# Parallel evolution of full-length genomes in a long-term evolution experiment with phage φX174

Eva Bons,† Hélène Chabas,† Hannelore MacDonald, Alejandro Escalera Ledermann, Nicolas Ochsner, Daniel C. Angst, Sebastian Bonhoeffer & Roland R Regoes*

Institute of Integrative Biology, ETH Zürich, Switzerland

† contributed equally

* To whom correspondence should be addressed; E-mail: roland.regoes@env.ethz.ch

April 30, 2025

**Corresponding author:**
Roland R Regoes
Integrative Biology, ETH Zurich
CH-8092 Zurich, Switzerland
tel: +41-44-632-6935
email: roland.regoes@env.ethz.ch

**Preprint servers:** bioRxiv

**Classification:** Biological Sciences/Evolution

**Keywords:** parallel evolution, φX174, long-read sequencing, experimental evolution

## Abstract

The study of evolution in organisms with high mutation rates requires detailed data on the genomic composition of the population. Here we report on an innovative use of high-throughput sequencing technology combined with rapid experimental evolution that allowed us to follow the genetic diversification of four independent bacteriophage populations that we evolved without imposing any direct selection pressure for 412 generations in unprecedented detail. Tracing over 80'000 bacteriophage genomes, of which 884 are unique, we find that evolution is largely neutral but document multiple instances of parallel evolution. Unlike previous observations of parallelism, we do not just see the parallel evolution of mutations at single sites, but on the level of full-length genomes. Computer simulations that recapitulate our experiments in great detail show that this degree of paralellism is inconsistent with neutral evolution. We further show that the observed extent of parallel evolution biases phylodynamic analysis of migration rates, and wrongly estimates significant migration between the independent evolution lines. Our approach is applicable to many other viruses and paves the way to advanced population genetic investigations that aim to shed light on phenomena that relate to genetic linkage across the genome.

## Significance Statement

Genetic sequences of organisms contain imprints of their population biology and migration history. To recover this history, phylodynamic inference has been developed that combines phylogenetics with population biological models. Although phylodynamic methods are being applied widely, they have not yet been experimentally validated. To do this, we evolved a bacteriophage in four independent evolution lines for over 400 generations. We did not impose any direct selection pressure in these evolution experiments because phylodynamic inference works best in neutrally evolving systems. We found that often the same mutations and combinations of mutations evolve in the independent evolution lines. As a result, phylodynamics wrongly estimates migration between the lines. We conclude that the effects of selection should be included into phylodynamic inference.

Evolution is driven by random mutations and subsequent changes in the frequency of organisms due to selection and drift. For complex organisms with low mutation rates these two processes typically occur on separate time-scales. For viruses with high mutation rates, however, the generation of new variants and their frequency changes are intertwined, leading to many concurrent variants (Gerrish and Lenski, 1998, Desai and Fisher, 2007), which makes it particularly difficult to disentangle the contribution of the two processes to viral diversification.

One technical issue that hampers the fine-grained tracing of the molecular evolution of viruses through time is the lack of sufficiently resolved data on the genomic composition of viral populations. One the one hand, full-genome sequences can be generated by picking single strains from a diverse population and sequencing them. This, however, detects at most a few dozen variants that have mostly high frequency in the population (Keele et al., 2008, Li et al., 2010). On the other hand, high-throughput sequencing can detect low frequency mutations, but at the cost of loosing linkage information (Acevedo et al., 2014), requiring to reconstruct haplotypes computationally (Zagordi et al., 2011, Prosperi and Salemi, 2012) which is confounded by the strong assumption that such procedures have to adopt. Here, we report on an approach involving long-read sequencing in combination with computer simulations that overcomes this issue, and successfully determine the compositional changes of four independently evolving populations of φX174, a bacteriophage, during a long-term evolutionary experiment.

We chose the bacteriophage φX174 for this proof-of-concept study on whole-population whole-genome tracking of viral evolution because of its short, well-annotated genome of 5.3kB (Sanger et al., 1977), its fast replication cycle of approximately 26 minutes (Baker et al., 2016), and its high mutation rate (Cuevas et al., 2009). For these reasons, φX174 has been one of the bacteriophages of choice in evolution experiments for over two decades (Bull et al., 1997, Wichman et al., 1999, Miller et al., 2016).

To generate data on the evolutionary dynamics, we evolved φX174 in 4 replicates in a constant and controlled environment of its host cells — *Escherichia coli C* (see Figure 1 and Materials and Methods). During the evolution experiment, we did not impose any external selection pressures, for example, by providing unusual host cells to which the phage was not adapted to, or by inhibiting phage replication. We could not rule out that selection pressures were operating though, and provide evidence for selection below. The rationale of not imposing selection was to obtain an insights into near neutral bacteriophage diversification. We seeded each evolution line with the same, genetically heterogeneous stock population of φX174. We then serially transferred 2% of the bacteriophage population into wells containing fresh host cells for 412 generations. We transferred approximately every 26 minutes, consistent with the duration of one replication cycle of φX174. With a newly developed experimental protocol (see Materials and Methods) we then extracted the DNA of 5% of the final bacteriophage population (generation 412), of an intermediate population (generation 196), as well as the original, "ancestor" bacteriophage population (generation 0), and sequenced this DNA using PacBio sequencing (Eid et al., 2009).

PacBio sequencing requires double-stranded DNA that is turned into a cirular template. Because the average read length of PacBio is above 25 kb, approximately five

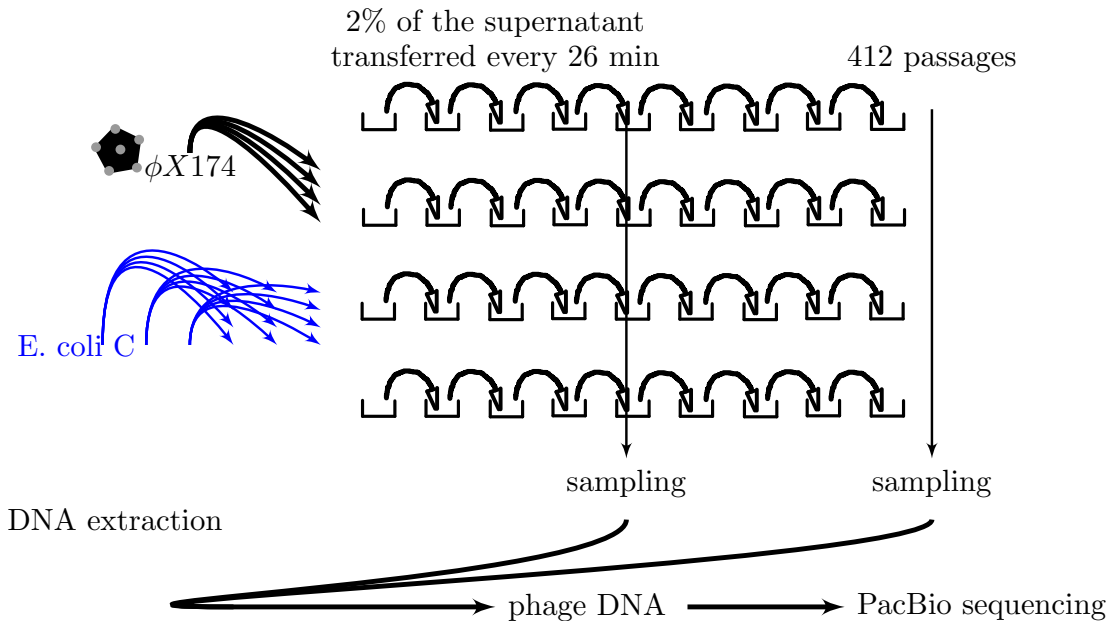



**Figure 1: Experimental design.** Using a liquid handling platform, we seeded wells containing rich medium with *E. coli C* — the host of φX174. After a delay of approximately 90 minutes that allows the host cells to enter the exponential growth phase, the first column of wells were inoculated with φX174. Subsequently, 2% of the supernatant was transferred every 26 minutes into wells of the next column that had been pre-seeded with host cells. The host bacteria in the wells were always seeded from the stock population, while bacteriophages (along with a few bacteria) were passaged from well to well. Thus only the bacteriophage populations could evolve during the experiment. After passages 196 and 412, samples were taken, the bacteriophage DNA was extracted, and submitted to PacBio sequencing. For more details see Materials and Methods.

times the genome length of φX174, the majority of reads contain multiple, (re-)sequenced copies of each sampled DNA molecule. By considering the consistency with which each nucleotide has been called at a given position, one can derive a quality score for each position, and generate the so-called circular consensus read (Acevedo et al., 2014, Wenger et al., 2019). Thus, we can detect and correct sequencing errors, which results in a strong reduction of the uncertainty associated with the PacBio protocol (Wenger et al., 2019). We obtain over 110'000 reads covering the entire genome of φX174, representing over 31'000 unique genotypes (see Figure 2). The majority of these genotypes are only present at very low frequency though. This raises the question whether they were present in the sample or whether they arose due to mutations introduced during the PCR or cell amplification involved in extracting the DNA and preparing it for sequencing, or due to errors introduced during PacBio sequencing that were not corrected through the circular consensus reads.

