DIGITAL EVOLUTION IN EXPERIMENTAL PHYLOGENETICS AND EVOLUTION EDUCATION

By

Cory Kohn

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Integrative Biology—Doctor of Philosophy
Ecology, Evolution, and Behavior—Dual Major

2021

ABSTRACT

DIGITAL EVOLUTION IN EXPERIMENTAL PHYLOGENETICS AND EVOLUTION EDUCATION

By

Cory Kohn

The creation and evaluation of known evolutionary histories and the implementation of student investigatory experiences on evolution are difficult endeavors that have only recently been feasible. The research presented in this dissertation is related in their shared use of digital evolution with Avidians as a model study system, both to conduct science research in experimental phylogenetics and to conduct education research in curricular intervention to aid student understanding.

I first present background discussions on the Avidian digital evolution study system—as implemented in Avida and Avida-ED—and its favorable use in experimental phylogenetics and biology education owing to its greater biological realism than computational simulations, and greater utility and generality than biological systems. Prior work on conducting experimental evolution for use in phylogenetics and work on developing undergraduate lab curricula using experimental evolution are also reviewed.

I establish digital evolution as an effective method for phylogenetic inference validation by demonstrating that results from a known Avidian evolutionary history are concordant, under similar conditions, to established biological experimental phylogenetics work. I then further demonstrate the greater utility and generality of digital evolution over biological systems by experimentally testing how phylogenetic accuracy may be reduced by complex evolutionary processes operating singly or in combination, including absolute and relative degrees of evolutionary change between lineages (i.e., inferred branch lengths), recombination, and natural selection. These results include that directional selection aids phylogenetic inference, while stabilizing selection impedes it. By evaluating clade accuracy and clade resolvability across treatments, I evaluate measures of tree support and its presentation in the form of consensus topologies and I offer several general recommendations for systematists.

Using a larger and more biologically realistic experimental design, I systematically examine a few of the complex processes that are hypothesized to affect phylogenetic accuracy—natural selection, recombination, and deviations from the model of evolution. By analyzing the substitutions that occurred and calculating selection coefficients for derived alleles throughout their evolutionary trajectories to fixation, I show that molecular evolution in these experiments is complex and proceeding largely as would be expected for biological populations. Using these data to construct empirical substitution models, I demonstrate that phylogenetic inference is incredibly robust to significant molecular evolution model deviations. I show that neutral evolution in the presence of always-occurring population processes, such as clonal or Hill-Robertson interference and lineage sorting, result in reduced clade support, and that selection and especially recombination, including their joint occurrence, restore this otherwise-reduced phylogenetic accuracy. Finally, this work demonstrates that inferred branch lengths are often quite inaccurate despite clade support being accurate. While phylogenetic inference methods performed relatively well in both theoretically facile and challenging molecular evolution scenarios, their accuracy in clade support might be a remarkable case of being right for misguided reasons, since branch length inference were largely inaccurate, and drastically different models of evolution made little difference. This work highlights the need for further research that evaluates phylogenetic methods under experimental conditions and suggests that digital evolution has a role here.

Finally, I examine student understanding of the importance of biological variation in the context of a course featuring a digital evolution lab. I first describe the Avida-ED lab curriculum and its fulfillment of calls for reform in education. Then I describe the specific education context and other course features that aim to address student conceptualization of variation. I present a modified published assessment on transformational and variational understanding and findings regarding student understanding of variation within an evolution education progression. Finally, I offer suggestions on incorporating course material to engage student understanding of variation.

To all the people that did or will affect the course of my light cone.

ACKNOWLEDGMENTS

So many people helped me reach this point. I am tremendously grateful for their caring, guidance, assistance, and steady presence.

Foremost thanks are owed to my two primary advisors – Barry Williams and Jim Smith. Barry plucked my grad school application out of a pile and he quickly wooed my evolutionary biology interests and passions. He was a highly personable mentor during the first half of my grad school years and always encouraged me to chase my research questions. Jim started as a committee member, but my work with him especially grew through our efforts on the Avida-ED team. He agreed to become my advisor, advocate, and good friend across my remaining years of work. Jim's advice and support were paramount in finishing my degree. I will be forever grateful for all the time and energy he has invested in my cause. Thank you, Jim and Barry!

The initial research design and data collection for the work in Chapter 1 was conducted by Barry and a summer undergraduate research student in the lab, Kyle Safran. Carlos Anderson also advised us and his knowledge of Avida laid the foundation for everything I would learn about the system. Others taught me a great more about Avida—especially various members of Charles Ofria's Digital Evolution lab—and to all those folks, thank you! Charles was also a valuable committee member, as was Shin-Han Shiu, who each gave actionable feedback especially in regard to communicating my research to various scientific communities.

The initial research design and survey distribution for the work in Chapter 4 was conducted by the Avida-ED curriculum development and assessment team, including Jim, Rob Pennock, Louise Mead, Mike Wiser. These folks, in addition to others that have joined us at various points, and our work together was my entry into biology education research, which fundamentally shaped by career trajectory. The Avida-ED team also sponsored my involvement with BioQUEST and SABER, communities that have been supportive and pedagogically invigorating. Rob and Diane Blackwood, as the Avida-ED PI and lead programmer, respectfully, have been integral to the success of Avida-ED and all work stemming from its use. Mike worked

closely with me on the transformational-variational scoring rubric, but moreover since my early days in graduate school I have cherished his wit and have always welcomed his conversation. I owe a lot to Louise, especially. She was the primary curriculum developer for *Integrative Biology* and my mentor for our FAST Fellowship work. Special thanks to our undergraduate teaching assistants, my graduate teaching assistant, and our IBIO150 students! Louise, along with Barry and Terri McElhinny, was also my primary undergrad educator role model.

Numerous other individuals and organizations made my years at MSU run smoother. Among others, Tom Getty, David Foran—and of course, Jim—made special arrangements for my continuation at MSU after my advising handover; I am fortunate to have had such support. Many administrative folks in IBIO, EEBB, and BEACON expertly handled my queries or business concerns over the years, and their assistance cannot go unremarked, so thank you, again! CNS, BEACON, EEBB, IBIO, COGS, and the Grad School via the FAST Fellowship provided research, travel, and/or living funds; without these monies, I never would have been able to afford this education. I also made many friends through these organizations and our time together was enlightening and entertaining! Finally, Kevin Hall and our friend Yvette Green will always hold esteemed places in my memories of Giltner Hall.

Before my time at MSU, my education was especially impacted by my undergraduate advisor, Susan Kalisz, and my teaching mentors of her, Tony Bledsoe, and Laurel Roberts at Pitt. I also want to thank all my other teachers across thirty years of education, most notably Eric Baltz who first formally introduced me to evolutionary biology at Freedom High School.

My family across the branches of Henry's pedigree have and continue to provide me with love and support. My Mom was especially integral in imbuing me with a love of learning, my Dad, Craig, always encouraged me to be adventurous and try new things, and my Dad, Mike, helped foster my "Why?" question-asking. Last but far from least come my you, Katie, and our joy, Henry, to whom I say this: Now that Daddy has finished his dissertation edits, which were years in the making, he and Mommy will joyously celebrate your first birthday. The two of you, and the minds of my students, are my meaning-making in the face of the Absurd.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1:

The Case for Using Digital Evolution in Experimental Phylogenetics and Evolution Education

Experimental Evolution and Digital Evolution

For most of its history, evolutionary biology research was impeded by a perceived inability to observe, measure, and experiment over evolutionarily relevant stretches of time (Garland and Rose, 2009). Even Darwin, the originator of innumerable insightful research avenues and methods of study, failed to understand that evolution may be investigated experimentally. Although long-term experiments were occasionally proposed (de Varigny, 1892), this inability largely persisted for three quarters of the history of biology research to date, from Darwin's publication of *On the Origin of Species* until experimental evolution studies began in earnest (see Rose et al. 2004). Experimental evolution is the study of populations across generations and under defined conditions imposed by the researcher, and its primary goal is to directly test evolutionary theory (Kawecki et al., 2012). Using this approach, a few populations have been studied in nature, although laboratory investigation is generally preferable due to greater control over environmental conditions and the ancestral population. Aspects of the evolutionary process can be observed in a laboratory setting over relatively limited amounts of time, and with particular study systems, most often phage, bacteria, yeast, and *Drosophila*. Even then, it is very difficult to record observations at the level of detail one might like. Still, experimental evolution, "evolutionary biology in its most empirical guise" (Garland and Rose, 2009), has been a very productive endeavor (Kawecki et al., 2012).

The meaning of the term "artificial life" was suggested by Lenski (2001). The quality of being "artificial" is straightforward – that which is not of nature. Although definitions or theories of life have been notoriously difficult to construct and defend (e.g., Ruiz-Mirazo et al. 2004), Lenski (2001) describes life as that which can evolve via natural selection; so, not only

the property of self-reproduction but also heritable variation and the propensity for variation in a population to change due to the benefit or detriment it confers.

Digital evolution

Digital evolution was inspired by early computer viruses, code that can reproduce although not evolve. Once computer scientists wrote code that could evolve they "domesticated" these programs in a controlled digital environment (Wilke and Adami, 2002), creating a form of artificial life. These self-reproducing computer programs, or digital organisms, differ from genetic algorithms and computational (also called numerical) simulations in that the organisms must, by themselves, reproduce and that no genotype is designated as the optimal, sought target by the experimenter. Natural selection occurs because the environment is computationally resource-limited and the reproduction process is designed to have random inaccuracies, or mutations, some of which may allow digital organisms to reproduce more efficiently, outcompeting other genotypes for resources. Complex computational metabolic processes in addition to self-reproduction may arise due to mutation, with organisms being able to perform computation using environmentally encountered numbers. These phenotypes can then evolve via selection if a suitable selective environment is provided – one that rewards such computation (Adami, 2006). It seems plausible that digital evolution systems exhibit the trait of open-ended evolution, as do biological systems (Lenski et al., 2015; Ruiz-Mirazo et al., 2004). Open-ended evolution is the capability of an evolving population to continually produce novel organisms rather than reaching a stable state. Although digital evolution's capability in this regard is an ongoing discussion among artificial life researchers (Taylor et al., 2016).

Digital organisms have the attributes sought for in experimental evolution model systems. Generation time is measured in seconds or less, population sizes can be massive, measurements can be taken with heretofore unprecedented ease and precision, and digital organisms readily tolerate human-influenced environments. Further, experiments can be highly controlled, easily replicated, and even identically repeated. Of course, the necessary and

sufficient computational hardware is needed, although these requirements are minimal with respect to modern machinery. Even "impossible evolutionary experiments" can be performed due to an experimenter having full control over the genetic and environmental conditions, e.g. disallowing all neutral and deleterious mutations, allowing biologists the ability to evaluate otherwise untestable ideas (O'Neill, 2003).

Avidians in Avida and Avida-ED

Avida is an artificial life platform designed to study broad questions in evolutionary biology via the evolution of digital organisms, called Avidians (Ofria et al., 2009; Ofria and Wilke, 2004). This highly manageable model system allows quick replicate experimentation and copious data output, leading to high-impact research regarding the nature of evolutionary processes (e.g., Lenski et al. 1999, 2003; Wilke et al. 2001; Chow et al. 2004; Goldsby et al. 2012; Covert et al. 2013). Avida-ED is the educational version of this research platform (Pennock, 2007a). Through its approachable graphical interface, simplified set of configurable experimental variables and output, and associated curriculum, Avida-ED allows students to draw connections between evolutionary processes operating in biological and digital systems, ask questions and conduct research involving biological theory, and engage in science and engineering practices in a similar manner to biologists using Avida or other digital or biological model study systems (Kohn et al., 2018). Avida-ED has garnered an award from The International Society for Artificial Life (2017) and is itself the subject of ongoing education research (Speth et al., 2009; Smith et al., 2016; Lark et al., 2018). Each program is freely available: Avida at http://avida.devosoft.org/ & https://github.com/devosoft/avida and Avida-ED at https://avida-ed.msu.edu/.

Avidians undergo computational metabolic processes to self-reproduce. An Avidian is a computer program consisting of a sequence of simple, modular computer instructions, the set of which constitutes its genome. The instruction set consists of 26 instructions and is Turing complete, which means that in principle any computer program could be encoded within the Avidian genomic language. An Avidian's genome encodes its ability to self-reproduce and

perhaps perform other computational tasks. Digital evolution experiments often begin with an organism capable of reproduction and nothing else.

During reproduction, each time a parental instruction is copied there is a chance the offspring will incorporate a different instruction at that position of its genome. This change occurs via random probabilistic means and is analogous to substitution point mutations in biological life. Although genome size is fixed in Avida-ED, the research version additionally allows other configurable genetic variables such as other types of mutation including insertions and deletions, as well as other instruction sets and even recombination. Mutations and, with sexual organisms, recombination result in the accumulation of genetic variation in a population and might allow an Avidian's genome to code for other features. The digital environment in which a population of Avidians exists is a grid-like lattice in which a single organism occupies a single space in the grid. By configuring this environment grid the researcher sets the maximum population size. Environments can be configured such that the performance of specific computational functions, most commonly bitwise Boolean logic tasks, are rewarded. This reward is in the form of additional computational resources such that the Avidian can execute its code quicker relative to others in the population, resulting in faster offspring production. An individual's fitness is measured as a function of its reproduction efficiency and ability to perform tasks. Organismal, population, and environmental data can be saved to track the evolutionary course of a population, and past experiments can be identically replicated through random number seed specification.

Avidian evolution as an instantiation of evolution

Avidian evolution results in evolutionary mechanisms and thus outcomes that are analogous to biological reality. Because the necessary and sufficient conditions for evolution—inheritance, variation, and differential reproduction (Dennet, 1995)—are inherent to the system, evolution is not simulated but rather actually occurs. Albeit digital, Avidian population change is an instantiation of evolution and neither a summary of established evolutionary patterns nor a simulation thereof (O'Neill, 2003; Pennock, 2007b). Users do not program what

will happen, but they can adjust initial genetic and environmental conditions and, after initiating the experiment, they can then observe and record what happens. Rather than being a model of evolution, Avidians undergo evolution; though, Avidian genetic and environmental complexity can be considered a model of that which exists in nature. Within their complex computational environments, Avidian adaptation often proceeds in creative ways the experimenter never would have predicted.

Because Avidians have a genomic sequence and exist in populations undergoing evolutionary change, many evolutionary processes or outcomes can be studied through experimentation. For molecular evolution, these have included epistasis and complexity (Adami et al., 2000; LaBar and Adami, 2016; Lenski et al., 1999a, 2003; Ofria et al., 2008; Ostrowski et al., 2015; Strelioff et al., 2010); genotype-phenotype mapping (Fortuna et al., 2017); phenotypic plasticity (Clune et al., 2007); genome size evolution (Gupta et al., 2016; Ofria et al., 2003); mutation rate evolution (Clune et al., 2008); mutational and drift robustness (de Visser et al., 2003; Elena et al., 2007; LaBar and Adami, 2017; Lenski et al., 2006; Wilke et al., 2001); clonal and Hill-Robertson interference (Adami, 2006; Covert et al., 2013; Ostrowski et al., 2007; Wilke and Adami, 2002); and Muller's ratchet, mutational meltdown, and the effects of recombination (Misevic et al., 2010, 2006, 2004). For ecology and macroevolution, these have included group selection (Beckmann et al., 2008; Clune et al., 2011; Goings et al., 2004; Goldsby et al., 2012, 2014a, 2014b; Knoester et al., 2007b); coevolution (Zaman et al., 2014); behavior, communication, and cooperation (Elsberry et al., 2009; Goldsby et al., 2008; Goldsby and Cheng, 2008; Grabowski et al., 2013; Knoester et al., 2013, 2007a; McKinley et al., 2008); ecological specialization, maintenance, and extinction (Chow et al., 2004; Connelly et al., 2010; Cooper and Ofria, 2003; Fortuna et al., 2013; Ostrowski et al., 2007; Yedid et al., 2009); historical contingency (Yedid et al., 2008); and even the origin of life (CG et al., 2017). Thus, there exist similarities between digital and biological organisms with respect to a remarkable array of evolutionary phenomena. In fact, it has been argued that "in terms of the complexity of their evolutionary dynamics, digital organisms can be compared with biochemical viruses and

bacteria" (Wilke and Adami, 2002). Accordingly, with Avida-ED students can perform experiments and collect actual research data amenable to hypothesis testing, learning about biology and practicing as scientists throughout their experience. Expert and novice scientists can study the power of evolution not just within the digital realm of Avida-ED, but also by analogy to the chemical and physical reality of biological life.

Experimental Evolution in Phylogenetics

The inference of historical relationships, i.e., phylogenetics, is a central goal in biology (Hillis, 1995). Phylogenies, representations of evolutionary relatedness among organisms or taxa generally, are usually inferred from molecular sequence data, although behavioral, morphological, and other characters can also be used. This inference process is crucially important because phylogenies are created for use across all of biology to support myriad research efforts. Our ability to infer phylogenies has consistently improved, or so we think, through the advancement in molecular evolution theory and modeling, the acquisition of molecular sequence data, and the implementation of sophisticated algorithms and computational tools.

Evaluating accuracy in phylogenetics

Evolutionary histories of appreciable degrees of evolutionary change cannot generally be observed, with few notable exceptions; thus, phylogenetic inference cannot be definitively tested, and the evaluation of phylogenetics methodologies and tools has largely relied on computational simulations (Hillis, 1995; Huang et al., 2017). Computer simulations are tools for algorithmically modeling molecular evolution under a set of assumptions. By necessity, simulations rely on relatively simple models of evolution (Arenas, 2012; Huelsenbeck, 1995). Basic simulations of molecular evolution generate probabilistic changes in characters such as nucleotides or amino acids according to a model of substitution rates between residues (Bull et al., 1993). More advanced simulations may incorporate the effects of the genetic code via codon evolution (Rambaut and Grass, 1997) or non-independence among characters

(Huelsenbeck and Nielsen, 1999). Even more advanced methods include coalescent approaches (Huang et al., 2010) or attempt to incorporate selection and linkage (Messer, 2013). Molecular evolution models can also aid, for example, in the simulation and analysis of tree topologies (Graybeal, 1998) and in relative and absolute rates of evolution (Kolaczkowski and Thornton, 2008) and speciation or extinction (Rabosky and Lovette, 2008). Simulations do provide valuable insight regarding a range of conditions that theory predicts are relevant to phylogenetic inference. They can be used to generate large amounts of data with relative ease, even providing exhaustive information within their defined parameters (Huelsenbeck, 1995). Overall, simulations are ideal for investigating the specific dynamic for which they are programmed.

While simulations have become increasingly sophisticated, the idealized conditions they model might, alone or in part, never truly exist in nature. Simulations are less useful in the analysis of combinations of complex factors, and when emergent or unknown properties are present in complex systems (Arenas, 2012). Simulations are limited in that they do not incorporate the full range of conditions operating in evolving populations, including those that we know to have potential in disrupting phylogenetic inference, those which we suspect might, and those which we have not yet discovered. For example, even the molecular evolution of coevolving sites remains very difficult to simulate (Arenas, 2012; Sousa et al., 2008). Simulations, as with all models, incorporate untested assumptions and fail to incorporate many other factors (Hillis, 1995). This gap in the credibility of simulations will always remain (Miyamoto and Cracraft, 1991) and it is unknown whether the assumptions and necessary simplifications in simulations reduce their relevance (Hillis et al. 1993). Thus, our understanding of the accuracy of phylogenetic methodologies is limited by a reliance on the evolutionary models implemented in simulations (Hillis et al., 1994). When evaluating phylogenetic methods using a simulation that makes explicit assumptions identical, or nearly so, to the assumptions in one of the methods (e.g., Jin and Nei 1990), then that method will prove superior – an observation that cannot be universalized to the broader range of conditions existing in nature

(Hillis, 1995). This makes it especially challenging to evaluate the robustness of inference methods using simulations. The only way to evaluate whether phylogenetic inference methodologies are sufficiently robust to these complexities is to evaluate their predictions using empirical, known evolutionary histories (Hillis, 1995). This alternative and complementary but rarely used approach—experimental phylogenetics—is the analysis of data from natural or experimental populations with a known evolutionary history.

The objective of experimental phylogenetics research is to use living systems to generate known evolutionary histories with which systematists can use to directly test phylogenetic methods (Bull et al., 1993). In contrast to approaches using computational simulations, experimental phylogenetics studies make substantially fewer untested assumptions regarding the evolutionary process. The result is an expectation that the evolutionary system incorporates a degree of complexity and reality otherwise unobtainable in computational simulations (Bull et al., 1993; Oakley, 2009). In fact, an aim of experimental phylogenetics research, and of experimental molecular evolution broadly, is the iterative creation of increasingly sophisticated models based on empirical data which can then be incorporated into theory and practice through simulations (Bull et al., 1993). Experimental phylogenetics "is not a substitute" for simulations, but rather complementary to their use (Hillis et al., 1992; Huelsenbeck, 1995). Researchers following the paradigm of experimental phylogenetics stress that the particulars of their study system or precise conclusions are not necessarily universally applicable in nature. Instead, an experimental study provides information about the evolution that actually occurred – instead of proposing what should happen with natural populations, it establishes what did happen with a set of evolving populations. All in all, experimental phylogenies are "a step closer to reality" (Hillis and Bull, 1993). This notion is illustrated by Hillis et al. (1993) using an analogy between weapons testing and phylogenetic inference testing:

> "The difference between experimental and simulated phylogenies is like the difference between experimental and simulated bombs: the explosion of an

experimental bomb does not indicate what will happen every time a bomb explodes, but it does provide information on one actual explosion."

History of experimental phylogenetics studies

The earliest studies directly evaluating phylogenetic histories were conducted with common research species, including animals (mice, Fitch and Atchley 1985), plants (oats, Baum et al. 1984), and viruses (T7 bacteriophage, Hillis et al. 1992). The first two concerned the results of artificial selection and as such were limited in that the populations experienced relatively minimal evolutionary change even over decades or centuries, timespans of great length to humans but hardly of note with respect to each taxon's rate of evolutionary change. Further, the histories were incompletely known. The work of Hillis et al. (1992) was revolutionary for the field of experimental phylogenetics in that theirs was the first "completely known" phylogeny, having been produced through careful laboratory experimentation for the purpose of testing phylogenetic methodologies.

Research using bacteriophage T7

The goals of Hillis et al. (1992) were twofold – to "establish the feasibility" of producing an experimental phylogeny using a system that could undergo considerable evolutionary change and to use the resulting phylogeny to test various methods of phylogenetic inference (Hillis et al., 1993). Because they knew the true evolutionary history in the lab, they could evaluate the accuracy of phylogenetic inference. Further, they argued that the performance of an inference method with an experimental phylogeny provides support on its performance with "other (natural) phylogenies" (Hillis et al., 1993). The primary results of Hillis et al. (1992) were that their rooted eight-taxon symmetric and approximately ultrametric tree topology was correctly inferred from restriction site data by parsimony in addition to neighbor-joining (NJ) and three other distance methods. On the other hand, branch lengths were incorrectly inferred by all methods, although parsimony was the least inaccurate of the five methods. Parsimony was the only method of those tested that can infer ancestral character states for internal nodes, and it did so with very high accuracy (98.6%). In a subsequent publication, Hillis et al.

(1994) reported analyses using sequenced DNA regions of the viral genomes. The nucleotide dataset had approximately one-third as many variable sites than that of the restriction site dataset. In a comparison, of maximum likelihood, parsimony, NJ, and two other distance methods, only parsimony inferred the correct tree topology. The researchers then controlled for the proportion of variant sites by creating bootstrapped samples of restriction site and nucleotide datasets. Overall, the restriction site data still outperformed the nucleotide sequence data for each analysis method except maximum likelihood, with parsimony being the most accurate among nucleotide-inferred trees and NJ performing best among restriction site analyses. The researchers attributed the better performance of restriction site data to its presumed greater independence among characters than in nucleotide sequences.

This phage study (Hillis et al., 1992) has received various criticism. Sober (1993) complained that it lacked a sufficient discussion regarding how the model of the evolutionary process used in the laboratory was similar to that found in nature. In response, Hillis et al. (1993) acknowledged that their system's molecular evolution, while not representative of most systems in nature, remained within the range of known processes occurring in nature. They concluded that their system "does not appear to be less representative than would any other single taxon chosen for study." Such a detailed description of the phage's molecular evolution over the course of this experiment was later expanded upon by the original researchers (Bull et al., 1993) and others (Oakley and Cunningham, 2000). Sober (1993) also suggested that the primary results of parsimony and distance methods agreeing on the true topology, might have arisen from the combination of uniform rates of evolution and equal branch durations in the evolutionary history, which would be sufficient to guarantee statistical consistency between these methods. He even suggested that their result of agreement between models might be evidence that the evolutionary processes operating in the laboratory environment are significantly different than in nature, since datasets from nature rarely produce agreement among phylogenetic inference methods. Hillis et al. (1993) responded that it is neither necessary nor sufficient for establishing the accuracy of an inference by showing agreement

among methods, as had been notably shown with the case of long branch attraction (Felsenstein, 1978). Hillis et al. (1994) advocated that a range of topologies and experimental conditions should be explored in future experimental phylogenetics studies, since no single set of conditions is representative of nature. Sober (1993) concluded that experimental phylogenetics studies may suggest the types of evolutionary processes that allow or inhibit methods from inferring the true history, however "it remains to be seen what experimental phylogenetics can teach us about the problem of phylogenetic inference." Since simulations cannot demonstrate that nature obeys its assumptions, a combination of simulation and empirical approaches are necessary to improve phylogenetics (Hillis, 1995; Hillis et al., 1994).

In follow-up studies in which these viral data (Hillis et al., 1992) were directly compared with simulated data, the general concordance between these investigatory approaches was demonstrated (Bull et al., 1993; Hillis and Bull, 1993; Oakley and Cunningham, 2000). By creating simulated datasets varying in topology, branch lengths, mutation rates, and number of characters, Hillis and Bull (1993) investigated the approximate range of conditions within which their viral populations evolved. This work provided the still cited benchmark of 70% for bootstrap support values as indicating true clade relationships, and therefore demonstrating that bootstrap values are conservative measures of phylogenetic accuracy (Sleator, 2011). Similarly, parametric bootstrapping was used to create simulated data modeled using detailed conversion and reversion rate estimates from their viral system (Bull et al., 1993). The primary results were that parsimony, NJ, and a second distance method inferred the correct tree with consistent success, though NJ did so the most often; and, as with their empirical data, no method produced accurate branch lengths, though all were close, and parsimony performed the best. Oakley and Cunningham (2000) further analyzed the evolved viral taxa to evaluate ancestor reconstruction methods under models of continuous phenotypic characters. The researchers quantified different measures for growth rate for terminal taxa and ancestral taxa at each bifurcation, and the terminal taxa were used to reconstruct the ancestral character values for each node. They found that inferred ancestral states were grossly inaccurate, even

when the known ancestor sequence was used to root the tree. This was due to egregious homoplasy in virulence, with convergent decreases due to selection. Computer simulations of continuous characters were consistent with these results.

Sousa et al. (2008) used the same viral system and similar methodology as Hillis et al. (1992) to evaluate an asymmetric topology with considerable branch length variation. A fourteen-ingroup taxon evolutionary history was created, with between 3 to 29 lytic cycles (a measure of generation lapse) occurring between bifurcations. Phylogenetic inference was conducted using Bayesian inference, minimum evolution, maximum likelihood, and the five methods used by Hillis et al. (1992), and with their same set of restriction enzymes in addition to DNA sequences constituting greater genomic coverage than Hillis et al. (1994). For the nucleotide data, methods assuming or enforcing a molecular clock model inferred the correct tree, and other methods produced trees with topological accuracy (i.e., the average of clade accuracy and clade resolvability) of 82%. This superiority of clock-based methods was attributed to a strict experimental bottlenecking regime thought to produce a constant rate of change due to genetic drift. For restriction site data, the distance methods inferred the correct tree, while criterion-based methods were at best 91% accurate. For both types of data, the presence of polytomies was the primary culprit for many, although not all, instances of inaccuracy. Unlike in Hillis et al. (1994), the superiority of restriction site data was not due to greater numbers of variable sites in the datasets, and Sousa et al. (2008) agree that the greater independence among restriction site characters may have contributed to its improvement over DNA sequence data. They additionally attribute the poor performance of their DNA data to the still-low genomic coverage of 12%.

Research on the effects of natural selection in other systems

Molecular convergence, its prevalence due to natural selection and its effect on phylogenetic inference, was evaluated by several research groups. This work was conducted with organisms evolving in nature (Leitner et al., 1996), in Petri dishes (Bull et al., 1997;

Cunningham et al., 1997; Fares et al., 1998), and in digital environments (Hagstrom et al., 2004; Hang et al., 2007, 2003).

Leitner et al. (1996) used two HIV genes sequenced from nine individuals with well-documented epidemiological relationships. In HIV evolution, one of the genes sequenced is generally under strong positive selection for missense mutations, while the other generally has purifying selection against changes. Seven inference methods were tested as well as several models of evolution, and datasets were constructed using each gene separately and concatenated. Considered alone, the gene under greater positive selection produced more accurate clade inference than the gene under purifying selection, and convergent molecular evolution rarely occurred and therefore was inconsequential. The concatenated dataset performed even better still, yet the most accurate trees differed by at least one set of clades. While no combination of phylogenetic method, model, and dataset produced the correct tree, the larger the character set (i.e., using concatenated genes) the more accurate the inferred tree proved to be for many methodological combinations. This demonstrated that differences in these methods' abilities were due to algorithmic efficiency rather than consistency (Hillis, 1995), and thus with sufficient data NJ, maximum likelihood (ML), and parsimony would each perform well. Branch length estimates varied and were not very accurate, with short branches overestimated and long branches underestimated, although it is not clear how the authors compared the known chronogram to the phylogram inferences, as the rate of evolution was presumably variable among viral populations and over time.

Bull et al. (1997) adapted a bacteriophage, φX174, to infect two different hosts in a high temperature environment. Using the ancestor as the outgroup and a seven-taxon ingroup evolutionary history, two different maximum likelihood analyses were conducted, one with a five-taxon evolved lineage ingroup and the other with a nine-taxon evolved lineage ingroup, including two isolates embedded within the history. Both analyses failed to resolve the true history for these sets of taxa, because convergent evolution, consisting of both parallelisms and reversals, resulted in over half of the observed substitutions being phylogenetically misleading.

Using the laboratory protocol of Hillis et al. (1992), Cunningham et al. (1997) produced twelve separate lineages of T7 phage evolved from bifurcations of six lineages that had the same ancestor. The bifurcations were performed after either 10, 20, or 30 lytic cycles and each final lineage evolved through three series of bottlenecks separated by 50 lytic cycles each, with isolates stored at every bottleneck. Therefore, the twelve-taxon star chronogram had six variously short internal branches and twelve much longer external branches. They observed multiple instances of parallel evolution, including of deletions and nonsense mutations, which the researchers attributed to the selective environment. From these viral sequences, both terminal and embedded isolates, Cunningham et al. (1998) modularly assembled various four-taxon phylogenies that varied in branch lengths. Taxa were chosen such that non-sister lineages showed greater convergent evolution and had long external branches, with the sister lineage of each of these having a shortened external branch. This would be a severe test of long-branch attraction. Using maximum likelihood, they evaluated the effect the model of evolution had on phylogenetic inference by evaluating six progressively more complex models. With short-to-long branch ratios of 1:3 or less, model choice had little influence on phylogenetic accuracy, whereas with more extreme ratios, model difference was significant and best-fit models were more successful in resolving the long-branch attraction effect.

Fares et al. (1998) used yet another viral system, foot-and-mouth disease virus, to create a known evolutionary history. Using parsimony, maximum likelihood, and distance method analyses, they found that no method produced the true tree, which they too attributed to convergence due to selection.

Digital evolution using the Avida platform was used to test the effect of natural selection on parsimony, and occasionally NJ analyses, with symmetrical unrooted four-taxon tree topologies. Varying the extent of evolution occurring along internal branches (i.e., a single branch when unrooted) and the external branches, Hang et al. (2003) evaluated different experimentally-evolved branch length combinations in addition to simulated data. Branches were of equivalent evolved duration per level, e.g., with all external branch lineages allowed to

evolve for an equivalent length of time. The researchers hypothesized that phylogenetic inference would be improved through the production of synapomorphic variation due to selection during internal branch evolution, and that this benefit would be especially pronounced for longer branches, with greater evolution causing more such variation to occur. Their hypothesis was supported for parsimony analyses, with shorter internal branch topologies being no better inferred than data simulated under genetic drift alone, and longer branches yielding substantially improved accuracy. The researchers summarily characterized selection's effect on sequence evolution, finding three sets of loci: fixed, slightly variant, and highly variant sites. By including this three-tiered proportional model in their genetic drift simulation, they demonstrated that the perceived phylogenetic inference benefit of selection was restored in the simulated data. How these findings pertain to NJ or with trees with a range of branch length combinations is unclear. Hagstrom et al. (2004) used five different selective regimes to evaluate the relative benefit natural selection had for phylogenetic inference. The branch lengths were such that neutral evolution was expected to swamp all phylogenetically informative signal, and that only selection could cause accurate inference. Their results included that NJ performed well for many different selective regimes and that parsimony only performed well when natural selection occurred along the internal branches. Hang et al. (2007) detailed how natural selection rescues phylogenetic inference, reproducing that significant adaptation along internal branches can cause this, especially if selection is strong and maintained. This effect was caused by the production of synapomorphic variation rather than non-uniform character substitution, as expected.

Research summary

The overall conclusion to be drawn from the entire body of published experimental phylogenetics research (summarized above) is that various inference methods generally perform well, though with considerable and inconsistent variation among methods and across studies. Further, many methods seem to be robust to deviations from the assumptions underlying their evolutionary models—although not necessarily character independence—and

long-branch attraction can have a strong effect. The benefit or detriment of using characters evolved under natural selection remains unclear; natural selection in biological systems tends to interfere with phylogenetic inference by producing more homoplasy (although see Leitner et al. 1996), and natural selection in digital systems tends to aid phylogenetic inference by producing more synapomorphy. Most studies remain confined to a limited range of taxa and phylogenetic-difficulty sample spaces (e.g., topologies and branch lengths), are nearly completely deficient in replication, and have rarely evaluated modern systematics methods and tools.

Hillis et al. (1992) concluded that experimental phylogenetics studies "will fill an important void in the science of phylogenetic reconstruction." However, after an initial flurry of research, experimental phylogenetics never became the "wave of the future in phylogenetics" (Oakley, 2009). To explain why the field has remained small after more than twenty-five years, having offered "few novel insights" (Oakley, 2009), we must examine the difficulties intrinsic to this methodological approach.

Experimental Evolution in Undergraduate Education

Undergraduate biology courses rarely address the processes of evolution (Alters and Nelson, 2002), and even more rarely focus on examples of these processes in action via experimentation. Rather, courses have tended to focus on the results of evolutionary processes, especially adaptation, and the various evidences for evolution, for example morphological, molecular, or developmental similarity in characters among taxa (Mead and Mates, 2009). Although experimental evolution has been touted as the perfect "eye-opener to people for whom 'seeing is believe'" (Kawecki et al., 2012), it has only recently been included in some introductory biology textbooks and laboratory curricula, with many still lacking this topic coverage (Burmeister and Smith, 2016; Hillis, 2007). Biological evolution can be difficult if not impossible to demonstrate or explore experimentally in the classroom. Providing opportunities for students themselves to participate in experimental evolution lab exercises is challenging

and finding a system in which they can conduct student-centered inquiry-based experimental evolution research is much more challenging still.

A few activities that allow students to engage with experimental evolution laboratory exercises have recently been produced. As with experimental evolution research generally, these activities are limited to a few model systems amenable to such work, namely microorganisms and, rarely, insects. Unfortunately, most of these activities are even further limited in that only one or a select few biology topics can be presented with each system. Labs with bacteria have been created to study mutation and adaptation, specifically antibiotic resistance (Krist and Showsh, 2007; Petrie et al., 2005), or adaptive radiation, niche colonization, and relative fitness measurement (Green et al., 2011); yeast have been used to study social evolution and inclusive fitness theory (Agren et al., 2017), or the origin of multicellularity either over many generations of experimental evolution (Ratcliff et al., 2014), or its limited initial evolution via rotifer predation (Pentz et al., 2015); and insects have allowed the study of allele frequency change in *Drosophila* (Plunkett and Yampolsky, 2010) and sexual selection and operational sex ratios in bean beetles (Cotner and Hebert, 2016). Finally, several activities explore the timing of mutation with altered Luria-Delbruck fluctuation tests or similar protocols with either bacteria or yeast (Green and Bozzone, 2001; Handelsman et al., 1997, p. 19997; Robson and Burns, 2011; Smith et al., 2015). Nearly all rely on inauthentic biological investigations with limited opportunities to learn—and engage in—the practice of science, from asking questions and proposing hypotheses, through gathering and analyzing data, and ultimately synthesizing and presenting research conclusions. A sole exception, the bean beetle system presented by Cotner et al. (2016), allows student inquiry and independent investigation, although the range of experiments possible are barely classifiable as experimental evolution since only one or very few generations can be observed over many weeks.

These activities have several disadvantages, some of which are common with most laboratory curricula. These drawbacks include detailed laboratory protocols that must be closely followed with little room for error, expensive equipment or profuse disposable

resources, and the maintenance of precise environmental conditions. These concerns can be of particular importance when contamination or inequivalent growth environments would impede fitness calculations (Green et al., 2011). Other drawbacks owe to the difficulties in navigating utility, realism, and generality with study systems used for classroom experimental evolution research, as discussed previously. Among the most consequential is the lack of suitability these activities and model systems have for engaging in student-centered inquiry-based experimental evolution research. These concerns are ameliorated with the Avida-ED system and its curriculum that intentionally integrates science and engineering practices.

Balancing Utility, Realism, and Generality in Experimental Evolution

For either the evaluation of phylogenetic methodologies or the observation of evolution within a classroom, the complexity of the task requires a considered balancing between the pairwise interplay of utility, realism, and generality. Although my specificity in describing these tradeoffs is novel, researchers employing or critiquing experimental phylogenetics, especially Oakley (2009), have argued similarly. Using my formulation, Oakley's distilled critique is as follows: While experimentally generated histories maintain greater biological realism than simulations, this comes at the significant expense of reduced utility in the form of time required to produce an evolutionary history. He also stipulates that with both simulations and experimental phylogenetics it is necessary to assume that the operating evolutionary processes apply with generality to other systems. In the analysis presented here, "utility" encompasses experimental feasibility, resource cost (e.g., time and money), and other difficulties such as technical expertise, "realism" entails the known and unknown complexities of biological evolution, and "generality," also called universality, is the ability for inferences drawn from scientific data to be applicable, through induction, more broadly. I also include within generality the ability of a system to display open-ended evolution and be useful to explore many avenues of research. These three factors produce tradeoffs that must be navigated within both experimental phylogenetics research and classroom experimental evolution labs.

