Interpreting the Tape of Life: Ancestry-Based Analyses Provide Insights and Intuition about Evolutionary Dynamics

Emily Dolson*
Michigan State University
BEACON Center for the Study of Evolution in Action
Department of Computer Science and Engineering
Ecology, Evolutionary Biology, and Behavior Program
dolsonem@msu.edu

Alexander Lalejini
Michigan State University
BEACON Center for the Study of Evolution in Action
Department of Computer Science and Engineering
Ecology, Evolutionary Biology, and Behavior Program

Steven Jorgensen
Michigan State University
BEACON Center for the Study of Evolution in Action
Department of Computer Science and Engineering

Charles Ofria
Michigan State University
BEACON Center for the Study of Evolution in Action
Department of Computer Science and Engineering
Ecology, Evolutionary Biology, and Behavior Program

Abstract Fine-scale evolutionary dynamics can be challenging to tease out when focused on the broad brush strokes of whole populations over long time spans. We propose a suite of diagnostic analysis techniques that operate on lineages and phylogenies in digital evolution experiments, with the aim of improving our capacity to quantitatively explore the nuances of evolutionary histories in digital evolution experiments. We present three types of lineage measurements: lineage length, mutation accumulation, and phenotypic volatility. Additionally, we suggest the adoption of four phylogeny measurements from biology: phylogenetic richness, phylogenetic divergence, phylogenetic regularity, and depth of the most-recent common ancestor. In addition to quantitative metrics, we also discuss several existing data visualizations that are useful for understanding lineages and phylogenies: state sequence visualizations, fitness landscape overlays, phylogenetic trees, and Muller plots. We examine the behavior of these metrics (with the aid of data visualizations) in two well-studied computational contexts: (1) a set of two-dimensional, real-valued optimization problems under a range of mutation rates and selection strengths, and (2) a set of qualitatively different environments in the Avida digital evolution platform. These results confirm our intuition about how these metrics respond to various evolutionary conditions and indicate their broad value.

Keywords
Phylogeny, lineage, metrics, evolutionary dynamics, phylogenetic metrics, data visualization, digital evolution

1 Introduction

Evolution is a collective effect of many smaller events—such as replication, variation, and competition—that occur on a fine-grained temporal scale. While evolution's emergent nature can be fascinating, it also presents challenges to studying the short-term mechanisms that, in aggregate, govern long-term results.

* Corresponding author.

© 2020 Massachusetts Institute of Technology

In computational evolutionary systems, we can theoretically collect data to help untangle these mechanisms. In practice, however, the sheer number of constituent events produces an overwhelming quantity of data. In response, we have developed a standardized suite of diagnostic metrics to summarize short-term evolutionary dynamics within a population by measuring lineages and phylogenies. Here, we describe these metrics and provide experimental results to develop an intuition for what they can tell us about evolution.

A lineage describes a continuous line of descent, linking parents and offspring in an unbroken chain from an original ancestor. A complete lineage can provide a post hoc, step-by-step guide to the evolution of an extant organism where each step involves replication and inherited variation. Indeed, lineage analyses are a powerful tool for disentangling evolutionary dynamics in both natural and digital systems; digital systems, however, allow for perfect lineage tracking at a level of granularity that is impossible in modern wet lab experiments. These data allow us to replay the tape of life in precise detail and to tease apart the evolutionary recipe for any phenomenon we are interested in [29]. In one notable example, Lenski et al. used the lineage of an evolved digital organism in Avida to tease apart, step by step, how a complex feature (the capacity to perform the *equals* logical operation) emerged [25].

Yet, tracking the full details of a single lineage, not to say a population of lineages, can be computationally expensive and will inevitably generate an unwieldy amount of data that can be challenging to visualize or interpret [28]. Summary statistics can help alleviate these difficulties by enabling the user to focus on aggregate trends across a population rather than needing to examine each individual's lineage. The question is how to effectively summarize a path through fitness space. One useful abstraction is to treat the path as a sequence of states. Here, we primarily use phenotypes and genotypes as the states in the sequence, but we could just as easily use some other descriptor of the lineage at a given point in time. With this abstraction in hand, a few metrics are easily formalized: the number of unique states, the number of transitions between states, and the amount of time spent in each state. Additionally, we may care about how the transitions between states happened. What mutations led to them? Were those mutations beneficial, deleterious, or neutral at the time? These mutations are particularly notable because they did not simply appear briefly, but stood the test of time, leaving descendants in the final population. Here, we explore a subset of these metrics that we expect will be broadly useful.

Whereas a lineage recounts the evolutionary history of a single individual, a phylogeny details the evolutionary history of an entire population. Measurements that summarize phylogenies can provide useful insight into population-level evolutionary dynamics, such as diversification and coexistence among different clades. A variety of useful phylogeny measurements have already been developed by biologists [38]. These measurements tend to treat the phylogeny as a graph and make calculations about its topology. Tucker et al. group them into three broad categories: assessments of the quantity of evolutionary history represented by a population, assessments of the amount of divergence within that evolutionary history, and assessments of the topological regularity of the phylogenetic tree. Such measurements can help quantify the behavior of the population as a whole, providing insight into interactions between its members. Thus, they are useful indicators of the presence of various types of eco-evolutionary dynamics.

Here, we present three types of lineage measurements and suggest adopting four phylogeny measurements from biology; these are lineage length, mutation accumulation, phenotypic volatility, depth of the most-recent common ancestor, phylogenetic richness, phylogenetic divergence, and phylogenetic regularity. For each metric, we discuss its application and our expectation for what it can tell us about evolution. We evaluate our intuition in two computational contexts: first, on a set of two-dimensional, real-valued optimization problems under a range of mutation rates and selection strengths, and second, on four qualitatively different environments in the Avida digital evolution platform (a minimal control environment, an environment that rewards evolving solutions for nine Boolean logic functions, an environment with limited resources, and a simple changing environment). For simplicity, we restrict our attention to asexually reproducing populations; however, we suggest how these metrics can be extended to sexual populations.

There are many visual analysis tools that operate on lineages and phylogenies. These data visualizations can provide further insight into evolutionary dynamics within a run. In this article, we discuss some approaches to visualizing lineages and phylogenies; we use these tools to build intuition about the behavior of our metrics.

In addition to demonstrating a range of analysis tools that are useful to digital evolution research, we intend for this work to begin a conversation within the artificial life community about how we quantify, interpret, and compare observed evolutionary histories. There have been extensive efforts to improve our ability to represent and visualize both lineages and phylogenies [6, 24, 28, 29, 37], which are indispensable for building intuitions and qualitatively understanding the dynamics embedded in a population's evolutionary history. However, we are unaware of efforts to formalize a suite of quantitative lineage- and phylogeny-based metrics for computational evolution. Lineage- and phylogeny-based analyses in artificial life are often a component of a larger set of experiment-specific analyses or are limited to qualitative descriptions and/or visualizations. Well-defined metrics not only provide valuable tools for teasing apart evolutionary dynamics, they also allow us to move away from exclusively qualitative descriptions of results toward a deeper quantitative understanding. This understanding, in turn, facilitates rigorous statistical analyses and hypothesis testing as well as comparison of evolutionary dynamics across different digital evolution systems.

2 Metrics

Code for all of our metrics is open source and available in the Empirical library [33]. Empirical is a C++ library built to facilitate writing efficient and easily sharable scientific software. Empirical is a header-only library, so adding these metrics to an existing project has minimal overhead. Standalone versions of these metrics that can be applied to any previously generated data are being developed as part of the Artificial Life Data Standards initiative [23].

2.1 Lineage Metrics

Each of the three lineage metrics that we discuss—lineage length, mutation accumulation, and phenotypic volatility—reduces a lineage to a linear sequence of states where each state represents an individual or sequence of individuals that share a common genotypic or phenotypic characteristic of interest; Figure 1 is given as a toy example to help guide our discussion of these metrics. While we limit our focus to three lineage metrics, this abstraction places lineages in a form suitable for a wide range of measurements, including the direct application of many data mining techniques designed to operate over sequences, such as sequential pattern mining and trend analysis [19].

Only asexual lineages where genetic material is exclusively vertically transmitted can be directly abstracted as a *linear sequence* of states. Sexual reproduction (or any form of horizontal gene transfer) complicates matters significantly, as such lineages are more appropriately represented by trees rooted at the extant organism, branching for each contributor of genetic material. For the metrics we present here, we limit our discussion to asexual populations; however, we suggest three approaches for generalizing these lineage metrics to sexual populations:

1. *Build lineages based on sites in a genome.* For genetic representations composed of multiple constituent parts that are inherited atomically (such as, sites in a genome), the complication of sexual lineages can be avoided by tracking the lineages of these parts. Genomic sites will only ever have a single parent, so their lineages are effectively asexual. An organism could then be viewed as a collection of lineages rather than a single one.
2. *Apply a lossy compression to reduce sexual lineages into linear sequences.* By modeling sexual reproduction events as asexual reproduction events, it is possible to compress sexual lineages into linear sequences of states. One parent is designated to be a part of the lineage, and the genetic contributions of other parents are considered to be sources of genetic variation (mutations). The primary downside to this approach is its lossiness (i.e., the fact that it discards potentially important parentage information).