In order to quantify the extent to which mutations are introduced during the DNA extraction and preparation, we performed *in silico* simulations. These simulations captured the population genetics of the serial passaging experiment, as well as the the methodological steps involved in sampling, preparing and sequencing the viral populations. The exact steps of the experimental setup incorporated in the simulations are shown in Figure 3A and are explained in detail in the Materials and Methods. In brief, the serial passaging experiment is simulated using a previously described sequence evolution simulation tool (Bons and Regoes, 2018), which includes the replication and genetic diversification of the viral population as well as the bottlenecking involved in passaging and has been adapted to φX174. We also incorporated fitness effects of mutations in φX174 that have been experimentally estimated by Domingo-Calap et al. (2009). The simulations also accounted for the methodological steps involved in sampling, preparing and sequencing the viral populations. The sampling of the viral population and the DNA extraction are implemented as a series of bottlenecking and amplification steps that are involved in the protocol. Bottlenecks can lead to the loss of genotypes, while amplification steps can introduce mutations and distort the relative frequencies of genotypes in the sample. Lastly, the simulations incorporate the introduction of sequencing errors at a rate of $2 \times 10^{-5}$ per nucleotide (consistent with restricting the sequencing data to reads with a PHRED score of 47 or higher). The detailed simulation allows us to understand in which exact step sequence variation is introduced, and to which extent this contributes to the variation observed in the final reads.

The simulations allow us to estimate how frequently mutations introduced during the DNA extraction protocol generate genotypes that are not present in the evolved populations. We find that such genotypes with errors do not reach high frequency. Most of the genotypes with errors (70-80%) are introduced during sequencing, while the different DNA extraction steps account for at most 30% of these genotypes. In none of our simulations did a genotype with errors occur more than three times (see Figure 3D).

Because our simulation showed that a genotype that occurs more than three times in the sequence data is unlikely to be the result of methodological error, we constrained our subsequent analysis to genotypes that are above this threshold. Of the 31'229 unique genotypes in the sequencing data only 884 different genotypes occur more than three

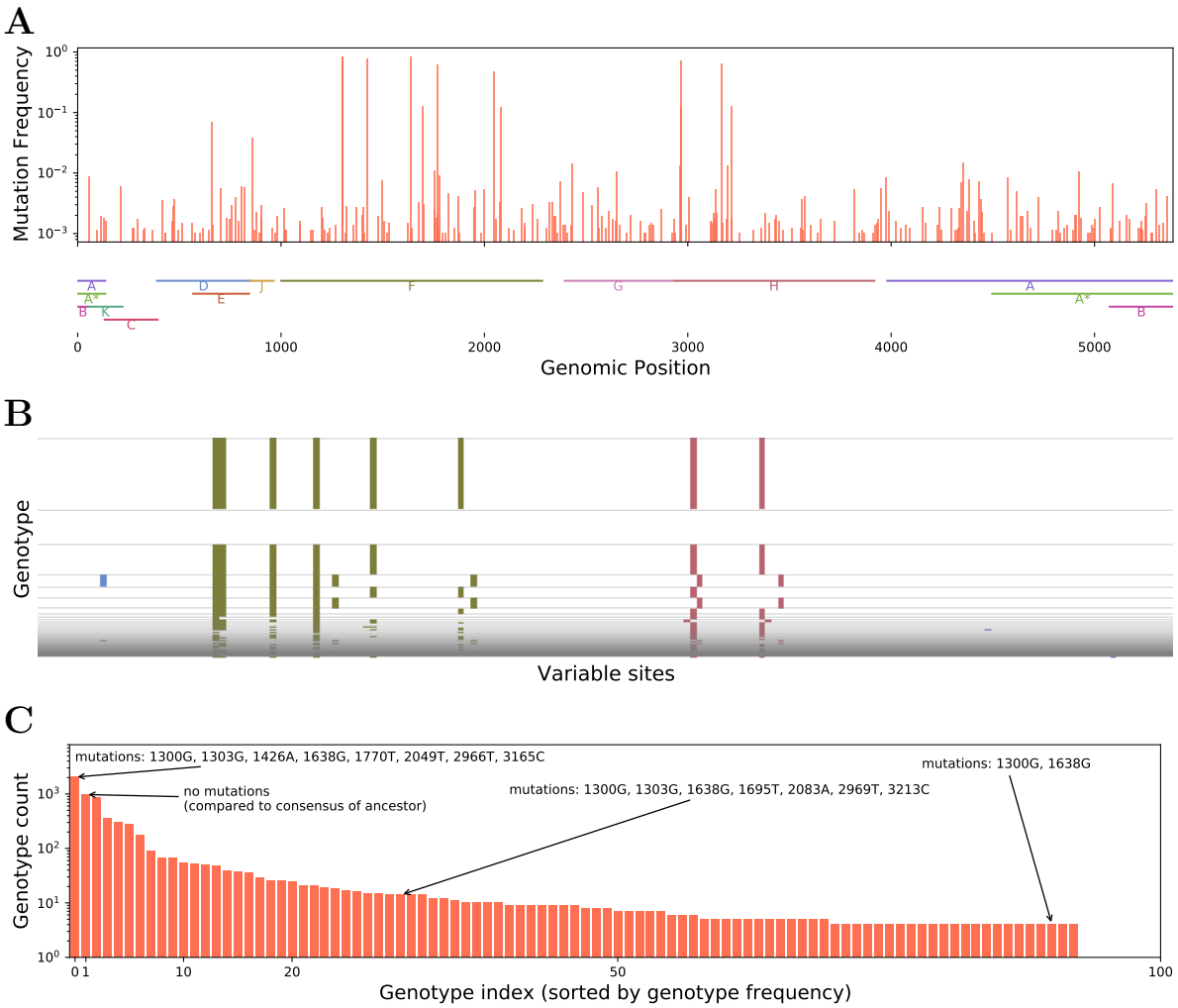



**Figure 2: Overview of the viral sequencing data of the first evolution line after 412 generations at different levels of abstraction. (A)** Frequency of mutations that occurred during the evolution experiment and their location in the genome. The locations of the genes A, A*, B, C, D, E, F, G, H, J, and K of φX174 are indicated below the plot. This representation of the data ignores the genetic linkage between mutations. **(B)** Representation of genotypes present in the population. Each row corresponds to a single genotype, the height of the rows indicate the frequency of the genotypes, The top row displays the most frequent, the second row from the top the second most frequent gentoypes, etc. The x-axis indicates the mutated positions. For the sake of visual clarity, we do show only genomic positions that were mutated at some point in the evolution experiment across all evolution lines and sampling times. Colored vertical lines indicate the positions where the genotype in question was mutated, the color indicates the gene in which this mutation occurs (same colors as in **(A)**). When a mutation occurs at a position with overlapping genes, only one gene is indicated.) This stack of genotypes is the most comprehensive representation, but makes very rare genotypes invisible. **(C)** Genotype count distribution. Genotypes in this plot are indexed by descending frequency in the population (not genetic similarity). This nicely summarizes the number and frequency of unique genotypes in a sample, but ignores the mutations a genotype carries. (We have annotated some of the genotypes with the mutations they carry to make this clear.) Rare genotypes become visible by using the log-scale, which would be very challenging to implement in **(B)**.

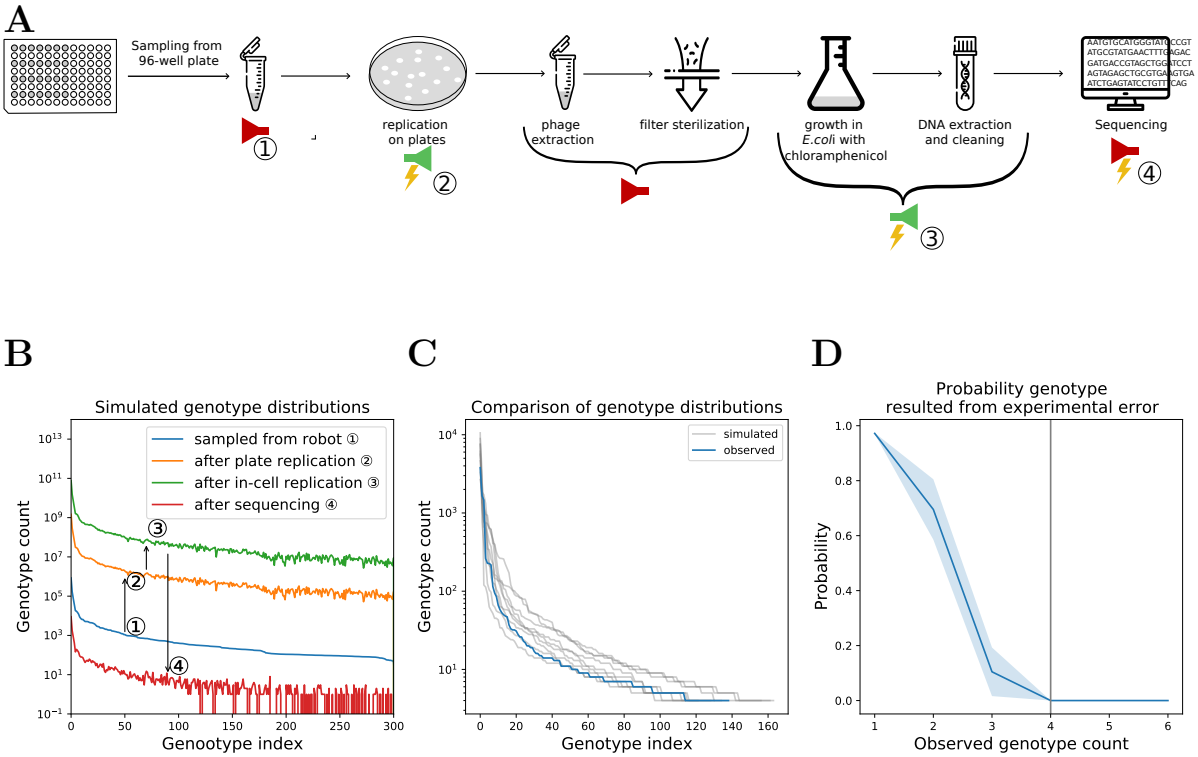



**Figure 3: *In silico* simulations of the passaging experiment, DNA extraction, and sequencing.** **(A)** Overview of experimental steps involved in the sampling, DNA extraction and sequencing of the bacteriophage population. The red symbols indicate experimental steps that were associated with a population bottleneck (e.g. sampling or filtering) The green symbols indicate experimental steps that were associated with an amplification of bacteriophage populations. Lastly, the yellow lightning bolts indicate experimental step during which mutations could have been introduced. The bottlenecking, amplification and mutation processes were implemented in a detailed simulation model (see Materials and Methods). **(B)** Effect of the different steps of the extraction process on genotype diversity. The circled numbers correspond to experimental steps shown in **(A)**. The first steps increase the population size without significantly affecting the genotype distribution. The final sequencing step forms a major bottleneck, reducing the population size with a factor $10^8$ and causing some genotypes to get lost. **(C)** Comparison of the observed to 10 simulated genotype distributions. The simulated genotype count distributions are very consistent with the observed one. **(D)** Probability that a genotype observed after sequencing (step ④) was present in the original population at step ①,based on 10 independent simulations of the experiment and DNA extraction. Genotypes that are present more than 3 times in the sequenced population were very likely present in the original population.

times, representing 80'886 individual genomes. In all samples pooled over every evolution line at both time-points and the ancestor population, there are 184 sites in the φX174 genomes that are mutated compared to the consensus. In the ancestral population, there are 53 genotypes and 38 mutated sites.

