Summarizing Populations: Characterizing the Effects of Sampling in Computational Evolutionary Replay Experiments

Nikolai Escondo[1] and Austin J. Ferguson[1]

[1]Grand Valley State University, United States
ferguaus@gvsu.edu

Abstract

When we sample an evolving population, how well do we capture its long-term evolutionary potential? This question underlies the validity of analytical replay experiments, which restart evolution from multiple points in a population's history to measure how long-term potential changed over time. Analytical replay experiments are becoming increasingly popular in both wet-lab and computational evolution studies. However, the population sampling method is still often picked via necessity, and no direct comparisons between methods exist. Here, we use computational evolution on NK landscapes to test the effects of population sampling techniques on genetic potentiation. We analyze four techniques: full population snapshots, random sampling, cloning the most abundant genotype, and tracing the dominant lineage. We find that cloning the most abundant genotype consistently results in lower potentiation than full snapshots, while tracing the dominant lineage consistently results in higher potentiation. Random sampling falls somewhere in the middle, where the sampling rate controls the variation across replays. We end by analyzing how an increase in mutation rate can counterintuitively stabilize the impacts of sampling on potentiation. This work leverages the speed of computational evolution to both provide insights into the results of previous replay experiments and to illuminate design implications for future experiments. While many methodological questions about replay experiments remain, this work demonstrates that computational studies can provide insights via experiments that are intractable in wet-lab settings.

Submission type: **Full Paper**; Data/Code available at: `https://github.com/FergusonAJ/replay_methods`

Introduction

Experimental evolution research has made astounding progress over the last several decades, with empirical studies employing new techniques and technologies to refine our understanding of evolution in the natural world (Ascensao and Desai, 2026). However, experimental evolution has its limitations (Kawecki et al., 2012), and every experimental design decision can impact the observed evolutionary dynamics. Researchers could hypothetically probe the effects of each decision, but this quickly becomes intractable when experiments can take months (Blount et al., 2008) or years (Lenski et al., 1991; Graves et al., 2017). Digital evolution and artificial life studies can fill this niche. Computational evolution allows for rapid generations across numerous replicates, while sacrificing population sizes and organismal complexity. Even with these limitations, computational studies have a storied history of improving our understanding of evolution (Lehman et al., 2020), and they can provide otherwise unobtainable insights into the design of evolution experiments.

In this work we leverage a computational evolution system to experimentally probe how population sampling techniques alter the outcomes of analytical replay experiments. We find that sampling methods dramatically and consistently alter a population's measured potentiation. Using full population snapshots as our baseline for comparison, we find that clonal populations founded with the dominant (most abundant) genotype from a given generation consistently underestimate the population's potential, while clonal population's founded from the final dominant lineage consistently overestimate potential. Additionally, we observe that randomly sampling a population slightly underestimates potential on average, though variation in potential increases as we collect fewer samples from the population. Overall, we provide additional context for interpreting previous replay experiments and empirically-backed considerations for designing new replay experiments.

Background

Evolutionary replay experiments

Observations of evolved populations in nature reveal life *as it is*, but evolutionary biologists have long been interested in life *as it could have been*. While we see one realization of evolutionary potential, what were the other possibilities? How likely was our observed outcome? Stephen Jay Gould once articulated this conundrum: we could learn so much about life, evolution, and their historical contingency if only we could "replay the tape of life", restarting evolution from a particular point to sample the other possibilities that could have been (Gould, 1990). While it is impossi-

ble to conduct Gould's experiment on the entire biosphere, experimental evolution researchers have developed methods for empirically investigating the role that contingency plays in the evolution of individual populations.

Using the terminology of Blount et al. (2018), for decades researchers have been conducting wet-lab *parallel replay experiments*, where multiple replicate populations evolve from a common ancestral clone or population (Lenski et al., 1991; Graves et al., 2017). Additionally, *historical difference experiments* evolve replicate populations under the same conditions after they have experienced varied evolutionary histories, whether artificially induced or naturally occurring (Travisano et al., 1995; Donofrio et al., 2026). Here, we focus on *analytic replay experiments*, where replay populations are founded from multiple time points along a population's history. By replaying evolution from different points in time, we can identify *when* the population experienced a change that altered its long-term evolutionary trajectory.

Analytical replay experiments were first employed to identify mutations in *Escherichia coli* that potentiated the evolution of citrate metabolism (Blount et al., 2008). Since then, most analytical replay experiments have been conducted on bacteria, where replays were used to investigate evolvability in *E. coli* (Woods et al., 2011), coevolution of Phage λ and *E. coli*, (Meyer et al., 2012; Gupta et al., 2022), *E. coli* clade extinction (Turner et al., 2015), antimicrobial resistance in *Pseudomonas aeruginosa* (Jochumsen et al., 2016), and diversity in *P. putida* (Al-Tameemi and Rodríguez-Verdugo, 2024). Recently, replay experiments have elucidated gene interactions in yeast (Vignogna et al., 2021) and mitochondrial evolution in the nematode *C. elegans* (Dubie et al., 2024). This near-exhaustive list demonstrates that analytical replay experiments are becoming more widespread, expanding to both new species and new evolutionary questions.