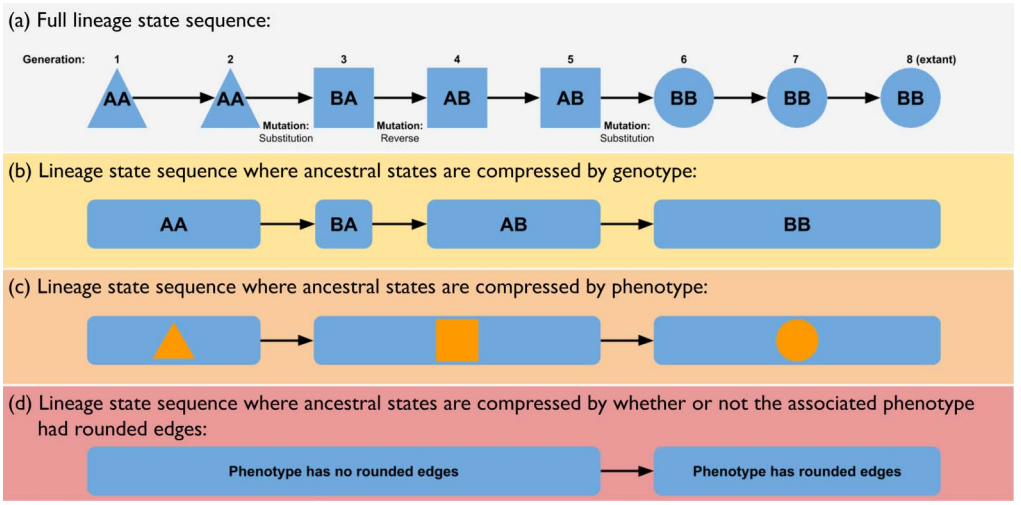

Figure 1. Four methods of representing a lineage. This example lineage has accumulated three mutations (one reverse mutation and two substitutions) and gone through three distinct phenotypes. In (a), each state along the lineage represents a single individual; the lineage length is the number of generations spanned by the lineage (eight). In (b), states represent the sequence of genotypes along the lineage, reducing the lineage length to four. In (c) states represent the sequence of phenotypes along the lineage; the lineage length is the number of times a different phenotype is expressed (three). In (d), states are a particular phenotypic characteristic; here, the lineage length is two.

3. *Extend our application of the metrics to operate over nonlinear (tree) structured lineages.* Alternatively, we can extend our metrics to operate over the more complex state sequences that constitute the lineages of sexually reproducing organisms. One such approach would be to consider all possible ancestor paths for an extant individual, calculate a given metric for each of them, and then average the resulting values together.

Assessing the efficacy of these and potentially other approaches would be a useful line of research to pursue in the future.

2.1.1 Lineage Length

Lineage length describes the number of states traversed by a lineage. If a state is defined as a single individual, lineage length is a count of the number of generations. The generation count is most useful in systems where generational turnover is not fixed, but instead determined by the life history strategies of organisms. For lineages that span equal lengths of time, more generations imply faster replication rates (e.g., *r*-selected lineage), while fewer generations imply slower replication rates (e.g., *K*-selected lineage).

Lineage length becomes a more flexible and informative metric if we consider more abstract definitions of states along a lineage. We might measure lineage length where a state represents a sequence of individuals that share a particular phenotypic or genotypic characteristic. In these cases, lineage length only increases when the characteristic of interest changes from parent to offspring. For example, in an environment where organisms must perform functions to be successful, we might define a state as the set of functions performed by an individual. In this scenario, lineage length would only increase when the set of functions performed by an ancestor changes; sequential ancestors that perform the identical sets of functions would be compressed into a single state in the sequence, even if other traits differ.

2.1.2 Mutation Accumulation

Mutation accumulation defines a set of measurements that track mutational changes across a lineage. These changes can be measured as the magnitude of the change (for real-valued genomes) or as the

total count of changes (for discrete-valued genomes). Mutation effects can also be tracked to gain insights about their distribution along a given lineage. Measures of mutation accumulation along the lineages of successful individuals can help tease apart the importance of different types of mutational events relative to what is expected by chance.

In conjunction with collected fitness information, the *class* of a mutation (beneficial, deleterious, or neutral) can also be tracked. Different evolutionary conditions are expected to cause different distributions of mutations along a lineage [3]; deviations revealed by measures of mutation accumulation can act as a barometer for unexpected evolutionary dynamics. The number and magnitude of deleterious mutations along a lineage can tell us both about the ruggedness of the fitness landscape and about a lineage's ability to cross fitness valleys [9]. Similarly, an elevated measure of neutral mutations relative to beneficial or deleterious mutations can suggest that the fitness landscape has neutral space in which the lineage is spending most of its time drifting around.

2.1.3 Phenotypic Volatility

Phenotypic volatility addresses the rate at which a phenotype changes as you move down a lineage (although the same concept can be applied to specific phenotypic traits or other types of state). In systems with discrete/categorical phenotypes, this can be measured by counting the number of times the phenotype changes. A related but subtly different measurement in such systems is the number of unique phenotypes on a lineage. In most cases, these values will be similar; a discrepancy would suggest that the lineage was cycling through a set of phenotypes. Such behavior could, for example, be indicative of some form of evolutionary bet-hedging [4].

In systems with continuous-valued phenotypes, a subtly different approach is needed to measure phenotypic volatility, because there are no discrete state transitions. Instead, we can measure the overall variance in phenotype along a lineage. In some cases, it may be desirable to smooth out the noise inherent in a real-valued phenotype. We can do so by instead taking the variance of the moving average of fitness, to more closely approximate the idea of measuring phase transitions.

2.1.4 Summary Statistics

Each of these metrics can be calculated for each member of the population at each time step. Doing so, however, would produce an amount of data so large that it would be difficult to make sense of. Instead, we need to come up with ways to generate useful summaries. There are two main approaches to doing so: (1) choose a small number of representative lineages from a given time point, or (2) collect summary statistics about the distribution of metric values across the population.

A single lineage can be chosen by selecting the lineage of a representative organism (either the most fit or the most numerous). In populations where diverse strategies coexist, this approach can be uninformative, as any one lineage is unlikely to be representative of all successful lineages. One alternative is to filter out lineages that do not have offspring some predetermined number of generations later, as such lineages are likely not representative of an important subset of the population. Still, any approach based on measuring only a subset of lineages can be challenging to interpret when the current dominant lineage (or lineages) is replaced with a different one; such changes can introduce a discontinuity if the value is being measured over time. If graceful responses to changes in which lineage is dominant are required, it can be advantageous to instead measure summary statistics (e.g., mean, variance, and range) across the entire population.

In scenarios with frequent selective sweeps, the dominant lineage will likely be similar to the average lineage, as most of the population will be closely related. When the population contains more phylogenetic diversity, however, the dominant lineage may differ from the mean. Of course, the nature of such differences is likely informative about the evolutionary dynamics occurring in the population.

2.2 Phylogeny Metrics

These metrics operate on entire phylogenies rather than single lineages within a population, eliminating the need to identify a representative organism or lineage. Because they use data from the entire

population, phylogeny metrics can be more computationally expensive to calculate than single lineage metrics. On the other hand, because most lineages tend to share substantial history, phylogeny metrics can usually be calculated more rapidly than full-population lineage metrics. Note that phylogenies can be constructed with regard to any taxonomic level of organization, be it individual, genotype, phenotype, or other. Thus, when we refer generally to items in a phylogeny, we will use the term *taxa*.

A standard technique for saving memory and time when working with phylogenies in computational systems is to *prune* them, removing dead (extinct) branches. Since all of the phylogeny metrics we discuss here are borrowed from natural systems (where we do not have information about taxa without offspring), they all are designed to work on pruned phylogenies. Thus, for the remainder of this article, we will assume we are working with pruned phylogenies.

In populations without ecological forces promoting coexistence, phylogenies should coalesce periodically, resulting in pruned lineages that mostly consist of a single path. When there is strong selection, this coalescence should happen even more rapidly. Thus, phylogenies with topologies that deviate from that expectation are an indication of ecological interactions within the population. The metrics discussed here can provide insight into the nature of those interactions and their long-term evolutionary effects [13]. As a result, they are often referred to as phylogenetic diversity metrics [38].

An important distinction between phylogenies in natural versus computational systems is that natural phylogenies are generally inferred from extant taxa, whereas computational phylogenies are directly recorded. Inferred phylogenies do not contain internal nodes except at branch points. They also do not contain history prior to the most recent common ancestor (MRCA) of all extant organisms. For consistency, we exclude pre-MRCA taxa from our analyses. Here, we will not remove non-branching internal nodes, as these only serve to make our phylogenies more informative.

In some cases, however, it may be advantageous to remove non-branching internal nodes. Doing so produces a more abstract summary of the phylogenetic process. Such summaries are useful in cases where trees produced under different conditions (e.g., different numbers of generations) are being compared.

Here we provide a high-level summary of phylogeny metrics that we expect will be particularly useful. Figure 2 gives two example phylogenies on which we demonstrate several of the metrics discussed in this subsection. For more metrics and more detail on all of these metrics, see [38, 43].