The sequencing data that we generated in our evolution experiments is so detailed — especially due to the fact that we have thousands of whole-genome sequences at each sampled time-point — that it presents a challenge for visualization. To overcome this challenge, we present several, partly newly-developed ways to visualize our whole-genome, whole population sequencing data.

A conventional summary plot of the sequencing data for evolution line 1 at generation 412 is shown in Figure 2. This plot shows the frequency of non-consensus nucleotides at each position in the genome. In the sample shown, polymorphisms more frequent than 1:1000 occur in 260 sites across the genome, of which the most frequent occur in the F gene that encodes for the major capsid protein (Figure 2A). This visualization of the sequencing data, however, ignores the linkage of mutations between different sites, and does not show the frequencies of each genotype.

Figure 2B displays the information on the diversity and linkage in a bacteriophage population sample by showing stacks of genotypes. This is similar to highlighter plots as used in the context of HIV, e.g. by Keele et al. (2008). The height of these stacks is proportional to the frequency of the genotype in the sample. The genotypes are displayed as a reduced genome, that contains only the sites that are polymorphic. (In this regard this plot differs from the highlighter plot that typically displays the entire genome.) This visualization, while containing information on the frequency of each genotype, does not reveal low frequency genotypes. The sample of the virus population in evolution line 1 at generation 412 that we display in Figure 2B consists of 6187 genomes, of which 93 are unique. The most prevalent genotype was observed 2035 times, the second most prevalent — the consensus in the ancestor population — was found 973 times. 23 genotypes in this sample were found 4 times (the lowest frequency above the cut-off defined above). These low frequency genotypes are not visible in Figure 2B.

To be able to visualize the full frequency distribution of genotypes in a sample, we developed what we refer to as *genotype count distribution* (Figure 2C). In this plot unique genotypes are indexed in order of decreasing frequency, and the logarithm of their frequency is shown as a histogram. This clearly displays the number and frequency of each unique genotype in a sample, but, unlike a highlighter plot, does not display the specific mutations each genotype carries.

Using these genotype count distributions, the diversification of the bacteriophage population over time can be visualized in great detail (Figure 4). The grey genotype count distributions at the left of each evolution line in Figure 4 show the ancenstor population, which is used to seed every evolution line and consist of 53 unique genotypes. The lightly colored genotype count distributions in each line show the composition of the bacteriophage populations at generation 196. They largely do not overlap with the ancestor population, indicating substantial shifts in the viral composition over time in each evolution line. There is also almost no overlap between the genotype count distributions between the evolution lines. Thus, most genotypes that evolve are new and unique to

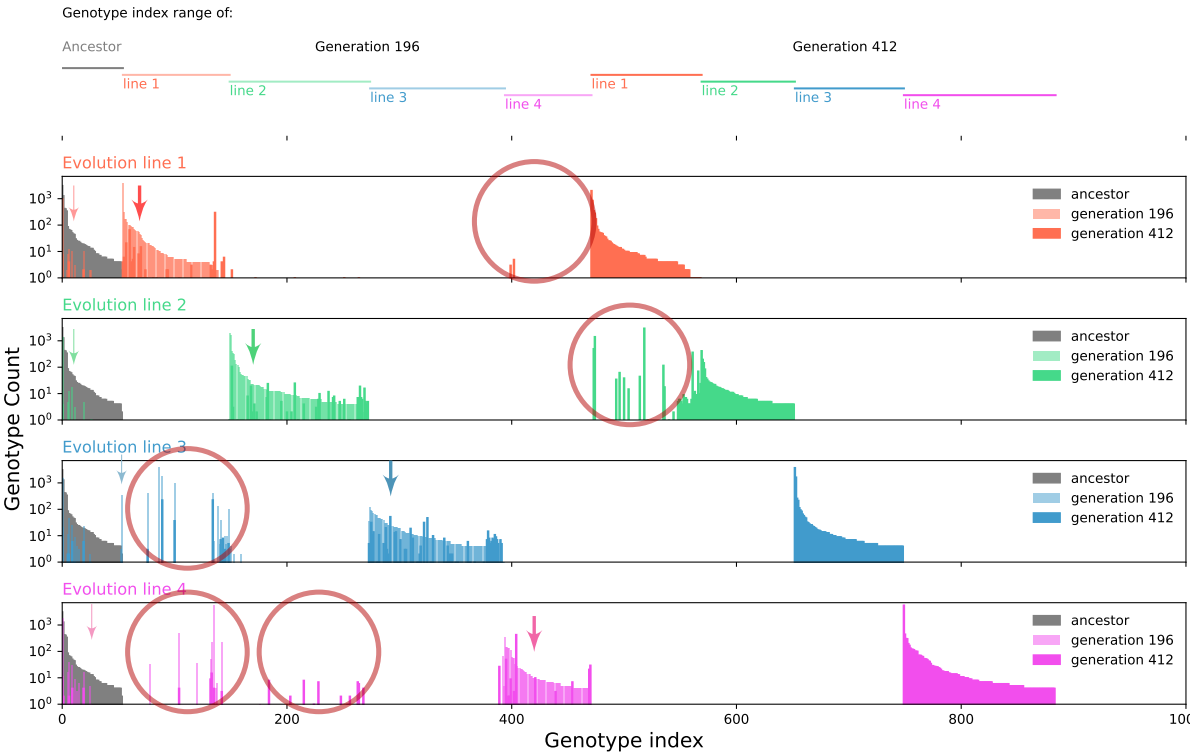



**Figure 4: Diversification of the φX174 populations in each evolution line over time.** Each row represents an evolution line, and shows the genotype count distributions for the ancestral (grey) and evolved viral population after 196 (light colors) and 412 (dark colors) generations. As in Figure 2C, we indexed the genotypes within the viral population at a given time-point by descending order of frequency in the sample. The grey distributions on the left of each line show the counts of the 53 unique genotypes in the ancestral population. The ancestral genotype count distribution is the same in each evolution line. New indices were added for genotypes in subsequent samples that did not occur, and were thus not indexed, in a previous sample. An overview of the index ranges for each evolution line and time-point is shown at the top of the graph. The shift of the genotype count distributions to the right in each line means that most of the genotypes differ between the viral populations from different time points. The low degree of overlap in the genotype count distributions across evolution lines is due to the fact that the newly evolved genotypes are largely unique to an evolution line. Not all genotype are unique to the line and time-point, however. Arrows point to examples of genotypes from the ancestral and mid-point populations that are still present in later samples. The red circles highlight genotypes that evolved in parallel in multiple evolution lines. Note these instances of evolution are parallel on the level of the whole-genome, not just on the level of mutations at single sites in the genomes.

the evolution line in which they evolved — as would be expected in a regime of neutral evolution.

Although, in our passage experiment, we generate mostly unique genotypes over time in each evolution line, there are genotypes that emerged in multiple evolution lines (indicated by red circles in Figure 4). These genotypes are very unlikely to have arisen by cross-contamination during passaging the bacteriophage. First, on the 96-well plate we used for passaging, we separated the wells in which the bacteriophage was passaged by control wells that were seeded with bacteria to detect spill-over (see Materials and Methods). None of these control wells was infected by spilling-over bacteriophage, making it unlikely that spill-over occured across two rows on the 96-well plates. Second, many of the genotypes that emerged in multiple lines were even further apart on the 96-well plate (quite a few between evolution lines 1 and 3, as well as 1 and 4 - see Figure 4). Thus these genotypes represent instances of parallel evolution.

It is important to emphasize that these instances of parallel evolution represent the parallel evolution on the level of whole genomes, rather than parallel evolution of single mutations. In total, we observe 26 genotypes that evolve in more than one evolution line. Compared with the ancestor genotype that is genetically closest, they aquired 1 to 5 additional mutations (mean 2.5). In total, these 26 genotypes carry 18 mutations that are not found in the ancestor population (see the file `stats-for-paper-rrr.html` in the supplementary code). Of these mutations 33% are synomymous, compared to 19% in the ancestor, and 25% in the entire dataset. Establishing parallelism on the whole-genome level requires the whole-population, whole-genome sequencing approach presented here that also captures genotypes of low frequency.