While the number of required evolutionary replicates make wet-lab replay experiments time- and labor-intensive, these methods are well suited for computational experiments. Indeed, replay experiments have been used in the Avida digital evolution platform to study re-evolution after an extinction event (Yedid et al., 2008), the evolutionary necessity of deleterious mutations (Covert III et al., 2013), and the evolution of associative learning (Ferguson and Ofria, 2023). Recently, replay experiments have been conducted in simple computational systems to study adaptive momentum (Ferguson et al., 2024) and developmental exaptations (Renner et al., 2024).

Population sampling methods in replay experiments

Whether conducted on natural or digital organisms, researchers employing analytical replay experiments must choose how to sample the original evolutionary record to found their replay populations. The nature of the experiment and involved organisms often dictates the possibilities for these founding events. Here, we describe four sampling strategies that attempt to summarize a population:

Full population snapshots. When sampling the historical record to found replay populations, the most accurate method is to not sample but instead *copy* the exact record. Recent computational studies were able to leverage this technique, replaying a perfect copy of the original population at every generation (Renner et al., 2024; Ferguson et al., 2024). In this work we use population snapshots as our baseline, as they perfectly capture the nuances of the population. However, full population snapshots are impossible to conduct in natural organisms, as they require perfect cloning of every organism. Further, snapshots work best in systems with discrete, synchronous generations where generations are clearly demarcated and organisms lack complex life histories that need captured.

Random sampling. In situations where population snapshots are infeasible, a simple solution is to randomly sample the reference population. By necessity, all replay experiments using natural organisms have employed a form of random sampling, including work in bacteria (Blount et al., 2008), yeast (Vignogna et al., 2021), and nematodes (Dubie et al., 2024). The details of random sampling vary between studies, as they are dependent on population size and the methods required to freeze, revive, and culture the organisms.

Extracting the dominant genotype each generation. We posit that random sampling of natural organism populations will often sample multiple genetically identical organisms. Indeed, some previous studies intentionally sample to the level of a single bacterial clone (Blount et al., 2008) or single hermaphroditic animal (Dubie et al., 2024). However, our computational evolution system experiences a much higher mutation rate than many natural organism populations, decreasing the probability that we sample only genetically identical organisms. To model this genetically identical sampling, we test "dominant genotype" sampling, where clonal replay populations are founded with the dominant (most abundant) genotype from that generation.

Lineage tracing. While computational systems allow perfect data tracking, experiments with asynchronous, overlapping generations raise the question of *when* to sample for replay populations when generational boundaries are unclear. Previous studies have instead relied on lineage tracing: researchers identify the dominant (most abundant) genotype at the end of the original evolutionary replicate, and then found clonal replay population with each ancestor along its lineage (Yedid et al., 2008; Ferguson and Ofria, 2023). This method sidesteps issues of generational overlap, but also relies heavily on survivorship bias. While some genotypes along the lineage may not have been populous in the original evolutionary record, this method founds replay experiments with clonal populations of these genotypes, completely ig-

noring population effects.

Methods

In this work, we characterize the effects of four population sampling techniques on the outcomes of evolutionary replay experiments. Wet-lab replay experiments are time- and labor-intensive, making it infeasible to empirically analyze these experimental design decisions. Here we employ a computational evolution system to evolve millions of replay replicates, allowing us to rigorously compare the effects of these sampling techniques. We use NK landscapes — simple fitness landscapes with tunable ruggedness — to balance speed and complexity. We analyze four population sampling techniques for replay experiments: 1) snapshotting the full population; 2) extracting the dominant (i.e., most abundant) genotype at each generation; 3) tracing the lineage of the final dominant genotype; and 4) randomly sampling the population at rates of 10%, 5% and 1%. Using the true historical record (full population snapshots) as a baseline, we then identify differences in evolutionary outcomes caused by the other sampling methods.

Evolution system

While computational evolution allows for a faster turnaround time compared to its web-lab counterparts, issues of time and complexity of a system still persist. To address this, we use NK landscapes — fast yet tunably rugged genotype-to-fitness mappings for bitstring organisms (Kauffman and Weinberger, 1989). N denotes the length of the bitstring and K is a proxy for the level of ruggedness in the landscape. Here, we needed landscapes that were difficult but not impossible to traverse to maximum fitness. We ran exploratory replicates to find such parameters, and all experiments in this work use $N = 20$ bits and $K = 9$.

All populations in this work are well-mixed and consist of 1,000 organisms. Each organism is a bitstring of 20 bits, and each reproduction event carries a mutation rate of 1% per bit. Initial (non-replay) replicates are seeded with a clonal population of the minimum-fitness genotype in the current NK landscape. Initial replicates experience 100 generations of evolution, with each generation being selected via tournament selection with a tournament size of seven (Goldberg and Deb, 1991). We chose this parameterization of tournament selection to generally favor the highest-fitness genotype in the population while still allowing lower-fitness genotypes some opportunity at reproduction.

Initial evolutionary replicates

Before running replay experiments, we first needed to identify viable fitness landscapes and our initial evolutionary replicates to replay. Because NK landscapes are randomly generated, we wanted to sample multiple landscapes to ensure our results were not dependent on a single landscape's idiosyncrasies. Additionally, replay experiments are traditionally used to measure a population's genetic "potentiation", its probability of evolving a particular genotype, trait, or behavior. Here we focus on a population's evolutionary potential to reach the genotype with maximum fitness in that NK landscape.

