in the data, to present a consensus sequence of the most common nucleotides at each position. This approach is particularly common in the sequencing of human viruses isolated from patients, where it makes sense to identify the primary strain present in an infection even though it may not capture the full picture if there is some viral diversity.

DEEP SEQUENCING: Contrary to consensus sequencing, deep sequencing provides a more complete picture of viral populations by sequencing at a depth that allows the detection of rare variants. This approach enables the study of viral diversity within a host, tracking

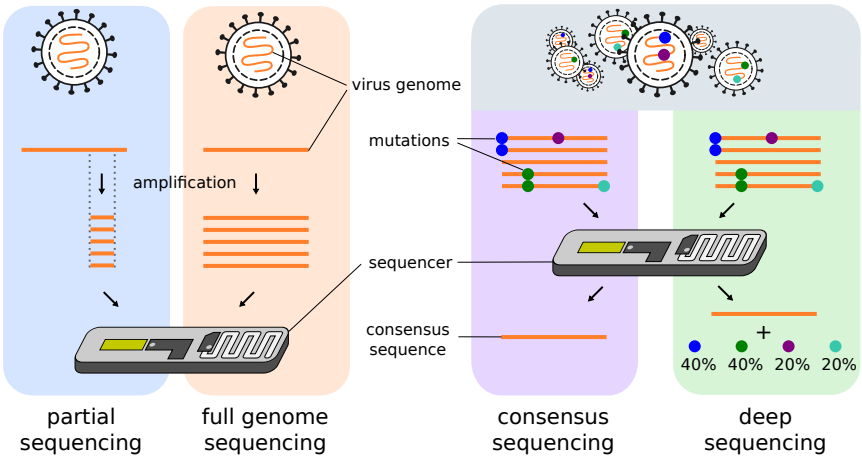


Figure 1.4: Illustration of sequencing approaches.

minor variants, and therefore studying the dynamics of viral evolution under different selective pressures. Deep sequencing offers a detailed picture of the viral population as a whole, an approach that is central to the work presented in chapters 2 and 3. This type of sequencing is more challenging than consensus sequencing, as one needs more sequencing depth to be able to detect minority variants in a population.

1.4.2 *Bioinformatics*

The advancements in sequencing technologies discussed previously have led to an exponential increase in the volume of genetic data available. This increase in sequencing data not only emphasized the importance of bioinformatics but also promoted growth in this field as more and more tools and analysis are required to extract actionable knowledge from the sequencing data.

Bioinformatics approaches have been central to the development of our understanding of viral evolution. A basic but essential example of its application is in reconstructing consensus genomes from viral samples taken from patients. This process involves the aggregation and comparison of many short DNA sequences coming from the sequencing process to create a representative consensus sequence of the virus infecting the patient. Such consensus sequences are at the center of our analyses of human virus spread and evolution. Further analyses based on these sequences enable the tracking of evolutionary patterns of viruses over time and predicting potential future changes, which is key for public health interventions or the design of effective vaccines. Such tasks, which would be unfeasible manually due to the sheer volume and complexity of the data, underscore the importance of bioinformatics in modern biological research.

In the following section we present some of the bioinformatics approach that are used to study viral evolution, with a focus on the methods relevant for the work presented in chapter 2 and 3.

1.4.3 *Studying natural evolution*

Studying viral evolution can be conducted through two primary methods: observing natural processes and conducting experimental research. In this section, we focus on the former, using the COVID-19 pandemic as an example to illustrate key aspects of studying natural viral evolution.

The COVID-19 pandemic provides an unparalleled case study for observing natural viral evolution. Coinciding with a time when advancements in sequencing technology (discussed previously) had made routine sequencing more feasible, the spread of SARS-CoV-2 was closely monitored on a global scale. Many countries and organizations rapidly implemented routine sequencing to track the virus's spread and evolution, sharing their findings with the global research community [38]. This resulted in an unprecedented collection of sequencing data in near-real time, which could be used to understand how SARS-CoV-2 evolves and spread. While the pandemic's impact is arguably devastating, it also presented an unparalleled opportunity to study viral evolution with a level of detail previously unattainable, making it a suitable example to illustrate the methodologies involved in viral evolution research presented below.

1.4.3.1 *Surveillance and sequencing*

Surveillance networks and organizations play a crucial role in the study of natural viral evolution as they provide the sequencing data which is at the core of the research. This data is often collected by medical practitioners from population samples and is fundamental for tracking the spread and evolution of viruses, including the identification of new variants. However, the effectiveness of bioinformatics analysis, which is usually done by researchers separate from those collecting the data, is dependent on the quality of this data. Inconsistencies in sampling or coverage of the pandemic can lead to blind spots and biases in our understanding of viral spread and evolution. The COVID-19 pandemic has shown that collaborations and public sharing of sequencing data is essential to provide a good basis on which viral evolution research can be developed.

1.4.3.2 *Epidemiological models*

The field of epidemiological modeling plays a central role in understanding and managing the spread of viruses. These models leverage the data collected from surveillance to help predict viral spread, which

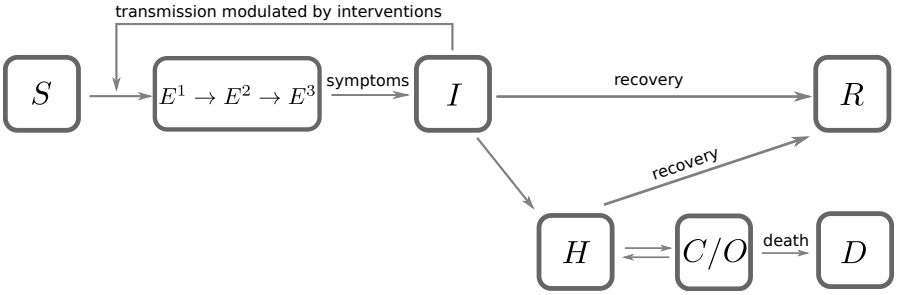


Figure 1.5: Diagram of the SIR model we used in [40]. S, E_i, I, R, H, C, O and D represent the susceptible, exposed, infectious, recovered, hospitalized, critical, overflow, and fatal compartments of the model. Each compartment was further stratified by age demographics. Figure adapted from [40].

is particularly important for public health strategies. The first phase of the COVID-19 pandemic highlighted this importance as epidemiological models were heavily used to inform mitigation strategies. We actively contributed to this work through two papers that were published in 2020 [39, 40].

There exist many types of epidemiological models, each adapted to different scenarios and with different complexities. One of the simplest types of models is the SIR (Susceptible, Infected, Recovered) model and its variants. These models use ordinary differential equations on population compartments to predict a deterministic average scenario based on fixed inputs for the epidemiological parameters. This is the type of model we used in [39] and [40] to monitor and predict the spread of COVID-19 in the initial phase of the pandemic. While deterministic models are somewhat limited in their scope, they are relatively simple to understand and implement. The accuracy of their prediction is dependent on the population structure chosen a priori, which can be complex when trying to be realistic, and the epidemiological parameters that are input into the model. The main challenge lies in the accurate inference of such parameters from real-world observations. For example, figure 1.5 shows a schematic of the model we used in [40] to simulate the early stages of the COVID-19 pandemic. It is an SIR model with additional compartments for improved realism.

Some of the other types of epidemiological models worth mentioning are:

- Stochastic models: these models incorporate randomness and variability into their parameters [41]. This approach is more realistic, as it acknowledges the inherent stochasticity in transmission chains and real-world scenarios. Stochastic models pro-

vide probability distributions of outcomes rather than a single predicted path.

- Agent-based models: these models take a different approach by simulating the actions and interactions of autonomous agents, which can represent individuals, small populations, or other entities [42]. These models can help capture complex social dynamics and individual behaviors. While they can offer a more detailed understanding of disease spread, their complexity can be a drawback as they require significant computational resources and data on agent behaviors.

- Spatial models: many of the models previously mentioned can be modified to introduce a spatial component to better understand how geographical factors and movement patterns influence disease spread. Such models can incorporate data on population density, transportation, and other spatial factors to predict how diseases will spread in different regions. A variant of such models would be the network models, which focus on the patterns of connections among individuals or groups and their impact on viral spread.

1.4.3.3 *Phylogenetics*

Phylogenetics is the study of the evolutionary relationships among biological entities, or subset of, like individual genes. It aims to construct a family tree, or phylogeny, that maps out these relationships, illustrating how different entities have evolved from common ancestors over time. In the context of viral evolution, phylogenetic trees are typically built from the DNA sequences of these viruses. This approach is heavily used as a tool to track, understand and potentially predict the evolutionary changes in viruses. Recent outbreaks like the COVID-19 pandemic highlight the importance of phylogenetics analysis as they are central in our understanding of how viruses evolve and help develop informed and effective public health interventions [43].

At their core, methods used for constructing phylogenetic trees involve the analysis of multiple sequence alignments (MSA) and try to find an evolutionary tree that can best explain the differences seen between the sequences of the MSA. The likelihood of a given phylogenetic tree is estimated based on a model of sequence evolution. Such models usually operate on DNA sequences as strings of characters (one for each site), each character being in one of four possible states: A, C, G and T. Such models can also be extended to protein evolution by using 20 states for the 20 amino acids.These models describe the probabilities of transitions between different states over time, denoted as $P_{ij}(t)$, where i and j represent different nucleotide states, and t represents the time. $P_{ij}(t)$ gives the probability that a site in state j will change to state i over a period of time t. A specific category of

substitution models are the time reversible models, which assume that the probability of a change from state i to j is the same as j to i at any given time. These models, operating under the assumption of an equilibrium in states concentration, randomness and independence of transitions are known as General(ised) Time Reversible (GTR) models [44]. This is the type of model we used in the work presented in chapter 2.

In such models, the transition matrix $\mathbf{P}(\mathbf{t})$ is described by the differential equation:

$$\frac{d\mathbf{P}}{dt} = \mathbf{P}(\mathbf{t})\mathbf{Q}$$

where $\mathbf{Q}$ is the rate matrix. This implies that:

$$\mathbf{P}(t) = \mathbf{e}^{\mathbf{Qt}}$$

The matrix element Q_{ij}^{α} describes the rate to go from nucleotide j to nucleotide i at site α. It can be generally described in this way:

$$Q_{ij}^{\alpha} = \mu^{\alpha} p_i^{\alpha} W_{ij}$$

In this equation, μ^{α} represents the overall rate of mutation for the site α , p_i accounts for the nucleotide preference at this site and W_{ij}, which is not dependent on the site (no α), accounts for the difference between transversion and transition. These parameters enable a precise description of the evolutionary dynamics in the model, but they can be challenging to infer from real data.

Phylogenetic tree builders use such substitution models to estimate the tree that fits the substitution model best. This assumes the transmission of such mutations is vertical, but these methods are often robust to some amount of horizontal gene transfer. There are two main computational approaches to building phylogenetic trees: maximum likelihood and Bayesian inference. Both methods use complex algorithms to analyze the MSA and estimate the most probable tree structure that explains the MSA. The maximum likelihood approach calculates the probability of observing the data given a particular tree structure and tries to find the tree that maximizes this probability. Bayesian methods are similar to maximum likelihood ones in the sense that they also use a probabilistic criterion to find the best tree, but this criterion is instead the probability of a tree conditional on the data and prior beliefs about the evolutionary process. The main difference is that maximum likelihood approaches give one optimal phylogenetic tree, while Bayesian methods sample many likely trees from the posterior distribution, providing a set of trees that represent the uncertainty in tree estimation. Given the importance of building phylogenetic trees in evolutionary research, several tools such as RAxML [45], IQTree [46] and BEAST [47] have been developed to help with tree inference.

Phylogenetic analyses are vital in retracing the evolutionary journey of viruses. They offer insights into key aspects like the origin and

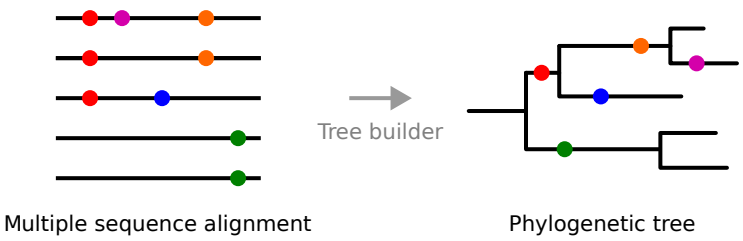


Figure 1.6: Simple example of phylogenetic reconstruction from a multiple sequence alignment. The DNA sequences are represented as lines and the mutations are shown as dots of different colors. The tree builder infers the phylogenetic tree from the mutations shared between sequences.

initial spread of a virus within human populations. The ability to reconstruct a virus history is not just about understanding its past, it can also help predict future trends. Many methods have been proposed to infer which viral strain is most likely to circulate in the future, with various degrees of success [48–52]. Such approaches effectively combine phylogenetics with epidemiology and help create viral spreading models that incorporate evolutionary information about the viruses. Such models are especially relevant in scenarios like influenza or SARS-CoV-2 where viral evolution is fast and interacts with a changing host immune landscape. Although this is a complex task, such models and prediction are crucial for guiding vaccine development as vaccine necessitate time to produce and often need to be strain specific to provide the best protection.

1.4.4 *Experimental approaches*

Experimental approaches are complementary to the observational methods discussed above as they make use of the controlled conditions of laboratory settings to test hypotheses and examine viral behaviors. They allow for precise manipulation of variables and conditions, enabling a more detailed exploration of viral behaviors, which is key for informing or validating models from observational studies. Some examples of how experimental approaches are utilized in viral evolution research are presented below.

VIRUS-HOST INTERACTION: Experimental approaches are extensively used to study virus-host interactions. An illustrative example is the use of human sera to explore cross-immunity. In these experiments, sera from hosts are exposed to different viral strains to observe how well the host antibodies are able to neutralize the virus. These experiments help in understanding how viruses like influenza evolve

to escape immunity from human hosts, providing insights into which strain is most likely to circulate in the future.

PHAGE TRAINING: In bacteriophage research, directed evolution experiments are used to 'train' phages to become more effective against bacterial targets [53]. This involves repeatedly exposing bacteriophages to their target bacteria under controlled conditions, observing their adaptation and selecting the most effective variants. This process is vital for the advancement of phage therapy as a viable alternative to traditional antibiotics, especially in the context of rising antibiotic resistance. Such phage training experiments are at the center of the work presented in chapter 3.

VIRAL DRUG RESISTANCE: Similarly to what is done to study the bacterial evolution of antibiotic resistance, experimental approaches are also used to understand how viruses evolve resistance to antiviral drugs. This can be done by either making experiments to directly evolve resistant viruses, or more commonly by studying viral samples from patients where treatments seem to be ineffective and studying the molecular changes that cause viral resistance. This knowledge is crucial for designing effective treatment strategies.

Such experimental approaches greatly benefit from the recent advancements in genetic engineering. The ability to create libraries of viral variants or to engineer phages in vitro greatly helps in testing and understanding viral evolution dynamics. Nonetheless, experimental approaches have limitations. Replicating the complex conditions of natural environments in a laboratory setting is challenging, and there are ethical concerns and risks associated with working with and evolving human-pathogenic viruses. However, these methods still contribute substantially to our understanding of viral evolution. They offer unique insights that complement natural surveillance and epidemiological studies of viral evolution.

1.5 AIM OF THE THESIS

Years of study of viral evolution have brought a lot to public health and molecular biology. From the creation of effective HIV therapies to the development and update of vaccines to manage outbreaks, understanding how viruses evolve has been instrumental in shaping modern healthcare and scientific knowledge. Building upon this, my thesis aims to dive deeper into the dynamics of viral evolution with the following aims:

1. Study and characterize how HIV-1 evolves both intra-host and inter-host, and explain how the evolutionary dynamics at the

pandemic level emerge from the peculiar evolution happening within-host.

2. Create a complete framework for high-throughput studies of bacteriophage evolution through directed evolution experiment. If successful, it will enable a better study and optimization of bacteriophage evolution but also provide general insights about viral evolution dynamics such as recombination between viruses.

2

UNDERSTANDING HIV-1 EVOLUTION: EVOLUTIONARY IMPACT OF REVERSION TO CONSENSUS

This chapter discusses our published work on HIV-1 evolution. This work focuses on the tracking and characterization of reversion to consensus mutations in patients infected with HIV-1, showing that these mutations are positively selected and largely responsible for the differences observed between within-host and between-host HIV-1 evolution. We start with an HIV specific introduction to provide motivation and context in section 2.1 followed by the publication in section 2.2.

2.1 INTRODUCTION TO HIV-1

2.1.1 *Historical background and relevance*

The Human Immunodeficiency Virus (HIV) was discovered in the early 1980s, primarily in the United States, where it was first recognized to be the cause of a rise of Acquired Immunodeficiency Syndrome (AIDS) cases in otherwise healthy young men. This marked the beginning of one of the most deadly pandemic in recent years. It is estimated that, since the beginning of the pandemic, 85 million people have become infected with HIV and 40 million have died [54]. Due to its significant health burden, HIV is a well studied and characterized virus.

HIV is thought to have been transmitted to human as a result of multiple spillover events from Simian Immunodeficiency Virus (SIV), which led to the emergence of two distinct types: HIV-1 and HIV-2 [2]. Among these, HIV-1 has diversified into several groups since its jump to humans, see figure 2.1. HIV-1 group M is thought to account for 90% or more of HIV infections and is therefore split into subgroups [55]. The most recent common ancestor of HIV-1 group M is estimated to be at the beginning of the 20th century, long before the recognition of the AIDS pandemic, and therefore had time to diversify into subtypes [56, 57]. Due to its prevalence, HIV-1 group M is the focus of our work.

2.1.2 *Characteristics of HIV-1*

HIV-1 is a lentivirus, a single-stranded RNA virus encapsulated by a lipid envelope. It primarily replicates by recognizing and entering human cells using the CD4 receptor and therefore targets key immune system cells such as the helper T-lymphocytes and macrophages. HIV-

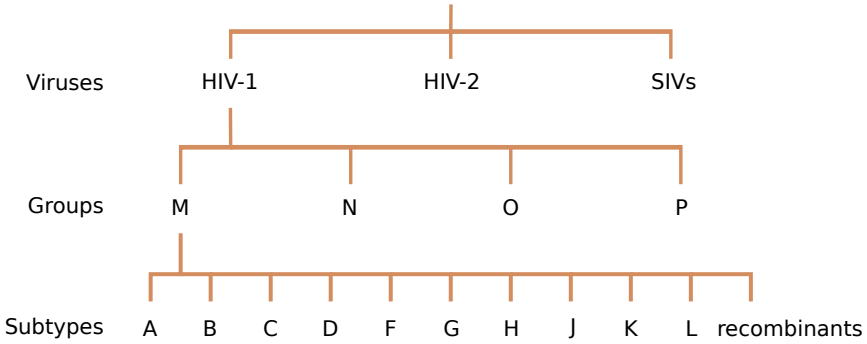


Figure 2.1: Schematic of HIV phylogeny. HIV-1 and HIV-2 are related to SIVs. HIV-1 is divided into 4 subgroups. HIV-1 group M causes the majority of infections and is consequently further divided into subtypes.

1 is a retrovirus, so upon entry into the host cell its RNA genome is reverse transcribed into DNA and inserted in the host own genetic material. The viral genes can then be transcribed to produce new virions immediately, or lie dormant within-host cells for extended periods of time before activation.

An HIV-1 virion has, like all retroviruses, two copies of its RNA genome inside its capsid. The reasons for having two copies of its genome are still unclear, but it is believed to be beneficial for the virus as it increases the chances for recombinations and potentially also increase genetic stability and fitness in case of deleterious mutations in one of the copies [58]. The capsid also contains some helper proteins as shown in figure 2.2. The viral genome is about 10 000 base pair long and encodes 9 genes over 3 reading frames [59]:

- ***gag***: Encodes structural proteins for the virus, crucial for virus assembly and maturation.
- ***pol***: Codes for viral enzymes like the reverse transcriptase, integrase, and protease.
- ***env***: Produces surfaces proteins, important for the virus's ability recognize and enter host cells.
- ***tat***: A regulatory gene enhancing viral transcription efficiency.
- ***rev***: Involved in RNA transport from the nucleus to the cytoplasm.
- ***nef***: Plays a role in immune evasion.
- ***vif***: Promotes virion maturation and infectivity.
- ***vpr***: Host cell cycle control.
- ***vpu***: Helps in new virus particle release and degrades CD_4.

Some of these genes code for large polyproteins that are later cleaved into smaller functional proteins. The three main genes, *env*, *pol* and *gag*, cover about 80% of the genome. Our publication focuses on the analysis of mutations in these three genes.

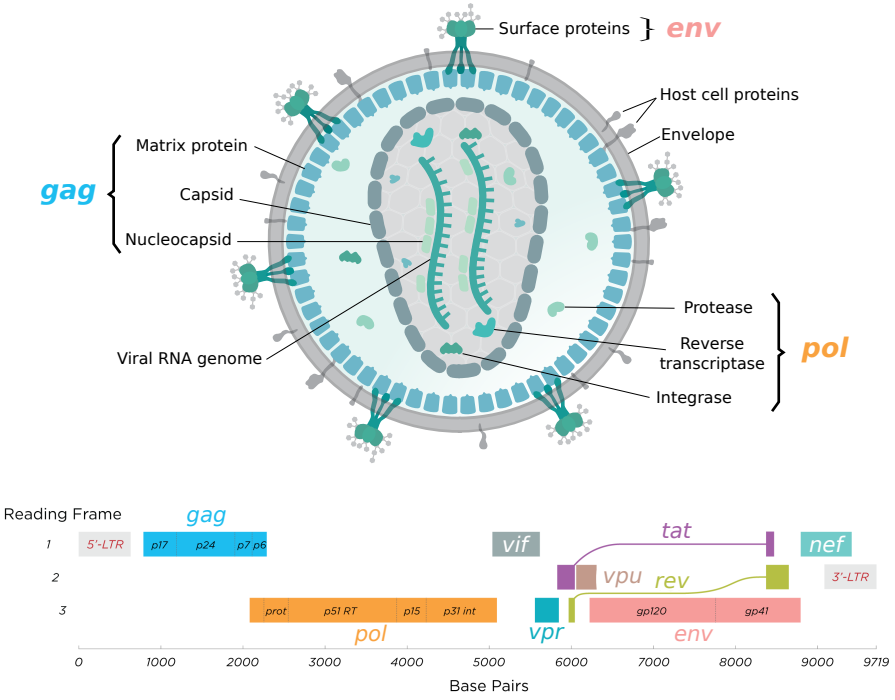


Figure 2.2: Schematic of the HIV virion and its RNA genome. Figure reproduced and adapted from [60] and [61] under the Creative Commons Attribution - ShareAlike License 4.0 [62].

2.1.3 *Unique challenges in HIV-1 evolution*

HIV-1 has some peculiar characteristics that greatly impact the evolutionary dynamics observed. Below are the main characteristics that make it special.

RAPID EVOLUTION: The mutation rate of HIV-1 is remarkably high, estimated at approximately $1.5 \cdot 10^{-5}$ to $3.5 \cdot 10^{-5}$ mutations per base per replication cycle from in vitro experiments [63, 64]. This is primarily due to the error-prone nature of its reverse transcriptase enzyme, a trait shared by many RNA viruses [65]. On a pandemic scale, this raw mutation rate translates to an evolution rate of around 10^{-3} mutation per site per year. However, accurately estimating this rate is complex, as it varies across different regions of the genome and is sensitive to the time scale used to measure it. The impact of the time scale used on the evolution rate measured is the central subject of our publication.

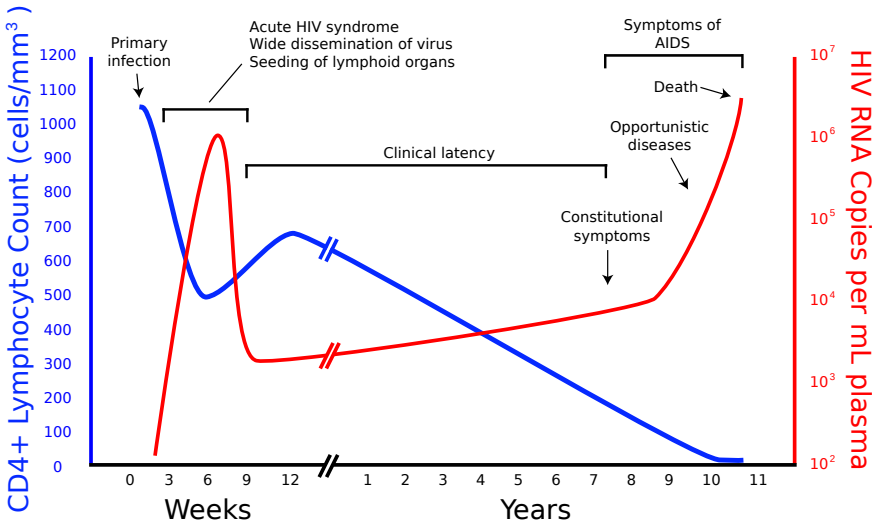


Figure 2.3: A generalized graph of the relationship between HIV copies (viral load) and CD4 counts over the average course of untreated HIV infection; any particular individual's disease course may vary considerably. Reproduced from [67] under the Creative Commons Attribution CC0 1.0 Universal Public Domain Dedication [68].

RECOMBINATIONS: Within-host recombination rates in HIV-1 are notably high, a phenomenon primarily attributed to the substantial viral load during infection [66]. Viral load, which is the concentration of virions in the bloodstream, typically varies from 10^3 to 10^7 virions per milliliter, depending on the stage of the disease and differences between hosts. The factors contributing to, and consequences of higher or lower viral loads are an active research field. The high number of virions circulating in a host gives plenty of opportunities for co-infection of a host cell, which can then result in recombinations. This is further increased by the high rate of template switching between the two genome copies during replication [58]. Although we did not study recombinations directly in this publication, it is important to acknowledge their role as they allow for the decoupling of mutations from one another and therefore have a big effect on HIV-1 evolution.

LIFE-LONG INFECTIONS: HIV-1's ability to evade the immune system is partially due to its high mutation and recombination rate, constantly adapting to stay ahead of the adaptive immune responses. As a retrovirus, HIV-1 also integrates its DNA into the host's genome, allowing it to become dormant and create latent reservoirs which can be reactivated later on. This aspect, combined with the targeting of long-lived memory immune cells, allows the virus to persist for extended periods in their hosts, as shown in figure 2.3. The killing of immune cells eventually leads to the development of AIDS without treatment. When under anti-retroviral therapy (ART) the production of new virions is stopped, which also halts the evolution of HIV-1 in

the host. Nevertheless, latent reservoirs can restart the infection if the therapy is stopped. We focused our work on non-treated patients.

VIRAL DIVERSITY: HIV-1 accumulates a substantial genetic diversity within an individual host over time. This can be attributed to several contributing factors. These include the daily production of a high volume of virions, estimated to be in the range of 10^8 to 10^{10} per day, the aforementioned rapid mutation and recombination rate, the persistence of infections that can last for years as shown in 2.3, and the continuous pressure exerted by the host's immune system. The diversity seen within hosts is also responsible for the large diversity observed at the between-host level and the many HIV-1 group M subtypes that exist and have about 10% to 20% sequence difference. This level of diversity is high when compared to other human viruses such as SARS-COV-2, and it is one of the reasons why vaccine development is challenging.

These characteristics of HIV-1, along with the significant public health threat it poses, have made it one of the first viruses where researchers could study evolutionary dynamics in depth, both within and between-hosts. This was the case even when sequencing capabilities were limited to only a few hundred base pairs at a time. The insights gained from these studies have been instrumental in shaping our current understanding of viral evolution and in guiding ongoing efforts to combat HIV-1.

2.2 PUBLICATION

VIRUS EVOLUTION

Virus Evolution, 2023, **9(1)**, 1–10
DOI: https://doi.org/10.1093/ve/veac118
Advance access publication date: 15 December 2022
Research article

Reversions to consensus are positively selected in HIV-1 and bias substitution rate estimates

Valentin Druelle[1,2,†] and Richard A. Neher[1,2,‡,*]

[1]Biozentrum University of Basel, Spitalstrasse 41, Basel 4056, Switzerland and [2]Swiss Institute of Bioinformatics, Spitalstrasse 41, Basel 4056, Switzerland
[†]https://0000-0002-2554-4982
[‡]https://orcid.org/0000-0003-2525-1407
*Corresponding author: E-mail: richard.neher@unibas.ch

Abstract

Human immunodeficiency virus 1 (HIV-1) is a rapidly evolving virus able to evade host immunity through rapid adaptation during chronic infection. The HIV-1 group M has diversified since its zoonosis into several subtypes at a rate of the order of 10^{-3} changes per site per year. This rate varies between different parts of the genome, and its inference is sensitive to the timescale and diversity spanned by the sequence data used. Higher rates are estimated on short timescales and particularly for within-host evolution, while rate estimates spanning decades or the entire HIV-1 pandemic tend to be lower. The underlying causes of this difference are not well understood. We investigate here the role of rapid reversions toward a preferred evolutionary sequence state on multiple timescales. We show that within-host reversion mutations are under positive selection and contribute substantially to sequence turnover, especially at conserved sites. We then use the rates of reversions and non-reversions estimated from longitudinal within-host data to parameterize a phylogenetic sequence evolution model. Sequence simulation of this model on HIV-1 phylogenies reproduces diversity and apparent evolutionary rates of HIV-1 in *gag* and *pol*, suggesting that a tendency to rapidly revert to a consensus-like state can explain much of the time dependence of evolutionary rate estimates in HIV-1.

1. Introduction

RNA viruses have low-fidelity polymerases, resulting in rapidly diversifying virus populations, which, in turn, facilitate the adaptation to changing environments. The human immunodeficiency virus 1 (HIV-1) is a prime example of such a rapidly evolving virus. The life-long infections it causes are characterized by a large viral population that accumulates diversity at a high rate to constantly evade host immunity (Coffin and Swanstrom 2013). This continuous evolution has led to a diverse viral population on the pandemic scale that is categorized into several viral subtypes (Brian Foley 2018; Li et al. 2015). Different lineages have accumulated diversity at a rate of about one substitution in 1,000 sites per year since its jump to human hosts at the turn of the 20th century (McCutchan 2006; Sharp, Hahn 2011; Korber et al. 2000).

Quantifying the rate of viral evolution, however, is surprisingly difficult and different approaches yield different answers. Most importantly, the timescale across which sequences are compared strongly affects the estimates, sometimes by orders of magnitude: the longer the timescale, the lower the estimate (Aiewsakun and Katzourakis 2016; Hanada et al. 2004; Worobey et al. 2010; Gilbert and Feschotte 2010; Ghafari et al. 2021). These discrepancies suggest that we lack a good understanding of how microevolutionary within-host (WH) processes—on the scales of days, months, and years—give rise to the diversity observed on longer timescales across hosts. In the case of chronic infections such as HIV-1, these microevolutionary processes are driven by selection to evade the host immune selection and mutations that reduce recognition can spread even if they reduce replication fitness. The pattern of immune selection changes at each transmission events and previously adaptive changes can become deleterious in the new host and sometimes revert (Leslie et al. 2004).

HIV-1 is an ideal system to study these effects in detail as the rate discrepancies among the WH, pandemic, and broader scales are well documented (Alizon and Fraser 2013; Worobey et al. 2010), the pandemic is well sampled, and high-resolution WH data exist. The evolutionary rate estimated on the pandemic scale is around two to five times lower than the one observed on the WH scale (Alizon and Fraser 2013). Several hypotheses have been put forward to explain this phenomenon. Two of the main hypotheses are the preferential transmission of ancestral HIV-1 variants, i.e. the 'store and retrieve' hypothesis (Lythgoe and Fraser 2012), and rapid reversion toward an ancestral-like state, i.e. the 'adapt and revert' hypothesis (Redd et al. 2012; Zanini et al. 2015; Leslie et al. 2004; Boutwell et al. 2010; Herbeck et al. 2006; Illingworth et al. 2020). The relative importance of these and possibly other processes for the discrepancy of rate estimates is not well understood (Raghwani et al. 2018).

We use WH longitudinal deep-sequencing data to explore how the rapid evolutionary processes within hosts can give rise to

apparently lower rates on longer timescales. These results suggest that the 'adapt and revert' mechanism can explain most of the rate mismatch observed at different timescales of the HIV-1 pandemic. We, firstly, show that HIV-1 sequence evolution shows strong signs of site saturation while distance relative to the root of the tree increases much more slowly than expected based on the rate of evolution. Similar signatures are observed in longitudinal WH data, suggesting that this saturation is independent of whether evolution is quantified along transmission chains or within hosts. Secondly, we investigate the cause of this saturation and find that WH reversion toward the HIV-1 consensus is more common than expected and that such reversions are positively selected. Lastly, we use simulations of evolution to quantify the impact of rapid reversions on rate estimates for timescales of decades or more. These simulations show that the degree of reversion observed within hosts can explain the phylogenetic patterns observed in the pandemic. More generally, our results highlight the evolutionary bias of viruses toward a state of high intrinsic fitness in a changing environment.

2. Results

We use (1) a set of sequences representative of the HIV-1 pandemic spanning multiple decades and (2) a longitudinal data set following the evolution of the virus within individual hosts to investigate patterns of evolution on multiple timescales. The former between-host (BH) data set contains 1,000 HIV-1 group M sequences from the Los Alamos National Laboratory (LANL) HIV database (Foley et al. 2013). This subsampling was performed to have the same number of sequences for each year to avoid sampling biases (except for early years, where fewer sequences are available) but otherwise randomly picked from the full data set. The phylogenetic tree was inferred using an IQ-TREE GTR+F+R10 model (Tavaré and others 1986; Yang 1995; Minh et al. 2020), which was found to be the best model according to the IQ-TREE ModelFinder and allows for rate variation (Kalyaanamoorthy et al. 2017). For more details on the phylogenetic analysis and the estimates of rates, see Sections 4.2 and 4.3.

Our WH analysis is based on the HIVEVO data set (Zanini et al. 2015), a whole-genome deep sequencing of HIV-1 populations in eleven patients during a 4–16-year follow-up without treatment. Between six and twelve samples are available per patient, which typically cover 5–7 years of infection. Sequencing depth and template input of all samples in this data set have been assessed and most samples allow a confident calling of frequencies of minor variation down to a few per cent (Zanini et al. 2016). See Section 4.1 for details.

We analyze the evolution of the *env*, *pol*, and *gag* genes of HIV-1 Section 2.1 to 2.4. They code for surface proteins, viral enzymes, and capsid proteins, respectively (Freed 2001). When combined, they cover approximately 80 per cent of the genome. We focus on the *pol* region in the main text and present analogous results for the *env* and *gag* regions in the Supplementary Materials.

2.1 Saturation and reversion effects are comparable between and within hosts

The 'adapt-and-revert' mechanism to explain the rate mismatch within and between hosts assumes that reversions during WH evolution 'shadow' previous changes, resulting in very low rate estimates. The 'store-and-retrieve' mechanism postulates that many WH changes are not transmitted and thus irrelevant for the evolution on longer timescales (Lythgoe and Fraser 2012). To look for such discrepancies between WH and BH evolutionary patterns, we compared the rates at which sequences diverge away from the root of the HIV-1 tree or their subtypes at the BH and WH scales, see Fig. 1A.

The rate at which divergence between sequences increases decreases with distance as more and more sites are hit multiple times by mutations (Felsenstein, 2004). For very similar sequences multiple hits are negligible and divergence increases linearly in time with a slope given by the evolutionary rate. If all sites evolve at the same speed, such saturation effects are only important once distances between sequences are large (the size of correction is proportional to the distance squared and thus substantial if distances are 0.25 or larger). However, if different sites evolve at drastically different rates, or reversions to a preferred state are common, such saturation effects set in much earlier and can lead to significant deviations even when sequences are still very similar (Puller et al. 2020; Ghafari et al. 2021). The 'adapt-and-revert' mechanism thus posits strong saturation effects of similar magnitude both within and between hosts when compared to distant references such as the root of the HIV-1 M tree.

Sequences in the HIV-1 pandemic differ from each other at about 10–20 per cent of sites and we would naively expect that saturation effects are small unless rate variation is very strong or reversion is a substantial contribution to evolution. Figure 1B explores the observable consequences of such saturation on the scale of the HIV pandemic for a simple substitution model with gamma-distributed rate variation. The panel shows the evolutionary distance to the root of the tree corrected for saturation effects in blue. The latter is simply the evolutionary rate times time and increases thus linearly with time. In addition, it shows the Hamming distance to the root of the tree in orange. Saturation effects are visible as reduced Hamming distances that increase more slowly over time, but the effects are small. As expected, saturation effects are even less pronounced when comparing sequences to the root of the subtypes (here assumed to be in 1965, compare Fig. 1A) or a 'founder' sequence in 1980.

Figure 1C shows the analogous patterns for HIV-1. The Hamming distance of HIV-1 sequences from the inferred root of the HIV-1 group M tree (orange) is substantially lower than the RTT distance (blue) and increases only at about half the rate, suggesting substantial saturation. Similarly, the Hamming distances to the subtype consensus (only done for Subtypes B and C) increase less rapidly than the RTT distance, despite the fact that at 2–5 per cent sequence divergences from the subtype root saturation effects are unexpected. Such rapid saturation can arise through rate variation (Soubrier et al. 2012) or heavily skewed site-specific equilibrium frequencies resulting in rapid reversion (Halpern and Bruno, 1998; Hilton and Bloom 2018; Puller et al. 2020; Ho et al. 2005; Wertheim and Kosakovsky Pond 2011).