To investigate to what extent the observed parallel evolution is consistent or inconsistent with an accumulation of unbiased and selectively neutral mutations, we performed randomization tests (see Materials and Methods). Specifically, we compared the frequency of the observed parallel individual mutations and combinations of mutations with the frequency predicted under unbiased neutral evolution. (Figure 5).

On the level of individual mutations, we found 137 that emerged in only a single evolution line, while 47 emerged in multiple lines. The fraction of synonymous mutations does not differ significantly between parallel and non-parallel mutations (see the file `stats-for-paper-rrr.html` in the supplementary code). Seven mutations were shared in all four evolution lines (Figure 5A - blue bars). Randomization tests show that this level of parallel evolution is not consistent with unbiased, selectively neutral mutations (Figure 5A - gray bars). This extent of parallelism is similarly high as in previous evolution experiments using the same bacteriophage (Bull et al., 1997, Wichman et al., 1999, 2005) or other viruses (Bertels et al., 2019, Bons et al., 2020).

The linkage information in our data puts us into the unique position to investigate the parallelism of combinations of mutations across the entire genome (Figure 5B–G). We observed the emergence of combinations of up to five mutations in multiple evolution lines (Figure 5E). Randomization tests show that the extent of parallelism is significantly higher than that predicted by neutral evolution. Randomization tests with an effectively shorter genome (see Materials and Methods) show that the parallelism of combinations of mutations cannot be fully explained by the excess parallelism of single mutations.

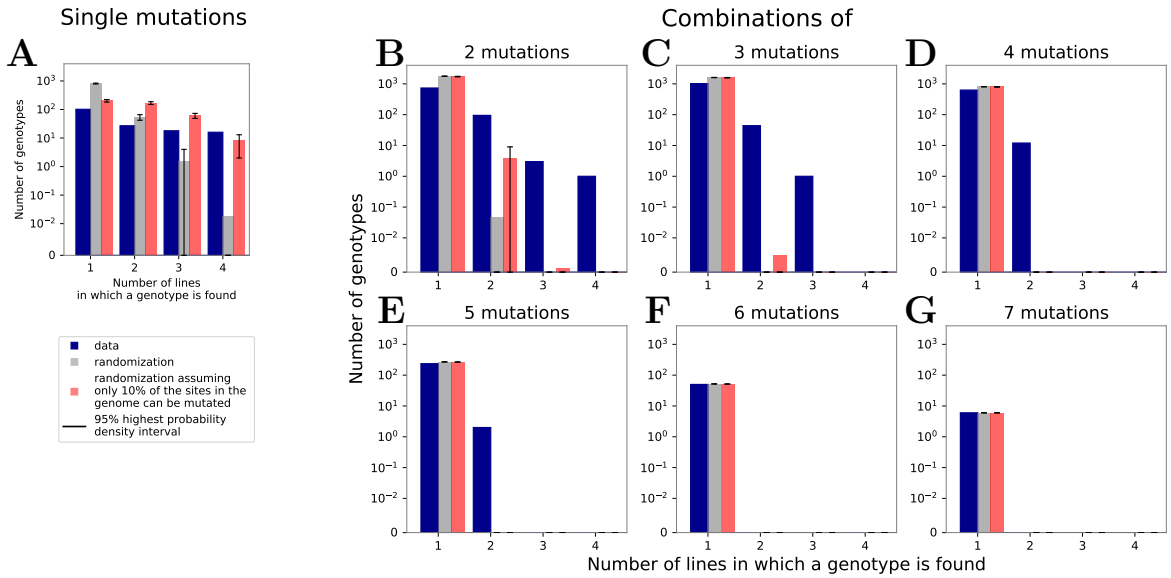



**Figure 5: Parallel evolution exceeds the level predicted by a model of neutral evolution.** The plots show the number of single mutations **(A)**, and combinations of mutations **(B–G)** against the number of evolution lines in which they appear. A mutation that occurs in more than one evolution line represents an instance of parallel evolution. The blue bars show the frequencies of mutations or their combinations observed in the experiments. The grey and red bars show expectations based on 10'000 randomizations (see Materials and Methods). We observe the parallel evolution of combinations of up to five mutations. The randomizations indicate that under neutral evolution combinations of more than two mutations are unlikely to emerge in more than a single evolution line.

This strongly suggests that the selective advantage of the mutations in combination must on average exceed that of the individual mutations involved, i.e. the mutations in many of the parallel-evolved combinations are positively epistatic. These combinations of mutations most likely arose by an accumulation of point mutations during our experiments as an analysis of genetic linkage (see Materials and Methods) did not find any evidence for signatures of recombination.

Viral sequences are widely used for phylodynamics inference which in turn has become key to study the spread and evolution of important viral pathogens (Fraser et al., 2009, Faria et al., 2014, Gire et al., 2014, Düx et al., 2020, Nadeau et al., 2021). However, parallel evolution can lead to biases in phylodynamic inference. To investigate to what extent the parallel evolution we observed in our experiment biases parameters estimates, we conducted a phylodynamic analysis, estimating the migration rates between the independent evolution lines (see Materials and Methods). Because the viral populations evolved independently in each line, there is no migration between the lines. According to the phylodynamic analysis, however, the migration rates between the lines are wrongly estimated to be significantly larger than zero (see Fig S2), even if we exclude the 26 genotypes from the analysis that evolved in parallel in multiple evolution lines. Thus, even in systems as ours, in which no external selection pressure is imposed on the viral population, selection is strong enough to bias phylodynamic inference.

In conclusion, we present an approach to capturing the molecular evolution of viruses with short genomes in much more detail than was previously possible. The approach gives rise to sequence reads spanning the entire genome without the typical loss of linkage information, and information on their frequency in the sample — data that, at this level of detail, do not exist for any organism to-date. The picture of bacteriophage evolution we document is characterized mostly by neutral evolution as most of the genotypes we find evolve *de novo* in each evolution line are unique to that line. But we also established instances of significant parallel evolution of many mutations and their combinations, and even of 26 genotypes that were identical across the entire genome. We also showed that this level of parallel evolution leads to an over-estimation of migration rates in phylodynamic analyses.

In addition to the detailed insights into the long-term viral evolution, this approach will open new avenues to trace viral population genetics on the whole-genome level, i. e. observe the emergence, drift and selection of mutations in combination with the genetic background on which they appear. It will also allow the direct observation of clonal interference of genotypes. This will lift the estimation of fitness and epistatic effects — that are at the heart of population genetics — to a new level.

# Materials and Methods

## Strains, growth conditions and bacteriophage titration

Except otherwise stated, *Escherichia coli* C (ATCC 13706) is grown in LB Miller broth at 37°C with aeration (180 rpm). When infected by bacteriophage φX174 (ATCC 13706-B1, ordered from the DSMZ collection), the broth is supplemented with 10 mM $MgCl_2$ and 2 mM $CaCl_2$.

The bacteriophage is amplified by inoculating 50 mL of broth with 500 $\mu$L of stationary phase bacteria and 250 $\mu$L of bacteriophage stock. Bacteriophage amplification occurs during overnight incubation at 37°C with aeration (180 rpm). Bacteriophage is extracted by centrifuging the tube for 5 minutes at maximal speed and filtrating the supernatant (pores diameter 0.22 $\mu$m).

Bacteriophage titration on plates is carried out using a standard plaque assay. Briefly, 6 mL of Soft Agar (LB Miller with 5 g/L of Agar, 10 mM $MgCl_2$ and 2 mM $CaCl_2$) containing 400 $\mu$L of stationary phase bacteria and 50 $\mu$L of lysate diluted in LB Miller, are poured on top of a petri dish containing Hard Agar (LB Miller with 15 g/L of Agar, 10 mM $MgCl_2$ and 2 mM $CaCl_2$). The plate is then incubated upside-down overnight at 37°C. Plaques are counted to determine the initial concentration of φX174.

Bacteriophage titration in liquid is carried out in 96-well plates. Each well of the plate is inoculated with 200 $\mu$L of LB Miller, 1% v/v bacteria in stationary phase and bacteriophages at a concentration of around 0.5 bacteriophages/well. The plate is incubated overnight at 37°C. After growth, the presence of the bacteriophage in each well is assessed by stamping 2 $\mu$L of each well on a bacterial lawn, similar to one used for plaque-assays without the phage. The plates are incubated at 37°C for 5 hours. The number of wells are counted that contain no bacteriophage growth. We back-calculated the viral concentrations from the frequency of wells without bacteriophage growth by assuming that the number of bacteriophages follows a Poisson distribution.

## Experimental evolution of bacteriophage φX174

Bacteriophage phiX174 is evolved for 412 generations in 4 parallel lines. The experimental evolution of bacteriophage φX174 was carried out in 96-well plates using a a Tecan Evo 200 automated liquid handling system (Tecan) with an integrated automated incubator (Liconic STX100, Liconic). At each transfer, 4 $\mu l$ of supernatant was passaged into the next well.

The passaging terminated in the liquid handling system 3 times at generation 95, 150 and 309. At generation 95, only 2 $\mu l$ of supernatant was passaged.

Bacteriophage were passaged from column to column on a 96 well-plate approximately every 26 minutes, allowing for one complete replication cycle in each well (Baker et al., 2016). On a single column, 4 replicated lines of bacteriophage φX174 evolves, each separated by a line containing only host bacteria. This allows the detection of contamination events: a delay in bacterial growth in these lines indicates a contamination by the bacteriophage. For each generation n, we realise a transfer (described in the next paragraph)

and then incubate the plates for 20 minutes at 37°C and 95% relative humidity. Once a day, we make a glycerol stock by mixing 100 $\mu$L of 40% glycerol with 100 $\mu$L of the n-1 bacteriophage generation. The stocks are then stored at -80°C. In case of bacteriophage contamination or other problems, the experiment is restarted using the last stock as a bacteriophage source.