To identify viable landscapes, we first generated 100 random NK landscapes and evolved 100 independent replicate populations on each landscape. We selected landscapes where at least one replicate evolved maximum fitness, but we discarded landscapes where 90 or more replicates evolved maximum fitness. As described below, we then replayed *all* replicates that evolved maximum fitness. The 90 replicate threshold ensured we did not waste computational effort on landscapes where potential could only slightly increase.

Replay experiments

Once suitable evolutionary replicates were found, we conducted replay experiments using the four sampling techniques: population snapshots, random sampling, dominant genotypes, and lineage tracing. During our initial evolutionary replicates, we saved a full population snapshot every generation and traced parent-offspring relationships, giving us perfect information about the population's historical record and underlying phylogeny. Due to computational limitations, we replayed only the first 25 generations of each replicate, which exploratory work showed to be sufficient for the vast majority of populations to reach their maximum-fitness genotype. All replay replicates evolved for the same number of *total* generations as the initial replicates (e.g., a replay started at generation 20 would experience 80 generations of evolution, for a total of 100 generations). Replay replicates were given new random number seeds to ensure they experienced unique mutations and selection events.

Three of our sampling methods are deterministic: taking full population snapshots, cloning the dominant genotype, and tracing the dominant lineage will always return the same set of organisms. For these methods, we founded 100 replay replicates using each method at each generation. While population snapshots already contain 1,000 organisms, dominant and lineage methods only give a single organism. We thus copied that organism to produce a full clonal population of 1,000 organisms, eliminating the possibility for adaptive momentum while the population filled (Bohm et al., 2024). To calculate the genetic potentiation of the population at a given generation, we calculated the percentage of replay replicates from that generation that evolved maximum fitness.

The fourth sampling method, random sampling, is stochastic. In this work, we sampled our populations of 1,000 organisms at three levels: 10% (100 organisms), 5% (50 organisms), and 1% (10 organisms). To create full populations, each sampled organism was copied multiple times

to ensure we had a total of 1,000 organisms (e.g., at 5% sampling, each of the 50 organisms were copied 20 times). When replaying the population at a particular generation using random sampling, we first sampled the population 50 times at each sampling rate. We then ran 100 replay replicates for each sample, allowing us to calculate the potentiation of each sample as the percentage of replay replicates that evolved maximum fitness. While this framework suffers from the exponential growth, it allows us to compare across samples to test our hypothesis that lowering the sampling rate will increase the inter-sample variance in potentiation.

Mutation rate experiment

We hypothesized that genetic diversity in the population plays an important role in replay sampling, particularly when randomly sampling. Thus we repeated our initial experiment with double the mutation rate (2% per bit). We again tested 100 replicates on each of 100 random NK landscapes, and replayed all replicates that reached maximum fitness with our four sampling techniques. We then compared results both within and across mutation rates to determine the role mutation rates play in potentiation dynamics.

Comparative analyses and statistics

Population snapshots capture all the information of our population at that point in time, so our 100 snapshot replay replicates give an accurate estimate of the population's true potentiation. We thus use our snapshot potentiation as a baseline, analyzing the difference in potentiation between other sampling methods and our snapshots. For dominant and lineage sampling, we first compare the treatment to the snapshot baseline and then compare the two treatments with each other, both comparison used a paired Wilcoxon signed rank test. For our random sampling treatments, we compare against snapshot replays and then we measure the standard deviation of sample potential for each replayed time point to quantify across-sample variance. We then perform a Kruskal-Wallis test to look for differences among standard deviation for 10%, 5%, and 1% random sampling. Given a significant difference among the three sampling rates, we then perform a pairwise Wilcoxon test to determine which sampling rates are significantly different. Finally, when comparing across mutation rates we switch from paired Wilcoxon tests to unpaired Mann-Whitney tests. We employ a Holm-Bonferonni correction for multiple comparisons (Holm, 1979). For all analyses, we focus on the dynamics before the population is fully potentiated. Thus, we halt our analysis of a replicate at the generation when its dominant and lineage replays both reach 100% potentiation.

All statistics and data visualization was conducted in the R version 4.1.2 (R Core Team, 2021) using the ggplot2 and dplyr packages (Wickham et al., 2020, 2022).

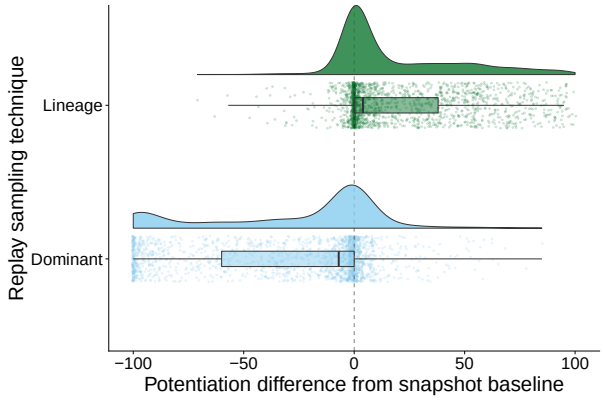

Figure 1: Raincloud plot showing the difference in potentiation between two sampling methods, dominant genotype and lineage tracing, and population snapshot replays. Each point shows the potentiation of a population at one point in time, calculated as the number of replicates (out of 100) that reached maximum fitness. The dashed line shows zero difference from snapshot replays. Each distribution shows N = 1,946 replayed populations.