Experimental utility versus biological realism

The primary critique of experimental phylogenetics focuses on this tradeoff (Oakley, 2009). For example, Hillis et al. (1992) sought a system in which they could manipulate and observe long-term evolution with expected population genetic dynamics, e.g. relatively low mutation rates and molecular sequence divergence due to both natural selection and genetic drift. They opted for a system with a high mutation rate and employed stringent, repeated bottlenecks (Bull et al., 1993). Through this aim of increasing the proportion of phylogenetically informative genetic variation, while maintaining selection for viability, the researchers approximated the molecular evolution that they desired but, reasonably, did not wish to expend the resources, e.g., time and cost, to produce in a more biologically realistic manner. Even so, this or similar approaches can take months or years of laboratory effort to generate a single instance of a known phylogeny (Hillis, 1995). At the extreme, an example of reducing realism in an experimental system for the sake of utility is the harnessing of hypermutagenic polymerase chain reaction to evolve DNA sequences entirely separate from their organismal context (Randall et al., 2016; Sanson et al., 2002; Vartanian et al., 2001). As have others before me, I consider this to be in vitro simulation and not experimental phylogenetics, per se.

Arguably the most awe-inspiring example of classroom experimental evolution is the laboratory system presented by Ratcliff et al. (2014) in which yeast cultures are repeatedly transferred after allowing time for cells to settle at the bottom of a test tube; a protocol that reliably produces clusters of cells that can be analogized to the origin of multicellularity. This is a fantastic means by which students may observe evolution in action. Yet, as argued by Pentz et al. (2015), this experiment suffers from two significant flaws related to the utility-realism tradeoff: The selective agent is contrived – a similar agent is thought to have played no role in the repeated evolution of multicellularity. And, even so, the experiment requires substantial investment in time and resources, requiring two weeks of daily transfers and hundreds of sterile media tubes. Pentz et al. (2015) produced an alteration that navigates the utility-realism tradeoff differently by using a rotifer predator to present a more plausible selective agent for

multicellularity, and thus greater realism. This protocol does reduce time and resource use, although rotifers require more technical expertise to manage. Overall, the balance between utility and realism is extremely tough to manage in experimental evolution.

Experimental utility versus generality

An ideal that has not yet been widely adopted in experimental phylogenetics is the production of replicate datasets. Doing so could provide crucial insight regarding the likelihood of phylogenetic methods to produce correct inferences more generally. Additionally, doing so could "detect subtle differences" in the rate of success among methods (Bull et al., 1993). As argued by Bull et al. (1993), the "dilemma that faces experimental phylogenetics" is this very notion that when replication is lacking, the reliance on a few datasets reduces the statistical utility of empirical observations. Unfortunately, replication has been prohibitively difficult to accomplish when a single replicate is costly to produce. Out of eighteen publications regarding experimental phylogenetics, only five unique experimental phylogenies using biological organisms have been produced for the purposes of this research (Bull et al., 1997; Cunningham et al., 1997; Fares et al., 1998; Hillis et al., 1992; Sousa et al., 2008), plus one additional study using a convenience dataset of naturally evolving taxa (Leitner et al., 1996), three publications employing digital evolution to evolve known evolutionary histories (Hagstrom et al., 2004; Hang et al., 2007, 2003), and the remainder either reexamining these datasets or commenting on them.

Perhaps an even more stark instance of the utility-generality tradeoff is the lack of model systems amenable to experimental phylogenetics or classroom experimental evolution. Oakley (2009) observes that experimental systems will only have utility if they can produce evolutionary histories within months or less; and such a system would require life history traits outside the norm, including short generation times and rapid rates of evolution. The taxonomic diversity between studies in which known non-experimentally evolved phylogenies are evaluated is high, including mice (Fitch and Atchley, 1985; Sage et al., 1993), oats (Baum et al., 1984), and HIV (Hillis et al., 1994; Leitner et al., 1996). Yet the range of taxa used in

experimental phylogenetics is extremely limited – viruses and Avidians. As would be expected, the systems used in biology classrooms to observe evolution over many generations are of similarly limited taxonomic diversity, with bacteria (Green et al., 2011; Krist and Showsh, 2007; Petrie et al., 2005), yeast (Ratcliff et al., 2014), and Avidians (Speth et al., 2009). Of course, other systems have been used to observe one or a few generations of evolution in the classroom, e.g. bean beetles (Cotner and Hebert, 2016), or experimental evolution research other than experimental phylogenetics, e.g. *Drosophila* (Burke et al., 2010) and *Arabidopsis* (Scarcelli and Kover, 2009). This suggests that the number of systems suitable to experimental phylogenetics research or extended classroom experimental evolution is greatly limited for reasons of utility.

Biological realism versus generality

As noted by Bull et al (1993) for their viral system, "the experimental organism is not of special interest by itself, so the value of the study must rest on its generality to other systems." The researchers continue by acknowledging that "generalities are not immediately apparent precisely because of the incorporation of genetic detail" into the specific model of molecular evolution applicable with their system. "The irony is that, by increasing the level of molecular resolution, we have discovered features that render the experiment unique, hence less general" (Bull et al., 1993). Depending on the inference methodologies under examination, increased modeling of molecular evolution complexity may necessitate increased taxonomic specificity and thus reduced general applicability to other systems.

This realism-generality tradeoff is highlighted by classroom experimental evolution protocols focusing on individual systems with extremely limited diversity in content exploration. In fact, model lab systems in which students can engage in student-centered inquiry-based experimental evolution research using biological organisms is entirely lacking. Thus, these classroom labs maintain realism for specific biology content while near-completely losing the generality of content exploration. In the modified evolution of multicellularity lab (Pentz et al., 2015), the incorporation of a plausible selection story increases the biological

realism of this major transition in evolution. However, neither the initial variation in the tendency to form multicellular clumps nor generations of change occur within this exercise. In this case, not only can students not experiment using a fully open-ended evolution system capable of facilitating multiple avenues of research, but they cannot even observe evolution in action.

Digital evolution offers a complementary balance

By using digital evolution systems, these utility-realism-generality tradeoffs can be navigated differently. Doing so fills a void between computational simulations and experimental evolution with biological organisms (Ofria, 2015). The use of digital evolution in experimental phylogenetics research or in classroom experimental evolution entails greater biological realism than simulations, and greater utility and generality than biological systems. Importantly, none of these approaches serve as a substitute to the others.

Experimentation with digital evolution entails much greater ***utility*** compared to biological systems but less so than simulations. Otherwise impossible experiments can be conducted in digital systems; for example, in the work of Hagstrom et al. (2004), a condition of two of their selection regimes was only feasible in a digital evolution system – offspring with non-neutral mutations were artificially sterilized by the environment, such that only neutral evolution could occur in the population. Generating an evolutionary history is relatively quick and easy using a system like Avida. Experiments can take minutes (e.g., Chapter 4), hours (e.g., Chapter 2), or weeks (e.g., Chapter 3) compared to much less and much greater time, respectively, for simulations and biological life, with approximately proportional cost in terms of computational resources with respect to simulations and overall costing much less than biological experiments. With respect to Avida-ED, resource cost is minimal since the program, currently version 3.2, is free to use and adequately runs on nearly all web-enabled devices, including computers, tablets, and smart phones, although its operation is quicker depending on available computing resources. This makes the observation of curious phenomena possible within a single class session (e.g., Chapter 4). As is apparent when new instructors first

implement Avida-ED in their classes or when scientists conduct their first experiment in Avida, the degree of human expertise in the form of technical knowledge and skill required for digital evolution experiments is more akin to biological experiments than simulations. The latter requires toggling established parameters, while the former each require specialized knowledge of the organismal system and manipulation of the organism and its environment. (No comparisons are intended regarding the initial creation/ discovery/ modification of each approach.) Bull et al. (1993) conclude that "the experimental approach is labor-intensive […] and it is not feasible to generate empirical data with near the ease of computer simulations." However, these authors were unaware of digital evolution systems that make this possible.

Digital organisms, such as Avidians, experience complex evolutionary and population genetic processes found in nature, thus achieving a degree of biological *realism* greater than that of simulations. The evolution that takes place with digital organisms is complex (e.g., Lenski et al. 1999, 2003; Wilke et al. 2001; Chow et al. 2004; Goldsby et al. 2012; Covert et al. 2013), and more so than with simulations of molecular evolution. For example, bounds on the presence and relative extent of mutation, recombination, and selection can be imposed upon the system by adjusting its genetic and environmental conditions. Yet the distribution of mutation effects, the character of mutations and substitutions, and the prevalence of linkage (under recombination) will organically change during population evolution. As with biological organisms, epistasis and genetic drift will always occur for Avidians. The researcher has control over aspects regarding organismal population size, yet less so regarding effective population size, which is a byproduct of the system. Other evolutionary factors such as migration and mate choice can be experimentally manipulated as well. Yet Avidian genetic and environmental substrates are much less complex than with biological life. For example, there is a lack of genetic modularity (e.g., chromosomes and well-defined genes), genetic expression and levels of information storage and use (e.g., transcription and translation), and intricately complicated environments (e.g., nuanced chemical and physical interactions). Aspects related to all the above can be readily measured and compared to biological ranges with the goal of attaining

greater biological realism (e.g., Chapter 3). Because the population of Avidians undergoes evolutionary change, experiments performed using Avida-ED result in actual research data amenable to hypothesis testing, allowing students to learn and engage in science practices (e.g., Chapter 4). For example, students can model their experiments after noteworthy biological counterparts, and compare their observations and evidence-based reasoning to that of practicing scientists.

While true with the study of any model system, the ***generality*** of research involving digital organisms must be approached carefully when designing experiments and drawing conclusions. The applicability of research produced using digital evolution systems can be case-specific and in large part owes to the degree of biological realism required. Fortunately, this can be measured with great utility because digital evolution approaches can generate vast quantities of data regarding the process and history of molecular evolution (e.g., Chapter 3). Similar data can be very difficult to produce in biological systems, if possible at all (Hang et al., 2003). For example, one can readily observe millions of generations of evolution while accurately recording copious measurements and maintaining the entire series of genealogical relationships and ancestral organisms, enough disk space withstanding. Students using Avida-ED have an opportunity to participate in the generation of large datasets from which they must extract salient information, a task common to many modern biological datasets. Classroom discussion is necessary to both introduce Avidians and explore how research using this system is applicable to biological life. An analogy that helps students at least initially understand Avidians is a direct comparison to bacteria (Johnson and Lark, 2018), although further discussion should treat Avidians as being their own system that shares generalities with all life, as with any model study system. Whereas replication has been prohibitively costly with biological systems, computational systems such as digital evolution and simulations alike are highly amenable to replication. The fact that Avidian evolution is an instantiation of evolution means that there is a richness and flexibility in the platform that cannot be matched by any tool

designed for a specific purpose. Avidians constitute an ideal study system to observe and experiment upon evolution in action in an open manner.

CHAPTER 2:

Digital Evolution Provides Direct Tests of Phylogenetic Accuracy

Introduction

Phylogenetics, the inference of evolutionary relationships, is a central goal in biology (Hillis, 1995). The accuracy of this inference process is crucially important because phylogenies are used to support research throughout all of biology. While phylogenetic methodologies have flourished, their accuracy has remained difficult to test since we lack true evolutionary histories observed in nature to compare to those inferred by phylogenetic methods. Therefore, computational simulations have been the primary means for evaluation of such methods (Hillis, 1995; Huang et al., 2017). While simulations provide insight regarding a range of theoretically important conditions, they are less suitable for evaluating combinations of complex factors and cannot address emergent or unknown properties of complex evolving systems (Arenas, 2012). The generation of known evolutionary histories—experimental phylogenetics—is a complimentary approach.

The first "completely known" phylogeny and follow-up research by David Hillis, James Bull, and colleagues in the early 1990s was revolutionary. By presenting a complete T7 bacteriophage evolutionary history, various molecular data sets, and phylogenetic analyses, this work inaugurated the field of experimental phylogenetics. First, Hillis et al. (1992) produced an experimental history in a carefully-controlled laboratory environment and evaluated various methods of phylogenetic inference. Then, other types of molecular evolution data from this experiment were used to investigate the accuracy of bootstrap support values (Hillis and Bull, 1993), characterize in detail the DNA sequence evolution (Bull et al., 1993), and evaluate clade resolution between various phylogenetic methods (Hillis et al., 1994), among other analyses. Together, this experiment and subsequent analyses stand as the foundational and most-impactful work produced in experimental phylogenetics, especially since this field has largely remained dormant for at least the past decade.

The experimental approach to evaluating phylogenetic methodologies has suffered from an imbalance of utility, realism, and generality compared with computational simulations. As argued in a critique by Oakley (2009), the greater biological realism found in experimentally generated evolutionary histories comes at the significant cost of reduced utility via time and resources and low generality, or universal applicability, to other systems. This critique seems to have been well-placed, as evidenced by a complete lack of published experimental phylogenetics research since the publication of Oakley's critique. In Chapter 1, I have argued that digital evolution provides an approach that navigates the tradeoffs among utility, realism, and generality differently, filling a void between computational simulations and experimental evolution with biological organisms. Specifically, digital evolution entails greater biological realism than simulations, and greater utility and generality than biological systems.

Here I present research to test the hypothesis that digital evolution using Avida is an effective model system for experimental phylogenetics research. I first evaluate the concordance of results from Avidian digital evolution treatments designed to reproduce the basic molecular evolutionary dynamics and phylogenetic inferences of the work of Hillis and colleagues using the T7 system. The resulting correspondence between these systems under comparable conditions is evidence that Avida is a satisfactory system for experimental phylogenetics. I then further demonstrate that digital evolution entails greater utility and generality than biological systems by presenting digital evolution research under a range of theoretical phylogenetically-challenging scenarios, the scope of which the T7 work did not address. Across 21 additional Avidian experimental evolution treatments, the effects of natural selection, recombination, and differing extents of lineage evolution within an evolutionary history are investigated. The results of these treatments were hypothesized to correspond with predications based on phylogenetics theory and prior research using computational systems. The completion of such a series of treatments is evidence of the greater utility and generality of digital evolution, because doing so with a biological system would be much more costly if not altogether impossible.

Taken together, concordance with a biological system's experimental phylogenetics results under similar conditions and a further demonstration of the experimental possibility of digital evolution with the correspondence of results with theory and computational systems constitute evidence that digital evolution using Avida is suitable for experimental phylogenetics research.

Phylogenetic analysis methods

The aim of phylogenetics is to produce hypotheses of evolutionary relatedness subject to falsification and/or statistical measure, and has progressed from distance-based methods to criterion and model-based methods (Felsenstein, 2004; Nei and Kumar, 2000; Yang, 2006; Yang and Rannala, 2012). An understanding of the philosophies and limitations underlying these methodologies is important for considering what makes an evolutionary history theoretically facile versus challenging to infer.

Distance-based methods use relatively simple algorithms to construct a phylogeny from a dataset based on overall similarity. With the need to analyze huge amounts of data, especially of molecular sequences, computationally efficient algorithms such as the phenetic distance-based algorithms of UPGMA (unweighted pair group method with arithmetic averages, Sokal and Michener 1958) and, especially, NJ (neighbor-joining, Saitou and Nei 1987) were adopted in the late 20th century. NJ is a clustering method that uses its highly efficient algorithm to produce a single result – the NJ tree. This method is purported to produce a very good approximation; in fact, if the distance matrix is an exact reflection of the true tree then NJ is guaranteed to determine it. Importantly, the algorithm is efficient enough to easily be manageable for hundreds or more taxa. For these reasons, NJ is very useful for analyzing large datasets that have a low degree of sequence divergence. However, a NJ tree is often not relied on to be a good hypothesis of an evolutionary history, since it is an approximation whose result cannot be directly compared to other trees within its framework (Felsenstein, 2004). Instead, more sophisticated approaches rely on the optimization of a statistical criterion.

Maximum parsimony (MP, Swofford 1998), maximum likelihood (ML, Felsenstein 1981), and Bayesian inference (BI, Huelsenbeck et al. 2001) rely on the comparison of phylogenetic trees using an optimality criterion. An optimality criterion is a characteristic upon which a comparison can be made, and phylogenetic inference is therefore an optimization search among the evolutionary hypotheses evaluated. These methods are generally superior to distance methods because instead of collapsing character state data into a single difference value, thus discarding substantial information, they preserve all available information by comparing sequences in the alignment, considering each site (i.e., character) at a time. However, criterion-based methods are not as computationally efficient as distance-based methods, especially because they rely on evaluating the optimality of all possible trees. This is generally cost-prohibitive, so a heuristic search algorithm is used to search within the space of all possible tree topologies.

The optimality criterion under MP is the parsimony score (Swofford, 1998). This is the total number of character-state changes necessary for a phylogenetic inference to explain the observed taxa-character dataset. This criterion provides a philosophically justifiable approach; homology should be assumed *a priori* and the hypothesis that requires the fewest *ad hoc* assumptions is the most preferable. This approach resulted in the distinction between apomorphy and plesiomorphy, and thus informative synapomorphies versus symplesiomorphies, and finally homology from homoplasy. If we are to deem phylogenetics as valuable, then we should want to maximize its utility. MP was shown to not have maximal utility (Felsenstein, 1978). When the amount of data gets larger, a statistically consistent method should converge on the correct answer, and MP does not have this property under certain circumstances. Model-based methods challenge MP regarding statistical justification, since it has been shown that a MP result is an approximation of a ML result only if the rates of change (i.e., branch lengths) are sufficiently small (Felsenstein, 2004).

Statistical phylogenetic inference methods, including ML and BI, use a probabilistic model of evolution to produce robust inferences. These models vary in their level of

parameterization with respect to differences in character state frequency and change. For example, with models of nucleotide evolution successively more complex, all state changes might have equal rates—the one-parameter JC model (Jukes and Cantor, 1969), unequal base frequencies—the four-parameter F81 model (Felsenstein, 1981), different transition and transversion rates—the five-parameter HKY model (Hasegawa et al., 1985), and unique rates between all character states—the ten-parameter generalized-time-reversible, or GTR, model (Tavaré, 1986). Additionally, among other inclusions to the model of evolution, each of these can have one or two parameters of site-to-site rate heterogeneity modeled by allowing a proportion of invariable sites and/or by using the gamma distribution (Yang, 1994). While ML and BI differ on how and to what degree these parameters are estimated, the aim is to identify the set of parameters that best fit the entire model: the model of evolution, tree topology, and branch lengths. The value that summarizes this fit is the optimality criterion.

Maximum likelihood's optimality criterion is the likelihood value (Felsenstein, 1981), which is the probability of the data given the tree and model of evolution. It is calculated using the probabilities for character state changes among all possible ancestral reconstructions, with branch lengths and model of evolution parameters optimized, and by assuming both character and branch independence throughout the tree. Statistically, each tree topology is a model, and the parameters are the branch lengths and substitution parameters. Thus, under ML inference, very many statistical models are iteratively compared, and classical confidence intervals cannot be constructed (Beerenwinkel and Siebourg, 2012; Yang and Rannala, 2012). ML has the desirable statistical properties of being unbiased, consistent in that it approaches the true value with greater data analyzed, and efficient in that it has the smallest variance among unbiased estimates, though these properties might not hold under all circumstances, especially if the substitution rate model is inaccurate (Yang, 2006).

Bayesian inference's optimality criterion is the posterior probability (Huelsenbeck et al., 2001), the probability of the tree and model of evolution given the data. The posterior is directly related, through the Bayes Theorem, to the product of the likelihood calculation and

the prior probability of the tree and model of evolution, and inversely to the probability of the data. Therefore, ML only concerns the data information, and its results are interpreted with respect to the data only, while BI uses additional information for the calculation of the posterior, a result that is interpretable with respect to the evolutionary inference. An essential difference between ML and BI is the prior. The prior probability of the tree and model is a set of probability distributions that are subject to the researcher's *a priori* ideas drawn from other data sources (Baum and Smith, 2013) and/or attempts to be as uninformative or objective as possible so that the calculation of the posterior is strongly inferred from the likelihood (Beerenwinkel and Siebourg, 2012). In practice, as with many phylogenetic methodologies, generally the program developer's suggested default settings are used (Brown et al., 2010; Yang and Rannala, 2012). The probability of the data term is challenging, since it requires synthesizing across all possible trees, branch lengths, and model parameters. Markov chain Monte Carlo (MCMC) methods, specifically using the Metropolis algorithm, allow the sufficient sampling of the probability distribution as long as mixing between algorithmic chains has been sufficiently conducted. Together, these distributions result in a major deviation from the ML approach in that under BI the posterior of every parameter is a probability distribution, or credible interval. This results in much greater information than a single ML estimate for each model parameter. For example, the probability of a clade, also called the clade credibility, can be expressed as a point estimate or as a distribution within a probability range, and, theoretically, with a uniform prior the mode of the posterior is equivalent to the ML estimate (Beerenwinkel and Siebourg, 2012). BI maintains the statistical consistency and other sought properties of ML (Steel, 2013).

The BI posterior probability is what one expects from statistics, since it lets us directly compare the probabilities of hypotheses given data. The intrinsic problem with classical statistics, including ML, is that these methods produce statements about the probability of the data or the method for analyzing the data, for example the probability that identical analyses of data drawn from a statistical population will contain the true parameter value within a

confidence interval. BI produces statements about the probability of the parameter of interest, for example the credible existence of an evolutionary relationship by a posterior probability density interval. The former is a statement on a population of parameter value analyses; the latter is a probabilistic statement about the actual value of the parameter. This philosophical distinction is exemplified by the construction and interpretation of measures of support for clade relationships in phylogenetic analyses.

Clade support evaluation

Clade support or uncertainty in phylogenetic inference is often measured with nonparametric bootstrap support values and posterior probabilities, although these metrics are quite distinct. Under classical statistical approaches, e.g., using MP and ML, nonparametric bootstrapping is used to craft statements regarding the statistical estimate of the clade given the data. In bootstrapping, the characters of the data are sampled with replacement to generate bootstrap pseudo-samples for the taxa-character dataset. Each bootstrap pseudo-sample replicate is analyzed identically to the actual dataset, and this sampling and analysis process is repeated hundreds or thousands of times. The proportion of trees among the bootstrap analyses that contain a particular clade is the bootstrap support value for that clade. This is used to determine the relative influence of the characters in the data. If the original data sample accurately reflects the phylogenetic information of the taxa, then the bootstrap will reflect experimental repeatability. As such, a clade's bootstrap support value is a function of the likelihood of the phylogenetic data for that clade.

The bootstrap method has remained difficult for many to interpret (Yang and Rannala, 2012), and is often erroneously considered a measure of clade accuracy (Hillis and Bull, 1993), while the BI posterior probability is easily and directly interpretable. A posterior probability is a measure of what proportion of the analysis a region of the multidimensional parameter-space is sampled within the BI search process. The entire space constitutes all the information of the prior probabilities for the tree topology, branch lengths, and free parameters in the model of evolution as well as how that information relates to the observed data. The more often a

specific region of parameter space, for example a clade in tree space, is sampled, the greater the probability it is accurate. The difference between bootstrap support and posterior probability is the fundamental distinction regarding the analysis outcome – the probability of the data (ML) and the probability of the hypothesis (BI). As bootstrapping is a resampling of the data, MCMC is a resampling of the hypothesis. A bootstrap value is the clade's support with respect to the experimental repeatability of the data, while a posterior probability is the clade's support with respect to the possible clades containing those taxa.

A researcher's interpretation or valuation of clade uncertainty is demonstrated by their presentation of either fully resolved or partially resolved trees. Fully resolved trees include the single best tree resulting from an analysis, such as the NJ result or the tree with the greatest likelihood under ML, lowest parsimony score under MP, or greatest posterior probability under BI. The latter is also termed the maximum clade credibility tree, and although this tree is often fully resolved, it is not guaranteed to be. Nonparametric bootstrap replicates or clade probabilities can be used to construct a fully resolved tree that considers clade uncertainty using the majority rule extended (MRe) algorithm, which produces a so-called greedy consensus tree (Bryant, 2003). This approach constructs a tree by starting with the most-supported clades and successively adding non-conflicting clades in order of greatest support. A MRe tree is fully resolved in that it lacks polytomies, although it may include clades that have very low support albeit without conflict to other, often greatly supported, clades. A different approach, the standard majority rule (MR) consensus algorithm, uses a threshold value with clades that have insufficient support being collapsed into polytomies. Any threshold of 50% or greater may be used, and MR trees with greater thresholds, in addition to the MRe tree, are always resolutions of the 50% MR tree, such that they will not include conflicting clades but might contain additional resolved clades (Degnan et al., 2009).

Hillis and Bull (1993) provided the threshold of 70% bootstrap support as indicating accurate clades by demonstrating that bootstrap values are conservative measures of phylogenetic accuracy (i.e., accuracy greater than 95% when bootstrap support is greater than

70%), and this result has become a commonplace rule of thumb (e.g., Sleator 2011). In simulation studies, bootstrap values have been shown to be highly conservative estimates of clade accuracy, and while posterior probability values have been shown to be closer estimates, they may in fact be overly liberal estimates (Cummings et al., 2003; Wilcox et al., 2002), with values being approximately 100% so often as to merit suspicion (Yang and Rannala, 2012). In practice, a variety of threshold values are used, and trees presented often display bootstrap support values and/or posterior probabilities for every clade, or at least those deemed significant.

Recently-published research shows a diversity in threshold values and trees presented. A review of the 19 empirical phylogenetics analyses, pre-published online, for the August 2018 volume of *Molecular Phylogenetics and Evolution* presents the diversity of support thresholds currently used in the literature. For example, Psonis et al. (2018) used a node coloration scheme to present six categories of joint disagreement between BI and ML analyses: four colors represented a posterior probability of 100% and bootstrap values of either 100%, ≥ 90%, ≥ 70%, or ≤ 70%; a fifth color for posterior probability ≥ 95% and bootstrap value ≤ 70%; and a final category of posterior probability ≤ 95% and bootstrap value ≤ 70%. A different study (Liu et al., 2018) presented posterior probabilities ≥ 0.95 as strong support and ≥ 0.85 as moderate support. Kim et al. (2018) considered bootstrap values ≥ 75% as showing strong support and ≥ 50% as moderate support, whereas Zhang et al. (2018) presented bootstrap values ≥ 90% as having strong support, ≥ 70% as moderate support, and < 70% as weak support. The trend of posterior probabilities being much higher than bootstrap support values is widely exhibited with empirical work, as is disagreement among qualitative descriptions of bootstrap support thresholds. These 19 studies also present trees with a diversity of clade support: Single best trees include maximum likelihood or maximum clade credibility trees in addition to MRe consensus trees. Trees exhibiting polytomies or with annotations that clades should be considered as such include majority rule thresholds of 50%, 70%, 95%, and 99%.

Overall, measures of clade support, including both clade accuracy and resolvability, appear to be highly valued among researchers, although with different weighting and interpretation, and largely without ground truthing other than the 70% bootstrap support threshold provided by Hillis and Bull (1993).

Overview of T7 phage experimental phylogenetics research

Experimental evolutionary history

The viral growth environment and experimental methodology of Hillis et al. (1992) were designed to provide ideal conditions to create phylogenetically informative variation between lineages. Phage were grown with a chemical to increase the mutation rate, and repeated bottlenecking occurred throughout the within-branch (i.e., anagenic) evolution of each lineage. These factors increased the extent of evolutionary change that occurred during the experiment. Additionally, new lineages at divisions (i.e., cladogenesis) were seeded with single-cloned populations, therefore eliminating lineage sorting or coalescent-type variation by fixing variation segregating in the population. The mutagen, anagenic bottlenecking, and cladogenic bottlenecking presumably increased the phylogenetically informative variation by increasing inferred internal branch lengths. The experimental population sizes, relatively small for a virus, and repeated bottlenecking also decreased phylogenetically misleading variation by increasing the influence of genetic drift and thus decreasing the possibility of homoplasious evolution via natural selection with parallel or convergent evolution.

The phage were serially grown, divided, and transferred at periodic intervals in a predetermined manner to create known evolutionary relationships between the resulting nine viral populations. The eight terminal ingroup lineages experienced division at equivalent intervals to create a symmetrical, binary tree-like evolutionary history, and the outgroup lineage evolved for nearly an equivalent total length of time, 105 lytic cycles compared to 120 total cycles for each ingroup (Bull et al., 1993). This desired ultrametric topology would be predicted to have exhibited uniform and constant rates of evolution, with all ingroup lineages experiencing the same mutagenic growth environment and having equal duration, and

therefore an equivalent likelihood of change (Sober, 1993). However this was not guaranteed to have occurred because the researchers fixed the extent of time between cladogenesis, rather than the degree of evolutionary change, which itself was a natural product of the evolving system (Hillis et al., 1993).

This ultrametric, symmetrical eight-ingroup taxon tree shape (i.e., topology) was chosen due to its predicted ease for phylogenetic inference, and with the intention that it could be used as a null model or best-case scenario for comparison with other topologies. If this best-case scenario viral growth environment and topology failed to produce data that resolved correctly then there would not be anything gained from attempting to evolve organisms in more realistic environments using topologies predicted to be more phylogenetically difficult (Bull et al., 1993; Hillis et al., 1993).

Phylogenetic and molecular analyses

The resulting evolutionary history was then used to evaluate phylogenetic methods. Hillis et al. (1992) used restriction-site mapping for 34 restriction enzymes to create a taxon-character dataset containing 202 characters, excluding sites invariant across all taxa. These data were used to infer and then compare the actual history with inferences produced using five phylogenetic methods. These five methods included MP, along with UPGMA, NJ, and two other distance methods, Fitch-Margoliash (Fitch and Margoliash, 1967) and Cavalli-Sforza (Cavalli-Sforza and Edwards, 1967). For this tree of eight ingroup and one outgroup taxa there are 135,135 possible rooted bifurcating tree topologies (Felsenstein, 2004), so a correct tree inference was unlikely to occur by chance alone. All five methods correctly inferred relationships among taxa. The methods varied in their prediction of branch lengths, and while MP performed best, no method inferred the correct branch lengths. The amount of homoplasy found in their dataset was approximately equivalent to levels found in empirical studies also involving nine taxa (Hillis et al., 1993).

As a follow-up study, Hillis and Bull (1993) evaluated how nonparametric bootstrap proportions compare to clade accuracy. To create very many pseudo-replicate datasets that

resemble the restriction site data of Hillis et al. (1992), they used jackknifing (i.e., sampling without replacement) to produce 500 subsamples of 50 characters each. Each of these were then bootstrapped (i.e., sampled with replacement) for 100 replicates each to construct bootstrap support values. Parsimony analyses for each bootstrap sample were conducted to produce a large set of trees from which the proportion of entire true trees as well as individual true clades were evaluated. They concluded that bootstrap support values of 70% or more are indicative of a high probability (> 95%) the clade is real. This finding indicated that bootstrap values may be a suitable albeit conservative measure of phylogenetic accuracy.

Additionally, Bull et al. (1993) examined a DNA data set collected from the viral lineages to characterize the molecular evolution and parameterize parametric bootstraps. These 665 sites, of the virus's 39,937 base pair (bp) genome, were chosen due to their likely high rate of substitutions and lack of deletions. Within the ingroup lineages' evolution, a total of 18 substitutions occurred in these DNA sequences from two genomic regions overlapping three genes. This amounted to approximately 0.0019 substitutions per site per ingroup branch, and a total approximate lineage evolution from ancestor to terminal population of 0.0058. Parametric bootstrapping was used with detailed conversion and reversion rate estimates to produce simulated datasets of restriction site evolution. While both parametric and nonparametric bootstrapping are used to produce datasets similar to the original, the former involves parameterizing a simulation using evolutionary rates to create independent datasets while the latter uses bootstrapping to create datasets that lack independence. Tree topology and branch length inference were then carried out using MP, NJ, and UPGMA analyses for these data. Bull et al. (1993) found that each method inferred the entirely correct tree topology with consistent success. Specifically, NJ outperformed MP and UPGMA, with success rates of 99.1%, 97.8%, and 97.3% respectively. The researchers concluded that each method would usually infer the correct topology upon repeated empirically generated phylogenies following their system, but that NJ would perform the best overall. They also determined that MP more accurately

predicted branch length than the two distance methods tested, although it did not perform perfectly.

Hillis et al. (1994) sequenced 1,091 bp across four genes, finding 63 variable sites among the taxa. This dataset had much less phylogenetic potential than the restriction site dataset of Hillis et al. (1992), as it had approximately one-third as many variable sites. MP analyses, using either weighted or unweighted characters, estimated the correct topology, although a second tree with a single clade difference was equally parsimonious. The other methods evaluated, ML, NJ, Fitch-Margoliash, and UPGMA, each found a single, incorrect topology differing from the correct tree by one clade. It is not clear whether each of these methods found the same incorrect topology as one another and whether it was the same topology as MP's other equally parsimonious tree. The researchers sought to compare these DNA sequence data with the restriction site data of Hillis et al. (1992) by creating 1,000 bootstrapped samples for each, while controlling for the amount of variant sites. The percentage of clades accurately resolved varied for each analysis method, with restriction site data performing better than DNA sequence data for each method except ML. Overall, the bootstrapped DNA sequence data produced the most accurate tree using MP (approximately 87% accurate clade resolution), and bootstrapped restriction site data produced the most accurate trees using NJ (approximately 95%). A simple model of evolution, the four-parameter F81 model (Felsenstein, 1981), was used for the ML analyses, and Hillis et al. (1994) suggested that the empirically determined extremely biased substitution matrix (Bull et al., 1993) likely contributed to its poor performance.

Digital evolution for experimental phylogenetics research

Digital evolution systems should, theoretically, be amenable for use with phylogenetic inference, for example in the inference of Avidian evolutionary relationships. Phylogenetics does not require biological life; these methods have been ported to cultural studies as "phylomemetics," with greater or lesser success, for example with the evolution of folktales such as "Little Red Riding Hood" (Tehrani, 2013) or the classification of plastic bag clips (Lehmer

et al., 2011). The fundamental requirements for phylogenetics include the following: ancestor-descendant relationships among evolving individuals (i.e., taxa), a series of multifurcations or at least bifurcations with respect to evolutionary relatedness (i.e., lineage division or cladogenesis), and heritable traits exhibiting variation (i.e., characters). Additionally, for more advanced modern methods, a set of assumptions regarding character evolution (i.e., model of evolution) is required. Avidians are organisms (taxa) that can self-reproduce by virtue of their computer instruction genomic sequence (characters) and undergo evolutionary change resulting in variation in evolutionary relatedness (cladogenesis). Established phylogenetics software is designed to handle biological sequence data such as DNA or amino acid sequences; importantly, the possible instructions used in Avidian genomes can be limited to or recoded as one of these sets of alphabetic characters, such that the software can be used without modification. The utility and degree of information afforded by Avida makes their mutational system of evolution known and configurable while allowing their substitutional system to be knowable by tracking fixation events (see Chapter 3). Thus, models of Avidian evolution can entail fewer untested assumptions than with biological evolution. Moreover, Avida has already had limited usage in experimental phylogenetics research (Hagstrom et al., 2004; Hang et al., 2007, 2003).

As Hillis et al. (1992) initially established the feasibility of their system using experimental conditions theorized to produce evolutionary patterns minimally challenging for phylogenetic inference, I do so by evaluating results of Avidian evolution from a pair of treatments designed to produce the basic molecular evolutionary dynamics and phylogenetic inferences of their work. Specifically, the molecular evolution of T7 presented by Bull et al. (1993) was used to design experimental conditions that in one Avidian evolution treatment should approximately produce the molecular evolution exhibited by the DNA dataset of Hillis et al. (1994), and in a second Avidian treatment that of the restriction site dataset of Hillis et al. (1992). The phylogenetic inference methods of NJ, MP, ML, and BI, and with both the best tree resulting from each analysis and various consensus support trees, are evaluated and compared

for both clade accuracy and resolvability. Basic signatures of molecular evolution such as the number of variable and parsimony informative sites as well as inferred internal and external branch lengths are also evaluated.

A series of 21 additional Avidian experimental evolution treatments are then used to evaluate more complex conditions that are predicted to be challenging to phylogenetic inference. Whereas the influence of natural selection was minimized to the extent possible for the T7 work, various selective regimes are investigated with Avidian evolution, including relatively weaker stabilizing selection and much stronger stabilizing selection occurring uniformly throughout an experimental evolutionary history as well as directional selection variously occurring throughout an evolutionary history. Treatments with lineages varying in their extent of evolution, and thus inferred branch length, were conducted to investigate such dynamics known to cause issues for phylogenetic inference. And finally, the effects of recombination, as implemented in Avida, are investigated in combination with these other conditions.

These experiments are used to examine how clade accuracy relates to support values within the context of individual clades and across entire trees. This large set of Avidian evolution phylogenetic data is used to reexamine the results of Hillis and Bull (1993) for bootstrap support values and additionally investigate how BI posterior support values correspond with clade accuracy. Finally, how clade accuracy and clade resolvability correspond with best, MRe, and various MR consensus threshold trees are compared across analyses to examine tree accuracy within a whole-tree context.

Methods

Experimental design

Base evolutionary history

A total of 23 experimental treatments, each with ten replicates, were conducted, with most treatments sharing a base design. In the base design, eight-taxon Avidian evolutionary

histories were produced, with each lineage in the history evolving for an equivalent number of generations for any given experimental treatment (Fig. 2.01a). A single genotype was used as an ancestor in Avida to initiate two independent experimental histories (i.e., branches A and B), together termed tree level 1 for the first set of lineages. Lineage evolution occurred for a set number of generations then from each lineage two new lineages were identically seeded (e.g., branch A leading to branches C and D, tree level 2). This was again repeated with each of these lineages producing two more descendent lineages (e.g., branch C leading to branches G and H, tree level 3). Finally, this third set of lineages evolved for that same set number of generations. The set of "extant" populations or taxa is two to the number of tree levels, 2^3, or eight. An equivalent length of time, as in the elapsed number of Avidian generations, occurred during each lineage's evolution, with the total number of generations from the ancestral genotype to any one extant taxon being its multiple of tree levels. For example, if each branch persisted for 100 generations then the number of generations from the ancestor to the extant taxon is 300 generations.