We then performed a similar analysis on WH data on a 5-year timescale to determine whether similar rates and saturation effects exist within hosts. We compute WH evolutionary rates by measuring the divergence over time in Fig. 2D. Specifically, we calculate the divergence $d(t)$ relative to a reference sequence, such as the root of the tree, according to:

$$d(t) = \frac{1}{N}\sum_{i}^{N} \delta_i^{ref} \cdot (1 - f_i(t)) + (1 - \delta_i^{ref}) \cdot f_i(t), \quad (1)$$

where N is the length of the region and $f_i(t)$ is the frequency of the founder nucleotide at position i and time t in the viral population. This founder nucleotide at each position i is approximated as the majority nucleotide at the first time point t_0, and

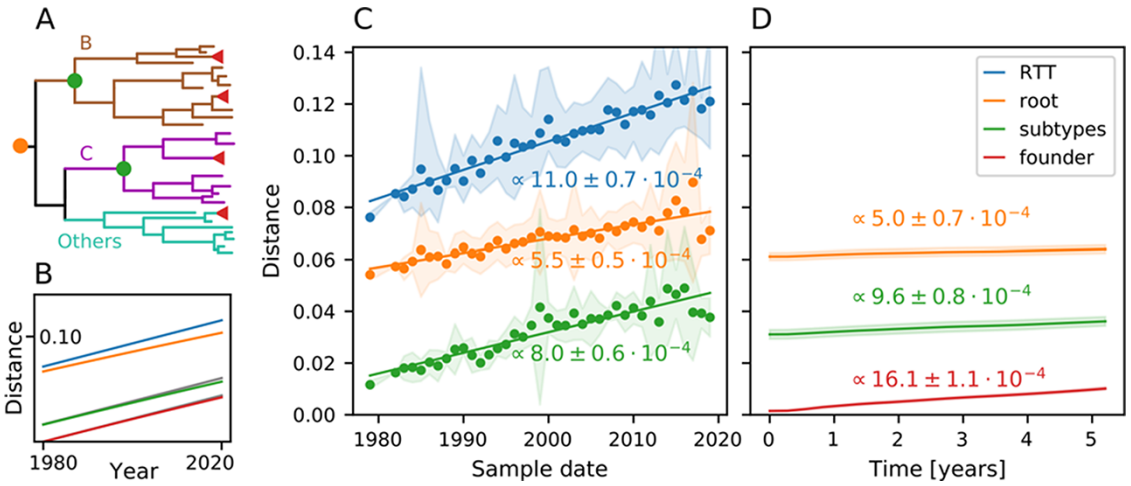


Figure 1. Divergence over time in the *pol* gene. (A) Sketch of the HIV-1 group M phylogenetic tree and its subtypes. Dots correspond to the position of the references used to compute distances in the other panels. WH evolution is indicated by red triangles. (B) Expected Hamming distances under a Jukes–Cantor (Jukes et al., 1969) evolution model with rate variation (gamma distributed, Parameter 2). Different curves show expected distance to the root of the tree (orange), subtype root (green), and WH founder (red). The blue and gray indicate linear growth of distance without saturations with a rate equal to the estimate from the root-to-tip (RTT) distance in Panel C. As expected, saturation effects are small since distances are around 10 per cent and multiple hits are rare. (C) Average Hamming distance from the root of the HIV-1 group M tree (orange), from the respective subtype (green, see dots in Panel A), or RTT distance in a phylogeny as a function of time. Each data point is the average of sequences from one year, lines are linear fits, and the shaded area indicates the 10–90 per cent range. (D) The WH divergence over time relative to the putative founder genotype, the HIV-1 group M root, and the subtype consensuses, averaged over all patients in the HIVEVO data set. Divergence is computed according to Equations (1) and (2). Standard estimates for the evolution rates BH and WH are the slopes of the RTT distance (blue) and divergence from founder sequence (red). There is an approximately 50 per cent difference between the evolution rates estimated while sequence distance is only a couple per cent. Comparing to the expectation (B), we can see that significant saturation of comparable magnitude can be seen on both BH (C) and WH (D) scales. Results for regions *env* and *gag* are shown in Supplementary Figs. S1 and S2.

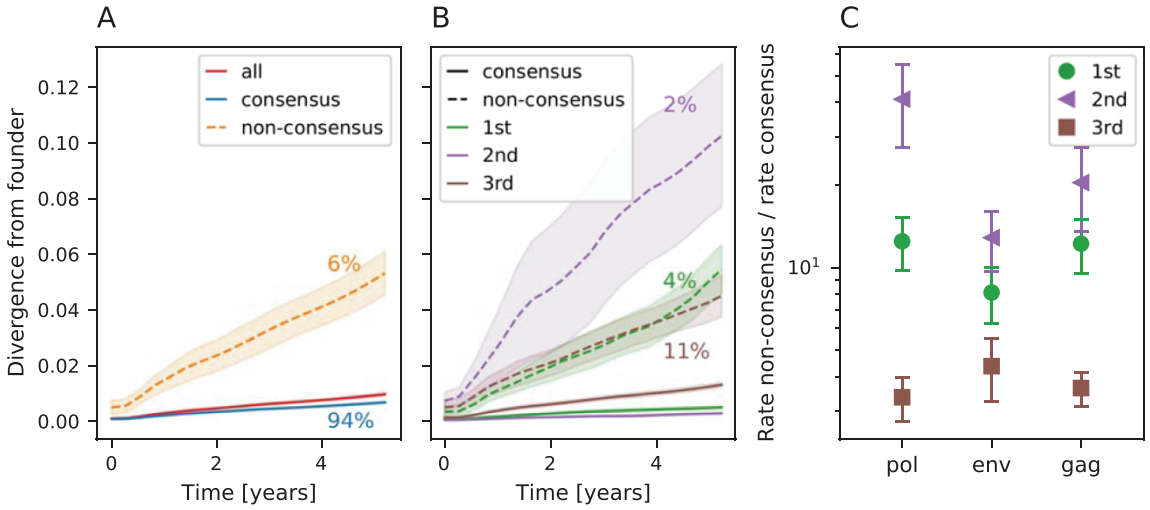


Figure 2. Divergence from founder sequence over time in the *pol* gene. (A) Divergence from founder overall and split for sites initially in consensus and non-consensus states. The reference used to define consensus and non-consensus sites is the HIV-1 group M consensus. Colored percentages are the fraction of sites corresponding to the related curve. Non-consensus sites represent only 6 per cent of the gene but diverge faster over time. Overall, 87 per cent of this divergence are due to reversions, while only 13 per cent are mutations toward another non-consensus nucleotide. (B) The data set from Panel A further split among the first, second, and third codon positions. The difference in evolution speed is greatest for nucleotides in the second position. (C) Ratio of non-consensus to consensus evolution rates computed from the curves in Panel B (Supplementary Figs. S3 and S4 for *env* and *gag*). The ratio is highest for second positions (triangles), where mutations can not be synonymous, followed by first and third positions.

its frequency at each time point t is used to compute $d(t)$. Details about the computation of the founder sequence can be found in Section 4.4. The Boolean δ_i^{ref} is such that $\delta_i^{ref} = 1$ if the founder nucleotide at position i is the same as in the reference sequence and $\delta_i^{ref} = 0$ otherwise. The first term $\delta_i^{ref} \cdot (1 - f_i(t))$ in Equation (1) accounts for the change away from the founder at positions where the founder sequence equals the reference sequence. The term $(1 - \delta_i^{ref}) \cdot f_i(t)$ accounts for the change at positions where the founder sequence differs from the reference sequence. In most cases, the founder nucleotide is replaced by the reference nucleotide and the population is getting more similar to the reference, and mutations to other states are ignored in this calculation (see below). When measuring the divergence relative to the founder sequence, Equation (1) simplifies to:

$$d(t) = 1 - \frac{1}{N}\sum_{i}^{N} f_i(t). \quad (2)$$

In all cases, the quantity $d(t)$ measures the Hamming distance to the reference sequence expected for a randomly chosen sequence from the viral population of a sample. We then averaged the divergence trajectories of different patients and estimated uncertainty by bootstrapping groups of samples from the same patient with replacement. In analogy to the BH analysis, we use the root of the HIV-1 group M tree and subtype consensuses as reference sequences, supplemented by the founder sequence of each patient. Results are shown in Fig. 1D over a period of 5.5 years as the follow-up of most patients stopped after this duration. The filled areas represent one standard deviation of the bootstrap replicates. For more details about the methodology, see Sections 4.2 and 4.3.

Figure 1D shows that the divergence increases the fastest relative to the founder sequence at approximately $(16.1 \pm 1.1) \cdot 10^{-4}$ mutations per site per year. This rate is significantly and substantially higher than the rate at which RTT distance increases on the pandemic scale in line with previous observations that WH rate estimates tend to be higher (Alizon and Fraser 2013). Hamming distances to the subtype consensus or the root of the HIV-1 (M) tree increase significantly more slowly. In fact, these rates are compatible with their corresponding estimates at the pandemic scale (compare Panels C and D).

A 'store-and-retrieve' mechanism to explain the discrepancy between rate estimates should not only result in differences between BH and WH rate estimates (the rates at which the RTT distance and the distance to the founder sequence increase), but also for the rates at which Hamming distances to HIV-1 root or subtype consensuses increase. Since divergence to reference sequences decades in the past is increasing at compatible rates within and between hosts, these analyses suggest similar modes of divergence accumulation and do not support 'store-and-retrieve' as a primary mechanism to explain the discrepancy in rate estimates. In contrast, rate variation or rapid reversion is not expected to affect Hamming distance dynamics to fixed reference sequences like the HIV-1 (M) root. RTT distance estimates, however, are expected to be biased downward since rapid back-and-forth mutations are unaccounted for by the substitution models and do not contribute to the RTT distance. The observations in Fig. 1, and analogous results for *env* and *gag* regions shown in Supplementary Figs. S1 and S2, are thus compatible with saturation effects not captured by substitution models. We will now investigate WH dynamics of polymorphisms to show that rapid reversion to consensus states is a major contributor to this saturation.

2.2 Non-consensus sites diverge faster

Next, we explored the evolution toward and away from consensus within hosts in Fig. 2. Panel A shows the WH divergence separately at sites where the founder sequence agrees with the HIV-1 group M consensus and where it differs from it. Filled areas show the standard deviation of the bootstrap estimate. The divergence at sites where the founder sequence differs from the global consensus increases approximately seven times faster than in the rest of the sequence. A mutation at a site that initially differs from consensus could either be a reversion to consensus or a mutation to one of the two remaining nucleotides. We found that 87 and 85 per cent of mutations at these sites are reversion toward consensus for *pol* and *gag*, while this figure is 76 per cent for *env*. Mutations to a third state are thus a minor contribution. The sevenfold increased rate at 6 per cent of the sites that are initially non-consensus (in *pol*) implies that about one in three mutations bring the sequence closer to the HIV-1 root sequence (the number of reversion mutations divided by the total number of mutations: $\frac{7 \cdot 0.06 \cdot \mu}{(0.94 + 7 \cdot 0.06)\mu} \approx \frac{1}{3}$ where μ is the observed evolution rate.). This strong tendency to revert can explain the difference in evolutionary rates observed on WH and BH scales and is consistent with the threefold difference in slope between the divergence relative to the founder or HIV-1 group M root shown in Fig. 1D.

This accelerated evolution could be due to (1) reversion to an ancestral state to increase fitness or (2) reduced purifying selection at sites with high levels of diversity in global HIV-1 population. In order to differentiate between these possibilities, Fig. 2B shows the divergence by codon position. The degree to which divergence is accelerated differs among the first, second, and third positions in a codon. In particular, sites in the second position diverge the fastest when in a non-consensus state, while they diverge the slowest in a consensus state. This is consistent with the fact that second positions tend to be most conserved as only 2 per cent of such sites differ from the consensus sequence in *pol*.

Figure 2C quantifies the ratio of divergence rates at sites initially in a consensus or non-consensus state for *pol*, *env*, and *gag*. Details on the computation of these rates can be found in Section 4.3. In all cases, evolution rates of non-consensus sites are higher than consensus ones. The difference is greatest for second codon position sites, followed by first codon position sites (see Supplementary Figs. S3 and S4 for divergence plots for other genes). Mutations at second codon position sites are always non-synonymous and often cause drastic amino acid changes, while mutations at third codon position sites are often synonymous and generally less impactful. Mutations at first position sites can be both synonymous and non-synonymous. The observation that divergence is fastest at non-consensus but otherwise strongly conserved sites suggests that reversion mutations are selected to increase fitness and are not the result of reduced purifying selection at sites of high diversity. These results are consistent with previous observations showing that conserved sites tend to revert more quickly (Zanini et al. 2015) and the notion that selection for reversion is probably driven by the fitness costs of mutations that enabled immune escape in a previous host (Leslie et al. 2004). Such rapid reversion is an example of adaptation within hosts, but the combined escape-reversion dynamics on timescales spanning several transmission events looks like purifying selection at conserved sites.

2.3 Reversion mutations are positively selected

If a lot of reversions are driven by selection, as the codon-position-specific analysis above suggests, effects of selection should be detectable in the dynamics of intra-host single nucleotide variants (iSNVs). Specifically, we expect to see a tendency of reversion mutations to increase in frequency and fix. We analyzed the frequency trajectories of iSNVs to look for such features. Similar to the previous analysis, we separate all trajectories into reversion and non-reversion groups and compare their evolutionary dynamics in Fig. 3. We select trajectories with at least one data point in a frequency interval $[f_{min}, f_{max}]$ for each group. We offset these trajectories in time so that $t = 0$ corresponds to their first data point seen in the frequency interval and compute the mean frequency of the trajectory group over time. The small minority of trajectories where both the initial nucleotide and the target nucleotide differ from the consensus sequences are classified as 'non-reversions' in

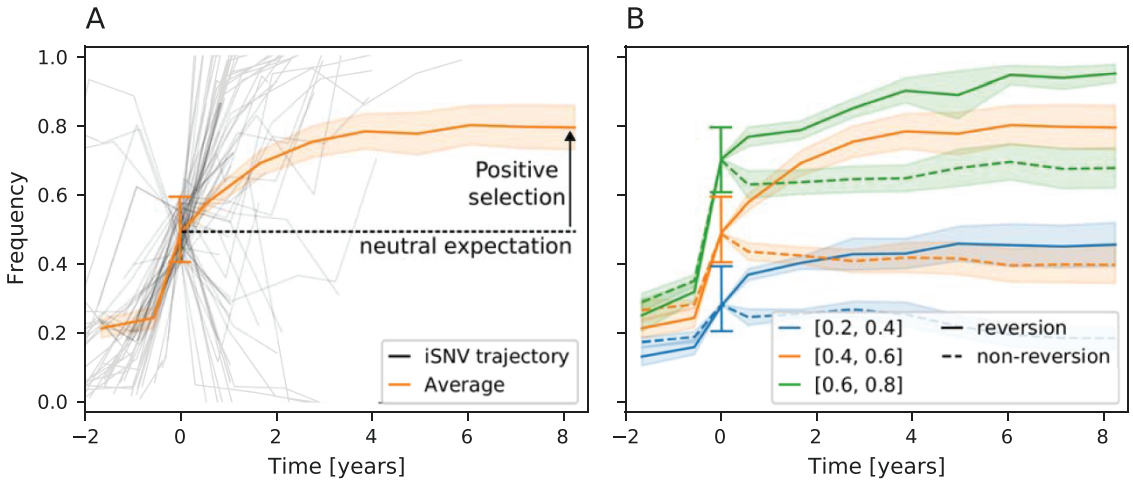


Figure 3. Positive selection on reversion mutations. (A) Frequency of reversion mutations seen between 0.4 and 0.6 frequency at one time point (offset to be $t = 0$) and their average over time. (B) Mean frequency over time for reversion (full lines) and non-reversion mutations (dashed lines) for different frequency windows (colors). Reversion trajectories are strongly selected for as their mean frequency increases over time. Non-reversion trajectories evolve close to the neutral expectation. The reference sequence used to define reversion mutations is the HIV-1 group M consensus. The solid orange line is the same in both panels.

this analysis. More details about the definition of trajectories and the methodology are in Supplementary Fig. S5 and Sections 4.5 and 4.6. We use the HIV-1 group M consensus sequence as a reference to define reversion mutations, but results are qualitatively similar when using subtype consensus or root sequence as a reference.

Figure 3A shows individual trajectories shifted to pass through the frequency interval [0.4, 0.6] at $t = 0$ along with their average. The mean frequencies for different initial conditions and groups of trajectories are shown in Fig. 3B. Since we condition the set of trajectories to start as minor variants and pass through a frequency interval at $t = 0$, we expect that trajectories tend to rise for $t < 0$, as is indeed observed. The dynamics at $t > 0$, i.e. after the time of conditioning, are informative about the selection of the iSNV. We do not expect any consistent trend to rise or fall in frequency for neutral mutations, hence their average frequency should be constant for $t > 0$. Contrary to that, we show in Fig. 3B that the frequency of reversion mutations increases on average over time. This suggests that these reversion mutations are beneficial on average and fix preferentially in the population, with probability given by the end point of the curves for each group of trajectories. This finding is consistent with the notion that the HIV-1 consensus sequence approximates a fitness optimum of HIV-1 (Zanini et al. 2017). On the other hand, non-reversion curves are flat or slightly decreasing for $t > 0$, suggesting that such mutations tend to be slightly selected against or are neutral—at least those that reach high frequency in the first place.

We note that the selection for reversion mutations is strongest for the *gag* region, see Fig. S6 for details. When splitting trajectories into synonymous and non-synonymous changes (irrespective of reversion/non-reversion), we observe that synonymous mutations tend to decrease in frequency for $t > 0$, while on average non-synonymous mutations increase, see Fig. S7. This suggests that high-frequency non-synonymous mutations tend to be beneficial, while synonymous mutations are slightly deleterious, consistent with earlier results (Zanini and Neher 2013). Common synonymous reversions, on the other hand, tend to further increase in frequency and fix preferentially, see Fig. S8.

2.4 Reversions can explain the rate mismatch

Over longer timescales, the rapid reversions we observe within hosts will lead to undetected substitutions along branches of the phylogeny whenever a mutation and its corresponding reversion happen on the same branch. When such reversion dynamics are not captured by the substitution models, the evolutionary rate inferred by phylogenetic methods will be too low (Halpern and Bruno 1998; Hilton and Bloom 2018; Puller et al. 2020). Here we explore how much of the discrepancy between evolutionary rates estimates at the WH and BH scales can be attributed to rapid reversions not being properly captured by substitution models.

We quantify the impact of reversions on the BH evolution rate using an evolutionary model that accounts for the reversion bias we observed within hosts. We use the TreeTime library (Sagulenko et al. 2018) to define a site-specific general time-reversible (GTR) model (Puller et al. 2020). We parameterize the mutation rate from nucleotide j to i at position α as:

$$Q_{ij}^{\alpha} = \mu p_i^{\alpha} W_{ij}, \tag{3}$$

where μ is the mean mutation rate per site per year, p_i^{α} describes the equilibrium probability of finding nucleotide i at site α, and W_{ij} accounts for the overall variation in rate between different nucleotide pairs i and j independent of position (i.e. the differences between transitions and transversions). We use $\mu = 16.1 \cdot 10^{-4}$, the overall WH evolution rate observed in Fig. 1D. In this model, the bias for reversion is introduced via the equilibrium frequencies p_i^{α}. These depend on the genome position α, enabling us to skew the frequencies toward the consensus nucleotide at each position. Contrary to common evolutionary models that include rate variation between sites, we keep the evolutionary rate constant across positions and vary p_i^{α} instead. However, our results show little change if a gamma-distributed rate variation is incorporated, especially when the shape parameter is greater than 2. We choose p_i^{α} such that the model reproduces the WH rates of reversions and evolution away from consensus. Specifically, we use

$$p_i^{\alpha} \quad amp; = \frac{\mu^{\alpha}}{\mu^{\alpha} + \mu_{+}^{\alpha}} \quad amp; \text{if} \qquad i = \text{consensus at } \alpha, \tag{4}$$

$$p_i^\alpha \quad amp;= \frac{\mu_+^\alpha}{\mu_-^\alpha + \mu_+^\alpha} \cdot r_i^\alpha \quad amp;\text{if} \qquad i \neq \text{consensus at } \alpha, \tag{5}$$

where μ_+^α and μ_-^α are the consensus and non-consensus divergence rates, respectively, computed from WH data shown in Fig. 2B. These rates reproduce the equilibrium frequencies in a model with two states (consensus and non-consensus). These rates are codon position-specific, meaning for every α that is a first codon position $\mu_+^\alpha = \mu_+^{1st}$ and $\mu_-^\alpha = \mu_-^{1st}$ and analogously for the second and third codon positions. The parameter r_i^α is used to specify the relative proportions of the three non-consensus nucleotides. It is chosen so that 85 per cent of the non-consensus nucleotides are the transitions from the consensus, while the two transversions contribute 7.5 per cent each. These values were inferred from the BH alignment and are consistent with the WH observations, see Section 2.2. Otherwise, this GTR model is purely informed by WH rates.

We then used this model to simulate evolution along an HIV-1 phylogeny and generate a multiple sequence alignment (MSA) using TreeTime and the inferred HIV-1 root sequence (as used in Fig. 1). We then inferred a tree from the MSA generated using IQ-TREE, as we did for the real data.

Figure 4 compares the diversity of original and generated MSAs and the length of the inferred trees to quantify the impact of reversions on phylogenetic inference. A model that does not account for reversions, i.e. where $p_i = 0.25$ for $i \in$ A,C,G,T for all sites, was included for comparison and is referred to as the naive GTR model. Figs. 4A and 4B show a comparison of the real and generated MSA characteristics. The MSA generated using our WH-informed GTR model (green) has a similar nucleotide content and distance to the root as the real BH data (blue). On the contrary, the naive GTR model that does not take reversions into account (orange) results in a more diverse MSA and overall nucleotide content that is less similar to the BH data.

Figure 4C shows that the evolutionary rate estimated from the RTT regression of the tree reconstructed from the MSA simulated using the naive GTR model is, as expected, very close to the WH evolution rate of $\mu = 16.1 \cdot 10^{-4}$ mutation per site per year we input into the model. Our custom GTR model, which uses the same μ but accounts for reversions, results in a RTT regression with a slope of $11.9 \cdot 10^{-4}$, substantially lower than the true rate and within 10 per cent of the rate estimate from the RTT regression for the original phylogenetic tree. This suggests that a substitution model parameterized by rates and reversion propensity of WH evolution can largely reconcile the discrepancy of rate estimates at different scales, even if it does not include rate variation between different sites.

We find qualitatively similar results for the *gag* (see Supplementary Fig. S10). In the case of *env*, the tree reconstructed from the data generated and subsequent analysis is unreliable due to excessive saturation in the model (see Supplementary Fig. S9).

3. Discussion

Evolutionary rate estimates depend strongly on the timescale over which they are measured (Ho et al. 2005; Aiewsakun and Katzourakis 2016). Here, we explored this effect on the scale of the HIV-1 pandemic, individual subtypes, and within hosts. We showed how observations on short timescales give rise to patterns on longer scales. Differences between rate estimates WH scale and on the pandemic scale can, to a substantial degree, be explained by a strong tendency to revert deleterious mutations to their preferred state. These unpreferred states are probably the result of escape from immune selection in a previous host, which gradually revert as the host-specific selection pressure is relaxed in future hosts. Microscopically, we thus observe evolutionary dynamics driven by the adaptation to a changing environment: both changes, escape and reversion, are beneficial in their respective environments. These transiently beneficial escape mutations are generally deleterious on longer timescales, such that the aggregate effect of this dynamic looks like slowly acting purifying selection (Wertheim and Kosakovsky Pond 2011).

Substitution models commonly used to reconstruct phylogenies and infer evolutionary rates do not account for rapid reversions, which would require site-specific preferences for different states (Halpern and Bruno 1998; Hilton and Bloom 2018;

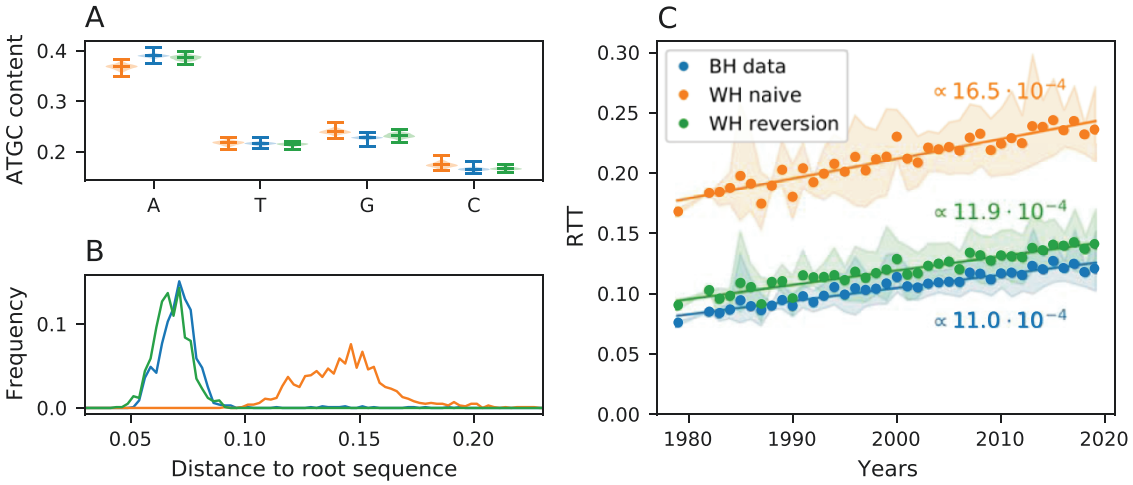


Figure 4. Substitution models that account for reversions can largely explain the rate mismatch. This figure shows the sequence diversity and RTT distances for simulated data generated with a substitution model that accounts for reversion, parameterized by WH data (WH reversion) and a model that does not account for reversion (WH naive) for the *pol* region. (A) Violin plot of the nucleotide content for the BH data and the MSAs generated. The naive model tends to equilibrate the nucleotide composition. (B) Histogram of Hamming distances to the root sequence. The reversion-informed model agrees well with the BH observations, while the naive one generates sequences very far from the root. (C) While the RTT distance estimated from data generated by the naive model is consistent with the evolution rate we used in the model (true value $16.1 \cdot 10^{-4}$ per site and year), the data generated using the model with reversion results in much lower estimates, similar to the rate estimated from BH data. The results for *env* and *gag* are shown in Supplementary Figs. S9 and S10.

Puller et al. 2020). We explored the effect of reversions on phylogenetic inference and rate estimates of HIV-1 by defining a simple site-specific model parameterized by reversion and non-reversion rate estimates from longitudinal data within hosts. Phylogenetic inference from data simulated using this model revealed that reversions during chronic infection can explain approximately 90 per cent of the apparent slowdown of evolution for the *pol* gene of HIV-1. A similar selection for reversion mutations has also been observed during acute infection (Boutwell et al. 2010; Leslie et al. 2004) or the transmission bottleneck (Carlson et al. 2014). Such preferential transmission of consensus-like variants could amplify the overall effect of incomplete reversions during chronic infection. Together, these results suggest that, among the hypotheses proposed to explain the difference in rates (Lythgoe and Fraser 2012; Redd et al. 2012; Zanini et al. 2015; Leslie et al. 2004), 'adapt and revert' is the main mechanism.

The strongest effects of unaccounted reversions in phylogenetic inference are expected on long branches in the phylogenetic tree, where mutations are masked by their corresponding reversions (Hilton and Bloom 2018; Puller et al. 2020). The well-known phenomenon of long-branch attraction can, in these cases, already set in for branches that are nominally quite short. A strong tendency to revert can lead to sites that are completely saturated, yet almost always are in the same state—an effect not captured well by rate variation.

Rapid reversions are probably essential to conserve global fitness by purging costly immune escape mutations acquired in individuals earlier in the transmission chain (Carlson et al. 2014; Zanini et al. 2017). In addition to reversion, fitness costs of escape mutations can, of course, also be mitigated by compensatory mutations (Crawford et al. 2007; Carlson and Brumme 2008). Although such compensatory mutations presumably slow down many reversions, we still observe a marked difference in iSNV frequency dynamics toward vs. away from consensus. In addition, compensatory evolution can change the preferred sequence to a new local fitness maximum to which mutations revert, adding an additional slow timescale to the evolutionary process. We expect the preferred sequence to slowly drift on timescales much longer than the typical serial interval along transmission chains. This effect has been observed in deep mutational scanning experiments in influenza viruses (Hilton and Bloom 2018; Doud et al. 2015). Such effects are also consistent with the 'Prisoner of War' model by Ghafari *et al.* (2021), where a slowly changing fitness landscape (through host switches, host adaptation, or compensatory evolution) gives rise to apparent rates of evolution that decrease with the timescale of observation over many orders of magnitude.

The star-like diversification of HIV-1 into multiple subtypes gives a clear notion of a consensus sequence that can be used to approximate a putative fitness peak toward which reversions occur. In other viruses, for example, influenza A viruses, the ladder-like or otherwise structured phylogenies do not allow a straightforward definition of a consensus sequence. Nevertheless, it is possible that adaptation to a changing immunity landscape and reversions contribute with a similar magnitude to sequence turnover.

4. Materials and Methods

4.1 Data set and filtering steps

4.1.1 *Between-host data sets*

Our BH data sets come from the LANL HIV databases. All HIV-1 group M sequences with exact sampling date were downloaded for the *pol*, *env*, and *gag* regions. Subtype O and N sequences were filtered out. Sequences with ambiguous nucleotides and sequences labeled as 'problematic' on LANL website were removed. Only one sequence was kept per patient. The data sets were downloaded on 14 July 2021. This gave us a total of 6,649 sequences for *pol*, 15,034 for *env*, and 8,948 for *gag*.

Regarding each genomic region, we subsampled the data set to have 1,000 sequences in each case, with the same number of sequences for each year where sequences were available (except for early years where fewer sequences were available). For each region, Subtype B represents approximately 40 per cent of all sequences, Subtype C approximately 15 per cent, and the rest encompass the other subtypes or unlabeled subtypes. Subtype B sequences are more common in early years while Subtype C sequences represent a larger proportion in recent years. We then performed an MSA, including the reference HIV-1 HXB2 sequence, using Multiple Alignment using Fast Fourier Transform (Katoh and Standley 2013) and the Nextstrain framework (Huddleston et al. 2021). Insertions relative to the reference HXB2 sequence were removed. We removed all positions of the alignment where more than 10% of sequences have a gap as the alignment can be unreliable in such positions. The alignment for the *pol*, *env*, and *gag* regions are the data sets used for our BH analyses. See the section Code and data availability for access to the data sets.

4.1.2 *Within-host data sets*

Our WH analysis leverages the time resolution of the HIVEVO data set (Zanini et al. 2015). This data set is freely available with tools made available to facilitate the analysis. We use these tools to obtain a three-dimensional matrix of nucleotide frequencies for each patient. The three axes of these tables are the HIV-1 genome position, the nucleotide, and the time since infection of the sample. Each entry in these matrices gives the frequency of a given nucleotide at a given position on the genome at this time point, relative to the total intra-patient HIV-1 population. These matrices form our WH data set. We excluded patients *p7* and *p10* from our analysis as their samples were very uneven in time or because there was evidence of multiple founder sequences.

The estimates of nucleotide frequencies are unbiased in the $[0.1, 0.9]$ range, while coverage and depth are globally sufficient (Zanini et al., 2016). We applied several filtering steps prior to analysis to avoid biases in our results. We masked data points with sequencing coverage inferior to 100 and/or where the depth was low. We also removed genome positions that were not mapped to the consensus sequence and/or seen to be too often gapped in the MSA of BH sequences. The alignment and mapping of such sites can be unreliable; thus, we removed them from our analysis. This filtering procedure is mainly relevant for the *env* gene, which is the region with the most noise.

4.2 Distance and divergence over time

The first result section gives an overview of the method used to compute the distance and divergence over time in Fig. 1 and Supplementary Fig. S1 and S2. Additional details are given below.

Hamming distances were computed by counting the number of sites that do not match the reference sequence for each sequence in the data set. We then divide this number by the length of the sequence to obtain the relative distance to the reference. Hamming distances were computed using three reference sequences. The first is the root sequence of the tree. The tree was inferred using the IQ-TREE GTR+F+R10 model (Minh et al. 2020), while the

root sequence was computed using TreeTime ancestral reconstruction on this tree (Sagulenko et al. 2018). We chose to use the root sequence instead of the consensus sequence of the alignment in Figs. 1 and 4 to avoid biases due to over-representation of Subgroup B and C sequences. The second and third reference sequences are Subgroup B and C consensus sequences. See Section 4.4 for details on the computation of consensus and founder sequences. To compute the Hamming distances to the subtype consensus, we averaged the distances computed for Subtype B and C sequences relative to their consensus. The average was then weighted by the relative number of each subtype sequence in each year.

The RTT distances shown in Figs. 1 and 4 and Supplementary Figs. S1, S2, S9, and S10 are computed directly from the tree generated via IQ-TREE. Such distances were computed for every leaf of the tree (i.e. every sequence in our data set) and then averaged for sequences sampled in the same year for visualization. Taking into account the phylogenetic information allows the detection of some mutations that occur and then revert along the tree. Consequently, the estimates of the RTT distance are higher than the Hamming distance ones.

4.3 Evolutionary rates

The evolutionary rates in Figs. 1 and 2C and Supplementary Figs. S1 and S2 are the slopes of linear fits of the data. For the BH plots (Fig. 1C and Supplementary Figs. S1A and S2A), the fit was done on the data from 1979 to 2022. For the WH plots (Figs. 1D and 2C and Supplementary Figs. S1B and S2B), we estimated a linear fit from 200 to 2,000 days in the infection. We removed the first 200 days from the fit as for most patients the first sample we have is in the 0–200 days window. This causes the small flat part of the founder curves near $t = 0$, which could bias our evolution rate estimates. Consequently, we decided to only use data starting from 200 days into the infection for the fit, which is more than enough to get an accurate estimate of the slope. For the WH rate estimates, we estimate the error by bootstrapping patients.

Estimating confidence intervals for evolutionary rates at the level of the pandemic is challenging because of the phylogenetic relationship and shared ancestry of the sequences. Instead of using probabilistic phylogenetic models, which suffer from residual recombination and model inadequacies, we opted for phylogenetic boot-strapping procedure for the BH rate estimates. Specifically, we cut all branches of the time-scaled phylogenetic trees at the year 1980 and thereby obtain a collection of subtrees. Sequences in the same subtree are correlated, but they are not correlated with sequences on another subtree (as evolution happens on different branches of the original tree). We performed bootstrapping to estimate distances and rates by sampling with replacement from sequences in these subclades. The errors provided for the rate estimates in Fig. 1 and Supplementary Figs. S1 and S2 are computed from these bootstrap estimates. The HIV-1 pandemic has undergone a large radiation in 1960s and 1970s, which makes such bootstrap estimates possible.

4.4 Consensus and founder sequence

Consensus sequences were computed from our BH data sets. We computed three consensus sequences for each region studied. The first is the HIV-1 group M global consensus, which is the majority nucleotide of the alignment at each position. The second and third are the Subtype B and C consensus sequences. These were computed in the same way, using a subset of the alignment that contains only the sequences of the subtype in question.

The founder sequence is an approximation of the sequence of the virus at the time of infection in a patient. They are computed from our WH data set for each patient separately. The founder sequence is the majority nucleotide in each position from the first sample of each patient. In this sense, it is the consensus sequence obtained from the first sample of each patient. For most patients in our data set, the first sample is taken at approximately 90 days after infection and no data are available on the early phase of infection. Consequently, the founder sequence computed is an approximation of the original virus.

4.5 Trajectory extraction and metadata

A trajectory is a sequence of nucleotide frequencies and associated time. Each trajectory corresponds to one genome position and one nucleotide only. We extracted trajectories from our WH data set according to several criteria. Firstly, every trajectory must be extinct before the first point, i.e. we consider only new mutations. This is to avoid biases that could be due to immune interaction existing already. Secondly, frequencies must be between 0.01 and 0.99 at all time points. The trajectory is considered extinct if it is below 0.01 and fixed if above 0.99. Lastly, we apply a mask to data points according to what is shown in Section 4.1. Trajectories that have their first and/or last points masked are removed from the analysis.

Every trajectory extracted according to the criteria above is coupled with its metadata. This contains all the relevant information, such as whether the mutation is a reversion or not and whether it fixed or was lost. This information is used to create subgroups of trajectories. From these subgroups, one can study the impact of a trait associated with a mutation for WH evolution, as shown in Fig. 3 and Supplementary Fig. S5 for reversion and non-reversion trajectories.

4.6 Mean frequency in time

While looking at divergence values informs us about the global evolution of the WH population, it cannot tell us whether the mutations we see on non-consensus sites are actually reversions to the consensus state or simply mutations to another nucleotide. This motivated us to look directly at the evolution of new mutations independently by observing their frequency trajectories in time. Trajectories were extracted and filtered according to Sections 4.1 and 4.5. Despite these filtering steps, our data are inherently biased toward small and/or low-frequency trajectories which are more common. In order to alleviate this bias, we compare reversion and non-reversion trajectories in the same manner. Accordingly, the resulting signal can be attributed to the effect of being a reversion (or not).

Due to the limited number of trajectories available and the often lack of information about trajectory fixation, for example, because it is still active at the last sample, the probability of fixation plots were not adequate for our analysis. We, thus, decided to pay attention to the evolution of the mean frequency in time for groups of trajectories. Trajectories were grouped in frequency bins, as described in the main text, to avoid bias toward positively or negatively selected trajectories. Supplementary Fig. S5 illustrates how this was done. Sometimes a trajectory's first pass through the frequency window is missed and only caught on the second pass, which results in a few trajectories that enter the frequency window from above. This happens when the frequency of a mutation changes drastically from one sample to the next, i.e. the reported

frequency jumps directly from below to above the window. Nevertheless, these are 'new' mutations as they were not seen in the first sample of the patient. We kept these trajectories to avoid potential bias, but including or excluding them does not have a big impact on the final results.

We then created time bins of 400 days from 600 days before up to 3,000 days after a trajectory is seen in a frequency window. We compute the average frequency of all trajectories belonging to the same group in each time bin. A trajectory contributes its current frequency if a data point is available at this time and does not contribute if no data are available in that time bin. Trajectories that fixed in the population contribute with a frequency of $f = 1$ to time bins subsequent to their fixation. Similarly, lost trajectories contribute $f = 0$ to time bins subsequent to their disappearance in the viral population. Trajectories that are still active after their last data point (because the study stopped before it could fix or be lost) contribute the frequency of their last data point to the following time bins.

Code and data availability

The code and data used for the analysis can be found at https://github.com/neherlab/HIVEVO_reversion. Due to issues with the data sets' size, only intermediate BH and WH data files in a compressed format are found in the github folder. A link to the full data set is available there. Scripts are present to reproduce the results shown in this paper.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Acknowledgements

We gratefully acknowledge the stimulating discussions with Pierre Barrat-Charlaix and Marco Molari. University of Basel; Swiss National Science Foundation (310030_188547); PhD Fellowship Program of the Biozentrum (to V.D.).

References

Aiewsakun, P. and K., Aris (2016) 'Time-Dependent Rate Phenomenon in Viruses', *Journal of Virology* VI, 90: 7184–7195.

Alizon, S. and C., Fraser (2013) 'Within-host and Between-Host Evolutionary Rates Across the HIV-1 Genome', *Retrovirology*, 10: 1–10.

Boutwell Christian, L. et al. (2010) 'Viral Evolution and Escape During Acute HIV-1 Infection', *The Journal of Infectious diseases*, 202: S309.

Brian Foley Cristian Apetrei et al Thomas Leitner. (2018) 'HIV Sequence Database: 2018 Compendium'.