2.2.1 Depth of Most-Recent Common Ancestor

The depth of the MRCA (i.e., the number of steps it is from the original ancestor) is an informative metric and is easy to calculate. A recent MRCA implies frequent selective sweeps and less long-term

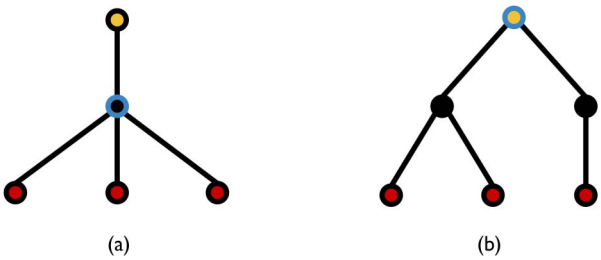

Figure 2. Both (a) and (b) show pruned phylogenetic trees on which we demonstrate several phylogenetic metrics. Nodes in each tree represent taxa, and edges between nodes represent ancestor-descendant relationships. In each tree, extant taxa (leaf nodes) are shaded red, the original ancestor (the root node) is shaded yellow, and the MRCA is outlined in blue. We are assuming that all branches have a length of 1 (in some contexts, branch lengths may be varied to more precisely reflect evolutionary distance between taxa). In (a), the phylogenetic richness (calculated as phylogenetic diversity [15]) equals 4, the phylogenetic divergence (calculated as the mean pairwise distance among extant taxa) equals 2, and the phylogenetic regularity (calculated as the variance of pairwise distances among extant taxa) equals 0. In (b), the phylogenetic richness equals 6, the phylogenetic divergence equals 3.33, and the phylogenetic regularity equals 1.33.

stable coexistence between clades. A downside to the depth of the MRCA as a metric is that any population that does have a stable ecology will likely never change its MRCA after the very beginning of evolution (although this property has the benefit of allowing us to detect stable coexistence in the population). Measuring the frequency with which the MRCA depth changes (i.e., the number of coalescence events) can also be informative, as some conditions can inflate the length of the lineage relative to other conditions without actually increasing the frequency of selective sweeps. This scenario is particularly likely when the population size is changing over time.

2.2.2 Phylogenetic Richness

Measurements of phylogenetic richness quantify the total amount of evolutionary history contained in a set of taxa. The most traditional metric of phylogenetic richness is *phylogenetic diversity*, which is calculated as the number of nodes in the minimum spanning tree from the MRCA to all extant taxa [15]. Another approach is to calculate the pairwise distances between all taxa and sum them [38]. A third approach is to sum evolutionary distinctiveness, a measure of a taxon's evolutionary uniqueness [20], across all extant taxa [38].

2.2.3 Phylogenetic Divergence

Measurements of phylogenetic divergence quantify how distinct the taxa in the population are from each other and are often averaged across individual taxa. For example, one option is to average the pairwise distances across all taxa in the population [41]. Similarly, phylogenetic divergence can be calculated by averaging the evolutionary distinctiveness across each taxon in the population.

2.2.4 Phylogenetic Regularity

Measurements of phylogenetic regularity quantify how balanced the branches are in a phylogeny and are often the variances of values calculated for individual taxa [38]. Just as the mean of the pairwise distances between all taxa in the population is a measure of phylogenetic divergence, their variance is a measure of phylogenetic regularity. The same is true of the variance of evolutionary distinctiveness across the population.

3 Visualizations

Data visualization is a critical component of data analysis. Summary measurements can be misleading [2], but we often have too many data to digest from purely numerical values. Visualizations allow us to structure our data in a way that is more intuitive to the viewer. Through this lens, we can confirm that our intuitions about the drivers of a system's behavior are correct. In this section, we will present some visualization techniques that are useful for understanding lineages and phylogenies. These visualizations have the power to substantially strengthen our intuitions about evolutionary systems. Later, we will use these visualizations to ensure that our metrics are capturing the information that we intend them to.

The idea that any lineage can be abstracted into a sequence of states (see Figure 1) is important for visualization as well as quantitative measurements. We must be careful about choosing the right level of abstraction; using too low a level can obscure broader patterns, whereas too high a level loses the pattern entirely. Furthermore, generating some of these visualizations at all becomes intractable with too little abstraction.

3.1 State Sequence Visualizations

A direct consequence of the fact that lineages are just sequences of states is the fact that we can visualize these sequences [24]. Each sequence can be represented as a series of stacked rectangles. The colors of these rectangles indicate which state they represent, and their heights represent how

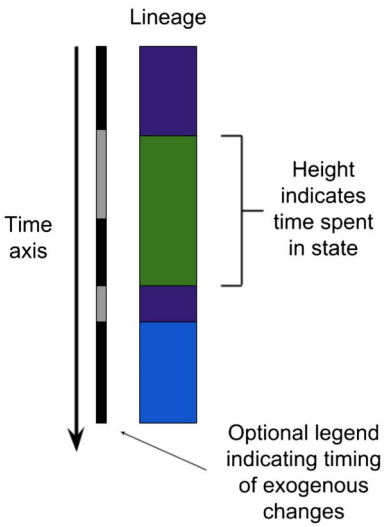

Figure 3. Schematic representation of state sequence visualization. Colors indicate different states. The optional legend along the side can be used to indicate any information relevant to understanding the drivers of the state changes.

long the lineage stayed in that state (see Figure 3). Placed side by side, these sequences provide an aggregate visual representation of the behavior of lineages from a given run or experiment.

3.2 Fitness Landscape Overlays

In order to understand how a particular evolutionary algorithm performs on a given problem, we can observe how that algorithm moves a population across the problem's fitness landscape [21]. A fitness landscape can be thought of as a map from solution representations into the quality of that solution (its fitness) [44]. Traditionally, fitness landscapes have most commonly been visualized as three-dimensional surface plots illustrating the effects of two continuously varying traits on fitness (e.g., [26], or Figure 7 of this article). One of these traits is the x-axis, one is the y-axis, and the fitness conferred by that combination of values is the z-axis. Although we will focus here on fitness landscapes that can be fully depicted in three dimensions, it is important to recognize that most real-world fitness landscapes have far more dimensions. Various dimensionality-reduction techniques can be applied to reduce them to three dimensions, but doing so may be misleading [16].

We can visualize the path that each lineage takes through the fitness landscape, mapping the x, y, and z (fitness) coordinates of each ancestor of each member of the population [40]. In some cases, it may also be helpful to overlay lineages on top of other spaces [10, 30]. Creating such a visualization entails condensing a large quantity of information into a limited space. When projected onto two dimensions, lineages can obscure and be obscured by parts of the fitness landscape (and each other).

Fortunately, we are entering an era where we no longer need to squeeze this information into two dimensions, confined to a computer screen or piece of paper. Over the past few years, virtual reality technology has advanced to the point where it is a viable tool for data visualization [35, 39]. To this end, we used the A-Frame framework [1] to build a three-dimensional data visualization that can be viewed in virtual reality (see Figure 8) [11].

A-Frame supports building 3D scenes and rendering them to a variety of platforms. In the simplest case, the visualization is rendered in WebGL, allowing it to be viewed in a standard web browser. Mouse interactions such as rotation make it possible to view the visualization from all angles, and WebGL's use of the graphics card allows it to render data-rich visualizations. A-Frame also supports rendering the page with WebVR, allowing it to be viewed using various virtual reality headsets. These platforms allow the user to explore the data in three dimensions. For the data interpretation in this

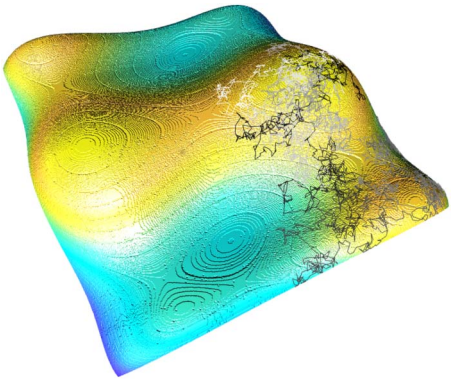

Figure 4. A lineage drawn on top of a fitness landscape.

article, we used an Oculus Rift to provide us with fine-grained control of which part of the visualization we were looking at.

Scenes in A-Frame are built by layering components on top of each other. The base component of our visualization is a three-dimensional surface plot of the fitness landscape. On top of this layer, we overlay a three-dimensional path from an individual that existed at the end of an evolutionary run all the way back to its earliest ancestor, passing through the locations of all intermediate ancestors along the way. A path can be added for each extant member of a final population to produce an entire overlaid phylogeny.

To make visualizations more interpretable, we introduced a color gradient along the lineage; in the visualization presented here, we use a grayscale lineage drawn onto a colorful landscape (see Figure 4). Our lineages transition from white to black as evolutionary time progresses, indicating when each portion of the lineage existed.

Our full visualization, complete with data, can be viewed on the web or using a virtual reality headset at https://emilydolson.github.io/fitness_landscape_visualizations.

3.3 Phylogenetic Trees

Phylogenetic trees are a time-honored data visualization in biology. They depict the parent-child relationships of taxa over the course of evolution (see Figure 5(a)). As with phylogeny metrics, memory usage is a primary concern in constructing these visualizations, as phylogenies can be very large. Applying pruning and choosing more abstract taxonomic units (as discussed in Section 2.2) are the best ways to mitigate this problem.

While simply drawing the phylogenetic tree is valuable in its own right, a lot of additional information can be added to these visualizations. Nodes in the tree can be colored based on features of the corresponding taxon, such as phenotype or fitness. Edges can be colored based on information about the transition between taxa, such as the types of mutations that occurred or the direction of fitness change. Note, though, that in large trees coloring nodes and edges the same way often makes patterns easier to see.