Data and software availability

Evolution experiments were conducted using the Modular Agent-Based Evolver 2 (MABE2) (https://github.com/mercere99/MABE2/). Simulation and experiment code, cleaned data, and analyses can be found in our supplemental material (https://github.com/FergusonAJ/replay_methods).

Computational cost

To demonstrate the need for computational studies of replay methods, we want to highlight the number of replay replicates required for this work. Snapshots, dominant genotypes, and traced lineages were replayed 100 times for each of 25 generations, requiring 2,500 replay replicates per sampling method per landscape. We replayed a total of 461 replicates across both mutation rate experiments for a total of 3,457,500 replay replicates. For random sampling, we additionally sampled each population 50 times for each of our three sampling rates. We thus conducted 172,875,000 replay replicates for our random sampling treatments, and 176,332,500 total replay replicates. In our system, each replay replicate takes only a fraction of a second, but these numbers demonstrate that computational experiments are currently the only way to test replay methodologies at scale.

Results and discussion

Here we describe the results of our replay experiments and situate them within the existing evolutionary replay literature.

Evolution of initial replicates

To find replicates to replay, we first ran 100 evolutionary replicates on each of 100 randomly-generated NK landscapes. With our base 1% per-bit mutation rate, only 20 landscapes had at least one replicate evolve maximum fitness. One of the 20 landscapes had 93 replicates evolve maximum fitness, so it was excluded due to computational costs and its narrow window for potentiation growth. Of the remaining 19 landscapes, the number of successful replicates ranged from 1 to 41, with a mean of 12.2 and a median of 4.5 replicates.

Dominant replays underestimate potentiation while lineage replays overestimate potentiation

We next ran replay experiments using our three deterministic sampling methods: full population snapshots, extracting the dominant (most abundant) genotype, and lineage tracing. Since population snapshots are our true baseline, we show the distribution of differences from that snapshot baseline (Figure 1).

Both dominant and lineage replays produce potentiation values significantly different than full population snapshots. Dominant genotype replays consistently result in lower potentiation values than population snapshots (Wilcoxon, adjusted $p < 1e\text{-}145$). Conversely, lineage tracing replays consistently result in higher potentiation values than snapshot replays (Wilcoxon, adjusted $p < 1e\text{-}166$). Transitively, lineage replays also show higher potentiation than dominant replays (Wilcoxon, adjusted $p < 1e\text{-}241$).

These results demonstrate that the sampling method we employ can dramatically alter our resultant potentiation values. If a mutation in the dominant lineage creates a genotype that has low fitness relative to the current population, but that genotype potentiates a path to maximum fitness in the long term, lineage tracing will immediately result in clonal populations full of the potentiated genotype. Dominant replays will only contain that genotype (or its descendants) when it reaches plurality. Both sampling methods ignore the population dynamics at play; a new genotype has to spread and reproduce to alter the long-term evolutionary trajectory. Population snapshots capture this spread, with each reproduction event influencing the population's potential, while dominant and lineage replays only capture large stepwise changes in potential.

These findings help us interpret the results of previous replay studies. Computational work that uses lineage tracing is likely overestimating potentiation on average (Yedid et al., 2008; Ferguson and Ofria, 2023). While an increase in lineage-based potentiation should result in an increase in snapshot-based potentiation, the magnitude of these changes can be drastically different. Similarly, previous work that samples a single organism or clone (most likely the dominant genotype) is likely underestimating the changes in potentiation (Blount et al., 2008; Dubie et al., 2024). In both cases, our evidence suggests that, while potentiation estimates are not far off on average, lineage replays may be reporting increases in potentiation early, while dominant replays will often lag major increases in population potential.

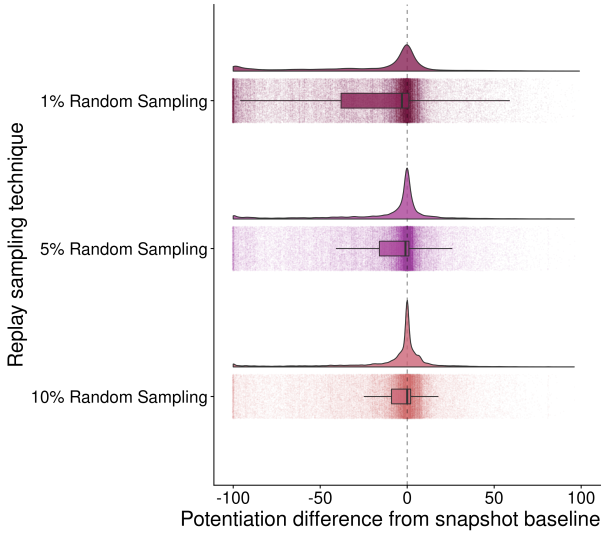

Figure 2: Raincloud plot showing the difference between random sampling replays at various sampling rates and population snapshot replays. Points are smaller than Figure 1 to account for the increased number of replicates due to sampling. Each point shows the potentiation of one sample, calculated as the number of replicates (out of 100) that reach maximum fitness. N = 97,300 samples for each treatment.