Each transfer is composed of several steps with pipette sterilisation after each step (see below). First, on column n+6, 5 $\mu$L of freshly grown overnight *E. coli* C are used to inoculate 200 $\mu$L of LB Miller supplemented with 10 mM $MgCl_2$ and 2 mM $CaCl_2$. Second, 4 $\mu$L of freshly mixed column n-1 are transferred into column n. As a result, bacteria are inoculated approximately 1h30 before bacteriophage infection: this assures that bacteriophage infection occurs on exponentially growing bacteria. In addition, this protocol aims to ensure that the multiplicity of infection stays largely below one, which limits bacteriophage recombination. Third, if needed, glycerol stock of bacteriophages are made as previously described.

The entire experiment is carried out with reusable tips, which require sterilisation between each step to inactivate any bacteria or bacteriophage particle. To do so, the tips are washed twice in 5.6% bleach, flushed from behind with water, washed in 70% ethanol and then in milliQ water.

To calculate the mulitiplicity of infection (MOI), we estimated the density of *E. coli* and φX174 in a well. We calculated *E. coli* density by performing a standard colony forming unit calculation as follows: 5 $\mu$L of stationary phase bacteria are inoculated into a well with 200$\mu$L of LB Miller. After 90 minutes of incubation at 37°C, a serial dilution is prepared from the culture to back-calculate the density in the well. 100$\mu$L of each dilution are then spread onto a petri dish containing Hard Agar (LB Miller with 15 g/L of Agar, 10 mM $MgCl_2$ and 2 mM $CaCl_2$). The plate is incubated upside-down overnight at 37°C. We counted bacterial colonies and determined that there were $1.84 \times 10^7$ bacteria/well. We calculated φX174 density using the plaque forming unit protocol described above at generations 0, 196 (midpoint) and 412 (endpoint). We report phage density in 4$\mu$L as this is the volume used to initiate the experiment and transfer between wells each generation. The phage densities are as follows: $1.44 \times 10^6$ at generation 0, $2.42 \times 10^7$ at generation 196 and $5.12 \times 10^7$ at generation 412. Overall, these densities correspond to roughly 1:1 ratio of bacteria:phage and thus an MOI of around 1.

## Extraction of the φX174 DNA

For our study, we developed two methods to extract the DNA of φX174 from our experimental samples. These two protocols extract φX174 DNA at different stage of its life cycle (dsDNA vs ssDNA) and use different technique to amplify and extract the DNA. We used the dsDNA extraction protocol on all samples.

The data we show in the main paper are based on this extraction protocol. To compare the dsDNA and ssDNA protocols, we have extracted the bacteriophage DNA using with both protocols for two samples: evolution line 3, generation 200 and evolution line 4, generation 412. Figure S1 shows a comparison of the consistency of the detected

genotypes and their frequency. The dsDNA extraction protocol generated more unique genotypes and higher counts. Due to its higher efficiency, we chose it as the main protocol and applied it to all the samples.

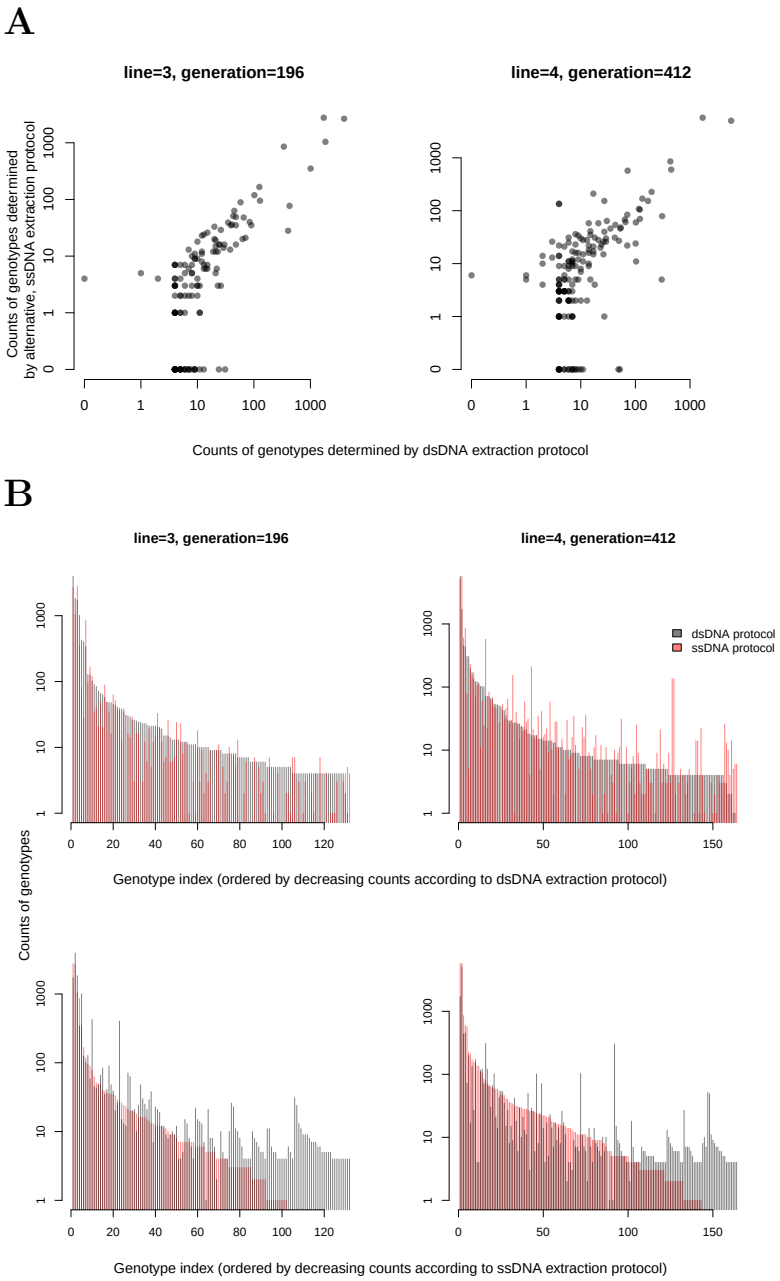



**Figure S1:** Comparison of the dsDNA extraction protocol and the alternative ssDNA extraction protocol in 2 of our 9 samples. **(A)** Comparison of the frequencies of the genotypes. **(B)** Comparison of the genotype counts distributions. The top and bottom rows display the genotype counts order by decreasing counts according to the dsDNA and ssDNA extraction protocol, respectively. The dsDNA extraction protocol was more efficient as it generates more unique genotypes and higher counts.

## dsDNA extraction

A ready to use version of this protocol is available as supplementary information.

As part of its life cycle, bacteriophage φX174 forms an intermediate circular dsDNA genome and we can block the bacteriophage cycle in this stage by adding chloramphenicol (Sinsheimer et al., 1962, Denhardt and Silver, 1966). We take advantage of this and use a plasmid DNA extraction protocol to extract this DNA. In details, we use a 5-step protocol.

First, we prepare a concentrated bacteriophage lysate deriving from our evolution experiment as described above.

Second, we infect bacteria with this lysate, block the bacteriophage life cycle and incubate for genome amplification. To do so, we inoculate 200 mL of broth with 2 mL of bacteria in stationary phase and incubate at 37°C for 1h30. After this incubation, we supplement the broth with 2 mM $CaCl_2$, 10 mM $MgCl_2$ and the 2 mL of bacteriophage lysate previously prepared. To decrease recombination, we confirm that the bacteriophage MOI remains below 1 (see below at the end of the protocol). We incubate at 37°C with aeration for precisely 5 minutes and then add 600 $\mu$L of chloramphenicol at 100 mg/mL. We then incubate at 37°C with aeration for 3 hours.

Third we extract φX174 DNA by using the Monarch Plasmid DNA Miniprep Kit (NEB Cat. T1010) with the following protocol. We split each sample in 4 50-mL subsamples. We centrifuge at max speed for 5 minutes each of these, discard the supernatent and resuspend the pellet in 200 $\mu$L of the resuspension buffer. Then, we lyse bacterial cells by adding 200 $\mu$L of the lysis buffer and gently mixing. As soon as the cells are lysed, we neutralyse the lysis buffer by adding 400 $\mu$L of the plasmid neutralization buffer. After a short incubation at room temperature, we centrifuge at 16 000g for 5 minutes. Then, we carefully transfer the supernatent of the first subsample in one column and centrifuge it for 1 minute at 16 000g. We discard the flowthrough and we reuse the same column for subsamples 2-4 with a 1 minute 16 000g centrifugation between each subsample. We then clean the DNA according to the manufacturer instruction and resuspend the DNA in DNAse free water.

Fourth, we open the circular DNA by adding 10% v/v of NEB Cutsmart buffer (NEB Cat. B7204) and 1 $\mu$L of the restriction enzyme Pst1-HF (NEB Cat. R3140). The digestion is carried at 37°C for 10 minutes.

Fifth, we clean the DNA using Ampure beads as described above (final resuspension is made in 20 $\mu$L) and assess its quality by measuring its concentration (Qubit measurements using the broad range dsDNA kit), its chemical purity (using nanodrop) and checking that no other DNA has been carried through (using an Agilent Tape Station with its genomic DNA screen tape). For these 3 measurements, we follow manufacturer instructions.

To minimize recombination during φX174 DNA amplification in 200 mL flasks, we confirmed that the MOI was below 1. Therefore, we quantified the phage titre after overnight agar plate amplification and the bacterial concentration after 1h30 minutes growth in 200 mL flasks. We titre the lysates obtained from two samples after overnight growth on Agar plate and we find a concentration of $1.2 \times 10^9$ PFU/mL and $4.4 \times 10^8$

PFU/mL. We also measure bacterial concentration after 1h30 of growth for 1 replicate and find a concentration of $2.94 \times 10^7$ CFU/mL. As the flask contains 200 mL, this gives us a total of $5.88 \times 10^9$ CFU. As for the dsDNA extraction protocol, we put maximum 2 mL of phages that has been amplified on Agar plate, the MOI is below 1.