Another way to convey additional information is in the placement of nodes. In biology, the length of the edge between nodes is generally used to indicate the amount of time that elapsed between the origins of taxa. This convention also results in contemporaneous nodes being placed at the same level of the tree; in essence, the tree has a time axis running from top to bottom. In artificial life systems, we can take this idea a step further. Since we know the complete time span that a taxon existed for, we can depict it on the phylogenetic tree by plotting rectangles in place of circular nodes (see Figure 5(b)). The tops and bottoms of these rectangles correspond to the birth and death times of the taxon along the time axis. While this approach runs the risk of producing an overly complicated plot, the boxes can yield wildly different intuitions than circles would.

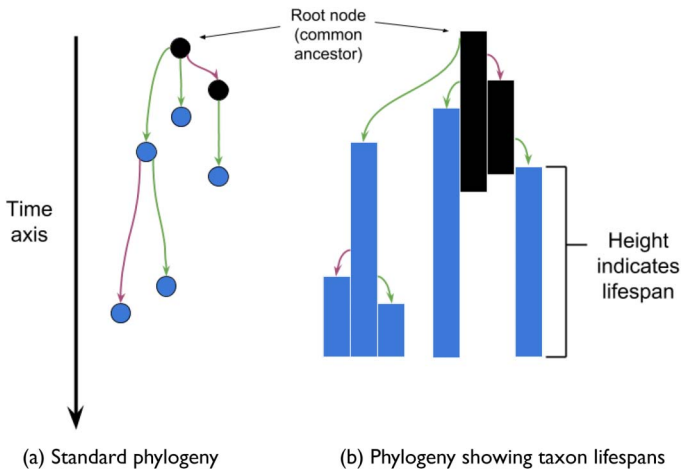

Figure 5. The same phylogeny depicted in two different ways. (a) A standard phylogeny, where nodes represent taxa and edges indicate parent-child relationships. In this phylogeny, node position along the time axis represents time of birth. Nodes are colored to distinguish taxa that are extant in the present (blue) from extinct taxa (black). Note that, because this is a pruned phylogeny, all leaf nodes are currently extant. Edges can be colored to convey whatever information is most useful. (b) The same phylogenetic tree, but with boxes indicating taxon life spans in place of circular nodes.

3.4 Muller Plots

Muller plots show the abundance and ancestry of taxa in a population over time [27, 31]. The percentage of a population occupied by each taxon at a given time point is depicted via a stacked-area plot—the vertical space of the plot is occupied by a stack of colored polygons, each with height proportional to the relative abundance of a corresponding taxon (see Figure 6). These polygons are connected over time (as in a streamgraph), showing the changes in each taxon's abundance.

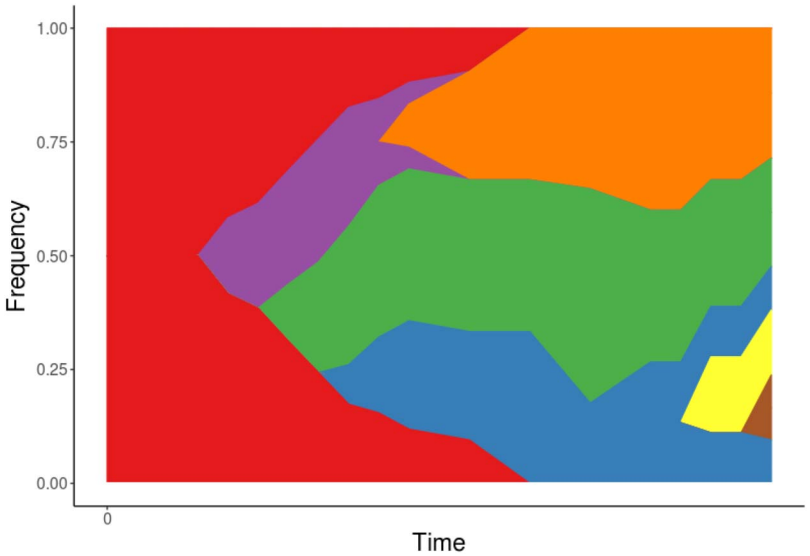

Figure 6. A Muller plot depicting the same phylogeny as the visualizations in Figure 5. The red region represents the root node. As it gradually goes extinct, the proportion of the figure it takes up gradually diminishes. Its three offspring are shown in purple, green, and blue.

Additionally, each taxon's polygon is depicted as emanating from its ancestor's polygon. Muller plots in this article were made using the ggmuller package [32] for the R statistical computing language [36].

When should you use a Muller plot versus a phylogenetic tree? In general, Muller plots are more useful when you care about the population dynamics of a relatively small number of fairly abstract taxonomic units. Phylogenetic trees can accommodate a larger number of less abstract taxonomic units and allow for greater flexibility in depicting evolutionary patterns.

4 Example Study Systems

We applied the metrics defined above to four benchmark functions from the GECCO competition on niching methods [26] and to four qualitatively different environments in the Avida digital evolution platform. Both of these evolutionary contexts are well understood and studied, making them particularly well suited for building our intuitions about what our proposed suite of ancestry-based metrics can tell us about evolutionary dynamics.

4.1 Niching Competition Benchmark Problems

To gain a broad understanding of our metrics, we applied them to a diverse subset of benchmark problems from the GECCO competition on niching methods: Himmelblau, Shubert, Composition Function 2, and Six-Humped Camel Back [26]. For each test problem, the x and y coordinates offered by a given organism are translated by the function into a fitness value. Because of their low dimensionality, we can fully visualize each problem's actual fitness landscape, allowing us to directly view how our ancestry-based metrics respond to the actual paths evolved lineages take through the fitness landscape under different conditions. We used the implementations of these problems at https://github.com/mikeagn/CEC2013 (C++ for fitness calculations during evolution, Python for post hoc analysis). Figure 7 shows the fitness landscapes defined by each of our four chosen test problems.

For each test problem, we evolved populations of 1000 organisms under a range of mutation rates and selection strengths for 5000 generations. Each organism's genome consisted of two floating-point numbers that defined its position in the fitness landscape. We initialized populations by randomly generating a number of organisms equal to the population size. To determine which organisms reproduced each generation, we used tournament selection. We evolved populations under five different tournament sizes: one, two, four, eight, and sixteen. Tournament size represents strength of selection: higher tournament sizes correspond to strong selection and lower tournament sizes correspond to weak selection [5]. A tournament size of one is equivalent to no selection pressure (i.e., every organism in the population has an equal chance of being selected to reproduce). Organisms selected to reproduce did so asexually. Values in an offspring's genome were mutated by adding noise given by a normal distribution with a mean of 0; the *mutation rate* of a treatment determined the standard deviation used to define this normal distribution and was given as a proportion of the test problem's domain. We prevented mutations from causing a value to exceed the valid domain of the given problem. For each problem and tournament size, we evolved populations at eight mutation rates: 1e−08, 1e−07, 1e−06, 1e−05, 1e−04, 1e−03, 1e−02, and 1e−01.

We also ran a second set of experiments using these benchmark problems to explore the influence of ecological dynamics on our suite of ancestry-based metrics. For these experiments, we generated a stable ecology using the Eco-EA algorithm as a selection technique [17]. Eco-EA is a technique for creating niches that promote stable diversification in the context of an evolutionary algorithm. In our test problems, we created niches associated with spatial locations across the fitness landscape. For all experimental conditions, we ran ten replicates, each with a unique random number seed. Our experiment is implemented using the Empirical library; our implementation is included in the supplemental material for this article [22].

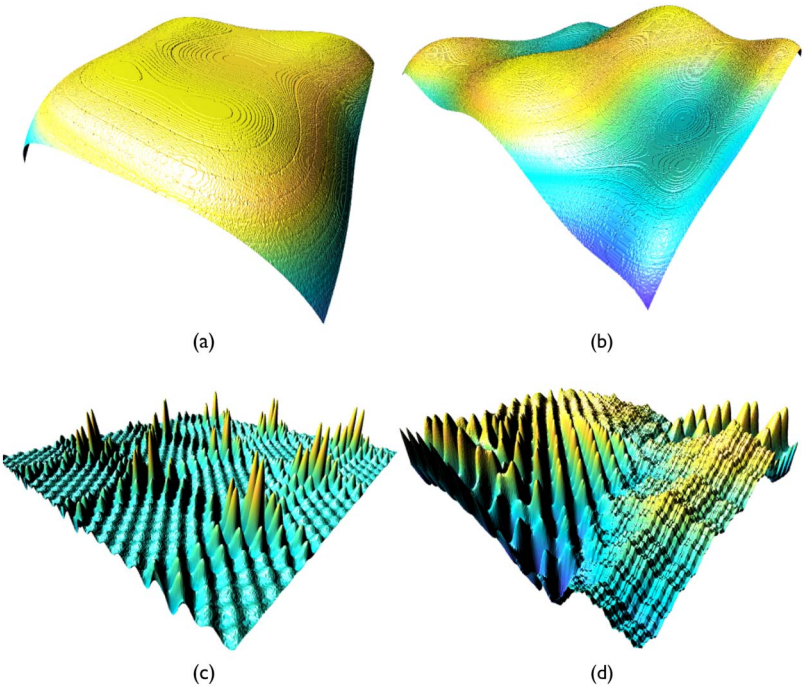

Figure 7. The fitness landscapes used in this experiment: (a) Himmelblau, (b) Six-Humped Camel Back, (c) Shubert, and (d) Composition Function 2. Interactive versions available at https://emilydolson.github.io/fitness_landscape_visualizations.