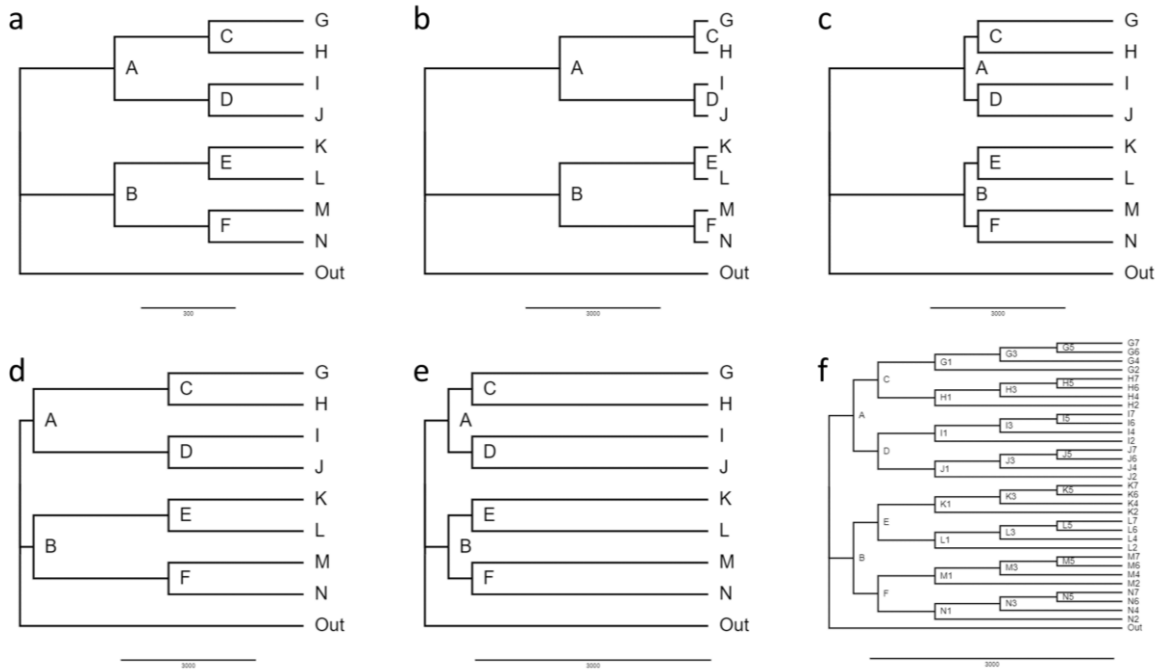


Figure 2.01. Representative evolutionary history topologies with branch lengths denotating the number of generations lineages evolved. The base design experienced equivalent generations of evolution per branch, either 100 (not shown), 300 (a), or 3,000 generations per branch (not shown). Four designs (b-e) had differing numbers of generations per tree level, with either short ("S") branches of 300 generations or long ("L") branches of 3,000 generations across all branches among a level. These designs are named by tree level length from external to internal tree levels, and include SLL (b), LSL (c), LLS (d), and LSS (e). An additional treatment, LSS^{B} (f), used the LSS branching pattern with additional external branches to "break-up" the long branch, resulting in a 32-taxon asymmetrical history. The scale bar in subplot a is 300 generations and the scale bar in subplots b-f is 3,000 generations. Internal branches are labeled at their terminal node, and all evolutionary histories were true polytomies at their origin.

The ancestor for branches A and B was a cloned 333-instruction long Avidian. For most experiments, the ancestor was a longer counterpart to the standard default Avidian genotype, which can replicate its genome to produce offspring but perform no other meaningful computation, for example a task rewarded by the environment. This default genome consists of two strings of instructions, necessary for reproduction, spaced apart by "blank tape" of *nop-C* instructions, for a total genome length of 100 instructions. This large *nop-C* region acts as genomic filler within which mutations can occur that might code for interesting computation, e.g., task performance. An additional 233 *nop-C* instructions were added to this filler region to create the 333-instruction ancestral genotype. This is referred to as the "naïve" ancestor to differentiate it from the "pre-adapted," task proficient, ancestor. Using a population size of

10,000 to increase the efficiency of selection, the pre-adapted ancestor reached full task proficiency in the logic-9 task resource environment. After the population contained numerous organisms capable of performing all nine tasks, it evolved an additional 100 generations to further improve task performance and reproductive efficiency. In all treatments, genome length was fixed by disallowing insertion and deletion mutations and further requiring a null genome size differential between parent and offspring; together, these settings ensured the preservation of homology at each position in the genome.

After the set number of generations elapsed for each internal branch, new lineages were seeded using the single most abundant genotype in the extant population. This occurred in all treatments that lacked recombination. In recombination treatments, lineages were seeding with the entire extant population to additionally include the effects of lineage sorting among segregating variation. Experiments that included recombination were configured with the number of "modules" in the genome set as 333, equal to the fixed genome size. This resulted in complete independent assortment among loci.

The per site mutation rate was determined by balancing adherence to the work of Hillis et al. (1992) while maintaining the relative likelihood that Avidians would adapt in selective environments by acquiring task performance within a reasonable amount of time. Using values reported in Bull et al. (1993), the substitution rate of the T7 phage history was calculated to be 18/332.5/14/100 = $3.867 * 10^{-5}$ substitutions/site/branch/generation: 18 substitutions were observed among the 14 ingroup internal and external branches. Effectively 332.5 sites were used, because mutations only affected the G/C sites, which were approximately half of the 665 sequenced DNA sites. Approximately 100 generations of evolution occurred along each branch, with an estimate of 2-3 generations per T7 lytic cycle and 40 cycles per branch. The aim was to use a mutation rate within one order of magnitude of this empirical T7 substitution rate, so the per site mutation rate was set to $3.867 * 10^{-4}$. Avidians reproduce with over-lapping generations such that the mutation rate affects an offspring genotype and not the parent's, with both organisms persisting following reproduction. Thus, the effective mutation rate is one

half this rate, or $1.9335 * 10^{-4}$, which is five times the observed substitution rate in the T7 study. With 100 generations per branch, the expected phylogenetic branch length under neutral evolution would therefore be approximately $2 * 10^{-2}$ substitutions/site/branch. Finally, with a 333-site genome, the expected number of substitutions/branch would on average be 6.4.

Non-default Avida settings shared by all experiments included the following: The instruction set disallowed seven instructions from being incorporated into the genome via mutation. One of these, *h-copy*, is required for Avidian genome replication and each genotype used as an ancestor included a single copy of this instruction; a mutation changing this position would cause the organism to be inviable. The remaining six disallowed instructions were *if-less*, *set-flow*, *shift-r*, *shift-l*, *dec*, and *add*. This resulted in a total of 20 possible instruction characters, one of which was effectively invariant. For use in phylogenetic analysis algorithms, these 20 unique instructions were coded using the conventional single-letter amino acid abbreviations, allowing the resulting genetic sequences to pass as biological amino acid sequence data. The birth method was set as "mass action," in which an offspring is placed randomly into the population instead of near their parent, resulting in a lack of spatially- and genetically-structured populations and increasing the effectiveness of selection because organisms were equally likely to compete for space with relatives and nonrelatives. All other settings were as the default, most notably including the default logic-9 task resource (i.e., selection) environment and "power" merit rewards (i.e., strength of selection); some treatments altered the presence/absence of a task resource, but none altered reward strength or included tasks outside this set. Avida (Ofria et al., 2009; Ofria and Wilke, 2004) version 2.9.0 was used with only minor modifications to produce custom population output files.

Stabilizing selection treatments

Six experimental treatments were conducted under stabilizing selection with all lineages evolved for an equivalent amount of time within each history. These experiments differed in having relatively weaker or stronger stabilizing selection by using either a naïve or a pre-adapted ancestor (termed "Stabilizing" and "Uniform" as treatment conditions, respectfully),

the absence or presence of recombination ("Asex" and "Sex"), and the number of generations the lineages evolved (100 or 300 per branch). For treatments using the pre-adapted ancestor, the logic-9 task resource environment was used, resulting in strong stabilizing selection for task maintenance. All other stabilizing selection treatments resulted in relatively weak selection because organisms remained under selection to reproduce efficiently, although without the possibility of drastic fitness reductions due to the loss of task function.

Four experimental treatments were conducted under the same relatively weak stabilizing selection but with differing numbers of generation per tree level throughout each history, although all branches within a level were of the same length (Fig. 2.01b-e). Short ("S") branches were 300 generations and long ("L") branches were 3,000 generations. These treatment conditions are named by their tree level size reading from the external branch to the most internal tree level. For example, treatment "LSS" had eight external branches of 3,000 generations in length, a set of four internal branches of 300 generations, then the basal two ingroup branches of 300 generations. Note that these treatments are named in reverse chronological order, as each lineage's evolution in the LSS treatment experienced an initial 300 generations, lineage division and 300 more generations, and lineage division with a final 3,000 generations of evolution.

An additional treatment used the LSS branching pattern but differed by having additional external branches to "break-up" the long branch, thereby named "LSS^{B}" (Fig. 2.01f). In this treatment, the external branches were bifurcated after 700 generations, and only one of the resulting two branches was then bifurcated after an additional 800 generations, and again one of the resulting two branches was bifurcated after 700 generations, with 800 final generations remaining before the experiment ceased. For each of these bifurcations, the lineage that underwent no further bifurcations evolved for an additional length of time such that all branches evolved for a total of 3,000 generations. Rather than resulting in eight extant taxa that evolved in a fully symmetrical pattern as with all other experimental histories, this treatment resulted in a 32-taxon asymmetrical history.

Directional selection treatments

Eight experimental treatments were conducted using one of four directional selection regimes (Fig. 2.02) and in either the absence or presence of recombination, with all branches lasting for 3,000 generations. Selection regime "1" was the full Logic-9 task resource environment. These tasks can be classified into five theoretical difficulty classes, with each class requiring a longer and/or more complex sequence of instructions. There are two tasks per each of the first four difficulty classes and one (i.e., EQU) in the hardest class. Selection regime "2" consisted of two environments, with each environment rewarding four tasks, one from each of the four easiest classes. Branch A and all its descendent lineages evolved in one environment, and branch B and its descendants in the other. In this way tasks were not selected in parallel between the GHIJ and KLMN extant clades. Regime "3" consisted of four environments and was identical to "2" except branch A only had one of the two easiest tasks and B the other. Each of their descendants had the remaining 3 tasks from the environment "2" sets. Finally, in selection regime "4" branches A and B had the same environment as in "3," their second tree level branches had one additional task of the two in the second difficulty class, and the final tree level branches had a further one additional task of the two in the third difficulty class. In this manner, for selection regime "4," all 14 ingroup branches experienced a unique selective environment. In every selection regime new tasks remained rewarded throughout the duration of a lineage's evolution, for example every task rewarded in branch C remained rewarded in branches G and H.

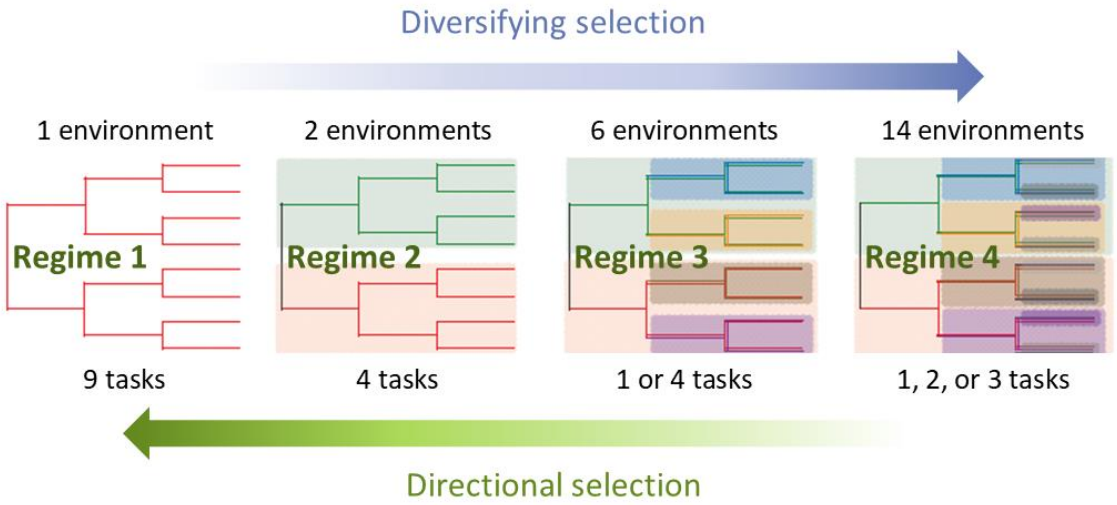


Figure 2.02. The four natural selection regimes visualized with respect to the base evolutionary history topology. Colored shading indicates shared tasks rewarded among one or more branches within a tree. Diversifying and directional selection are inversely related in these designs, with regime "1" having the same selective environment across all branches as well as the strongest directional selection, and regime "4" having a unique selective environment on each branch but the weakest directional selection compared to the other regimes.

In these environments, the strength of selection, as in the difference in potential fitness between an organism capable of performing none or all tasks selected in the environment, decreased from regime "1" through "4." Environment "1" had all nine tasks rewarded per branch and the maximal advantage for performing all versus no tasks present in the environment was 33,554,432x; "2" had 4 tasks per branch and the maximal advantage was 1,024x; "3" had either 1 or 4 tasks per branch and the maximal advantage was either 2x or 1,024x depending on the branch; and "4" had 1, 2, or 3 tasks rewarded per branch, with maximal advantages of 2x, 8x, or 64x, respectively. Since task performance usually entails a nominal reduction in offspring cost relative to merit and since Avidians tend to evolve tasks sequentially, the realized differences in fitness between contemporaneous organisms is most likely reduced from these values. Note that even still, the relatively weaker selection here is still very strong selection compared to biological organisms in most environments. All directional selection treatments used a population size of 1,000 organisms to further lessen the influence of genetic drift.

Finally, four experimental treatments were conducted with directional selection (regime "2" or "3"), with or without recombination, and with the LSS tree level pattern of generations of evolution per tree level.

Analyses

Taxon-character datasets

For each of the ten replicates per experimental treatment, a taxon-character dataset was created using the eight extant populations, or 32 in the case of LSS^B. A random organism was sampled from each population, and therefore more abundant genotypes had a greater likelihood of being sampled. These sampled genotypes constituted the ingroup taxa for the replicate. The outgroup taxon was a randomly sampled organism from an extant population of a different replicate of that same experimental treatment; therefore, the outgroup taxon evolved for the same total number of generations and experienced similar evolutionary conditions to the ingroup taxa. Note that the base of the evolutionary history is a true polytomy of the outgroup, branch B, and branch C, that is, branch B and C do not share a more recent common ancestor than either does with the outgroup (Fig. 2.01).

The complete genomic sequence of the sampled organism was used. Each sequence consisted of 20 single-letter characters that represented the computational instructions that were available to mutation. Three of these characters, J, O and B, did not have amino acid abbreviation counterparts and were therefore translated to W, Y, and V, respectively. Therefore, phylogenetic programs designed to handle amino acid character data would treat Avidian sequences as such. Since genome length was fixed to be a constant 333 characters, sequence alignment was not required, and each locus had perfect homology. Across the 23 experimental treatments with 10 replicates per treatment a total of 220 eight-taxon by 333-character datasets and, for the LSS^B treatment, 10 thirty-two-taxon by 333-character datasets were created. Each of these datasets was used to conduct four different phylogenetic analyses.

Phylogenetic analyses

Analyses were conducted using settings that were as simplistic as possible, i.e., using default settings and minimally parameterized models. This was done to avoid biasing the results in favor of one or other inference method. For example, several settings could be altered for each individual phylogenetic analysis with the aim of improving the inference accuracy; however, this could become untenable when inferring multiple trees from each of 230 datasets. Unless otherwise indicated, default settings were used for each program. In all cases, analyses were identically conducted across all experimental treatments and replicates.

The simplest amino acid model of molecular evolution was used, the Poisson model. The Poisson model is a fixed rate model that assumes equal rates and state frequencies among all 20 characters (Bishop and Friday, 1987), and is analogous to the JC model for nucleotide evolution. This is, in fact, the mutational model of molecular evolution as implemented in Avida, where each instruction has an equal probability of mutating to any other. Although higher-parameterized models were variously suggested by such programs as ProTest (Abascal et al., 2005), there is no theoretical reason why, for example, a *nop-c* instruction should behave like a cysteine because both are abbreviated as C. Further, I wanted to conduct analyses as similarly as possible across treatments. Because rate heterogeneity was expected to occur in this Avidian evolution, with at least one site being invariable (the *h-copy* site necessary for reproduction, as discussed previously), the model of evolution additionally included rate heterogeneity among sites. The model did so by allowing both a proportion of invariable sites and a discrete Gamma model with four rate categories, since this combination of rates is very commonly used (Stamatakis, 2016), despite criticism that these parameters cannot be optimized independently (Yang, 2006).

Trees were inferred using NJ, MP, ML, and BI, and fully-resolved “best” and majority rule consensus trees were created for each analysis, as possible. Neighbor joining trees were constructed using QuickTree, version 2.0 (Howe et al., 2002). Since this algorithm produces only a single tree inference, when included in figures comparing consensus trees, the NJ tree is

clarified as "¬Con.," i.e., not a consensus tree. MP was implemented using MPBoot, version 1.1.0 (Hoang et al., 2018, 2017). The number of initial parsimony trees evaluated was increased to 10,000 to better search the set of possible trees, and 1,000 ultrafast bootstrap replicates were conducted. These bootstrap trees were used to construct 50%, 70%, 95%, and 99% consensus trees. The "best" tree in the analyses reported here is a random selection among the equally parsimonious trees identified in the heuristic search. The source code of MPBoot was modified to enable the printing in the log file of the parsimony score for each of the trees evaluated in the set of candidate trees, which was used to calculate the minimum number of equally parsimonious trees. This value is a minimum in that since an exhaustive search was not conducted, there is a possibility that a region of tree space with better or equivalent parsimony scores was not explored. ML was implemented using IQ-TREE, version 1.5.5 (Minh et al., 2017; Nguyen et al., 2015). For each analysis, a total of 100 nonparametric bootstrap replicates were also produced and used to construct consensus trees with the four above thresholds in addition to a MRe cladogram. The "best" tree reported here is the single phylogram with the greatest likelihood value. BI was implemented using MrBayes, version 3.2.5 (Ronquist et al., 2012, 2011), and specifically using the parallel processing implementation and the BEAGLE library (Altekar et al., 2004; Ronquist et al., 2012). The single non-default setting was that the Markov chain was sampled every 100 generations, to provide a greater number of cladogram samples. The post-burn-in sampled trees were used to construct consensus trees with the above thresholds as well as a MRe cladogram. The "best" tree reported here is the single cladogram with the greatest tree posterior probability, also termed the maximum clade credibility tree. Among other uses, Newick utilities (Junier, 2011; Junier and Zdobnov, 2010) were used to root and produce consensus trees for each threshold. The Consense program from the PHYLIP package (Felsenstein, 2005) was used to produce MRe trees. FigTree (Rambaut, 2018) and the Iroki web application (Moore et al., 2020) were used for tree visualization. Finally, Python, version 2.7, and the following packages, among other general modules, were used to organize and present data: Jupyter, version 0.27.0 (Kluyver et al., 2016; Perez and Granger, 2007); Matplotlib, version

1.3.1 (Hunter, 2007); ETE2, version 2.2.1 (Huerta-Cepas et al., 2016); and DedroPy, version 3.12 (Sukumaran and Holder, 2010).

Phylogenetic measures

Topological accuracy between the true tree and inference tree was calculated using variants of the Robinson-Foulds (RF) distance. Also termed the symmetric difference, RF distance is the number of internal branches (also called edges, partitions, or splits) that are present in one tree and not the other, and vice versa (Robinson and Foulds, 1981). RF distance is therefore the sum of false positive branches (FP, those appearing in the inferred tree and not the true tree) and false negative branches (FN, those appearing in the true tree and not the inferred tree). These can be further calculated as rates. The FP rate is FP divided by the number of internal branches in the true tree, and the FN rate is FN divided by the number of internal branches in the inferred tree. The arithmetic mean of the FP and FN rates is termed the average topological error (Swenson et al., 2010). RF distance can be normalized by dividing by the maximal possible RF distance, and when the trees under comparison are binary trees (i.e., are fully resolved by lacking polytomies) the average topological error is equivalent to the normalized RF distance. Since I want to emphasize the accuracy of inference methods rather than their error, I report complement values. I have termed the FP rate complement as "Clade Accuracy," since this metric indicates the percentage of clades correctly inferred; the FN rate complement as "Clade Resolvability," since this metric indicates the percentage of clades correctly resolved; and average topological error as "Average Topological Accuracy," since this metric indicates the overall accuracy of the phylogenetic inference. Examples of these metrics are shown for a comparison of a known (or correctly inferred) cladogram (Fig. 2.03a) to four variously-inaccurate inferred cladograms (Fig. 2.03b-e). Note that the best tree topology produced by an analysis is always fully resolved although it may include incorrect clades, so for these trees each false positive clade requires a counterpart false negative clade, and thus clade accuracy, clade resolvability, and average topological accuracy are necessarily equivalent values (e.g., Fig. 2.03b,e). This holds for all best trees across analyses; even though BI maximum clade

credibility trees are not necessarily fully resolved, they were for all instances analyzed here. Trees that are not fully resolved, for example as resulting from a majority rule consensus algorithm, contain at least one false negative clade and may contain zero (e.g., Fig. 2.03c) or multiple false positive clades (e.g., Fig. 2.03d). Each analysis's best tree as well as the variously constructed consensus trees were evaluated for these metrics.

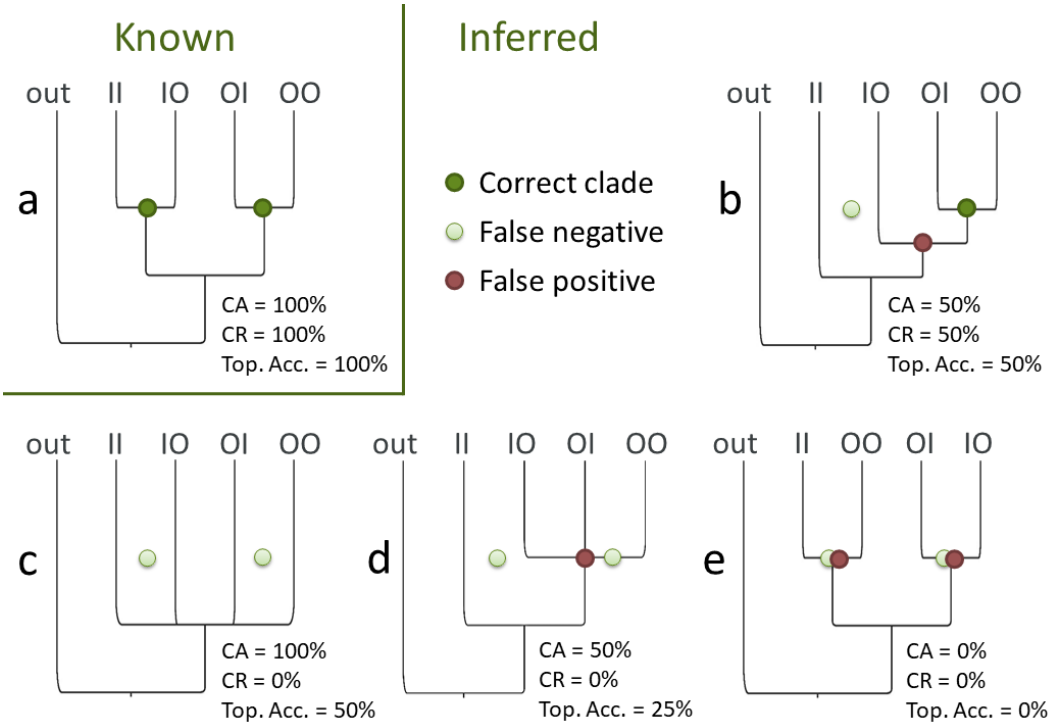


Figure 2.03. Example uses for the clade support metrics of Clade Accuracy (CA), Clade Resolvability (CR), and Average Topological Accuracy (Top. Acc.) for comparing a known cladogram to four inaccurate cladograms. This four-ingroup-taxon rooted tree has two true clades to be inferred (II + IO and OI + OO, dark green circle) and therefore as many as two false negative (light green circle) and two false positive (red circle) clades per cladogram. Note that the actual topologies of the Avidian evolution experiments constitute 8-taxon or 32-taxon trees, with many more potential combinations of false negative and false positive clades.

Additional metrics include comparisons of inferred branch lengths and the amount of variable and parsimony informative sites. Branch lengths are summarized as the median length across all internal branches and, separately, across all external branch lengths. Only branch lengths as inferred for each analysis's best tree will be shown, so that polytomies, as may arise on consensus trees, are not present and therefore all potential internal branches are included. As with the model of evolution, digital evolution allows the comparison of branch lengths to the

actual number of observed substitutions occurring within lineage evolution. However, such an analysis was not conducted for this set of experiments (although, see Chapter 3). Positions in the taxon-character dataset are considered variable if at least two types of characters exist among the set of taxa. Additionally, positions are informative under a MP analysis (and other phylogenetic approaches) if at least two sets of characters are found among at least two taxa each. Sites which lack informativeness are either invariant across taxa, are autapomorphic in that variation is unique to single taxa, or otherwise such that every possible cladogram would score equivalent under MP. These basic measures of taxa-character data richness are directly comparable across phylogenetic studies and provide a basic estimate of the phylogenetic signal of the data.

Statistics

Recognizing that phylogenetic tree metrics lack independence, the measures reported here are limited to comparisons of median values or aggregate sums across treatments and replicates, as indicated. Various statistical tests have been used to compare phylogenetics simulation results (e.g., Hall 2005; Wang et al. 2011); although other studies eschew their use (e.g., Kuhner and Felsenstein 1994; Barbançon et al. 2013). It is not clear that the Central Limit Theorem and its derivative parametric statistical tests and even non-parametric tests are applicable in such work since clade and branch inclusion (and therefore branch length) do not exhibit independence. Further, even if statistical tests were suitable for simulation studies, it is not clear that they would remain so for experimental phylogenetics research where each replicate is a separate instance of evolution.

Results

Treatment comparisons

Treatment conditions are labeled in Figures 2.04-2.15 by their distinctive selection, recombination, and branch length conditions. They are also arrayed in the order described in the experimental design section, above. Selection conditions include the following: starting

with the naïve ancestor in an environment rewarding no resources, resulting in relatively weak *stabilizing* selection; starting with a pre-adapted ancestor in the full selective environment, resulting in relatively strong and *uniform* stabilizing selection; and selection regimes "*1*", "*2*", "*3*", and "*4*," as described above. Recombination conditions include its absence as *asex*ual reproduction or its presence as *sex*ual reproduction. The number of generations per branch include topologies with uniform lengths for all tree levels of *100*, *300*, or *3,000* generations. Topologies of branch lengths differing in length per tree level have labels reflecting *s*hort (300 generations) or *l*ong (3,000 generations) lengths and are denoted from extant to ancestral tree levels, for example "LSS" had only long external branches; and the final of these treatments ("LSS^{B}") had long external branches *b*roken-up with additional branches, resulting in a 32-taxon asymmetrical history.

The number of variable and parsimony informative sites found for the taxon-character dataset for each experimental treatment replicate are shown in Figure 2.04. For the first six treatments, i.e., those with stabilizing selection and topologies with uniform 100- or 300-generation branches, the number of variable sites strongly differed per treatment, and for each treatment the number of informative sites was approximately half the number of variable sites. The trends among these treatments were that longer-evolved branches yielded greater numbers of informative and variable sites, and that treatments starting with the pre-adapted ancestor resulted in fewer numbers of sites. For all other treatments, i.e., those with at least one set of branches that was 3,000 generations long, nearly the entire genomic sequence (333 loci) exhibited variation, with only 10-20 fixed sites. The number of informative sites was nearly this numerous for SLL and LSS^{B} branch length treatments, with treatments LSL and LLS having a third less, and treatment LSS having less than half as many. The twelve treatments conducted under directional selection and with or without recombination did not exhibit a clear pattern with respect to the number of informative sites, although it was reduced for the last four of these treatments, that is, those with the LSS branch pattern.

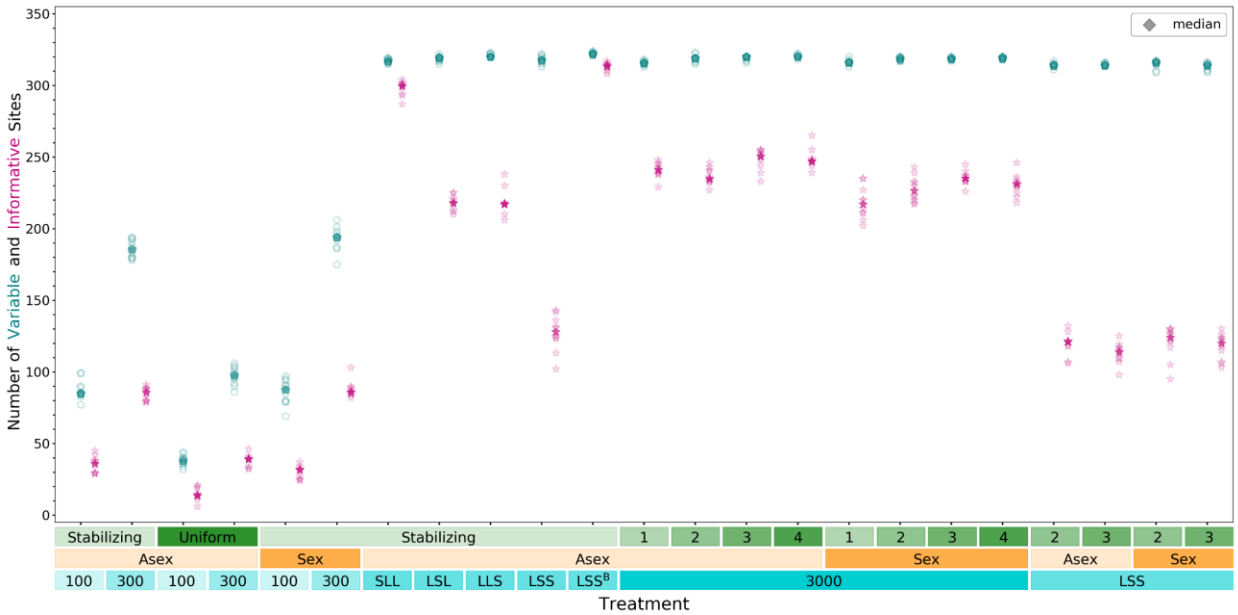


Figure 2.04. Number of variable sites (blue pentagons) and parsimony informative sites (purple stars) for all experimental treatments. Open symbols are for individual treatment replicates, and closed symbols for the treatment median. Experimental treatments are denoted by their selective condition (green labels), recombination condition (orange labels), and number of generations per branch (cyan labels); see text for further information on condition notation.

Inferred branch lengths for each experimental treatment are summarized in Figure 2.05. Only analyses that resulted in a phylogram with branch lengths expressed as the number of inferred substitutions per site are shown, leaving MP excluded here. The median of internal branch lengths and the median of external branch lengths are presented for the single best tree of each analysis for each of the ten replicates per treatment. Although differences between the first six treatments, i.e., those with stabilizing selection and uniform branch lengths, appear slight at this scale, the trends match that for the number of variable and informative sites, with longer inferred lengths for topologies with a greater number of generations per branch and shorter inferred lengths for treatments starting with the pre-adapted ancestor. While it is barely noticeable for these first six treatments, the trend of NJ underpredicting branch lengths relative to the other methods is readily apparent for all further treatments and is especially the case for conditions with longer branches. There is also a general trend across nearly all 23 treatments of ML inferring slightly shorter branch lengths than BI.

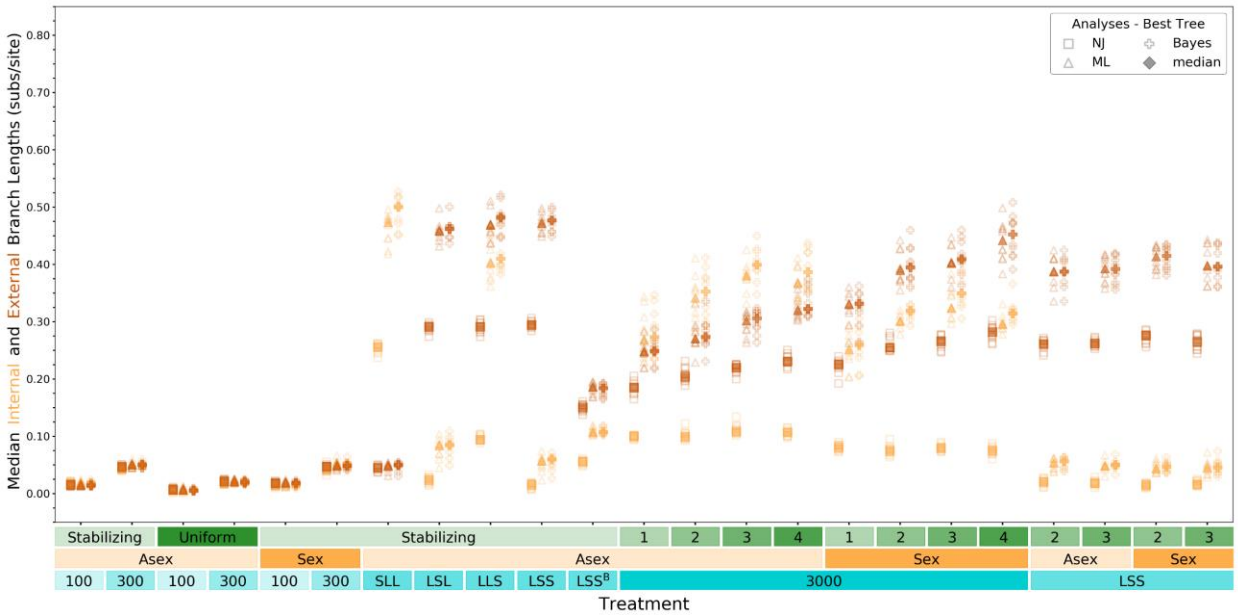


Figure 2.05. Inferred median internal branch lengths (yellow) and median external branch lengths (orange) for the single best tree resulting from NJ (square), ML (triangle), and BI (plus) analyses. Open symbols are for individual replicates, and closed symbols for the median across treatment replicates. Note that MP is not included since it does not report trees with branch lengths expressed as the number of substitutions per site.

For the four treatments of stabilizing selection with differing numbers of generations per tree level, the pattern of inferred median internal and external branch lengths matches the treatment design (Fig 2.05): long internal branches for treatments SLL and LLS, short internal branches for treatments LSL and LSS, short external branches for treatment SLL, and long external branches for treatments LSL, LLS and LSS. Note that since internal branches include the two branches at the basal tree level (labeled A and B in Fig. 2.01) and the four branches at the middle tree level (labeled C-F in Fig. 2.01), the median across these branches is dominated by the middle tree level. Across these treatments, with ML and BI analyses, inferred long branches are slightly less than ten-fold longer than short branches. For treatment LLS, NJ especially underpredicted internal branch lengths compared to ML and BI. The treatment with broken-up long external branches (LSS^B) had relatively short external branches, consistent with its additional taxa with fewer generations per branch.

For all directional selection treatments, branches were generally reduced in length from that observed for stabilizing selection treatments that also had branches of 3,000 generations,

for example the long external branches of treatment LSS (Fig. 2.05). For ML and BI analyses with treatments either with or without recombination, external branch lengths tended to increase from selection regimes 1 through 4, and internal branch lengths tended to increase from regimes 1 through 3 then slightly decrease in regime 4 compared to 3. While NJ maintained its trend of underestimating branch lengths, it did not exhibit this pattern of change with respect to selection regimes. Internal branch lengths are greater for asexual reproduction treatments and external branches are greater for sexual reproduction treatments. For the final four directional selection treatments with the LSS branch pattern, inferred internal and external branch lengths were similar to although reduced from the stabilizing selection LSS treatment. External branch lengths for these four treatments, while similar to one another, were increased compared to the standard selection regimes 2 and 3 without recombination.

Clade accuracy and clade resolvability results are presented separately in Figures 2.06-2.13 for sets of experimental treatments. Within each figure, treatments are separated by dotted blue lines, and conditions are labeled as previously indicated. Within each treatment, all ten replicates of each of sixteen phylogenetic analysis and consensus threshold combinations are presented: The NJ tree, then for each of MP, ML, and BI analyses, the analysis' best tree and then consensus trees of thresholds 50%, 70%, 95%, and 99%, with dashes differing in width at the figure top indicating the relative ordering of these trees within the set of trees per analysis. Analyses are denoted by separate symbols; and sets of trees per analysis are further visually distinguished by a thin vertical line.

Clade accuracy was near-perfect for five of the six treatments conducted under stabilizing selection with lineages evolved for an equivalent amount of time (Fig. 2.06). One replicate each for the 100-generation per branch treatments under weak stabilizing selection resulted in either a NJ or MP best tree inferring a single incorrect clade. The 100-generation treatment under strong, uniform stabilizing selection had decreased clade accuracy for multiple replicates of each analysis's best tree, although NJ and BI best trees performed better overall as indicated by their medians retaining 100% accuracy. For this same treatment, ML continued to

have decreased accuracy for at least one replicate up to and including the 70% consensus threshold. These treatments, stabilizing selection with lineages evolved for an equivalent amount of time, did not score as highly in terms of clade resolvability (Fig. 2.07). Treatments with 300 generations per branch maintained greater clade resolvability than their 100-generation equivalents. And compared to those with weaker stabilizing selection, treatments with strong stabilizing selection had greatly reduced resolvability, with a few replicates having a complete comb- or rake-like topology of 0% resolvability (as in the example of Fig 2.03c). Comparing the medians across consensus thresholds, BI maintained the best resolvability followed by MP then ML for each treatment demonstrating variation in clade resolvability. Finally, recombination appears to have slightly increased clade resolvability across consensus thresholds for each analysis when comparing the otherwise equivalent 100-generation treatments.