Carlson Jonathan, M. and L., Brumme Zabrina (2008) 'HIV Evolution in Response to HLA-Restricted CTL Selection Pressures: A Population-Based Perspective'. *Microbes and Infection* IV', 10: 455–461.

Carlson Jonathan, M. et al. (2014) 'HIV Transmission. Selection Bias at the Heterosexual HIV-1 Transmission Bottleneck', *Science* VII, 345: 1254031.

Coffin, J. and S., Ronald (2013) 'HIV Pathogenesis: Dynamics and Genetics of Viral Populations and Infected Cells', *Cold Spring Harbor Perspectives in Medicine* I, 3: a012526.

Crawford, H. et al. (2007) 'Compensatory Mutation Partially Restores Fitness and Delays Reversion of Escape Mutation within the Immunodominant HLA-B* 5703-Restricted Gag Epitope in Chronic Human Immunodeficiency Virus type 1 Infection', *Journal of virology*, 81: 8346–8351.

Doud Michael, B., A., Orr and D., Bloom Jesse (2015) 'Site-Specific Amino Acid Preferences Are Mostly Conserved in Two Closely Related Protein Homologs', *Molecular Biology and Evolution* XI, 11: 2944–2960.

Felsenstein, J. (2004) '*Inferring phylogenies*'.

Foley, B. et al. (2013) 'Theoretical Biology and Biophysics Group', *Los Alamos: Los Alamos National Laboratory*, 13.

Freed Eric, O. (2001) 'HIV-1 Replication', *Somatic Cell and Molecular genetics*, 26: 13–33.

Ghafari, M. et al. (2021) 'Prisoner of War Dynamics Explains the Time-Dependent Pattern of Substitution Rates in Viruses', *bioRxiv*, II: 2021.02.09.430479.

Gilbert, C. and F., Cédric (2010) 'Genomic Fossils Calibrate the Long-Term Evolution of Hepadnaviruses', *PLoS biology* IX, 8: e1000495.

Halpern, A. L. and W. J., Bruno (1998) 'Evolutionary Distances for Protein-Coding Sequences: Modeling Site-Specific Residue Frequencies', *Molecular Biology and Evolution*, 15: 910–917.

Hanada, K., S., Yoshiyuki and G., Takashi (2004) 'A Large Variation in the Rates of Synonymous Substitution for RNA Viruses and its Relationship to a Diversity of Viral Infection and Transmission Modes', *Molecular Biology and Evolution* VI, 6: 1074–1080.

Herbeck Joshua, T. et al. (2006) 'Human Immunodeficiency Virus Type 1 env Evolves Toward Ancestral States upon Transmission to a New Host', *Journal of Virology* II, 80: 1637–1644.

Hilton Sarah, K. and D., Bloom Jesse (2018) 'Modeling Site-Specific Amino-Acid Preferences Deepens Phylogenetic Estimates of Viral Sequence Divergence', *Virus Evolution* VII, 4: 2.

Ho Simon, Y. W. et al. (2005) 'Time Dependency of Molecular Rate Estimates and Systematic Overestimation of Recent Divergence Times', *Molecular Biology and Evolution*, 22: 1561–1568.

Huddleston, J. et al. (2021) 'Augur: A Bioinformatics Toolkit for Phylogenetic Analyses of Human Pathogens', *Journal of Open Source Software* I, 6: 2906.

Illingworth Christopher, J. R. et al. (2020) 'A de Novo Approach to Inferring within-host Fitness Effects During Untreated HIV-1 Infection', *PLoS pathogens*, 16: e1008171.

Jukes Thomas, H., and R., Cantor Charles and others (1969) 'Evolution of Protein Molecules', *Mammalian Protein metabolism*, 3: 21–132.

Kalyaanamoorthy, S. et al. (2017) 'ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates', *Nature methods*, 14: 587–589.

Katoh, K. and M., Standley Daron (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution* IV, 30: 772–780.

Korber, B. et al. (2000) 'Timing the Ancestor of the HIV-1 Pandemic Strains', *science*, 288: 1789–1796.

Leslie, A. J. et al. (2004) 'HIV Evolution: CTL Escape Mutation and Reversion after Transmission', *Nature medicine*, 10: 282–289.

Li, G. et al. (2015) 'An Integrated map of HIV Genome-wide Variation from a Population Perspective', *Retrovirology*, 12: 18.

Lythgoe Katrina, A. and F., Christophe (2012) 'New Insights into the Evolutionary Rate of HIV-1 at the within-host and Epidemiological Levels', *Proceedings of the Royal Society B: Biological Sciences*, 279: 3367–3375.

McCutchan Francine, E. (2006) 'Global Epidemiology of HIV', *Journal of Medical virology*, 78: S7–S12.

Minh Bui, Q. et al. (2020) 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic era', *Molecular Biology and evolution*, 37: 1530–1534.

Puller, V., S., Pavel and A., Neher Richard (2020) 'Efficient Inference, Potential, and Limitations of Site-Specific Substitution Models', *Virus Evolution* VII, 6: 2.

Raghwani, J. et al. (2018) 'Evolution of HIV-1 within Untreated Individuals and at the Population Scale in Uganda', *PLOS Pathogens* VII, 14: 7.

Redd Andrew, D. et al. (2012) 'Previously Transmitted HIV-1 Strains Are Preferentially Selected During Subsequent Sexual Transmissions', *The Journal of Infectious Diseases* XI, 206: 1433–1442.

Sagulenko, P., V., Puller and A., Neher Richard (2018) 'TreeTime: Maximum-Likelihood Phylodynamic Analysis', *Virus Evolution* 01, 4: vex042.

Sharp Paul, M. and H., Hahn Beatrice (2011) 'Origins of HIV and the AIDS pandemic', *Cold Spring Harbor Perspectives in medicine*, 1: a006841.

Soubrier, J. et al. (2012) 'The Influence of Rate Heterogeneity Among Sites on the time Dependence of Molecular Rates', *Molecular Biology and evolution*, 29: 3345–3358.

Tavaré, S. and others (1986) 'Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences', *Lectures on Mathematics in the Life sciences*, 17: 57–86.

Wertheim Joel, O. and L., Kosakovsky Pond Sergei (2011) 'Purifying Selection Can Obscure the Ancient Age of Viral Lineages', *Molecular Biology and Evolution* XII, 28: 3355–3365.

Worobey, M. et al. (2010) 'Island Biogeography Reveals the Deep History of SIV', *Science* IX, 329: 1487–1487.

Yang, Z. (1995) 'A Space-Time Process Model for the Evolution of DNA Sequences', *Genetics*, 139: 993–1005.

Zanini, F. et al. (2016) 'Error Rates, PCR Recombination, and Sampling Depth in HIV-1 Whole Genome Deep Sequencing', *Virus research*, 65: 201–204, December.

Zanini, F. et al. (2015) 'Population Genomics of Intrapatient HIV-1 Evolution', *Elife*, 4: e11282.

Zanini, F. and A., Neher Richard (2013) 'Quantifying Selection Against Synonymous Mutations in HIV-1 env Evolution', *Journal of Virology* XI, 84: 11843–11850.

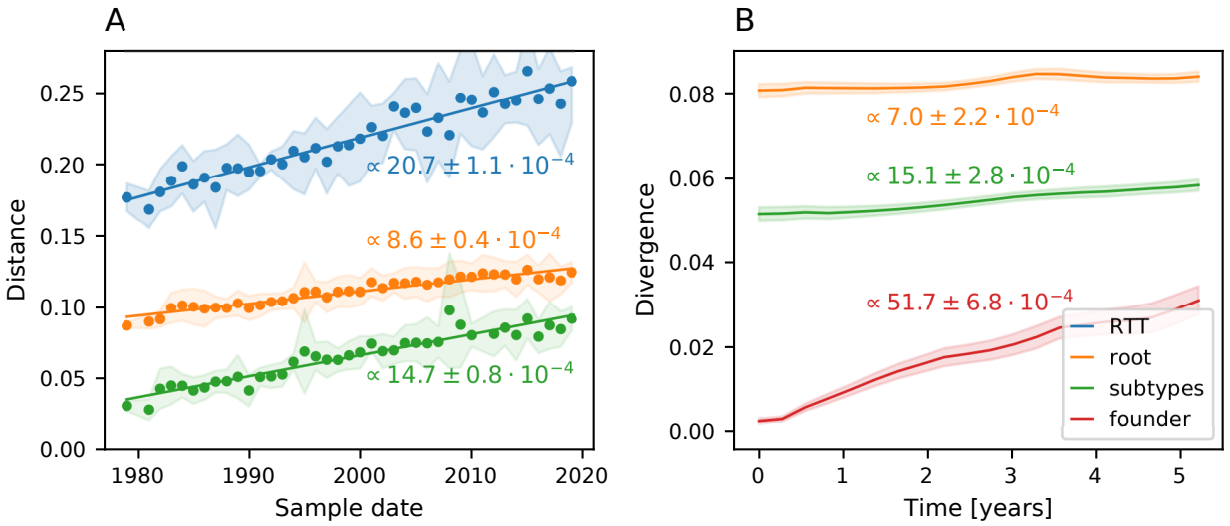
Zanini, F. et al. (2017) 'In vivo Mutation Rates and the Landscape of Fitness Costs of HIV-1', *Virus evolution*, 3: 1.

Supplementary materials: Reversions to consensus are positively selected in HIV-1 and bias substitution rate estimates

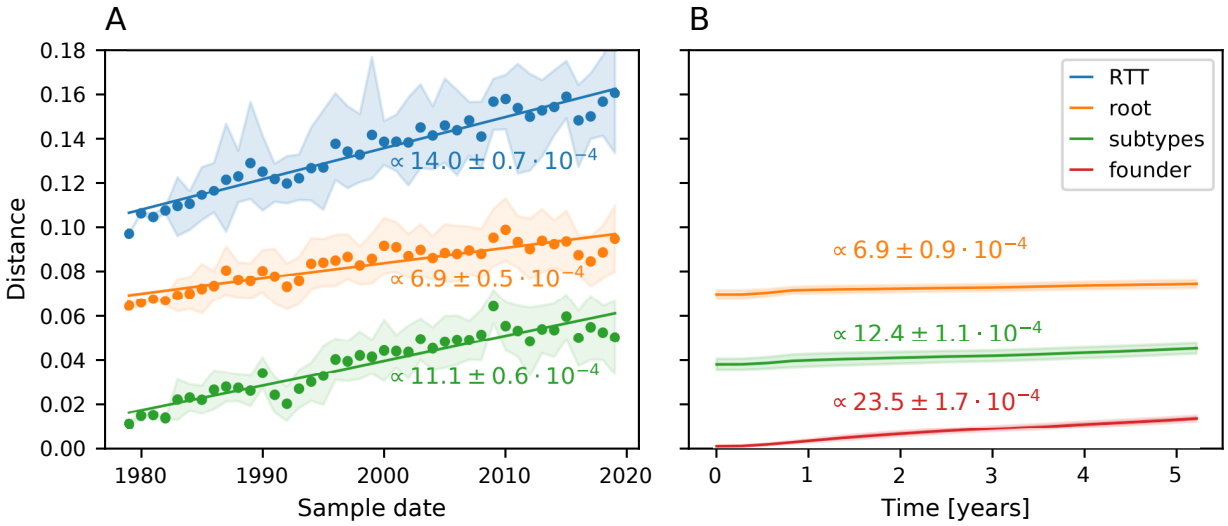
Valentin Druelle[1,2] and Richard A. Neher[1,2]

[1]Biozentrum, University of Basel, Basel, Switzerland
[2]Swiss Institute of Bioinformatics, Basel, Switzerland

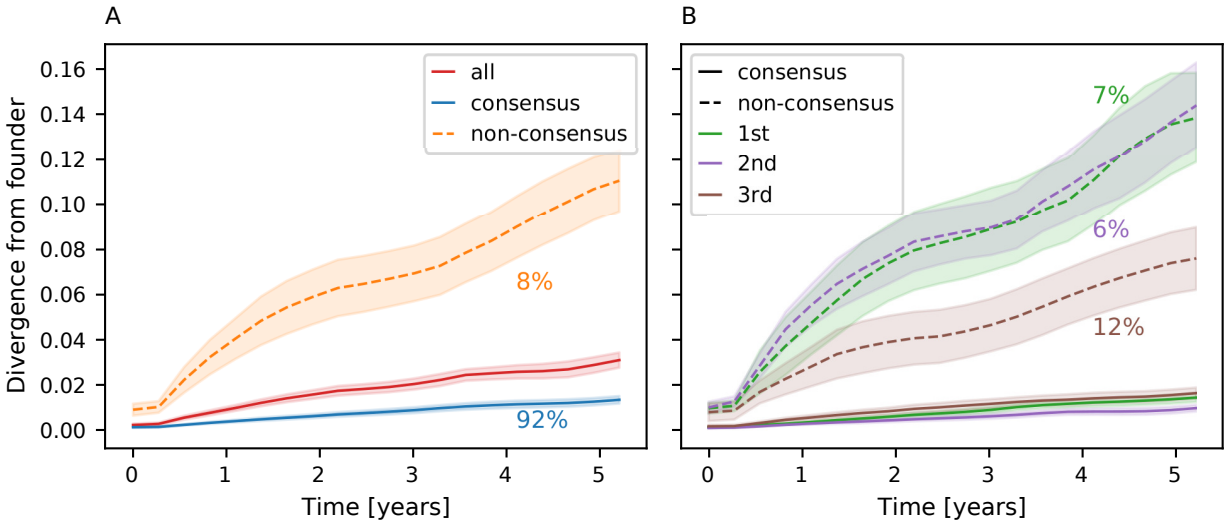
December 7, 2022



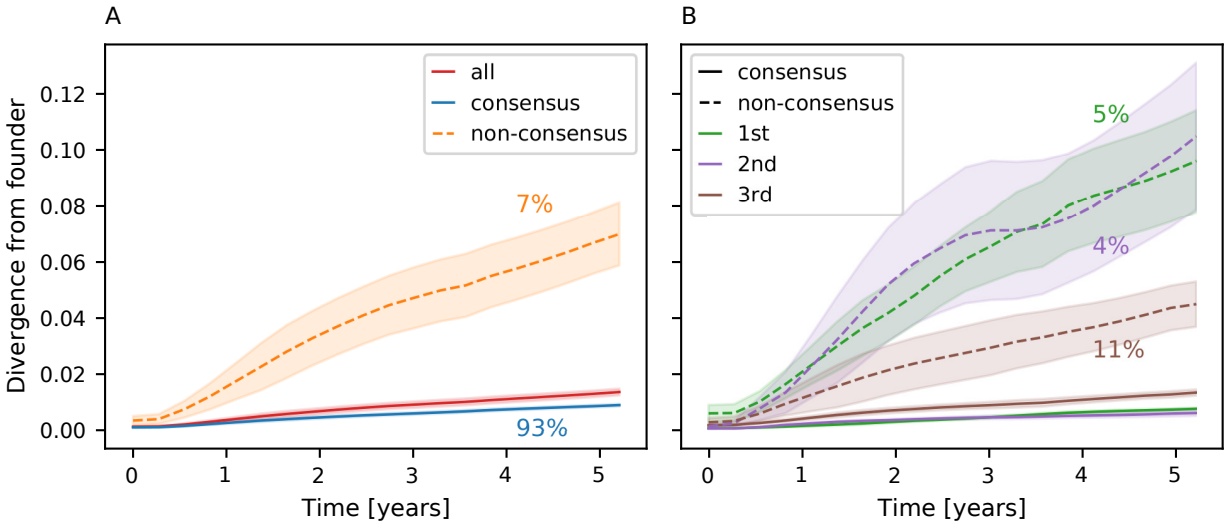
Supp. Fig. S 1: Corresponds to Figure 1, for the *env* gene. The Y axes of panel A and B are not shared in this case as the RTT distance is much higher than what we observe within host. The relative difference between the rates is higher than what is seen for the *pol* and *gag* genes. This is consistent with the fact that *env* mutates faster overall, which would also lead to more reversions.



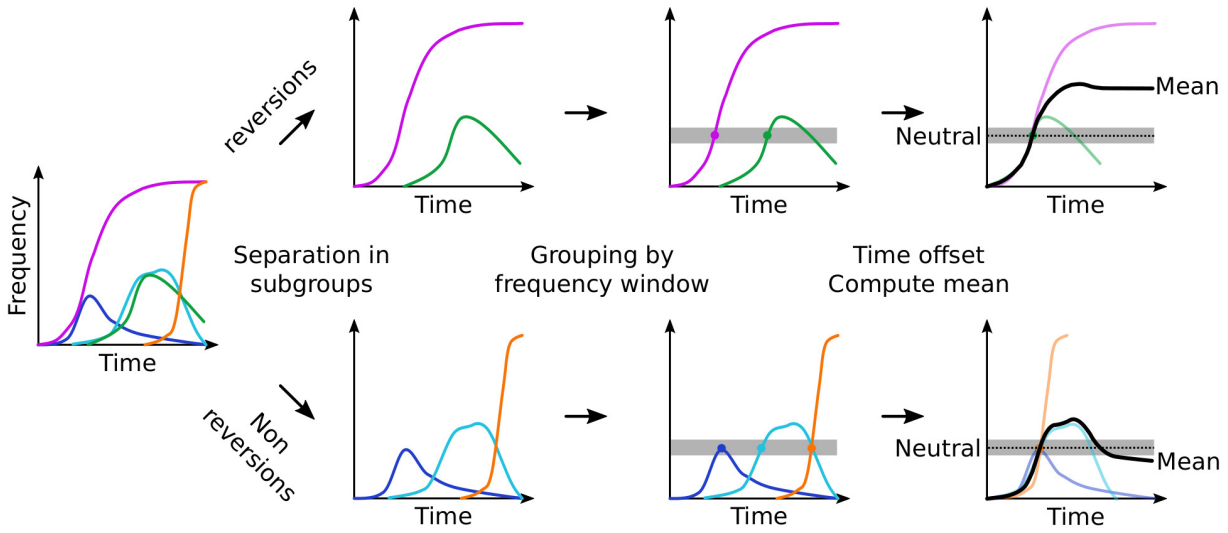
Supp. Fig. S2: Corresponds to Figure 1, for the *gag* gene. The overall mutation rate is slightly higher than for the *pol* gene but the relative difference between the rates is similar.



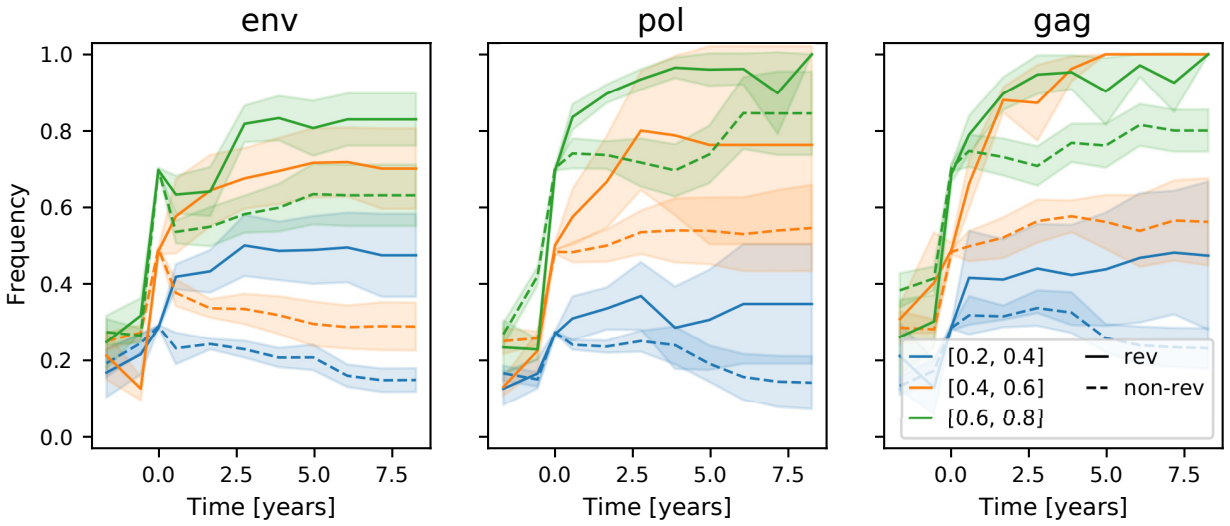
Supp. Fig. S3: Corresponds to Figure 2, for the *env* gene. In this gene, non-consensus sites at the 1st and 2nd codon position seem to diverge at a similar rate, suggesting a comparable selection for such mutations. 3rd codon position sites in a non-consensus state still diverge the slowest.



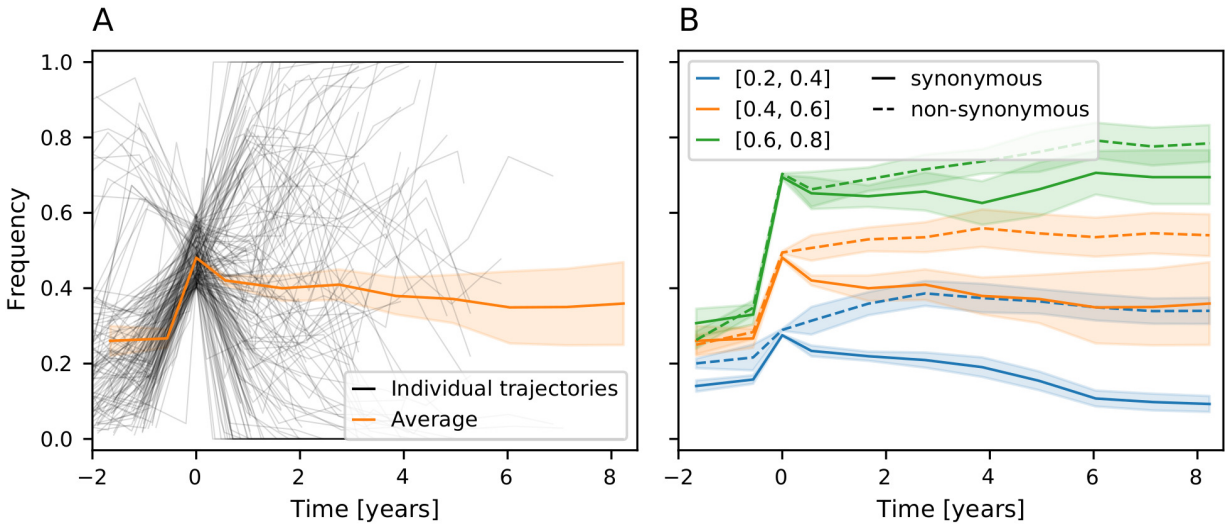
Supp. Fig. S4: Corresponds to Figure 2, for the *gag* gene. Similar to the *region*, non-consensus sites at the 1st and 2nd codon position diverge at similar rates. Non-consensus sites in the 3rd codon position still diverge the slowest, consistent with the fact that mutations at such sites are often synonymous and consequently under less selection pressure.



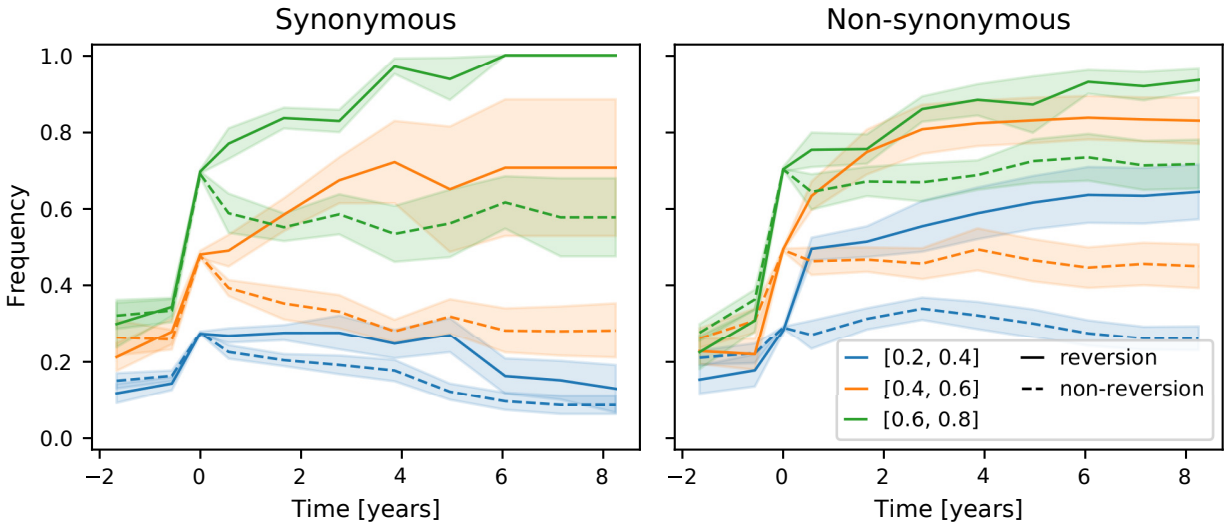
Supp. Fig. S5: Sketch of the methodology used to compute the curves shown in Figure 3 as described in the main text and section M&M 1.2. Trajectories are divided into reversion and non-reversion mutations. From each of these subgroups, trajectories that have one data point in the given frequency window are grouped together and offset in time so that this data point corresponds to t=0. We compute the mean of these trajectories and plot it in Figure 3.



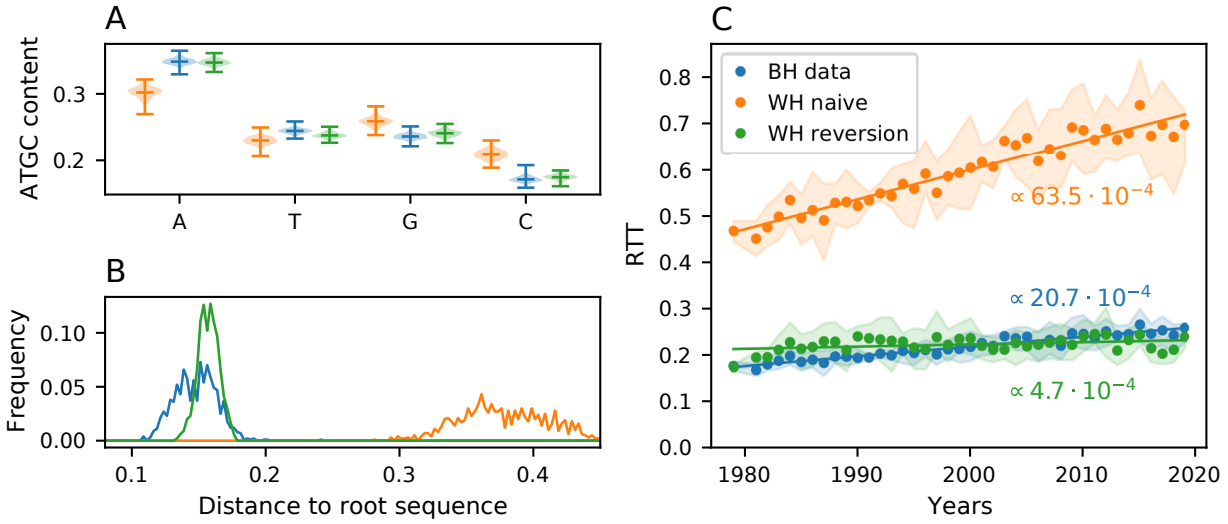
Supp. Fig. S 6: Corresponds to Figure 3B, split by gene. Selection for reversion is strongest in the *gag* region and weakest in the *env* region.



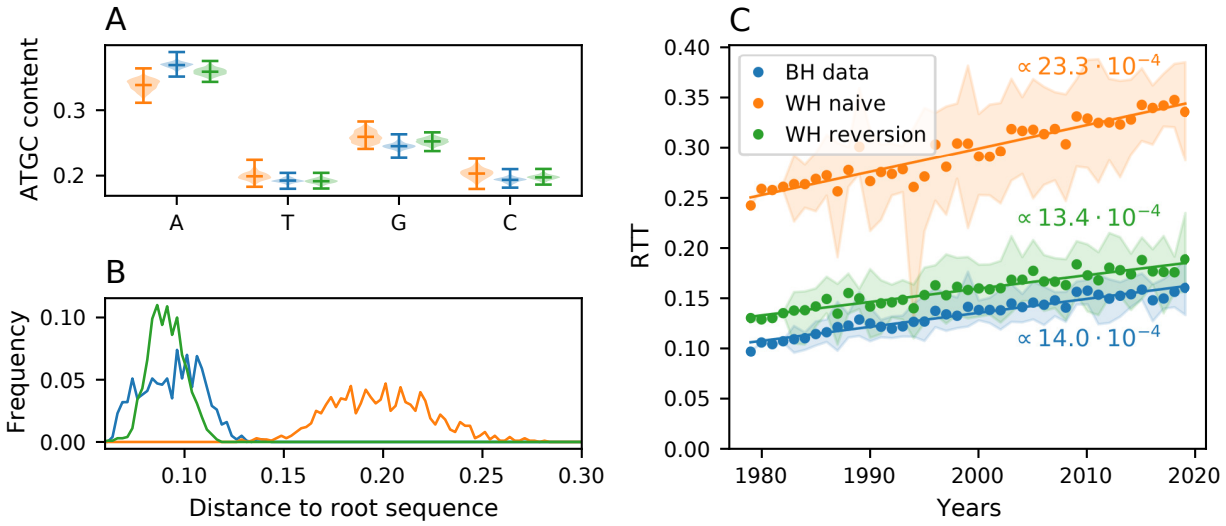
Supp. Fig. S 7: Corresponds to Figure 3, for synonymous and non-synonymous trajectories. Overall synonymous mutations are selected against and non-synonymous mutations seem to be selected for, but the effect is smaller than what we see for reversions and non-reversions in Figure 3.



Supp. Fig. S8: Corresponds to Figure 3B, with a split between synonymous trajectories only (left) and non-synonymous trajectories only (right). Overall reversion mutations are selected. Interestingly, synonymous reversions seem to be selected more strongly than non-synonymous ones at higher frequencies, but the opposite is true at lower frequencies.



Supp. Fig. S9: Corresponds to Figure 4, simulated for the *env* gene. The WH mutation rate in this region is so high that the reversion model attenuates most of the clock signal. This leads the tree reconstruction to fail and underestimates the evolution rate for the WH reversion model in this case.



Supp. Fig. S 10: Corresponds to Figure 4, simulated for the *gag* gene. The WH reversion model matches the between host observations better in this region as well, with relative differences in observed evolution rates that are similar to the *pol* region.

3

CREATION OF A HIGH-THROUGHPUT FRAMEWORK FOR BACTERIOPHAGE DIRECTED EVOLUTION

This chapter discusses our not yet published work on the creation and use of a high-throughput framework for studying bacteriophage evolution. This work focuses on the creation of an autonomous continuous culture machine, named the Aionostat, its usage and what it can bring to phage evolution research. We start with an introduction specific to bacteriophages and their evolution in section 3.1. We briefly cover the main areas in bacteriophage research and the current limitations as well. We follow by presenting the BASEL phage collection in section 3.2. This is a diverse collection of well-characterized *E.coli* phages which aims to help fix some of the current limitations of phage research that I helped to create. Then we present the high-throughput framework for bacteriophage evolution in section 3.3 and its central piece the Aionostat in section 3.4. Finally we present the results of the showcase experiments that we performed within this framework in section 3.5. These experiments were performed using phages from the BASEL collection and prove that the framework and the Aionostat are an effective way to study bacteriophage evolution.

3.1 INTRODUCTION TO BACTERIOPHAGES

3.1.1 *Historical background and relevance*

The discovery of bacteriophages dates back to the early 20th century, when Frederick Twort and Félix d'Hérelle independently discovered viruses parasitic on bacteria in 1915 and 1917 respectively. Félix d'Hérelle gave them the name of bacteriophages after noticing they were reliant on killing bacteria to amplify [69]. At the time, bacterial infections were a major cause of mortality and there was no consistently effective method to treat such infections. With his research, Félix d'Hérelle was the first to introduce the concept of using bacteriophages as antimicrobial: phage therapy. The prospect of having an antimicrobial agent that was mostly safe for humans in a time when antibiotics had not been invented yet was a small revolution. Therefore, bacteriophage research focused on therapeutic use.

However, the initial enthusiasm for phage therapy encountered several limitations, including a lack of understanding of phage biology, lack of reliability, inconsistent treatment results, and technical challenges. The advent of antibiotics in the 1940s led to a decline in phage research in the Western world as shown in figure 3.1. Nevertheless,

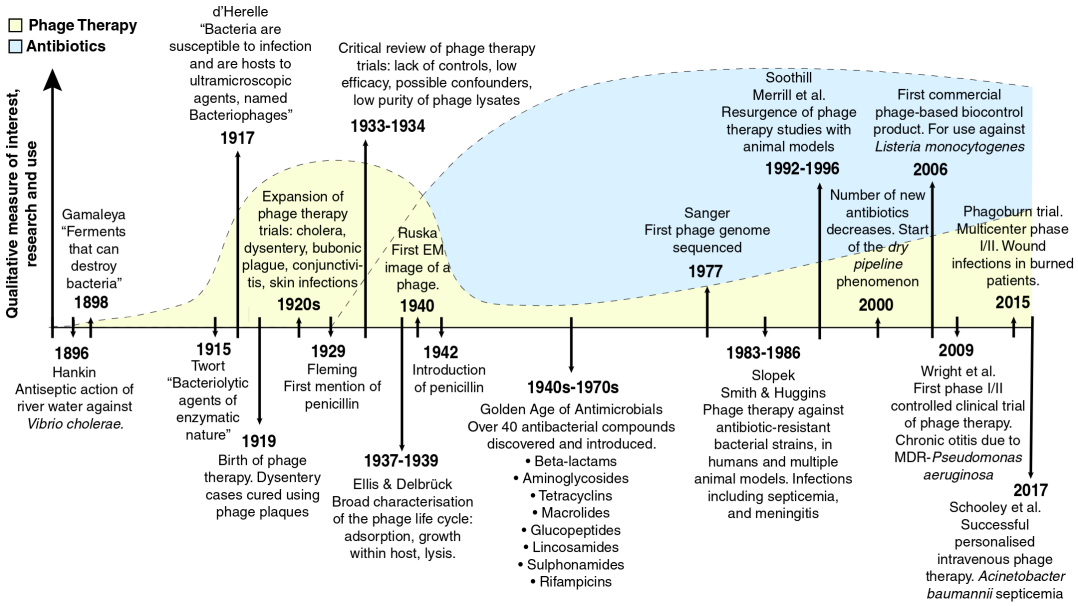


Figure 3.1: Timeline of major events in the history of research on phages, phage therapy, and antibiotics. Background curves represent a qualitative measure of the overall interest, research, and use of phage therapy (yellow) and antibiotics (blue), showing how the introduction of antibiotics and the critical review of the early phage therapy studies coincided to bring phage therapy research and development to an almost complete standstill around the 1940s. Figure and caption reproduced with permission from [6].

the rising threat of antibiotic-resistant bacteria has reignited interest in phages as potential alternatives to traditional antibiotics in recent years. This resurgence is bolstered by success stories in phage therapy, promoting a new wave of phage research which goes beyond phage therapy alone [6]. The main areas of bacteriophage research are presented in section 3.1.4.

3.1.2 *Bacteriophage biology*

Diversity

Bacteriophages are extremely abundant biological entities on Earth, and accordingly there is a huge diversity in phage biology, structure, genetic makeup, size and lifestyle. Figure 3.2 presents an overview of the main phage groups that we know of based on their genome type and morphology. The taxonomy presented in this figure and which we discuss in this section is the "traditional" phage taxonomy, which is partially based on morphology. It has recently been abolished but has not been replaced with something similarly comprehensive yet, so we chose to use this taxonomy nonetheless [70].

Like other viruses, bacteriophage genomes are made of DNA or RNA, which can be either single-stranded or double-stranded. There

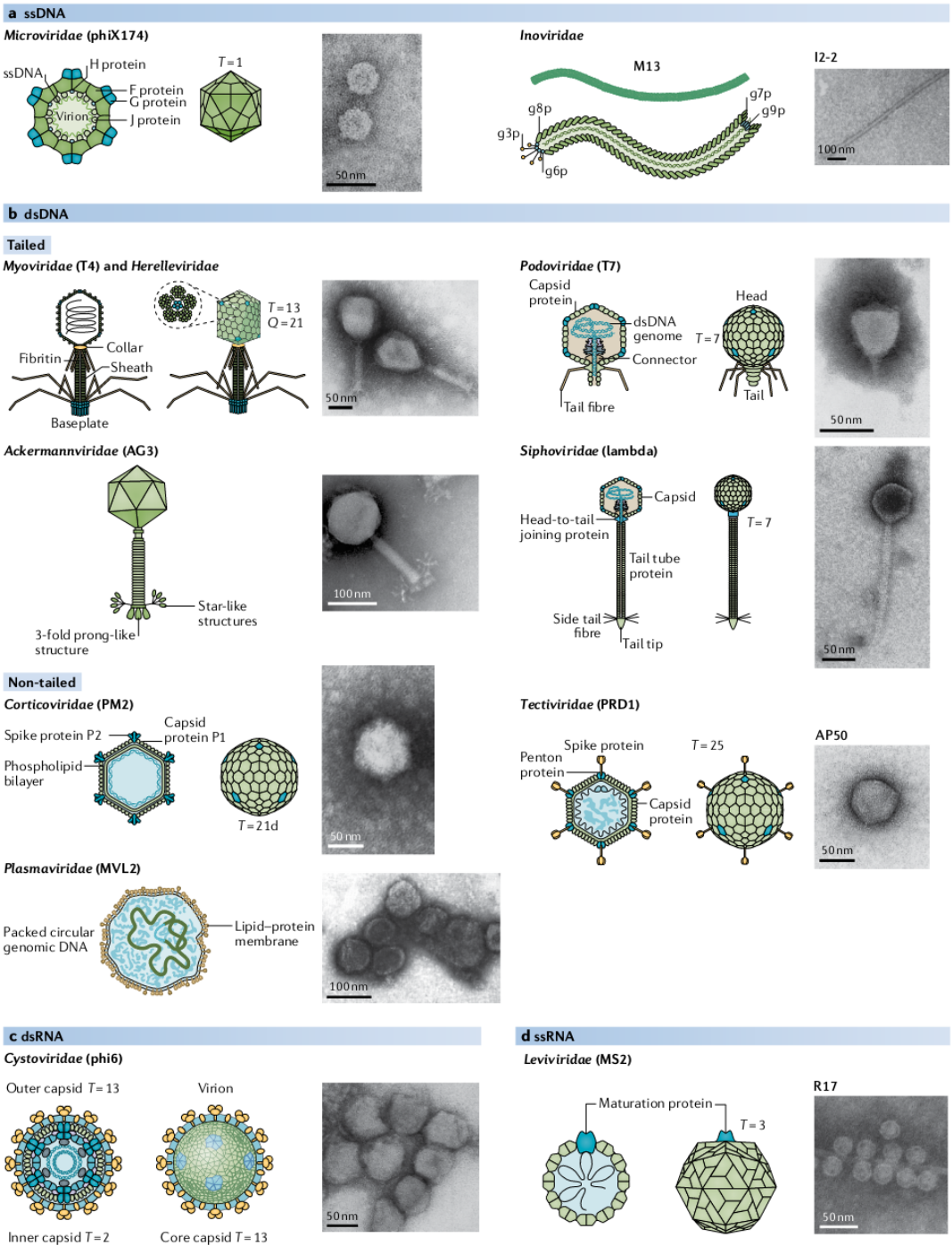


Figure 3.2: Overview of phage diversity and defining traits. T values correspond to triangulation numbers, a measure of the complexity of the capsid shape, defined as the number of proteins per asymmetric unit. Figure and caption reproduced and adapted with permission from [71]. Refer to original publication for more details and image source.

are two known families of phages using ssDNA as genetic material: *Microviridae* and *Inoviridae*. The former is a family of very small bacteriophages, hence the name, whose virions are around 25-30nm in size. Accordingly the genome of these bacteriophages is also very small, about 4 to 6 kbp. The latter have a drastically different morphology. These bacteriophages house their ssDNA genome in a long flexible cylindrical protein shell, earning them the name of filamentous bacteriophages. They are about 6 to 8nm in diameter and can be up to 2000nm long [72]. These phages also have a peculiar way of exiting their host cell once assembled, they are extruded through bacterial membranes without lysis [73].