Random sampling increases variance in potentiation

Next we replayed the same populations using random sampling to found our replay replicates. We ran three treatments, randomly sampling 10%, 5%, and 1% of the original population. All three sampling rates resulted in a significant decrease in potentiation compared to snapshot replays (Wilcoxon, adjusted $p < 1e\text{-}305$ for each). Lowering the sampling rate from 10% of the population to 1% saw a larger spread of potentiation values, with 1% sampling resulting in some samples with substantially higher potentiation and some samples with much lower potentiation than the snapshot baseline (Figure 2).

These findings match our expectations. Randomly sampling 10% of the population generally provides a fairly accurate representation of the population, though some samples still see extremely low or, very occasionally, extremely high potentiation. These extremes are exacerbated as we decrease our sampling rate. At 5% and especially 1% sampling, samples will most often consist of many copies of the dominant genotype and potentiation will be underestimated as in dominant genotype-based replays. However, when a random sample includes genotypes with high potential, such

as those along the dominant lineage, these lower sampling rate experiments can overestimate the population's potentiation. Indeed, Figure 5 shows that the standard deviation of potentiation for a replayed population increases significantly as we decrease the sampling rate (Kruskal-Wallis, adjusted $p < 1e\text{-}20$; all pairwise Wilcoxon adjusted $p < 1e\text{-}20$).

All replay experiments conducted on natural organisms must use some form of random sampling. As such, our results provide evidence that even when our sampling rate is as small as 1%, the replay experiments produce potentiation results fairly consistent with snapshot replays. However, we have also shown that the exact sample can dramatically affect the observed potentiation. Even at 10% sampling, a sample can dramatically over- or underestimate potentiation. Replication is thus vital for experiments using random sampling; even though replay experiments are time- and labor- intensive, multiple samples are needed to lower the chance of all replays being outliers. Even early replay work leveraged six replay replicates per population (Blount et al., 2008), which substantially decreases the probability the results are based on outlying samples.

A case study on the interconnected effects of mutation rate

We hypothesized that non-snapshot replay results were highly contingent on genetic diversity in the population, and that increased genetic diversity would increase the differences between our sampling methods and snapshot replays. To test this hypothesis, we repeated our experiment with 100 new NK landscapes and double the mutation rate (from 1% to 2% per bit).

Of the 100 new landscapes tested with the higher mutation rate, 41 one of them had at least one of 100 initial replicates evolve to reach maximum fitness. Across the 41 landscapes, the number of successful replicates ranged from 1 to 50 with a mean of 7.7 and a median of 2 replicates. No landscapes were over our 90-replicate threshold, so all 41 were included in our replay experiments. This increase in the number of replicates that evolved maximum fitness is likely due to increased evolutionary exploration. To reach the maximum fitness genotype in an NK landscape, an evolving population must explore the fitness landscape without converging on a local optimum. Increasing the mutation rate would increase the exploratory power of the evolving populations, leading to more populations reaching maximum fitness.

To verify that doubling the mutation rate increased the genetic diversity in the population, we calculated the genotypic diversity of every population that was sampled to be replayed. Genotypic diversity was calculated as the Shannon entropy of all 1,000 genotypes in the population (Figure 3). We find that doubling the mutation rate significantly increased the genotypic diversity in the population (Mann-Whitney, adjusted $p < 1e\text{-}91$).

Like the original experiment, we begin by examining

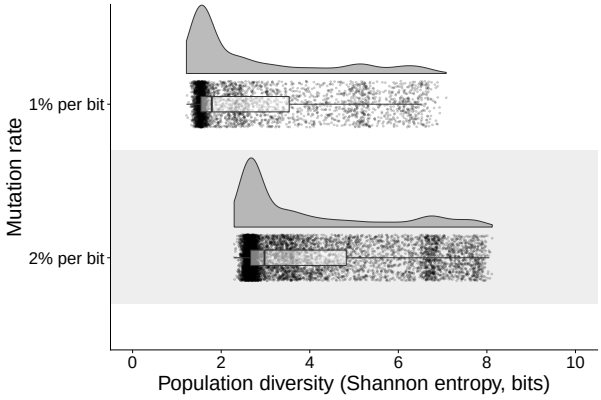

Figure 3: Raincloud plot showing genotypic diversity (calculated as Shannon entropy, in bits) for each replayed population. The shaded background shows results from the increased mutation rate experiment.