## ssDNA extraction

A ready-to-use version of this protocol is available as supplementary information.

Bacteriophage φX174 genome is composed of a 5.3 kb ssDNA genome. Here, we use a 5-step ssDNA extraction protocol.

First, we prepare a concentrated bacteriophage lysate deriving from our evolution experiment. To do so, we amplify 10% of each sample on a Petri dish by pouring 6 mL of Soft Agar supplemented with 400 $\mu$L of stationary phase bacteria and 20 $\mu$L of bacteriophage sample on top of a Hard Agar lawn. Each plate is incubated overnight at 37°C and we extract the bacteriophage by adding 3 mL of LB on top of it and careful mixing. We retrieve 2 mL of this LB and filter it (pore size: 0.22 $\mu$m). The bacteriophage concentration after this step is available in the supplementary material.

Second, we precipitate the bacteriophage, remove bacterial nucleic acid and open the bacteriophage capsid. To do so, we mix 1 mL of the lysate with 500 $\mu$L of a solution of PEG30%, NaCl 1.5M and we let this solution rest at 4°C for 24 hours. Then, we centrifuge at 13 000g for 10 minutes at 4°C and discard 1 mL of the supernatant. We centrifuge again the remaining part for 5 minutes in the same conditions. We discard the supernatant and resuspend the pellet in 200 $\mu$L of DNase buffer A (New England Biolabs (NEB), Cat. B0303S). We leave on ice for 10 minutes and add 2 $\mu$L of DNase 1 (NEB, Cat. B0303S) and RNase A (Thermo Scientific, Cat. EN0531) and we incubate for 10 minutes at 37°C. To inactivate the DNase, we add 20 $\mu$L of 50 mM EDTA, pH8 and incubate for 10 minutes at 75°C. We open the bacteriophage capsid by adding 5 $\mu$L of 20% SDS and 5 $\mu$L of pronase (from Sigma-Aldrich, CAS Number: 9036-06-0) and incubate at 40°C for 2 hours and 30 minutes.

Third, we ethanol precipitate the DNA to remove all the chemicals from step 2. To do so, we precisely measure the volume of our sample and add 10% v/v of a NaCl 5 M solution and 2 volume of ice-cold 100% ethanol. The solution is then incubated overnight at -20°C. Then, we centrifuge the samples at max speed for 10 minutes at 0°C, remove the supernatent, add 200 $\mu$L of ice cold 70% ethanol and centrifuge for 2 minutes in the same conditions. We discard the supernatent and let the pellet dry for 15 minutes at room temperature before we resuspend it in 100 $\mu$L of DNase free water.

Fourth, we amplify the DNA using a whole genome PCR protocol. For one PCR tube, the PCR mix is composed of 21.5 $\mu$L of nuclease-free water, 10 $\mu$L of Q5 polymerase buffer (NEB Cat. M0941), 2 $\mu$L 100% DMSO, 2.5 $\mu$L of both forward (ATC GCT TCC ATG ACG CAG AA) and reverse (TGC AGG TTG GAT ACG CCA AT) primers at 10 $\mu$M, 1 $\mu$L of 10 mM dNTP, 0.5 $\mu$L of Q5 polymerase (NEB Cat. M0941) and 10 $\mu$L of bacteriophage DNA. For each sample, we conducted 5 independent PCRs. The PCR program is composed of an initial denaturation at 98°C for 30 seconds and 32 cycles of the followings steps: 1) denaturation at 98°C for 10 seconds, 2) annhealing at 68°C for

30 seconds and 3) elongation at 72°C for 108 seconds. At the end of these cycles, there is an additional final elongation at 72°C for 2 minutes before a storage at 4°C. To confirm correct amplification, we use an Agilent Tape Station with its genomic kit according to manufacturer instruction.

Fifth, we clean the DNA and confirm its quality prior to PacBio sequencing. To do so, we combine all the PCR tubes coming from the same sample, measure precisely the total volume and mix with 1 volume of Ampure beads. We incubate with agitation at room temperature for 10 minutes before putting the tube on a magnetic rack until the beads separate from the solution ($\approx$ 3 minutes). We remove the supernatent and add 1 mL of freshly made 70% ethanol (wash solution). We remove this wash solution and add another 1 mL of fresh wash solution. This step is repeated 2 times. Then, we remove as much as possible of the supernatent and resuspend the beads in 80 $\mu$L of EB Buffer (NEB Cat. T1016) that has been preheated at 50°C. We then remove the tubes from the magnetic rack and incubate it for at least 15 minutes at 37°C. After incubation, we put it back on the magnetic rack, wait for the solution to become clear and keep 75 $\mu$L of the supernatent. Finally, we assess the quality of the DNA and its concentration using a Nanodrop and a Qubit (br dsDNA kit) according to manufacturers instruction.

## PacBio sequencing

The DNA is sent for long read sequencing to the Functional Genomics Center in Zürich. Sequencing was performed on a PacBio Single molecule sequencer Sequel SMRT Cell 1M.

The long-read sequence data generated in the context of this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB80636 (`https://www.ebi.ac.uk/ena/browser/view/PRJEB80636`), and will be made publicly available upon acceptance.

## Simulations of the passaging experiment, DNA extraction, and sequencing

The simulation of our entire experiment, entails first a simulation of the viral populations and its diversification throughout 400 passages. This results in approximately $2\times10^6$ virions with potentially different genotypes (although many virions carry identical genomes).

Subsequently, the simulated virions are sampled binomially according to the volume sampled from the wells in our experiments, which leads to a loss of rare genotypes and a distortion of their original frequencies. The sampled virions are then plated, their DNA is extracted, and, when simulating the ssDNA extraction protocol (see above), PCR amplified, or when simulating the dsDNA extraction protocol (see above), replicated in cells. The simulated plating, PCR amplification, and replication in cells can introduce errors into the genomes. Lastly, the simulated genomes are sequenced, which is implemented as binomial sampling and a worst case sequencing error rate.

For all the details on these simulations please consult the file "code/simulations.py" on the git repository.

## Simulating sequences as expected under neutral evolution

To generate genotypes that would be expected under neutral evolution, we randomized the locations of mutations according to the following procedure.

First, we infer which of the mutations of a genotype have newly emerged during the passage experiment, rather than having been already present in the ancestral population. We do this by identifying the genotype in the ancestral population that has most mutations in common with the genotype in question. We assume this specific ancestral genotype to be the evolutionary origin, and consider all the mutations of the genotype not already present in the specific ancestor to have newly emerged. In case of multiple ancestral genotypes with the same number of shared mutations from the genotype in question, we assume the evolutionary origin to be the one that is most frequent in the ancestral population.

Second, we randomize the positions of these likely newly emerged mutations. Formally, if we have $n$ newly emerged mutations, we sample without replacement n out of $\{1, \ldots, \lambda\}$, where $\lambda = 5386$ is the length of the genome of φX174. (Thus, if, for example, a genotype has likely acquired a mutation at positions 1770 and 2966, this randomization could generate positions 3328 and 4599.) For the sake of simplicity, we randomize only the positions, rather than the positions and specific nucleotides substitutions. As a result, parallel evolution becomes more likely in these simulations, and the resulting randomization test is conservative, i.e. it underestimates the extent and significance of a deviation from neutrality.

Third, we adopted this procedure to generate viral populations with mutational patterns that would be expected under neutral evolution for each evolution line. Repeating this 10'000 times for all 31'229 unique genotypes, we can estimate the probability of a mutated position being shared across multiple evolution lines, and compare it to the observed frequency (Figure 5A). In addition to considering single mutated positions, we also investigated combinations of mutated positions (Figure 5B–G). The clear conclusion from this analysis is that there is much more sharing of single mutated positions, and combinations of mutated positions than expected under neutral evolution (see main text). This could be due to a selective advantage that these mutations confer, or to a higher mutation probability at the positions that are shared across the evolution lines.

Last, we investigated if combinations of mutations are shared more than expected given that already single mutations are shared more than expected under neutral evolution. To this end, we set the effective genome length to $\lambda = 539$ instead of $\lambda = 5386$. As a consequence, shared mutations become more likely. At this particular effective genome length, the observed number of shared mutated positions across four evolution lines is comparable to the neutral expectation (Figure 5A). The number of mutated positions shared across two or three evolution lines even exceeds this expectation, which makes the following analysis conservative.

In contrast to single mutated positions, the observed number of shared combinations

of mutated positions is still much lower than the expectation at this low effective genome length (Figure 5B–G). This is strong evidence that these combinations have a selective advantage that exceeds the potential advantage that the individual mutations confer, i.e. interact under positive epistasis. Note that, while we cannot disentangle potential mutational bias from selective advantage for single mutations, once the effect of single mutations is subsumed by shortening the effective genome length, deviations from the neutral expectation must be due to positive epistatic interactions between the mutated sites.

## Linkage disequilibrium analysis to probe for genetic signatures of recombination

To investigate if recombination occurs to an extent that it leaves a detectable trace in the viral genomes, we performed a bioinformatic analysis of the sequencing data. We specifically searched for patterns of genetic linkage that are indicative of recombination.

The well-known relationship between linkage disequilibrium and recombination rate by Hill and Robertson (1968) (Hartl and Clark, 2007, Hill and Robertson, 1968) cannot be used for our purpose because of some of the assumptions that go into the derivation of the relationship. The formula applies only if there are no new mutations (which occur during our experiment), if there is no selection (which we assume is acting), and, most importantly, when an evolutionary equilibrium has been reached (which is not the case in our experiments either).