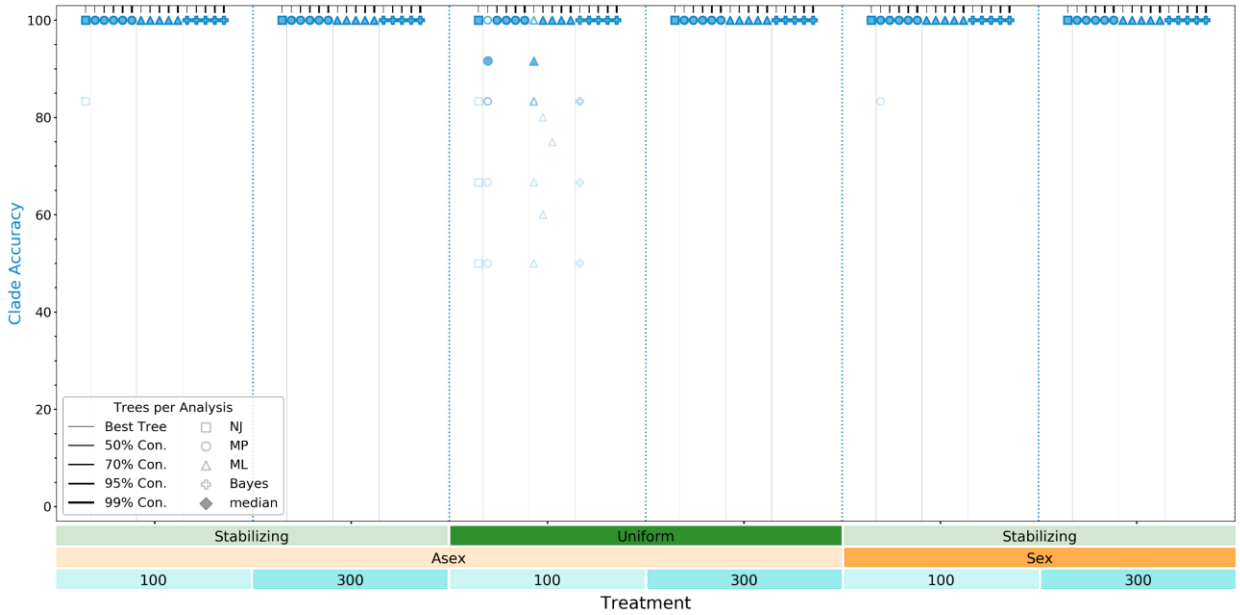


Figure 2.06. Clade accuracy, the percentage of clades correctly inferred, for the set of treatments with stabilizing selection and lineages evolved for equivalent generations per branch. Best and consensus trees are included for each analysis, including NJ (square, best trees only), MP (circle), ML (triangle), and BI (plus), with open symbols for individual replicates and closed for the median across replicates. Within each analysis, the best tree then consensus trees in the order of increasing threshold strictness are arrayed with tick marks at the top indicate these distinctions; see the text for further information on tree type notation.

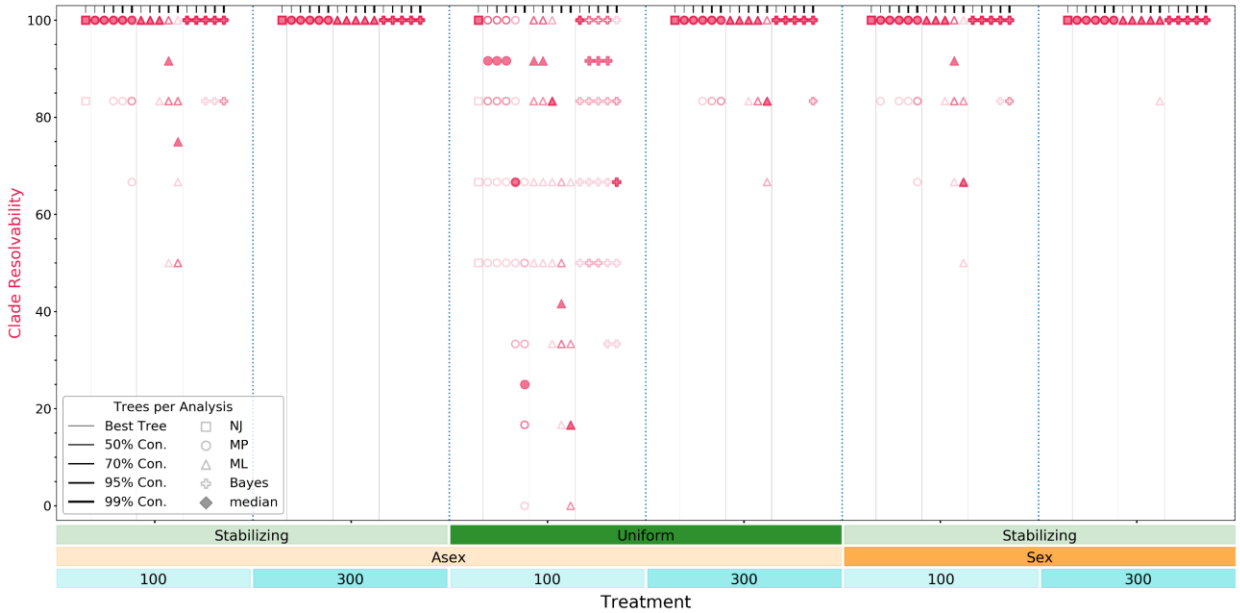


Figure 2.07. Clade resolvability, the percentage of clades correctly resolved, for the set of treatments with stabilizing selection and lineages evolved for equivalent generations per branch. Best and consensus trees of increasing threshold strictness are included for each analysis, except for NJ which only includes the best tree, with open symbols for individual replicates and closed for the median across replicates.

Weak stabilizing selection with differing numbers of generations per tree level resulted in mixed results for clade accuracy (Fig. 2.08). Clade accuracy was perfect for all analyses and consensus thresholds when all internal branches were long (i.e., SLL) and when external long branches were disrupted by increased taxon sampling, (i.e., LSS^{B}). However, with long external branches and at least one set of short internal branches clade accuracy is reduced for many if not all replicate best trees across analyses and does not reach 100% for all replicates until consensus threshold 70% or 95% for MP and ML and threshold 95% for BI. Clade resolvability is largely perfect for treatment SLL across analyses and consensus thresholds, and treatment LSS^{B} has reduced resolvability for stricter thresholds (Fig. 2.09). Clade resolvability is quite low for treatments LSL, LLS, and LSS. Treatment LSL had about half the resolution across analyses compared to LLS, and treatment LSS had very low resolution, with stricter thresholds having replicates of 0% resolvability. Comparing across consensus thresholds for both clade accuracy (Fig. 2.08) and resolvability (Fig. 2.09), BI performed better than ML, and ML better than MP.

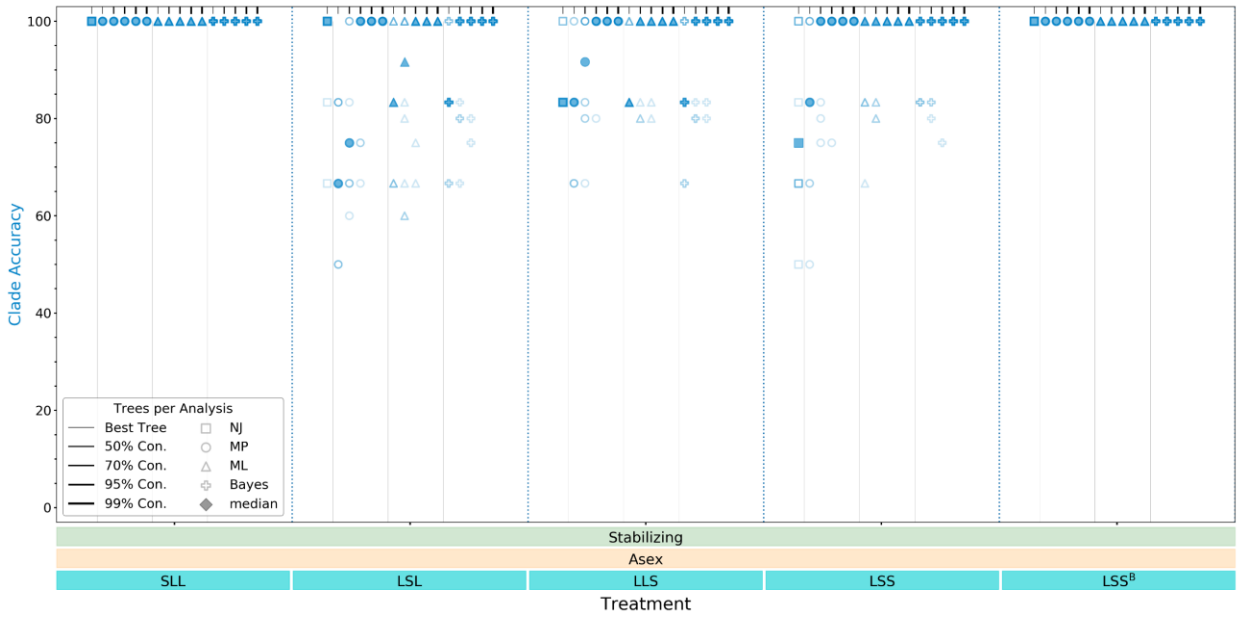


Figure 2.08. Clade accuracy for the set of treatments with stabilizing selection and lineages evolved for differing generation per tree level. Best and consensus trees of increasing threshold strictness are included for each analysis, except for NJ which only includes the best tree, with open symbols for individual replicates and closed for the median across replicates.

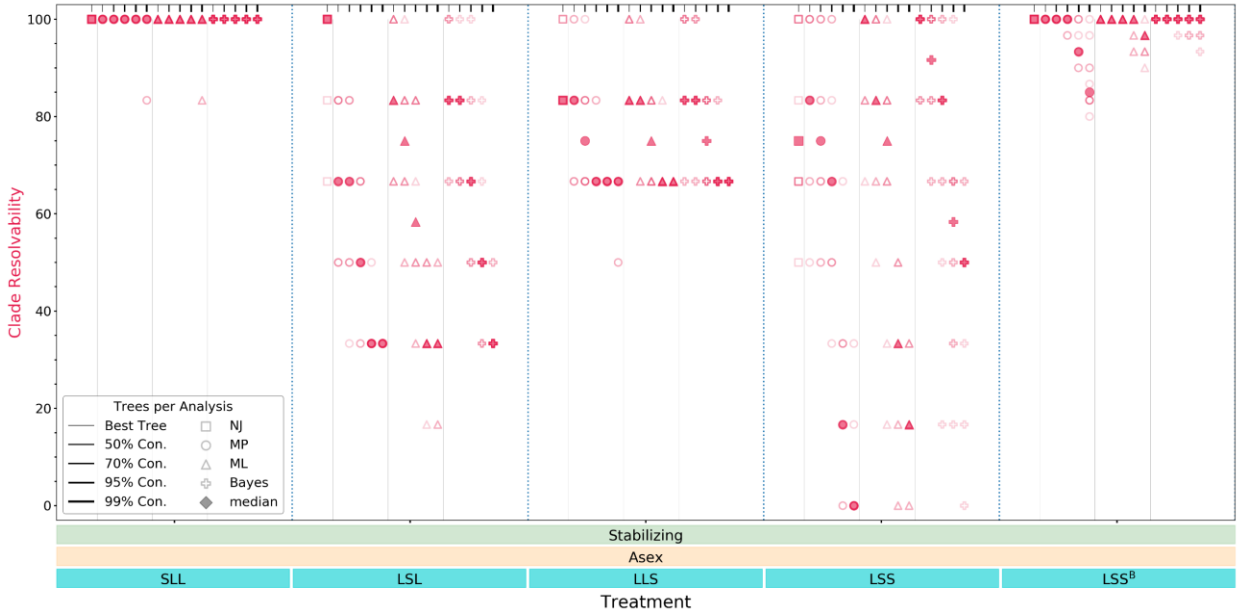


Figure 2.09. Clade resolvability for the set of treatments with stabilizing selection and lineages evolved for differing generation per tree level. Best and consensus trees of increasing threshold strictness are included for each analysis, except for NJ which only includes the best tree, with open symbols for individual replicates and closed for the median across replicates.

The treatments conducted under directional selection and with a consistent 3,000 generations per branch had essentially perfect clade accuracy across analyses and consensus

thresholds (Fig. 2.10). The single exception was one MP best tree replicate that had one clade incorrect, although an equally parsimonious tree was identified although not selected for inclusion here due to random chance. Clade resolvability was generally quite high for most replicates (Fig. 2.11). For both the set of sexual and asexual reproduction treatments, clade resolvability was reduced from regime 1 through 3, with the slight distinction between regimes 3 and 4. Comparing across consensus thresholds, BI maintained greater clade resolvability than ML, with MP having the lowest. The set of treatments with recombination had slightly lower clade resolvability at stricter consensus thresholds than their equivalent selective regime treatments without recombination.

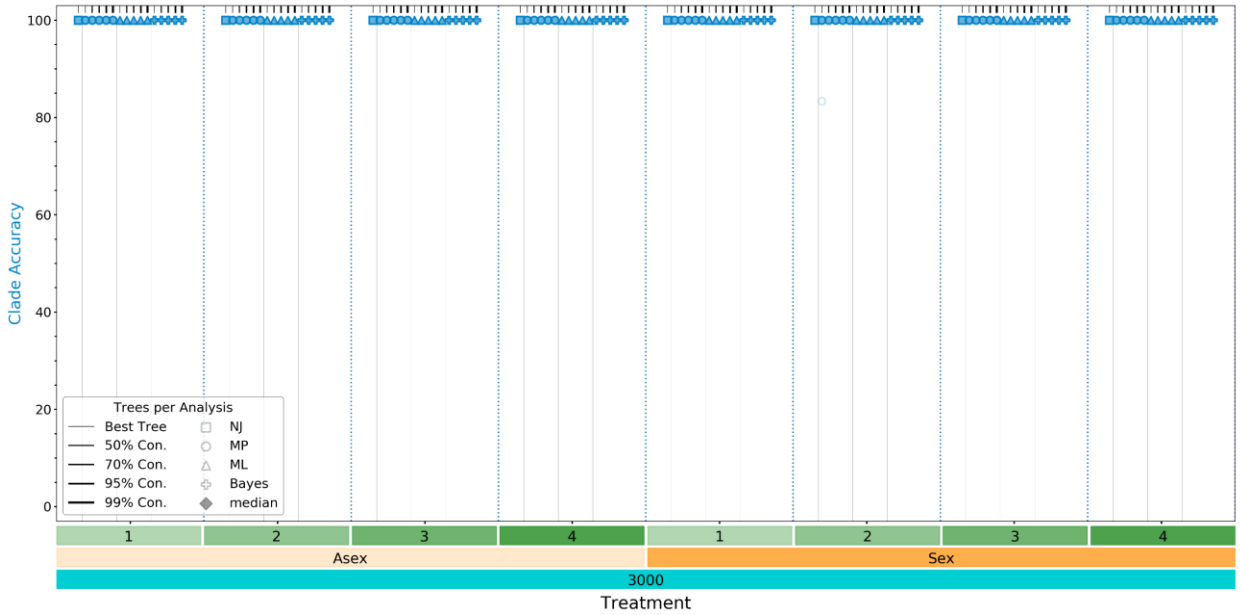


Figure 2.10. Clade accuracy for the set of treatments with directional selection and lineages evolved for equivalent generations per branch. Best and consensus trees of increasing threshold strictness are included for each analysis, except for NJ which only includes the best tree, with open symbols for individual replicates and closed for the median across replicates.

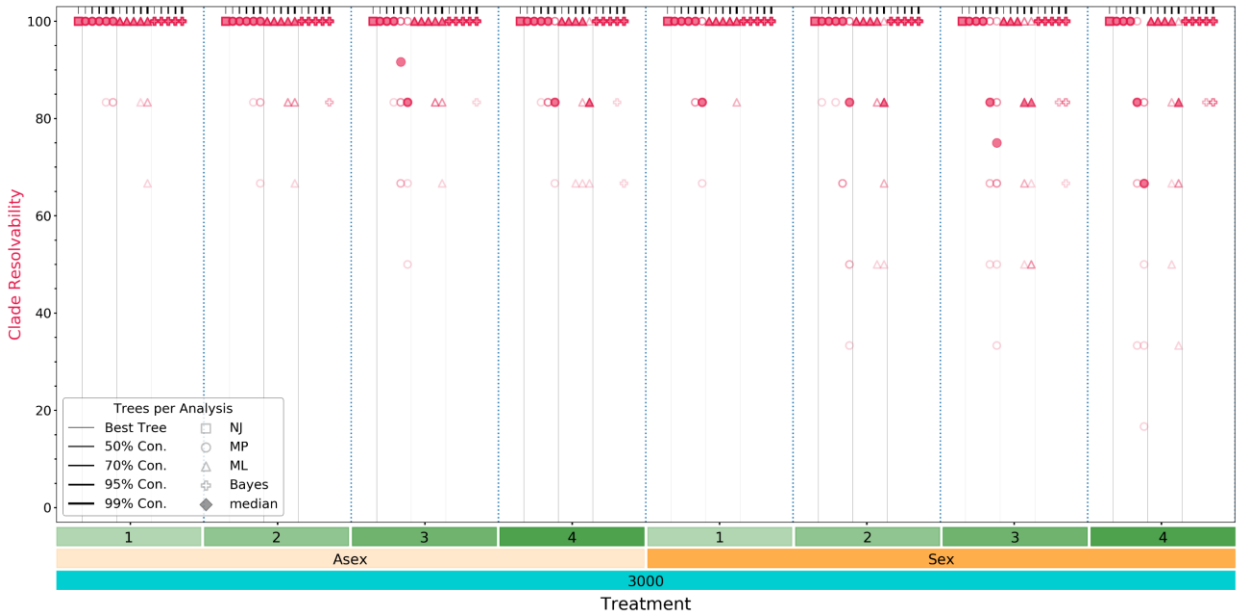


Figure 2.11. Clade resolvability for the set of treatments with directional selection and lineages evolved for equivalent generations per branch. Best and consensus trees of increasing threshold strictness are included for each analysis, except for NJ which only includes the best tree, with open symbols for individual replicates and closed for the median across replicates.

Clade accuracy shows interesting patterns among the treatments with directional selection regime 2 or 3 and/or the LSS pattern of long external branches, with the presence of recombination being a further complicating factor (Fig. 2.12). For any pair of otherwise identically conducted treatments, selection regime 2 had slightly decreased clade accuracy compared to regime 3. Topologies with equivalent generations per tree level had the greatest accuracy irrespective of recombination, as discussed previously. With comparatively short internal branches (i.e., relatively long external branches) of topology LSS, clade accuracy was greatest when directional selection was present and recombination absent, was slightly reduced when stabilizing selection was present, and was lowest under combined directional selection with sexual reproduction. While the trends are more difficult to observe with clade resolvability (Fig. 2.13), since resolvability was generally greatly reduced, all the trends highlighted with respect to clade accuracy were repeated.

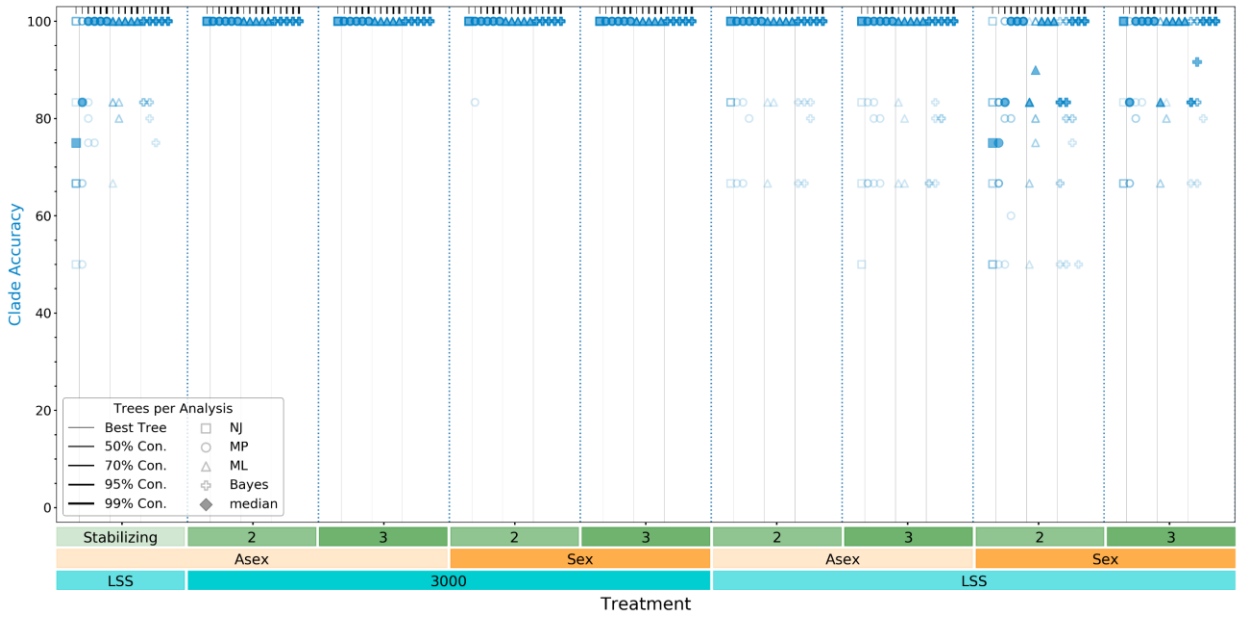


Figure 2.12. Clade accuracy for treatments with directional selection regimes 2 or 3 and/or the LSS pattern of generations per branch. Best and consensus trees of increasing threshold strictness are included for each analysis, except for NJ which only includes the best tree, with open symbols for individual replicates and closed for the median across replicates. Note that the first five treatments were shown in prior figures and are included here for comparison.

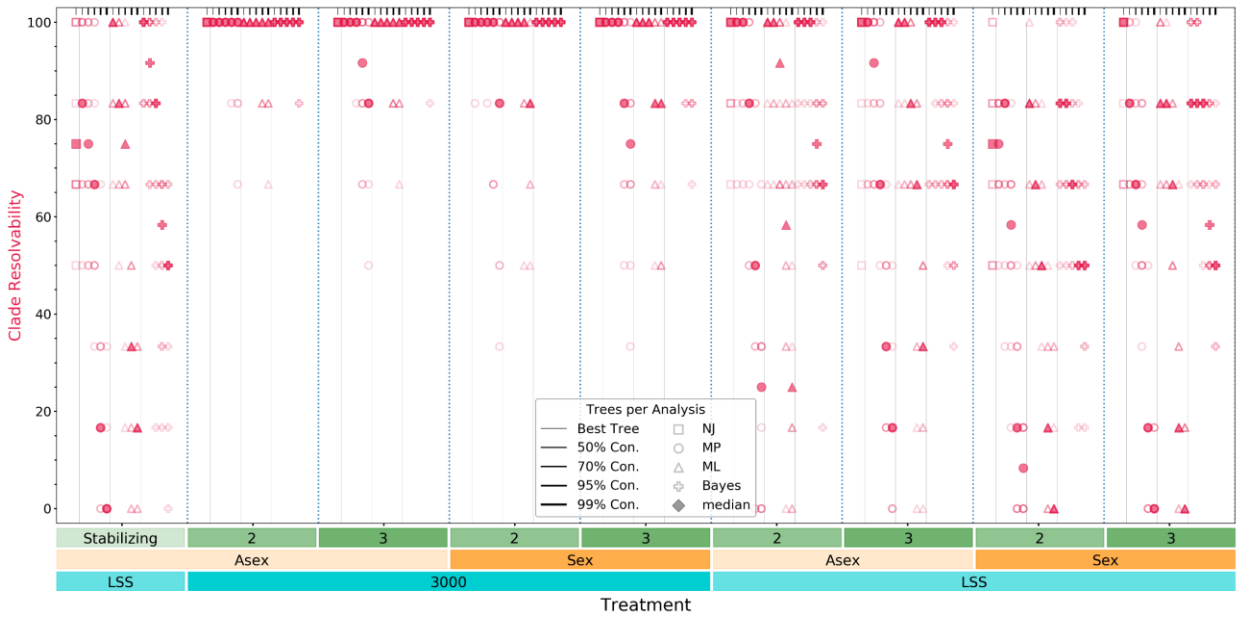


Figure 2.13. Clade resolvability for treatments with directional selection regimes 2 or 3 and/or the LSS pattern of generations per branch. Best and consensus trees of increasing threshold strictness are included for each analysis, except for NJ which only includes the best tree, with open symbols for individual replicates and closed for the median across replicates. Note that the first five treatments were shown in prior figures and are included here for comparison.

Across all treatments, the best tree reported by each analysis often exhibited very high accuracy (Fig. 2.14). Recall that since the best tree produced by each analysis lacks polytomies although may include incorrect clades instead, each false positive clade necessarily creates a false negative clade, and so for these fully-resolved trees clade accuracy is equal to clade resolvability, and thus their average as topological accuracy (as in the examples of Fig 2.03b,e). For twelve treatments, all phylogenetic analyses inferred the true tree for all replicates. An additional three treatments had a single replicate each that resulted in a tree inference with one incorrect clade under either NJ or MP only. The remaining eight treatments had multiple replicates that each had up to three incorrect clades under multiple analyses; these included the 100-generation per branch treatment starting with the pre-evolved ancestor, the three stabilizing selection eight-ingroup taxon treatments with at least one set of short internal branches (i.e., treatments LSL, LLS, and LSS), and the final four directional selection treatments with short internal branches. To simplify further comparisons, results from these eight treatments are collated as "*troublesome treatments*," with the treatments producing well-resolved best tree inferences as "*non-troublesome treatments*."

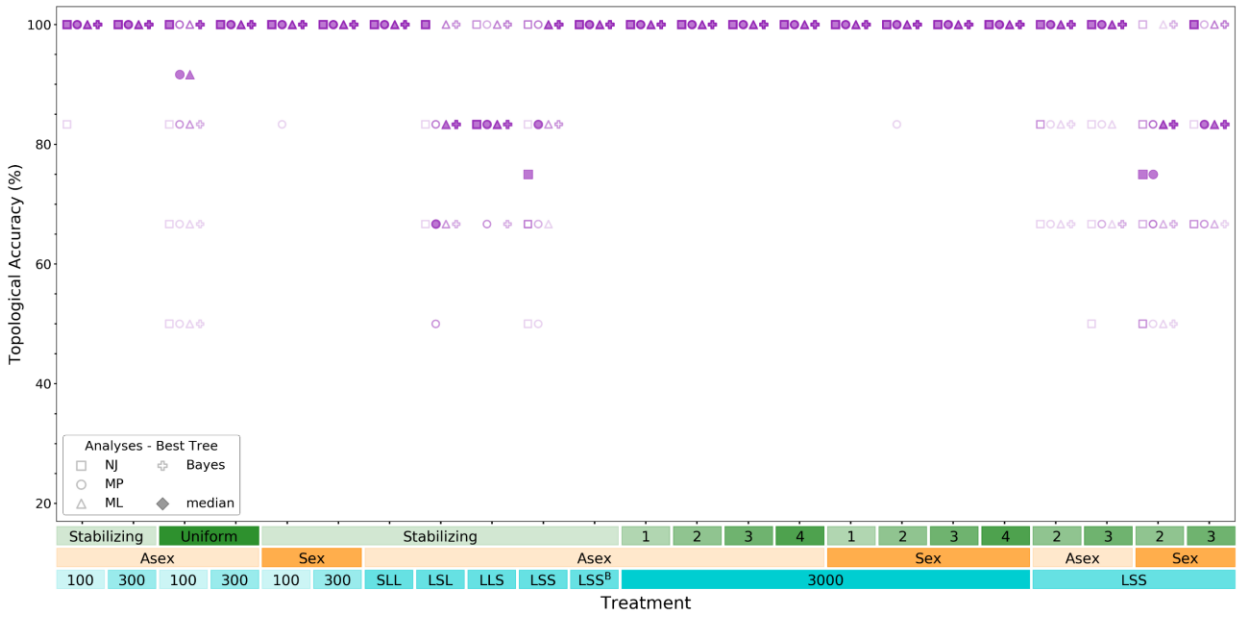


Figure 2.14. Average topological accuracy for each analysis' best tree for each treatment, with open symbols for individual replicates and closed for the median across replicates. "Troublesome treatments" are the eight treatments demonstrating considerable variation.

Both two- and four-taxon clades were responsible for decreased resolvability among the eight troublesome treatments, and there is considerable clade support variation between replicates (Fig. 2.15). While only BI analyses are presented in Figure 2.15, ML bootstrap support is similar albeit with magnified trends due to generally lower clade proportions. The clades responsible for decreased resolvability, that is, true clades appearing in low proportions of bootstrap or posterior samples, are always those denoted by short branches. For example, when clade resolvability was low in treatment LSL, it was always due to true two-taxon clades – the clades whose successful inference depended on the middle tree level (Fig. 2.15b). Likewise, treatment LLS lacked only true four-taxon clades (Fig. 2.15c), and replicates for all LSS treatments lacked two- and/or four-taxon clades (Fig. 2.15d-h). Note that in the eight-ingroup taxon topology there are twice as many two-taxon clades compared to four-taxon clades; this disparity of clades denoted by the middle versus basal tree level is responsible for the clade resolvability pattern between treatments LSL and LLS (Fig. 2.15b,c), with the latter's short branch responsible for providing support to fewer overall clades. In the case of the treatment with strong stabilizing selection and 100 generations per branch, decreased clade resolution was due to both two- and four-taxon clades (Fig. 2.15a). While previous figures also demonstrated variation among replicates for any given treatment, this variation is especially evident in these target plots (Fig. 2.15). For example, in Figure 2.15a only five replicates had one or more clades with resolution less than 50% and none included false clades, while in Figure 2.15f two replicates had clades with resolution less than 50% and each also had false clades with greater than 50% support. Note that while many additional false clades were present within the trees sampled, only those with support of 50% or more are presented, since only these contribute to decreased clade accuracy.

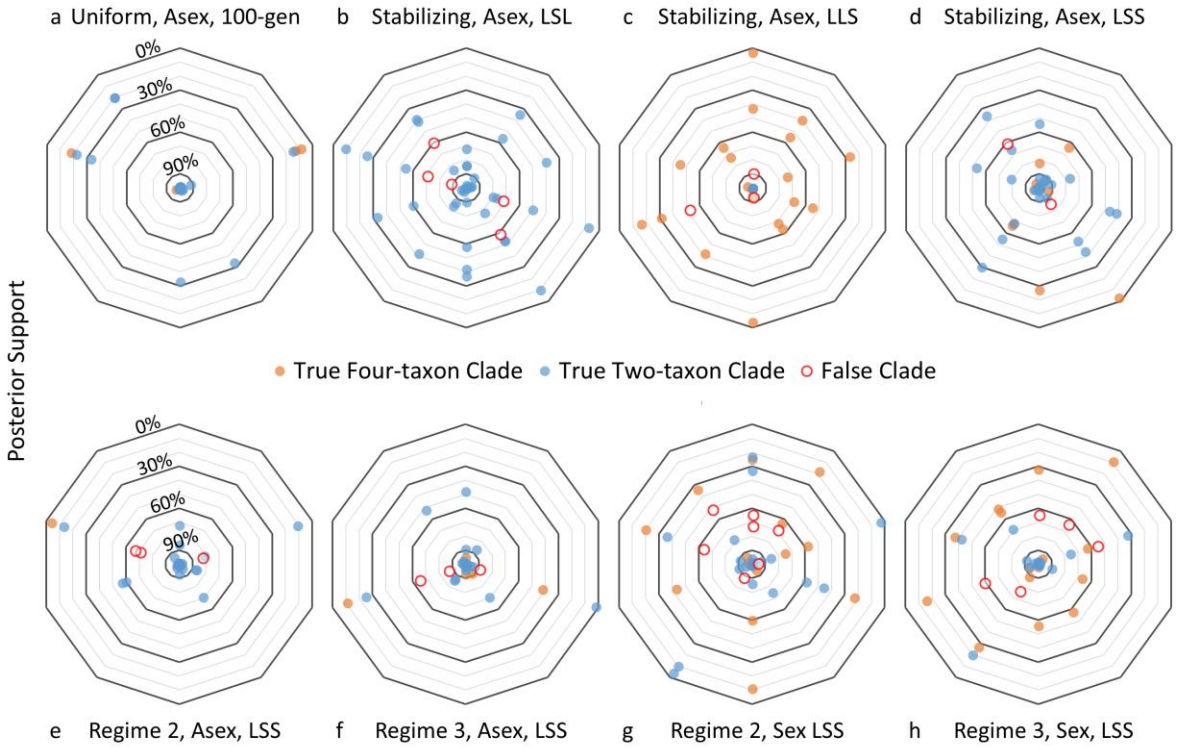


Figure 2.15. BI posterior support for true four-taxon clades (orange), true two-taxon clades (blue), and false clades (red outline) for all eight troublesome treatments. Each spoke per target plot is a replicate and each eight-ingroup taxon tree replicate had 2 four-taxon clades and 4 two-taxon clades, therefore 6 true clades are arrayed linearly from the outer to inner ring for each spoke. The thick target rings are in intervals of 30% and thin rings by 10%, with the absolute center of the target plot as 100% support. Only false clades with 50% or greater support are shown. Experimental treatment labeling is as indicated in the text.

Accuracy and consensus thresholds

The proportion of true and false clades identified by ML and BI analyses show a stark distinction between troublesome and non-troublesome treatment sets (Fig. 2.16). For both ML and BI, the non-troublesome treatments exhibited a perfect relationship around the 50% bootstrap or posterior support threshold, with no true clade represented in fewer than half the trees sampled and no false clades in greater than half (Fig. 2.16a,c). Troublesome treatments had many more, as well as more highly represented, false clades (Fig. 2.16b,d), and several infrequently produced true clades, with some even having nil support. For both troublesome and non-troublesome tree sets, BI greater proportions of clades with high support.

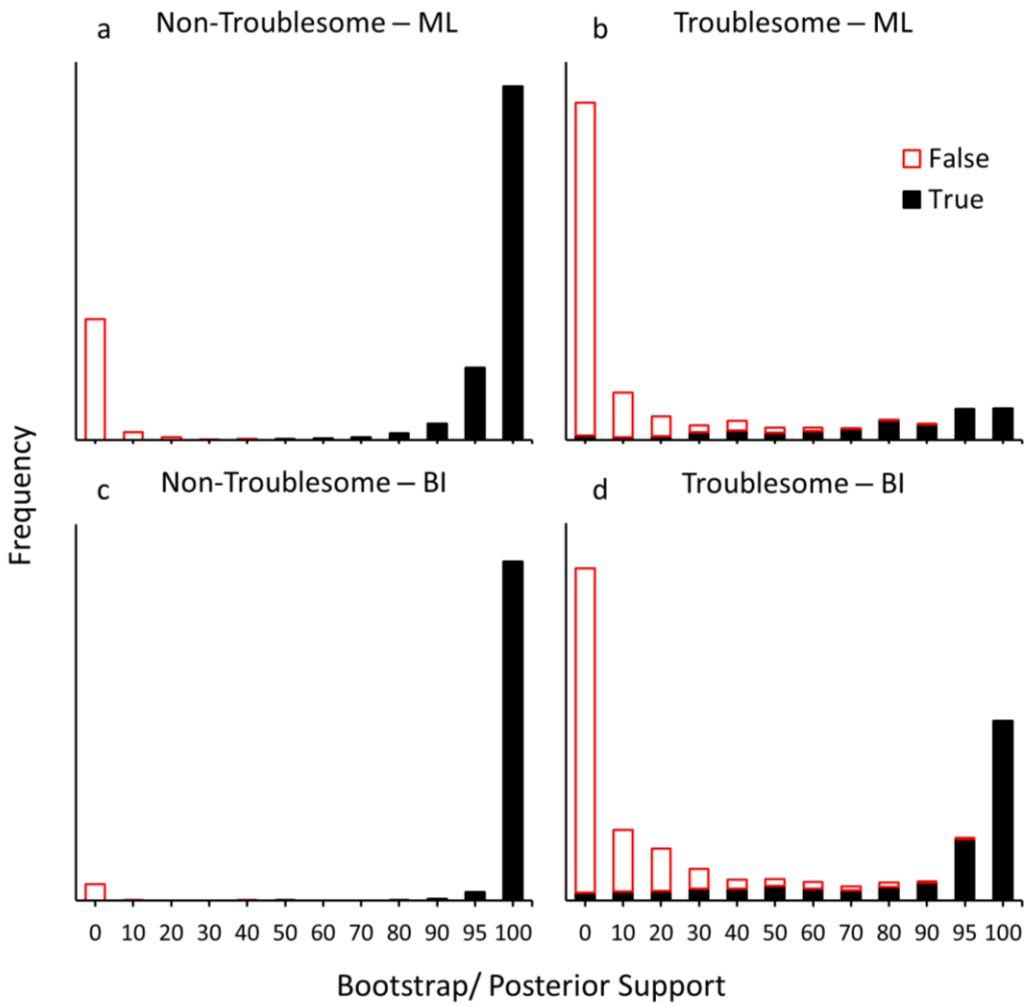


Figure 2.16. Frequency of all true and all false clades per treatment set by relative bootstrap support for ML analyses and by posterior probability support for BI analyses. Note the x-axis is not to scale, including the 95% support category. Troublesome treatments are the eight treatments demonstrating considerable variation among best trees across phylogenetic analyses; and non-troublesome treatments are the fifteen treatments demonstrating near-perfect clade accuracy among best trees across phylogenetic analyses. Treatment LSS^{B} is excluded so that only eight-ingroup taxon trees are compared.

As a measure of the probability of a clade being correct, the relative proportion of true versus false clades is compared across support values in Figure 2.17, following an analysis provided by Hillis and Bull (1993). These researchers used jackknifing followed by bootstrapping to determine MP bootstrap proportions for the restriction site dataset of Hillis et al. (1992). Hillis and Bull (1993) observed that bootstrap proportions are lower than the probability of being correct for all proportions above 35%, and that proportions 70% or greater indicate a

very high probability (>95%) that the clade is real. For ML analyses with the troublesome treatments, bootstrap proportions greater than 30% were conservative in accuracy, and very highly so for proportions greater than 70%. BI analyses for the troublesome treatments produced posterior proportions conservative with respect to accuracy for all proportions greater than 30% except for 80%, which was slightly liberal in its representation of clade accuracy. BI was only more conservative than ML for support values of 40% and 50%. Within the range of 30-50%, BI posterior support does not closely approximate clade accuracy, with all other support values having a posterior probability ±5% of the probability the clade is true.

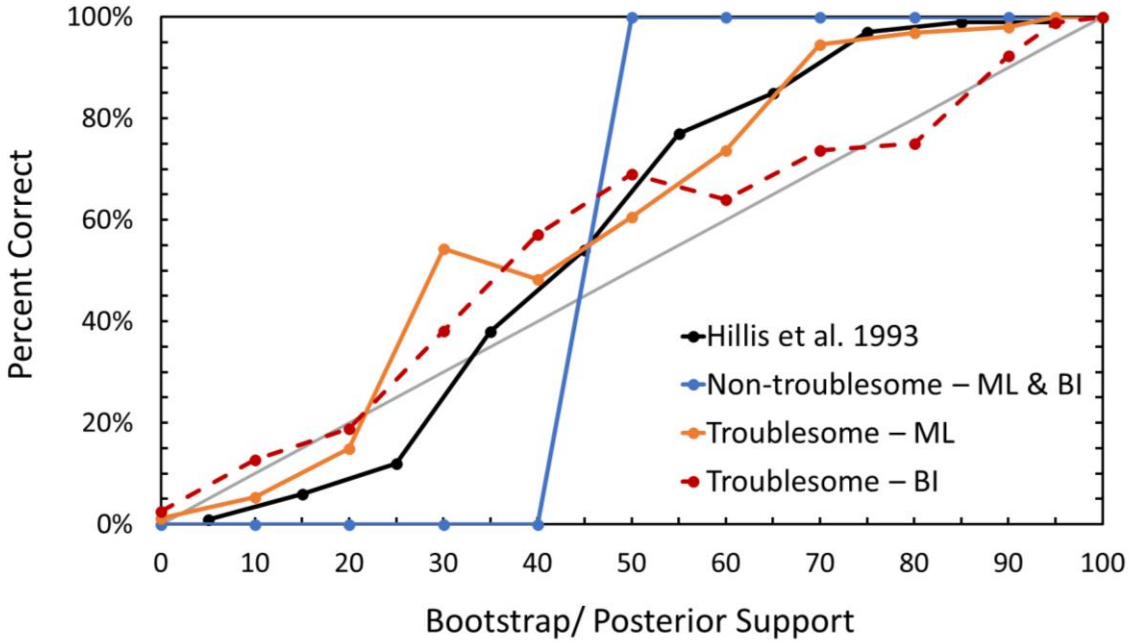


Figure 2.17. Relationship between clade accuracy as the percent of correct clades for values of bootstrap support for ML analyses and posterior probability support for BI analyses. Results include data from Hillis and Bull (1993) as estimated from their Figure 4a (black), data from Avidian evolution non-troublesome treatments for both ML and BI analyses (blue, identical relationship), and data from Avidian evolution troublesome treatments for ML analyses (orange) and BI analyses (red dashed). The thin grey line is the one-to-one accuracy-to-support relationship; values above the line are conservative as being an underestimation of accuracy and values below are liberal as being an overestimation of accuracy. Troublesome treatments are the eight treatments demonstrating considerable variation among best trees across phylogenetic analyses; and non-troublesome treatments are the fifteen treatments demonstrating near-perfect clade accuracy among best trees across phylogenetic analyses. Treatment LSS^{B} is excluded so that only eight-ingroup taxon trees are compared.