4.2 Avida

The Avida digital evolution platform [34] is a well-established artificial life system that has been used to study a wide range of evolutionary dynamics [12, 18, 25, 45]. Avida's track record and popularity make it a good next step in testing our intuitions for how our suite of lineage and phylogeny metrics will respond under a range of evolutionary dynamics.

In Avida, populations of self-replicating digital organisms compete for space in a finite, toroidal grid. Each digital organism has a set of virtual hardware and a circular genome composed of assembly-language-like instructions. An organism's virtual hardware contains components such as a central processing unit (CPU) for executing instructions, registers used for computation, memory stacks, and input and output buffers. The instruction set of Avida is Turing-complete and enables organisms to perform basic computations, control their own execution flow, and replicate. Further, the Avida instruction set is syntactically robust—all possible genetic sequences are syntactically valid, even if they do not perform a meaningful computation.

Organisms in Avida replicate asexually by copying themselves instruction by instruction and dividing; the copy and divide operations, however, are not perfect and can result in mutated offspring. When an organism successfully replicates, the resulting offspring is placed randomly in the world, replacing that location's former occupant. Thus, there is selection pressure for organisms to replicate quickly before being copied over by others.

A digital organism's replication speed can be improved by reducing the number of instruction executions required for an organism to copy itself (e.g., by optimizing the self-copy genetic machinery) or by increasing its metabolic rate. The metabolic rate determines the speed at which an organism executes instructions in its genome; a higher metabolic rate allows an organism to execute its genome faster, which, in turn, allows the organism to copy itself faster. Initially, an organism's metabolic rate is approximately proportional to its genome length; however, organisms can influence their metabolic rate by performing particular functions, such as mathematical computations. In this way, we can differentially reward or punish the performance of different functions.

We evolved thirty replicate populations of size 500 for 200,000 generations in four different environments: a minimal environment, the logic-9 environment, a limited-resource environment, and a simple changing environment. Across all environments, instruction-copy operations erred at a rate of 0.0075, resulting in substitution mutations. Divide operations could result in single-instruction insertions and deletions at a per-divide rate of 0.05.

In the *minimal environment*, we did not reward any functions. Because these digital organisms cannot influence their execution speed by performing functions, selection pressure is entirely focused on optimizing the efficiency at which organisms can self-replicate. Thus, we expected lineages with short generation times to be most successful, resulting in long lineages comprising many individuals. This minimal environment has only a single niche, so we expected to see low phylogenetic diversity and frequent population-wide selective sweeps, each leading to a change in most recent common ancestor (MRCA).

In the *logic-9 environment*, we rewarded the performance of all nontrivial one- and two-input Boolean logic functions: NAND, NOT, OR-NOT, AND, OR, AND-NOT, NOR, XOR, and EQUALS (for more information on these logic functions in Avida see [25]). In addition to selection pressure for efficient self-replication, there is selection pressure for organisms to improve their instruction execution speed by performing logic functions. Logic-9, like the minimal environment, is a single-niche environment, so we expected to see low phylogenetic diversity and frequent selective sweeps. However, because this environment rewards performing functions, we expected to see lineages with longer generation times than those in the minimal environment. In addition to looking at individuals along a lineage, we compressed lineages into sequences of phenotypic function profiles. The nine rewarded Boolean logic functions are of different computational complexities, NAND and NOT being the simplest, and EQUALS being the most complex. Further, in the logic-9 environment, simpler functions are building blocks for the more complex functions. Thus, we expected to see functions appear on lineages in roughly the order of function complexity. We can also use lineages compressed into sequences of function profiles to measure the rate of function acquisition (phenotypic volatility) or to look for trends in how long lineages spend in a particular phenotypic state, which can help untangle which phenotypic transitions might be the most challenging.

Like the logic-9 environment, the *limited-resource environment* rewards the performance of all nontrivial one- and two-input Boolean logic functions; each of these computational functions, however, is associated with a limited pool of resources. When an organism performs one of the nine logic functions, it consumes resources associated with that particular function (lowering its concentration) in proportion to the resource's availability. Rather than adjusting organisms' execution speeds based directly on function performance, in the limited-resource environment an organism's execution speed is adjusted as a function of the amount of resources it has collected. The limited-resource environment has been shown to support stable ecologies via negative frequency-dependent selection [8]. Here, we expected the phylogeny metrics to reveal high phylogenetic diversity relative to the other three environments. Because previous work has shown that the limited-resource environment supports the stable coexistence of multiple ecotypes, we expected infrequent changes in the MRCA.

In the *changing environment*, the rewarded functions cycled between two opposing states (at a rate of 200 updates[1]): ENV-NAND (where we rewarded NAND and punished NOT) and ENV-NOT (where we rewarded NOT and punished NAND). ENV-NAND and ENV-NOT were configured so that no phenotype can be optimal across both environment states. To achieve maximal fitness in either ENV-NAND or ENV-NOT, organisms must perform *only* the focal function. In such environments, successful lineages should move to occupy a region of genotype space such that it is easier to mutationally toggle between a phenotype of only performing NAND and a phenotype of only performing NOT [7]. In this treatment, we expected to see a high number of phenotype changes but a low total number of unique phenotypes along evolved lineages. Further, we expected the rate of phenotypic change along successful lineages to

1 One update in Avida is defined as the amount of time it takes for the average organism to execute 30 instructions. See [34] for more details.

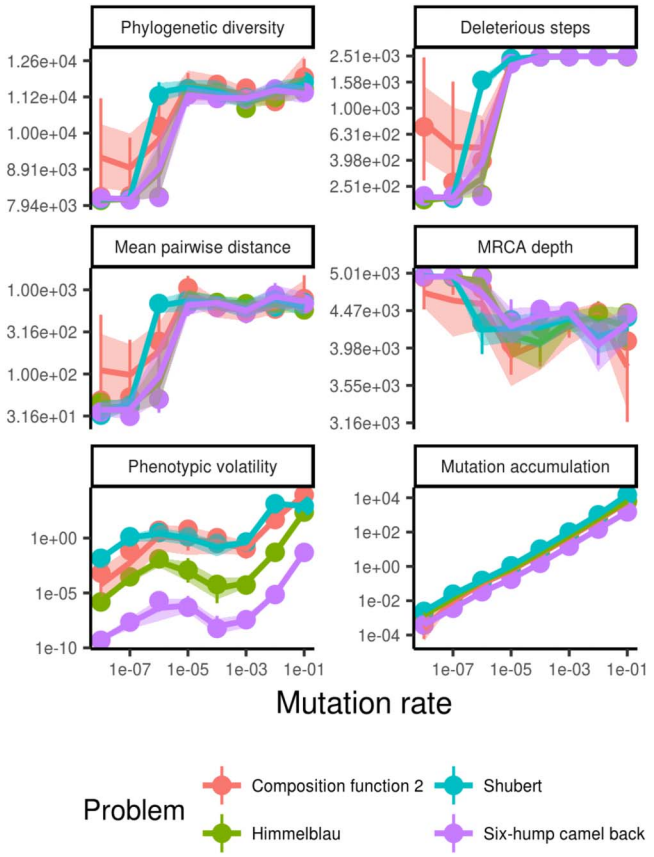

Figure 8. Values of example metrics across different mutation rates for each of the four problems. For these problems, all lineage-based metrics are calculated on the lineage of the fittest organism at the final time point; population-level means behaved similarly. All experiments shown here used a tournament size of 4. Circles are medians, vertical lines show interquartile range, and the shaded area is a bootstrapped 95% confidence interval around the mean. Note that both axes are on log scales.

approach the rate of environmental change, as these lineages would have been more likely to produce offspring well adapted to the next environment.

The specific configurations used in these experiments can be found in the supplemental material for this article [22].

5 Data Analysis

We analyzed trends in our metrics using the R statistical computing language [36]. Specifically, we used the ggplot2 package for all graphs included in this article [42]. All analysis scripts are available in the supplemental material for this article [22].

6 Results and Discussion

6.1 Niching Competition Benchmark Problems

Overall, our results were consistent with evolutionary theory. As the mutation rate increases, coalescence takes longer, as evidenced by the fact that the MRCA is farther back in time at higher

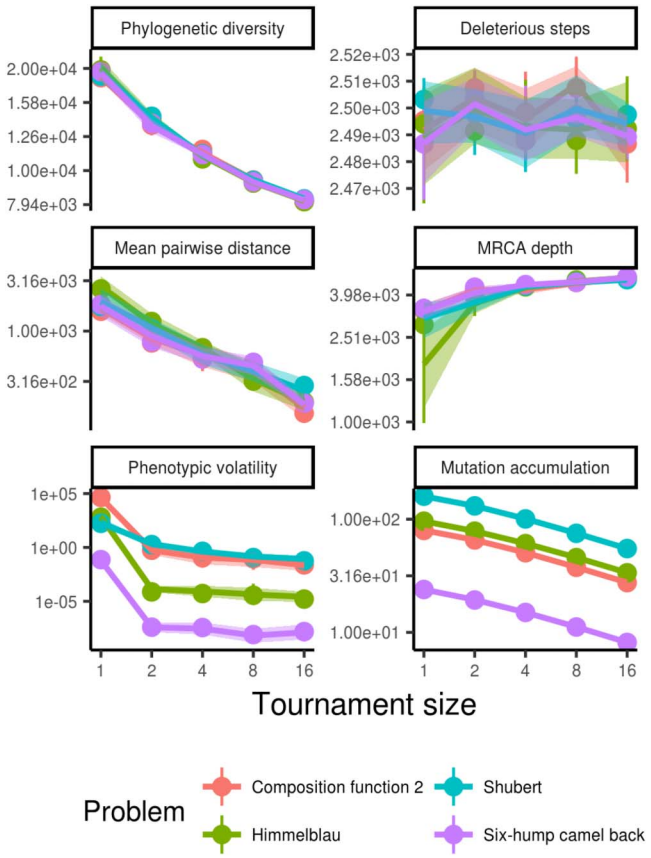

Figure 9. Values of example metrics across different tournament sizes for each of the four problems. All experiments shown here used a mutation rate of 0.001. For these problems, all lineage-based metrics are calculated on the lineage of the fittest organism at the final time point; population-level means behaved similarly. Circles are medians, vertical lines show interquartile range, and the shaded area is a bootstrapped 95% confidence interval around the mean. Note that both axes are on log scales.