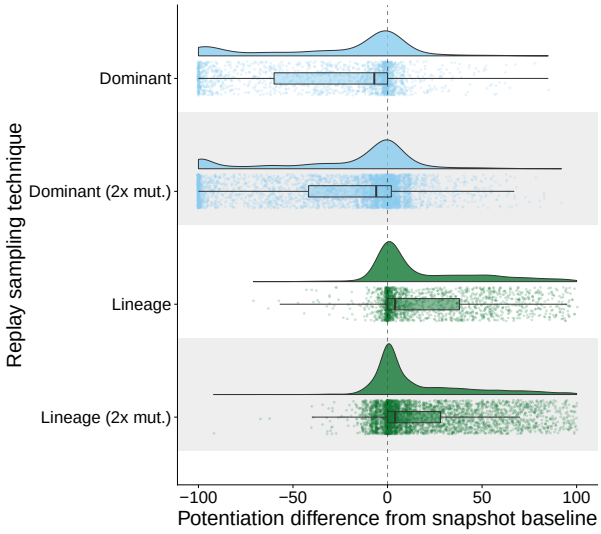

Figure 4: Raincloud plot showing potentiation results for dominant and lineage sampling for both mutation rates. The shaded background indicates results from the increased mutation rate experiment. Results for the base mutation rate are the same as Figure 1, shown again for easy comparison (N = 1,946 replicates each). The doubled mutation rate plots show potentiation from N = 4,418 replay replicates each.

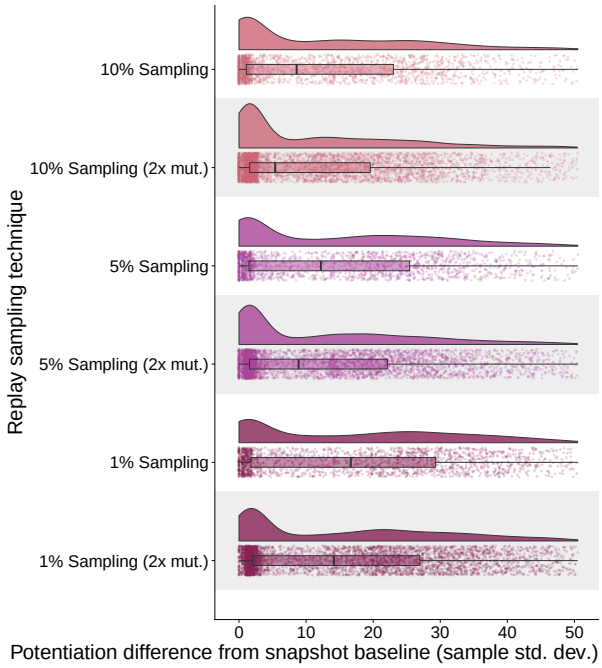

Figure 5: Raincloud plot showing the standard deviation of potentiation for all random sampling replays. Each point represents the standard deviation of potentiation across 50 samples. The shaded background indicates results from the increased mutation rate experiment. Each doubled mutation rate plot shows potentiation from N = 4,418 replay replicates, while the base mutation rate plots show N = 1,946 replicates each.

how dominant genotype and lineage tracing replays differed from the snapshot baseline (Figure 4). Again, we find that dominant replays significantly underestimate potentiation (Wilcoxon, adjusted p<1e-223) and lineage replays significantly overestimate potentiation (Wilcoxon, adjusted p<1e-305). We found significant, but not substantial, differences in potentiation across mutation rates for both dominant (Mann-Whitney, adjusted p $<$ 1e-5) and lineage sampling (Mann-Whitney, adjusted p $\approx$ 0.015). Interestingly, doubling the mutation rate resulted in potentiation values that were closer to the snapshot baseline for both dominant and lineage methods.

Our original hypothesis was that an increase in genetic diversity would cause more variation in the potentiation of randomly-sampled replay populations. However, at all sampling levels we failed to detect a significant difference in standard deviation of potentiation between the two mutation rates (Mann-Whitney, adjusted p>0.09 for all). Indeed, while the differences are not statistically significant, doubling the mutation rate *lowered* the median and mean variance in across-sample potentiation for all sampling rates.

While initially counterintuitive, our new hypothesis is that increasing the mutation rate increased evolutionary exploration, increasing the probability for any population to evolve to maximum fitness. Increased genotypic diversity may add more genetic variation to the random samples, but those samples were generally more likely to reach maximum fitness, decreasing the variation in potentiation across samples. We include this case study to demonstrate that even small changes to an experiment can alter replay experiment results in ways we might not expect. Since replay experiments measure the *potential* of a population to evolve a certain trait, we must consider all aspects of that population's evolution. More studies are needed to solidify our shared understanding of potentiation and replay experiments.

Conclusion

Population sampling techniques drastically alter replay outcomes

We have provided the first empirical analysis of how different sampling methods for seeding analytical replay experiments alter the potentiation measured for each population. By using full population snapshots as a baseline for comparison, we demonstrated that the survivorship bias implicit in lineage tracing leads to consistently higher potentiation values, while clonal populations founded with the dominant (most abundant) genotype result in consistently lower potentiation values. Further, we demonstrated that random sampling can be an effective method for founding replay replicates, though lower mutation rates are prone to higher variance. Finally, we presented a case study that showcases how a simple change to the system, such as increasing the mutation rate, can result in unexpected potentiation results.