Clade accuracy and clade resolvability, as well as their averaged score, topological accuracy, are compared within a whole-tree context in Figures 2.18 and 2.19. Results from each

phylogenetic analysis and consensus method are shown for the fifteen non-troublesome treatments (including treatment LSS^{B}) in Fig. 2.18 and for the eight troublesome treatments in Fig. 2.19 as the proportion of trees that attained thresholds for clade accuracy (subplot a), clade resolvability (subplot b), and average topological accuracy (subplot c). These measures provide overall evaluations of the information provided in the trees resulting from the Avidian evolution treatments presented. Consensus methods presented include the thresholds previously evaluated as well as the MRe result. Note that ten replicates were performed for each treatment and that each phylogenetic analysis and consensus method was performed for all such treatments and replicates. Therefore, the data shown within each column of these plots represent 150 trees and 80 trees, respectively, in Figures 2.18 and 2.19.

Non-troublesome treatment trees had high clade accuracy yet were often not fully-resolved and therefore had low overall accuracy for more strict consensus support values (Fig. 2.18). Clade accuracy was perfect (i.e., meeting the 100% threshold) for all consensus trees for each analysis and for best trees for ML and BI analyses (Fig. 2.18a); of course, these were the criteria for collating these treatments as "non-troublesome." Clade accuracy was greater than 95% for NJ and the MP best trees, although for the latter an equally parsimonious tree was consistently the true tree. A 70% consensus support threshold maintained very high clade resolvability (Fig. 2.18b), with at least 95% of trees being fully resolved for each analysis (i.e., meeting the 100% CR threshold). While MP performed slightly better than ML for thresholds of 95% and 99%, BI maintained a much greater clade resolvability at these high support thresholds. When weighting clade accuracy and clade resolvability equally as average topological accuracy (Fig. 2.18c), BI produced a large proportion of accurate trees even at very strict support thresholds.

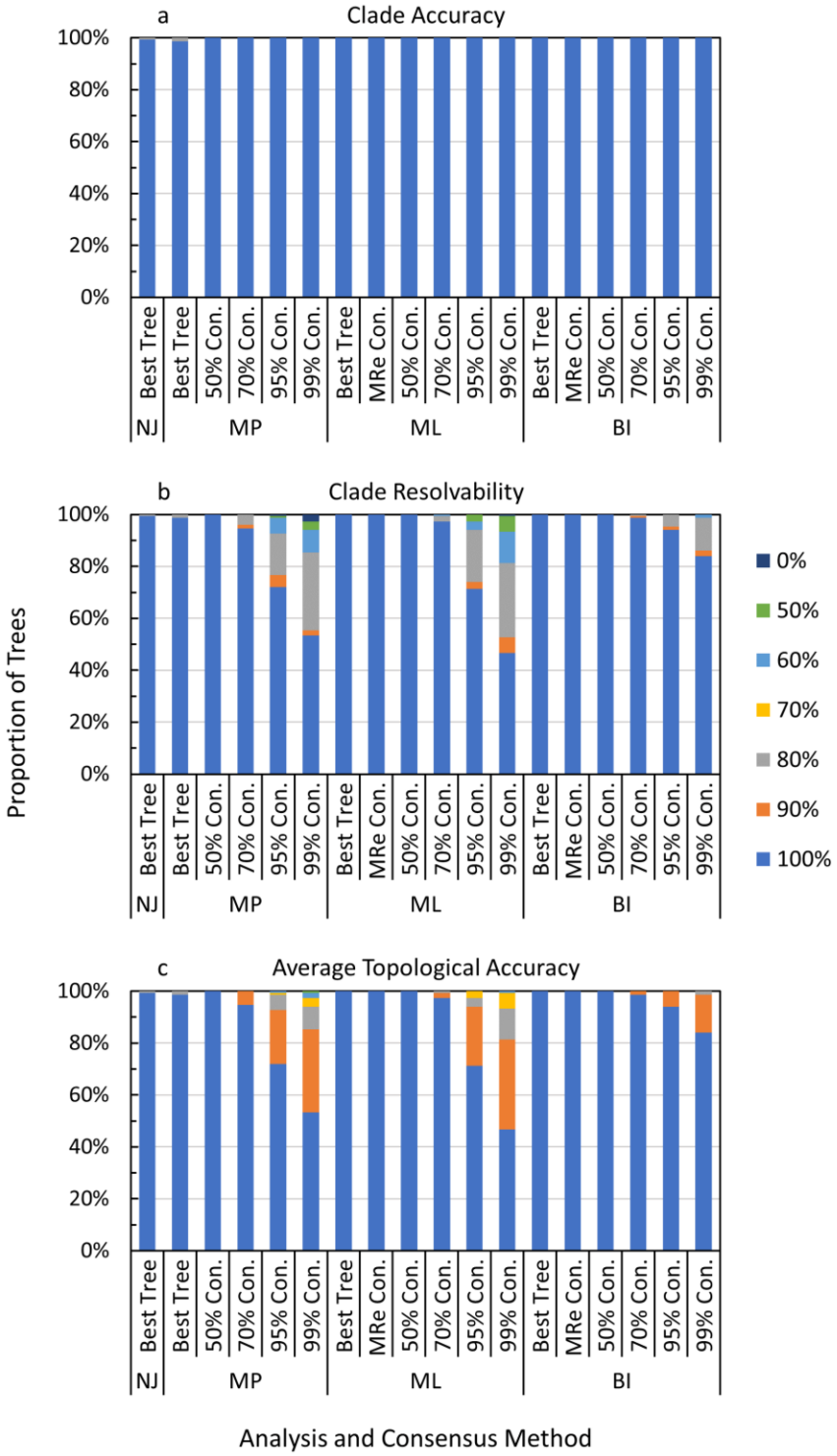


Figure 2.18. Proportion of trees across all non-troublesome treatments that met or exceeded clade accuracy (a), clade resolvability (b), and average topological accuracy (c) thresholds for each analysis and consensus method evaluated. Non-troublesome treatments are the fifteen treatments demonstrating near-perfect clade accuracy among best trees across phylogenetic analyses.

Trees from troublesome treatments had high clade accuracy but were poorly resolved, with low average topological accuracy (Fig. 2.19). Clade accuracy was perfect for only 50% of the best trees produced by each analysis, and it was at least 80% for nearly 80% of the best trees produced by each analysis, with MP performing a bit worse (Fig. 2.19a). As the strictness of the majority rule consensus threshold increased, clade accuracy increased at different rates among analyses, with ML approaching 95% of trees as being perfect at the 70% consensus threshold and both MP and BI requiring a threshold of 95%. For trees of increasing consensus threshold strictness, clade resolvability decreased substantially, from approximately half the trees being fully resolved to less than 5% of trees (Fig. 2.19b). MP and ML reached this low resolvability at a 95% consensus threshold while BI reached it only at 99%. Across consensus thresholds, ML consistently produced better-resolved trees than MP, but less so than BI. This trend of BI maintaining greater clade resolvability stands out when examining trees with at least a 60% clade resolvability (light blue, Fig. 2.19b). For 99% consensus trees, MP and ML produced only about 15% of trees with at least 60% clade resolvability while BI produced 45%. This increased resolvability caused BI consensus trees to have greater overall accuracy, and ML consensus trees median accuracy compared to MP (Fig. 2.19c). For MP and BI, 50% majority rule trees had slightly lower overall accuracy compared to best trees by having greater clade accuracy and slightly lesser resolvability, although for ML over 15% more trees had reduced overall accuracy. MRe consensus trees for ML and BI had slightly reduced clade accuracy or clade resolvability, with minimal increase in the other metric, if any; therefore, these trees tended to have near-equivalent or reduced overall accuracy compared to an analysis' best tree.

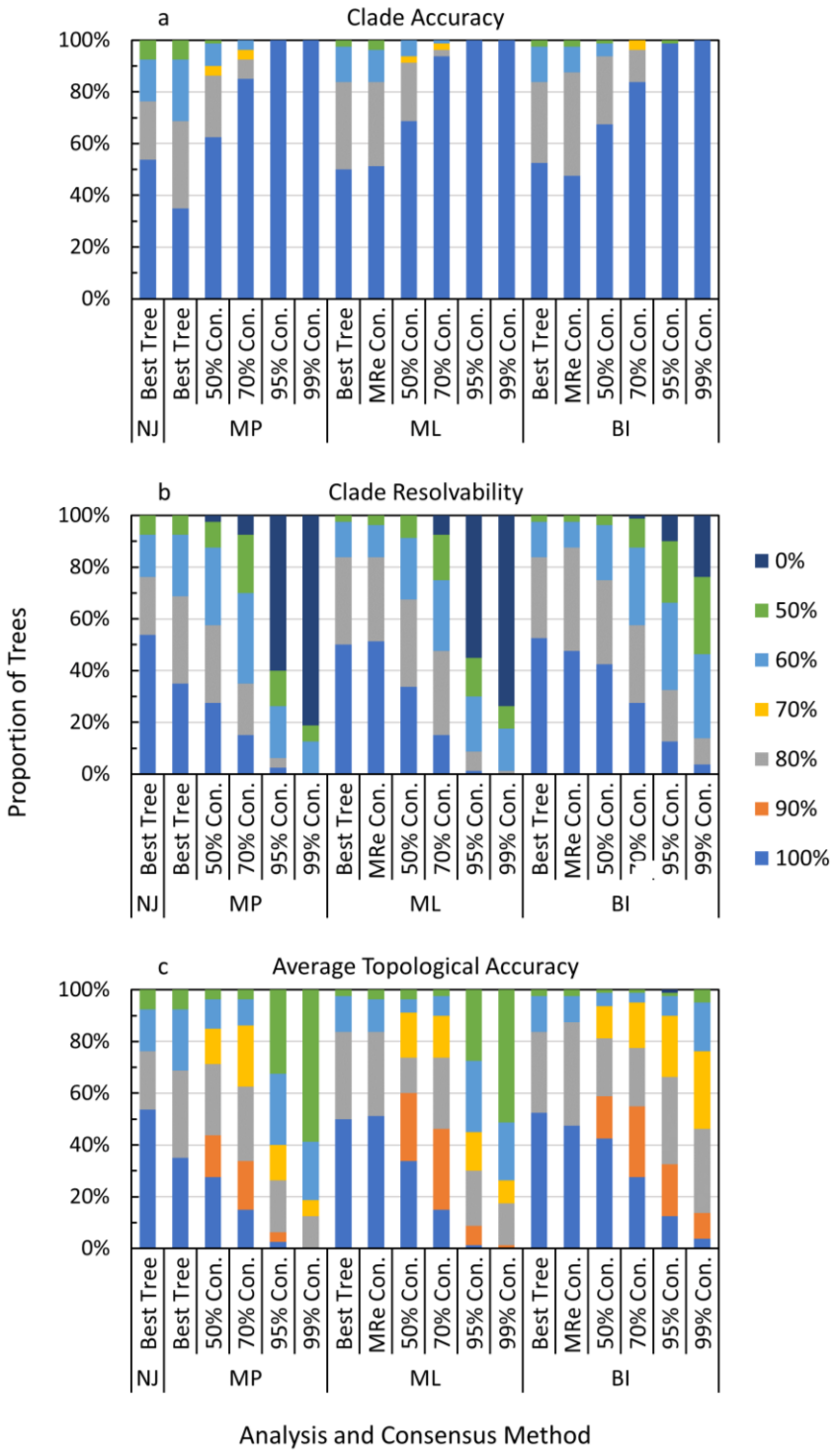


Figure 2.19. Proportion of trees across all troublesome treatments that met or exceeded clade accuracy (a), clade resolvability (b), and average topological accuracy (c) thresholds for each analysis and consensus method evaluated. Troublesome treatments are the eight treatments demonstrating considerable variation among best trees across phylogenetic analyses.

Discussion

Treatment comparisons

The basic pair of treatments exhibited similar molecular evolution and comparable phylogenetic inference accuracy in Avida compared to that in the T7 phage research. These 100- and 300-generation per branch treatments under asexual evolution and stabilizing selection with the naïve ancestor favorably compared to sets of molecular T7 data exhibiting distinct rates of evolution. Hillis et al. (1994), using a DNA dataset, observed 63 variable sites and a total of 69 substitutions across internal and external ingroup branches. In comparison, the 100-generation per branch treatment using Avida resulted in a median of 85 variable sites (Fig. 2.04) and a total of approximately 74 substitutions, as extrapolated from inferred branch lengths across ML and BI best trees (Fig. 2.05). The NJ analysis Hillis et al. (1994) performed had a tree with a single clade as incorrect, which was found for one of the ten NJ replicates, with the others having perfect accuracy (Fig. 2.06). Their MP analysis resulted in the true tree and one other as equally parsimonious tree, and this result also occurred for one replicate here, with the other replicates identifying only the true tree as most parsimonious. Whereas the T7 researchers found ML as being incorrect by one clade, all replicates of ML and BI for the 100-generation per branch treatment produced trees with perfect clade accuracy. Hillis et al. (1994) performed nonparametric bootstrapping and reported a single clade resolvability score for each analysis, whereas here I report clade resolvability separately for consensus trees of increasing threshold strictness. They found that MP maintained much higher resolvability than ML, with NJ slightly outperforming them all. The consensus trees analyzed for the 100-generation per branch treatment confirm these results (Fig. 2.07), although a bootstrap analysis with NJ was not performed. Consensus thresholds of 70% and greater resulted in one or more replicates producing trees with at least one polytomy, and MP and BI were not as greatly affected by increasing threshold strictness as was ML. Hillis et al. (1992), using a restriction site dataset, observed 202 variable sites and a total of 220 substitutions across internal and external ingroup branches, which is comparable to the 300-generation treatment which had a median of 186

variable sites and approximately 236 substitutions. Hillis et al. (1992) found that their dataset produced correct tree inferences under each method evaluated, and this result is reproduced here, with each of these analyses and consensus thresholds having 100% clade accuracy and resolvability. Comparing the Avidian evolution treatments, it is reasonable to suspect that the three-fold increased evolution between lineage division allowed greater fixations to occur, providing greater support for clade relationships, and therefore resulting in greater variable and parsimony informative sites and thus increased topological accuracy. These treatments approximated the molecular evolution characteristics of dataset size and degree of evolutionary change and found similar phylogenetic success as the T7 experimental phylogenetics research. This demonstrates that Avida was successfully used to reimplement prior experimental phylogenetics work using biological organisms.

Three of the four remaining treatments conducted under stabilizing selection with lineages evolved for an equivalent amount of time within each history produced results like the prior treatments, with the remaining treatment being phylogenetically problematic. When an ancestor pre-adapted to the environment was used, the extent of molecular evolution decreased dramatically. Presumably, strong stabilizing selection in maintaining ancestral adaptation to the selective environment caused many fewer fixations to occur, resulting in reduced variable and parsimony informative sites (Fig. 2.04), shorter branch lengths (Fig. 2.05), and thus lower clade accuracy and resolvability (Figs. 2.06-2.07). As with the basic pair of stabilizing selection treatments, increased lineage evolution between cladogenesis allowed greater phylogenetic information and inferred topological accuracy for the 300-generation treatments. A notable trend was that clade resolvability was improved by the addition of sex to the 100-generation treatments. This was a curious result, as recombination in these treatments appears to increase clade resolvability while not greatly increasing the number of variable or informative sites. Together the set of six stabilizing selection treatments with equivalent lineage evolution per branch demonstrates that phylogenetically informative sequence variation and

thus topological accuracy is reduced due to relatively stronger stabilizing selection and relatively reduced lineage evolution between cladogenic division, as would be expected.

Five treatments were designed to evaluate the effects of differing branch lengths. Evolution along internal branches can produce evidence of the shared ancestry among latter lineages via the production of synapomorphies. This is especially likely to occur if the population has sufficient opportunity to acquire substitutions, i.e., on relatively longer branches. On the other hand, evolution along external branches can only produce such evidence due to lineage sorting (i.e., the segregation of variant characters following cladogenesis). However, evolution along both internal and external branches may produce homoplasy, which is more likely to occur along relatively longer branches (Rokas and Carroll, 2006). Of course, both internal and external branches can also negate evidence of shared ancestry when substitutions occur at sites that previously exhibited such evidence. Therefore, internal branches are more important in positively contributing to phylogenetic inference and this contribution is relative to their length, and external branches are likely to negatively contribute with increased length. For example, Hang et al. (2003) used digital evolution with Avida to demonstrate that relatively longer internal branches improved phylogenetic inference through the generation of synapomorphic variation. Three Avidian evolution treatments presented here evaluated this through the placement of a single short ("S," 300-generation) set of branches per tree level among the remainder of long ("L," 3,000-generation) sets of branches: SLL had eight external branches of 300 generations each, LSL had four internal branches of 300 generations, and LLS had two basal-most internal branches of 300 generations. Felsenstein (1978) numerically demonstrated that long external branches can "attract" one another through the production of homoplasious evolution swamping out positively informative cladogenic signal. The fourth Avidian evolution treatment, LSS, had only external long branches to test this long branch attraction effect. Graybeal (1994) used simulations to show that breaking up such long external branches by adding targeted taxa resolves those

difficulties. Finally, Avidian evolution treatment LSS^B did just that, resulting in a 32-taxon asymmetrical history.

This set of treatments, conducted under weak stabilizing selection with differing numbers of generations per tree level, supported this prior research. Since the primary distinction between parsimony informative sites and variable sites is that the latter additionally includes autapomorphic variation, the observed pattern of informative sites in Figure 2.04 for these treatments makes sense. Treatment LSS would be expected to show the greatest proportion of such sites due to greater evolution along the long external branches. Although treatments LSL and LLS have the same long external branches, they also have a set of long internal branches, which provided a greater opportunity for parsimony informative variation among the resulting taxa. Treatment SLL did not have long external branches during which substantial evolution could take place. And treatment LSS^B had longer external branches than SLL, although it also had four times as many taxa, allowing greater opportunity for informative yet misleading variation to occur due to parallelism and convergence. NJ underpredicting long branches for these and other treatments is as expected since the algorithm's distance metric is sensitive to deviations from its model and generally undercounts rates of change (Tateno et al., 1994). As expected, clade accuracy was perfect for all analyses and consensus thresholds when all internal branches were long and when external long branches were disrupted by increased taxa breaking up long branches yet was reduced with treatments of long external branches and at least one set of short internal branches (Fig. 2.08). When clade resolvability was decreased, the responsible clades were consistently those denoted by short branches (Fig. 2.15b-d). Overall, this set of treatments produced the phylogenetic trends as predicted by theory, simulation, and prior digital evolution research.

The eight treatments with directional selection and long branches throughout the tree support prior experimental phylogenetics research on natural selection using digital evolution. While it is undetermined whether selection aids or hurts phylogenetic inference generally, experimental phylogenetics research using biological systems tends to show that selection

produces homoplasy (Bull et al., 1997; Cunningham et al., 1997; Fares et al., 1998), although see Leitner et al. (1996), while in digital systems it primarily produces greater synapomorphies (Hagstrom et al., 2004; Hang et al., 2007, 2003). Four different selective regimes simultaneously altered the strength of selection and the possibility of selection producing parallelism or convergent homoplasious evolution, with the strength of selection decreasing from regimes 1 through 4 and environmental diversity among branches increasing from regimes 1 through 4 (Fig. 2.02). For example, in regime 1 all branches experienced very strong selection for the same tasks, and in regime 4 each branch experienced relatively weaker selection for distinct combinations of tasks. Recombination is expected to increase the efficiency of natural selection by fixing beneficial and removing deleterious alleles, so its presence is expected to magnify the effects of selection. The pattern of branch lengths increasing from regimes 1 to 4 for ML and BI inferred best trees (Fig. 2.05) is evidence that selection for new tasks during later periods of lineage evolution was driving greater evolution. Since NJ did not show this pattern, this suggests greater deviation from its distance metric's model of evolution for the diversifying selection treatments. Since branches were reduced in length from that observed for stabilizing selection treatments that also had branches of 3,000 generations (e.g., long external branch treatments in treatment LLS), selection presumably caused fewer fixations to occur. Other than increased selection fixing beneficial alleles and removing deleterious alleles, presumably stronger selective sweeps were occurring in stronger selection regimes, further causing a decrease in evolutionary change. Internal branch lengths appeared inflated without recombination and external branch lengths inflated with recombination because recombination treatments included the seeding of new lineages with the entire population instead of the most abundant genotype. This should cause lineage sorting dynamics, resulting in fewer overall substitutions, and delaying some fixations until later in the evolutionary history. The trend observed across all 23 treatments of ML inferring slightly shorter branch lengths than BI was especially prevalent for internal branch lengths in these treatments and is best exemplified by the selection regime 3 treatments with or without recombination. While it is not clear what is

causing this near-systematic difference, perhaps either ML or BI is more sensitive to deviations from the model of evolution than the other, although not nearly as sensitive as NJ's distance metric. Comparing the set of treatments with recombination to those without, recombination appears to decrease clade resolvability (Fig. 2.11). This is the opposite trend from that observed for 100-generation per branch treatments (Fig. 2.07). Overall, it appears that stronger selection in these treatments increases clade resolvability, with homoplasious evolution presumably not being as significant of a concern in this digital evolution system.

The final four treatments evaluated the combinatorial effects of selection, recombination, and varying branch lengths in the long branch attraction (LSS) design. Selection and a lack of recombination increased both clade accuracy and resolvability compared to the standard LSS treatment (Figs 2.12 and 2.13). Whereas selection in the presence of recombination decreased both clade accuracy and resolvability compared to the standard LSS treatment. With either recombination present or absent, both clade accuracy and resolvability slightly increased in the less selective environment, regime 3, countering the trend observed for the 3,000-generation per branch treatments. As with the stabilizing selection LSS treatment, clade resolvability was diminished by both two- and four-taxa clades (Fig. 2.15), owing to the presence of internal branches needing to provide phylogenetic support for clades of both tree levels. Overall, complex combinatorial dynamics were observed with these treatments.

Accuracy and consensus thresholds

Arguably, the most impactful individual result from the T7 work is the connection of MP bootstrap support values with clade accuracy, with the threshold of 70% indicating very high (> 95%) clade accuracy (Hillis and Bull, 1993). The expansive set of experimental phylogenetics data presented here was used to revisit this clade-level accuracy result within the context of ML and BI, and to further consider it within a tree-level context.

The relative proportions of true and false clades were directly compared across support values as a measure of the probability of a clade being correct (Fig. 2.17). Avidian evolution ML analyses with the troublesome treatments supported the results of Hillis and Bull (1993), with

bootstrap proportions greater than 30% being conservative in accuracy, and very highly so for proportions greater than 70%. BI analyses for the troublesome treatments provide strong evidence that posterior probabilities are very close estimates of clades accuracy, confirming analyses of simulation studies by demonstrating that, at least for these "troublesome" experimental conditions, BI posterior support is a closer estimate to clade accuracy than is ML bootstrap support for support values greater than 60%. Curiously, the range of 30-50% in which BI posterior support fails to closely track clade accuracy does not appear to be especially under-sampled, and so it is unclear why these median values are overly conservative. While a few simulation studies suggest that BI posterior probabilities may be overly liberal estimates, only two support values (20% and 80%) were indeed liberal, and neither were overly so. This result suggests that BI posterior probabilities are a much better reflection of clade accuracy than bootstrap support, at least for most support values of interest to systematists. It also suggests that the commonplace 100% posterior support value may be regarded as a significant indication of clade accuracy rather than as an overly liberal estimate.

While the individual accuracy of a clade is important, clades are rarely just evaluated alone but also in the context of one another in a tree, and therefore it is important to examine clade accuracy, as well as resolvability, in a whole-tree context (Figs. 2.18-16). Greater assurance of clade accuracy (for BI, anyway, based on results in Fig. 2.17) using more strict consensus support thresholds entailed a stark trade-off with clade resolvability even for the non-troublesome treatments (Fig 2.18a-b), although it had extreme consequences in troublesome treatments (Fig 2.19a-b). For both sets of treatments, BI maintained greater clade resolvability at stricter thresholds and therefore produced the most overall accurate trees even at very strict consensus thresholds. This increased clade resolvability caused BI consensus trees to have much greater average topological accuracy than MP and ML. A systematist might use the MRe method to produce a fully-resolved tree that considers bootstrap consensus support information. The data presented here suggest that this would be misguided, as it sacrifices clade accuracy and/or clade resolvability with little gain.

Across both clade-level and tree-level measures, a ML 70% consensus support tree is approximately as accurate as a BI 95% consensus support tree. As shown in Fig. 2.17, since ML bootstrap support is so conservative, at the 70% threshold clades have approximately a 95% probability of being correct. This compares to the much closer measure of accuracy provided by BI posterior probability, with a 95% posterior clade support having approximately a 95% probability of being correct. For conditions in which phylogenetics should perform well, as with the non-troublesome treatments, a 70% consensus ML tree may be approximately equally expected to exhibit perfect clade accuracy, clade resolvability, and therefore average topological accuracy as a 95% posterior support BI tree (Fig. 2.17). And for more difficult conditions, these trees are similar yet with a few differences; specifically, a 70% consensus ML tree may have a slightly lower chance of having perfect clade accuracy, a slightly higher chance of being fully-resolved, and overall a slightly greater chance of having perfect topological accuracy than a 95% posterior support BI tree (Fig. 2.19).

Conclusions

As the most ambitious examination of phylogenetic accuracy in an experimental system to date, this work demonstrates the use of digital evolution in experimental phylogenetics as a powerful tool for the evaluation of phylogenetic inference.

This work has shown that Avida can be used to approximate the basic molecular evolution characteristics and phylogenetic inference results obtained in the foundational T7 experimental phylogenetics research of Hillis et al. (1992) and Bull et al. (1993), and supports the bootstrap-accuracy conclusion of Hillis et al. (1994). Avida can also be used to expand upon these earlier experiments by evaluating a range of designs expected to cause phylogenetic inference complication, thereby demonstrating the greater utility and generality of the digital evolution approach over biological systems for experimental phylogenetics.

A fundamental phylogenetics phenomenon has been demonstrated using this system across many treatments – that greater evolution aids phylogenetic inference except in

conjunction with lesser evolution among internal branches; although, uninvestigated here, too much evolution along branches will eventually swamp out positively informative signal.

Directional selection has been shown to aid phylogenetic inference in this system, and does so the stronger selection occurs, although directional selection in combination with recombination and differing extents of lineage evolution produces conflicting results. Strong stabilizing selection in this system can hurt tree accuracy if it sufficiently reduces the magnitude of evolutionary change. It is not clear what the effects of sexual recombination, as implemented in Avida, are on phylogenetic inference, as it appears to increase clade resolvability in some circumstances and lower both clade resolvability and accuracy in others.

This research suggests several general recommendations for systematists. BI posterior clade support probabilities are very close estimates of clade accuracy at least throughout the range of 60–100%, inclusive. Bootstrap support values in ML analyses are highly conservative measures of clade accuracy for values of 70% and greater. Since higher consensus thresholds fail to substantially improve accuracy, although they may drastically reduce clade resolvability, ML 70% bootstrap support values represent an ideal trade-off between accuracy and resolvability. BI maintains greater clade resolvability at greater consensus support thresholds than does MP or ML. A 50% majority rule consensus tree provides a fair trade-off between clade accuracy and resolvability, having potentially greater clade accuracy than an analysis's best tree without sacrificing too much resolvability for most analyses. MRe consensus trees are not useful, as they are either equal to or worse than an analysis's best tree in terms of clade accuracy, resolvability, or both. If one wishes to have a fully-resolved tree, then the best tree resulting from the analysis should be used instead of the MRe tree. Under either phylogenetically facile or challenging circumstances, ML 70% consensus trees are approximately equivalent to BI 95% consensus trees for both clade-level and tree-level measures of accuracy.

Finally, there are a few outstanding questions that were not addressed in this study. Overall, the effects of natural selection and whether its phylogenetic inference effects are

different in biological versus digital systems stands to be further investigated. Since treatments reported here provided conflicting results, the influence of recombination (as implemented in Avida) on phylogenetic inference, separately and together with other factors, remains to be evaluated, in addition to its underlying mechanistic effects. The molecular genetics occurring within these Avidian populations, especially instruction frequencies and rates of change, was surely different from the Poisson model of evolution used for these analyses. Deviations from this model presumably accounted for differences in branch length estimation between analyses and potentially contributed to differences in clade support. The relative robustness between ML and BI to deviations from its model of evolution should be further explored. A characterization of the Avidian molecular evolution occurring in these treatments and the effects of deviation from its model of evolution might provide insight into differences between analyses. While molecular genetics mechanisms, such as selective sweeps and lineage sorting, have been suggested here as likely explanations for certain results, a characterization of such patterns in these evolutionary contexts remains to be shown. Finally, it is unclear whether trees have a higher likelihood of being accurate if multiple phylogenetic analyses, e.g., ML and BI, produce identical topologies, or, in the case of polytomies, at least produce non-conflicting topologies. If not, then these results suggest that BI analyses with posterior probability tree support of 95% or greater should be used by systematists desiring an assurance of high clade accuracy and reasonably well-resolved trees.

CHAPTER 3:

Digital Evolution Addresses Intractable Research Questions in Phylogenetics

Introduction

The lack of experimental phylogenetics research following the revolutionary work of Hillis et al. (1992) in constructing and evaluating a known biological evolutionary history has been remarkable. After an initial flurry then trickle of related research, critiques against this means of evaluating phylogenetic accuracy seem to have brought the field's extinction (Oakley, 2009). In Chapter 1, I argued that the experimental phylogenetics approach had so far suffered from an imbalance of utility, realism, and generality compared with computational simulations due to the historical use of biological study systems. Instead, digital evolution may preserve the greater biological realism found in experimentally generated evolutionary histories while not sufficiently reducing the utility in time, resources, and expertise and not sufficiently reducing the generality or universal applicability to other systems compared to simulations. In this manner, digital evolution may fill the current void between computational simulations and biological experimental evolution for the evaluation of phylogenetic methodologies.

In Chapter 2, I presented the digital evolution system Avida as a suitable system for experimental phylogenetics. I did so by demonstrating the correspondence between prior research using biological organisms (Bull et al., 1993; Hillis et al., 1994, 1992; Hillis and Bull, 1993) with digital evolution experiments designed to produce similar results. By extending this work to evaluate a greater range of predicted phylogenetically troublesome conditions, I then further demonstrated the utility and generality of the digital approach to experimental phylogenetics. Those results provided a few outstanding areas for investigation, including the impact of natural selection, sexual recombination, and deviations from the model of evolution on phylogenetic accuracy—both clade support and branch length inference—and a lack of a

detailed understanding regarding the molecular evolutionary circumstances of such experimental treatments.

The work presented in this chapter addresses those research questions: Can experimental evolution conditions in Avida be configured to closely approximate the biological reality of molecular evolutionary dynamics? And if so, does natural selection aid phylogenetic inference, as suggested in Chapter 2? Does sexual reproduction always aid this, or are its effects not easily predictable, as suggested in Chapter 2? Finally, are deviations from the model of molecular evolution used in a phylogenetic analysis highly impactful on phylogenetic accuracy or are deviations well-tolerated?

With respect to the impact of selection and sexual recombination, the experimental treatments investigated here varied in their potential for neutral evolution or natural selection to occur and whether asexual or sexual reproduction occurred. And to even further differentiate the molecular evolutionary trends as summarized in models of evolution, experiments were initiated with either a naïve or pre-adapted ancestor genotype. A fully factorial design was implemented to investigate the effects of all eight combinations of these experimental treatments on the phylogenetic accuracy, both topological and branch length accuracy, of Avidian evolutionary histories. Deviations from the model of evolution were investigated by comparing phylogenetic analyses using the Poisson model of molecular evolution to empirical substitution models created from the detailed tracking of substitutions that occurred across experimental replicates. For all experiments, experimental evolution conditions were set to approximate biological reality, and this too was evaluated using population genetic analyses to determine whether evolution proceeded as expected.

The work presented here demonstrates the greater utility and generality of digital evolution over biological systems for evaluating phylogenetic inference. For example, to create the eight empirical models of molecular evolution, substitutions across a total of 1,637,600,000 generations of evolution were tracked, and to characterize the population genetic dynamics of this evolution, the identified 3,291,266 substitutions were evaluated for their fitness effect

throughout each's population frequency trajectory. The collection and characterization of similar molecular evolution data in biological systems would be prohibitively burdensome if not impossible. Additionally, the design of these experiments strove for biological realism, for which analyses conducted here provide more so than those in chapter 2. For example, the population sequence diversity, the distribution of mutational effects, and the extent of evolution per lineage between cladogenic events were well within the range observed in biological systems (e.g., Eyre-Walker and Keightley, 2007; Li, 1997; Pin et al., 2001; Simmons, 2012).

The impacts on phylogenetic inference by model divergence, selection, recombination, and complex population dynamics have been underexplored in phylogenetics research. Digital evolution has potential utility in exploring such phylogenetically difficult sample spaces, and the increased utility and generality of digital evolution can be harnessed to explore questions that simulations have aimed to address. The following work was performed within the context of providing specific case-study examples of the digital experimental phylogenetics approach, and it has produced several surprising findings that are at least relevant under the evolutionary conditions investigated. This research shows that recombination may have a beneficial role in phylogenetic inference by encouraging substitutions to occur gradually throughout lineage evolution; that neutral evolution can pose greater difficulty for phylogenetic inference than directional selection; that using a more accurate model of evolution in the phylogenetic analysis may not offer improvement; that inferred branch lengths may often be quite inaccurate despite clade support being accurate; and, that metrics like bootstrap support and posterior clade support may not be close estimates of clade accuracy. The aim of this work is that it may show the phylogenetics and experimental evolution communities alike the potential for future uses of digital evolution to investigate research questions that are intractable with evolving biological populations.

Methods

Experimental design

Base evolutionary history

Eight experimental treatments were conducted in a fully factorial design of neutral evolution or natural selection, and naïve or pre-adapted ancestor, and asexual or sexual reproduction. Each treatment was replicated ten times, for the generation of a total of eighty 1,024-taxon fully symmetrical Avidian evolutionary histories.

Unlike in Chapter 2, the base of the ingroup taxa was not a true polytomy but was a root lineage uniting all taxa. This first lineage was initiated with a full population of the starting organism, and each subsequent lineage began with the cloned population from the end of the previous lineage. Therefore, lineage seeding in these experiments was consistent with the selection treatments in Chapter 2, and not like the other Chapter 2 treatments that used the single most abundant genotype to initiate lineages. In this manner, population growth did not occur, and the population size remained very close to 1,000 organisms for the entire evolutionary history. Under this rooted and fully symmetrical design, ten tree levels of evolution proceeded, producing 1,024 (i.e., 2^{10}) external branches and 1,023 (i.e., 2^{10}-1) internal branches.

The per site mutation rate was set sufficiently high to maintain the relative likelihood that Avidians would adapt in selective environments, yet low enough to be within the range of biological reality (Li, 1997). Mutation rates from $3 * 10^{-5}$ to $5 * 10^{-6}$ were initially evaluated using the default organism, population size, and under neutral evolution or a similar selective environment as described in the conditions here. These populations evolved for 10,000 generations and the sequence diversity of the population every 1,000 generations was evaluated. Variation at each locus was transcoded using the DNA bases for each instructional variant so that population genetics software could be used to infer nucleotide diversity, π (Nei, 1987). Rather than using a genetic code or having some other biological meaning, this was simply a means to characterize locus diversity; for example, the four most common variants

were coded as A, T, G, or C, and for loci with greater than four variants, all others were coded with a dash, although this rarely occurred. Using the program MEGA, version 6.06 (Tamura et al., 2013), sequence diversity was calculated for each timepoint and compared to the biologically realistic range of 0.005 to 0.02 (Pin et al., 2001; Stone et al., 2002). The only mutation rate evaluated that was within this range for nearly all conditions and generations observed was $1 * 10^{-5}$. Since Avidians reproduce with over-lapping generations such that the mutation rate affects an offspring genotype and not the parent's, the effective mutation rate was $5 * 10^{-6}$. For comparison, this rate is approximately forty times lower than in Chapter 2.

The number of generations each lineage evolved per branch was set so that branch lengths were reasonably short. The objective was to produce branches of lengths on the shorter range of biological reality, so that future work may use taxon sampling on the resulting evolutionary histories to evaluate longer branch lengths. Of the phylogenetic simulation studies surveyed, many evaluated branches as small as 0.01 substitutions per site and as large as 0.4 (e.g., Wiens and Cannatella 1998; Wiens and Soltis 2005; Simmons 2012). A total of 5,000 generations of evolution per branch was chosen, yielding an expected neutral evolution branch length of 0.025 substitutions per site, or expectedly lower under selection. This is approximately the length of the shortest branches inferred in Chapter 2 treatments. The total extent of evolution from root-to-tip experienced by a lineage evolving across eleven tree levels was 55,000 total generations and expected to be 0.275 substitutions per site under neutral evolution, which simulation analyses suggest should be within optimal ranges across phylogenetic difficulty conditions (Klopfstein et al., 2017).