The vast majority of bacteriophage described to date are dsDNA bacteriophages. There are several families of bacteriophages that use dsDNA as genomic material, which can be grouped into tailed and tailless bacteriophages. Tailed bacteriophages belong to the *Caudovirales* order and are by far the most commonly studied phages. They represent more than 85% of the phage genomes in public databases [71]. These phages have their dsDNA packed into a protein capsid which is attached to the tail. The size of their genome is variable, between 20kp and 500kpb, as is the size of their capsid. Receptor binding proteins located on prolate structures at the tip of the tail, such as tail fibers, are responsible for host receptor recognition and initiate the genome injection into the bacteria [13]. In particular, phages belonging to the *Myoviridae* family have a contractile tail which pierces through the bacterial membrane to deliver the phage genome into the cytoplasm of the host. These phages are the focus of our research.

Tailless dsDNA bacteriophage are much less studied. One notable feature of the *Corticoviridae*, *Tectiviridae* and *Plasmaviridae* family is the presence of a lipid membrane in the virions, lacking a hard capsid in the case of *Plasmaviridae*.

Finally there are two known families of RNA bacteriophages. These families of phages are far less studied than DNA phages. First we have the *Cystoviridae* family, whose genome is double stranded RNA of about 14kbp in length. This genome is segmented in 3 smaller parts [74, 75]. These viruses have an outer lipid membrane with two-layered inner capsid. Most of the identified bacteriophages belonging to this group infect *Pseudomonas*, but this could be due to a bias in screening method [76]. Lastly there is the *Fiersviridae* family (previously named *Leviviridae*), which are small single stranded RNA viruses with a genome of about 4kbp.

Life cycle and lifestyle

Bacteriophages, or phages, exhibit a variety of lifestyles that are intricately tied to their morphology and their interaction with their bacterial hosts. The life cycle of tailed bacteriophages, which are the focus of our research, can be broadly categorized into two main types:

lytic and lysogenic. Obligatory lytic phages are called virulent phages, while the those capable of either a lytic or lysogenic lifestyle are called temperate phages [77].

In the lytic cycle, phages infect bacterial cells and immediately begin the process of replication. Well-studied models such as the T phages belong to this group. The infection usually involves destroying the hosts defense systems and genome followed by hijacking the host's cellular machinery to synthesize new phage components. These are then assembled into new phage particles, at which point the host cell is lysed, releasing the progeny viruses. This cycle can be relatively rapid, with the entire process taking around 25 minutes or less for rapid phages like T7 under optimal conditions [16]. The lytic cycle duration is very dependent on the phage and the host state. The cycle is typically faster when bacteria are in a fast growing state. The number of virions produced per infected cell, or burst size, also varies significantly. Some phages like T7 can release more than a hundred new virions upon lysis.

The lysogenic cycle, on the other hand, involves the integration of the phage genome into the host's genome under specific conditions. A well-studied model for this lifestyle is phage λ, where the lysis-lysogeny decision of the phage upon infection is at the center of many studies [78–80]. This integration into the host genome is similar to what is observed for retroviruses such as HIV-1. In this state, known as a prophage, the viral genome can be replicated along with the host's DNA during cell division. This results in a coexistence of the phage and bacterial genome that can be stable over many generations. Since temperate phages can integrate into bacterial genomes, they can form a transiently beneficial symbiotic relationship with their host. Consequently it is not uncommon for temperate phages to have genes that provide a fitness advantage to their host. These can be genes encoding toxins such as Shiga or Cholera toxins [81, 82], or even immunity systems against other bacteriophages [83]. However, under specific triggers such as stress or UV radiation, the prophage can be induced and will excise itself from the host chromosome to enter the lytic cycle. This initiates the production of new phage particles and the eventual lysis of the host cell. The lysogenic lifestyle brings some interesting evolutionary dynamics. There exists many triggers that govern prophage induction [77].

Although the lytic and lysogenic lifestyles describe broadly two categories of bacteriophages, recent research seems to suggest that the difference in phage lifestyle in the environment is not a dichotomy, but rather a continuum which also includes inefficient lytic and chronic infection lifestyles [84].

3.1.3 *Bacteriophage evolution*

The omnipresence of bacteriophages in natural environment stems from their ability to evolve and adapt to changing conditions. Like many viruses, they are fast evolving biological entities. This ability has given rise to the wide variety of phage and phage lifestyle we observe today [85]. The evolutionary dynamics observed in bacteriophage populations are very diverse due to the many structures, hosts and lifestyles. Understanding the evolution of bacteriophages has practical relevance in various domains of research. For instance, their ability to constantly adapt to evolving bacteria is extremely relevant in healthcare, where the rise of resistant bacterial strains is a real concern.

Similar to the evolution of HIV-1, bacteriophages also evolve both vertically and horizontally [86]. Vertical evolution occurs through random errors in the replication process when creating new virions, a process which is reasonably well understood. On the other hand, horizontal evolution is a much more intricate as bacteriophages have the ability to exchange genome fragments in different contexts. A lytic phage, for example, can exchange genetic material during an infection in the following ways:

- **Exchange with a lytic phage:** In the case of a co-infection of a host cell by two different phages, there can be recombination events between the two phage genomes. This can create chimeric offspring phages, with a genome composed of a mix of the two parents' genomes [87].
- **Exchange with the host:** When infecting a bacteria, lytic phages usually chop the host genome to reuse these resources for creating new virions. During this process there is a chance that pieces of the host's genome recombine with the bacteriophage genome.
- **Exchange with a prophage:** In the case of an infection of bacteria that contains integrated prophage(s), there can be a recombination between the lytic phage genome and the prophage genetic material which can again create a chimeric offspring. Recent research suggests that this is the main driver of horizontal gene transfer in bacteriophages. There are two main reasons for this. First, it is more likely to have two phage genomes simultaneously in the same host cell if one of them is carried for extended periods of time on the host genome. Second, recombination rate of homologous sequences is likely higher than for non homologous sequences, so it is more likely to recombine with other phage genes rather that the host's genome. Considering that prophages are found in many bacteria, one could expect that this is the most likely way of horizontal evolution.

The mosaic nature of bacteriophage genomes and the presence of recombinases in many of them suggests that horizontal transfer is

the main driver of phage evolution [88, 89]. The ability to exchange and acquire large chunks of genetic materials gives the opportunity to acquire or lose whole biological functions, which seems essential to adapt rapidly to changing environments. Bacteria also evolve over time, and the evolutionary pressure imposed by phage predation can drive the evolution of anti-phage defense systems [90]. Bacteria develop anti-phage systems while bacteriophages evolve to counter or bypass such systems. A prime example is the widespread presence of CRISPR systems in bacteria while anti-CRISPR systems are found in some bacteriophages [91, 92]. In reality the situation is more complex than this, as the co-evolution of bacteriophages and their hosts can also be mutually beneficial, as is the case with some temperate phages.

An integral aspect of bacteriophage evolution is their ability to perform 'host jumps', a phenomenon where phages evolve to infect a new bacterial strain. Host jumping is a complex evolutionary process involving a series of genetic adaptations that enable a phage to recognize and bind to new host receptors, which are typically highly specific, and then successfully hijack the host machinery for a productive infection. From an evolutionary perspective, this is a complex task, but bacteriophages seem extremely capable at performing such jumps [93]. This ability appears closely linked to their ability to exchange genetic elements, which can for example provide new receptor binding proteins from another bacteriophage that promote binding to a new host.

Host jumps are essential for many aspects of bacteriophage success in nature, but they are particularly interesting to us in the context of phage therapy. It is common that phage therapy centers do not have effective bacteriophages against the strain causing a patient's infection. Thankfully one can leverage the ability of bacteriophage to perform host jumps to evolve a new phage tailored for that specific patient. This is typically performed using variants of the Applemans protocol, which involves exposing several phages to a series of bacterial strains, encouraging them to recombine, adapt and potentially expand their host range in the search of evolved bacteriophages that would be suitable to treat the patient's infection [94]. Although this is empirically shown to be effective, the precise experimental parameters and evolutionary mechanisms to optimize phages are not yet fully understood [95]. Broadly speaking, the amazing ability of bacteriophage to evolve is one of their defining trait and is tied to most of the research areas presented in the following section.

3.1.4 *Current research*

Ecology

Viruses are the most abundant and diverse biological entity on Earth, with an estimated 10^{31} virions existing at any given moment [96]. Most of these viral particles are bacteriophages, or phages in short, the viruses targeting bacteria, and current research seems to show that they play a key role in ecology. Bacteria are present in most environments found on the planet, and it is estimated that bacteriophages outnumber their bacterial hosts in these environments by an order of magnitude on average, initiating approximately 10^{23} infections per second. The sheer number of bacteriophages that exist and the number of infection they cause make them the top predators of the microbial world [85]. The omnipresence of bacteria in diverse ecological environments makes bacteriophages key ecological players due to the predatory pressure they exert on bacterial communities. The interaction between phages and bacteria influences the microbial dynamics in various habitats and often promotes stability in such environments. For example in the case oceanic ecosystems, the predation of bacteria by phages contribute significantly to the carbon cycle as illustrated in 3.3. It is estimated that around 50% of bacterial deaths is caused by bacteriophages, the other half being due to grazing protists [97]. By infecting and lysing bacteria they release dissolved organic matter (DOM) in the water, which can be either reused by other microbes or aggregate as particulate organic matter (POM) and sinks to the deep ocean. This viral shunt maintains the marine environment by encouraging nutrient cycling and promotes the carbon shunt of the ocean. The role of bacteriophages in ecology is not restricted to marine ecosystems, they are also key players in various other environments such as plants [98] and even our own gut microbiota [99].

Viral models

The widespread presence of bacteriophages out there and the diversity observed is a result of their amazing ability to evolve and adapt to various hosts and environments [100]. Their high mutation rates, fast generation time, diverse lifestyles and their ability to exchange genomic material provide a unique opportunity to study a wide range of evolutionary dynamics. Thanks to recent advances in sequencing technologies and the renewed interest in phage research, the number of individually published bacteriophage genomes has doubled in the last 5 years (see `https://millardlab.org/bacteriophage-genomics/`), and the numbers are even higher for viral genomes from metagenomic studies [101]. This surge of data enables broad study of bacteriophages using phylogenetics. Unlike many viruses infecting eukaryotes, phages can also be easily manipulated and studied in various laboratory con-

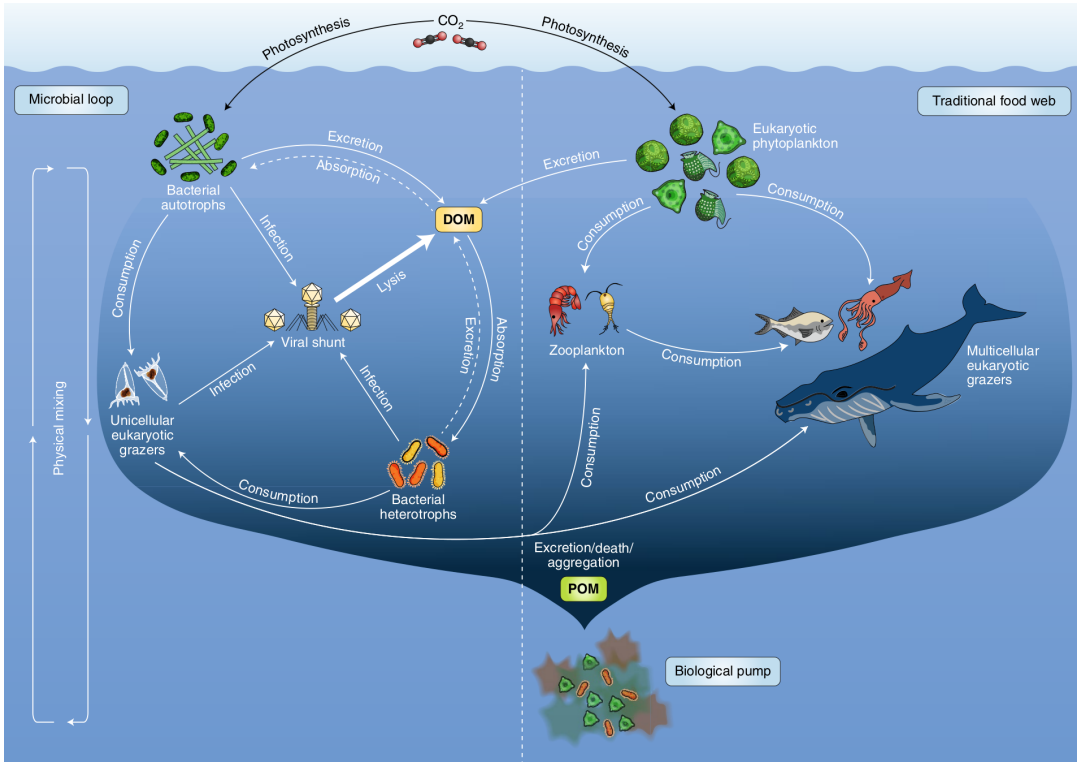


Figure 3.3: Schematic of the carbon cycling in the ocean. When bacteria are lysed by phages, carbon and nutrients are released in the water, some of which is reused while another part sinks to the deep ocean. The viral shunt caused by bacteriophages function as both the ocean's recycling system and carbon sink. Figure and caption reproduced and adapted with permission from [97].

ditions using relatively straightforward and harmless bacterial cultures and plates. Therefore it is comparatively simple to test evolutionary hypothesis from analysis by direct experiments. This ease of use and speed combined with the wide variety of evolutionary strategies observed make bacteriophages an excellent model for studying viral evolution as a whole, both experimentally and with a bioinformatics approach.

Outside of bacteriophage evolution per se, these viruses can also serve as a valuable tool for broader studies of protein and RNA evolution [102, 103]. Techniques such as phage display exemplify how attributes tied to phage fitness can be enhanced through the evolutionary processes of the phages themselves [104].

Phages in healthcare

The historical focus of bacteriophage research has been largely driven by their potential in phage therapy, leveraging their inherent ability to target, kill and replicate on bacteria. This approach, initiated in the early 20th century, was faced with many challenges due to limited understanding of phage biology and technical limitation at the time,

therefore research in this area was mostly stopped with the advent of antibiotics. The rise of antibiotic resistant bacteria is already a real threat in many parts of the world, and this is only getting worse over time. Consequently, developing alternative ways of dealing with bacterial infection is essential to uphold current health standard, and possibly improve them. In this context, the use of bacteriophages for phage therapy is regaining interest as it is seen as a promising solution to fight resistant bacteria [7]. This new enthusiasm is amplified by the possibilities offered by modern molecular biology tools as well as a few successful phage therapies that made the headlines in the last years.

In most cases, phage therapy involves finding bacteriophages that are effective on the bacterial strain(s) causing the infection, potentially training them, and then using such bacteriophages in a cocktail administered to the patient. We rely on the ability of phages to find their target bacteria, infect and kill it while self-replicating to clear the infection. In most cases, phages are combined with antibiotics in a synergistic manner [6]. Such combined treatments seem to perform best as the effectiveness of phage killing seems to be inversely correlated to the resistance level of bacteria, imposing a challenging evolutionary tradeoff to the bacteria [105]. Currently, the main challenge of phage therapy is finding the right phages and training them to treat the infection. This is no easy task as bacteriophages have a narrower host range compared to most antibiotics, but also because bacteria often carry anti-viral systems [106]. A good overview of the current effectiveness of bacteriophage therapy is given in the retrospective of the first 100 cases of phage therapy in the leading institute for phage therapy in Europe [107].

Outside of direct use of natural bacteriophages, there is also an emerging field that looks at engineering bacteriophages to enhance their effect, or simply to use them as payloads to deliver drugs to specific targets, opening new avenues for medical interventions. Since then we have realized that bacteriophages also have a natural role in our health. Bacteriophages contribute to maintaining the delicate balance of our microbiota, particularly in the gut, where they play a crucial role in modulating and driving microbial diversity [108, 109]. This diversity is essential for various aspects of health, ranging from nutrient absorption to immune system modulation [110].

Our current knowledge of bacteriophage biology and phage-bacteria interactions, especially in the patient's body, limits our ability to successfully predict treatment outcome from the phage characteristics. Currently phage therapy can be compared to personal medicine, where treatment is designed on a per patient basis. This severely limits the potential of phage therapy for broader use, like is the case for antibiotics. Developing our understanding to improve these aspects is one of the main drivers of current phage research.

Reservoir of biological functions

Bacteriophages and their hosts are among the most diverse biological entities on the planet. They represent an immense reservoir of biological functions that is at the source of many molecular biology discoveries. For instance, phages have been central to the understanding of DNA as the genetic material with the Hershey-Chase experiments, or the discovery of restriction modification systems [111, 112] and the CRISPR-Cas adaptive immune system in bacteria [113, 114]. These are just a few examples of the many novel enzymes, proteins and functions discovered via the study of bacteriophages. Figure 3.4 illustrates the main discoveries linked to the study of phages. These discoveries not only advanced our fundamental knowledge but also provided essential tools for modern molecular biology research. The CRISPR-Cas system, for example, has been repurposed into a powerful genome editing tool with widespread applications in medicine, agriculture, and research. Recent estimates suggest that we have just begun to uncover the tip of the iceberg of viral diversity. For instance, it is estimated that around 60% of the annotated genes in bacteriophage genomes are hypothetical proteins with no characterized homologs [115]. It is therefore tempting to speculate that many more exiting biology remains to be uncovered by studying bacteriophages [105, 116].

Limitation of current research

Improving our understanding of phage evolution is essential for the research fields presented above, like phage therapy, where a better knowledge of these evolutionary dynamics could help optimize bacteriophages to cure patients. Even when it is not the direct focus of the research, bacteriophage evolution is central to all areas of research presented above. Be it for ecological reasons, where it is essential to understand how phages adapt and diversify to shape ecosystems, or for studying their molecular biology, where evolutionary cues provide meaningful information regarding the role of unknown genes.

However, studying phage evolution poses multiple challenges. The current state of phage evolution research is limited to a handful of well-characterized bacteriophages like the T-series of phages infecting *E. coli* [117], or to broad environmental metagenomics phage studies where the phages themselves are rarely isolated and poorly studied [118–120]. The former limits the scope of the findings, while the latter prevents detailed analysis that would require experimental intervention. To better understand and manipulate phage evolution to our benefit we need to bridge this gap, which requires both a better understanding and characterization of phage diversity as well as methods that are high-throughput, rapid, reproducible, and cost-effective to perform evolution experiments on bacteriophages and infer general principles.

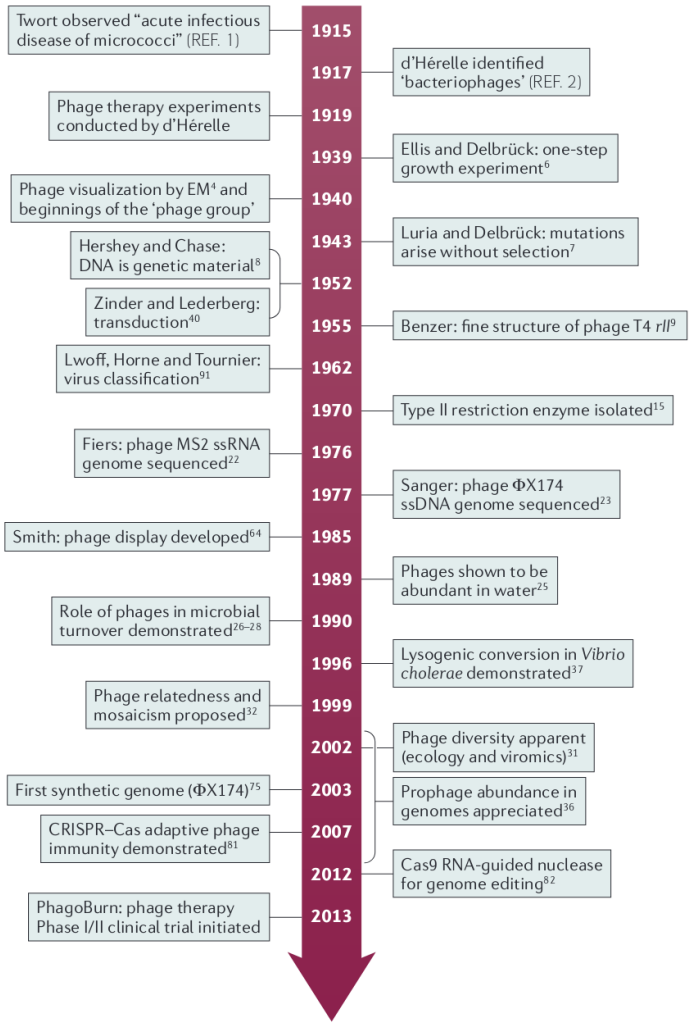


Figure 3.4: Timeline of major events in the first 100 years of phage research. EM, electron microscopy; ssDNA, single-stranded DNA; ssRNA, single-stranded RNA. Figure and caption reproduced and adapted with permission from [116].

It is in this context and to help solve these issues that we performed the work presented in section 3.2, 3.3, 3.4 and 3.5.

3.2 BASEL PHAGE COLLECTION

One of the current shortcomings of bacteriophage research is that studies are often restricted to a handful of well-characterized bacteriophages, or to newly isolated phages that are often poorly characterized and difficult to obtain for further studies. To help alleviate this issue we created a new bacteriophage collection: the BActeriophage SElection for your Laboratory, or BASEL collection for short. This work was published in November 2021 [121].

In this publication, we isolated *de novo* 68 bacteriophages and built a collection by characterizing them along with 10 well-studied phage references. In total, the collection is composed of 78 phages with in-depth phenotypic characterization of their host receptors and sensitivity to several defense systems of bacteria alongside high-quality hand curated and annotated genomes. Figure 3.5 gives an overview of the collection and how it was constructed. The BASEL phage collection is largely representative of the natural diversity of bacteriophages that infect *E. coli*, and their characterization and curated genomes provides a solid foundation for bacteriophage research. The patterns observed in phage phenotypes are clearly indicative of evolutionary trade-offs between traits like broad host range and resistance to bacterial immunity which likely explain the wide diversity observed.

Although it has been published relatively recently, the BASEL collection is already widely used as a reference. At the time of writing, the manuscript has been cited 83 times (Google scholar metrics) and the collection has been shared with more than 50 research groups around the world. This work was spearheaded by Prof. Harms and the collection and characterization of bacteriophages involved the work of many students. I contributed to this work mainly by providing bioinformatics analysis alongside the isolation and characterization of a few bacteriophages.

The research group of Prof. Harms is currently working on an "expansion-pack" of the BASEL phage collection. Currently this collection contains exclusively phages that were isolated on the laboratory strain *E. coli* K-12. This strain has lost its O-antigen over the years of evolution in laboratory environments [122]. Although this does not affect the biology of the bacteria, the presence of O-antigen on the bacterial surface, or lack thereof, impacts drastically the ability of bacteriophages to infect the bacteria. Most of the bacteriophages from the BASEL collection cannot infect the same bacterial strain with restored O-antigen, so it is likely that the isolation of phages was somewhat biased due to the bacterial strain used. To improve the diversity of bacteriophages in the collection, this "expansion-pack" focuses on phages that are reliant on the presence of O-antigen. The publication is currently under preparation.

The role of O-antigen in bacteriophage infection is central to the experiments that we present in section 3.5. To perform these experiments we used bacteriophages from the BASEL collection which belong to the *Vequintavirinae* group and relatives, and evolved them for increased infectivity on an *E. coli* K-12 strain with restored O-antigen. Details about the *Vequintavirinae* bacteriophages from the BASEL collection are shown in figure 3.6. These bacteriophages have a genome of 131kb to 140kb which characteristically encodes 3 different sets of lateral tail fibers that are coexpressed [123]. Because of this unusual feature, such bacteriophages can be compared to "nanosized Swiss army knife".

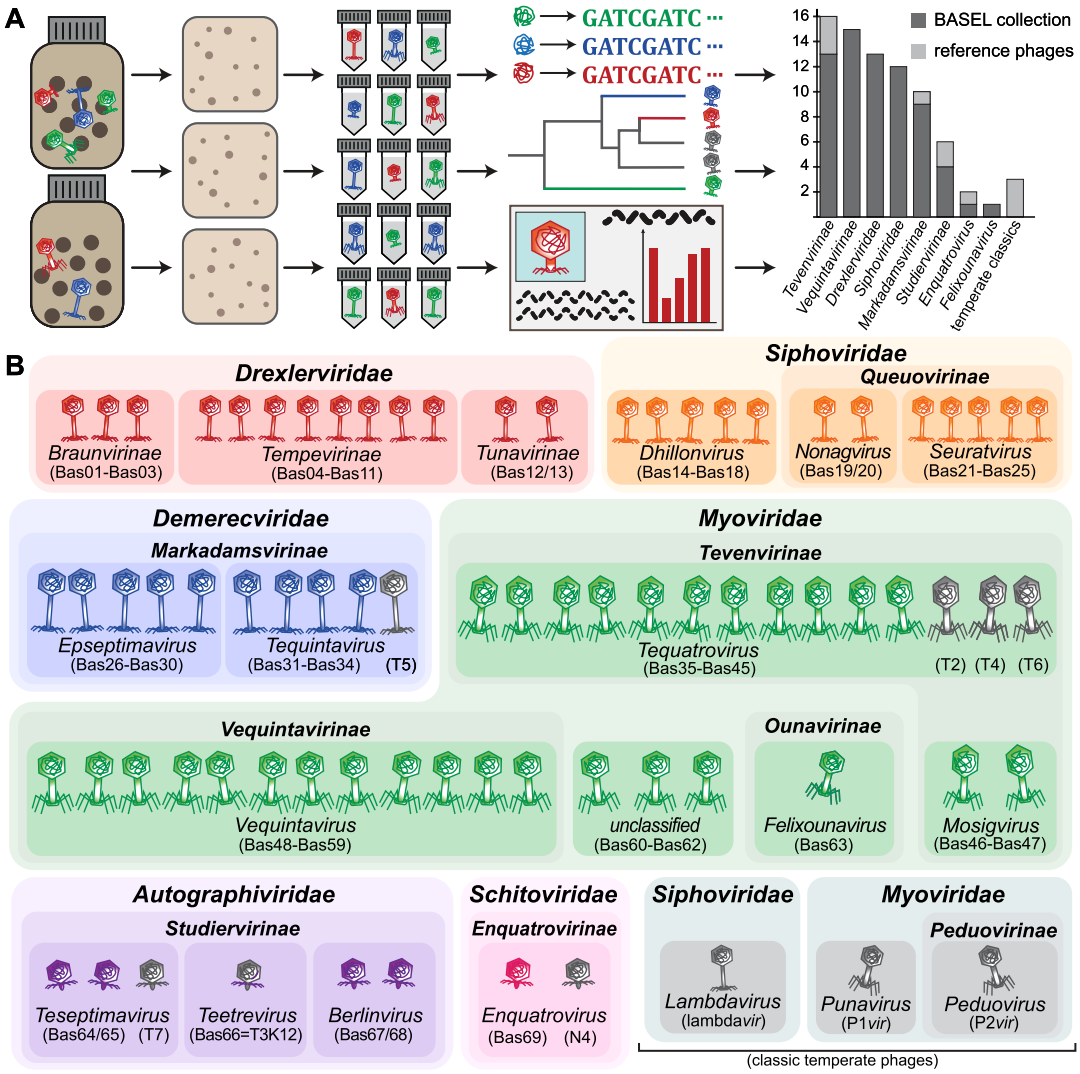


Figure 3.5: Overview of the BASEL phage collection. **A**: Illustration of the workflow of bacteriophage isolation, characterization, and selection that resulted in the BASEL collection (the bar diagram includes phi92-like phages in Vequintavirinae for simplicity). **B**: Taxonomic overview of the bacteriophages included in the BASEL collection and their unique Bas identifiers. Newly isolated phages are colored by current taxonomic classification, while reference phages are shown in gray. Figure and legend reproduced from [121] under Creative Commons Attribution License.

This distinctive feature might help recognize and bind to several host surface structures, which could be responsible for the exceptionally broad range of these phages.

There are several reasons why the *Vequintavirinae* bacteriophages and their relatives were used as models in our experiments:

- These bacteriophages come from the BASEL collection, therefore they are well-characterized and readily available.
- There are many phages in this group with various amount of genetic similarity. This gives us the chance to perform evolutionary experiments and probe the impact of genetic similarity or lack thereof.
- These phages can infect *E. coli* K-12 with restored O-antigen originally, but are not efficient at doing so. This leaves room for improvement of this phenotype via evolution.
- These bacteriophages do not have large DNA modifications, which simplifies the sequencing using Nanopore. We originally planned to work with bacteriophages from the *Tevenvirinae* group, but such bacteriophages have large DNA modifications that impair the sequencing via Nanopore [124].

3.3 FRAMEWORK FOR BACTERIOPHAGE EVOLUTION

Current research about bacteriophages and their evolution has some limitations, as discussed in 3.1.4. On one side of the spectrum there is "low throughput" research that looks at a few bacteriophages and characterizes them well, while on the other side you find "high-throughput" studies that look at phage metagenomics broadly but fall short in the characterization of these phages. There is a gap between these two sides of bacteriophage research, and work such as the BASEL phage collection help bridge the two sides by providing a large and representative collection of bacteriophages alongside a high level of characterization. Nonetheless, to understand phage evolutionary dynamics in a meaningful way one must also be able to evolve and study large amounts of phages. This requires methods that are high-throughput, rapid, reproducible and cost-effective. It is in this context that we developed a complete high-throughput framework to evolve bacteriophages and analyze their evolution. This framework is illustrated in figure 3.7 and involves both experimental work and bioinformatics analysis.

The goal of this framework is to perform and analyze the results of bacteriophage evolution experiments at scale, rapidly and with minimal amount of manual labor. Bacteriophage evolution experiments are typically performed by hand using daily serial passages, which

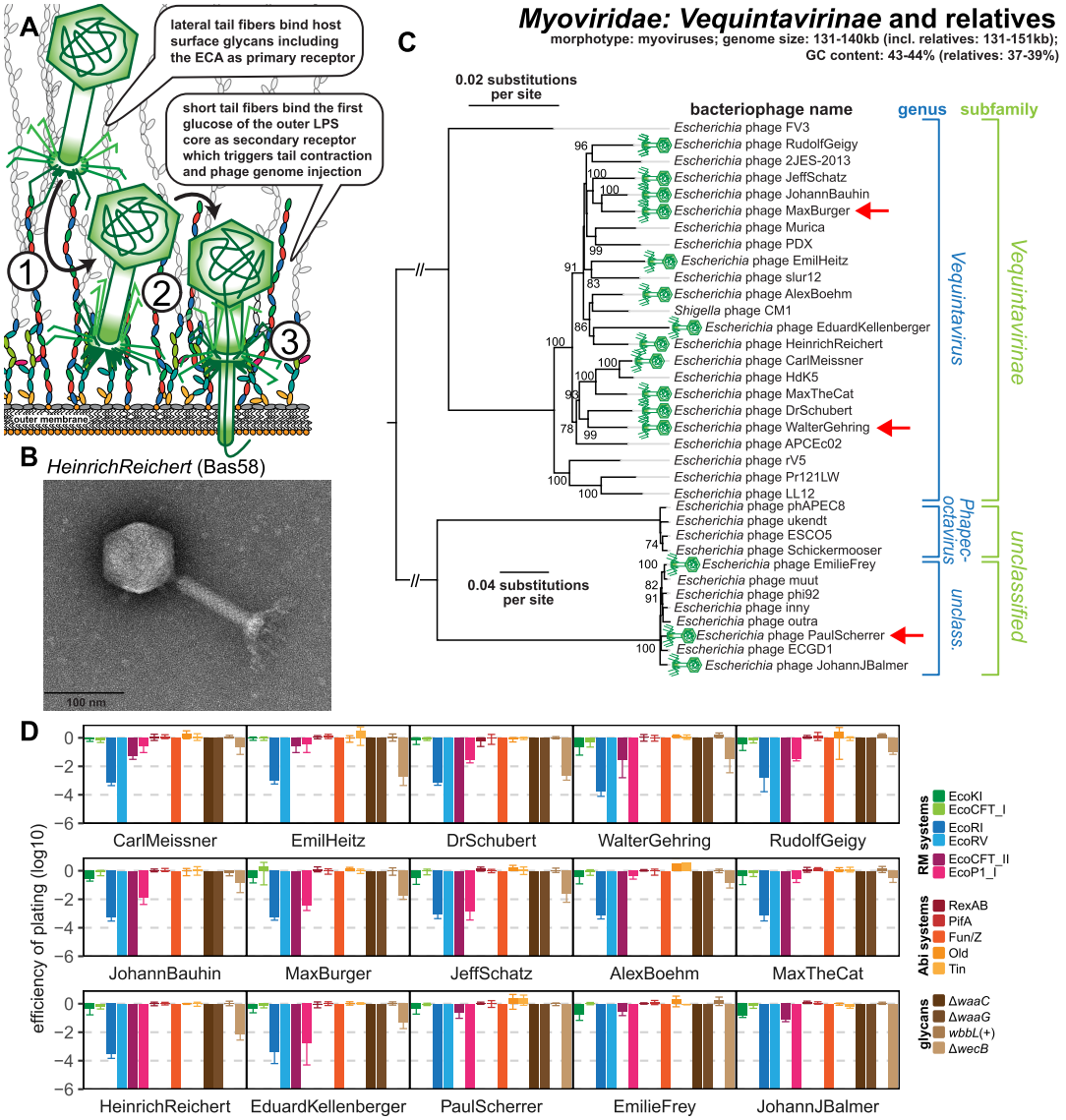


Figure 3.6: Overview of the Myoviridae subfamily Vequintavirinae and relatives. A: Schematic illustration of host recognition by Vequintavirinae and related myoviruses. B: Representative TEM micrographof phage HeinrichReichert (Bas58). C: Maximum-Likelihood phylogeny of the Vequintavirinae subfamily of Myoviridae and relatives based on a curated whole-genome alignment with bootstrap support of branches shown if >70/100. The phylogeny was rooted between the Vequintavirinae sensu stricto and the 2 closely related, formally unclassified groups at the bottom. Newly isolated phages of the BASEL collection are highlighted by green phage icons, red arrows are the phages used in section 3.5. D: The results of quantitative phenotyping experiments with Vequintavirinae and their phi92-like relatives regarding sensitivity to altered surface glycans and bacterial immunity systems are presented as efficiency of plating. Data points and error bars represent average and standard deviation of at least 3 independent experiments. Figure and legend reproduced from [121] under Creative Commons Attribution License.

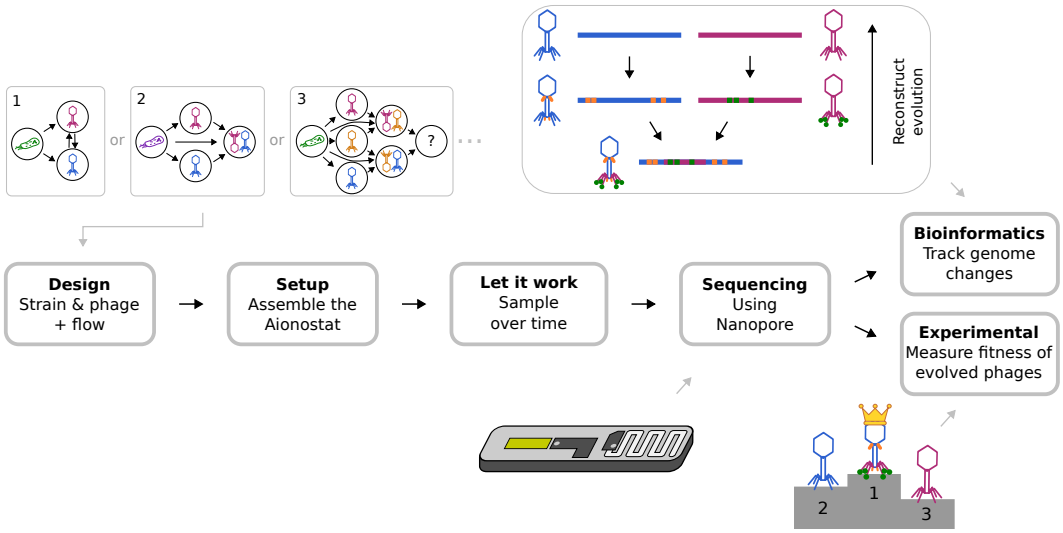


Figure 3.7: Overview of the experimental framework for high-throughput bacteriophage evolution. The central piece of the framework is the Aionostat, a continuous culture machine that is presented in 3.4 and enables to do a wide variety of evolution experiments with minimal amount of manual labor. Streamlined processes have also been developed for the sequencing, genomic and phenotypic analysis of these experiments.

is a time consuming process prone to experimental error like cross contamination. We created this framework to improve on these aspects. The central part of this framework is the Aionostat. We designed and built this machine to perform continuous culture experiments on bacteriophages autonomously. The Aionostat is presented in details in section 3.4, we focus on the general framework here.