Overall, this work serves a dual purpose. First, it grants us additional insight into the results of previous replay experiment studies. For example, we now have evidence that studies using lineage-based sampling methods are likely consistently overshooting the population's true potentiation. Second, this work provides an empirical baseline for those making methodological decisions about future replay experiments. While these replay sampling methods had been used before, this work is the first to empirically demonstrate their differences. Future replay experiments that can leverage population snapshots should do so. Replay experiments that cannot employ snapshots should consider random sampling if they have the resources to conduct multiple samples to avoid outliers.

Outlook and future work

We view this work as an initial step in using computational evolution to empirically investigate the methodological choices in replay experiments. Studies like this will inform future replay experiment in both natural and digital systems. However, many such methodological choices remain unexplored, and thus we describe multiple avenues for future work in this area.

Repetition We have only provided evidence of how sampling techniques affect replay experiments in one system. We expect general trends, such as higher potentiation observed in lineage-based replays, to hold across systems. However, the details are likely to change based on system parameters. For example, we have shown that mutation rate can substantially alter replay outcomes. Population size, spatial reproduction, and the selection method of the system are also likely to alter potentiation dynamics. Similarly, we measured potentiation as the population's probability of reaching the single maximum-fitness genotype. Even slight modifications, such as measuring potentiation as reaching a genotype with a fitness in the top 1%, may have substantial impacts on replay outcomes.

Further, we conducted this work using NK landscapes due to their speed and tunability. Future work should detect how potentiation dynamics change as N and K increase, creating landscapes that are more difficult for evolving populations to traverse. Beyond NK landscapes, future work should expand beyond simple fitness landscapes to include full digital evolution systems such as Avida or Aevol.

Empirical investigation of other replay experiment methods We have shown that the speed and observability of computational evolution allow us to empirically test replay experiment sampling techniques via millions of evolved populations. Computational evolution is currently the only testbed for such questions, as the scale required for such methodological studies is intractable in experiments with natural organisms. Future work should empirically measure the effects of other design decisions in replay experiments. For example, the original analytical replay experiments in (Blount et al., 2008) ran all replay replicates for the same number of generations, regardless of the time point from which the replay was founded. Computational replay experiments have traditionally altered the length of replays to ensure each replayed population experiences the same *total* number of generations. While this experiment decisions in wet-lab work are typically made out of necessity (e.g., we cannot replay 30 years of *E. coli* evolution for a replay experiment), computational experiments provide us the means to identify the implications of these decisions.

Advancing genetic bottleneck theory Genetic bottlenecks occur when a large portion of a population perishes and the extant individuals represent only a fraction of the population's previous genetic diversity. Bottlenecks are well studied, including many empirical studies, due to their dire implications in conservation biology (see Bouzat (2010) for review). Yedid et al. (2008) leveraged replay techniques to study re-evolution of a complex trait after its loss, but we are unaware of studies using replay experiment to test the the effects of genetic bottlenecks at multiple points in a population's history. In this work, randomly sampling the population is effectively reducing the genetic diversity of the population to that found in the X% of sampled organisms. By replaying a population from different generations, we are testing how a population's accumulated history buffers it from disastrous genetic bottlenecks. Simply by shifting the framing, future work should leverage similar replay experiments to empirically determine if populations experience periods of increased resilience from genetic bottlenecks.

Acknowledgments

We thank the other members of the Computational Evolution Laboratory at GVSU for thoughtful comments. Computational resources were provided by Grand Valley State University's Clipper system and Michigan State University's Institute for Cyber-Enabled Research.

References

Al-Tameemi, Z. and Rodríguez-Verdugo, A. (2024). Microbial diversification is maintained in an experimentally evolved synthetic community. *mSystems*, 9(11):e01053–24.

Ascensao, J. A. and Desai, M. M. (2026). Experimental evolution in an era of molecular manipulation. *Nature Reviews Genetics*, 27(1):81–95.

Blount, Z. D., Borland, C. Z., and Lenski, R. E. (2008). Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. *Proceedings of the National Academy of Sciences*, 105(23):7899–7906.

Blount, Z. D., Lenski, R. E., and Losos, J. B. (2018). Contingency and determinism in evolution: Replaying life's tape. *Science*, 362(6415):eaam5979.

Bohm, C., Ragusa, V. R., Ofria, C., Lenski, R. E., and Adami, C. (2024). Reduced selection during sweeps lead to adaptive momentum on rugged landscapes. *bioRxiv : the preprint server for biology*.

Bouzat, J. L. (2010). Conservation genetics of population bottlenecks: The role of chance, selection, and history. *Conservation Genetics*, 11(2):463–478.

Covert III, A. W., Lenski, R. E., Wilke, C. O., and Ofria, C. (2013). Experiments on the role of deleterious mutations as stepping stones in adaptive evolution. *Proceedings of the National Academy of Sciences*, 110(34):E3171–E3178.

Donofrio, M. A., Blasius, H. L., Nguyen, C. C., Schnell, A. L., and Turner, C. B. (2026). Antibiotic susceptibility of Escherichia coli is affected by evolutionary history but not by history of elemental limitation. *mSphere*, 0(0):e00538–25.

Dubie, J. J., Katju, V., and Bergthorsson, U. (2024). Dissecting the sequential evolution of a selfish mitochondrial genome in Caenorhabditis elegans. *Heredity*, 133(3):186–197.