Non-default Avida settings shared by all experiments included the following. Both the neutral and default sexual instruction sets allowed only 20 possible characters to mutate. The default instruction did so by disallowing the same set of instructions from being incorporated into the genome via mutation as in Chapter 2, except for *h-copy*, required for Avidian genome replication, which could mutate freely here. The birth method was set as "mass action," in which an offspring is placed randomly into the population instead of near their parent, resulting

in a lack of spatially- and genetically-structured populations and increasing the effectiveness of selection because individuals were equally likely to compete for space with relatives and nonrelatives alike. Population details, including genotype frequencies, were recorded every 100 generations, resulting in 50 sampling timepoints per branch lineage. All other settings were as the default, and Avida (Ofria et al., 2009; Ofria and Wilke, 2004) version 2.14.0 was used with modifications including the neutral instruction set and the extra computational tasks available for reward, as explained below.

Neutral evolution versus natural selection

Treatments varied in whether neutral or adaptive evolution occurred. Avida was modified to incorporate a new instruction set that consisted of 20 copies of the *nop-X* instruction (e.g., *nop-X*^{A}, *nop-X*^{B}, *nop-X*^{C}, etc.), varying only in their alphabetic abbreviation. This instruction does not take any computational resources to perform unlike other Avidian instructions, including *nop-C*, that do to various degrees. Therefore, there should not be any stabilizing selection or the potential for any form of adaptive evolution with Avidian genotypes using this set of instructions. One additional instruction other than a *nop-X* analogue, called *repro-sex*, was used that was necessary and sufficient for reproduction for these genotypes; while this instruction's locus was effectively invariant, it was the 1,001st position in the genome and was excluded from all analyses. While the reproduction instruction was excluded from being introduced via mutation, it had the possibility to mutate away causing inviability. In this manner, very, very weak stabilizing selection occurred as it was limited to one locus.

Selective environments were designed to produce very gradual adaptive evolution throughout the evolutionary history while additionally minimizing the potential of homoplasy. Each of the 2,047 lineages within the evolutionary history were exposed to a distinct selective environment to limit homoplasious character evolution, like selection regime "4" in Chapter 2 (Fig. 2.02). As with those treatments, a nesting environmental structure promoted continual adaptation, with all ancestrally rewarded tasks also rewarded in derived branches. Here, every branch additionally rewarded ten new tasks. To accomplish this, 215 new tasks were coded,

bringing the total number of available Avidian tasks to 348, including one-, two-, and three-input logic or mathematical tasks. Using these, relative task difficulty was roughly approximated by evolving replicate Avidian populations in the full-task environment and recording how many generations elapsed before a mutation conferring it arose. A total of 277 tasks were identified from these approximated difficulty classes and environments were configured such that two tasks from each class were awarded per branch to promote steady adaptation across lineages.

Since realistic population genetics were sought, selective environments were designed to decrease the magnitude and frequency of selective sweeps, e.g., fixation within fewer than 100 generations. Each task was rewarded with a 10% increase in merit, which was thought to be approximately minimally sufficient to promote phenotypic evolution. Note, there was no tradeoff in merit for performing computationally more difficult tasks, as each was equivalently rewarded. The ten new tasks rewarded each branch therefore allowed a maximal selective advantage of 1.1^{10} (i.e., 2.6x). In practice, however, Avidian populations rarely if ever reached full proficiency in rewarded tasks. At the very most, a difference between an Avidian performing no tasks versus performing all 110 tasks rewarded in an external lineage environment was 35,743x. Since task performance usually entails a nominal reduction in offspring cost relative to merit and since Avidians tend to evolve tasks sequentially, the realized differences in fitness between contemporaneous individuals are most likely greatly reduced from these values. These conditions varied from those of Chapter 2, in which all treatments experienced at least relatively weak stabilizing selection, and in which the most selective environment rewarded nine tasks with a maximal advantage of performing all versus no tasks of over 30 million.

Naïve versus pre-adapted ancestor

The ancestor used to initiate the evolutionary histories varied by being either naïve or pre-evolved. Each ancestor had a genome of 1,000 loci and this size was fixed by disallowing insertion or deletion mutations. Rare Avidian programmatic circumstances that may otherwise disassociate positional homology among loci was also controlled by mandating offspring size to

be identical to its parent and disallowing unstable genotypes from reproducing. As in Chapter 2, the naïve ancestor was equivalent to the default Avidian genotype although with additional *nop-C* instructions as genomic filler between the two strings of instructions necessary for reproduction. In this case 900 such instructions were added. As before, this genotype can reproduce but perform no other meaningful computation, for example a task rewarded by the environment.

Whereas the sought distinction between the naïve and pre-adapted ancestor in Chapter 2 was that the latter would be task proficient in the environments encountered during the evolutionary history, this was not the rationale between the naïve and pre-evolved ancestors used here. Recognizing the tremendous difference between the model of evolution experienced by the naïve ancestor and a biological organism, the goal for the pre-evolved ancestor was to create a genotype that had instruction frequency comparable to typical amino acid profiles in biological organisms. The was accomplished by evolving ten replicate Avidian populations from the default ancestor through millions of generations of evolution. The selective environment rewarded the remaining 71 tasks that were not rewarded among any branch environment in the selection treatments, although since only five tasks were performed by the identified genotype, the environment largely fostered stabilizing selection.

After periodically evaluating random organisms for instruction frequencies comparable to biological taxa, the genotype chosen as the pre-evolved ancestor had experienced 2,038,000 generations of evolution. Table 3.01 presents summary statistics for the pre-evolved and naïve Avidians in addition to several biological taxa, whose empirical amino acid frequencies were obtained using a diversity of methods and with taxa across the tree of life. While the naïve ancestor had 98.8% of its genotype as *nop-C* instructions and several instructions unrepresented, the pre-evolved ancestor had a frequency distribution approximately comparable to biological organisms, with only 9.9% of its genome as *nop-C*. Ancestors did not substantively differ in sequence variation between neutral and selection treatments, with the same genotypic sequence of instructions used for each but under differing instruction sets (i.e.,

the C instruction coded for *nop-C* in selection treatments and *nop-X*^{C} in neutral treatments). The only sequence difference was the presence of the additional final instruction for neutral treatments, which was necessary for reproduction, and was excluded from all phylogenetic analyses.

Table 3.01. Summary statistics for empirical amino acid frequencies for ancestor Avidian instruction frequencies and sets of biological taxa, with values scaled to 0–1,000 for ease of comparison to a 1,000-loci genome. For example, a minimum of 8 indicates that the lowest frequency for any of the twenty amino acid frequencies in the dataset was 8 out of 1000 (i.e., 0.8%). Some empirical frequencies were accessed via the ExPASy resource portal (2012).

	Pre-evolved ancestor	Naïve ancestor	Brooks et al. (2002)	UniProt Consortium (2017)	McCaldon and Argos (1988)	Hormoz (2013)	King and Jukes (1969)
Minimum	10	0	8	11	13	9	13
Median	55	0	48	54	53	53	47
Maximum	99	988	89	97	90	102	81
Standard Deviation	23	215	25	22	21	24	20

Asexual versus sexual reproduction

Recombination in sexual reproduction treatments occurred differently than in Chapter 2. In Avida there are four settings that determine the genetical recombination process, with additional settings related to other factors like mate choice or the two-fold cost of sex, which are of no concern here. The first setting is RECOMBINATION_PROB, which is the probability that recombination will occur between a pair of mates. A value of 0 was used for the asexual reproduction treatments and 1 for sexual treatments both here and in Chapter 2. Note that this setting is different from the recombination rate value used for biological organisms, where a rate of 0 means there is a nil chance of recombination between two specified loci (i.e., complete genetic linkage disequilibrium) and the maximum rate is 0.5 in which there is a 50% chance of recombination between two loci, (i.e., independent assortment). The second setting is MODULE_NUM, or the number of "modules" in the genome, which determines the fixed position of recombination breakpoints. Modules divide the genome evenly such that 25 modules for a 100-length genome would be four loci each. The end of each module is a

potential recombination breakpoint. Each module is evaluated independently when recombination occurs between a pair of mates, and each module has a 50% chance that recombination will occur. The other two settings, CONT_REC_REGS and CORESPOND_REC_REGS govern how modules are swapped, and are unimportant here, with values of 0 used for each. In Chapter 2, recombination occurred via independent assortment among all loci, with the number of modules set as the genome size. Here, the number of modules was set to 0, which causes two breakpoints to be chosen at random within the genome. Thus, linkage disequilibrium between two loci is negatively associated with greater genomic distance and will degrade over generations of reproduction as recombination probabilistically disassociates alleles. Therefore, Chapter 2 recombination is biologically analogous to each locus being like genes far apart on a chromosome or as different chromosomes altogether, while recombination in these experiments is more analogous to a sequence of DNA for which only very far apart loci associate independently.

Analyses

Distribution of mutational effects

The fitness of all one-step mutants was calculated with respect to a randomly chosen Avidian to evaluate whether the distribution of mutational effects was approximately like that observed for biological organisms. Since the genome was 1,000 characters in length and with 19 possible mutant instructions at each position, the relative fitness of 19,000 Avidians was recorded for each genotype evaluated. A systematic examination of the distribution of mutational effects was not conducted across an entire evolutionary lineage nor for all extant organisms at a population timepoint. Instead, genotypes were evaluated at random across a few treatments to roughly gauge adherence to biological examples. The genotypes chosen for inclusion here are approximately representative of others evaluated from similar evolutionary histories evolved under selection and include the following: The genotype of the pre-evolved ancestor was evaluated in both its ancestral (71-task) environment and in the 10-task environment at the root of the selection tree. The naïve ancestor's genotype was also

evaluated in the root selective environment. A descendant of the pre-evolved ancestor at the last timepoint on one external branch was evaluated in that branch's unique 110-task environment, as well as that for a descendant of the naïve ancestor in the same selective environment. In this manner, the distribution of mutational effects was recorded at the beginning and very end of the evolutionary history for both naïve and pre-evolved treatment conditions in addition to the final state of the pre-evolved ancestor in its ancestral environment.

Characterization of observed substitutions and empirical models of evolution

All alleles that reached fixation in a population were identified and tracked throughout their frequency trajectory during segregation. Alleles were considered as segregating within a population if their frequency was between 5% and 95%, above which they were effectively fixed and below which they were too infrequent to track. A substitution occurred when an ancestral fixed allele was supplanted by a derived allele that reached fixation, assessed as greater than 95% frequency. The fifty 100-generation population timepoints for each branch were evaluated for all such occurrences. If a substitution entirely occurred between data samples, i.e., within 100 generation, then it was considered a quick selective sweep, or "quick sweep." Otherwise, the frequency of the substitution and all other segregating alleles at that locus were recorded for each timepoint during which it was segregating. For each substitution, its locus, derived and ancestral state, origin and fixed branch, and number of generations to reach fixation were recorded. Data were used to evaluate various distributions, including the frequency of fixations per branch, frequency of fixations per generation per branch timepoint, substitutions per locus, number of generations to fix, and the generation on the fixed branch.

For selection treatments, the relative fitness of each substitution was then evaluated throughout its frequency trajectory. This was calculated for each timepoint by dividing the average fitness of all extant organisms with the derived state by the average fitness of all other extant organisms in the population. If at any timepoint there were no organisms that did not have the derived state, then the relative fitness was considered as 1, although such timepoints

were not considered when classifying by fitness type, below. If there was only one such timepoint then it was considered a quick sweep substitution.

The minimum and maximum relative fitness values across measured timepoints were used to classify substitutions by selection type (Table 3.02), using thresholds calculated following nearly neutral theory (Ohta and Kimura, 1971). Alleles of sufficiently small selection coefficients, s, have a probability of fixation primarily due to genetic drift when $|s| \leq 1/N_e$. The effective population size, N_e, is less than the maximum population size 1,000, so the threshold 0.001 used here is conservative in classifying substitutions as neutral. For example, beneficial and deleterious substitutions were never measured as having a relative fitness within the range 1 ± s, s = 0.001 for which genetic drift should be the dominant evolutionary determinant of their probability of fixation. Neutral-beneficial and neutral-deleterious substitutions had at least one timepoint for which their relative fitness was within this range and at least one timepoint above or below, respectively, and beneficial-deleterious substitutions may or may not have had one or more timepoints within the neutral range although they had at least one timepoint with a fitness above and at least one below.

Table 3.02. Classification of substitutions by the minimum and maximum relative fitness recorded across measurable timepoints throughout their frequency trajectory. Threshold values were determined based on nearly neutral theory (Ohta and Kimura, 1971) for a population size of 1,000, and fitness types are mutually exclusive. See the text for a description of a seventh type, quick sweep.

Substitution Fitness Type	Minimum Relative Fitness	Maximum Relative Fitness
Beneficial	> 1.001	–
Neutral–Beneficial (“neu-ben”)	≥ 0.999 & ≤ 1.001	> 1.001
Neutral	≥ 0.999	≤ 1.001
Neutral-Deleterious (“neu-del”)	< 0.999	≥ 0.999 & ≤ 1.001
Deleterious	–	< 0.999
Beneficial-Deleterious (“ben-del”)	< 0.999	> 1.001

The tracking of ancestral and derived states for all substitutions also allowed the creation of empirical fixed-rate models of Avidian instruction substitution. The structure of the

models resembles empirically-derived amino acid models (e.g., Dayhoff et al. 1978; Jones et al. 1992; Whelan and Goldman 2001; Le and Gascuel 2008) of unequal, fixed state frequencies for the 20 amino acid characters and the 190 general time-reversible substitution rates. Each model was parameterized using substitution rates and state frequencies pooled across all ten replicates for each of the eight experimental treatments. Note that substitution rates were treated as time-reversible, i.e., not differentiating between ancestor and derived character states in pooling X-to-Y and Y-to-X substitutions together, since greater-parameterized models were not compatible with the phylogenetic programs.

Phylogenetic analyses

For each of the eighty experimental treatment replicates, taxon-character datasets were created using all 1,024 extant populations. As in Chapter 2, a random organism was sampled from each population and the outgroup taxon was a randomly sampled organism from an extant population of a different replicate of the same experimental treatment. The complete sequence of a sampled organism was used, except for the final 1,001st character for neutral treatment organisms. As in Chapter 2, three characters, J, O, and B, were translated to amino acid abbreviation counterparts W, Y, and V, respectively, for compatibility with phylogenetic programs designed to handle amino acid abbreviations. Perfect positional homology was maintained with fixed organism genome lengths, so alignment was not required. Each of these datasets was used to conduct six different phylogenetic analyses.

Phylogenetic analyses generally occurred identically to those in Chapter 2, although no consensus methods were used, with only the “best” tree reported for each analysis. Neighbor-joining (NJ) trees were constructed using QuickTree, version 2.0 (Howe et al., 2002), and maximum parsimony (MP) was implemented using MPBoot, version 1.1.0 (Hoang et al., 2018, 2017). As before, the number of parsimony trees evaluated was 10,000 to conduct a more thorough heuristic search, and the “best” tree reported here is a random choice among the identified equally parsimonious trees. ML was implemented using IQ-TREE, version 1.5.5 (Minh et al., 2017; Nguyen et al., 2015), and the “best” tree was the phylogram with the greatest

likelihood value. A total of 100 nonparametric bootstrap replicates were also produced and used to determine the relative proportions of true and false clades with support values over 10%, a cutoff that excluded the excessive number of very infrequently inferred clades. BI was implemented using MrBayes, version 3.2.5 (Ronquist et al., 2012, 2011), with the parallel processing implementation and the BEAGLE library (Altekar et al., 2004; Ronquist et al., 2012), with the Markov chain sampled every 100 generations. The "best" tree reported here is the single phylogram with the greatest posterior probability, also termed the maximum clade credibility tree. The post-burnin sampled trees were used to determine the relative proportions of true and false clades with support values over 10%. For maximum likelihood (ML) and Bayesian inference (BI), separate analyses were performed using the Poisson fixed-rate model and the calculated empirical model corresponding to the treatment. For example, analyses of taxon-character datasets from the neutral treatment under asexual reproduction and starting from the naïve ancestor used the empirical model as calculated from the substitution data pooled across all replicates from that treatment. Python, version 2.7, and the following packages were used to organize and present these data: Jupyter, version 0.27.0 (Kluyver et al., 2016; Perez and Granger, 2007); Matplotlib, version 1.3.1 (Hunter, 2007); ETE2, version 2.2.1 (Huerta-Cepas et al., 2016); and DedroPy, version 3.12 (Sukumaran and Holder, 2010).

Topological accuracy between the true tree and inference tree was calculated using the variants of the Robinson-Foulds (RF) distance discussed in Chapter 2. Briefly, the RF distance is the sum of false positive (FP) branches and false negative (FN) branches, which can also be calculated as rates. To emphasize the accuracy of these inference methods, I report complement values. "Clade Accuracy" is the complement of the FP rate (i.e., FP divided by the number of internal branches in the true tree), and "Clade Resolvability" is the complement of the FN rate (i.e., FN divided by the number of internal branches in the inferred tree). The arithmetic mean of the FP and FN rates is the average topological error (Swenson et al., 2010) and I report its complement here, "Average Topological Accuracy." This metric indicates the overall accuracy of the phylogenetic inference by equally weighting correctly resolved clades

and unresolved clades. Examples of these metrics are shown for a comparison of a known (or correctly inferred) cladogram (Fig. 2.03a) to four variously inaccurate inferred cladograms (Fig. 2.03b-e). Note that the trees reported here, i.e., the best tree topology produced by each analysis, is fully resolved for each analysis except BI (since maximum clade credibility trees are not necessarily fully resolved). Since each non-BI best tree may include incorrect clades, each false positive clade requires a counterpart false negative clade, and therefore clade accuracy, clade resolvability, and average topological accuracy are equivalent values (e.g., Fig. 2.03b,e).

Additional measures include comparisons of inferred branch lengths and the amount of variable and parsimony informative sites in the taxon-character dataset. Branch lengths are summarized as the median length across all internal branches and, separately, across all external branch lengths. Tracking all substitutions that occurred within each evolutionary history allows the calculation of the true median substitutions per site per branch. These rates are compared to the median inferred branch lengths, as well as the expected value based on neutral theory (Kimura, 1983). Locus positions are variable if at least two types of characters exist among the set of taxa, and parsimony informative if at least two sets of characters are found among at least two taxa each.

Statistics

Statistical tests were used to evaluate the significance of differences between treatments and population genetic expectations. When comparing variation across replicates per treatment with a null hypothesis derived from population genetics theory, single-sample t-tests were conducted. For example, the number of substitutions per evolutionary history was evaluated using single-sample t-tests with the null hypothesis being that they were equivalent to the product of the mutation rate, genome size, number of generations per branch, and number of branches per tree, and this was conducted separately for each treatment. To evaluate significant differences among treatments, one-way ANOVA analyses were used, e.g., if there was a significant difference in the number of substitutions per evolutionary history among neutral treatments. Post-hoc analyses to determine the pairwise treatments driving the

statistical significance were conducted using t-tests with Bonferroni corrections for multiple comparisons. For the relative proportions of substitutions by fitness type, Chi-square tests were used. These included a four-by-two test to evaluate a difference among neutral treatments for quick sweep versus those taking greater than 100 generations to fix, and a four-by-seven test to evaluate a difference among all fitness types for selection treatments. Following this, the adjusted standardized residuals were evaluated for significance using a Bonferroni adjustment to the *z* critical value for the table size, and a four-by-two Chi-square test was used with a "ransacked" portion of the full table (Sharpe, 2015).

As in Chapter 2, the focus of the phylogenetic results is on comparisons of median values and trends across treatment replicates due to the recognized lack of independence in phylogenetics, including among topology metrics such as clade accuracy and resolvability and for branch lengths, since the presence of a non-zero branch length is dependent on the clade's inclusion in the tree.

Results

Empirical models of evolution

Separate empirical fixed-rate models of Avidian instruction substitution were created for each treatment and parameterized with rates and state frequencies pooled across all ten replicates per treatment. The two models presented in Table 3.03 represent the extremes of two trends of lesser versus greater frequency variation among the full set of eight models constructed. Within neutral evolution treatments, as shown in the first model in Table 3.03, the character-to-character substitution rates and per character state frequency values were approximately equalized, at least for all non-C characters, since all mutations were equally likely to occur and therefore result in substitution since fitness effects were nil. Within naïve ancestor treatments (e.g., the first model in Table 3.03), substitution rates and character state frequencies remained high for all values involving character C (i.e., *nop-C* in selection or *nop-*X^C in neutral treatments) due to its prevalence within the naïve ancestor genotype (i.e.,

constituting 988 out of 1,000 characters; Table 3.01). Conversely, and as shown in the second model presented in Table 3.03, selection treatments had greater variation among substitution rates and character state frequencies, at least for all non-C characters, since fitness effects were not equivalent among character states. And pre-adapted ancestor treatments (e.g., the second model in Table 3.03) had lesser variation among substitution rates and character state frequencies for values involving character C due to greater uniformity among character state frequencies in the pre-adapted ancestor genotype (Table 3.01). For reference, the Poisson model has equivalent rates and state frequencies for all values (Bishop and Friday, 1987). Using the scaling in Table 3.03, substitution rates are uniformly 2.63 (i.e., $\frac{500}{20*19/2}$) and character state frequencies are 5 (i.e., $\frac{100}{20}$) under the Poisson model.

Table 3.03. Two examples of empirical fixed-rate models of Avidian instruction substitution out of the eight models constructed, one per experimental treatment, by pooling substitutions observed across all ten replicates. Character-to-character substitution rates for the neutral, asexual evolution treatment starting with the naïve ancestor are above the diagonal and character state frequencies are the *first line of values below. Substitution rates for the selection, asexual evolution treatment from the pre-evolved ancestor are below the diagonal and state frequencies are the **second line. Note that the instructions are arrayed by the alphabetical order for full amino acid names, as is the convention for phylogenetic programs. Substitution rates are specified relative to a total of 500 per treatment, and state frequencies are out of 100 per treatment.

Neutral, Asexual, Naïve Treatment Model of Substitution Rates and Character State Frequencies*

Selection, Asexual, Pre-evolved Treatment Model of Rates and Frequencies**

	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
A		0.6	0.6	0.6	21.4	0.6	0.6	0.6	0.5	0.6	0.5	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
R	4.6		0.5	0.5	21.2	0.6	0.6	0.5	0.5	0.6	0.5	0.6	0.6	0.6	0.6	0.5	0.6	0.6	0.6	0.6
N	4	5		0.5	21.3	0.6	0.6	0.6	0.5	0.6	0.5	0.5	0.5	0.6	0.5	0.5	0.6	0.5	0.5	0.5
D	3.9	5	4.7		21.2	0.6	0.6	0.6	0.5	0.5	0.5	0.5	0.5	0.6	0.6	0.5	0.5	0.5	0.6	0.6
C	5.6	5.5	4.9	5.5		21.6	21.4	21	21.3	21.5	21.4	21	21.5	21.3	21.3	21.3	21.5	21.4	21.3	21.5
E	4.7	5.1	4.7	4.7	5.2		0.6	0.5	0.5	0.6	0.5	0.6	0.5	0.6	0.6	0.6	0.6	0.5	0.6	0.6
Q	4.1	4.6	4.3	4.2	4.7	4.3		0.6	0.6	0.5	0.6	0.6	0.6	0.6	0.6	0.5	0.6	0.6	0.5	0.6
G	0.6	1	0.9	1.1	0.9	0.8	0.5		0.5	0.6	0.5	0.6	0.5	0.6	0.6	0.5	0.6	0.6	0.6	0.6
H	4	4.7	4.6	4.4	4.8	4.4	4	0.9		0.5	0.5	0.5	0.5	0.6	0.5	0.5	0.6	0.5	0.6	0.6
I	3.8	4.5	5	4.2	4.8	4.2	3.8	1	4.2		0.5	0.5	0.5	0.6	0.6	0.5	0.6	0.5	0.6	0.6
L	1.1	1.3	1.2	1.4	1.4	1.3	1.2	0.3	1.2	1.3		0.5	0.5	0.5	0.5	0.4	0.6	0.5	0.5	0.6
K	2.7	3	3.1	2.6	2.9	2.9	2.7	0.8	2.8	3	1.6		0.5	0.7	0.6	0.6	0.6	0.6	0.6	0.5
M	3.3	3.7	3.9	3.4	4.4	3.3	3.2	0.6	4.2	3.6	0.9	2.3		0.6	0.6	0.6	0.5	0.5	0.5	0.6
F	0.4	0.9	0.8	1.1	0.7	0.7	0.5	0.6	0.7	0.9	0.1	0.5	0.6		0.6	0.6	0.6	0.6	0.6	0.6
P	1.3	1.7	1.6	1.4	1.8	1.5	1.5	0.3	1.4	1.1	0.3	0.8	1.2	0.5		0.6	0.6	0.5	0.6	0.6
S	4.2	4.7	5.2	3.9	5.3	4.4	3.7	0.8	4.8	4.1	1.1	4	3.3	1.2	1		0.5	0.5	0.6	0.5
T	0.9	0.8	1	0.9	1.1	0.8	0.7	0.4	0.9	1.1	0.4	0.9	0.8	0.3	0.2	0.9		0.6	0.6	0.6
W	2.9	3.1	2.8	2.7	2.9	2.9	2.6	0.8	2.5	2.5	1.4	2.9	1.9	0.5	0.9	2.5	0.7		0.6	0.6
Y	4.1	4.3	5.5	4.2	4.8	4.3	3.8	0.9	4.4	4.7	1.3	2.8	4	0.8	1.3	5.8	1	2.5		0.6
V	4.7	5	4.6	4.7	5.7	4.8	4.3	0.9	4.7	4.3	1.2	2.7	3.6	0.8	1.9	4.6	0.8	2.9	4.3	
*	3.2	3.1	3.1	3.1	40.5	3.2	3.2	3.1	3.1	3.1	3.1	3.1	3.1	3.2	3.2	3.1	3.2	3.1	3.1	3.2
**	6.1	6.8	6.8	6.4	7.3	6.5	5.9	1.4	6.4	6.2	2	4.5	5.2	1.3	2.1	6.5	1.5	4.2	6.5	6.6

Distribution of mutational effects

Clear trends were found when the fitness of all mutants with a single difference was evaluated for a randomly chosen Avidian from a specified population in its selective

environment. The proportion of lethal mutations (i.e., relative fitness equal to 0) is largely consistent across genotypes and environments, averaging 11% per genotype (Fig. 3.01). Using the conservative nearly neutral theory threshold of mutations being fitness neutral within the fitness range of 1 ± 0.001 (Table 3.02), neutral mutations were the largest category of possible mutations, with 42–83% per genotype. Neutral mutations were more common for the two root genotypes than the three genotypes evaluated at the termination of their evolutionary history in the selective environment, at 78% and 48% on average respectfully. Terminal genotypes include descendants in the tip environments of the evolutionary history as well as the pre-evolved ancestor in its ancestrally adapted environment. Small effect beneficial mutations, having up to 1% fitness effect, were extremely rare for all organisms, averaging less than 0.09% of possible mutations, and large effect beneficial mutations (i.e., relative fitness greater than 1.01) were rare, averaging less than 0.7%. Small effect deleterious mutations were relatively more common than beneficial mutations, with 10% on average for root genotypes and 3% for terminal genotypes. Finally, moderate and large effect deleterious mutations (i.e., relative fitness greater than 0 and less than 0.99) were more prevalent for the three terminal genotypes, averaging 37% compared to the root genotypes at 4%. Note that although relative fitness was calculated with high precision, values are collected here in bins of varying size to highlight proportions of mutants of certain effects.

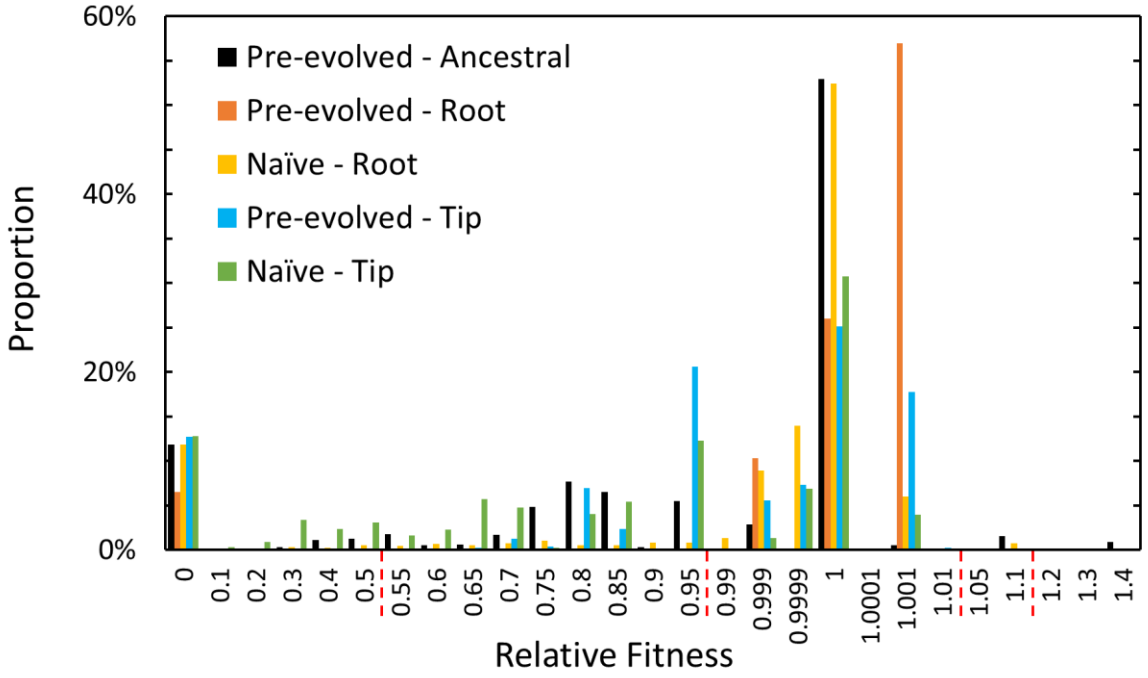


Figure 3.01. The distribution of mutational effects for five genotypes with the relative fitness proportions of all genotypes differing from a randomly chosen genotype by one mutation. Relative fitness bin size varies between 0.1, 0.5, and a logarithmic scale of base 10, with differences in scaling demarcated by red dashed lines below the x-axis. The genotypes evaluated included the pre-evolved ancestor in its ancestral selection environment (black), the same genotype in the selective environment at the root of the evolutionary history (orange), the naïve ancestor in the same root environment (yellow), a final descendant of the pre-evolved ancestor at the tip of the evolutionary history (blue), and a final descendant of the naïve ancestor in the same tip environment (green).

Characterization of observed substitutions

Fewer total substitutions occurred than expected under the neutral theory for all treatments. According to neutral theory (Kimura, 1983), the per locus per generation mutation rate is equal to the substitution rate for neutral alleles. Therefore, across 2,047 lineages (i.e., evolutionary history branches) each of 5,000 generations with a mutation rate of $5 * 10^{-6}$, the genome of 1,000 loci should experience 51,175 substitutions. Each of the eight treatments were evaluated for a significant difference from this expected value for neutral alleles and was significantly lower ($t_9 > 3$, $p < 0.05$; Fig. 3.02).

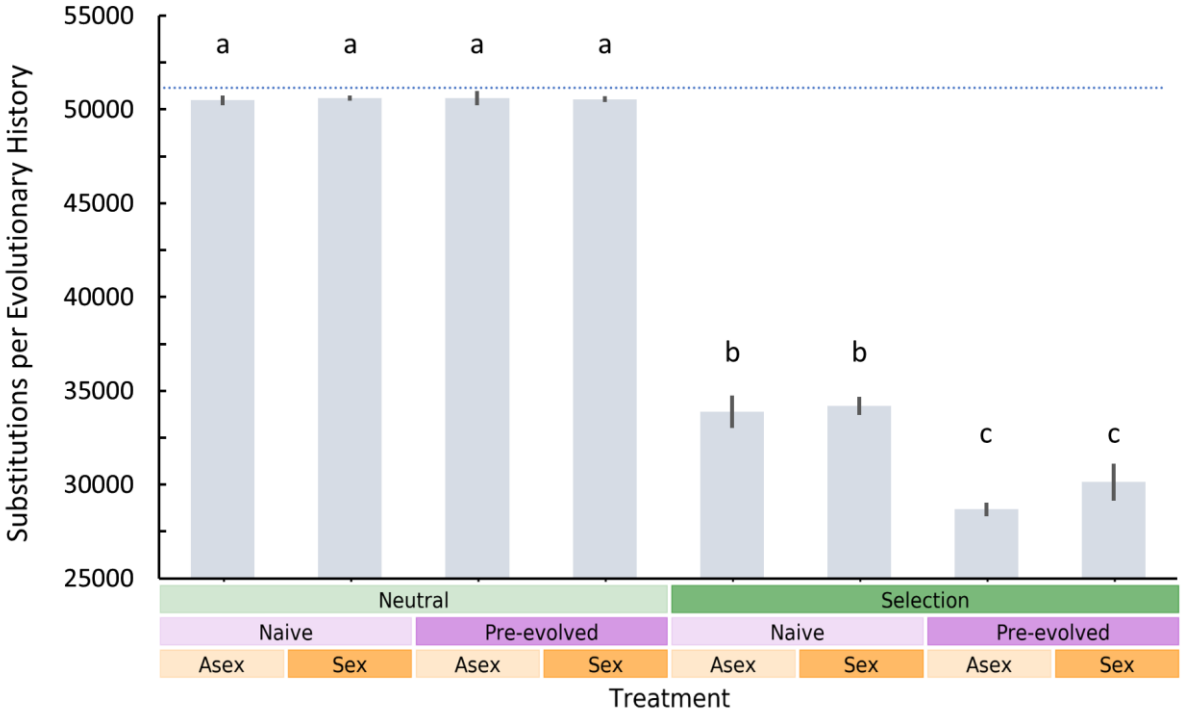


Figure 3.02. Number of recorded substitutions per treatment (mean ± 95% CI) relative to the expectation from neutral theory (dashed line). Experimental treatments are denoted by their selective condition (green labels), starting ancestor genotype (pink labels), and recombination condition (orange labels).

Evolutionary histories exhibited variation in total substitutions recorded per treatment condition. Compared to neutral evolution treatments that had a mean of approximately 50,500 substitutions per treatment, the selection treatments had at least 30% fewer total fixations, and within selection treatments those with the pre-evolved ancestor had about 15% fewer fixations than with the naïve ancestor (Fig. 3.02). Among the four neutral treatments there was no significant difference between the mean number of fixations per replicate ($F_{3,36} = 0.141$, $p = 0.935$), as also shown in the confidence intervals labeled “a” in Figure 3.02, and therefore neither the mode of reproduction nor ancestor genotype had an effect. Among the four selection treatments there was a highly significant difference between the mean number of fixations per replicate ($F_{3,36} = 41.28$, $p < 0.001$). Post-hoc analyses demonstrated a highly significant difference ($t_{36} > 7$, adjusted $p < 0.001$) between four of the six pairwise treatments. The exceptions were in the comparison of treatments otherwise identical except for having

asexual or sexual recombination. There was a non-significant difference in mean number of fixations per replicate between the recombination treatments with selection and the naïve ancestor (t_{36} = 0.556, adjusted p = 1), shown in Figure 3.02 with confidence intervals labeled "b", and the recombination treatments with selection and the pre-evolved ancestor (t_{36} = 2.775, adjusted p = 0.052), although the latter was significant for 95% normal-distribution confidence intervals, labeled "c". Thus, selection treatments starting with the pre-evolved ancestor fixed significantly fewer substitutions than those starting with the naïve ancestor, and while treatments with asexual reproduction fixed fewer substitutions than under sexual reproduction the difference was not significant for the naïve ancestor and borderline significant for the pre-evolved ancestor.

Within selection treatments, natural selection influenced nearly all substitutions, both positively and negatively, for at least one generation timepoint sampled across each substitution's population frequency trajectory. Across all treatments, as shown in Figure 3.03, 95–99% of all substitutions were beneficial for at least one timepoint (i.e., pooling beneficial, neutral-beneficial, and beneficial-deleterious fitness type categories) and 77–90% were deleterious for at least one timepoint (i.e., pooling deleterious, neutral-deleterious, and beneficial-deleterious). Most substitutions were of fitness type beneficial-deleterious (74–89% across treatments) or beneficial (12–20%). There was inconsistent ordering among the next three fitness classes, although neutral-beneficial substitutions were usually the third most common (0.75–4%), followed by deleterious (0.28–1.64%), and neutral-deleterious (0.4–0.67%). Quick sweep substitutions, whose relative fitness was indeterminable, since the allele's mutation and fixation happened within the 100-generation data collection window, was the second least common class (0.04–0.26%) and neutral the very least (0.02–0.09%). There was a significant difference in these proportions by fitness type among treatments ($\chi^2_{18,1268850}$ = 37028, p < 0.01). Evaluation of the adjusted residuals compared to the Bonferroni adjusted threshold indicated that all but four of the 28 factors drove this significance. Two of the exceptions are not especially enlightening (deleterious substitutions for the asexual

reproduction using the naïve ancestor treatment, and neutral substitutions for the sexual reproduction using the pre-evolved ancestor treatment); and the other two concern the observed proportion of quick sweeps in those same treatments, indicating that while they are not different from one another (0.13%), they are different with respect to sexual versus asexual reproduction treatments with the same starting ancestor. In comparison, for the neutral treatments only 0.001–0.002% were found to be quick sweeps (i.e., fixation within 100 generations) and there was no significant difference among treatments for the proportion of quick sweeps versus substitutions taking longer to fix ($\chi^2_{3,2022416} = 1.996$, $p = 0.573$). Comparing among other experimental conditions for the selection treatments, asexual treatments had approximately 10% relatively fewer beneficial-deleterious substitutions than their counterpart for the same starting ancestor. This was compensated with twice to four times greater proportions of each other class, especially neutral-beneficial substitutions.