Performing an experiment in this framework starts with the design of such experiment. This includes selecting the appropriate strain(s) of bacteria and phage(s), defining the conditions of the experiment such as media, temperature and duration, and planning the flow of liquid between the different vials of the Aionostat which defines its configuration for the experiment. Figure 3.7 shows examples of flows that can be used in an experiment.

Once the design of the experiment is chosen it is time to the setup of the Aionostat. It is assembled in the right configuration, sterilized, programmed with the right experimental parameters and then loaded with the bacteria and phages in the appropriate vials. Fresh media is then connected to the input of the machine and the experiment is started.

Over the duration of the experiment, the work of the operator is limited to refilling the input media bottle when they are empty and replacing the waste bottle when it gets full. Additionally, one can take samples from the vials for storage and later analysis if interme-

diate time points of evolution are of interest. More details about the Aionostat and these manipulations is be given in section 3.4.

Once the experiment is over, the Aionostat is cleaned and disassembled while the samples are processed to extract the DNA. This DNA can be from both bacteria and bacteriophage, and from population samples as well as clonal samples. The DNA is then directly sequenced on site using the Oxford Nanopore technology. We are using the rapid barcoding kit 24 which enables to sequence 24 samples at a time with a total yield from 15 to 30Gbp, which is enough to have deep sequencing depths for the samples. The data is then basecalled using the pipeline here: `https://github.com/vdruelle/nanopore_basecalling`. The sequencing takes 3 days and costs approximately 28CHF per sample, so it is both fast and relatively cheap, which is in line with the goal of this framework.

The final steps of the framework involve both experimental and computational work to characterize the evolutionary changes observed. The genomic changes are analyzed from the sequencing data using a Snakemake pipeline, which is publicly available at `https://github.com/mmolari/evo-genome-analysis`. The phenotypic changes are measured by performing killing curve experiments of the ancestral and evolved phages using a plate reader. The comparison of these killing curves inform on the fitness advantage of the evolved bacteriophages, which can then be linked to the genomic changes observed from the sequencing and bioinformatics analysis.

All in all, within this framework a trained user can reliably perform an evolution experiment and characterization of several bacteriophages in less than a month with limited manual labor. This represents a methodological advancement over traditional manual approaches that are still standards in this field. This framework offers a scalable approach to investigate bacteriophage evolution and contributes to the advancement of bacteriophage research.

3.4 THE AIONOSTAT

3.4.1 *Overview*

In the previous section we presented the general framework that we developed for high-throughput study of bacteriophage evolution. The central piece of this framework is the continuous culture device that performs the evolution experiment. This device has been named the Aionostat, in reference to the greek deity Aion associated to cyclic time. A picture of this machine and its main components is shown in figure 3.8. In this section, we dive into the details of the Aionostat to present how it is able to perform such evolution experiments. We present the showcase experiments performed with this machine in section 3.5.

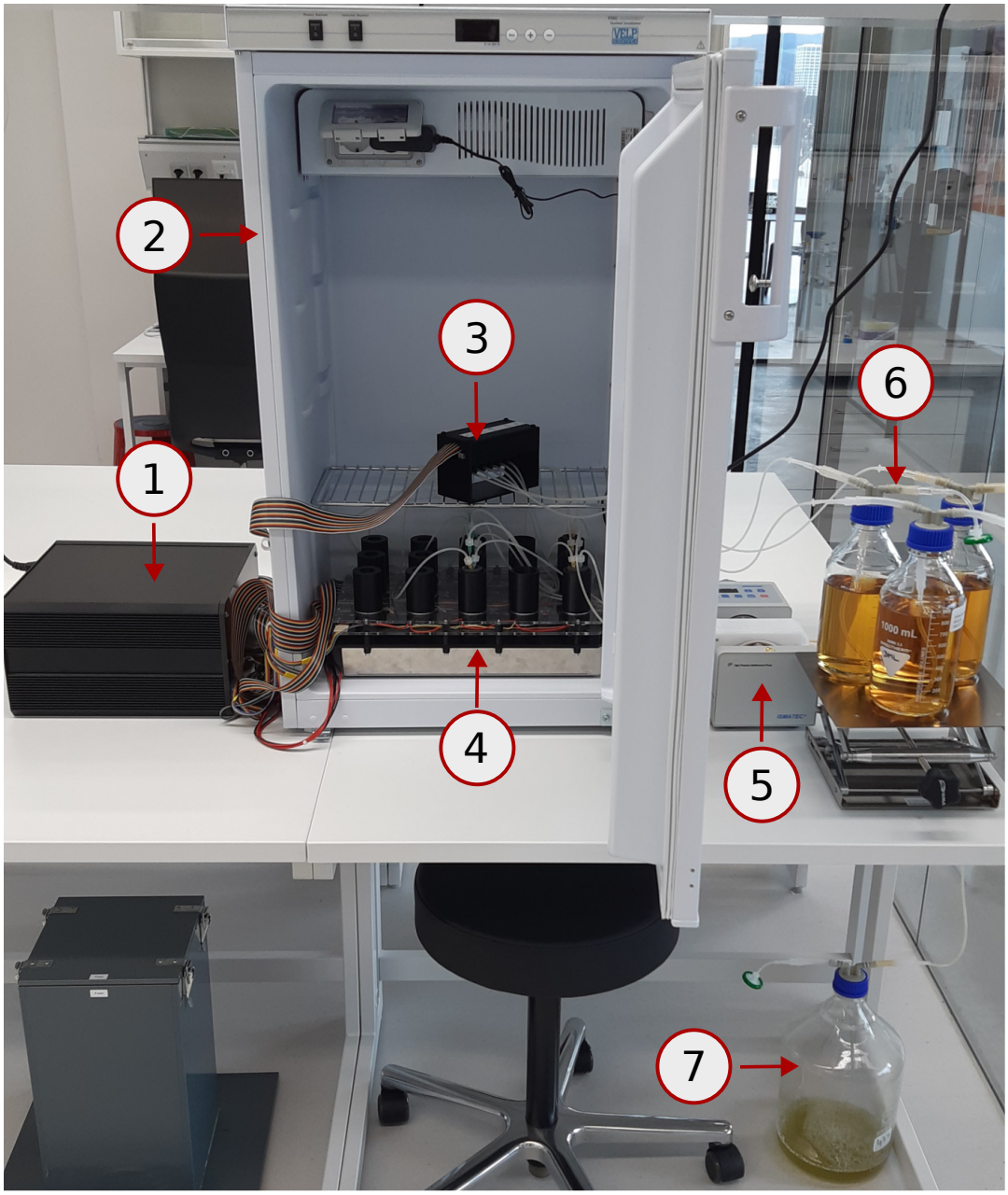


Figure 3.8: Picture of the Aionostat showing the setup for the experiments presented in section 3.5. For more details about the components refer to figure 3.10. **1.** Electronic components' enclosure. Contains the central computer and custom circuits for the electric components. **2**. Incubator for temperature control. **3**. Single channel piezoelectric pump array. **4**. Array of experiment vials on stirrer plate. **5**. Peristaltic exhaust pump. **6**. Input media bottle. **7**. Waste bottle.

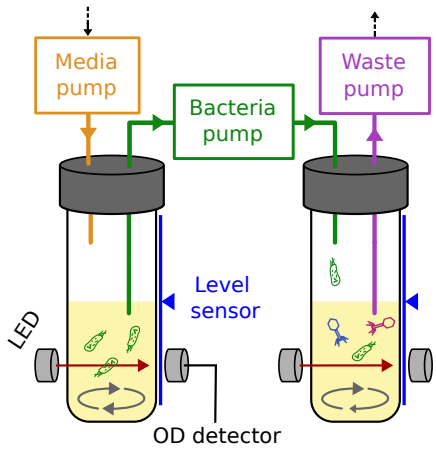


Figure 3.9: Schematic of the working principle of the Aionostat. One vial is used to grow bacteria in exponential phase. These bacteria are then continuously transferred to the second vial and get infected by phages, which can evolve over time.

The Aionostat operates similarly to a morbidostat [125], 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 [126, 127]. The most basic setup is a dual-vial system as shown in figure 3.9. The first vial maintains bacteria in their exponential growth phase without phages. These bacteria are then channeled to the second vial where they encounter and get infected by phages. This enables evolution of the phages over time. The constant dilution of the phage vial with new bacteria imposes a selective pressure on the phages, as the fitter bacteriophages will outcompete the other ones.

3.4.2 *Build*

The Aionostat is an autonomous continuous culture machine that has two main parts. The first is composed of the components that handle the liquids, the structure for such components as well as the pumps that move the liquids around. This part sits inside an incubator for temperature control. The second part act as the brain and power source for the machine, which can be programmed to perform a wide variety of experiments. Overall, the Aionostat was made from commercially available components, as well as custom 3D printed parts, electric circuits and wiring.

Vials

The experimental setup utilizes vials of two sizes, specifically 8 mL and 40 mL total volume, as depicted in Figure 3.10.B and 3.10.C. These vials are interchangeable depending on the experimental requirements.

For the experiments showcased in section 3.5, the larger vials were employed for both the bacterial and phage cultures. Alternatively, the smaller vials offer an option to intensify selective pressure on the phages. Utilizing a reduced volume for the phage culture effectively increases the dilution rate, which enhances selective pressure. This increased pressure can accelerate the sweeping of beneficial mutations, potentially leading to more rapid phage evolution.

Each vial is equipped with a magnetic stir bar to ensure consistent mixing. The vials are sealed with an open cap, fitted with a PTFE-coated silicon septum. This design allows for the sterile transfer of liquids in and out of the vials using needles and tubing. These components are shown in Figure 3.10.A.

Vial holders and magnetic stirrer plate

The vials are placed within custom 3D-printed holders. These holders not only secure the vials but also position the OD and liquid level sensors in close proximity to the vials, as illustrated in Figure 3.10.E. The vial holders, along with their respective vials, are positioned on a 15 position magnetic multistirrer. They are held in place using acrylic panels crafted via laser cutting and are assembled with screws and 3D-printed spacers, as depicted in Figure 3.10.I. It is in this vials that the experimental evolution happens over time. This is the central part of the Aionostat, label 4 in figure 3.8.

Liquid handling

Liquid transfer into and out of the vials is facilitated by commercially available needles, which go through the silicon septum. These needles are connected on one end to silicon tubing using Luer connectors, and to piezoelectric pumps on the other end. These pumps offer more control over the flow of liquid in each tube and are more compact than single channel peristaltic pumps. This gives great versatility to the experiments that can be performed with the Aionostat. The pumps are organized in custom 3D-printed arrays, available in various configurations (5, 8, or 15 pumps), ensuring stable positioning of the pumps and separation of electrical connections from the tubing and potential leaks. The 5 pumps version is shown in Figure 3.10.G and 3.10.H. The downside of the piezoelectric pumps is that, unlike peristaltic ones, they are sensible to the pressure difference in the tubes. This means that they cannot pump liquid with more than a 50-100cm difference between inlet and outlet tube, and that their flow rate varies depending on this difference in height.

Liquid from the input solution is pumped from the sterile bottles sitting outside of the incubator using pass through caps connected to the tubing and the piezoelectric pumps (Figure 3.10.F). Additionally, a 15-channel peristaltic pump is used as exhaust and overflow protection

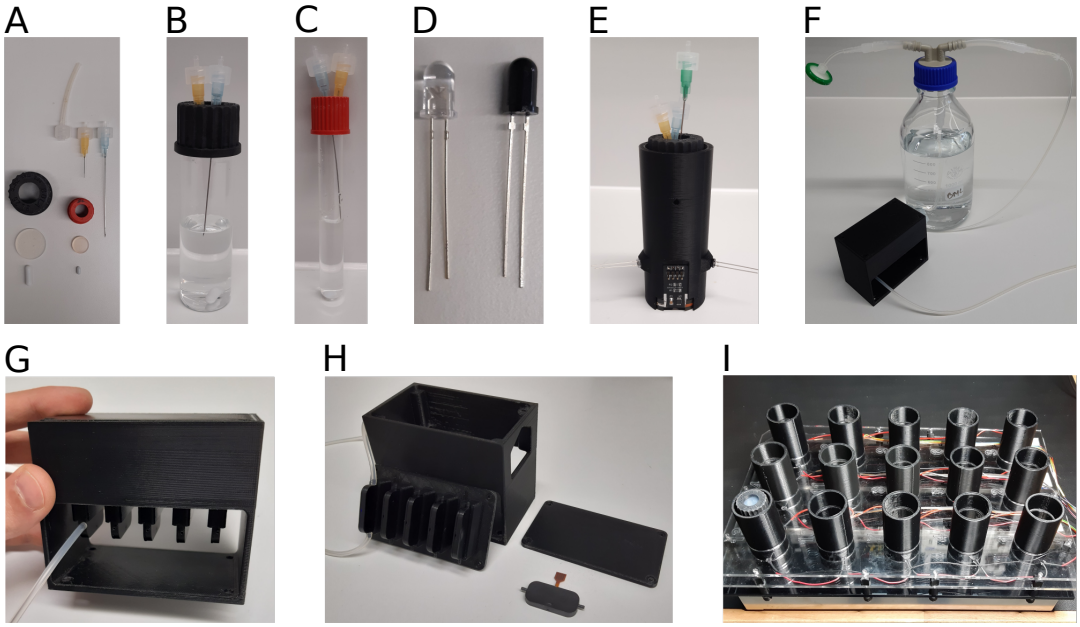


Figure 3.10: Components of the Aionostat. **A**: Needles, silicon tubing, Luer connector, open vial caps, silicon septum and magnetic stir bars used in the assembly of the vials. **B**: Big vial assembly (40mL). **C**: Small vial assembly (8mL). **D**: LED and phototransistor for the bacterial density measurement. **E**: Full assembly of one vial in its vial holder. **F**: Pump connection to the input bottle. **G-H**: piezoelectric pump array and housing (5 pumps version). **I**: Vial holders positioned on the magnetic stirring plate.

for the vials. The depth of the needle attached to the peristaltic pump is what sets the working volume in the vials. We used 60mm needles, which set the working volume of the vials to half of their total volume. This pump sits outside of the incubator.

The flow rate of the piezoelectric and peristaltic pumps is calibrated before each experiment as explained in the protocols section.

Optical density measurement

Bacterial density within the vials is assessed using an optical setup involving an LED and a phototransistor positioned on opposite sides of the vial at a 135° angle. This configuration allows for the detection of light diffracted by bacteria within the vial, rather than relying on direct absorption, thus enhancing sensitivity at low bacterial densities. The 135° angle between the light source and detector is the optimal angle to measure maximum diffraction. To accurately translate the phototransistor's signal into optical density values, calibration against standards with known optical densities is essential. The detailed calibration procedure is outlined in the protocols section.

The LED and phototransistor are positioned using designated holes on the sides of the vial holders as shown in Figure 3.10.E. The specific models used are the MT5880-IR LED from Marktech Optoelectronics

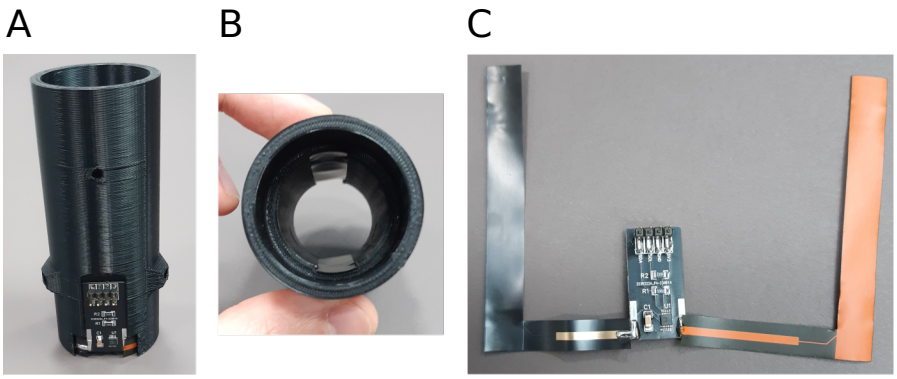


Figure 3.11: Custom capacitive liquid level sensor. **A**: Vial holder and capacitive sensor custom PCB. **B**: Sensing electrodes inside the vial holder. **C**: PCB and electrodes outside of the vial holder.

and the SFH 300 FA-3/4 phototransistor from ams OSRAM (Figure 3.10.D). Several combination of LEDs and phototransistors were tested, this particular pairing was found to offer the best dynamic range and signal-to-noise ratio for our experiments.

Liquid level sensing

To enhance control and safety in liquid handling within the Aionostat, custom capacitive sensors have been developed for monitoring the liquid level in the vials. One vial holder and its sensor is shown in figure 3.11. The electrodes of this sensor are positioned in the inner section of the vial holders, ensuring direct contact with the vials as shown in 3.11.B. Given the fixed diameter of the vials, the volume of liquid present can be accurately determined from the readings provided by these sensors. The addition of these sensors is one of the main improvements relative to other continuous culture devices such as the one presented in [125]. The ability to measure the volume in vials enables experiments at variable volumes and greatly improves the reliability of the machine. It also makes the calibration of the pumps much easier. The experiments presented in section 3.5 were performed at constant volume, so we used these sensors primarily as a mean of overflow protection.

The controller

The Aionostat is controlled using electronic components that are outside of the incubator and connected to the components inside via wires. Its central processing unit is a Raspberry Pi 4B board, equipped with HATs for analog-to-digital conversion of voltage readings and additional GPIOs. Custom Python scripts were developed to automate the experimental procedures and are accessible at `https://github.com/vdruelle/Morbidostat_phage`. The analog to digital readings are used to measure the sensors' signal, which provides

feedback on the state of the experiment. The GPIOs control the activation of piezoelectric and peristaltic pumps, based on the feedback from the sensors and the ongoing experimental requirements. The pumps have a constant flow rate, so the volume pumped is controlled by the time the pump is running.

3.4.3 *Protocols*

Initial tests and calibrations

Calibration of the Aionostat's components is a critical pre-experimental step. Firstly, the optical density sensors are calibrated. To perform the calibration, vials are prepared with bacterial dilutions of different optical density measured externally on a spectrophotometer. We also include a vial with raw media to cover optical densities between 0 and 1. Each vial holder is tested sequentially with these OD standards, recording the phototransistor's voltage output. A linear fit between these voltages and OD values is computed and saved for each vial holder. During experiments, these fits are used to deduce the OD from the sensor signals. Each vial holder's calibration accounts for component variability.

We continue with the calibration of the level sensors. This process involves recording sensor voltages with vials at varying liquid levels: empty, full, and intermediate volumes. The sensor readings and level of the liquid have a linear relationship, enabling the interpolation of liquid height through a linear fit.

Finally we perform the calibration of the piezoelectric and peristaltic pumps. Given their constant flow rate, calibration involves running the pumps for a set duration and measuring the output volume. This can be done directly in the vials using the readings from the level sensors, or simply by weighing the vials on a scale. The flow rate is determined by dividing the volume by the time. These rates also enable calculation of the dilution rates in the vials by factoring the working volume in the vials. The flow rate of piezoelectric pumps is impacted by the pressure in the tubes, so it is recommended to perform this calibration with a configuration similar to the one that will be used for the experiment.

Sterilisation of the Aionostat

Sterilization of the Aionostat is conducted post-calibration and pre-experiment to prevent contamination. This is achieved in two stages. First, all tubing, pass-through caps, and vials are washed and then autoclaved at 120°C for 20 minutes. The vials are assembled and sealed as depicted in Figures 3.10.B and 3.10.C, while the tubing and caps are wrapped in aluminum foil prior to autoclaving. These components are

then installed in the morbidostat, and the tubing, needles, and pumps are connected as they will be during the experiment.

The second stage involves chemical sterilization using 3% sodium hypochlorite (bleach) and 3% citric acid solutions, applied sequentially throughout the entire setup. Beginning with the sodium hypochlorite solution, all pumps are activated for one minute, five times sequentially over a 30-minute period. The vials are then emptied using the pumps, and this procedure is repeated with the citric acid solution. After this, the vials are emptied again, and the system is rinsed with MiliQ water, which is run through the vials and tubing. This process is automated, requiring manual input only for changing the input bottles.

At this point the inside of the Aionostat is sterile, and it is crucial to not disconnect any tubing.

Setting up

The next step is programming the Aionostat for the experiment, which involves using and modifying the pre-written control code. This can also be performed before or during the previous steps.

After programming and ensuring the Aionostat is set up and sterile, the MiliQ water in the vials is replaced with fresh sterile media. This replacement is carried out by changing the input bottle (near a flame for sterility) with fresh sterile media and using the pumps to exchange the liquid in the vials. Lysogenic broth was used in the experiments presented in section 3.5.

The final preparation step is the inoculation of the sterile media with bacteria and bacteriophages. This is done manually using a syringe and needle to pierce the septum and introduce the appropriate bacterial strain or bacteriophage into each vial. We finish by closing the incubator and setting it to the desired temperature. Typically, the entire preparation of the Aionostat takes 4 to 8 hours.

Running the experiment

At this stage the experiment is fully prepared, and we start the run by launching the code for the experiment from the Rasberry Pi. This can be done directly from the board, or more conveniently via remote connection to the Raspberry Pi by SSH. Once started, the program provides real-time updates on the experiment's progress, pump actions, and sensor readings, allowing for monitoring to ensure the experiment's smooth operation. The sensor data is also recorded and saved over time.

During the following days, routine maintenance involves changing the input media bottles before depletion and replacing the waste bottle when it is full. A 5-liter bottle pre-filled with some disinfectant, is used as the waste container. Manual sampling from the vials is

conducted using a syringe and needle, piercing through the septum for sample collection. It is recommended to remove bacteria from phage samples before storage. Regular checks on the experiment's status are also recommended, for example when changing media or collecting samples.

Upon completion of the experiment, the vials are emptied and disassembled, and the setup can be prepared for the next experiment following the previously outlined steps.

3.5 EVOLUTION EXPERIMENTS

3.5.1 *Overview*

In this section we present the two experiments performed with the Aionostat. These experiments are meant to showcase the abilities of the Aionostat and demonstrate that it can be used to efficiently evolve bacteriophages. The principle of these experiments is illustrated in figure 3.12. In these experiments, we used 3 bacteriophages from the BASEL phage collection presented in 3.2. The 3 bacteriophages used are phage WalterGehring (bas51, NCBI GenBank accession MZ501111.1), phage MaxBurger (bas54, accessionMZ501093.1), and phage PaulScherrer (bas60, accession MZ501100.1). These bacteriophages are well-adapted to their isolation strain E.coli K12 BW25113, the parental strain from the Keio collection [128].

To assess their directed evolution against a novel and more challenging strain, we used *E.coli* K12 BW25113 *wbbl(+)* as a model [121]. This particular strain is a derivative of BW25113 which has a restored O16-type O-antigen glycan barrier [122], they are otherwise genetically similar. The restored O-antigen adds long chains on the lipopolysaccharide (LPS), which acts as a protective barrier on the bacterial surface by shielding the cell surface as shown in Figure 3.12. This inhibits infection from bacteriophages that do not bind the LPS or other glycans [121]. The phages used in this study are impaired by the O-antigen, but infection is not completely inhibited. This is likely due to their ability to bind another surface glycan [129].

The goal of these two experiments was to evolve phages for better infectivity on *E.coli* K12 BW25113 *wbbl(+)*. In the first experiment, a linear evolution approach was used. The phages were evolved in separate vials for better infectivity on a challenging E.coli strain, resulting in evolved phages which fitness was measured and compared to their ancestors. Second, a phage "cocktail" experiment where these phages were first mixed and then evolved on the same challenging E.coli strain, which resulted in the appearance of recombinant phages.

Both of these experiment necessitated half a day of work to prepare and launch, and around 15 min of work a day in the subsequent days to collect samples from the vials and refill the media bottles. Samples

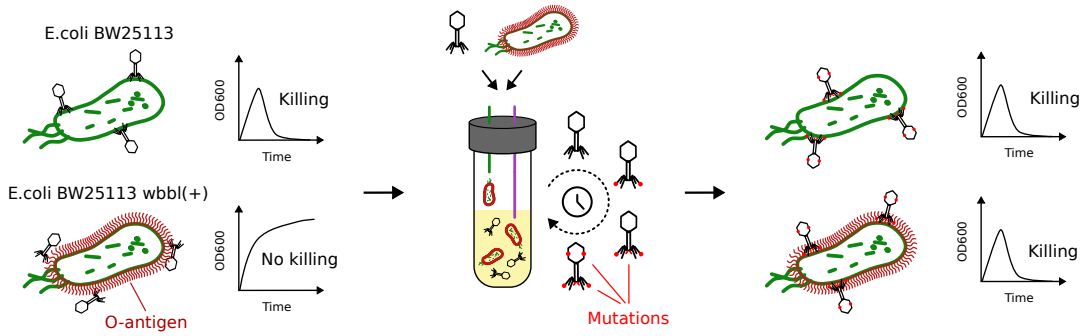


Figure 3.12: Schematic of evolution experiments. Originally the phage does not infect the strain with restored O-antigen well, which does not causes significant bacterial killing when the bacteria are grown in presence of the phage. After evolution using the Aionostat, the evolved phages infect both the ancestor and the O-antigen strain, causing significant bacterial killing.

taken from the bacterial cultures were streaked on bacterial lawn, showing that the cultures stay phage free for the whole duration of the experiment. The phage population size remained stable over time, and no cross contamination was observed between the phage samples. For both experiments we present results for phage bas51 and bas54 only as no significant evolution was observed for phage bas60.

3.5.2 *Linear evolution experiment*

The experiment was conducted using the Aionostat, as depicted in schematic Figure 3.13, over a duration of five days. Each phage vial was paired with a culture vial of *E.coli* BW25113 *wbbl(+)* kept in exponential phase in lysogeny broth medium (LB) at 37°C as shown on the schematic 3.9. Vials were seeded from bacterial and phage stocks right before the start of the experiment. The bacterial culture's dilution rate with raw LB was adjusted to maintain a constant optical density at 600nm (OD600) of 0.5. The excess liquid resulting from the bacterial culture dilution was transferred to the phage vial, where infection and replication of the phages occur. This transfer of bacterial culture to the phage vial provides fresh bacteria for infection, while the volume in the vial is maintained by discarding any surplus. Consequently, the phage solution becomes more diluted over time, which selects for bacteriophages with higher fitness.

Throughout the experiment, daily samples were extracted from each vial. To prepare these phage samples for storage and later analysis, bacteriophage population samples were cleared of bacteria. The ancestor phages, along with the phage populations from day 1, 3, 5 and phage clones from the day 5 populations, were sequenced using in-house Nanopore sequencing as detailed in section 3.5.4. Genomic changes were tracked over the experiment's duration using the sequencing data. Lastly, to discern differences in phage fitness, turbidity based killing

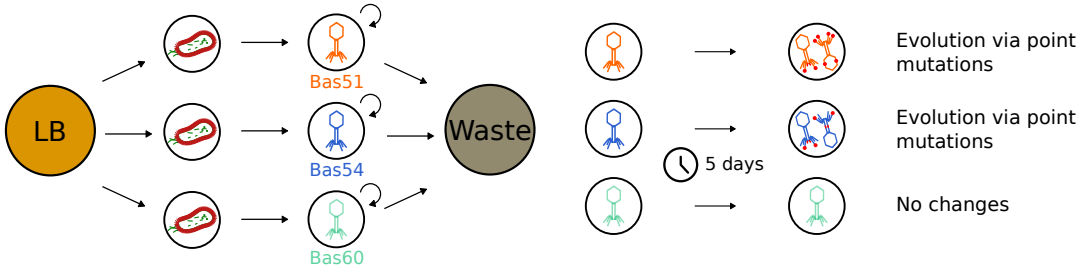


Figure 3.13: Schematic illustration of the linear evoltution experiment. Three phages were evolved in parallel, in separate vials, on a challenging *E.coli* strain with O-antigen restored. The experiment lasted for 5 days. Samples of the phage population were taken once a day.

curves of both *E.coli* K12 BW25113 and *E.coli* K12 BW25113 *wbbl(+)* were done with the ancestral phages, evolved phage populations and isolated clones. Differences in phage fitness cause different killing dynamics, enabling phenotypic comparison between the phages.

A copy of this experiment has been done by hand using daily serial transfer to provide comparison to a more established approach. Details about the methodology are described in section 3.5.4.

Results

Firstly, we focus on the phenotypic changes in the evolved phages. Figure 3.14 shows the killing curves of the bacterial strains used in the experiment by the ancestor and evolved phages. When observing the interaction with the isolation bacteria, both the ancestor (dashed blue line) and evolved phages (solid blue line) show a sharp decline in OD600 values around the 3-hour mark, going down to the detection limit. This suggests that both ancestor and evolved phages kill the isolation bacteria at comparable rates.

In contrast, when these phages interact with the *wbbl(+)* strain, differences between the ancestral and evolved phages become apparent. The evolved phages (solid orange lines), cause a steeper decline in OD values between 3 to 7 hours than their ancestral counterparts (dashed orange lines). This indicates that the evolved versions of both bas51 and bas54 are more efficient at killing the *wbbl(+)* bacteria than their predecessors.

In summary, while the evolved bas51 and bas54 populations maintain similar killing rates as their ancestors on *E.coli* BW25113, they show enhanced efficiency against the *wbbl(+)* strain. The killing curves were also performed with clones isolated from the evolved population as shown in figure 3.15. The results are similar to the ones observed with the phage population, suggesting that phage mutants have the ability to infect both bacterial strains, and that this is not an effect of the phage diversity present in the evolved phage population.

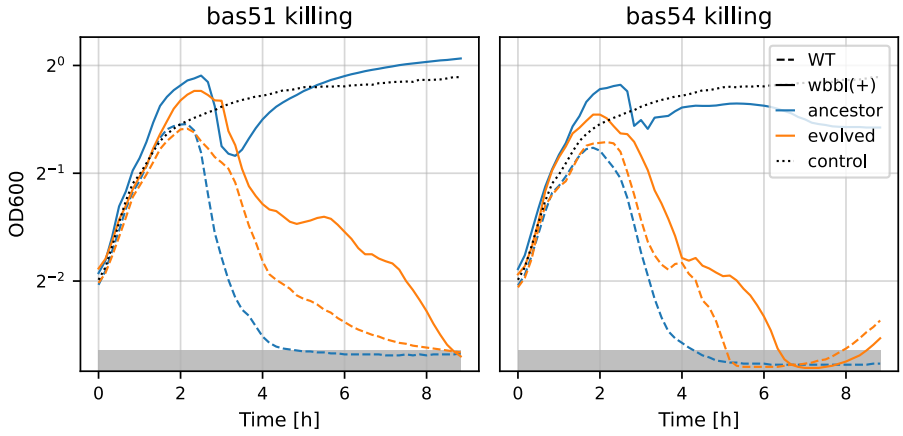


Figure 3.14: Killing curve of ancestor and end population of evolved phages on *E.coli* K12 BW25113 (denominated WT) and *E.coli* K12 BW25113 *wbbl(+)*. Evolved phages kill *wbbl(+)* better while retaining their ability to kill WT.

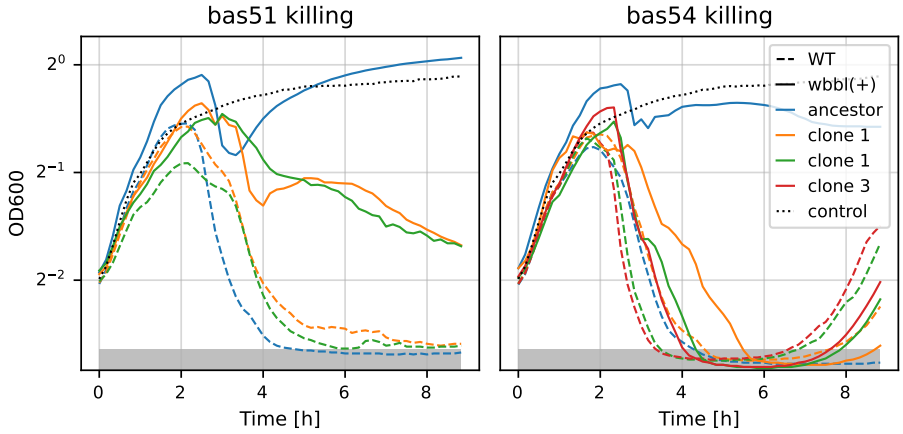


Figure 3.15: Killing curves of phage clones. The full blue lines are the same as in 3.14. The killing of the *wbbl(+)* strain is better than the ancestor phages for all clones, but clones from bas54 seem to be better in this regard like what is seen for the population killing curves. Killing on WT is still comparable to the ancestor phages.

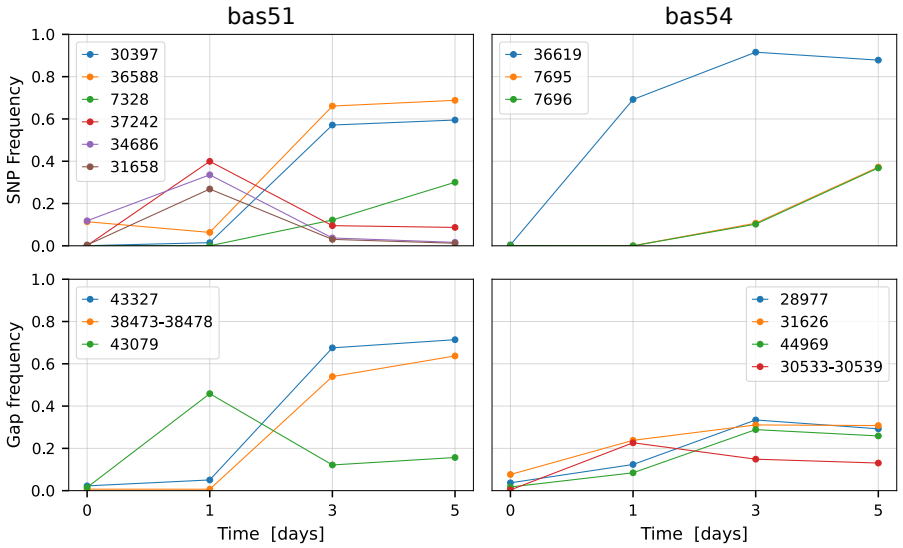


Figure 3.16: Frequency in time of genomic changes observed in the phage populations.

Secondly, we analyze the genomic changes from this evolution experiment. Figure 3.16 shows the frequency trajectories of single nucleotide polymorphisms (SNPs) and gaps for the evolved phage populations over the span of the experiment. Observing the bas51 panels, distinct trajectories of several SNPs are evident. Notably, the SNPs at position 30397 (lateral tail fiber), 36588 (lateral tail fiber with fibronectin type 3 domain), and 7328 (major capsid protein) exhibit sweeping behaviour over time, while some other mutations seem to appear earlier but then disappear. In terms of gaps it seems that a gap in position 43327 (lateral tail fiber with intimin domain) and 38473-38478 (putative protein) show sweeping behaviour.

Transitioning to the bas54 panels, the SNP at position 36619 (lateral tail fiber with glycosidase and deacetylase domains) stands out, reaching near fixation by day 5. Mutations in position 7695 and 7696 (putative protein) increase over time as well, but seem to have appeared later in the population. In terms of gaps there does not seem to be a clear pattern for bas54. We see some variants with gaps in their genome, but only at low frequency.

Additionally, for both phages, we see the appearance of a phage subpopulation with low frequency of around 5% that have a big deletion in their genome by day 5. This is from position 43300 to 51800 for bas51 and from position 46680 to 50600 for bas54. There seem to be a bit of diversity on the exact position of this deletion. This subpopulation of phages appears early in the experiment but does not seem to take over the population. No other big rearrangements are observed in the phage genomes.

Overall it is clear from Figure 3.16 that what is originally a clonal population of phages at the start of the experiment diversifies in the

phage	clone	position	mutation	gene
bas51	1, 2	**30397**	SNP	Lateral tail fiber
		36588	SNP	Lateral tail fiber with fibronectin type 3 domain
		43327	Gap	Lateral tail fiber with intimin domain
		38473-38478	Gap	Putative protein
bas54	1, 2, 3	**36619**	SNP	Lateral tail fiber with glycosidase and deacetylase domains
bas54	2, 3	**30533-30539**	Gap	Lateral fail fiber with fibronectin type 3 domain
bas54	1	43080	SNP	Lateral tail fiber with intimin domain
bas54	1	**28977**	Gap	Lateral tail fiber with fibronectin type 3 domain
bas54	1	105486-105492	Gap	Putative N6 adenine methyltransferase
bas54	2	**7695**	SNP	Putative protein
bas54	3	**7696**	SNP	Putative protein

Table 3.1: Mutations of the clones extracted from the phage populations shown in Figure 3.16. Positions in bold are the sites where a mutations was also observed from the population sequencing. Killing curves of these clones are shown in Figure 3.15.

following days. We do see a fair number of mutations rising and then decreasing in frequency, which suggests that the mutations providing the most increase in fitness take a while to appear. Interestingly, we also see that newly acquired mutations do not seem to fix completely, keeping some diversity in the population. Phage clones were also isolated from these populations. Unsurprisingly, most of the phages picked are genetically identical, and have the mutations seen at high frequency in Figure 3.16. Details about the phage clones' mutations can be found in table 3.1.

Linking these observations to the killing curves in Figure 3.14, it can be assumed that the genomic changes observed play a pivotal role in the phages' enhanced capability to kill the *wbbl(+)* strain. Mutations seen may offer insight into genetic changes that confer advantages to these phages, allowing them to efficiently combat both the isolation bacteria and the *wbbl(+)* strain. A good portion of the mutations observed are focused on the tail fibers of the phages, which 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. The phages are likely well adapted to their isolation host, hence the only challenge that they would face with the *wbbl(+)* strain would be linked to binding and absorption since the bacteria are otherwise genetically identical.

Further in-depth genetic analysis could be pursued to pinpoint the exact role these mutations in the phages' improved predation capacities and confirm that it is linked to absorption efficiency on the bacteria. The goal of this experiment being to showcase the ability of the Aionostat, we leave the molecular biology details for future work.