Ferguson, A. J. and Ofria, C. (2023). Potentiating Mutations Facilitate the Evolution of Associative Learning in Digital Organisms. In *ALIFE 2023: Ghost in the Machine: Proceedings of the 2023 Artificial Life Conference*. MIT Press.

Ferguson, A. J., Ofria, C., and Bohm, C. (2024). Predicting the Unpredictable: Using replay experiments to disentangle how evolutionary outcomes are altered by adaptive momentum. In *ALIFE 2024: Proceedings of the 2024 Artificial Life Conference*. MIT Press.

Goldberg, D. E. and Deb, K. (1991). A Comparative Analysis of Selection Schemes Used in Genetic Algorithms. In Rawlins, G. J. E., editor, *Foundations of Genetic Algorithms*, volume 1, pages 69–93. Elsevier.

Gould, S. J. (1990). *Wonderful Life: The Burgess Shale and the Nature of History*. WW Norton & Company.

Graves, Jr, J., Hertweck, K., Phillips, M., Han, M., Cabral, L., Barter, T., Greer, L., Burke, M., Mueller, L., and Rose, M. (2017). Genomics of Parallel Experimental Evolution in Drosophila. *Molecular Biology and Evolution*, 34(4):831–842.

Gupta, A., Zaman, L., Strobel, H. M., Gallie, J., Burmeister, A. R., Kerr, B., Tamar, E. S., Kishony, R., and Meyer, J. R. (2022). Host-parasite coevolution promotes innovation through deformations in fitness landscapes. *eLife*, 11:e76162.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Jochumsen, N., Marvig, R. L., Damkiær, S., Jensen, R. L., Paulander, W., Molin, S., Jelsbak, L., and Folkesson, A. (2016). The evolution of antimicrobial peptide resistance in Pseudomonas aeruginosa is shaped by strong epistatic interactions. *Nature Communications*, 7(1):13002.

Kauffman, S. A. and Weinberger, E. D. (1989). The *NK* model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology*, 141(2):211–245.

Kawecki, T. J., Lenski, R. E., Ebert, D., Hollis, B., Olivieri, I., and Whitlock, M. C. (2012). Experimental evolution. *Trends in Ecology & Evolution*, 27(10):547–560.

Lehman, J., Clune, J., Misevic, D., Adami, C., Altenberg, L., Beaulieu, J., Bentley, P. J., Bernard, S., Beslon, G., Bryson, D. M., Cheney, N., Chrabaszcz, P., Cully, A., Doncieux, S., Dyer, F. C., Ellefsen, K. O., Feldt, R., Fischer, S., Forrest, S., Fŕenoy, A., Gagńe, C., Le Goff, L., Grabowski, L. M., Hodjat, B., Hutter, F., Keller, L., Knibbe, C., Krcah, P., Lenski, R. E., Lipson, H., MacCurdy, R., Maestre, C., Miikkulainen, R., Mitri, S., Moriarty, D. E., Mouret, J.-B., Nguyen, A., Ofria, C., Parizeau, M., Parsons, D., Pennock, R. T., Punch, W. F., Ray, T. S., Schoenauer, M., Schulte, E., Sims, K., Stanley, K. O., Taddei, F., Tarapore, D., Thibault, S., Watson, R., Weimer, W., and Yosinski, J. (2020). The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities. *Artificial Life*, 26(2):274–306.

Lenski, R. E., Rose, M. R., Simpson, S. C., and Tadler, S. C. (1991). Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2,000 generations. *The American Naturalist*, 138(6):1315–1341.

Meyer, J. R., Dobias, D. T., Weitz, J. S., Barrick, J. E., Quick, R. T., and Lenski, R. E. (2012). Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science*, 335(6067):428–432.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Renner, A., Marois, É., Fiorito, J., Ashworth, J., and Yoder, J. A. (2024). A Computational Model of Developmental Exaptations. In *ALIFE 2024: Proceedings of the 2024 Artificial Life Conference*. MIT Press.

Travisano, M., Mongold, J. A., Bennett, A. F., and Lenski, R. E. (1995). Experimental tests of the roles of adaptation, chance, and history in evolution. *Science*, 267(5194):87–90.

Turner, C. B., Blount, Z. D., and Lenski, R. E. (2015). Replaying evolution to test the cause of extinction of one ecotype in an experimentally evolved population. *PLoS One*, 10(11):e0142050.

Vignogna, R. C., Buskirk, S. W., and Lang, G. I. (2021). Exploring a local genetic interaction network using evolutionary replay experiments. *Molecular biology and evolution*, 38(8):3144–3152.

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., and Dunnington, D. (2020). *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*.

Wickham, H., François, R., Henry, L., and Müller, K. (2022). *Dplyr: A Grammar of Data Manipulation*.

Woods, R. J., Barrick, J. E., Cooper, T. F., Shrestha, U., Kauth, M. R., and Lenski, R. E. (2011). Second-order selection for evolvability in a large Escherichia coli population. *Science*, 331(6023):1433–1436.

Yedid, G., Ofria, C. A., and Lenski, R. E. (2008). Historical and contingent factors affect re-evolution of a complex feature lost during mass extinction in communities of digital organisms. *Journal of evolutionary biology*, 21(5):1335–1357.