mutation rates (see Figure 8). Consequently, phylogenetic richness (as measured by phylogenetic diversity) is higher at high mutation rates. Phylogenetic divergence, measured here as mean pairwise distance between taxa, is similarly higher at high mutation rates. Evolutionary distinctiveness, being another measure of phylogenetic divergence, behaved almost identically [22]. The variance of evolutionary distinctiveness and the pairwise distance between taxa (phylogenetic regularity metrics) behaved similarly to the phylogenetic divergence metrics. This pattern makes sense, as most phylogenetic divergence on these landscapes will produce unbalanced phylogenetic trees. If there were stable coexistence between multiple clades, we would expect to see a reduced correlation between the phylogenetic divergence metrics and the phylogenetic regularity metrics. Increased mutation rate also increases the number of deleterious steps taken, a logical consequence of increasing mutation relative to strength of selection. The relationship between phenotypic volatility and mutation rate appears to fluctuate. This phenomenon is worthy of further study, but appears to be related to the probability of mutations moving a lineage between peaks of equal height. Completely unsurprisingly, mutation accumulation increases linearly with mutation rate.

Similarly, increasing tournament size generally increases the rate of coalescence, as higher tournament sizes correspond to stronger selection (see Figure 9). As a result, all of the measures of

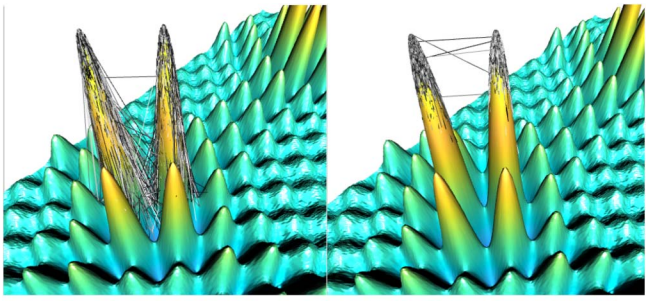

Figure 10. A close-up on two adjacent peaks in the Shubert function fitness landscape. Lineages are depicted as paths fading from white to black over evolutionary time. The lineages shown here evolved under a mutation rate of 0.01. (a) was evolved using a tournament size of 2, whereas (b) was evolved using a tournament size of 16. These figures neatly illustrate how increased tournament size keeps the lineage near the tops of the peaks.

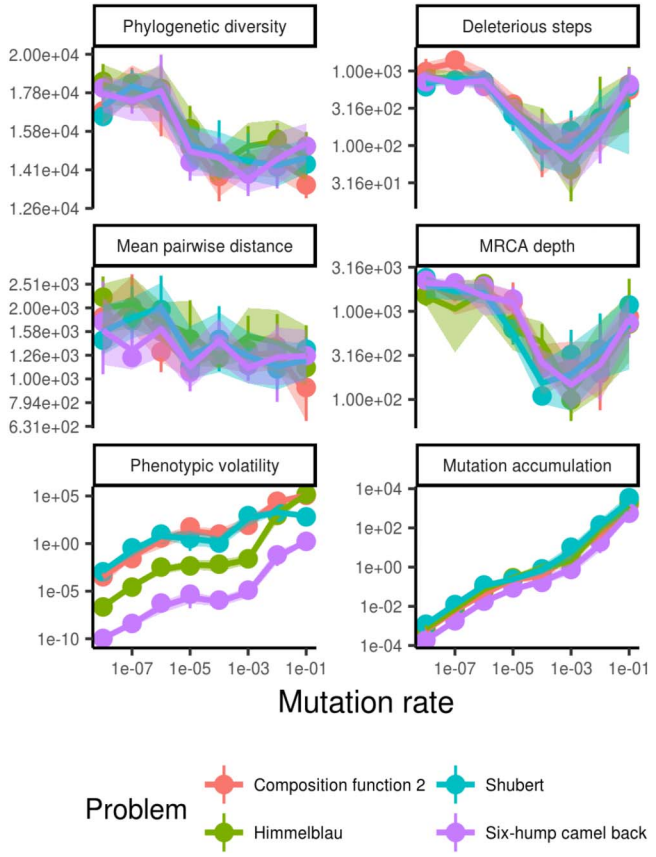

Figure 11. Values of example metrics across different mutation rates for each of the four problems under a diversity-preserving selection regime, Eco-EA. For these problems, all lineage-based metrics are calculated on the lineage of the fittest organism at the final time point; population-level means behaved similarly. All experiments shown here used a tournament size of 4. Circles are medians, vertical lines show interquartile range, and the shaded area is a bootstrapped 95% confidence interval around the mean. Note that both axes are on log scales.

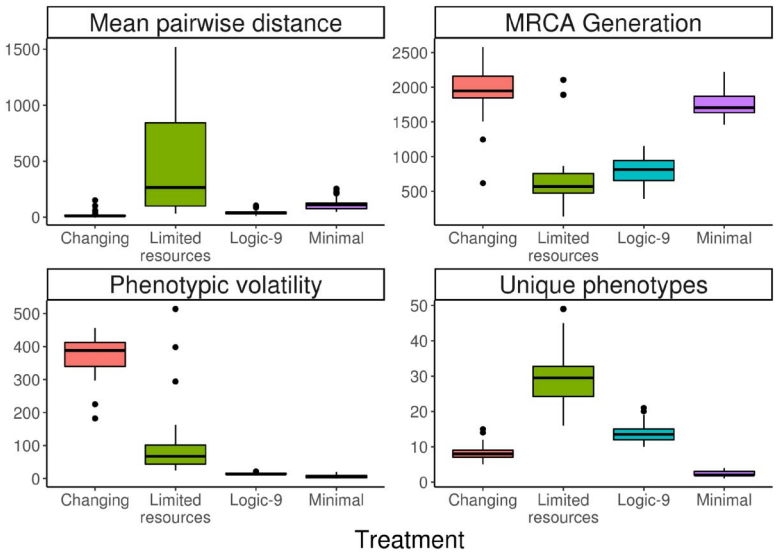

Figure 12. A sampling of informative lineage metrics from Avida calculated after approximately 10,000 generations. "Mean pairwise distance" is a measure of phylogenetic divergence; higher values imply greater phylogenetic diversity. "MRCA generation" is the generation at which the current most recent common ancestor occurred (lower numbers mean it was longer ago). For Avida, all lineage-based metrics are calculated on the lineage of the most numerous phenotype. "Unique phenotypes" is the count of unique phenotypes that occurred along the dominant lineage.

phylogenetic richness and divergence decrease as tournament size increases. MRCA depth, on the other hand, increases, directly reflecting the increased frequency of selective sweeps.

Surprisingly, there is no clear effect of tournament size on the count of deleterious steps along the dominant lineage (as evidenced by the fact that the confidence intervals all overlap). Values for all selection schemes and tournament sizes hover near 2500, meaning that a deleterious step is taken in roughly half of the 5000 generations. This result is partially an effect of mutation rate; at the lowest mutation rate, there is a clear trend toward fewer deleterious steps as tournament sizes increase [22]. However, the effect of the mutation rate on the relationship between tournament size and dominant deleterious steps is complex, particularly for Composition Function 2 [22]. These trends likely share a common cause with the thresholding effect evident in Figure 8, where the number of deleterious steps along the dominant lineage abruptly climbs between mutation rates of 10^{-7} and 10^{-5} and remains relatively flat over other mutation rates. Based on an inspection of the 3D fitness landscape visualizations, we can see that this is not an effect of lineages moving from peak to peak; at most mutation rates, they tend to remain on a single peak (with the exception of adjacent peaks in the Shubert function; see Figure 10). Thus, we can infer that this effect is the result of a driftlike phenomenon where, at sufficiently high mutation rates, all members of the population are constantly somewhat displaced from their local fitness peak.

Having reinforced our intuition about these metrics in a simple system, we can now expand them to a slightly more complex system. A large proportion of interesting short-term evolutionary dynamics relate to interaction between individuals in the population (i.e., ecological dynamics). In particular, such interactions often promote the stable coexistence of clades occupying different niches. Thus, it is important to establish a baseline for how our metrics respond to ecological coexistence.