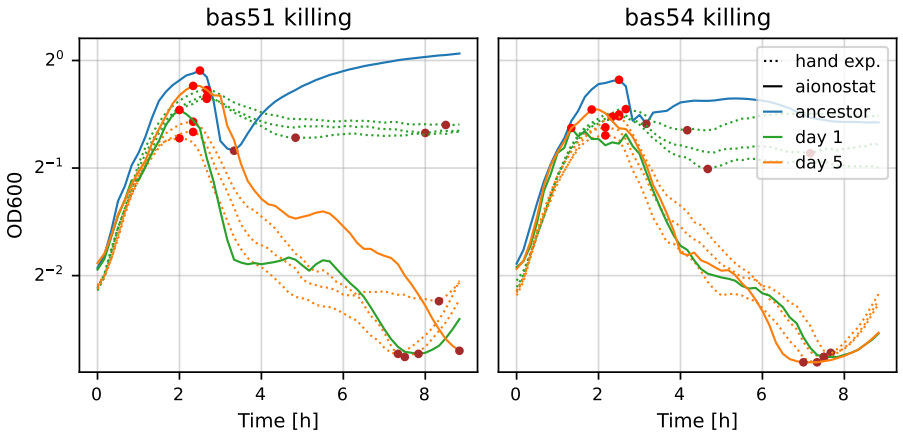


Figure 3.17: Comparison of killing curves for the linear evolution experiment made by hand VS made with the Aionostat. The full blue lines and full orange lines are the same as in 3.14. Red dots highlight the depth of the bacterial crash. Overall the killing of the *wbbl(+)* strain is good in all cases after 5 days of evolution (orange curves). The Aionostat phage population already kills well after 1 day of evolution, which is not the case for the hand made experiment (green curves). Ultimately phages from the manual experiment end up with some of the same mutations as observed in the Aionostat experiment, suggesting that the Aionostat is faster at exploring the mutational landscape to find the highest fitness mutations.

Comparison with manual serial dilution evolution

The evolution experiment performed as described in section 3.5.4 shows similar results to the Aionostat one. The evolved phages kill BW25113 *wbbl(+)* like is observed for the ones evolved with the Aionostat, see Figure 3.17. One main difference observed is that the phages evolved using the Aionostat evolve better killing on BW25113 *wbbl(+)* faster, as can appreciated by the crash depths of the curves. One can see that after one day of evolution, the phages evolved with the Aiononstat already have improved killing on *wbbl(+)*, while it is unclear for the phages evolved manually. Eventually they all achieve strong killing at day 5, showing that we get similar results in both cases, but that the Aionostat seems to promote faster evolution. The mutations observed in the hand evolution experiment are also mostly on the genes linked to phage absorption as shown in Table 3.2. Interestingly, the jump in killing efficiency seen between day 1 and day 5 seems to be linked to the appearance of mutations at the same locus as seen in the Aionostat experiment for phage bas51.

3.5.3 *Recombination experiment*

In this section we present the phage cocktail experiment that we performed with the Aionostat. It provides a proof of concept for phage

phage	replicate	position	mutation	gene
bas51	A	65786	SNP	Non-coding
		33761	SNP	baseplate protein
		100460	SNP	hypothetical protein
		32116	SNP	tail fiber with fibronectin type III domain
		36588	SNP	tail fiber with glycosidase and deacetylase domains
		31658	SNP	Lateral tail fiber with fibronectin type III domain
bas51	B	37767	SNP	Short tail fiber protein
		36588	SNP	Lateral tail fiber with fibronectin type III domain
		52820	SNP	putative metallopeptidase
		82858-82938	gap	non coding
		31900	insertion	Lateral tail fiber with fibronectin type III domain
bas51	C	1568	SNP	putative endonuclease
		34869	SNP	tail fiber with glycosidase and deacetylase domains
bas54	A	57616	SNP	ATPase
		29163	Gap	Lateral tail fiber with fibronectin type III domain
bas54	B	129548	SNP	sirtuin domain NAD-dependent deacetylase
		29163	Gap	Lateral tail fiber with fibronectin type III domain
bas54	C	99437	SNP	hypothetical protein
bas60	A	27327	SNP	baseplate protein
bas60	B	-	-	-

Table 3.2: Genomic changes observed in the linear evolution experiment done by hand. Positions in bold are the sites where a mutations was also observed in the experiment made with the Aionostat.

directed evolution through recombination using the Aionostat, a topic that has been introduced in 3.1.3. This experiment and its outcomes are illustrated in figure 3.18. The experimental parameters are the same as for the linear evolution experiment, with the exception of the initial content of the vials and the duration of the experiment. For the experiment a mix of equal amount of the 3 phages is used to seed two phage vials at the start of the experiment. The third vial served as a negative control, aiming to validate that it remained phage-free throughout the experiment.

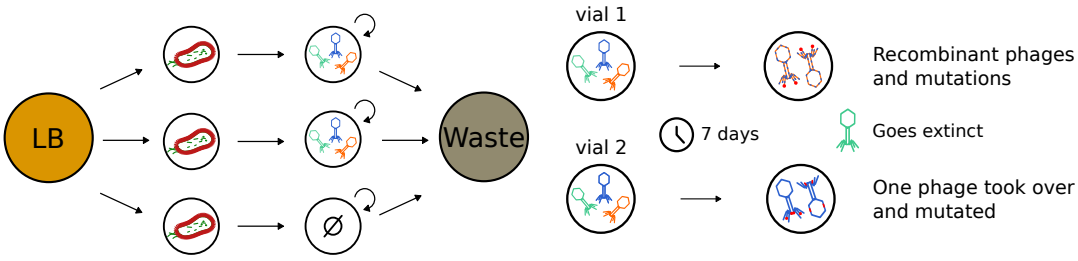


Figure 3.18: Schematic of the recombination experiment. The three phages were mixed in equal amount and then spread in two evolution phage vials. This cocktail is evolved on E.coli BW25113 wbbl(+). Samples of the phage population were taken once a day. After 7 days of evolution, one of the vials was overtaken by recombinant phages, while the other one was overtaken by one of the ancestral phages with additional mutations.

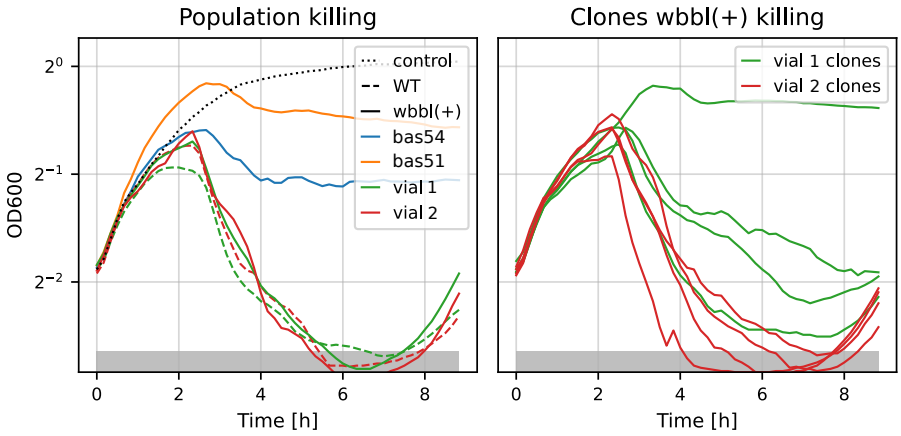


Figure 3.19: Killing curve of ancestors, evolved populations and evolved phage clones (day 7) on *E.coli* K12 BW25113 (denominated WT) and *E.coli* K12 BW25113 *wbbl(+)*. Evolved population kill *wbbl(+)* better in both cases and so is the case for individual clones of vial 2. This is not the case for all clones isolated from vial 1.

Results

Figure 3.18 summarizes the results of this experiment. We observed the rapid extinction of phage bas60 in both vials. In vial 1, the final phage population was composed of evolved recombinants of bas51 and bas54, along with several point mutations. A different scenario happened in vial 2, bas51 took over the phage population, out-competing the two other phages. The final phages are evolved version of bas51 with some mutations, similar to the ones observed in the linear evolution experiment discussed above.

Figure 3.19 shows the phenotypic differences in bacterial killing at MOI 1:1000 between the evolved phage populations and their ancestors. Once again we observe that, unlike the ancestor phages, the evolved phage populations cause a decline in bacterial density around the 4 hour mark on both BW25110 and the *wbbl(+)* mutant (green and red lines). Vial 3 clones (red) showed killing efficiencies on *wbbl(+)* comparable to the whole evolved population. However, the recombinant phages from vial 1 (green) showed superior killing of *wbbl(+)* compared to their ancestors, but not as effectively as the entire evolved population. There are two main reasons why this might be the case. We could have, by chance, missed the best phages in the population while picking and isolating clones. Another possibility is that phages in the evolved population have "specialized" in different ways, and that they kill best when together due to synergistic effects. Notably, all clones isolated from vial 1 were recombinant, suggesting a fitness gain from recombination that allowed them to outcompete non-recombinant phages.

Focusing on the genomic alterations, vial 2 exhibited evolution patterns similar to the linear evolution experiment. This is expected

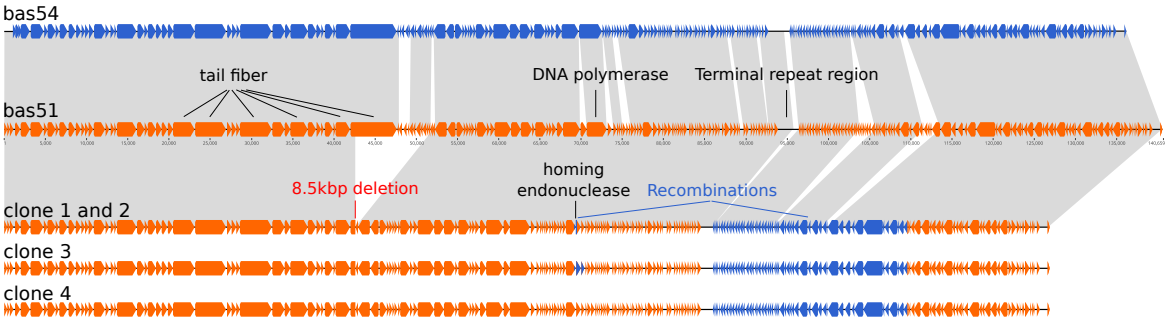


Figure 3.20: Genome comparison of ancestral phages and evolved recombinant phages isolated from vial 1. The genome structure of the evolved phages is mostly identical between clones, except for the region of the homing endonuclease. They differ in a few point mutations.

considering one phage took over the population and then evolved linearly. Consequently focus our attention on vial 1. The evolved phages present in this vial are recombinant phages between bas51 and bas54 as shown in figure 3.20. The main recombination event is 25kbp of bas54 that got inserted into bas51 genome. This starts at the terminal repeat region and ends in the middle of a gene 25kbp later and likely happened as shown in Figure 3.21. This region primarily contains hypothetical proteins, making the fitness benefits of this recombination unclear, but it must have been beneficial enough to out-compete the other phages.

Additional genomic changes in vial 1 clones include a secondary, smaller recombination event involving a homing endonuclease, which likely jumped from bas54 to bas51 during co-infection. Last is a big deletion of 8.5kpb in bas51's genome, which was seen in all clones of the first vial and 2 out of 4 clones in the second vial. This deletion starts in a lateral tail fiber protein and covers 12 hypothetical proteins after that. About 3/4th of the lateral tail fiber gene is deleted, which likely completely stops its function. A large deletion in the same area was also observed in the linear evolution experiment at a smaller frequency, which means it has evolved in a convergent manner several times. This suggests it hinders infection of the *wbbl(+)* strain and was lost. The evolved recombinant phages from the first vial have genomes approximately 10% smaller than their ancestors, which might also contribute to their improved fitness. Point mutations similar to those in the linear evolution experiment likely also contribute to the improved bacterial killing efficiency.

Further analysis could be done to understand the role of the genomic changes and their effect on killing efficiency, but we leave the molecular biology details for future work as the goal of this experiment was to showcase the ability of the Aionostat.

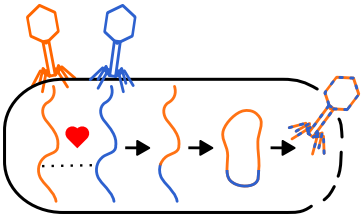


Figure 3.21: Schematic of the recombination process that happened in vial 1. Since there is a transition from bas51 to bas54 genes over the terminal repeat region it is likely that the recombination happened while genomes where in a linear state, causing only one cut.

3.5.4 *Material and methods*

Manual linear phage evolution experiment

This experiment is similar to the experiment done with the Aionostat presented in the main text but done by hand, to provide a comparison with a more standard approach to phage evolution experiments using daily serial passages for 5 days. The experiment was performed with the same phages: bas51, bas54 and bas60 in triplicates. Only the results of bas51 and bas54 are presented as not much evolution was observed for bas60 (like in the Aionostat experiment).

On the first morning, 5mL of LB in glass tubes were seeded with 1µL *E.coli* BW25113 *wbbl(+)* from an overnight culture and infected with the respective phages at an MOI of 1:100, allowing for several replication cycles before saturation in the vials. The cultures were then grown on a shaking incubator at 37°C for 8 hours, which was shown to be enough time for the culture to either lyse or saturate. It was decided to limit the growth to 8 hours to avoid evolutionary dynamics that would involve the rise of resistant bacteria and could cause bias in the results.

After 8 hours, 1mL samples were taken from each tube and cleared of bacteria using 1% chloroform plus strong vortexing, followed by a 2 minutes spin at 20 000g. The supernatant was extracted and stored at 4°C to seed the cultures on the next morning and later the analysis of the phages at that time point.

The next morning, 10µL of a 10^5 dilution of the supernatant was used to infect the daily cultures prepared as was done on the first day. The amount of supernatant used ensured between 100 and 10 000 phages are transferred to the vial, preventing the extinction of the phages by over diluting while keeping the MOI low. These serial passages were repeated every day until the end of the experiment.

Measuring phage concentration

Bacteriophage concentration in solution were measured using the serial dilution spotting method. This process starts with the preparation of top agar bacterial lawn. To do so, round LB agar plates (9.4cm diameter) were overlayed with top agar (LB agar containing only 0.5% agar) supplemented with 100µL of bacteria. While the top agar is solidifying, 10-fold serial dilutions of the bacteriophage solutions were prepared in a 96-well plate using phosphate-buffered saline (PBS), resulting in dilutions 10 to 10^8 of the original phage solution.

Then, 2.5µL of each dilutions are spotted on the top agar lawn and left to dry by the flame for 10 minutes. The plate is then moved to a 37°C incubator for 4 hours. The titer of the solution is inferred by counting the plaques in the bacterial lawn and factoring the dilution factor. This whole process was done with both *E.coli* BW25113 and *E.coli* BW25113 *wbbl(+)* to ensure there are no discrepancies between the two.

Phage amplification

The samples taken from the evolution experiment usually contained between 10^6 and 10^9 PFU/mL. When more phages were needed, like in the case of DNA extraction for sequencing, the phage samples were amplified in liquid culture. The amplification step was designed to minimize bias from the original sample by using an initially high amount of phages and a short incubation time to limit the number of replication rounds.

For each amplified phage stock, tubes were inoculated with 1mL LB and 300µL of *E.coli* BW25113 *wbbl(+)* from overnight culture and then put for 20min at 37°C 600RPM to restart the growth of bacteria. Subsequently, 100µL of the phage sample to amplify was added to the tubes and then incubated for 3 hours at 37°C 600RPM. The tubes were then cleared of bacteria by adding 1% chloroform, vortexing and spinning the tubes for 10min at 8000g. The surpernatant was extracted and titered, usually achieving between 10^{10} to 10^{12} PFU/mL. These amplified samples were stored at 4°C until they were used.

Phage geneomic DNA extraction

Genomic DNA of bacteriophages was prepared from high-titer stocks produced as explained above. 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 below.

Plate reader killing curves

Killing curves presented in Figures 3.14, 3.15 and 3.17 were generated using an Epoch2 plate reader in absorbance (OD600) mode. The phages were tested on both *E.coli* BW25113 and *E.coli* BW25113 *wbbl(+)* at a target multiplicity of infection of 1 to 1000.

Each well was prepared with 180µL of a bacterial dilution in LB of $5 \cdot 10^8$CFU/mL. 20µL of diluted phages with a concentration of $5 \cdot 10^6$PFU/mL was then added to their respective wells, achieving a final phage concentration of $5 \cdot 10^5$PFU/mL and a volume of 200µL in the wells. The phage stocks were titered on the same day as the experiment to ensure as much accuracy as possible, and then diluted in PBS to hit the target MOI of 1 to 1000. The phage dilutions used to prepare the plate were also titered right after the plate was loaded to the plate reader to ensure the MOI was correct.

Once prepared, the plate was then moved to the Epoch2 plate reader and run for 15 hours at 37°C degrees, which was long enough to observe bacteria killing and eventual regrowth of resistant bacteria. The experiment was performed with 450RMP double orbital rotation and OD600 readings every 10 minutes. The lid of the plate was removed, and replaced with an "easy breathe" membrane porous to dioxygen but not to water.

DNA sequencing

The DNA samples extracted as explained previously were sequenced in-house using the Oxford Nanopore sequencing technology. We utilized the MinION Mk1B device for sequencing, employing V14 chemistry coupled with R10.4.1 pores. The flow cells used in this procedure were of the type FLO-MIN114. To facilitate the sequencing, we utilized 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. The basecalling pipeline used is available here: `https://github.com/vdruelle/nanopore_basecalling`.

Sequencing analysis

The analysis of the sequencing data was performed using a Snakemake pipeline, which is publicly available at`https://github.com/mmolari/evo-genome-analysis`. This pipeline takes as input the raw reads from the samples and the reference genomes. It maps the reads from each sample to the references using Minimap2 [130] from which we extract trajectories of genomic changes over time, encompassing single nucleotide polymorphisms, gaps, insertions, clips, and rearrangements. These trajectories were then filtered and plotted as shown in Figure 3.16. These mutations were then manually inspected when additional information was needed.

Recombination were detected by mapping the recombinant phage genome's to the parental strain genomes and plotting the mutation density along the genome. Jumps in the mutation density clearly identified recombination regions. The breakpoint can only be inferred to about a 50-100bp region since the parental bacteriophages have high homology.

4

CONCLUSION AND OUTLOOK

Evolution is a fundamental force that shapes and guides the development of all living organisms, from the simplest organism to the most complex. Although this process of change and adaptation was first described in relation to the flora and fauna that populates our world by the work of Charles Darwin [131] and subsequent work, it has since been appreciated for the huge driving force it exerts of the microscopic world. Landmark studies such as those by Woese and Fox in 1977, which identified Archaea as a distinct form of life by studying 16S ribosomal RNA evolutionary relations [132], or work from Luria, Delbrück, Lederberg, Tatum, Beadle, Hayes and Zinder in the 1940s and 1950s, with the discovery of spontaneous genetic mutations and horizontal gene transfer in bacteria [133–136], have all contributed to showing the key role of evolutionary dynamics for life on our planet.

Since their discovery at the end of the 19th century through the work of Dmitri Ivanovsky [137], viruses have also been heavily studied for their ability to evolve. Although these biological entities straddle the line between life and non-life, their ability to replicate and mutate inside their hosts, passing down these genetic changes to their progeny, gives them the ability to evolve like prokaryotes and eukaryotes. The sheer abundance and diversity of viruses on Earth proves that viruses are master evolvers [71]. Their rapid replication rates and high mutation frequencies enable them to adapt rapidly to new environments and hosts. The evolutionary prowess of viruses, particularly evident in entities like HIV-1 and bacteriophages, is a defining trait of the viral lifestyle and has many implications in healthcare, ecology and biology. From the creation of effective vaccines to manage viral outbreaks, to the conception and improvement of bacteriophage therapy, understanding viral evolution is central to pushing the boundaries of modern biology and healthcare.

With this background in mind, the objectives of this thesis were the following (reported from section 1.5):

1. Study and characterize how HIV-1 evolves both intra-host and inter-host, and explain how the evolutionary dynamics at the pandemic level emerge from the peculiar evolution happening within-host.

2. Create a complete framework for high-throughput studies of bacteriophage evolution through directed evolution experiment. If successful, it will enable a better study and optimization of

bacteriophage evolution but also provide general insights about viral evolution dynamics such as recombination between viruses.

In this section we discuss the outcomes of the work performed in relation to to these two goals, their broader implications and limitations that could be addressed in the future.

4.1 HIV-1 bias for reversions and impact on its evolution

HIV-1 is a widespread virus which causes serious health burden around the world. The ability of HIV-1 to evolve has led to the diverse viral population that we see nowadays. Although it is widely recognized as a fast evolving virus, understanding the speed at which HIV-1 evolves and diversifies remains a challenging task. Different methods used to measure this rate often produce varying results [138–142]. This inconsistency indicates that our understanding of how within-host HIV-1 evolution translates to the evolution seen on the scale of the pandemic is incomplete.

In the case of HIV-1, evolution rate on a pandemic scale is about two to five times slower than within a host [143]. One can expect a saturation of evolution speed in the case of very diverged sequences, but HIV-1 sequences are too similar for that to contribute significantly. Recent HIV-1 research has proposed several other ideas to explain this mismatch. The two major theories are the "store and retrieve" hypothesis, which suggests that older variants of HIV-1 are more likely to be transmitted [144], and the "adapt and revert" hypothesis, which proposes that the virus quickly returns to a state similar to its original form after transmission [66, 145–149]. Both of these theories would "favor" older versions of the virus on the between-host scale, explaining why the evolution rate observed would be smaller. However, the exact impact of these and possibly other factors on the differences in estimated rates of evolution is still not well understood [150].

It is in an effort to shed light on this topic that we have performed and published the research presented in chapter 2. For this work, we used both between-host sequencing data covering many years of the pandemic as well as deep sequencing data from a longitudinal within-host evolution study to explain this evolution rate saturation. The between-host data was used to obtain a snapshot of HIV-1 evolution on a large scale, while the within-host data enabled the study of evolutionary dynamics inside hosts with a much smaller focus. We showed that HIV-1 evolution has a strong tendency to revert mutations to their globally preferred state during, approximated as the HIV-1 group M consensus sequence for this study. These unpreferred states are probably the result of escape from immune selection in a previous host. They are then reverted to the optimal state in the subsequent hosts as immune selection pressure changes. Both types

of mutations, escape and reversion, are beneficial in their respective environments. Therefore, we observe along chains of transmission evolutionary dynamics driven by the adaptation to changing environments, and the global dynamic observed on the between-host scale looks like slowly acting purifying selection [151]. This finding suggests that evolution rates observed along chains of transmission would be intermediate between the within-host and between-host rates, which is indeed what was observed in previous studies [152]. We therefore showed that the "adapt and revert" hypothesis is the main cause of the evolution slowdown observed on the pandemic scale, an effect that could nonetheless be amplified by evolution during the acute phase or by the transmission bottleneck [146, 147, 153].

Substitution models commonly used to study viral phylogenetics do not account for the rapid reversions highlighted in this work as it would require site-specific rates for the consensus and non-consensus states, an effect that is not captured by standard rate variation models [154–156]. These rapid mutations away and back to the consensus are not perceived on the phylogenetic tree due to the sparse sampling. This effectively shortens the branch lengths in the tree, an effect which is stronger for longer branches which are already susceptible to long branch attraction. Consequently, omitting this bias for reversions likely hurts our ability to extract information from phylogenetic analyses such as the evolution rate or the time to the most recent common ancestor, which will appear too slow or too close to the present respectively.

HIV-1 is an excellent model to study the bias for reversions mutations due to the life long infection it causes. This leaves ample time for the virus to adapt to a specific host, and then revert after transmission. Although this effect is probably stronger in HIV-1 than it is for other viruses, it likely plays a significant role for the evolution of viruses that are endemic, like influenza A, and have to adapt to changing immune landscapes of the population over longer time scale. When looking at viral evolution over time scales of thousands of years, this reversion bias makes viral evolution as slow as the viral consensus evolution, which seems to be directly linked to their long-term host relationships [157, 158].

Although the work presented in chapter 2 highlights and quantifies the impact of reversions to consensus on HIV-1 evolution, and consequently explain most of the discrepancies between with host and between-host evolution, HIV-1 evolution remains intricate and several aspects are still poorly understood. These include the role of recombination in HIV-1 evolution, the specific dynamics of HIV-1 subtypes, and the multifaceted impact of host factors such as HLA type, $CD4$ count, and viral load, particularly in the context of transmission chains [66, 159–161]. Current and future research in these areas is crucial

for a comprehensive understanding of HIV-1's adaptability and for developing effective strategies to combat its spread.

4.2 High-throughput framework for bacteriophage evolution with the Aionstat

Since their discovery at the beginning of the 20th century, bacteriophages have been shown to be one of the most abundant and diverse biological entity on Earth [96]. Their prevalence places them at the heart of many research domains such as ecology, healthcare and molecular biology [85]. The remarkable diversity of bacteriophage is tightly linked to their ability to evolve, which makes them an excellent model to study evolutionary dynamics in various contexts [100]. Even when it is not the direct focus of the research, bacteriophage evolution is inherently linked to all these research areas. Be it for ecological reasons, where it is essential to understand how phages adapt and diversify to shape ecosystems, or for healthcare, where the evolution of phages can be leveraged to cure patients. Unfortunately, the current limitations in phage evolution research does not allow a comprehensive understanding of this field. The focus of studies is often narrow, centering on a handful of well-characterized bacteriophages [117], or very broad, like for environmental metagenomics studies where phages are poorly studied [118–120]. This results in two main limitations for studying phage evolution: either the findings are too specific and would not hold for the broader diversity of phages, or they lack detailed characterization due to the absence of experimental investigation which limits our understanding of the relevant processes. There is a need to bridge this gap to push bacteriophage research further.

It is with this goal in mind that we designed and performed the work presented in chapter 3. The first approach to solve this dichotomy was to create a bacteriophage collection that would be both representative of the large diversity of phages infecting *E.coli* K-12, but also provide in depth characterization of each phage, including their bacterial receptors, sensitivity to different bacterial immune systems and well curated genomes. This collection - the BASEL phage collection [121] - highlights differences between phage groups that are informative of evolutionary trade-offs for these phages. Although it was recently published, it has already been shared widely with many groups around the world, providing a solid foundation and reference for future bacteriophage research.

The BASEL phage collection focused on bacteriophages that infect *E.coli* K12. This laboratory strain does not possess an O-antigen like many *E.coli* strains found in nature or in patient's infections. Consequently this collection likely miss a whole pan of *E.coli* bacteriophages that are dependent on the presence of this O-antigen.To improve the

diversity of bacteriophages in the collection, an "expansion-pack" is currently being created that focuses on such bacteriophages.

The BASEL collection and similar work such as [162] contribute greatly to expand our knowledge of bacteriophage diversity and characterization. Nonetheless such projects are very labor intensive due to the time needed to isolate and characterize bacteriophages individually. Although we can strive to better understand bacteriophage diversity and evolution this way, it is unlikely that this type of approach alone will be enough to do so. Fortunately, recent advancements in sequencing technologies, along with progress in bioinformatics analysis, are well positioned to enhance our understanding of bacteriophage evolution. The yearly increase in the number of sequenced bacteriophage genomes is a testament to the potential of these methods. However, there is still a significant amount of work to be done. Bacteriophage research, particularly from a bioinformatics perspective, does not receive as much attention as research on other viruses like HIV-1. This disparity is partly due to the limited insights gained from relying on metagenomics analysis only, as bacteriophages are not as extensively characterized as many human viruses. Therefore, integrating both traditional and bioinformatics approaches seem crucial for a more comprehensive understanding of bacteriophage evolution.

Although studying existing diversity of bacteriophages in nature gives great insights about the long term evolution and diversification of phages, it is but a snapshot of the results of many years of evolution. To understand the microscopic processes from which these evolutionary dynamics emerge one must look at bacteriophage evolution in finer details. This is done via experimental work and is complementary to the approach taken with the BASEL collection to provide the full evolutionary picture. Evolution experiments on bacteriophages are typically performed via manual serial passages over several weeks, a process which is both labor intensive and prone to experimental errors such as cross contamination [163, 164]. Therefore, evolution experiments are usually limited in scale, resulting in narrow insights into the mechanisms of evolution. To improve this aspect we need a way to perform evolution experiments on a large number of phages, rapidly and reliably.

It is precisely to fulfill that role that we created the high throughput framework for bacteriophage evolution. Within this framework, one can design, perform and analyze the results of an evolution experiment on multiple phages in a matter of weeks, with a limited amount of repetitive labor. At the heart of this framework is the continuous culture device we crafted to perform the bacteriophage evolution: the Aionostat. This machine is inspired by turbidostats and morbidostats such as [125–127], but has been heavily modified to increase performance, versatility, ease of use, reliability while keeping costs reasonable.

Two evolution experiments performed with the Aionostat were showcased in section 3.5. These experiments showed that this machine can effectively train bacteriophages to increase infectivity on a challenging bacterial strain in a matter of days, via both vertical evolution and recombination of the phages. Although limited in scope in terms of biology, these experiments show that the framework and the Aionostat perform well. We hope this work will pave the way for future large scale evolution experiments that aim to answer challenging biological and evolutionary questions.

A particularly interesting topic that could be characterized in depth using the Aionostat is the evolutionary dynamics linked to the recombination of phages [86–89]. These dynamics have been empirically shown to be central to the evolution of bacteriophage via methods such as the Applemans protocol, a technique which is often used to train bacteriophages in phage therapy contexts [94]. Nonetheless, we know comparatively little about the underlying processes and principles that govern the evolution of bacteriophages in such contexts. The Aionostat is well suited to reproduce such protocols at scale, and the repeatability achievable by this machine would allow in depth characterization of the underlying processes. We could deduce and fine-tune relevant parameters to improve future phage therapy, potentially advancing our understanding of phage biology in the process. It is worth reminding us that the perfect therapeutical phage, i.e. a phage with broad host range that would be exceptionally good at infecting and killing bacteria, does not exist in nature because it would quickly drive its hosts to extinction. Therefore, the ability to train and potentially modify phages for our purposes is of utmost importance, and the insights that the Aionostat can provide in this context are essential.

The Aionostat is a versatile machine, and it could be used in other contexts than directed phage evolution. Although it has been designed for phages, it can still perform bacterial-only experiments and provides many improvements over similar devices [125, 165]. In its current state it is already a valuable addition to many laboratories, but it is not as user friendly as standard laboratory equipments despite being in the same price range. We think the machine itself could be further improved to provide better user experience while reducing costs, but this would likely require product design expertise.

Finally, the laboratory environment associated with the Aionostat provides some benefits but does have drawbacks as well. Bacteriophages strive in a variety of natural environments which are poorly represented by laboratory conditions, and our limited comprehension of bacteriophage biology may stem from these artificial conditions to some extent [166, 167]. For instance, bacteria are more often than not in a non-growing state in natural habitats, which is known to have a significant impact on the ability of phages to productively infect and

replicate [168]. Research on bacteriophages in more natural conditions is becoming more common but there remains a lot to be done.

APPENDIX

A.1 DETAILS OF THE AIONOSTAT

In this section we give some details about the Aionostat and its components. The Aionostat is an intricate machine and the resources shown here will likely not be enough to duplicate it. These are meant for long term safe keeping. For more information, get in contact with Valentin Druelle or Richard Neher.

A.1.1 *Models*

The 3d printed models used to make the Aionostat are accessible at https://github.com/vdruelle/Aionostat_ressources. The 3d printed components were printed in black tough PLA, and the laser cut components are made from 4mm acrylic panes.

A.1.2 *Electronic*

The electronic components used in the Aionostat are detailed below.

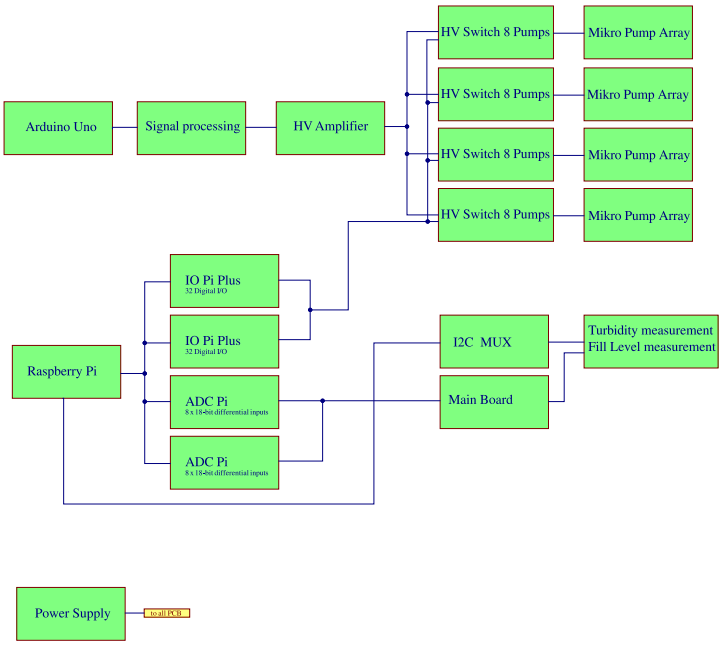


Figure A.1: Block diagram of the electric components of the Aionostat.

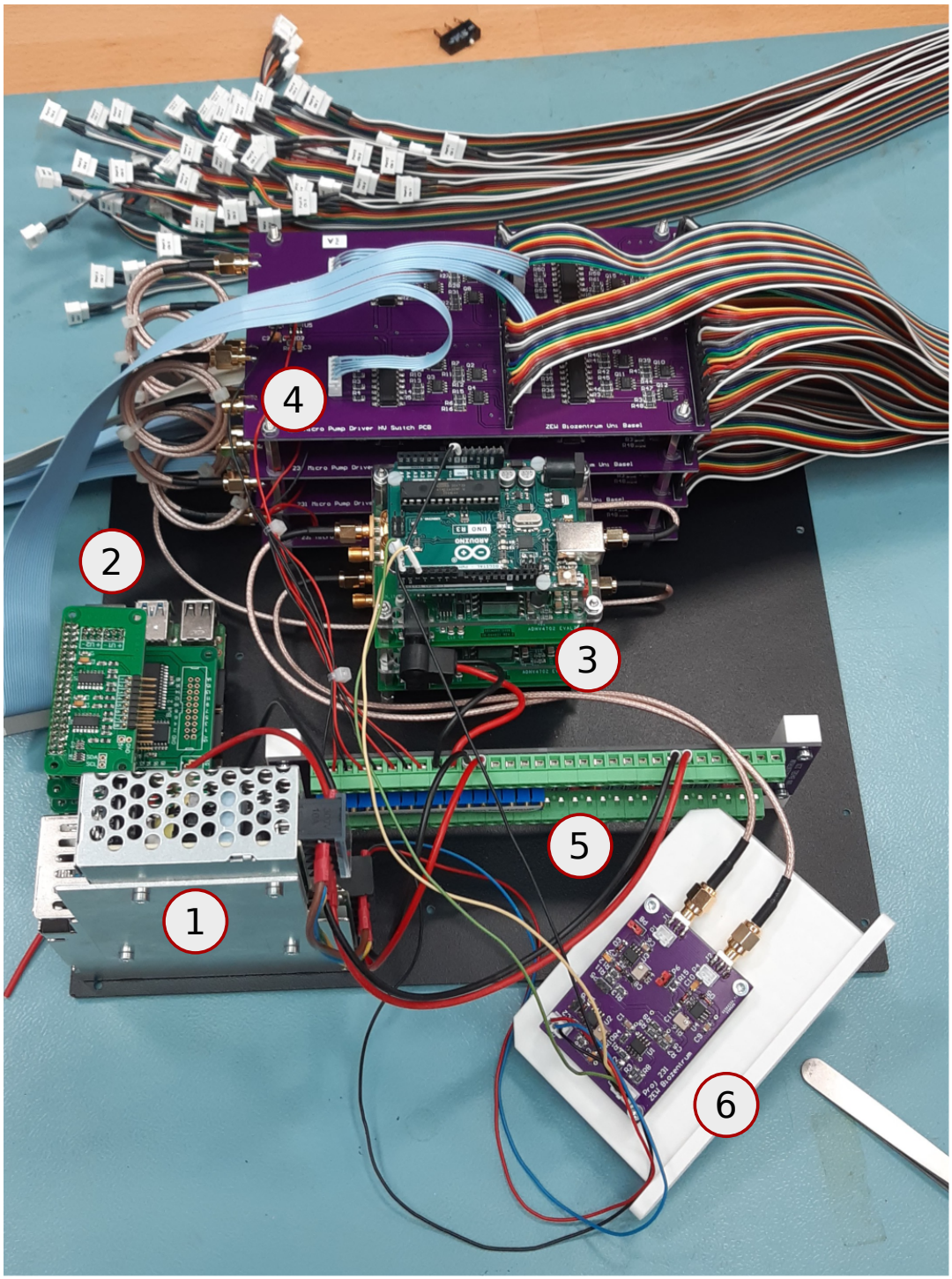
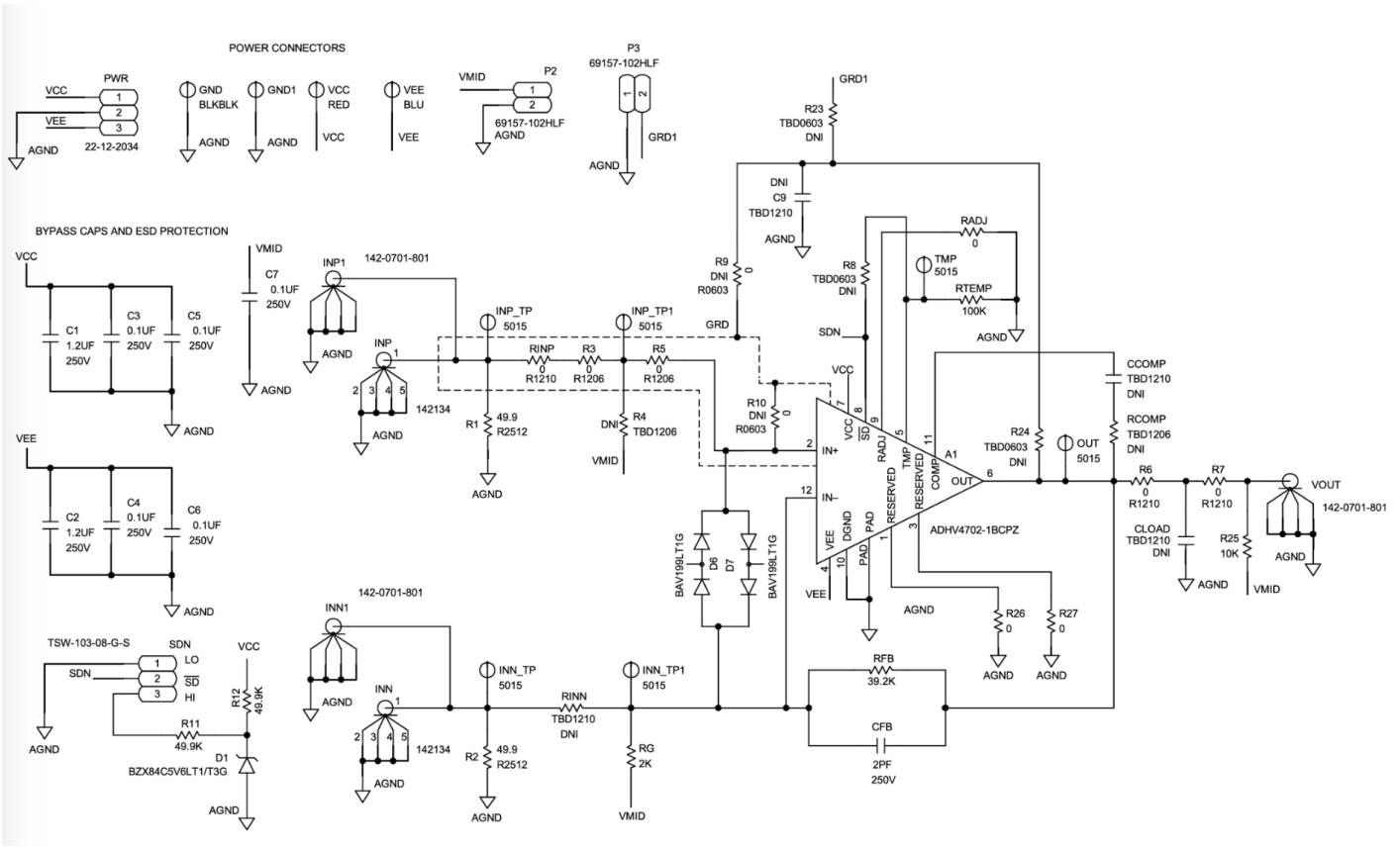
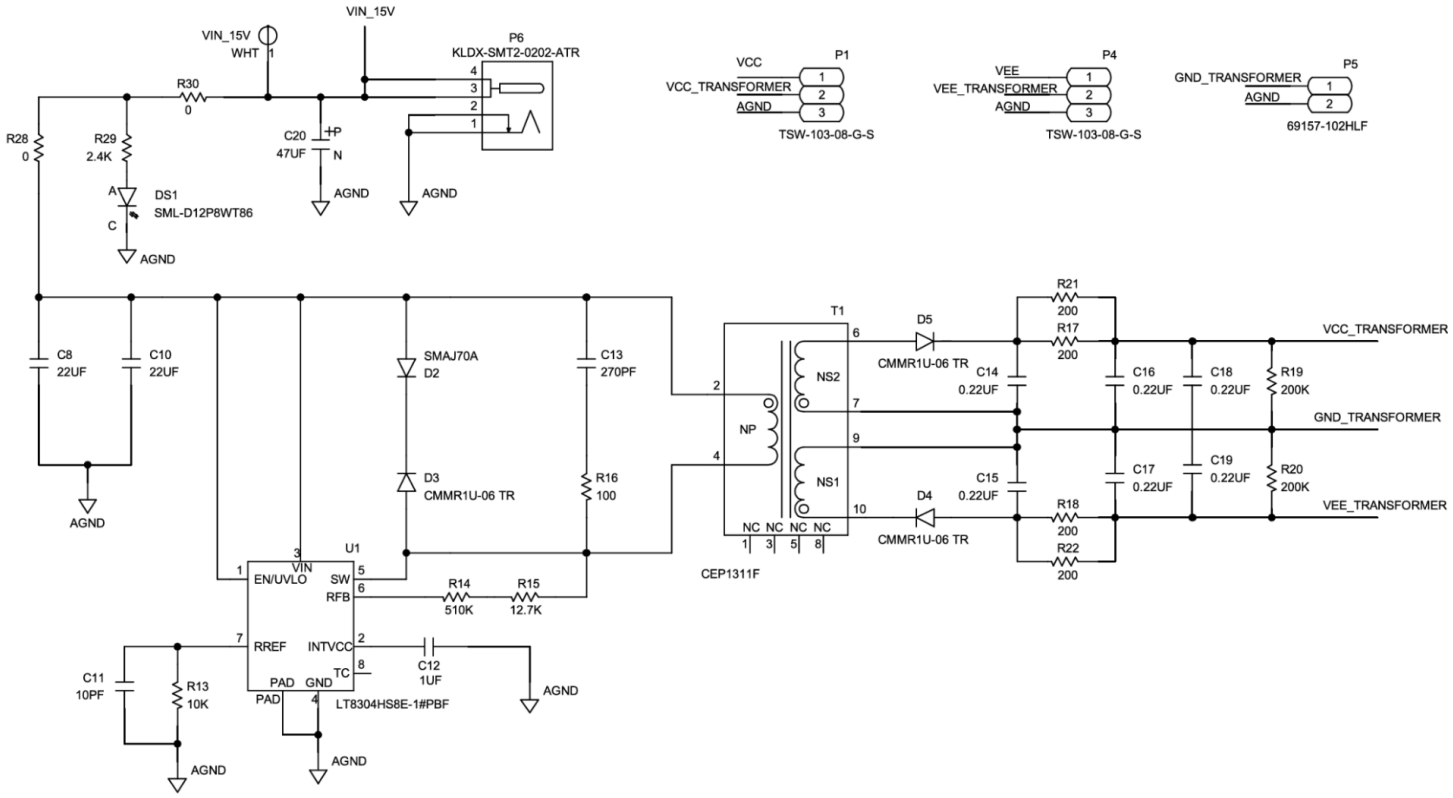


Figure A.2: Picture of the electronic components of the Aionostat. These are found in the electronic enclosure shown in figure 3.8. **1.** Power supply. **2.** Raspberry Pi 4B with HATs. **3.** ADHV4702-1CPZ (HV amplifier) and adruino for clock. **4.** Micro pump driver HV switch. **5.** Main board. **6.** Signal processing.