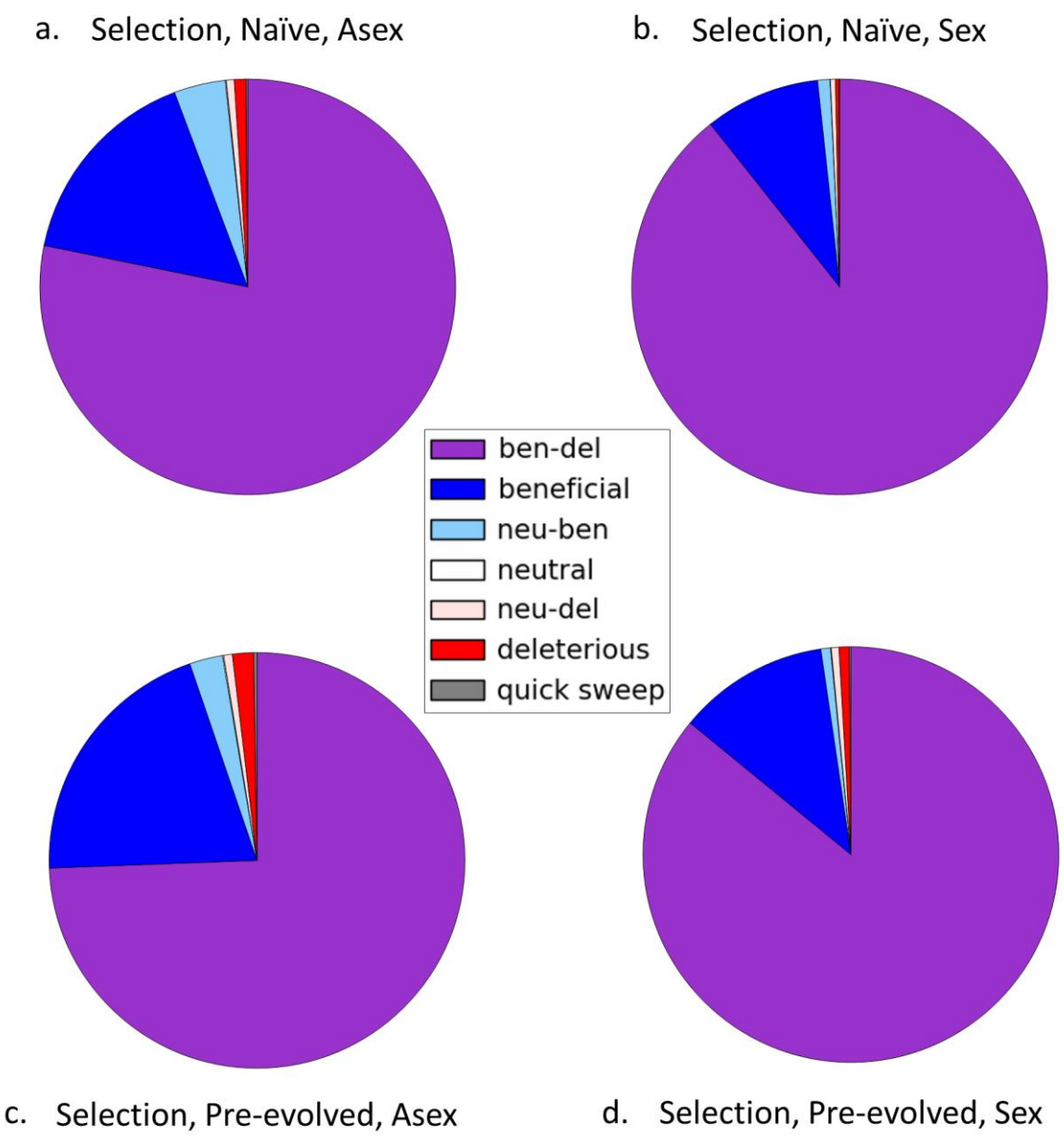


Figure 3.03. Number of observed substitutions by fitness type per selection experiment. Experiments began with either the naïve ancestor under asexual (a, N = 338,936) or sexual (b, N = 341,842) reproduction, or with the pre-evolved ancestor under asexual (c, N = 286,780) or sexual (d, N = 301,292) reproduction. For each treatment, data from all replicates are pooled, and substitutions are colored by fitness type as per Table 3.02.

The number of fixations were random with respect to position in the genome for neutral treatments and highly variable for each individual treatment replicate for selection experiments starting with the naïve ancestor, and there were greater regions of invariance across treatment replicates for selection experiments starting with the pre-evolved ancestor. All neutral

treatments exhibited random variation in which loci experienced substitution, as shown in Figure 3.04a for a single replicate. All selection treatment replicates exhibited a diversity in the number of substitutions per locus, with many sites being invariant or nearly invariant (e.g., Fig. 3.04c). The identity of sites with exceptionally low substitution rates differed for each replicate of the treatment starting with the naïve ancestor; for example, the variation shown in Figure 3.04c for a single replicate is largely averaged out when all replicates of that treatment are pooled (e.g., Fig. 3.04b). Selection treatment replicates initiated by the pre-evolved ancestor did not have as much variation in which sites had exceptionally low substitution rates, as shown in the pooled replicate data of Figure 3.04d, with many of the same sites remaining invariant for each replicate. In all selection treatments, the regions approximately corresponding to the reproduction machinery of the original naïve ancestor—the approximately five positions at the very beginning and very end of the genome—experienced few substitutions; further, the pre-evolved ancestor's genotype shared the first four and last five instruction sequences with the naïve ancestor, whose genotype was also its own ancestor prior to millions of generations of descent. After the first few positions, approximately the next fifty exhibited relatively higher rates of beneficial as well as deleterious substitutions. Only asexual reproduction treatments are shown because these trends do not appear to vary due to reproductive mode alone (i.e., treatments otherwise identical except for sexual versus asexual reproduction have similar distributions).

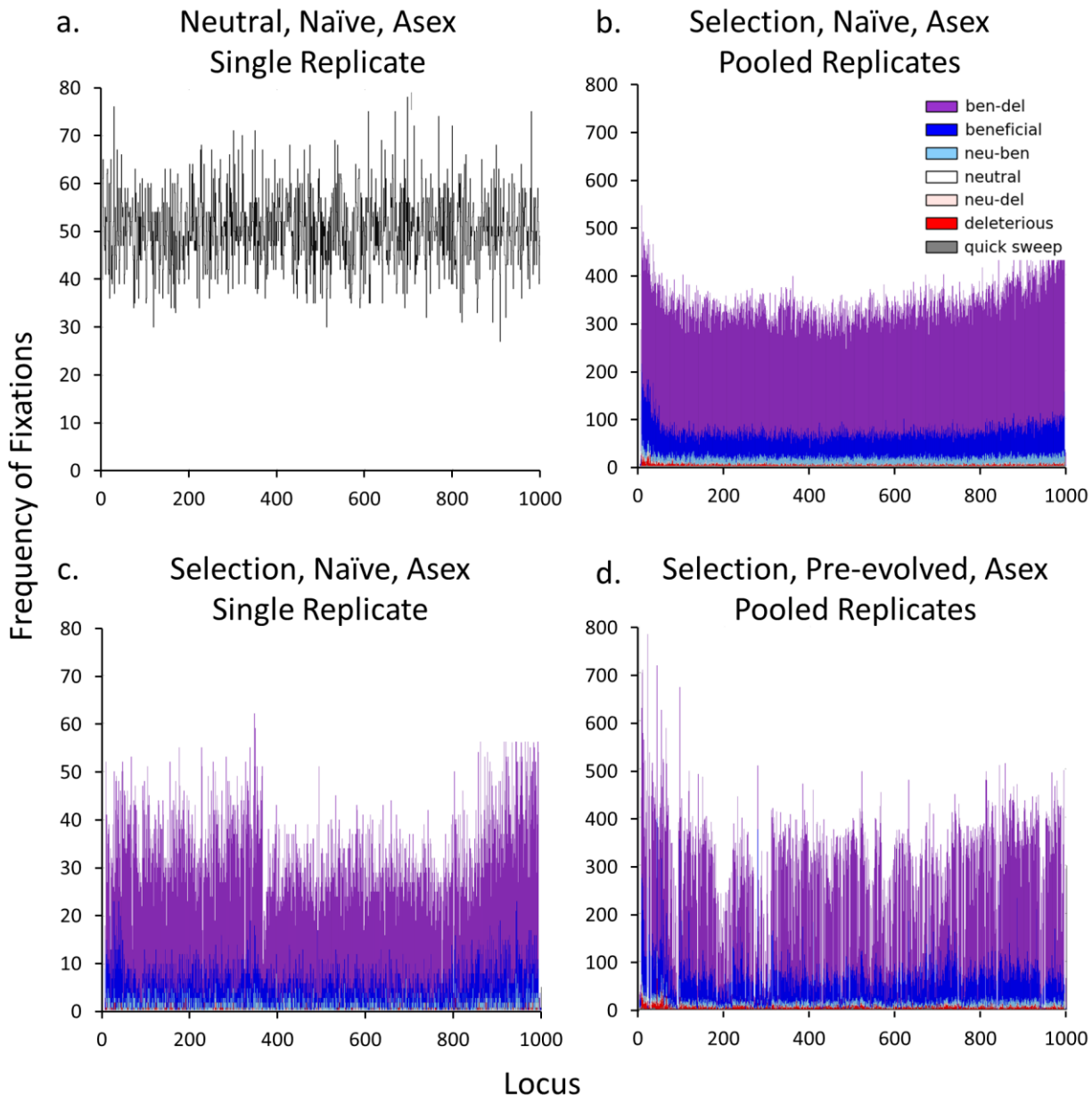


Figure 3.04. Representative treatment patterns for the number of fixations by Avidian locus. Selected patterns include single treatment replicates (a and c) or pooled treatment replicates (b and d), under neutral evolution (a) or selection (b–d), and starting with the naïve ancestor (a–c) or pre-evolved ancestor (d). Substitutions are colored by fitness type as per Table 3.02.

Recombination strongly affected how many substitutions fixed within the first couple generations after a population was introduced to a new selective environment. For asexual selection treatments (e.g., Fig. 3.05a), a much greater proportion of substitutions fixed every 100 generations during the first 600 generations than over the remainder of the lineage's evolution in that branch's selective environment. Sexual treatment replicates exhibited a

similar trend but with a lesser magnitude of difference across the branch (Fig. 3.05b). For either reproduction condition, the sampling interval which exhibited the greatest fixations was the second 100 generations, during which a much greater proportion of beneficial substitutions fixed. Only treatments initiated with the naïve ancestor are shown because these trends do not appear to vary due to ancestor genotype alone (i.e., otherwise identical treatments have similar distributions). Neutral evolution treatments did not exhibit variation with respect to the generation alleles fixed within a lineage.

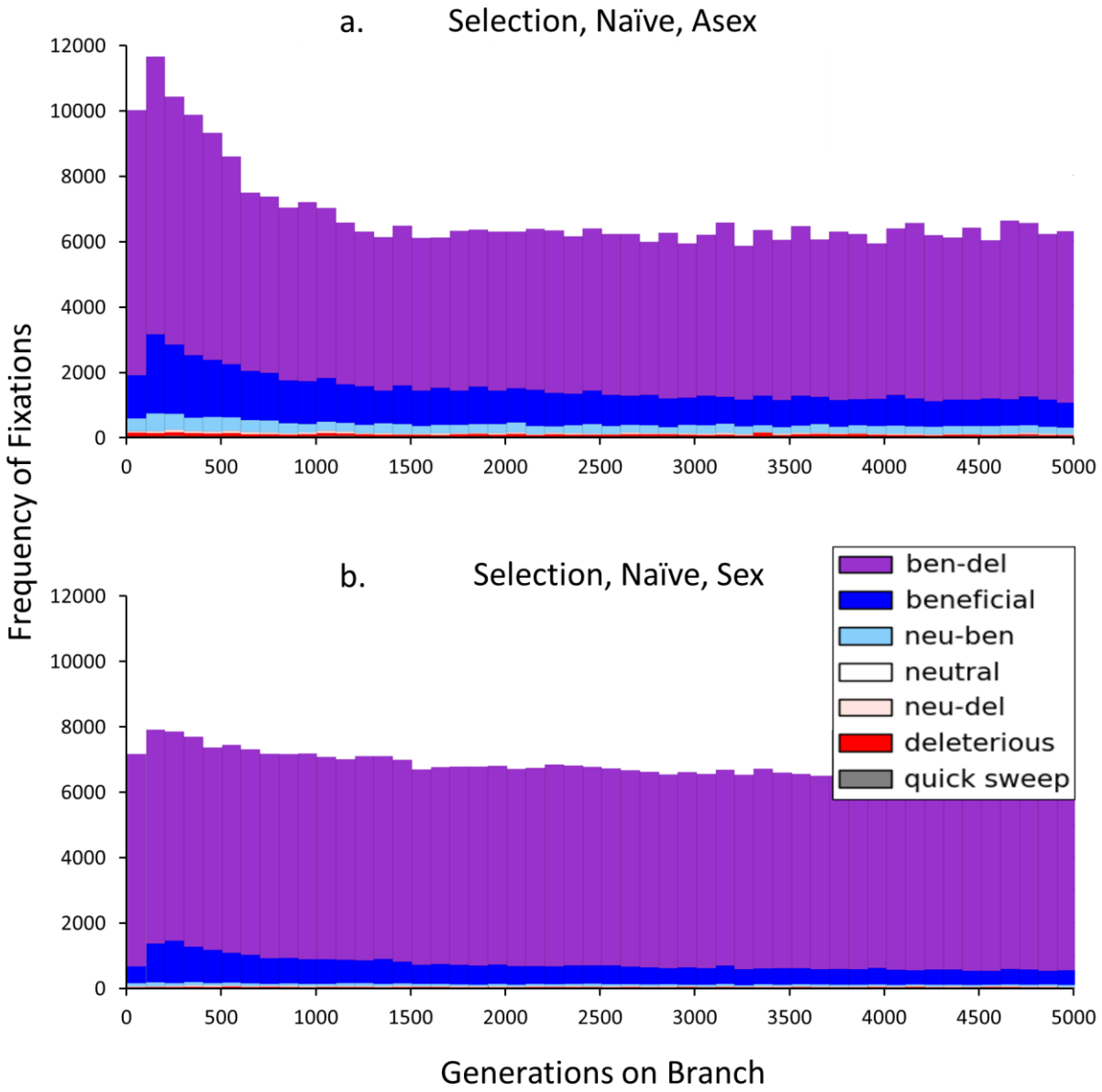


Figure 3.05. Representative patterns for asexual (a) and sexual (b) treatments under selection for the number of observed substitutions fixed per 100-generation population timepoint across all lineages (i.e., branches) in the evolutionary history. For each treatment, data from all replicates are pooled, and substitutions are colored by fitness type as per Table 3.02.

Alleles fixed slower than expected under neutral evolution. The time to fixation for neutral alleles in a haploid population should be $2N_e$ generations (Kimura, 1983), or at most 2,000 for this experiment. Each of the four neutral evolution treatments was evaluated for a significant difference from this expected value and was significantly higher ($t_{>500000} > 124$, $p < 0.001$), each with a mean of approximately 2,230 generations and a Poisson-like distribution.

Among these treatments there was a significant difference between the mean time to fixation ($F_{3,2022412} = 4.23$, $p < 0.01$). Post-hoc analyses demonstrated that this was driven by a significant difference in two of the six pairwise comparisons: between asexual treatments using the pre-evolved versus naïve ancestor ($t_{2022412} = 3.17$, adjusted $p < 0.01$) and between asexual versus sexual reproduction treatments starting with the naïve ancestor ($t_{2022412} = 2.66$, adjusted $p < 0.05$), however the largest difference in treatment means was less than 9 generations. Owing to this relatively miniscule effect size and with no clear pattern with respect to the effect of reproductive mode or starting genotype, the statistically significant differences between neutral treatments likely bears little importance.

Alleles fixed much slower with sexual compared to asexual reproduction when under selection. For each selection treatment, beneficial, neutral, deleterious, and quick sweep substitutions peak in fixing within their first hundred generations (Fig. 3.06); neutral-beneficial and neutral-deleterious alleles peak in fixing within their second hundred generations; and the mean fixation of beneficial-deleterious alleles is not until after approximately 1,100 generations under asexual reproduction and 1,300 under sexual reproduction after they are introduced to the population. Treatments with the pre-evolved ancestor (Fig. 3.06c,d) also fixed beneficial-deleterious alleles quicker compared to counterpart treatments with the naïve ancestor (Fig. 3.06a,b).

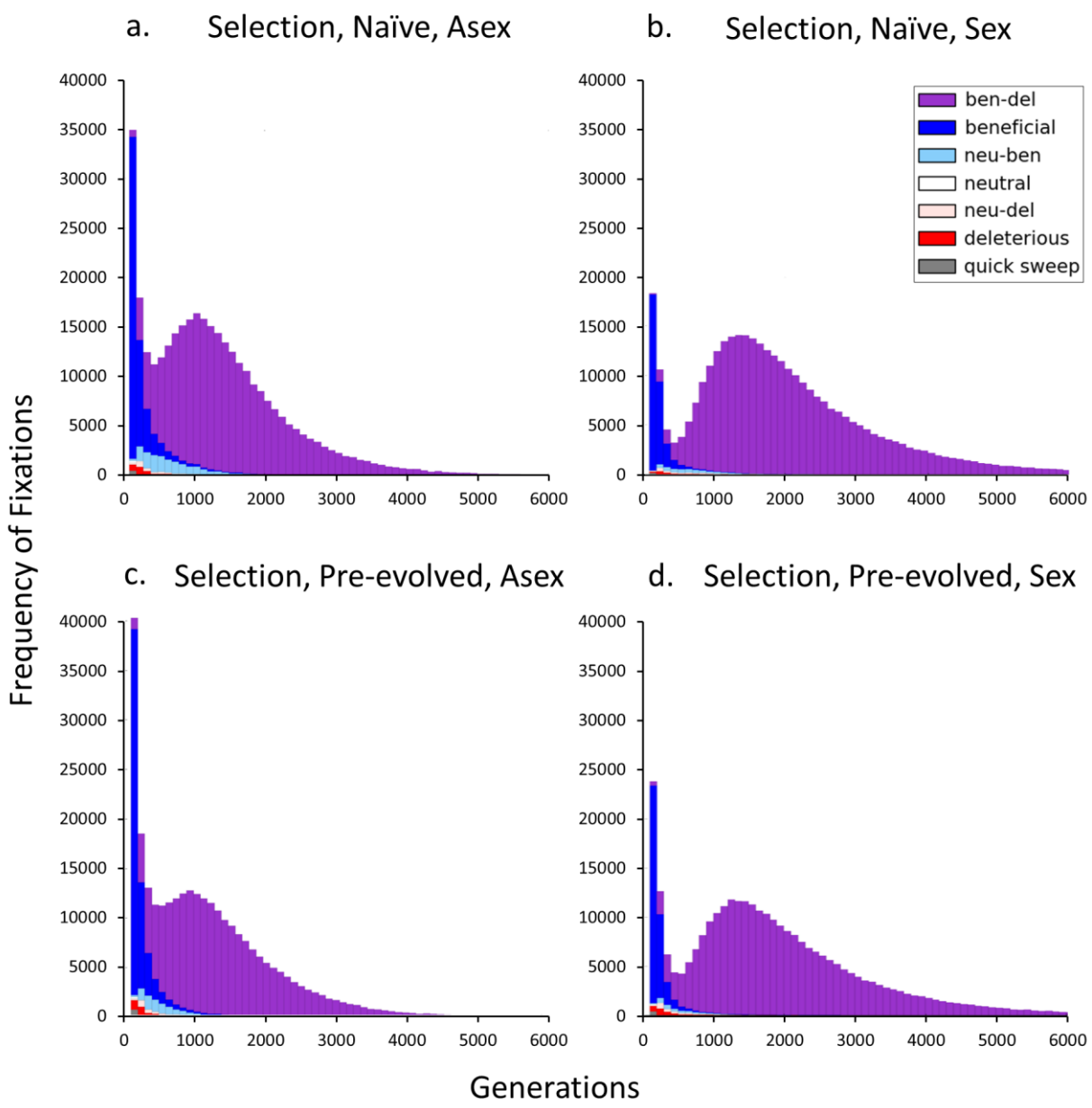


Figure 3.06. Number of observed substitutions by how many generations elapsed between entry in the population until fixation for selection treatments. Under sexual reproduction (b and c) the tail for beneficial-deleterious substitutions extends until 12,000 generations. For each treatment, data from all replicates are pooled, and substitutions are colored by fitness type as per Table 3.02.

Fewer fixations occurred every 100 generations under selection, and especially with sexual reproduction. For example, the selection with sexual reproduction treatment (Fig. 3.07d) exhibited fewer fixations per sampled timepoint compared to under neutral evolution (Fig. 3.07b), or with asexual reproduction (Fig. 3.07c), and far fewer when under neutral evolution

with asexual reproduction (Fig. 3.07a). Only treatments initiated with the naïve ancestor are shown because these trends do not appear to vary due to ancestor genotype alone (i.e., otherwise identical treatments have similar distributions).

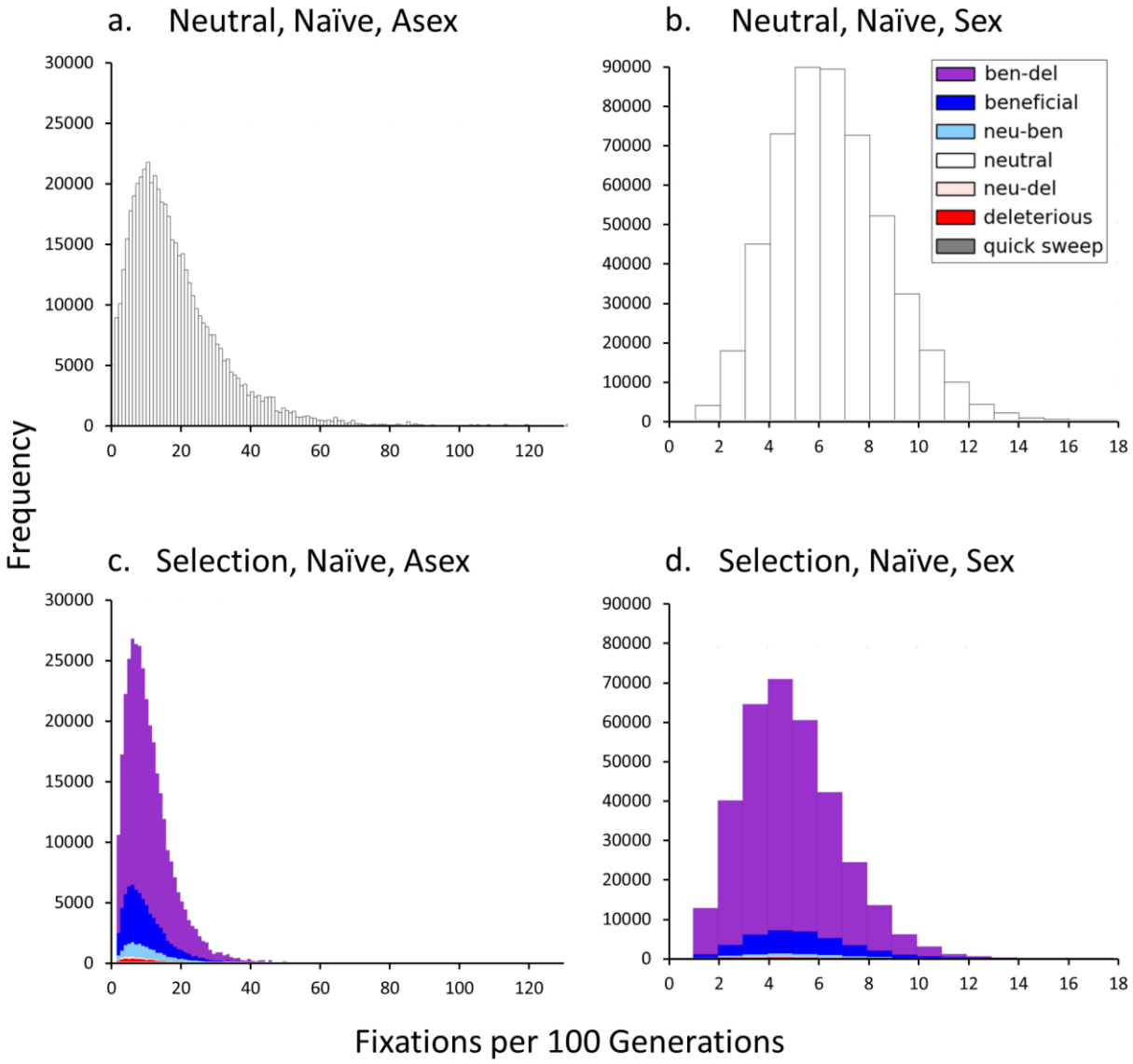


Figure 3.07. Representative treatment patterns for the number of observed substitutions fixed per 100 generations. Selected patterns include asexual (a and c) or sexual reproduction (b and d), and under neutral evolution (a and b) or selection (c and d). For each treatment, data from all replicates are pooled, and substitutions are colored by fitness type as per Table 3.02.

Neutral evolution caused stochastic frequency trajectory patterns, and asexual reproduction caused the periodic simultaneous fixation of more numerous substitutions compared to sexual reproduction. Population frequency change appears stochastic for

substitutions under neutral evolution, for example, as shown in Figure 3.08 for all substitutions occurring within one evolving population (i.e., one branch) in the evolutionary history. A greater number of fixations occurred together under asexual versus sexual reproduction, as represented by relative line thickness linking observed substitution frequencies in Figure 3.08. For example, at generation 500 and 3,200 of the asexual population (Fig. 3.08a), many alleles fixed together, while nearly all alleles in the sexual population fixed alone (Fig. 3.08b). Additionally, fewer unique genotypes at any given time contained an allele that would eventually reach fixation under asexual reproduction compared to sexual reproduction. For example, at the end of the branch (i.e., generation 5,000) in the asexual population, only one genotype of all that existed in the population is shown because only it contains alleles that reached fixation on a subsequent lineage following cladogenesis in the evolutionary history. In contrast, at the end of the branch in the example sexual population, more than ten genotypes contain one or more alleles that reach fixation following segregation on one or both subsequent branches.

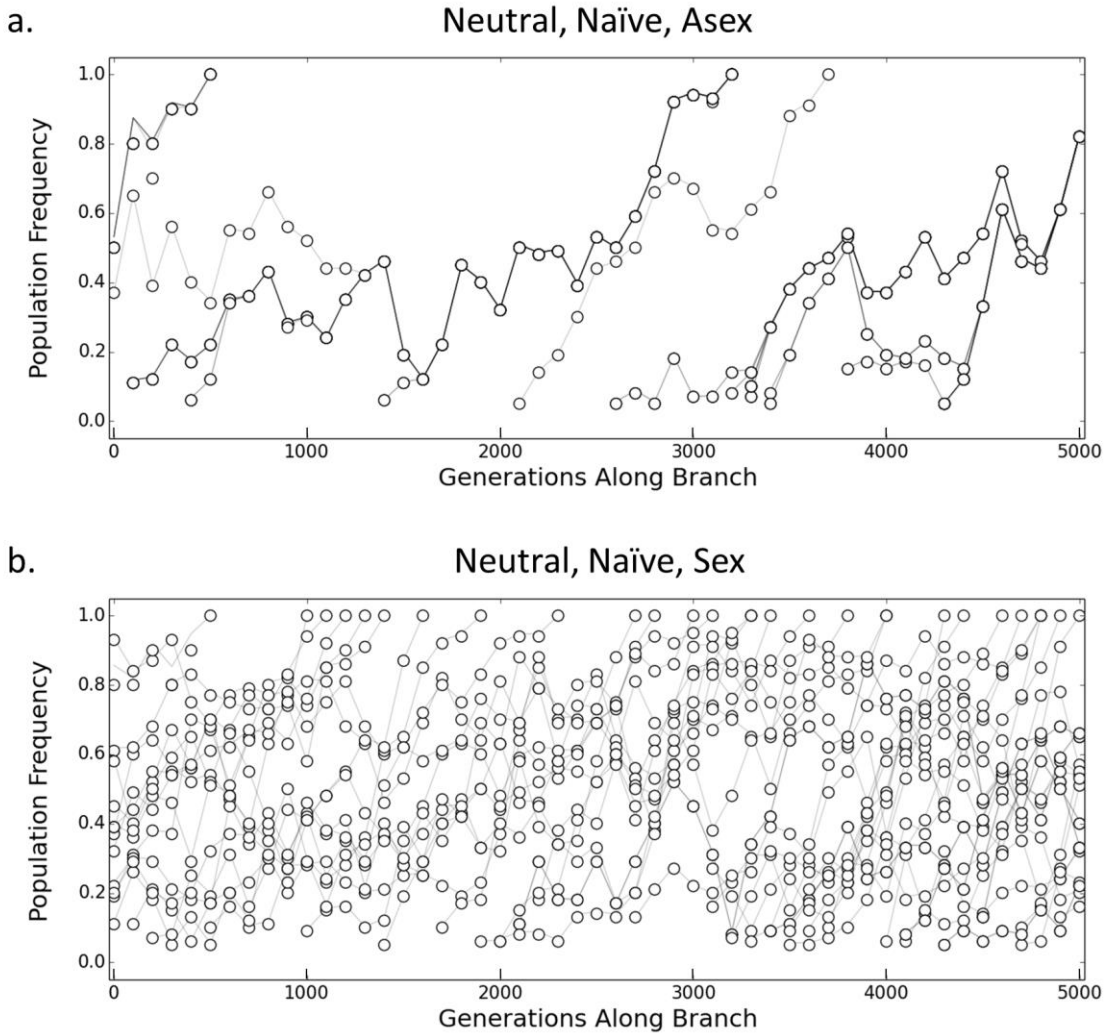


Figure 3.08. Representative patterns for asexual (a) and sexual (b) treatments under neutral evolution for how alleles that became substitutions changed in frequency between entry in a population (measured as ≥ 5% frequency) until fixation (≥ 95% frequency). Each panel shows a single population (i.e., one branch from one treatment replicate) on a middle tree level in the evolutionary history, and all substitutions that were segregating before, during, or after this lineage between cladogenic events. Observed substitution frequencies are shown as circles, and straight lines connect measurements every 100 generations. Line thickness indicates the relative number of substitutions on the same fixation trajectory.

Natural selection caused punctuated equilibrium patterns for average population fitness, which tended to be caused by beneficial alleles quickly fixing. Average population fitness tended to plateau before increasing by approximately 10%, for example, as shown in one evolving population in Figure 3.09d at generation 1,500. Such large increases in average population fitness tended to be caused by a beneficial allele quickly fixing and eliminating all variation at that locus, which is why fitness was indeterminable at that timepoint for the allele

(i.e., orange star in circle at 100% frequency at generation 1,500 in Figure 3.09c). Under either mode of reproduction, genotypes that eventually reached fixation were often influenced by positive and negative selection for at least one timepoint in their population frequency trajectory, in addition to neutral evolution caused by genetic drift. For example, the genotype in the example asexual population (Fig. 3.09a,b) that entered the population by generation 1,200 experienced strong negative selection within its first couple hundred generations, was within the 1% fitness range around neutrality for most of its trajectory and was under strong positive selection for at least the 200 generations before fixing by generation 2,800. As under neutral evolution (Fig. 3.08), asexual reproduction with selection caused many more alleles to fix together, and fewer unique genotypes at any given time contained an allele that would reach fixation (Fig. 3.09a) compared to sexual reproduction (Fig. 3.09c).

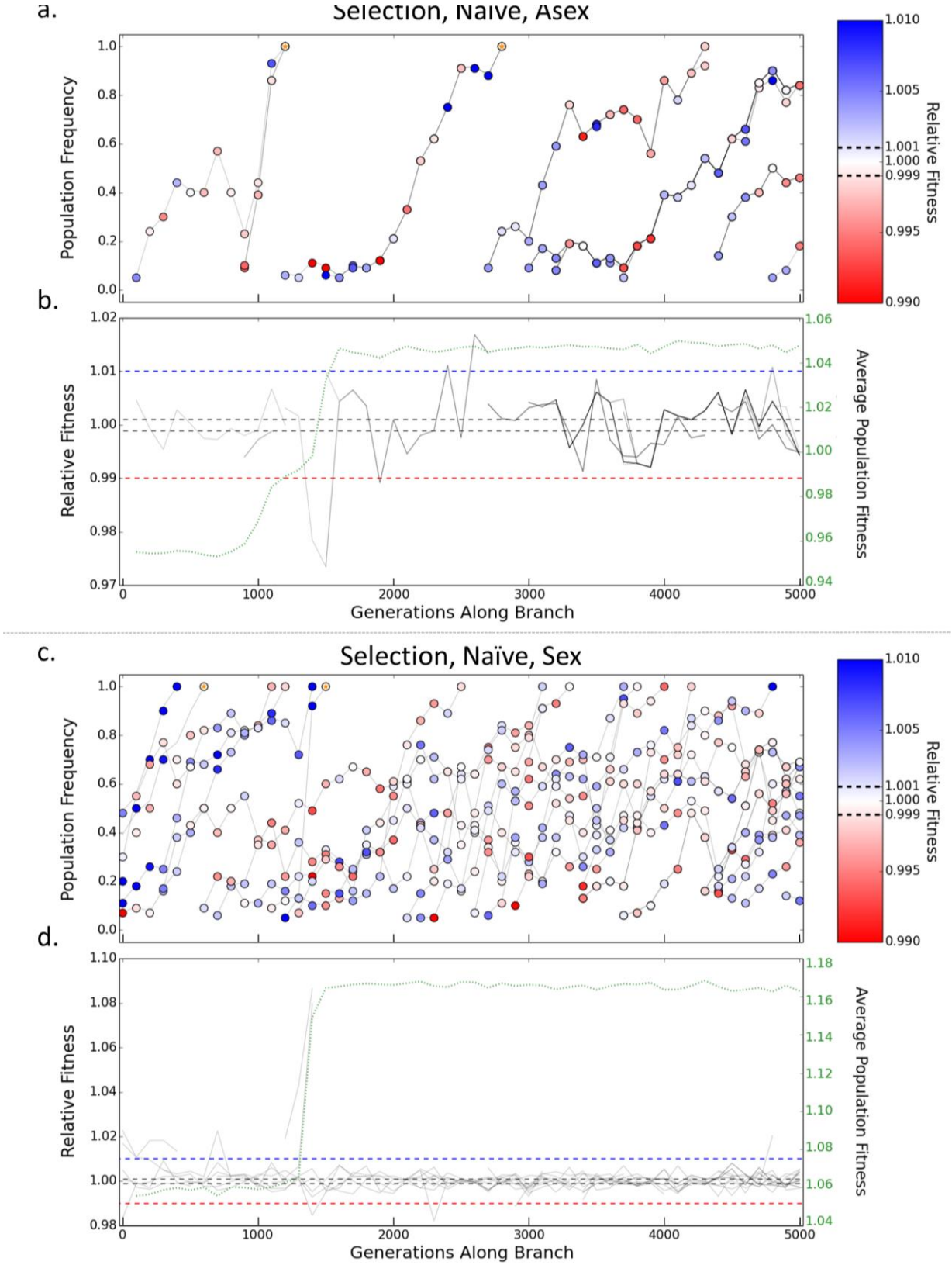


Figure 3.09. Representative patterns for asexual (a and b) and sexual (c and d) treatments under natural selection for how alleles that became substitutions changed in population frequency (a and c) and fitness (a–d), and their effects on average population fitness (b and d).

Figure 3.09 (cont'd). Panels a and b show a single population and panels c and d show a single population, each on a middle tree level in the evolutionary history, and include all substitutions that were segregating before, during, or after this lineage between cladogenic events. Panels a and c: Observed substitution frequencies are shown as circles, and straight lines connect measurements every 100 generations. Line thickness indicates the relative number of substitutions on the same fixation trajectory. Relative fitness is shown by coloration with values as indicated in the color bar, which includes the nearly neutral theory thresholds for neutrality (Table 3.02) as dashed lines. Orange stars within circles at 100% frequency indicate that fitness was indeterminable. Panels b and d: Relative fitness of substitutions are shown as solid black lines, with thickness indicating the relative number of substitutions on the same fixation trajectory. Dashed black lines indicate the nearly neutral threshold and dashed colored lines indicate a 1% selection advantage (blue) and disadvantage (red). Average population fitness is shown by the dotted green line and with values on the secondary y-axis.

Phylogenetic accuracy

Topological accuracy was high across treatments, including perfect for sexual reproduction treatments, and asexual reproduction treatments yielded improved accuracy when selection occurred. All analyses performed with greater than 96% topological accuracy (Fig. 3.10), although note that for these 1,024-ingroup taxon topologies, a 1% difference in accuracy is approximately equivalent to 10 clades being incorrectly inferred. Topological accuracy was perfect in sexual recombination treatments, reduced to about 99.8% (i.e., around 2 incorrect clades) for asexual selection treatments, and reduced to approximately 98% (i.e., around 20 incorrect clades) for asexual neutral evolution treatments. Ancestor genotype did not seem to affect topological accuracy.

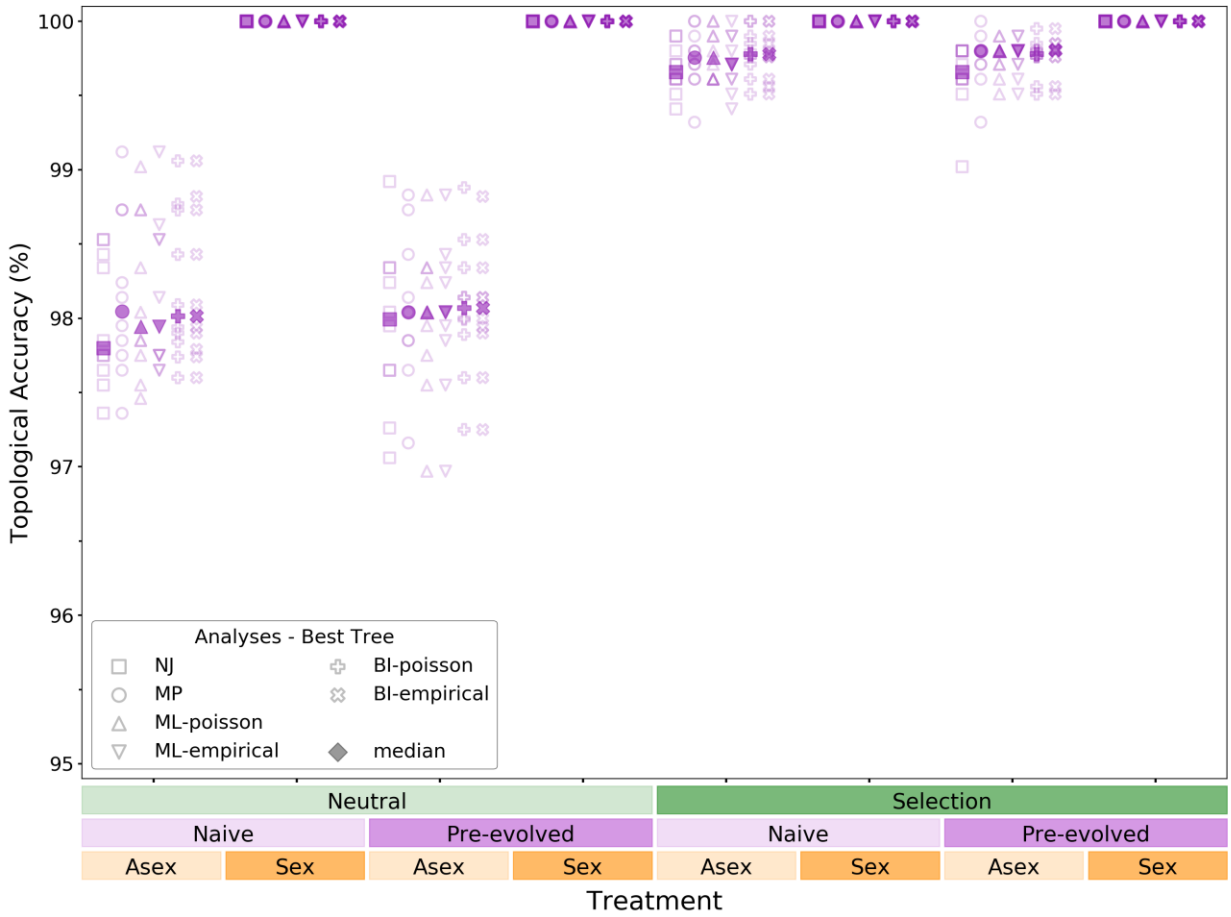


Figure 3.10. Topological accuracy, the average percentage of clades correctly inferred and correctly resolved, for each phylogenetic analysis' best tree per treatment. Analyses include neighbor joining (NJ), maximum parsimony (MP), and maximum likelihood (ML) and Bayesian inference (BI) with the Poisson or empirical model of evolution; open symbols for individual replicates and closed for the median across replicates. Experimental treatments are denoted by their selective condition (green labels), starting ancestor genotype (pink labels), and recombination condition (orange labels).