Indeed, the presence of stabilizing ecological dynamics substantially changes the values we observe for most metrics (see Figure 11). Perhaps the least surprising of these is that the MRCA depth is far lower than it was for tournament selection, reflecting the rarity of coalescence events under these conditions. Consequently, phylogenetic diversity is higher, as the extant population represents a greater amount of

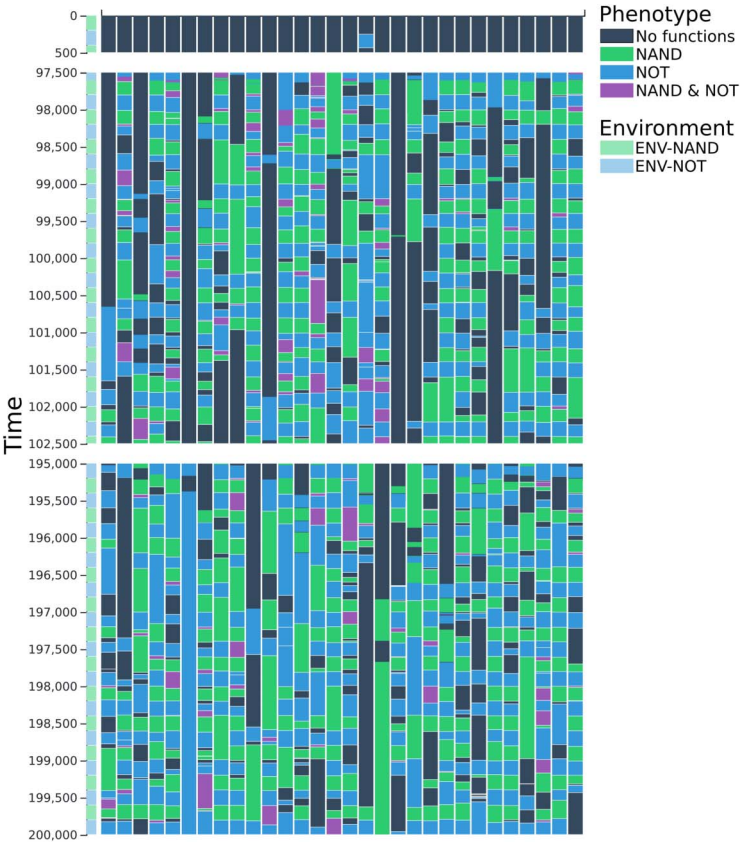

Figure 13. State sequence visualization of phenotypes along the dominant lineage in the changing environment condition. The key along the left indicates the environment's state over time. Each column is one replicate. From this visualization, we can see that most successful lineages are able to adapt to a change in the environment relatively quickly. We can also see that there is a fair amount of variation in the precise timings of these changes and whether they occur for every change of the environment.

evolutionary history. Relatedly, the mean pairwise distance among extant taxa is higher in the presence of ecology, as clades in different niches continue to diverge. Interestingly, the relationship of many metrics (e.g., deleterious steps and phylogenetic diversity) to the mutation rate is reversed in the presence of ecology. Explaining the underlying mechanisms behind these distinctions is beyond the scope of this article, but the ease with which the metrics identified their presence clearly indicates their power.

6.2 Avida

The data from Avida are also consistent with our expectations and additionally point towards some interesting directions for future investigation. As predicted, the only environment that preserved phylogenetic diversity was the limited-resource environment. Lacking diversity-preserving dynamics, the other environments all repeatedly lose phylogenetic diversity due to selective sweeps (see Figure 12).

Looking at phenotypic volatility, we see that the changing environment condition has dramatically higher volatility than the others. This result is precisely what we hypothesized, as the phenotype should change approximately every time the environment does (see Figure 12). Indeed, from Figure 13, we

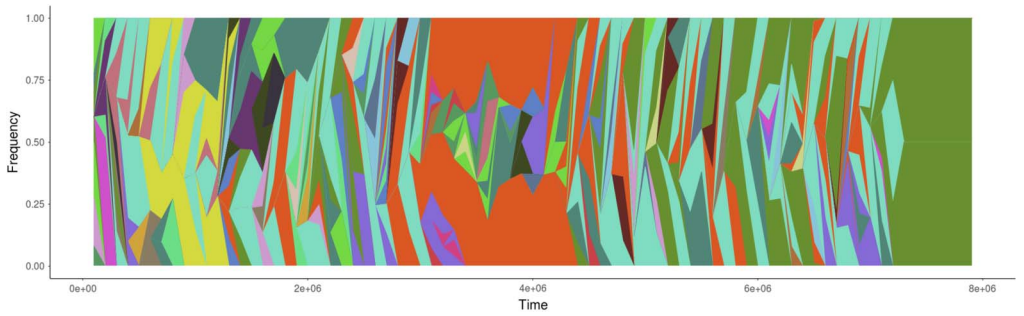

Figure 14. Muller plot from a representative run of Avida in the limited-resource environment. This plot shows the population dynamics of different phenotypes over 8,000,000 updates (a unit of time in Avida that is proportional to the number of CPU cycles used thus far). Only phenotypes that make up at least 5% of the population at at least one time point are shown. Note the region in the middle of the plot where we can see repeated selective sweeps that are contained to a single niche.

can see that this predicted mechanism is accurate. The limited-resource environment has lower phenotypic volatility than the changing environment, but higher phenotypic volatility than the other two environments. This observation is consistent with the fact that the limited-resource environment is also a changing one; competitive pressures shift with the composition of the population. This interpretation is corroborated by Muller plots from the limited-resource condition, which show frequent shifts in the abundance of various phenotypes (see Figure 14). We can gain further insight by comparing phenotypic volatility to the count of unique phenotypes along a lineage. Despite its high phenotypic volatility, the changing environment condition produces lineages with relatively few unique phenotypes. This finding makes sense, as the changing environment is only creating pressure to toggle back and forth between two phenotypes. The limited-resource condition, however, has a high count of unique states relative to its phenotypic volatility. From this comparison, we can conclude that the

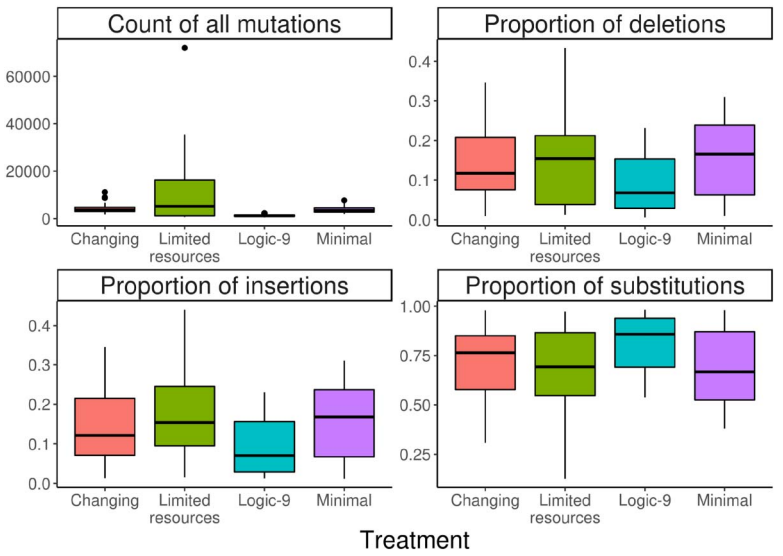

Figure 15. Mutations along the dominant lineage across treatments in Avida, calculated after approximately 10,000 generations. The upper left shows the total count of mutations of all types. The other plots show the quantity of each mutation type as a proportion of the total.

competition in this environment is creating pressure for the population to explore new phenotypic states, rather than simply cycling among a fixed set.

From the mutation accumulation data, we can begin to glean information about the precise dynamics of adaptation (see Figure 15). Across the board, the limited-resource environment has more mutations than any other condition, which is consistent with the greater number of unique phenotypes explored. The changing environment has the second greatest mutation accumulation, presumably due to the mutations that are required to change phenotype as the environment changes. A relatively high percentage of the mutations that accumulated were substitutions. Determining whether this is the result of chance or an indication that substitutions are particularly valuable will require further study.

7 Conclusions

We have demonstrated that these analysis techniques conform to our intuitions in well-understood systems. In more complex environments, they provide a quick way to identify interesting behavior to study further. We believe they are now ready to be adopted more broadly.

Our goals for this work are twofold: (1) to suggest a set of analyses that will improve our capacity to quantitatively understand evolutionary histories in digital evolution experiments, and (2) to spark a conversation in the computational evolution community about how to quantify, interpret, and compare observed evolutionary histories. With feedback from the community, we will expand our suite of lineage and phylogeny metrics, compiling accessible descriptions and examples of each metric.

More generally, we suggest that artificial life researchers should keep phylogeny- and lineage-based analyses in mind when trying to answer specialized questions as well. Recently, we proposed a suite of metrics for studying open-ended evolution that leverage the power of perfect phylogenetic information to screen systems for interesting evolutionary dynamics [14]. By comparing observed phylogenies with those that we would expect under certain theoretical assumptions, we can screen for specific behavior. This general approach should be adaptable to a variety of dynamics of interest.

Acknowledgments

We thank members of the MSU Digital Evolution Lab for helpful comments and suggestions on this manuscript. This research was supported by the National Science Foundation (NSF) through the BEACON Center (Cooperative Agreement DBI-0939454), Graduate Research Fellowships to E. D. and A. L. (Grant No. DGE-1424871), and NSF Grant No. DEB-1655715 to C. O. Michigan State University provided computational resources through the Institute for Cyber-Enabled Research and the Digital Scholarship Lab. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF or MSU.