We therefore used the even older formula for the decline of the linkage disequilibrium over time (Slatkin, 2008, Jennings, 1917). From one generation to the next, the linkage disequilibrium between two mutations, $D = p_{12} - p_1\ p_2$, declines by a factor $(1 - c)$. Hereby $p_1$, $p_2$, and $p_{12}$ denotes the frequencies of genotypes carrying mutation 1, 2, or both mutations, respectively, and $c$ denotes the probability of recombination. The recombination rate is commonly assumed to scale with the distance between the mutations on the genome, and can hence be rewritten as $c = c_{1bp}\ d$, with $c_{1bp}$ denoting the recombination rate at a genomic distance of a single base pair, and $d$ denoting the genomic distance in base pairs.

Extrapolating this decline of $D$ over time, we obtain the following relationship between the linkage disequilibrium $D$, the genomic distance of two mutations $d$, the recombination rate $c_{1bp}$, and the generation $g$ in our experiments:

$$D(g) = D(g = 0)\ (1 - c_{1bp}\ d)^g \tag{1}$$

We fitted this relationship (Equation 1) to all the mutation pairs that have a non-zero $D$ in the ancestor population. We normalized all measurements of $D$ by dividing them by the non-zero $D$ in the ancestor population. We obtain an estimate for $c_{1bp}$ that is not significantly different from 0 ($c_{1bp} = 0.0004 \pm 0.0015$). Thus, there is no signal for linkage disequilibrium decline over generations, and hence no evidence for recombination in our experiments. While this statistical test does not rule out any instance of recombination, it supports the assumption that recombination has not affected the evolutionary dynamics in a major way.

## Phylodynamics estimation of migration rates between the four evolution lines

To investigate to what extent the instances of parallel evolution that occur in our evolution experiment can bias the estimation of migration rates between the evolutionary lines, we conducted a phylodynamic analysis. As the evolution lines in our experiments evolve independently, i. e. there is no migration between these lines, the migration rates should be estimated to be zero.

In our experiments, we obtained a total of 80'886 whole-genome sequences that appeared more than 3 times in one of the samples. Of these sequences 884 are unique, i.e. many of the sequences are identical. This leads to the problem of polytomies in the reconstruction of the phylogenies, on which a phylodynamic analysis is built.

The high number of sequences and the polytomies make a one-step Bayesian estimation of the migration rates using the software BEAST (Drummond and Rambaut, 2007, Bouckaert et al., 2014, Vaughan et al., 2014, Bouckaert et al., 2019) prohibitive. We therefore resorted to a simpler, non-Bayesian two-step estimation procedure, ignoring the phylogenetic uncertainty inherent in the reconstructure of the tree. Technically, we conducted this analysis using *treetime* (Sagulenko et al., 2018), an efficient algorithm to generate timed phylogenetic trees from sequence data. Treetime allows a two-step *mugration* analysis, which estimates migration rates from a phylogenetic tree analogously to mutation rates by adding a locus to the genomes that has its own mutation rate and designates the deme in which the sequences was sampled. The mutation rates between the, in our case, five demes then represent estimates of the migration rate between them.

To circumvent having to analyze all 80'886 sequences, we adopt the common strategy to subsample. We subsample 900 sequences, 100 from the ancestor population and another 100 from generation 196 and 412 of each of the four evolutiom lines, and repeat this analysis 100 times. This subsampling approach captures parts of the phylogenetic uncertainty that we did not incorporate directly.

Another complication that arises in the phylodynamic analysis of our experimental data is that the ancestor population, with which the four evolution lines were seeded, consists of multiple genotypes that arose through a common evolutionary history. This "panmictic past" will leave a signal for migration even in the independent lines. Our solution of this issue is to assign the ancestor its own deme, from which migration to each evolution line is possible. (Each evolution lines is also assigned a deme of its own.) This way, the common evolutionary history of the genotypes in the ancestor population can be separated from the potential migration of genotypes between evolution lines in the analysis (see Table S1).

Figure S2A shows the distribution of estimated overall migration rates between evolution lines. In this distribution, the migration from the ancestor to the evolution lines has been subtracted. In a nutshell, the migration rate estimates are significantly different from zero, which they should not be because there is no migration between the evolution lines in the experiment. The point estimate is 0.00275 per generation, which corresponds to a per generation probability of 0.275% for a genome to migrate from one evolution line to the other. Over the course of the 412 generation experiment, this

**Table S1:** Migration rate estimate from each deme to each other deme (average over estimates for the 100 subsamples). Deme A is the ancestor population, demes EL1–4 are evolution lines 1–4. The direction of migration goes from column to row. The estimated rates of migration from the ancestor population to the evolution lines can be seen in the first column of the matrix, and has consistently high values. This is consistent with the fact that each evolution line has been seeded with the ancestor population, i. e. migration actually happened in the experiment. The other four columns show estimates of migration between the evolution lines, and from the evolution lines back to the ancestor. These migration rates are zero in the experiment, but are wrongly estimated to be non-zero. This bias arises because of parallel evolution between the evolution lines. Between evolution lines 1 and 2, for example, there are particularly many genotypes that evolved in parallel by generation 412 (see Figure 4), leading to the largest migration rate estimates between these lines (0.00168 and 0.00182). The basis to calculate the overall migration rate estimate between the lines were columns 2–5 in the migration rate estimate matrices. Leaving out the first column in calculating the overall migration rate ensures that the migration rate is not confounded by the "panmictic past" of the ancestor population.

| | A | EL1 | EL2 | EL3 | EL4 |
|---|---|---|---|---|---|
| A | 0.00000 | 0.00021 | 0.00020 | 0.00022 | 0.00019 |
| EL1 | 0.00153 | 0.00000 | 0.00182 | 0.00095 | 0.00098 |
| EL2 | 0.00138 | 0.00168 | 0.00000 | 0.00065 | 0.00059 |
| EL3 | 0.00132 | 0.00077 | 0.00056 | 0.00000 | 0.00062 |
| EL4 | 0.00145 | 0.00104 | 0.00069 | 0.00084 | 0.00000 |

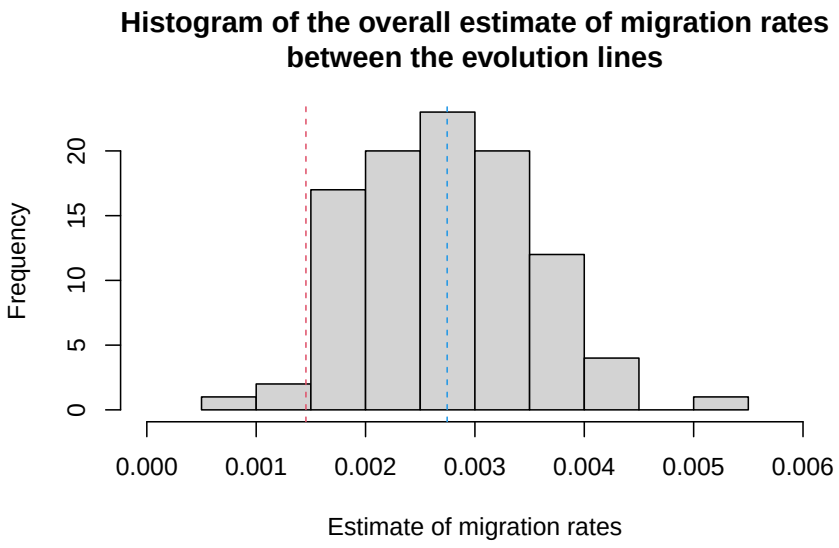
migration rate amounts to a 68% probability for each single genome to migrate to a different evolution line. Qualitatively, these insights also hold when we exclude the 26 genotypes that evolved in parallel Figure S2B.

This positive migration rate, rather than reflecting real migration, constitutes a bias that arises because the analysis does not account for the selection pressures that generate the patterns of parallel evolution in our experiment.

## Code

The computer code developed and used for simulating phage evolution, analysing linkage disequilibrium, and conducting the migration rate estimation with treetime has been submitted as a zip-compressed file along with this manuscript, and will be deposited at `https://github.com/RolandRegoes/PhagePhylodynamics` upon acceptance.

**A**




**B**

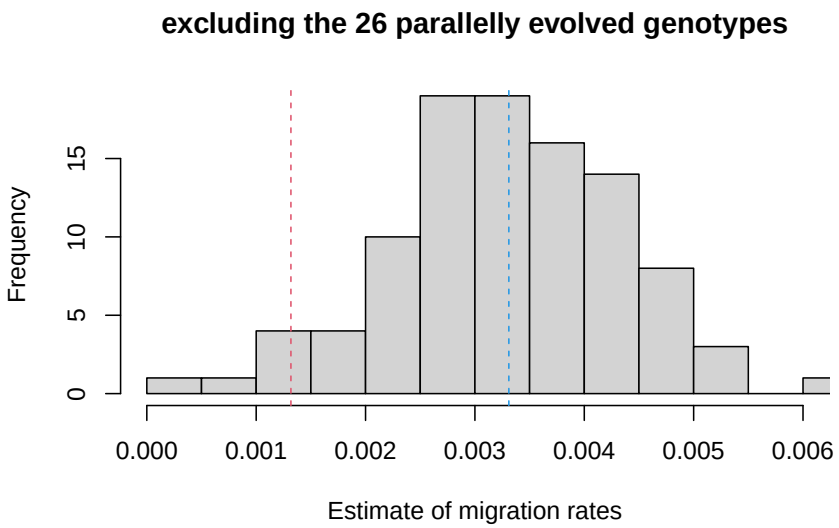



**Figure S2:** Distribution of migration rate estimates from the treetime analysis. The red dashed line shows the lower bound of the 95% confidence interval of the estimates, and the blue dashed line their mean. **(A)** The estimates were generated by subsampling 900 sequences from the total of 80'886 whole-genome sequences we obtained in the experiment, reconstructing the phylogeny, and performing the mugration analysis implemented in treetime. Migration rates between the ancestor population and the evolutionary lines have been subtracted from the overall migration rate estimate to obtain an estimate of the overall migration between the evolution lines only (see Materials and Methods). **(B)** Same analysis as in (A) except that we excluded from the data set the 26 genotypes that evolved in parallel in multiple line. In both cases, (A) and (B), the overall migration rate between the evolution lines is wrongly estimated to be significantly larger than zero.