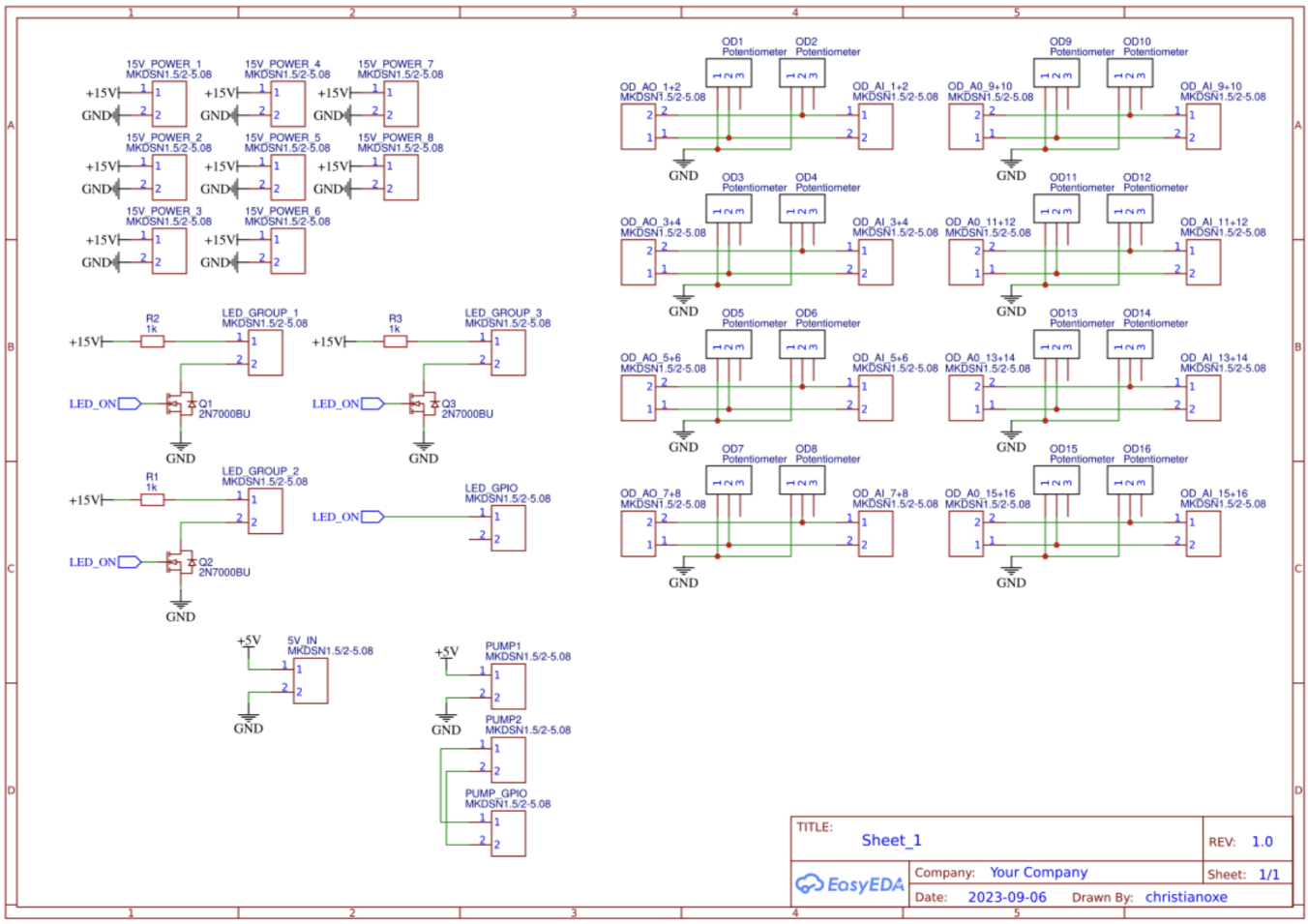
EVAL-ADHV4702-1CPZ Evaluation Board Device Under Test Crcuit Schematic



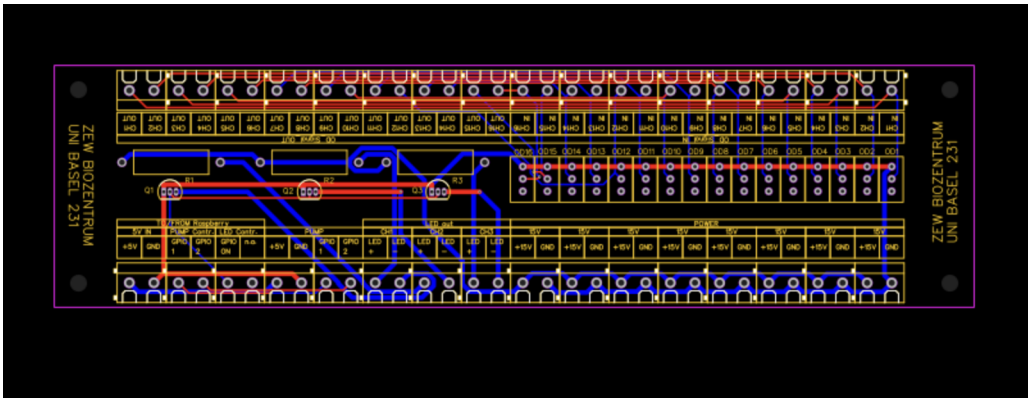
Changes: RFB = 75kOhm / R1 = Open / P2 = bridge / P3 Pin1 bridge P5 Pin1 / P1 = bridge Pin1 Pin2 / P4 = bridge Pin1



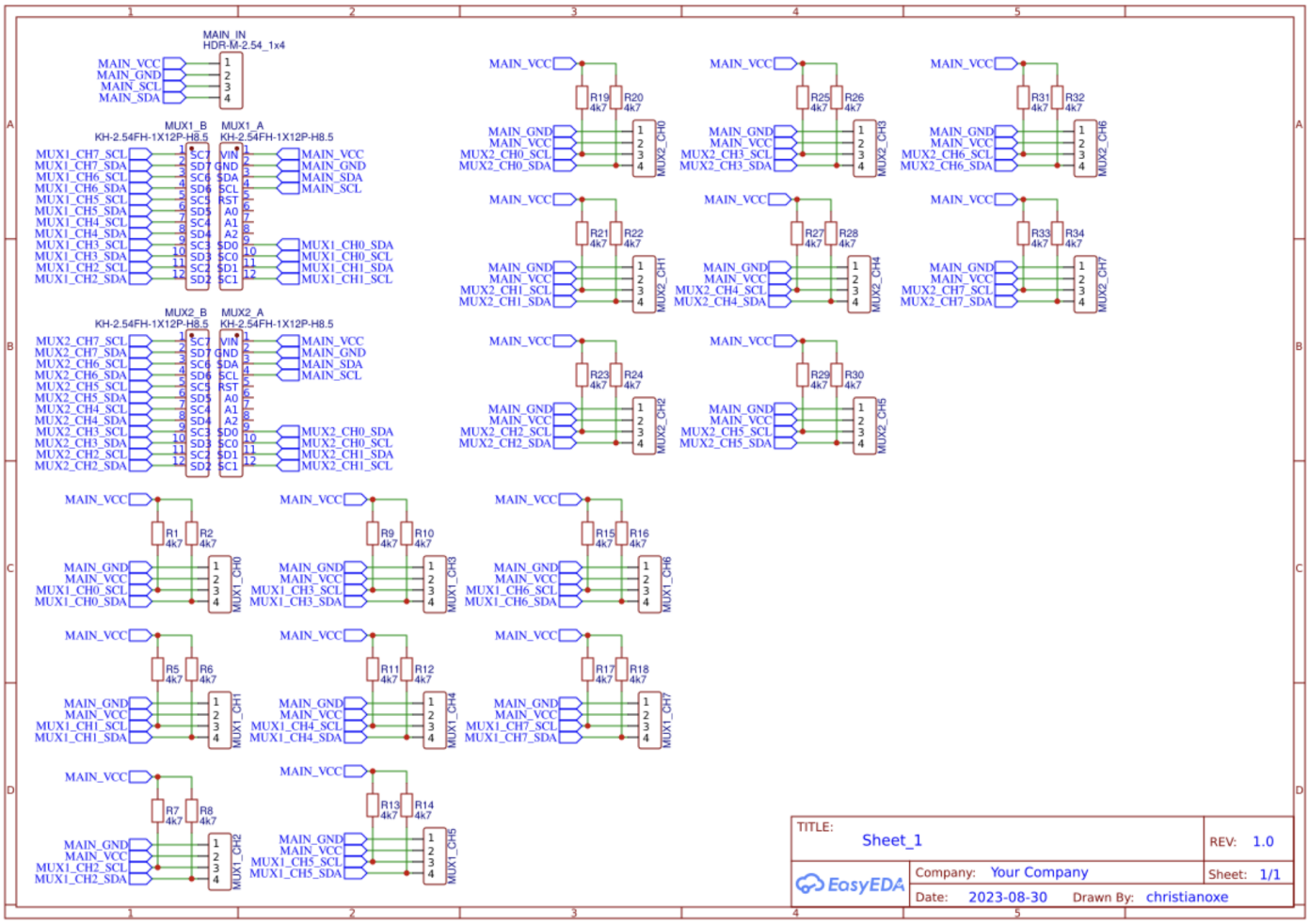
Main Board PCB



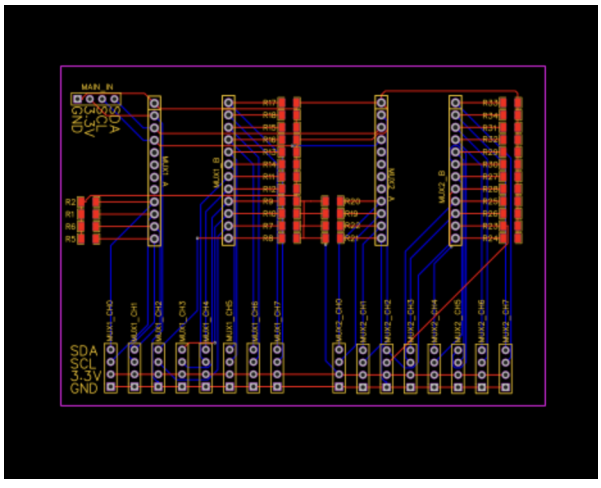
PCB



I2C MUX PCB



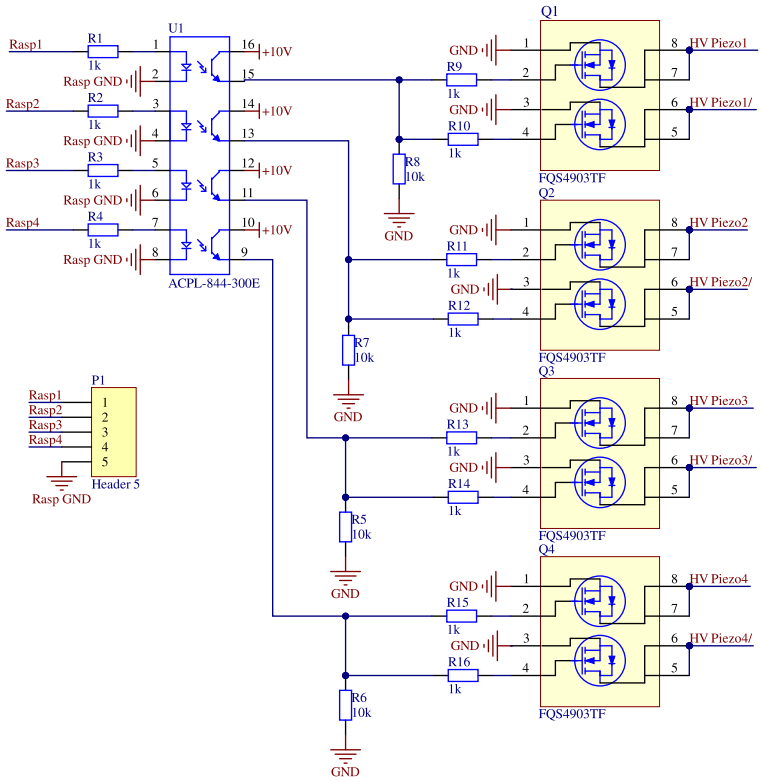
PCB

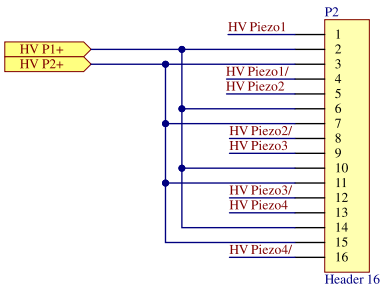


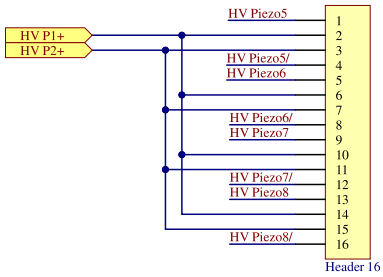
Micro pump driver

Only outputs change for the 4 other channels

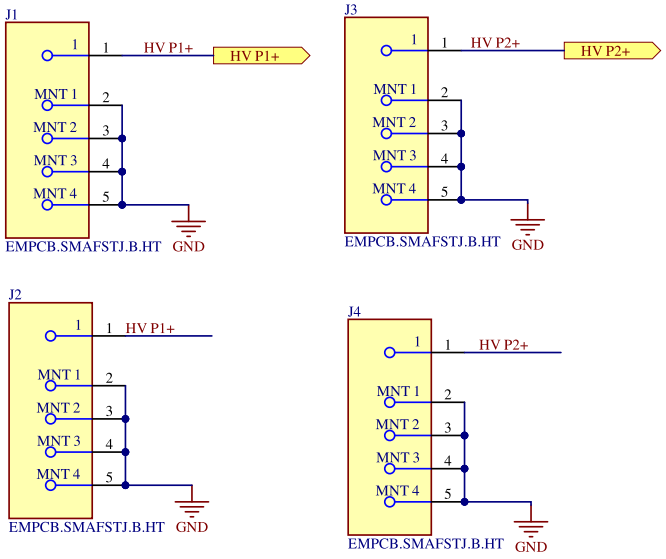
Channel 1

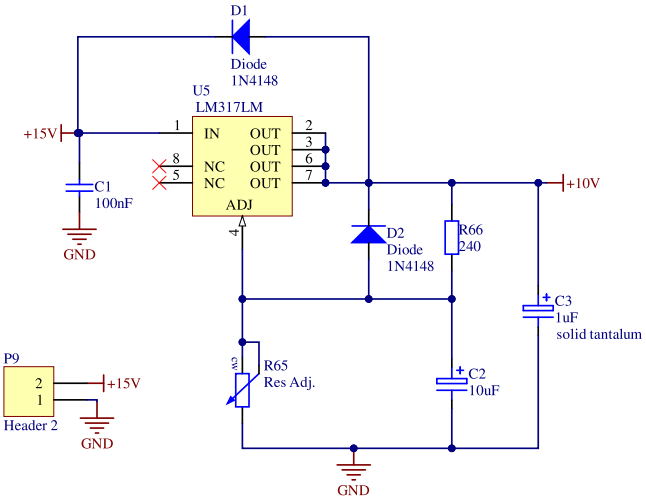




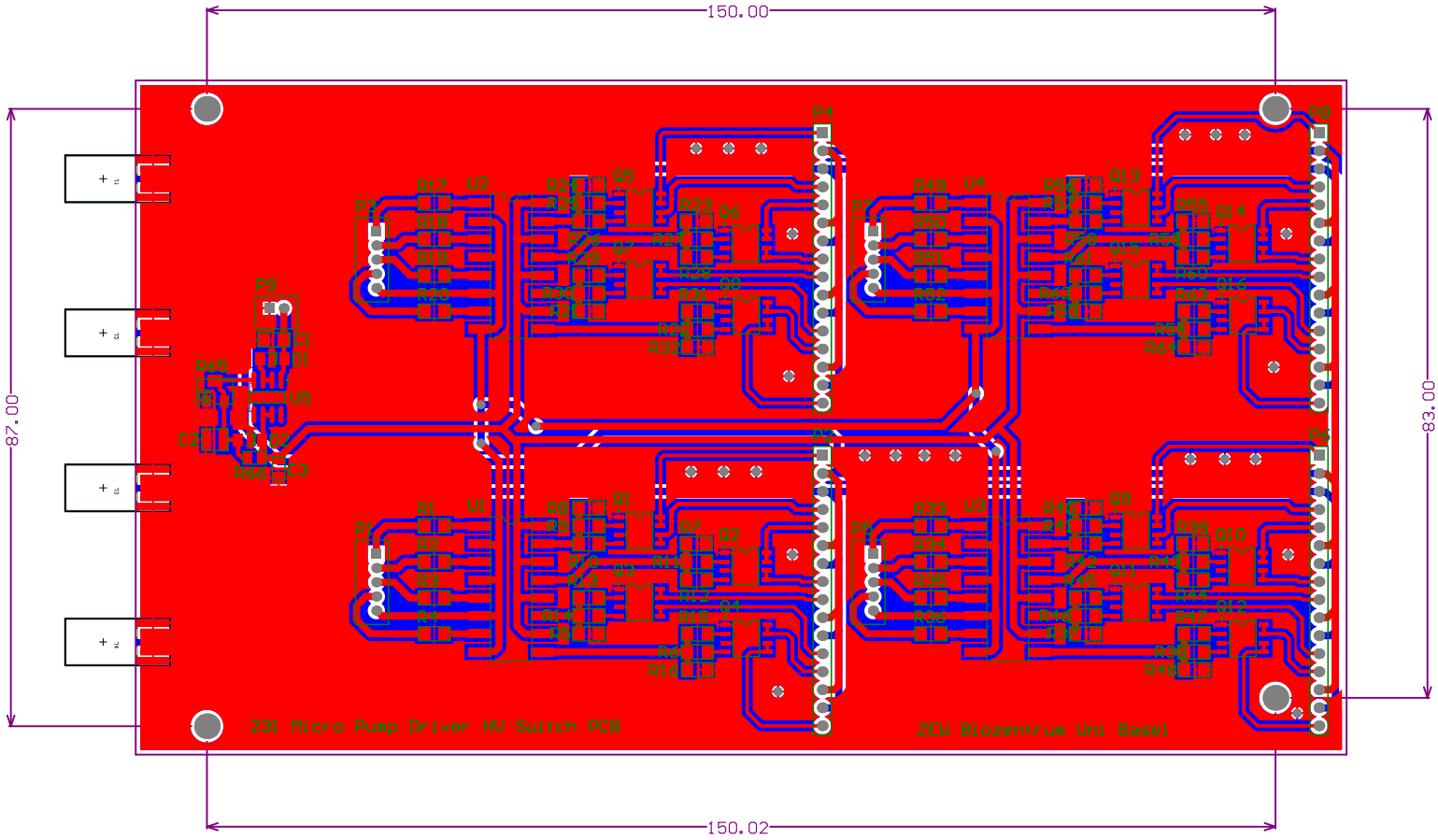


SMA Buchsen für HV In und Out zum schlaufen



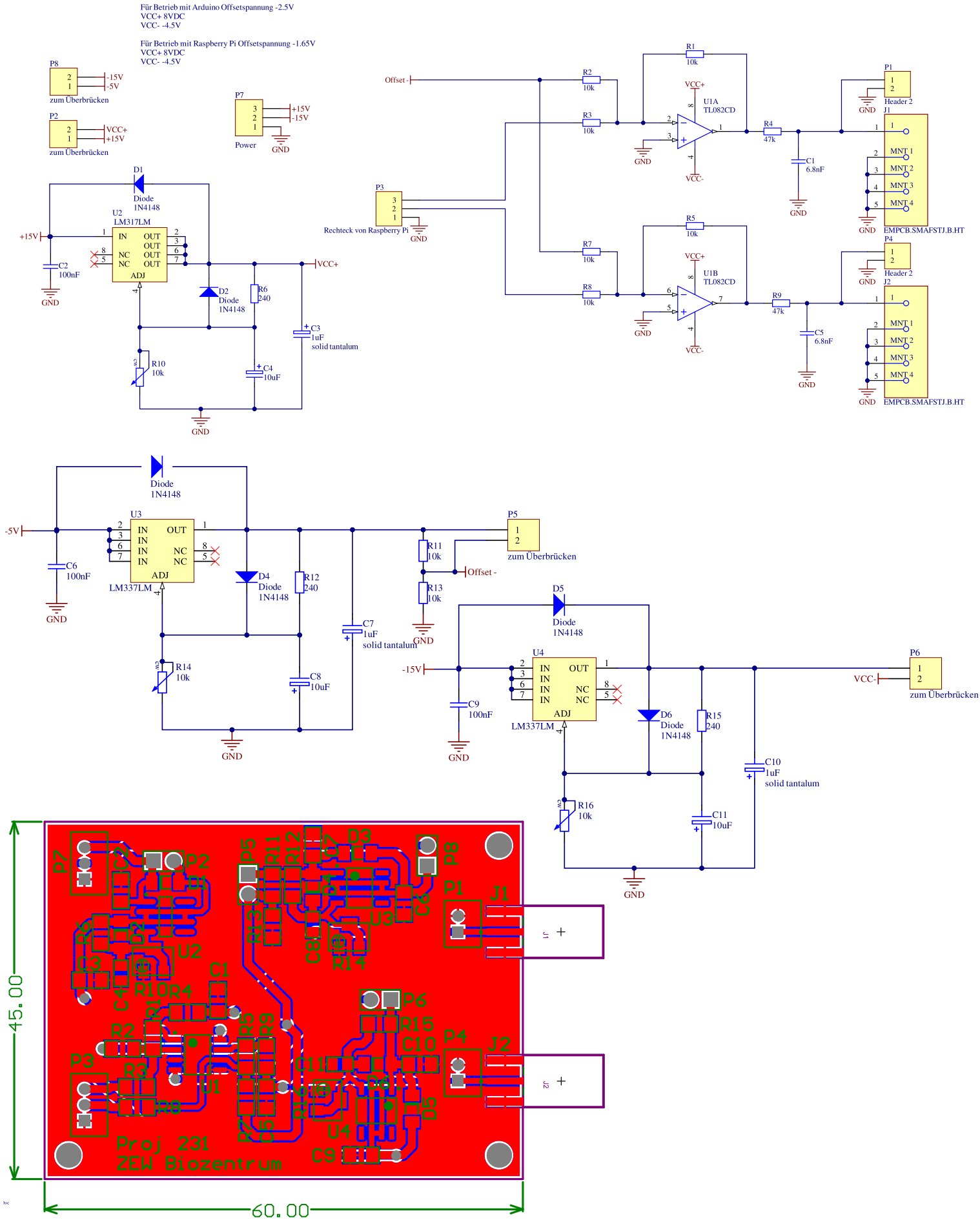


Title 231 Micro Pump Driver HV Switch Channel 1			Biozentrum Universität Basel Zentrale Elektronikwerkstatt Klingelbergstrasse 70 CH-4056 Basel Simon Saner Tel: +41 61 267 04 75 Christian Ox Tel: +41 61 207 73 97
Size: A3	Number: 1	Revision: 1.0	
Date: 23.01.2024 Time: 08:23:39	Sheet 1 of 5		



Comment	Description	Designator	Footprint	LibRef	Quantity
KOND KERAM	Capacitor (Semiconductor SIM Model)	C1	1206	KOND KERAM	1
KOND ELKO		C2	3528	KOND ELKO	1
KOND ELKO		C3	3216	KOND ELKO	1
Diode		D1, D2	SOD-80	Diode	2
EMPCB.SMAFSTJ.B.HT	CONN SMA JACK STR 50OHM EDGE MNT	J1, J2, J3, J4	FP-EMPCB_SMAFSTJ_B_HT-MFG	CMP-111093-000001-1	4
Header 5	Header, 5-Pin	P1, P3, P5, P7	B5B-PH	Header 5	4
Header 16	Header, 16-Pin	P2, P4, P6, P8	HDR1X16	Header 16	4
Header 2	Header, 2-Pin	P9	B2B-PH	Header 2	1
FQS4903TF		Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8, Q9, Q10, Q11, Q12, Q13, Q14, Q15, Q16	SOP-8	FQS4903TF	16
Res	Resistor	R1, R2, R3, R4, R5, R6, R7, R8, R9, R10, R11, R12, R13, R14, R15, R16, R17, R18, R19, R20, R21, R22, R23, R24, R25, R26, R27, R28, R29, R30, R31, R32, R33, R34, R35, R36, R37, R38, R39, R40, R41, R42, R43, R44, R45, R46, R47, R48, R49, R50, R51, R52, R53, R54, R55, R56, R57, R58, R59, R60, R61, R62, R63, R64	1206	Res	64
Res Adj.		R65	Pot 3214W	Res Adj.	1
Res	Resistor	R66	0805	Res	1
ACPL-844-300E	Optoisolator Transistor Output 3000Vrms 4 Channel 16-SO	U1, U2, U3, U4	AVAGO_SMD_16	ACPL-247-500E	4
LM317LM	3-Terminal Adjustable Regulator, 8-pin Narrow SOIC	U5	M08A_M	CMP-0062-01002-2	1

Signal processing



BIBLIOGRAPHY

[1] K David Patterson and Gerald F Pyle. "The geography and mortality of the 1918 influenza pandemic." In: *Bulletin of the History of Medicine* 65.1 (1991), pp. 4–21.

[2] Paul M Sharp and Beatrice H Hahn. "Origins of HIV and the AIDS pandemic." In: *Cold Spring Harbor Perspectives in Medicine:* 1.1 (2011).

[3] Indranil Chakraborty and Prasenjit Maity. "COVID-19 outbreak: Migration, effects on society, global environment and prevention." In: *Science of the total environment* 728 (2020), p. 138882.

[4] Vicente Javier Clemente-Suárez, Eduardo Navarro-Jiménez, Libertad Moreno-Luna, María Concepción Saavedra-Serrano, Manuel Jimenez, Juan Antonio Simón, and Jose Francisco Tornero-Aguilera. "The impact of the COVID-19 pandemic on social, health, and economy." In: *Sustainability* 13.11 (2021), p. 6314.

[5] Velislava N Petrova and Colin A Russell. "The evolution of seasonal influenza viruses." In: *Nature Reviews Microbiology* 16.1 (2018), pp. 47–60.

[6] Fernando L Gordillo Altamirano and Jeremy J Barr. "Phage therapy in the postantibiotic era." In: *Clinical microbiology reviews* 32.2 (2019), pp. 10–1128.

[7] Janis Doss, Kayla Culbertson, Delilah Hahn, Joanna Camacho, and Nazir Barekzi. "A review of phage therapy against bacterial pathogens of aquatic and terrestrial organisms." In: *Viruses* 9.3 (2017), p. 50.

[8] Alison Buchan, Gary R LeCleir, Christopher A Gulvik, and José M González. "Master recyclers: features and functions of bacteria associated with phytoplankton blooms." In: *Nature Reviews Microbiology* 12.10 (2014), pp. 686–698.

[9] Mei-Yue Wang, Rong Zhao, Li-Juan Gao, Xue-Fei Gao, De-Ping Wang, and Ji-Min Cao. "SARS-CoV-2: structure, biology, and structure-based therapeutics development." In: *Frontiers in cellular and infection microbiology* 10 (2020), p. 587269.

[10] Alan Engelman and Peter Cherepanov. "The structural biology of HIV-1: mechanistic and therapeutic insights." In: *Nature Reviews Microbiology* 10.4 (2012), pp. 279–290.

[11] H. Agut, A. Fillet, and V. Calvez. "[What is a virus?]." In: *La Revue du praticien* 47 6 (1997), pp. 602–7. DOI: 10.2307/j.ctv1ghv4rm.8.

[12] Arun Gupta, S. Kaushik, S. Kapoor, Gurmesh S. Sabarwal, S. Bobdey, and Kamalpreet Singh. "Disinfection by 1an innovative appropriate technology against COVID-19 in public health." In: *International Journal of Research in Medical Sciences* (2021). DOI: 10.18203/2320-6012.ijrms20215045.

[13] Franklin L Nobrega, Marnix Vlot, Patrick A de Jonge, Lisa L Dreesens, Hubertus JE Beaumont, Rob Lavigne, Bas E Dutilh, and Stan JJ Brouns. "Targeting mechanisms of tailed bacteriophages." In: *Nature Reviews Microbiology* 16.12 (2018), pp. 760–773.

[14] Filippo Scialo, Aurora Daniele, Felice Amato, Lucio Pastore, Maria Gabriella Matera, Mario Cazzola, Giuseppe Castaldo, and Andrea Bianco. "ACE2: the major cell entry receptor for SARS-CoV-2." In: *Lung* 198 (2020), pp. 867–877.

[15] Yinon M Bar-On, Avi Flamholz, Rob Phillips, and Ron Milo. "SARS-CoV-2 (COVID-19) by the numbers." In: *elife* 9 (2020), e57309.

[16] Richard H Heineman and James J Bull. "Testing optimality with experimental evolution: lysis time in a bacteriophage." In: *Evolution* 61.7 (2007), pp. 1695–1709.

[17] Elisa Visher, Shawn E Whitefield, J. McCrone, W. Fitzsimmons, and A. Lauring. "The Mutational Robustness of Influenza A Virus." In: *PLoS Pathogens* 12 (2016). DOI: 10.1371/journal.ppat.1005856.

[18] K. Narayanan and E. Procko. "Deep Mutational Scanning of Viral Glycoproteins and Their Host Receptors." In: *Frontiers in Molecular Biosciences* 8 (2021). DOI: 10.3389/fmolb.2021.636660.

[19] Alan F. Rubin, Hannah Gelman, Nathan Lucas, Sandra M. Bajjalieh, Anthony T. Papenfuss, Terence P. Speed, and Douglas M. Fowler. "A statistical framework for analyzing deep mutational scanning data." In: *Genome Biology* 18 (2017). DOI: 10.1186/s13059-017-1272-5.

[20] Trevor Hinkley, João Martins, Colombe Chappey, Mojgan Haddad, Eric Stawiski, Jeannette M Whitcomb, Christos J Petropoulos, and Sebastian Bonhoeffer. "A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase." In: *Nature genetics* 43.5 (2011), pp. 487–489.

[21] Fabio Zanini, Vadim Puller, Johanna Brodin, Jan Albert, and Richard A Neher. "In vivo mutation rates and the landscape of fitness costs of HIV-1." In: *Virus evolution* 3.1 (2017), vex003.

[22] John Steel and Anice C Lowen. "Influenza A virus reassortment." In: *Influenza Pathogenesis and Control-Volume I* (2014), pp. 377–401.

[23] Anice C Lowen. "Constraints, drivers, and implications of influenza A virus reassortment." In: *Annual review of virology* 4 (2017), pp. 105–121.

[24] Donald S Burke. "Recombination in HIV: an important viral evolutionary strategy." In: *Emerging infectious diseases* 3.3 (1997), p. 253.

[25] Bertha Cecilia Ramirez, Etienne Simon-Loriere, Roman Galetto, and Matteo Negroni. "Implications of recombination for HIV diversity." In: *Virus research* 134.1-2 (2008), pp. 64–73.

[26] Marcos Pérez-Losada, Miguel Arenas, Juan Carlos Galán, Ferran Palero, and Fernando González-Candelas. "Recombination in viruses: mechanisms, methods of study, and evolutionary consequences." In: *Infection, Genetics and Evolution* 30 (2015), pp. 296–307.

[27] Takashi Gojobori, Etsuko N Moriyama, and MOToo KIMuRA. "Molecular clock of viral evolution, and the neutral theory." In: *Proceedings of the National Academy of Sciences* 87.24 (1990), pp. 10015–10018.

[28] Thijs Kuiken, Edward C Holmes, John McCauley, Guus F Rimmelzwaan, Catherine S Williams, and Bryan T Grenfell. "Host species barriers to influenza virus infections." In: *Science* 312.5772 (2006), pp. 394–397.

[29] Marc P Girard, John S Tam, Olga M Assossou, and Marie Paule Kieny. "The 2009 A (H1N1) influenza virus pandemic: A review." In: *Vaccine* 28.31 (2010), pp. 4895–4902.

[30] Devika Singh and Soojin V Yi. "On the origin and evolution of SARS-CoV-2." In: *Experimental & Molecular Medicine* 53.4 (2021), pp. 537–547.

[31] B. Linz, Longhuan Ma, Israel Rivera, and E. Harvill. "Genotypic and phenotypic adaptation of pathogens: lesson from the genus Bordetella." In: *Current Opinion in Infectious Diseases* 32 (2019), pp. 223–230. DOI: 10.1097/QCO.0000000000000549.

[32] F. Brodsky, L. Lem, A. Solache, and E. M. Bennett. "Human pathogen subversion of antigen presentation." In: *Immunological Reviews* 168 (1999). DOI: 10.1111/j.1600-065X.1999.tb01294.x.

[33] N. Stephenson and J. Foley. "Parallelisms and Contrasts in the Diverse Ecologies of the Anaplasma phagocytophilum and Borrelia burgdorferi Complexes of Bacteria in the Far Western United States." In: *Veterinary Sciences* 3 (2016). DOI: 10.3390/vetsci3040026.