References

1. A-Frame authors. (2018). A-Frame: A web framework for building virtual reality experiences. https://github.com/aframevr/aframe.
2. Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21.
3. Barrick, J. E., & Lenski, R. E. (2013). Genome dynamics during experimental evolution. *Nature Reviews Genetics*, *14*(12), 827.
4. Beaumont, H. J., Gallie, J., Kost, C., Ferguson, G. C., & Rainey P. B. (2009). Experimental evolution of bet hedging. *Nature*, *462*(7269), 90.
5. Blickle, T., & Thiele, L. (1995). A mathematical analysis of tournament selection. In *Proceedings of the Sixth International Conference on Genetic Algorithms* (pp. 9–16). Citeseer.

6. Burlacu, B., Affenzeller, M., Kommenda, M., Winkler, S., & Kronberger, G. (2013). Visualization of genetic lineages and inheritance information in genetic programming. In C. Blum (Ed.), *Proceedings of the 15th Annual Conference Companion on Genetic and Evolutionary Computation* (pp. 1351–1358). New York: ACM.
7. Canino-Koning, R., Ofria, C., & Wiser, M. (2018). Fluctuating environments select for short-term phenotypic variation leading to long-term exploration. *bioRxiv*.
8. Cooper, T. F., & Ofria, C. (2003). Evolution of stable ecosystems in populations of digital organisms. In R. K. Standish, M. A. Bedau, & H. A. Abbass (Eds.), *Artificial life VIII: Proceedings of the Eighth International Conference on Artificial Life* (pp. 227–232). Cambridge, MA: MIT Press.
9. Covert, A. W., Lenski, R. E., Wilke, C. O., & Ofria, C. (2013). Experiments on the role of deleterious mutations as stepping stones in adaptive evolution. *Proceedings of the National Academy of Sciences of the U.S.A.*, *110*(34), E3171–E3178.
10. Dolson, E., & Ofria, C. (2017). Spatial resource heterogeneity creates local hotspots of evolutionary potential. In C. Knibbe, G. Beslon, D. Parsons, D. Misevic, J. Rouzaud-Cornabas, N. Bredeche, S. Hassas, O. Simonin, & H. Soula (Eds.), *ECAL 2017: The Fourteenth European Conference on Artificial Life* (vol. 29), (pp. 122–129). Cambridge, MA: MIT Press.
11. Dolson, E., & Ofria, C. (2018). Visualizing the tape of life: Exploring evolutionary history with virtual reality. In H. Aguirre (Ed.), *Proceedings of the Genetic and Evolutionary Computation Conference Companion, GECCO '18* (pp. 1553–1559). New York: ACM.
12. Dolson, E., Wiser, M. J., & Ofria, C. A. (2016). The effects of evolution and spatial structure on diversity in biological reserves. In C. Gershenson, T. Froese, J. M. Siqueiros, W. Aguilar, E. J. Izquierdo, & H. Sayama (Eds.), *Artificial life XV: Proceedings of the Fifteenth International Conference on Artificial Life* (pp. 434–440). Cambridge, MA: MIT Press.
13. Dolson, E. L., Banzhaf, W., & Ofria, C. (2018). Ecological theory provides insights about evolutionary computation. *PeerJ Preprints*, *6*, e27315v1.
14. Dolson, E. L., Vostinar, A. E., Wiser, M. J., & Ofria, C. (2019). The MODES toolbox: Measurements of open-ended dynamics in evolving systems. *Artificial Life*, *25*(1), 50–73.
15. Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, *61*(1), 1–10.
16. Gavrilets, S. (2010). High-dimensional fitness landscapes and speciation. In M. Pigliucci & G. B. Müller (Eds.), *Evolution—the extended synthesis* (pp. 45–80). Cambridge, MA: MIT Press.
17. Goings, S., Goldsby, H. J., Cheng, B. H., & Ofria, C. (2012). An ecology-based evolutionary algorithm to evolve solutions to complex problems. *Artificial Life*, *13*, 171–177.
18. Goldsby, H. J., Dornhaus, A., Kerr, B., & Ofria, C. (2012). Task-switching costs promote the evolution of division of labor and shifts in individuality. *Proceedings of the National Academy of Sciences of the U.S.A.*, *109*(34), 13686–13691.
19. Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Cambridge, MA: Elsevier.
20. Isaac, N. J. B., Turvey, S. T., Collen, B., Waterman, C., & Baillie, J. E. M. (2007). Mammals on the EDGE: Conservation priorities based on threat and phylogeny. *PLoS ONE*, *2*(3), e296.
21. Kauffman, S., & Levin, S. (1987). Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, *128*(1), 11–45.
22. Lalejini, A., & Dolson, E. (2019). Code, analysis, and configurations for interpreting the tape of life. https://doi.org/10.5281/zenodo.2576497.
23. Lalejini, A., Dolson, E., Ferguson, A., Bohm, C., & Rainford, P. F. (2019). Alife data standards, version 1.0-alpha. https://doi.org/10.5281/zenodo.2577410.
24. Lalejini, A., & Ofria, C. (2016). The evolutionary origins of phenotypic plasticity. In C. Gershenson, T. Froese, J. M. Siqueiros, W. Aguilar, E. J. Izquierdo, & H. Sayama (Eds.), *Proceedings of the Artificial Life Conference 2016* (pp. 372–379). Cambridge, MA: MIT Press.
25. Lenski, R. E., Ofria, C., Pennock, R. T., & Adami, C. (2003). The evolutionary origin of complex features. *Nature*, *423*(6936), 139–144.
26. Li, X., Engelbrecht, A., & Epitropakis, M. G. (2013). *Benchmark functions for CEC'2013 special session and competition on niching methods for multimodal function optimization* (Technical Report). Melbourne, Australia: RMIT University.

27. Maddamsetti, R., Lenski, R. E., & Barrick, J. E. (2015). Adaptation, clonal interference, and frequency-dependent interactions in a long-term evolution experiment with *Escherichia coli*. *Genetics*, *200*(2), 619–631.

28. McPhee, N. F., Casale, M. M., Finzel, M., Helmuth, T., & Spector, L. (2016). Visualizing genetic programming ancestries. In T. Friedrich (Ed.), *Proceedings of the 2016 Genetic and Evolutionary Computation Conference* (pp. 1419–1426). New York: ACM.

29. McPhee, N. F., Donatucci, D., & Helmuth, T. (2016). Using graph databases to explore the dynamics of genetic programming runs. In R. Riolo, W. Worzel, M. Kotanchek, & A. Kordon (Eds.), *Genetic programming theory and practice XIII* (pp. 185–201). Cham, Switzerland: Springer International.

30. Mouret, J.-B., & Clune, J. (2015). Illuminating search spaces by mapping elites. arXiv:1504.04909 [cs, q-bio].

31. Muller, H. J. (1932). Some genetic aspects of sex. *The American Naturalist*, *66*(703), 118–138.

32. Noble, R. (2017). *ggmuller: Create Muller plots of evolutionary dynamics*. R package version 0.5.2.

33. Ofria, C., Dolson, E., Lalejini, A., Fenton, J., Moreno, M. A., Jorgensen, S., Miller, R., Stredwick, J., Zaman, L., Schossau, J., Gillespie, L., Nitash, C. G., & Vostinar, A. (2019). Empirical. https://doi.org/10.5281/zenodo.2575607.

34. Ofria, C., & Wilke, C. O. (2004). Avida: A software platform for research in computational evolutionary biology. *Artificial Life*, *10*(2), 191–229.

35. Olshannikova, E., Ometov, A., Koucheryavy, Y., & Olsson, T. (2015). Visualizing big data with augmented and virtual reality: Challenges and research agenda. *Journal of Big Data*, *2*(1), 22.

36. R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

37. Standish, R. K., & Galloway, J. (2002). Visualising Tierra's tree of life using Netmap. In *ALife VIII Workshop proceedings* (p. 171). Cambridge, MA: MIT Press.

38. Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., Grenyer, R., Helmus, M. R., Jin, L. S., Mooers, A. O., Pavoine, S., Purschke, O., Redding, D. W., Rosauer, D. F., Winter, M., & Mazel, F. (2017). A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological Reviews*, *92*(2), 698–715.

39. van Dam, A., Laidlaw, D. H., & Simpson, R. M. (2002). Experiments in immersive virtual reality for scientific visualization. *Computers & Graphics*, *26*(4), 535–555.

40. Virgo, N., Agmon, E., & Fernando, C. (2017). Lineage selection leads to evolvability at large population sizes. In C. Knibbe, G. Beslon, D. Parsons, D. Misevic, J. Rouzaud-Cornabas, N. Bredeche, S. Hassas, O. Simonin, & H. Soula (Eds.), *ECAL 2017: The Fourteenth European Conference on Artificial Life* (pp. 420–427). Cambridge, MA: MIT Press.

41. Webb, C. O., & Losos, A. E. J. B. (2000). Exploring the phylogenetic structure of ecological communities: An example for rain forest trees. *The American Naturalist*, *156*(2),145–155.

42. Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

43. Winter, M., Devictor, V., & Schweiger, O. (2013). Phylogenetic diversity and nature conservation: Where are we? *Trends in Ecology & Evolution*, *28*(4), 199–204.

44. Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In *Proceedings of the Sixth International Congress of Genetics* (pp. 356–366). Brooklyn, NY: Brooklyn Botanic Garden.

45. Zaman, L., Meyer, J. R., Devangam, S., Bryson, D. M., Lenski, R. E., & Ofria, C. (2014). Coevolution drives the emergence of complex traits and promotes evolvability. *PLoS Biology*, *12*(12), e1002023.