# Acknowledgements

The authors thank James J. Bull for a decade of guidance and support of this study, Sylvain Gandon for his help designing the liquid titration of the bacteriophage in liquid, Marie-Agnes Petit, Claudia Igler, Silvia Kobbel, Léa Frachon for very fruitful discussions for the design of the DNA extraction protocols, Louis du Plessis for guidance on estimating migration rates with phylodynamic methods, and Claudia Igler and Carolin Wendling for comments on the manuscript. RRR acknowledges the financial support of the Swiss National Science Foundation (grant number 31003A_179170).

# References

Acevedo A, Brodsky L & Andino R (2014). **Mutational and fitness landscapes of an RNA virus revealed through population sequencing**. *Nature* **505**(7485): 686–690.

Baker CW, Miller CR, Thaweethai T, Yuan J, Baker MH, Joyce P & Weinreich DM (2016). **Genetically Determined Variation in Lysis Time Variance in the Bacteriophage $\phi X174$.** *G3 (Bethesda)* **6**(4):939–55.

Bertels F, Leemann C, Metzner KJ & Regoes R (2019). **Parallel evolution of HIV-1 in a long-term experiment.** *Mol. Biol. Evol.* **36**(11):2400–2414.

Bons E, Leemann C, Metzner KJ & Regoes RR (2020). **Long-term experimental evolution of HIV-1 reveals effects of environment and mutational history**. *PLoS Biol* **18**(12):e3001010.

Bons E & Regoes RR (2018). **Virus dynamics and phyloanatomy: Merging population dynamic and phylogenetic approaches.** *Immunol. Rev.* **285**(1):134–146.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A & Drummond AJ (2014). **BEAST 2: a software platform for Bayesian evolutionary analysis.** *PLoS Comput. Biol.* **10**(4):e1003537.

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N & others (2019). **Beast 2.5: An advanced software platform for bayesian evolutionary analysis**. *PLoS computational biology* **15**(4):e1006650.

Bull JJ, Badgett MR, Wichman HA, Huelsenbeck JP, Hillis DM, Gulati A, Ho C & Molineux IJ (1997). **Exceptional convergent evolution in a virus.** *Genetics* **147**(4):1497–507.

Cuevas JM, Duffy S & Sanjuán R (2009). **Point mutation rate of bacteriophage PhiX174.** *Genetics* **183**(2):747–9.

Denhardt DT & Silver RB (1966). **An analysis of the clone size distribution of $\varphi$x174 mutants and recombinants**. *Virology* **30**(1):10–19.

Desai MM & Fisher DS (2007). **Beneficial mutation–selection balance and the effect of linkage on positive selection**. *Genetics* **176**(3):1759–1798.

Domingo-Calap P, Cuevas JM & Sanjuán R (2009). **The fitness effects of random mutations in single-stranded DNA and RNA bacteriophages.** *PLoS Genet.* **5**(11):e1000742.

Drummond AJ & Rambaut A (2007). **BEAST: Bayesian evolutionary analysis by sampling trees.** *BMC Evol. Biol.* **7**:214.

Düx A, Lequime S, Patrono LV, Vrancken B, Boral S, Gogarten JF, Hilbig A, Horst D, Merkel K, Prepoint B & others (2020). **Measles virus and rinderpest virus divergence dated to the sixth century bce**. *Science* **368** (6497):1367–1370.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B & others (2009). **Real-time dna sequencing from single polymerase molecules**. *Science* **323**(5910):133–138.

Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, Tatem AJ, Sousa JaD, Arinaminpathy N, Pépin J, Posada D, Peeters M, Pybus OG & Lemey P (2014). **HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations.** *Science* **346**(6205):56–61.

Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, Griffin J, Baggaley RF, Jenkins HE, Lyons EJ, Jombart T, Hinsley WR, Grassly NC, Balloux F, Ghani AC, Ferguson NM, Rambaut A, Pybus OG, Lopez-Gatell H, Alpuche-Aranda CM, Chapela IB, Zavala EP, Guevara DME, Checchi F, Garcia E, Hugonnet S & Roth C (2009). **Pandemic potential of a strain of influenza A (H1N1): early findings.** *Science* **324**(5934):1557–61.

Gerrish PJ & Lenski RE (1998). **The fate of competing beneficial mutations in an asexual population**. *Genetica* **102**:127–144.

Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, Wohl S, Moses LM, Yozwiak NL, Winnicki S, Matranga CB, Malboeuf CM, Qu J, Gladden AD, Schaffner SF, Yang X, Jiang PP, Nekoui M, Colubri A, Coomber MR, Fonnie M, Moigboi A, Gbakie M, Kamara FK, Tucker V, Konuwa E, Saffa S, Sellu J, Jalloh AA, Kovoma A, Koninga J, Mustapha I, Kargbo K, Foday M, Yillah M, Kanneh F, Robert W, Massally JLB, Chapman SB, Bochicchio J, Murphy C, Nusbaum C, Young S, Birren BW, Grant DS, Scheiffelin JS, Lander ES, Happi C, Gevao SM, Gnirke A, Rambaut A, Garry RF,

Khan SH & Sabeti PC (2014). **Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak.** *Science* **345**(6202):1369–72.

Hartl D & Clark A. **Principles of Population Genetics**. Sinauer, 2007. ISBN 9780878933082.

Hill W & Robertson A (1968). **Linkage disequilibrium in finite populations**. *Theoretical and applied genetics* **38**:226–231.

Jennings HS (1917). **The numerical results of diverse systems of breeding, with respect to two pairs of characters, linked or independent, with special relation to the effects of linkage**. *Genetics* **2**(2):97.

Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, Sun C, Grayson T, Wang S, Li H, Wei X, Jiang C, Kirchherr JL, Gao F, Anderson JA, Ping LH, Swanstrom R, Tomaras GD, Blattner WA, Goepfert PA, Kilby JM, Saag MS, Delwart EL, Busch MP, Cohen MS, Montefiori DC, Haynes BF, Gaschen B, Athreya GS, Lee HY, Wood N, Seoighe C, Perelson AS, Bhattacharya T, Korber BT, Hahn BH & Shaw GM (2008). **Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection.** *Proc Natl Acad Sci U S A* **105**(21):7552–7.

Li H, Bar KJ, Wang S, Decker JM, Chen Y, Sun C, Salazar-Gonzalez JF, Salazar MG, Learn GH, Morgan CJ, Schumacher JE, Hraber P, Giorgi EE, Bhattacharya T, Korber BT, Perelson AS, Eron JJ, Cohen MS, Hicks CB, Haynes BF, Markowitz M, Keele BF, Hahn BH & Shaw GM (2010). **High Multiplicity Infection by HIV-1 in Men Who Have Sex with Men.** *PLoS Pathog.* **6**(5):e1000890.

Miller CR, Nagel AC, Scott L, Settles M, Joyce P & Wichman HA (2016). **Love the one you're with: replicate viral adaptations converge on the same phenotypic change.** *PeerJ* **4**:e2227.

Nadeau SA, Vaughan TG, Scire J, Huisman JS & Stadler T (2021). **The origin and early spread of SARS-CoV-2 in europe**. *Proc. Natl. Acad. Sci. U. S. A.* **118**(9):e2012008118.

Prosperi MCF & Salemi M (2012). **QuRe: software for viral quasispecies reconstruction from next-generation sequencing data.** *Bioinformatics* **28**(1): 132–3.

Sagulenko P, Puller V & Neher RA (2018). **Treetime: Maximum-likelihood phylodynamic analysis**. *Virus evolution* **4**(1):vex042.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM & Smith M (1977). **Nucleotide sequence of bacteriophage phi X174 DNA.** *Nature* **265**(5596):687–95.

Sinsheimer RL, Starman B, Nagler C & Guthrie S (1962). **The process of infection with bacteriophage $\varphi$x174: I. evidence for a "replicative form"**. *Journal of molecular biology* **4**(3):142–160.

Slatkin M (2008). **Linkage disequilibrium—understanding the evolutionary past and mapping the medical future**. *Nature Reviews Genetics* **9**(6):477–485.

Vaughan TG, Kühnert D, Popinga A, Welch D & Drummond AJ (2014). **Efficient Bayesian inference under the structured coalescent.** *Bioinformatics* **30**(16):2272–9.

Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND & others (2019). **Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome**. *Nature Biotechnology* **37**(10):1155–1162.

Wichman HA, Badgett MR, Scott LA, Boulianne CM & Bull JJ (1999). **Different trajectories of parallel evolution during viral adaptation.** *Science* **285**(5426):422–4.

Wichman HA, Millstein J & Bull JJ (2005). **Adaptive molecular evolution for 13,000 phage generations: a possible arms race.** *Genetics* **170**(1):19–31.

Zagordi O, Bhattacharya A, Eriksson N & Beerenwinkel N (2011). **ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data.** *BMC Bioinformatics* **12**:119.