[34] K. A. Fields, R. Heinzen, and R. Carabeo. "The Obligate Intracellular Lifestyle." In: *Frontiers in Microbiology* 2 (2011). DOI: 10.3389/fmicb.2011.00099.

[35] Frederick Sanger, Steven Nicklen, and Alan R Coulson. "DNA sequencing with chain-terminating inhibitors." In: *Proceedings of the national academy of sciences* 74.12 (1977), pp. 5463–5467.

[36] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, et al. "Accurate whole human genome sequencing using reversible terminator chemistry." In: *nature* 456.7218 (2008), pp. 53–59.

[37] Yunhao Wang, Yue Zhao, Audrey Bollas, Yuru Wang, and Kin Fai Au. "Nanopore sequencing technology, bioinformatics and applications." In: *Nature biotechnology* 39.11 (2021), pp. 1348–1365.

[38] Zhiyuan Chen, Andrew S Azman, Xinhua Chen, Junyi Zou, Yuyang Tian, Ruijia Sun, Xiangyanyu Xu, Yani Wu, Wanying Lu, Shijia Ge, et al. "Global landscape of SARS-CoV-2 genomic surveillance and data sharing." In: *Nature genetics* 54.4 (2022), pp. 499–507.

[39] Richard A Neher, Robert Dyrdak, Valentin Druelle, Emma B Hodcroft, and Jan Albert. "Potential impact of seasonal forcing on a SARS-CoV-2 pandemic." In: *Swiss medical weekly* 150 (2020), w20224.

[40] Nicholas B Noll, Ivan Aksamentov, Valentin Druelle, Abrie Badenhorst, Bruno Ronzani, Gavin Jefferies, Jan Albert, and Richard A Neher. "COVID-19 Scenarios: an interactive tool to explore the spread and associated morbidity and mortality of SARS-CoV-2." In: *MedRxiv* (2020), pp. 2020–05.

[41] Linda JS Allen. "An introduction to stochastic epidemic models." In: *Mathematical epidemiology*. Springer, 2008, pp. 81–130.

[42] Elizabeth Bruch and Jon Atwell. "Agent-based models in empirical social research." In: *Sociological methods & research* 44.2 (2015), pp. 186–221.

[43] Salathé Marcel, Althaus L Christian, Neher Richard, Stringhini Silvia, Hodcroft Emma, Fellay Jacques, Zwahlen Marcel, Senti Gabriela, Battegay Manuel, Wilder-Smith Annelies, et al. "COVID-19 epidemic in Switzerland: on the importance of testing, contact tracing and isolation." In: *Swiss medical weekly* 150 (2020), w202205.

[44] Ziheng Yang. *Computational molecular evolution*. OUP Oxford, 2006.

[45] Alexandros Stamatakis. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." In: *Bioinformatics* 30.9 (2014), pp. 1312–1313.

[46] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt Von Haeseler, and Robert Lanfear. "IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era." In: *Molecular biology and evolution* 37.5 (2020), pp. 1530–1534.

[47] Alexei J Drummond and Andrew Rambaut. "BEAST: Bayesian evolutionary analysis by sampling trees." In: *BMC evolutionary biology* 7.1 (2007), pp. 1–8.

[48] Thorsten R Klingen, Susanne Reimering, Carlos A Guzmán, and Alice C McHardy. "In silico vaccine strain prediction for human influenza viruses." In: *Trends in microbiology* 26.2 (2018), pp. 119–131.

[49] Marta Łuksza and Michael Lässig. "A predictive fitness model for influenza." In: *Nature* 507.7490 (2014), pp. 57–61.

[50] Richard A Neher, Colin A Russell, and Boris I Shraiman. "Predicting evolution from the shape of genealogical trees." In: *Elife* 3 (2014), e03568.

[51] Lauren A Castro, Trevor Bedford, and Lauren Ancel Meyers. "Early prediction of antigenic transitions for influenza A/H3N2." In: *PLoS computational biology* 16.2 (2020), e1007683.

[52] Pierre Barrat-Charlaix, John Huddleston, Trevor Bedford, and Richard A Neher. "Limited predictability of amino acid substitutions in seasonal influenza viruses." In: *Molecular Biology and Evolution* 38.7 (2021), pp. 2767–2777.

[53] Abdallah S Abdelsattar, Alyaa Dawooud, Nouran Rezk, Salsabil Makky, Anan Safwat, Philip J Richards, and Ayman El-Shibiny. "How to train your phage: The recent efforts in phage training." In: *Biologics* 1.2 (2021), pp. 70–88.

[54] UNAIDS. *Global HIV & AIDS statistics — Fact sheet*. 2022. URL: https://www.unaids.org/en/resources/fact-sheet.

[55] Shalom Spira, Mark A Wainberg, Hugues Loemba, Dan Turner, and Bluma G Brenner. "Impact of clade diversity on HIV-1 virulence, antiretroviral drug sensitivity and drug resistance." In: *Journal of Antimicrobial Chemotherapy* 51.2 (2003), pp. 229–240.

[56] Bette Korber, Mark Muldoon, James Theiler, Fei Gao, Radhika Gupta, Alan Lapedes, Beatrice H Hahn, Steven Wolinsky, and Tanmoy Bhattacharya. "Timing the ancestor of the HIV-1 pandemic strains." In: *science* 288.5472 (2000), pp. 1789–1796.

[57] Michael Worobey, Marlea Gemmel, Dirk E Teuwen, Tamara Haselkorn, Kevin Kunstman, Michael Bunce, Jean-Jacques Muyembe, Jean-Marie M Kabongo, Raphaël M Kalengayi, Eric Van Marck, et al. "Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960." In: *Nature* 455.7213 (2008), pp. 661–664.

[58] Deborah Cromer, Andrew J Grimm, Timothy E Schlub, Johnson Mak, and Miles P Davenport. "Estimating the in-vivo HIV template switching and recombination rate." In: *Aids* 30.2 (2016), pp. 185–192.

[59] Brian Foley, Thomas Leitner, Cristian Apetrei, Beatrice Hahn, Ilene Mizrachi, James Mullins, Andrew Rambaut, Steven Wolinsky, and Bette Korber. "HIV sequence compendium 2018." In: *Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM, LA-UR* 18 (2018), p. 25673.

[60] Thomas Splettstoesser. *Diagram of the HIV virion*. URL: https://en.wikipedia.org/wiki/HIV#/media/File:HI-virion-structure_en.svg.

[61] Thomas Splettstoesser. *Structure of the RNA genome of HIV-1*. URL: https://en.wikipedia.org/wiki/HIV#/media/File:HIV-genome.png.

[62] Creative Commons. *Creative Commons Attribution - ShareAlike License 4.0*. URL: https://creativecommons.org/licenses/by-sa/4.0/deed.en.

[63] Michael E Abram, Andrea L Ferris, Wei Shao, W Gregory Alvord, and Stephen H Hughes. "Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication." In: *Journal of virology* 84.19 (2010), pp. 9864–9878.

[64] Louis M Mansky and Howard M Temin. "Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase." In: *Journal of virology* 69.8 (1995), pp. 5087–5094.

[65] Bradley D Preston, Bernard J Poiesz, and Lawrence A Loeb. "Fidelity of HIV-1 reverse transcriptase." In: *Science* 242.4882 (1988), pp. 1168–1171.

[66] Fabio Zanini, Johanna Brodin, Lina Thebo, Christa Lanz, Göran Bratt, Jan Albert, and Richard A Neher. "Population genomics of intrapatient HIV-1 evolution." In: *Elife* 4 (2015), e11282.

[67] Thomas Splettstoesser. *Relationship between HIV copies and CD4 count over the course of untreated HIV infection*. URL: https://en.wikipedia.org/wiki/HIV#/media/File:Hiv-timecourse_copy.svg.

[68] Creative Commons. *Creative Commons CCo 1.0 Universal Public Domain Dedication*. URL: https://creativecommons.org/publicdomain/zero/1.0/deed.en.

[69] M Roux. "On an invisible microbe antagonistic to dysentery bacilli. Note by MF d'Herelle, presented by M. Roux." In: *Comptes rendus academie des sciences* 165 (1917), pp. 373–5.

[70] Dann Turner, Andrey N Shkoporov, Cédric Lood, Andrew D Millard, Bas E Dutilh, Poliane Alfenas-Zerbini, Leonardo J van Zyl, Ramy K Aziz, Hanna M Oksanen, Minna M Poranen, et al. "Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee." In: *Archives of Virology* 168.2 (2023), p. 74.

[71] Moïra B Dion, Frank Oechslin, and Sylvain Moineau. "Phage diversity, genomics and phylogeny." In: *Nature Reviews Microbiology* 18.3 (2020), pp. 125–138.

[72] Anne Mai-Prochnow, Janice Gee Kay Hui, Staffan Kjelleberg, Jasna Rakonjac, Diane McDougald, and Scott A Rice. "Big things in small packages: the genetics of filamentous phage and effects on fitness of their host." In: *FEMS microbiology reviews* 39.4 (2015), pp. 465–487.

[73] Belinda Loh, Andreas Kuhn, and Sebastian Leptihn. "The fascinating biology behind phage display: filamentous phage assembly." In: *Molecular Microbiology* 111.5 (2019), pp. 1132–1138.

[74] Leonard Mindich, Xueying Qiao, Jian Qiao, Shiroh Onodera, Martin Romantschuk, and Deborah Hoogstraten. "Isolation of additional bacteriophages with genomes of segmented double-stranded RNA." In: *Journal of bacteriology* 181.15 (1999), pp. 4505–4508.

[75] Sari Mäntynen, Lotta-Riina Sundberg, and Minna M Poranen. "Recognition of six additional cystoviruses: Pseudomonas virus phi6 is no longer the sole species of the family Cystoviridae." In: *Archives of virology* 163.4 (2018), pp. 1117–1124.

[76] Olin K Silander, Daniel M Weinreich, Kevin M Wright, Kara J O'Keefe, Camilla U Rang, Paul E Turner, and Lin Chao. "Widespread genetic exchange among terrestrial bacteriophages." In: *Proceedings of the National Academy of Sciences* 102.52 (2005), pp. 19009–19014.

[77] Cristina Howard-Varona, Katherine R Hargreaves, Stephen T Abedon, and Matthew B Sullivan. "Lysogeny in nature: mechanisms, impact and ecology of temperate phages." In: *The ISME journal* 11.7 (2017), pp. 1511–1520.

[78] Amos B Oppenheim, Oren Kobiler, Joel Stavans, Donald L Court, and Sankar Adhya. "Switches in bacteriophage lambda development." In: *Annu. Rev. Genet.* 39 (2005), pp. 409–429.

[79] John W Little and Christine B Michalowski. "Stability and instability in the lysogenic state of phage lambda." In: *Journal of bacteriology* 192.22 (2010), pp. 6064–6076.

[80] François St-Pierre and Drew Endy. "Determination of cell fate selection during phage lambda infection." In: *Proceedings of the National Academy of Sciences* 105.52 (2008), pp. 20705–20710.

[81] Herbert Schmidt. "Shiga-toxin-converting bacteriophages." In: *Research in microbiology* 152.8 (2001), pp. 687–695.

[82] E Fidelma Boyd, Brigid M Davis, and Bianca Hochhut. "Bacteriophage–bacteriophage interactions in the evolution of pathogenic bacteria." In: *Trends in microbiology* 9.3 (2001), pp. 137–144.

[83] François Rousset, Julien Dowding, Aude Bernheim, Eduardo PC Rocha, and David Bikard. "Prophage-encoded hotspots of bacterial immune systems." In: *bioRxiv* (2021), pp. 2021–01.

[84] Adrienne MS Correa, Cristina Howard-Varona, Samantha R Coy, Alison Buchan, Matthew B Sullivan, and Joshua S Weitz. "Revisiting the rules of life for viruses of microorganisms." In: *Nature Reviews Microbiology* 19.8 (2021), pp. 501–513.

[85] Anne Chevallereau, Benoît J Pons, Stineke van Houte, and Edze R Westra. "Interactions between bacterial and phage communities in natural environments." In: *Nature Reviews Microbiology* 20.1 (2022), pp. 49–62.

[86] Roger W Hendrix. "Bacteriophages: evolution of the majority." In: *Theoretical population biology* 61.4 (2002), pp. 471–480.

[87] Hiroya Kunisaki and Yasunori Tanji. "Intercrossing of phage genomes in a phage cocktail and stable coexistence with Escherichia coli O157: H7 in anaerobic continuous culture." In: *Applied microbiology and biotechnology* 85 (2010), pp. 1533–1540.

[88] Graham F Hatfull and Roger W Hendrix. "Bacteriophages and their genomes." In: *Current opinion in virology* 1.4 (2011), pp. 298–303.

[89] TN Mavrich and GF Hatfull. *Bacteriophage evolution differs by host, lifestyle and genome. Nat Microbiol 2: 17112*. 2017.

[90] Héloïse Georjon and Aude Bernheim. "The highly diverse antiphage defence systems of bacteria." In: *Nature Reviews Microbiology* 21.10 (2023), pp. 686–700.

[91] Sarah Camara-Wilpert, David Mayo-Muñoz, Jakob Russel, Robert D Fagerlund, Jonas S Madsen, Peter C Fineran, Søren J Sørensen, and Rafael Pinilla-Redondo. "Bacteriophages suppress CRISPR–Cas immunity using RNA-based anti-CRISPRs." In: *Nature* 623.7987 (2023), pp. 601–607.

[92] Alexander P Hynes, Geneviève M Rousseau, Daniel Agudelo, Adeline Goulet, Beatrice Amigues, Jeremy Loehr, Dennis A Romero, Christophe Fremaux, Philippe Horvath, Yannick Doyon, et al. "Widespread anti-CRISPR proteins in virulent bacteriophages inhibit a range of Cas9 proteins." In: *Nature communications* 9.1 (2018), p. 2919.

[93] Patrick A de Jonge, Franklin L Nobrega, Stan JJ Brouns, and Bas E Dutilh. "Molecular and evolutionary determinants of bacteriophage host range." In: *Trends in microbiology* 27.1 (2019), pp. 51–63.

[94] Ben H Burrowes, Ian J Molineux, and Joe A Fralick. "Directed in vitro evolution of therapeutic bacteriophages: The Appelmans protocol." In: *Viruses* 11.3 (2019), p. 241.

[95] Gal Ofir and Rotem Sorek. "Contemporary phage biology: from classic models to new insights." In: *Cell* 172.6 (2018), pp. 1260–1270.

[96] Curtis A Suttle. "Viruses in the sea." In: *Nature* 437.7057 (2005), pp. 356–361.

[97] Mya Breitbart, Chelsea Bonnain, Kema Malki, and Natalie A Sawaya. "Phage puppet masters of the marine microbial realm." In: *Nature microbiology* 3.7 (2018), pp. 754–766.

[98] Sarah L James, Mojgan Rabiey, Benjamin W Neuman, Glynn Percival, and Robert W Jackson. "Isolation, characterisation and experimental evolution of phage that infect the horse chestnut tree pathogen, Pseudomonas syringae pv. aesculi." In: *Current microbiology* 77 (2020), pp. 1438–1447.

[99] Eugene V Koonin and Natalya Yutin. "The crAss-like phage group: how metagenomics reshaped the human virome." In: *Trends in Microbiology* 28.5 (2020), pp. 349–359.

[100] Graham F Hatfull. "Dark matter of the biosphere: the amazing world of bacteriophage diversity." In: *Journal of virology* 89.16 (2015), pp. 8107–8110.

[101] Antonio Pedro Camargo, Stephen Nayfach, I-Min A Chen, Krishnaveni Palaniappan, Anna Ratner, Ken Chu, Stephan J Ritter, TBK Reddy, Supratim Mukherjee, Frederik Schulz, et al. "IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata." In: *Nucleic acids research* 51.D1 (2023), pp. D733–D743.

[102] Nathalie Turgeon, Marie-Josée Toulouse, Bruno Martel, Sylvain Moineau, and Caroline Duchaine. "Comparison of five bacteriophages as models for viral aerosol studies." In: *Applied and environmental microbiology* 80.14 (2014), pp. 4242–4250.

[103] Andreas K Brödel, Mark Isalan, and Alfonso Jaramillo. "Engineering of biomolecules by bacteriophage directed evolution." In: *Current Opinion in Biotechnology* 51 (2018), pp. 32–38.

[104] William GT Willats. "Phage display: practicalities and prospects." In: *Plant molecular biology* 50 (2002), pp. 837–854.

[105] Kaitlyn E Kortright, Benjamin K Chan, Jonathan L Koff, and Paul E Turner. "Phage therapy: a renewed approach to combat antibiotic-resistant bacteria." In: *Cell host & microbe* 25.2 (2019), pp. 219–232.

[106] Florian Tesson, Alexandre Hervé, Ernest Mordret, Marie Touchon, Camille d'Humières, Jean Cury, and Aude Bernheim. "Systematic and quantitative view of the antiviral arsenal of prokaryotes." In: *Nature communications* 13.1 (2022), p. 2561.

[107] Jean-Paul Pirnay, Sarah Djebara, Griet Steurs, Johann Griselain, Christel Cochez, Steven De Soir, Tea Glonti, An Spiessens, Emily Vanden Berghe, Sabrina Green, et al. "Retrospective, observational analysis of the first one hundred consecutive cases of personalized bacteriophage therapy of difficult-to-treat infections facilitated by a Belgian consortium." In: *medRxiv* (2023), pp. 2023–08.

[108] Luis F Camarillo-Guerrero, Alexandre Almeida, Guillermo Rangel-Pineros, Robert D Finn, and Trevor D Lawley. "Massive expansion of human gut bacteriophage diversity." In: *Cell* 184.4 (2021), pp. 1098–1109.

[109] Andrey N Shkoporov and Colin Hill. "Bacteriophages of the human gut: the "known unknown" of the microbiome." In: *Cell host & microbe* 25.2 (2019), pp. 195–209.

[110] Marion Dalmasso, Colin Hill, and R Paul Ross. "Exploiting gut bacteriophages for human health." In: *Trends in microbiology* 22.7 (2014), pp. 399–405.

[111] Werner Arber and Daisy Dussoix. "Host specificity of DNA produced by Escherichia coli: I. Host controlled modification of bacteriophage λ." In: *Journal of molecular biology* 5.1 (1962), pp. 18–36.

[112] Hamilton O Smith and KW Welcox. "A restriction enzyme from Hemophilus influenzae: I. Purification and general properties." In: *Journal of molecular biology* 51.2 (1970), pp. 379–391.

[113] Francisco JM Mojica, Cesar Díez-Villaseñor, Elena Soria, and Guadalupe Juez. "Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria." In: *Molecular microbiology* 36.1 (2000), pp. 244–246.

[114] Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A Doudna, and Emmanuelle Charpentier. "A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity." In: *science* 337.6096 (2012), pp. 816–821.

[115] Gita Mahmoudabadi and Rob Phillips. "A comprehensive and quantitative exploration of thousands of viral genomes." In: *Elife* 7 (2018), e31955.

[116] George PC Salmond and Peter C Fineran. "A century of the phage: past, present and future." In: *Nature Reviews Microbiology* 13.12 (2015), pp. 777–786.

[117] Eric C Keen. "A century of phage research: bacteriophages and the shaping of modern biology." In: *Bioessays* 37.1 (2015), pp. 6–9.

[118] Jorge A Moura de Sousa, Eugen Pfeifer, Marie Touchon, and Eduardo PC Rocha. "Genome diversification via genetic exchanges between temperate and virulent bacteriophages." In: *bioRxiv* (2020), pp. 2020–04.

[119] Germán Bonilla-Rosso, Théodora Steiner, Fabienne Wichmann, Evan Bexkens, and Philipp Engel. "Honey bees harbor a diverse gut virome engaging in nested strain-level interactions with the microbiota." In: *Proceedings of the National Academy of Sciences* 117.13 (2020), pp. 7355–7362.

[120] Michael J Tisza and Christopher B Buck. "A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases." In: *Proceedings of the National Academy of Sciences* 118.23 (2021), e2023202118.

[121] Enea Maffei, Aisylu Shaidullina, Marco Burkolter, Yannik Heyer, Fabienne Estermann, Valentin Druelle, Patrick Sauer, Luc Willi, Sarah Michaelis, Hubert Hilbi, et al. "Systematic exploration of Escherichia coli phage–host interactions with the BASEL phage collection." In: *PLoS Biology* 19.11 (2021), e3001424.

[122] Dan Liu and Peter R Reeves. "Escherichia coli K12 regains its O antigen." In: *Microbiology* 140.1 (1994), pp. 49–57.

[123] Imke HE Korf, Jan P Meier-Kolthoff, Evelien M Adriaenssens, Andrew M Kropinski, Manfred Nimtz, Manfred Rohde, Mark J van Raaij, and Johannes Wittmann. "Still something to discover: novel insights into Escherichia coli phage diversity and taxonomy." In: *Viruses* 11.5 (2019), p. 454.

[124] Tue Kjærgaard Nielsen, Laura Milena Forero-Junco, Witold Kot, Sylvain Moineau, Lars Hestbjerg Hansen, and Leise Riber. "Detection of nucleotide modifications in bacteria and bacteriophages: Strengths and limitations of current technologies and software." In: *Molecular Ecology* 32.6 (2023), pp. 1236–1247.

[125] Erdal Toprak, Adrian Veres, Sadik Yildiz, Juan M Pedraza, Remy Chait, Johan Paulsson, and Roy Kishony. "Building a morbidostat: an automated continuous-culture device for studying bacterial drug resistance under dynamically sustained drug inhibition." In: *Nature protocols* 8.3 (2013), pp. 555–567.

[126] Tzvi Holtzman, Rea Globus, Shahar Molshanski-Mor, Adam Ben-Shem, Ido Yosef, and Udi Qimron. "A continuous evolution system for contracting the host range of bacteriophage T7." In: *Scientific Reports* 10.1 (2020), p. 307.

[127] Andreas K Brödel, Rui Rodrigues, Alfonso Jaramillo, and Mark Isalan. "Accelerated evolution of a minimal 63–amino acid dual transcription factor." In: *Science Advances* 6.24 (2020), eaba2728.

[128] Tomoya Baba, Takeshi Ara, Miki Hasegawa, Yuki Takai, Yoshiko Okumura, Miki Baba, Kirill A Datsenko, Masaru Tomita, Barry L Wanner, and Hirotada Mori. "Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection." In: *Molecular systems biology* 2.1 (2006), pp. 2006–0008.

[129] Benjamin Sellner, Rūta Prakapaitė, Margo van Berkum, Matthias Heinemann, Alexander Harms, and Urs Jenal. "A new sugar for an old phage: a c-di-GMP-dependent polysaccharide pathway sensitizes Escherichia coli for bacteriophage infection." In: *Mbio* 12.6 (2021), e03246–21.

[130] Heng Li. "Minimap2: pairwise alignment for nucleotide sequences." In: *Bioinformatics* 34.18 (2018), pp. 3094–3100.

[131] Charles Darwin. "Origin of the Species." In: *British Politics and the Environment in the Long Nineteenth Century*. Routledge, 2023, pp. 47–55.

[132] Carl R Woese and George E Fox. "Phylogenetic structure of the prokaryotic domain: the primary kingdoms." In: *Proceedings of the National Academy of Sciences* 74.11 (1977), pp. 5088–5090.

[133] Salvador E Luria and Max Delbrück. "Mutations of bacteria from virus sensitivity to virus resistance." In: *Genetics* 28.6 (1943), p. 491.

[134] Joshua Lederberg and Edward L Tatum. "Gene recombination in Escherichia coli." In: *Nature* 158.4016 (1946).

[135] William Hayes. "Recombination in bact. coil K 12: Unidirectional transfer of genetic material." In: *Nature* 169.4290 (1952), pp. 118–119.

[136] Norton D Zinder and Joshua Lederberg. "Genetic exchange in Salmonella." In: *Journal of bacteriology* 64.5 (1952), pp. 679–699.

[137] Dmitri Iwanowski. *Über die Mosaikkrankheit der Tabakspflanze*. Glagoslav Publications, 2020.

[138] Pakorn Aiewsakun and Aris Katzourakis. "Time-dependent rate phenomenon in viruses." In: *Journal of virology* 90.16 (2016), pp. 7184–7195.

[139] Kousuke Hanada, Yoshiyuki Suzuki, and Takashi Gojobori. "A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes." In: *Molecular biology and evolution* 21.6 (2004), pp. 1074–1080.

[140] Michael Worobey, Paul Telfer, Sandrine Souquière, Meredith Hunter, Clint A Coleman, Michael J Metzger, Patricia Reed, Maria Makuwa, Gail Hearn, Shaya Honarvar, et al. "Island biogeography reveals the deep history of SIV." In: *Science* 329.5998 (2010), pp. 1487–1487.

[141] Clément Gilbert and Cédric Feschotte. "Genomic fossils calibrate the long-term evolution of hepadnaviruses." In: *PLoS biology* 8.9 (2010), e1000495.

[142] Mahan Ghafari, Peter Simmonds, Oliver G Pybus, and Aris Katzourakis. "Prisoner of War dynamics explains the time-dependent pattern of substitution rates in viruses." In: *bioRxiv* (2021), pp. 2021–02.

[143] Samuel Alizon and Christophe Fraser. "Within-host and between-host evolutionary rates across the HIV-1 genome." In: *Retrovirology* 10 (2013), pp. 1–10.

[144] Katrina A Lythgoe and Christophe Fraser. "New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels." In: *Proceedings of the Royal Society B: Biological Sciences* 279.1741 (2012), pp. 3367–3375.

[145] Andrew D Redd, Aleisha N Collinson-Streng, Nikolaos Chatziandreou, Caroline E Mullis, Oliver Laeyendecker, Craig Martens, Stacy Ricklefs, Noah Kiwanuka, Phyu Hninn Nyein, Tom Lutalo, et al. "Previously transmitted HIV-1 strains are preferentially selected during subsequent sexual transmissions." In: *The Journal of infectious diseases* 206.9 (2012), pp. 1433–1442.

[146] AJ Leslie, KJ Pfafferott, P Chetty, R Draenert, MM Addo, M Feeney, Y Tang, EC Holmes, T Allen, JG Prado, et al. "HIV evolution: CTL escape mutation and reversion after transmission." In: *Nature medicine* 10.3 (2004), pp. 282–289.

[147] Christian L Boutwell, Morgane M Rolland, Joshua T Herbeck, James I Mullins, and Todd M Allen. "Viral evolution and escape during acute HIV-1 infection." In: *The Journal of infectious diseases* 202.Suppl 2 (2010), S309.

[148] Joshua T Herbeck, David C Nickle, Gerald H Learn, Geoffrey S Gottlieb, Marcel E Curlin, Laura Heath, and James I Mullins. "Human immunodeficiency virus type 1 env evolves toward ancestral states upon transmission to a new host." In: *Journal of virology* 80.4 (2006), pp. 1637–1644.

[149] Christopher JR Illingworth, Jayna Raghwani, David Serwadda, Nelson K Sewankambo, Merlin L Robb, Michael A Eller, Andrew R Redd, Thomas C Quinn, and Katrina A Lythgoe. "A de novo approach to inferring within-host fitness effects during untreated HIV-1 infection." In: *PLoS pathogens* 16.6 (2020), e1008171.

[150] Jayna Raghwani, Andrew D Redd, Andrew F Longosz, Chieh-Hsi Wu, David Serwadda, Craig Martens, Joseph Kagaayi, Nelson Sewankambo, Stephen F Porcella, Mary K Grabowski, et al. "Evolution of HIV-1 within untreated individuals and at the population scale in Uganda." In: *PLoS pathogens* 14.7 (2018), e1007167.

[151] Joel O Wertheim and Sergei L Kosakovsky Pond. "Purifying selection can obscure the ancient age of viral lineages." In: *Molecular biology and evolution* 28.12 (2011), pp. 3355–3365.

[152] Bram Vrancken, Andrew Rambaut, Marc A Suchard, Alexei Drummond, Guy Baele, Inge Derdelinckx, Eric Van Wijngaerden, Anne-Mieke Vandamme, Kristel Van Laethem, and Philippe Lemey. "The genealogical population dynamics of HIV-1 in a large transmission chain: bridging within and among host evolutionary rates." In: *PLoS computational biology* 10.4 (2014), e1003505.

[153] Jonathan M Carlson, Malinda Schaefer, Daniela C Monaco, Rebecca Batorsky, Daniel T Claiborne, Jessica Prince, Martin J Deymier, Zachary S Ende, Nichole R Klatt, Charles E DeZiel, et al. "Selection bias at the heterosexual HIV-1 transmission bottleneck." In: *Science* 345.6193 (2014), p. 1254031.

[154] Aaron L Halpern and William J Bruno. "Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies." In: *Molecular biology and evolution* 15.7 (1998), pp. 910–917.

[155] Sarah K Hilton and Jesse D Bloom. "Modeling site-specific amino-acid preferences deepens phylogenetic estimates of viral sequence divergence." In: *Virus evolution* 4.2 (2018), vey033.

[156] Vadim Puller, Pavel Sagulenko, and Richard A Neher. "Efficient inference, potential, and limitations of site-specific substitution models." In: *Virus Evolution* 6.2 (2020), veaa066.

[157] Peter Simmonds, Pakorn Aiewsakun, and Aris Katzourakis. "Prisoners of war—host adaptation and its constraints on virus evolution." In: *Nature Reviews Microbiology* 17.5 (2019), pp. 321–328.

[158] Pakorn Aiewsakun and Aris Katzourakis. "Endogenous viruses: Connecting recent and ancient viral evolution." In: *Virology* 479 (2015), pp. 26–37.

[159] Alice Moisan, Fabienne Tombette, Manon Vautrin, Elodie Alessandri-Gradt, Thomas Mourez, and Jean-Christophe Plantier. "I n vitro replicative potential of an HIV-1/MO intergroup recombinant virus compared to HIV-1/M and HIV-1/O parental viruses." In: *Scientific Reports* 14.1 (2024), p. 1730.

[160] Paul J McLaren and Jacques Fellay. "HIV-1 and human genetic variation." In: *Nature Reviews Genetics* 22.10 (2021), pp. 645–657.

[161] Mariana Santa-Marta, Paula Matos de Brito, Ana Godinho-Santos, and Joao Goncalves. "Host factors and HIV-1 replication: clinical evidence and potential therapeutic approaches." In: *Frontiers in immunology* 4 (2013), p. 343.

[162] Pauline C Göller, Tabea Elsener, Dominic Lorgé, Natasa Radulovic, Viona Bernardi, Annika Naumann, Nesrine Amri, Ekaterina Khatchatourova, Felipe Hernandes Coutinho, Martin J Loessner, et al. "Multi-species host range of staphylococcal phages isolated from wastewater." In: *Nature Communications* 12.1 (2021), p. 6965.

[163] Alex Betts, Marie Vasse, Oliver Kaltz, and Michael E Hochberg. "Back to the future: Evolving bacteriophages to increase their effectiveness against the pathogen P seudomonas aeruginosa PAO 1." In: *Evolutionary Applications* 6.7 (2013), pp. 1054–1063.

[164] Adair L Borges. "How to train your bacteriophage." In: *Proceedings of the National Academy of Sciences* 118.28 (2021), e2109434118.

[165] Erdal Toprak, Adrian Veres, Jean-Baptiste Michel, Remy Chait, Daniel L Hartl, and Roy Kishony. "Evolutionary paths to antibiotic resistance under dynamically sustained drug selection." In: *Nature genetics* 44.1 (2012), pp. 101–105.

[166] Elizabeth Kutter and Alexander Sulakvelidze. *Bacteriophages: biology and applications*. Crc press, 2004.

[167] Samuel L Díaz-Muñoz and Britt Koskella. "Bacteria–phage interactions in natural environments." In: *Advances in applied microbiology* 89 (2014), pp. 135–183.

[168] Enea Maffei, Anne-Kathrin Woischnig, Marco R Burkolter, Yannik Heyer, Dorentina Humolli, Nicole Thürkauf, Thomas Bock, Alexander Schmidt, Pablo Manfredi, Adrian Egli, et al. "Phage Paride can kill dormant, antibiotic-tolerant cells of Pseudomonas aeruginosa by direct lytic replication." In: *Nature Communications* 15.1 (2024), p. 175.

Valentin Druelle

Thannerstrasse 68 - 4054 Basel - Switzerland

+3368 953 62 34 • valentin.druelle@unibas.ch

27 years old (13.03.1996) - French

Education

PhD in computational Biology — **University of Basel, Biozentrum - Switzerland**
Viral evolution, HIV-1 and bacteriophages — *October 2019 - March 2024*

Master of applied Physics (Physics engineer) — **EPFL Lausanne - Switzerland**
Physics of complex systems and biophysics — *Sept 2017 - July 2019*

Bachelor of Physics — **EPFL Lausanne - Switzerland**
Major in statistical physics — *Sept 2014 - July 2017*

Baccalauréat Scientifique — **Lycée Champollion Grenoble - France**
Major in SSVT — *Sept 2011 - July 2014*

Skills & Competencies

Research topics:

- Bioinformatics - Sequence analysis, phylogenetics, epidemiology
- Viral evolution - HIV-1, Bacteriophages
- Numerical simulation - Pathogen spread and evolution, physical models

Computer Skills:

- Programming - Python, Julia, C++, Matlab
- CAD and PCB design - Fusion360 (2D, 3D), EasyEDA
- Deep learning - Keras TensorFlow, Pytorch
- Biomedical imaging, Scientific writing in LateX

Additional skills:

- Laboratory skills - Biological and physics laboratories
- Tinkering - Integrated systems, 3D printing, laser cutting, water jet, electronics and PCBs

Languages:

- French - Native
- English - Fluent, C2
- German - Beginner, A2
- Spanish / Portuguese - Basics

Professional Experience

Biozentrum - Richard Neher and Alexander Harms laboratories **Basel, Switzerland**
PhD student *October 2019 - March 2024*
Bioinformatics analysis of HIV-1 evolution. Isolation and characterization of bacteriophages. Creation of an autonomous continuous culture machine for high-throughput directed evolution of bacteriophages. Experimental lab creation and maintenance. Creation and magagement of the Nanopore facility of the institute.

EPFL Physics of Complex Systems laboratory **Lausanne, Switzerland**
Master Thesis *Feb 2019 - August 2019*
Statistical physics of living tissues. Creation of a vertex model for epithelial tissue simulation (C++, Python). Analysis of phase transition impact and comparison with amorphous materials.

EPFL/UNIL **Lausanne, Switzerland**
Student Assistant *Sept 2017 - July 2019*
Supervision and teaching of physics student laboratory work. Tutoring general physics lectures for medicine and physics students.

GE Healthcare **Paris, France**
R&D Internship *July 2018 - Jan 2019*
Project in long-term development research team in medical imaging. Literature review and creation of a physics-based and deep learning-based model for realism improvement (Python, Matlab).

EPFL **Lausanne, Switzerland**
Academic Research Projects *Sept 2017 - Jan 2018*
Master projects in protein folding statistics and inference (C++). Projects in machine learning and deep learning classification (Python).

Publications

1 Enea Maffei, Aisylu Shaidullina, Marco Burkolter, Yannik Heyer, Fabienne Estermann, **Druelle, Valentin**, Patrick Sauer, Luc Willi, Sarah Michaelis, Hubert Hilbi, et al. Systematic exploration of escherichia coli phage–host interactions with the basel phage collection. *PLoS Biology*, 19(11):e3001424, 2021.

2 Richard A Neher, Robert Dyrdak, **Druelle, Valentin**, Emma B Hodcroft, and Jan Albert. Potential impact of seasonal forcing on a sars-cov-2 pandemic. *Swiss medical weekly*, 150:w20224, 2020.

3 Nicholas B Noll, Ivan Aksamentov, **Druelle, Valentin**, Abrie Badenhorst, Bruno Ronzani, Gavin Jefferies, Jan Albert, and Richard A Neher. Covid-19 scenarios: an interactive tool to explore the spread and associated morbidity and mortality of sars-cov-2. *MedRxiv*, pages 2020–05, 2020.

4 Marko Popović, **Druelle, Valentin**, Natalie A Dye, Frank Jülicher, and Matthieu Wyart. Inferring the flow properties of epithelial tissues from their geometry. *New Journal of Physics*, 23(3):033004, 2021.

5 **Druelle, Valentin** and Richard A Neher. Reversions to consensus are positively selected in hiv-1 and bias substitution rate estimates. *Virus Evolution*, 9(1):veac118, 2023.

6 Nicolas Wenner, Anouk Bertola, Louise Larsson, Andrea Rocker, Nahimi Amare Bekele, Chris Sauerbeck, Leonardo F Lemos Rocha, **Druelle, Valentin**, Alexander Harms, and Mederic Diard. Phenotypic heterogeneity drives phage-bacteria coevolution in the intestinal tract. *bioRxiv*, pages 2023–11, 2023.

Presentations

2023: "Aionostat:viral recombination unleashed - Let it do the work !" - Poster presentation at the Biozentrum PhD retreat

2023: "Aionostat:viral recombination unleashed - Let it do the work !" - Poster presentation at the Biozentrum symposium

2023: "Directed evolution with a Morbidostat" - Oral presentation in Leossner and Harms seminar, ETHZ

2022: "Reversion mutations are positively selected and slow down evolution of HIV-1 on the pandemic scale" - Oral presentation at the Biozentrum PhD retreat, awarded best talk prize

2022: "Reversion mutations are positively selected and slow down evolution of HIV-1 on the pandemic scale" - Poster presentation at the Biozentrum retreat

2022: "Reversion mutations are positively selected and slow down evolution of HIV-1 on the pandemic scale" - Poster presentation at Swiss Institute of Bioinformatics (SIB) days conference, Biel

2021: "The contribution of reversions to within and between host HIV-1 evolution" - Oral presentation at the Dynamics and evolution of human virusees conference

2019: "Inferring the flow properties of epithelial tissues from their geometry" - Oral presentation in the Physics of living systems seminar, EPFL

Awards and Achievements

August 2022: Best talk prize (1^{st}) at the Bio- and Pharmazentrum PhD retreat.

August 2019: Awarded the Biozentrum PhD "Fellowship for Excellence" for a duration of 3 years.

July 2019: Congratulation for obtaining the best grade for master thesis manuscript and defence.

January 2018: Academic research project acknowledged with congratulations of supervisor and professor.

July 2014: Jury's congratulations award for Baccalauréat Scientifique with an average of 18/20

Referees

Prof. Richard Neher: Head research group, Biozentrum - University of Basel. richard.neher@unibas.ch, +41 61 207 58 34

Prof. Alexander Harms: Head of Molekulare Phagen-Biologie group, ETHZ. alexander.harms@hest.ethz.ch, +41 44 633 89 33