Differences between phylogenetic analysis performance among the asexual treatments, which did not reach perfect topological accuracy, was slight. The overall trend was that the median NJ replicate performed slightly worse than other analyses and BI tended to perform the best (Fig. 3.10). The exception is the most inaccurate treatment (i.e., asexual reproduction neutral evolution with the naïve ancestor), for which the median MP tree was more accurate than all other analyses, and this is especially surprising considering that the MP best tree selection was chosen from between six to 100 equally parsimonious trees across this treatment's ten replicates. With respect to median topological accuracy, ML and BI analyses

only slightly varied when conducted with the Poisson versus empirical model of evolution. The greatest difference was in the selection treatment with asexual reproduction and the naïve ancestor for which the Poisson model had 0.05% greater median accuracy.

Unlike in Chapter 2, the number of taxa analyzed here contributed to a lack of full resolvability in BI tree inferences. All non-BI analyses produced fully-resolved trees, so for these trees, topological accuracy is equivalent to clade accuracy and clade resolvability. Clade resolvability was equivalent for each BI model comparison per treatment and very nearly so for each treatment comparison among neutral and among selection treatments, so the very slight differences in topological accuracy between such comparisons are attributable to slight differences in clade accuracy alone. The greatest difference in accuracy between any such comparisons is for the selection treatment with asexual reproduction and the pre-evolved ancestor, for which the empirical model had 0.05% greater median accuracy and therefore 0.025% greater median topological accuracy. Since this lack of full resolution among BI trees caused such minor difference in median topological accuracy, only topological accuracy is presented here to ease comparison among analyses.

Bootstrap and posterior probability clade support values are overly liberal for neutral evolution treatments, and often overly conservative for selection treatments, with neither type of clade support value being a close approximation of clade accuracy. Clade support values greater than the 50% threshold, which is the minimum value of relevance for consensus support, are conservative estimates of accuracy for selection treatments (filled lines, Fig 3.11). The exception to this trend is that the 90% BI posterior support threshold produced liberal estimates of accuracy. For neutral treatments, clade support values greater than 50% are consistently liberal estimates of accuracy (unfilled lines, Fig 3.11). For either analysis type, ML versus BI (solid versus dashed lines), or the model of evolution used, empirical versus Poisson (blue versus orange), no consistent trends were observed. Treatments otherwise identical except for the starting ancestor produced very similar results, so such treatment pairs are pooled here. Unlike the analyses of asexual reproduction treatments, which resulted in both

true and false clades across most support thresholds, sexual reproduction treatments resulted in near-perfect sets of bootstrap and posterior sample topologies (i.e., exceedingly low frequencies of false clades), so the treatments shown here with asexual reproduction.

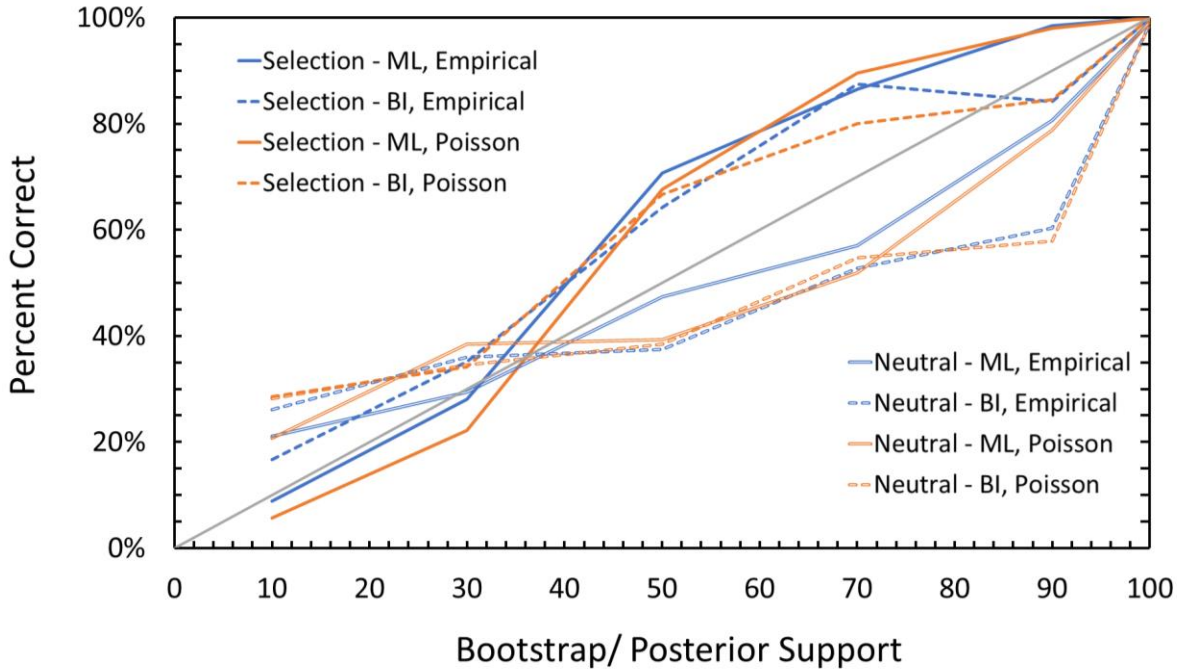


Figure 3.11. Relationship between clade accuracy as the percent of correct clades for values of bootstrap support for ML analyses and posterior probability support for BI analyses of asexual reproduction treatments pooled across naïve and pre–evolved ancestor conditions. Results include comparisons to bootstrap support for ML analyses (solid lines) or posterior probability support for BI analyses (dashed), using the empirical model of evolution (blue) or Poisson model (orange), and under selection (filled lines) or neutral evolution (unfilled lines). The grey line is the one-to-one accuracy-to-support relationship; values above are conservative as being an underestimation of accuracy and values below are liberal as an overestimation of accuracy. Note only support greater than 10% was assessed due to the large number of very low-supported clades.

The observed substitution rate (i.e., the empirical equivalent to branch length) per internal and external evolutionary lineage was less than the neutral theory expectation for all treatments. The expected substitution rate for neutral alleles (Kimura, 1983) is the product of the mutation rate and the number of generations per branch, so a lineage (i.e., evolutionary history branch) of 5,000 generations with a mutation rate of $5 * 10^{-6}$ should have 0.025 substitutions per site (Figure 3.12, solid line). Observed internal branch substitution rates (Figure 3.12, dashed lines) were found to be significantly lower than this expectation for seven

treatments (t_{1022} > 2, p < 0.01), with the exception being the treatment of neutral evolution with asexual reproduction starting with the naïve ancestor (t_{1022} = 1.784, p = 0.0747). Observed external branch substitution rates were also found to be highly significantly lower than the neutral theory expectation for seven treatments (t_{1023} > 3, p < 0.001), and although the exception—neutral evolution with asexual reproduction starting with the pre-evolved ancestor—was also significantly different, it was not as highly significant (t_{1023} = 2.210, p = 0.0273). Note that the observed substitution rates shown in Figure 3.12 are median values for consistency with the inferred branch lengths, while all statistical tests performed are for means.

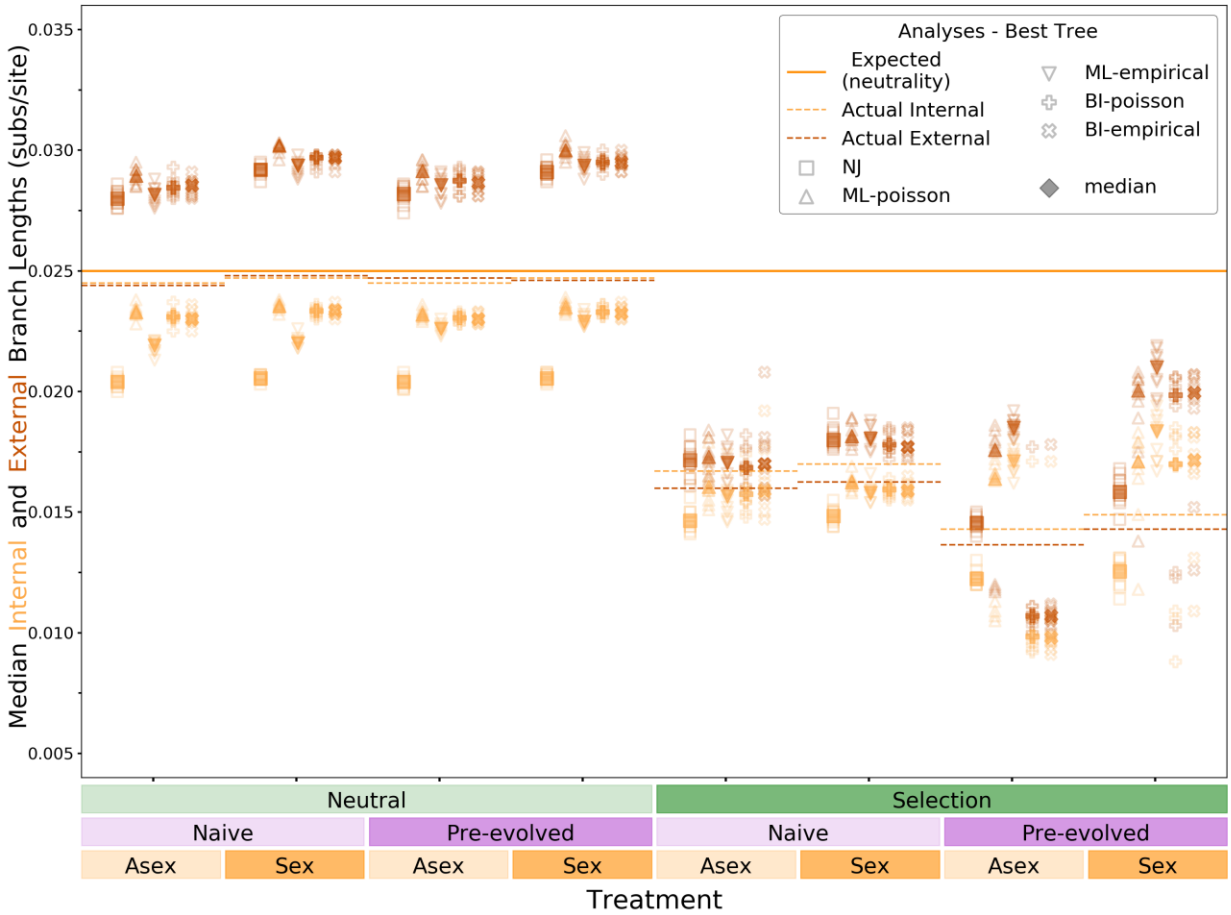


Figure 3.12. Empirical, theoretical, and inferred median internal (yellow) and external (orange) branch lengths per treatment. The empirically observed rates of substitutions per site (dashed lines) and the expectation under neutral theory (solid line) are included for comparison to the branch lengths inferred for the single best tree resulting from each analysis and model of evolution. Analyses include NJ (square), ML with Poisson model (triangle pointing up), ML with empirical model (triangle pointing down), BI with Poisson model (plus), and BI with empirical model (cross); open symbols for individual replicates and closed for median across replicates.

Only for selection treatments was the observed substitution rate, on average, higher for internal branches than external branches (Fig. 3.12). For neutral treatments, there was no significant difference in mean substitution rate between internal and external branches for three treatments ($t_{2045} < 1.5$, $p > 0.1$), with the exception being the treatment with sexual reproduction and naïve ancestor, which had significantly higher rates for external branches than internal branches ($t_{2045} = 2.7$, $p < 0.01$). For each selection treatment, the mean substitution rate for internal branches was higher than for internal branches ($t_{2045} > 6$, $p < 0.001$).

Ancestor genotype and mode of reproduction did affect the observed substitution rate for either internal or external branches for neutral evolution treatments, however for selection treatments, ancestor genotype and, less often, mode of reproduction affected observation substitution rates for both internal and external branches (Fig. 3.12). Among neutral treatments, there was no significant difference in mean substitution rate for internal branches ($F_{3,4088} = 0.36$, $p = 0.785$) and external branches ($F_{3,4092} = 1.45$, $p = 0.226$). Selection treatments demonstrated a highly significant difference in mean substitution rate for internal branches ($F_{3,4088} = 409.6$, $p < 0.001$) and external branches ($F_{3,4092} = 406.1$, $p < 0.001$). Post-hoc analyses for internal branches demonstrated a highly significant difference ($t_{4088} > 8$, adjusted $p < 0.001$) between all pairwise treatments except for the comparison of asexual and sexual reproduction treatments with the naïve ancestor ($t_{4088} = 1.397$, adjusted $p = 0.975$). Similarly, post-hoc analyses for external branches also showed a highly significant difference ($t_{4092} > 7$, adjusted $p < 0.001$) between all pairwise treatments except for the comparison of asexual and sexual reproduction treatments with the naïve ancestor ($t_{4092} = 1.952$, adjusted $p = 0.306$).

For each treatment and phylogenetic analysis, external branches were inferred to be longer than internal branches (Fig. 3.12). And for most treatments and analyses, as demonstrated by the observed median substitution rates (i.e., the empirical equivalent to branch length) for both internal and external branches being in between these sets of inferences, internal branches were inferred as shorter and external branches as longer than

what occurred during lineage evolution. Among neutral treatments, sexual reproduction produced slightly greater inferred branch lengths across analyses, while the ancestor genotype did not have an effect. NJ tended to infer shorter branches than did ML and BI, and this was especially the case for internal branches. Selection treatments starting with the naïve ancestor produced more accurate branch length inferences, and this was especially true for asexual reproduction. For this treatment pair, BI outperformed NJ and ML with respect to external branches, while ML slightly outperformed NJ and BI for internal branches. For selection treatments starting with the pre-evolved ancestor, the branch length inference accuracy of NJ was greater than BI and ML. This treatment regime with asexual reproduction was the only treatment in which ML and BI markedly differed, with BI even inferring shorter branches than NJ; although, several ML replicates using the Poisson model inferred similarly short internal and external branches as the median BI and NJ inference, and a few BI replicates using either model inferred similarly long external branches as median ML values. ML consistently inferred more accurate lengths than did BI for the sexual reproduction treatment, while each inferred either internal or external branch lengths better than the other for the asexual treatment. There was much greater variation among inferred branch lengths for selection treatments, with at least a few treatments per analysis and model of evolution for ML or BI being very different from the others.

The model of evolution made a difference in branch length inference most prominently among ML analyses compared to among BI analyses, which inferred very similar lengths using either model (Fig. 3.12). For neutral treatments, analyses using the Poisson model inferred greater median branch lengths than did those using the treatment's empirical model, and this was especially true for ML analyses of treatments starting with the naïve ancestor. This trend was not consistent among selection treatments, as only selection with sexual reproduction starting from the naïve ancestor produced this trend for both ML and BI, and ML alone did so for the otherwise equivalent treatment under asexual reproduction. For all others, the empirical model inferred greater branch lengths. Since both analyses consistently inferred

shorter internal and longer external branch lengths than that the observed substitution rates, model use impacted branch length accuracy differently for external and internal branches. For example, with neutral treatments the Poisson model was relatively more accurate at inferring internal branch lengths and less so for external branches compared to the empirical model.

Under neutral evolution every locus in the taxon-character datasets used for phylogenetic inference was parsimony informative, and under selection most loci were parsimony informative, but with fewer especially among the descendants from the pre-evolved ancestor. There was no variation among neutral treatments for either the number of variable or informative sites, with all replicates having the entire genomic sequence exhibiting variation and each character being informative (Fig. 3.13). One-way ANOVA analyses among selection treatments demonstrated a highly significant difference between the mean number of variable sites per replicate ($F_{3,36} = 16.52$, $p < 0.01$) as well as the number of informative sites ($F_{3,36} = 28.71$, $p < 0.01$). Post-hoc analyses demonstrated a highly significant difference between four of the six pairwise treatments for both the number of variable sites ($t_{36} > 5$, adjusted $p < 0.01$) and informative sites ($t_{36} > 7$, adjusted $p < 0.01$). For both measures, the exceptions were in the comparison of treatments otherwise identical except for having asexual or sexual recombination ($t_{36} < 1$, adjusted $p = 1$). Thus, selection treatments starting with the pre-evolved ancestor had significantly fewer, on average 21, variable and significantly fewer, on average 40, parsimony informative sites than those starting with the naïve ancestor, and recombination did not cause a difference.

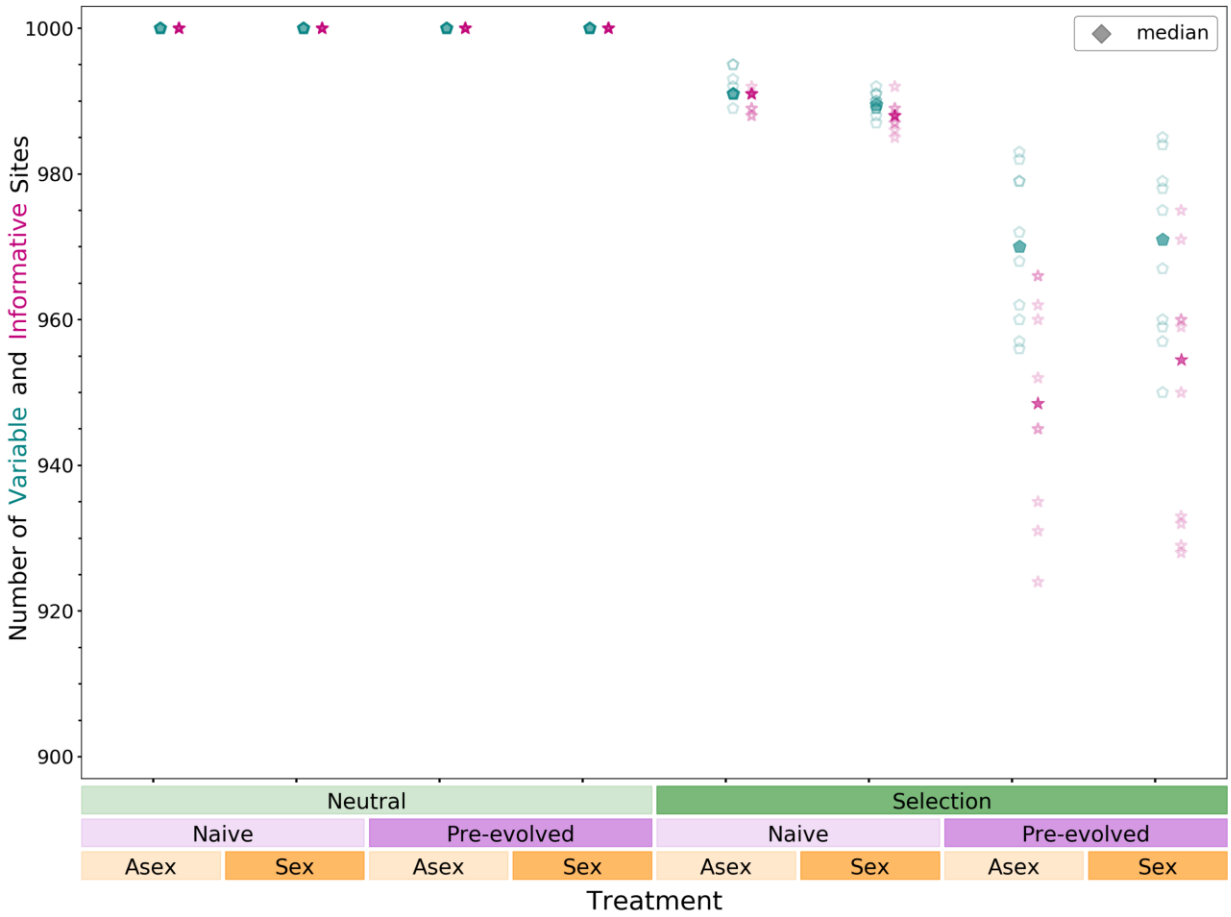


Figure 3.13. Number of variable sites (blue pentagons) and parsimony informative sites (purple stars) for all experimental treatment taxon-character datasets, with open symbols for individual replicates and closed for the median across replicates.

Discussion

The experimental evolution conditions were successfully configured so that digital evolution closely approximated the biological reality of molecular evolutionary dynamics. Establishing the congruity in molecular dynamics allows a stronger evaluation via analogy of whether natural selection aids phylogenetic inferences. Indeed, selection, at least directional and stabilizing selection, does increase topological accuracy and may increase branch length accuracy. Sexual reproduction, at least as implemented in this system, aided phylogenetic inference even more so than natural selection. And deviations from the model of molecular evolution used in the phylogenetic analysis made minor impact. Finally, neutral evolutionary dynamics did not entirely proceed as expected in neutral evolution treatments, and branch

length inference was especially problematic and clade support measures especially overestimated accuracy in these treatments.

Distribution of mutational effects

The distributions of mutational effects for Avidians evaluated after adaptation to their selective environment were similar to those observed for biological organisms (Fig. 3.01). All potential genotypes accessible by a single mutational change were evaluated for five genotypes and classified as lethal, large (i.e., > 1%) or small effect deleterious or beneficial, and neutral (i.e., relative fitness effect of 0.999 through 1.001) according to the nearly neutral framework (Ohta and Kimura, 1971) for the maximum population size of 1,000 organisms. As with biological organisms (Eyre-Walker and Keightley, 2007), most potential mutations for these genotypes are neutral with respect to fitness, a large collection are deleterious, a reduced proportion are lethal, and a small proportion are beneficial. The pre-evolved ancestor evaluated in its ancestral environment, having been evolved for millions of generations in that selective environment, most closely fit this expected profile. The descendant of either the naïve or pre-evolved ancestor at the tip of the 55,000-generation evolutionary history also fit this profile quite well. This is evidence that these Avidians adapted to their selective environment in contrast to the ancestors at the root of the tree, which did not have a history of adaptive evolution in the root environment. Compared to genotypes evaluated after many generations of evolution in their selective environment, the potential single-step mutations available to ancestor organisms evaluated at the root of the tree were proportionally more neutral (78% versus 48%), more small effect deleterious (10% versus 3%), and less non-lethal large effect deleterious (4% versus 37%, Fig. 3.01). As Avidian adaptation proceeds, greater numbers of loci are necessary for the completion of environmentally rewarded tasks (i.e., fitness-increasing phenotypes); further, some loci evolve to be pleiotropic in that they are necessary for the completion of multiple tasks. A mutation that confers a loss of a task within these selective environments yields a merit (and approximate fitness) loss of 9.1% per task, and a mutation that confers the gain of a task yields a 10% increase in merit (and approximately so for fitness).

The pools of potential deleterious and beneficial mutations observed here are in accord with these values or multiple thereof, with the latter being evidence that numerous loci are pleiotropic. The very large proportion of slightly beneficial albeit neutral mutations for the pre-evolved ancestor in the root environment were likely due to slight offspring cost optimization when not requiring the maintenance of previously phenotype-conferring loci in the ancestral environment.

Characterization of observed substitutions

Compared to the neutral treatments, the decreased number of fixed alleles (Fig. 3.02) and reduced branch lengths (Fig. 3.12) for the selection treatments agrees with the population genetics expectation and other data shown here. As expected, and shown by the distribution of mutation effects (Fig. 3.01), many mutations were deleterious and therefore eliminated by selection, causing a net loss of fixations compared to neutrality. The statistically significant difference in the total number of fixations between selection treatments with the naïve versus pre-evolved ancestor (Fig 3.02) would be expected due to greater mutations being deleterious for the latter; that is, stabilizing selection was more strongly occurring for experiments starting with the pre-evolved ancestor. This was not shown in Figure 3.01, where each tip-environment descendant had approximately 55% deleterious or lethal mutations, although the statistically significant greater proportion of invariant loci for the pre-evolved treatments (Fig. 3.12) suggest that a more thorough examination of mutational effects might demonstrate this as not uniformly the case among extant Avidians across tip environments.

The relative proportions of substitutions by fitness type for selection treatments broadly agrees with the population genetics expectation and prior research using Avida (Fig. 3.03). Since recombination decreases linkage disequilibrium, it is also expected to magnify the influence of selection in that an allele's evolution is more directly a consequence of its own effect on fitness rather than that of the broader genetic background to which it is linked (Felsenstein, 1974). This effectively makes selection more efficient by increasing the fixation of beneficial alleles and decreasing that for deleterious alleles. And neutral alleles have a decreased chance of fixation

due to being associated with beneficial alleles. The relative proportions shown in Figure 3.03 do exhibit fewer deleterious, neutral, and neutral-deleterious substitutions fixed in sexual treatments. While these treatments show fewer strictly beneficial substitutions, for each pair of otherwise identical treatments, this proportion is more than compensated by an even greater increase in beneficial-deleterious alleles.

There is a remarkable prevalence of individual substitutions being so prominently influenced by a diversity of evolutionary forces (i.e., positive selection, genetic drift, and/or negative selection), as suggested by fitness type proportions (Fig. 3.03). This finding is likely evidence of the pervasiveness of epistasis within evolved Avidian genomes (Lenski et al., 1999b; Strelioff et al., 2010; Valverde et al., 2012). For example, Covert et al. (2012) has demonstrated that sign-epistasis, that is, the beneficial-deleterious substitutions under the formulation here (Table 3.02), is common in Avida and may greatly contribute to adaptation. The prevalence of sign-epistasis has been predicted for biological systems too (Kvitek and Sherlock, 2011), however it is incredibly difficult to collect precise fitness data for individual alleles over multiple generations of population frequency change. It should be noted that the calculated prevalence of substitutions with mixed fitness effects throughout their frequency trajectory is likely an undercount since relative genomic fitness was evaluated only every 100 generations. Calculating fitness effects more often (e.g., every generation) likely would have shown that substitutions defined as purely beneficial, neutral, or deleterious substitutions as sampled every 100 generations were influenced by additional evolutionary forces between these point estimates.

The nearly neutral threshold used here is conservative in that the effective population size is not as great as the maximum population size. Yet it is unlikely conservative enough to account for the large proportions of substitutions that cross selection boundaries throughout their frequency trajectory, i.e., those that are not strictly neutral, beneficial, or deleterious. For example, increasing the neutral threshold from 0.001 to 0.005 for these analyses would be purposefully liberal in overestimating neutrality in a population of 1,000 individuals, because

0.005 is the neutrality threshold expectation for an effective population size of 500 organisms. Using a 0.005 threshold still results in very large proportions of substitutions such as beneficial-deleterious, with between 45–67% per treatment (not shown) compared to the 74–89% shown in Figure 3.03. With this liberal threshold, the proportion of strictly neutral substitutions in selection treatments is still quite low, with between 2.8–6.2% per treatment, although much higher than the 0.02–0.09% shown in Figure 3.03.

A few of the results in neutral evolution treatments exhibited deviations from neutral theory expectations and have an as-yet unidentified cause. The number of substitutions for entirely neutral alleles in the neutral evolution treatments was significantly less than expected (Fig. 3.02). The neutral theory expectation is that the mutation rate, here $5 * 10^{-6}$, should be equal to the substitution rate, which was found to be approximately $4.94 * 10^{-6}$ pooled across treatments and branches. And a similarly reduced rate of evolution was found with respect to substitutions per site per lineage (i.e., branch length) across internal branches and, separately, external branches, with medians and the neutral expectation shown in Figure 3.12. One suggestion is that since allele frequencies were sampled every 100 generations, these approximately 615 "missing" substitutions per evolutionary history may have swept to fixation and gone uncounted if more than one substitution occurred at the same locus within those 100 generations. This is impossible, however, since only 34 substitutions or 0.002% across all 80 neutral treatment evolutionary histories were identified as quick sweeps, which is over 700 times less frequent than would need to occur, let alone the improbability of two sweeping at the same site within 100 generations.

Another deviation from neutral theory involves the time to fixation of neutral alleles, which in a haploid population should be $2N_e$ generations (Kimura and Ohta, 1969). Across neutral treatments, the average time to fixation was approximately 2,230 generations, which was significantly different from the 2,000-generation expectation. This value would yield an effective population size of 1,115, which would be impossible with the maximum population size of 1,000 experimentally limited here. Conservatively, with an estimated effective

population size of the maximum population size, fixation time was observed to take 230 generations longer than expected. In fact, it took even longer than this since segregation below 5% and greater than 95% in the population was not tracked. While allele frequencies were sampled at population timepoints every 100 generations, this should systemically undercount both their entrance and fixation into the population, and therefore not affect the observed average fixation times of neutral alleles. Note that although there was a significant difference between neutral evolution treatments in time to fixation, the relatively miniscule effect size of 9 generations and lack of a clear pattern with respect to the effect of reproductive mode or starting genotype suggests that this difference bears little importance compared to the overall difference from the neutral theory expectation.

The fixation trends exhibited by the neutral and selection treatments agree with the population genetics expectations. For selection treatments, each lineage or branch of the evolutionary history experienced a new selective environment. The population tended to experience a burst of adaptive evolution, with many beneficial and beneficial-deleterious substitutions being fixed after a brief lag from its introduction to the environment (Fig. 3.05). For the remaining approximately 75% of the branch, the population underwent further adaptive evolution although with no trend in when alleles fixed, and with little difference in the relative proportion of substitutions by fitness type. Beneficial substitutions fix rapidly and appear to sweep to fixation neutral, deleterious, and quick sweep alleles, which all peak in fixing within their first hundred generations in the population (Fig. 3.06). And the frequency trajectory of beneficial alleles within a population was highly dependent on their relative fitness (Fig. 3.09). For comparison, evolution in neutral treatments seemed uniform and stochastic in that there was no variation in which generation along a lineage a fixation occurred, treatment conditions did not alter their time to fixation, and their frequency trajectory in the population appeared to be textbook (Hartl and Clark, 2007). Curiously, the time to fixation of beneficial-deleterious alleles in selection treatments seems distinct from alleles of other fitness types within those same treatments (Fig. 3.06). While these alleles did fix after fewer generations

than for alleles in the neutral evolution treatments, their fixation dynamics in aggregate appear to be much closer to neutral in effect size rather than beneficial or deleterious.

The fixation trends exhibited by the sexual and asexual treatments also agree with the population genetics expectation. In comparison to asexual treatments, there is two- to three-fold fewer quick sweep alleles fixed in the sexual treatments (Fig. 3.03). Recombination caused a much lower increased rate of fixation when a population was first introduced to a new selective environment (Fig. 3.05), although this burst involved a relatively greater proportion of beneficial alleles compared to other fitness types (Fig. 3.06). There was a much greater distinction between substitutions fixed quickly versus slowly upon their introduction to the population, with beneficial and hitchhiking alleles fixing rapidly and quickly decaying in their time to fixation. There was also a more distinct proportion of beneficial-deleterious alleles with respect to their time to fixation, and they took longer to fix on average than their asexual treatment counterparts. The starkest difference between sexual and asexual treatments was in the number of alleles that hitchhiked to fixation among linkage groups, with many fewer alleles fixing every hundred generations under sexual reproduction (Fig. 3.07).

Clonal interference in asexual population and the Hill-Robertson effect in sexual populations are nicely illustrated in these evolving Avidian populations. Figure 3.09a illustrates the dynamic under adaptive asexual reproduction in which new advantageous alleles can only become substitutions once a prior linkage group has fixed (e.g., at 1,100 and 2,700 generations) unless they arise on a background that is already fixing in the population (e.g., 3,000, 3,200, 4,400 generations, etc.). This is clonal interference (Gerrish and Lenski, 1998) and while we cannot see the frequency trajectories of competitors that had relatively high fitness but went extinct because only alleles that fixed were tracked, the decrease in relative fitness that made the fixing allele deleterious or neutral from generations 1,900 through 2,300 was likely caused by one or more relatively higher fit competitors. And finally, Figure 3.09b illustrates the Hill-Robertson effect (Hill and Robertson, 1966) in which advantageous alleles can be combined onto the same linkage block and increase in frequency together. For example, the very highly

advantageous new mutant at generation 1,200 quickly rose in frequency and fixed in the population. At generation 1,500, when the allele fixed, at least four other alleles that later fixed in the population were segregating; their maintenance in the population required recombination onto the genetic background that included the highly beneficial allele. In contrast, under clonal interference in an asexual population, at the time of fixation, only at most one other allele destined for fixation would be segregating in the population (e.g., at 2,700 generations, Fig. 3.09a).

Phylogenetic accuracy

All phylogenetic analysis methods for all treatments with sexual reproduction produced the correct 1,024-taxa topology, and when reduced under asexual reproduction, selection improved accuracy (Fig. 3.12). It's not entirely clear why sexual reproduction improved phylogenetic inference compared to asexual reproduction irrespective of natural selection or other deviations to the model of evolution. The few trends that varied due to sexual versus asexual reproduction alone are that fewer fixations per 100 generations occurred with sexual reproduction (Fig. 3.07), which was at least in part driven by fewer fixations occurring simultaneously (e.g., Fig. 3.08 and Fig. 3.09), even though there tended to be no difference in the number of total substitutions (Fig. 3.02) or the per lineage substitution rate (Fig. 3.12) for otherwise identical treatment conditions. Together these trends suggest that phylogenetic inference is improved by substitutions occurring gradually throughout lineage evolution, instead of in bursts (e.g., Fig. 3.08). Note that this is a different phenomenon than a molecular clock (Kimura, 1968), because under neutral evolution an asexual population would still have a molecular clock; this is more akin to how loud versus quiet the clock ticks (i.e., fewer substitutions fixing simultaneously) rather than the clock ticking regularly versus irregularly or ticking faster versus slower. This hypothesis could help support why natural selection aided phylogenetic inference under asexual reproduction (Fig. 3.12), because selection itself causes fewer fixations per 100 generations (Fig. 3.07).

While topological accuracy was very high, the population processes and genomic evolution processes occurring in this digital evolution system likely contributed to the inaccuracy observed. The lowest performing analyses had about 2% of clades incorrectly inferred, which is about 20 clades; surprisingly, this relatively poor performance was the neutral evolution treatments without sex (Fig. 3.10). Since coalescent and other population genetic processes are commonly evaluated with respect to simulations of neutral evolution, this result demonstrates that such modeling is missing important population processes that decrease phylogenetic algorithm performance. And this modeling is additionally failing to account for other processes that may promote accuracy, such as selection as evidenced here, which had an order of magnitude less inaccuracy, about 0.2% (Fig. 3.10). One explanation for selection, particularly stabilizing selection, improving topological accuracy over neutral evolution may be the maintenance of synapomorphic loci important in contributing to rewarded tasks (i.e., phenotypes). As shown in Figures 3.04 and 3.13, these treatments had greater frequencies of invariant sites and sites that otherwise experienced reduced rates of change.

Branch length inferences were highly inaccurate across most treatments, and there was no relationship between topological accuracy and branch length accuracy for a given treatment. For example, the median external branch for the selection treatment with sexual reproduction and the pre-adapted ancestor had an inferred branch length about 40% longer (for analyses other than NJ) than the median observed substitution rate for those lineages (Fig. 3.12); and this was despite clade accuracy being perfect for this treatment (Fig. 3.10). In contrast, the median external branch for the selection treatment with asexual reproduction and the naïve ancestor had an inferred branch length of only about 5% longer than the median observed substitution rate (Fig. 3.12); and clade accuracy was about 99.8% or around two incorrect clades (Fig. 3.10). Inferred median external branch lengths for neutral treatments were about 40% longer than the observed substitution rates (Fig. 3.12), and although this was consistent across treatments differing by mode of reproduction or ancestor genotype, topological accuracy varied between 98% and 100% (Fig. 3.10).

Specific explanations to explain the inaccuracy of inferred branch lengths are lacking. With random genotype population sampling for the taxa, the overestimation of external branches agrees with such random sampling of segregating variation as being interpreted as fixed differences. However, this should not have been the case for internal branches, which were generally underestimated. This underestimation may be attributable to multiple substitutions occurring at the same loci and the models not sufficiently accounting for this (Sullivan and Joyce, 2005). However, the treatments that produced overestimated internal branches, those starting from the pre-evolved ancestor under selection, were the treatments that had many more invariant sites, as shown in Figures 3.04d and 3.13. While a few sites especially near the beginning of the genome had a much greater turnover rate, there did not seem to be an overall tendency to have a greater number of sites with higher rates of change than on average for these treatments with the naïve ancestor (Fig. 3.04). NJ produced similar internal branch lengths for this pair of treatments, selection with the pre-evolved ancestor, although it showed lower and more accurate branch lengths under asexual reproduction (Fig. 3.12). For this pair of treatments, ML and BI exhibited a great deal of variation in inferred branch lengths across replicates, and for the treatment under asexual evolution, this variation was substantial enough that median values for BI were about 22% lower than the observed substitution rate while ML produced values about 32% greater (Fig. 3.12). Looking across experimental conditions, it is unclear which of these analyses has greater resiliency in inferring branch lengths under the complex biologically realistic phenomena explored here, although all are affected.

The inaccuracy of branch length inference was especially surprising for the neutral evolution treatments. With a lack of natural selection occurring, the molecular evolutionary dynamics should be closer to the models of evolution upon which the phylogenetic analyses are based, and therefore lead to greater accuracy, but this was not the case. Especially considering other found deviations from neutral theory expectations, such as substitution rate and time to

fixation, branch length inaccuracy adds to the evidence that complex population processes in an evolving system are not being captured by models of evolution used in phylogenetics.

This work has demonstrated the possibility of evaluating phylogenetic analyses using highly customized fixed-rate models of evolution constructed from close observations of population genetic history across hundreds of thousands of generations. The model used for each experiment was parameterized using the recorded substitutions from across the ten replicates for that experiment, so substitution rates and state frequencies were highly accurate; much more accurate than biological amino acid model counterparts that estimate values using ancestor reconstruction for molecular datasets (e.g., Dayhoff et al. 1978; Jones et al. 1992; Whelan and Goldman 2001; Le and Gascuel 2008) due to the infeasibility of collecting observed substitution data in biological populations. And the empirical models were as precise as possible given the general time-reversibility assumptions necessary for use with the phylogenetic programs, despite the character polarity for each recorded substitution being known. However, at least for the phylogenetic difficulty evaluated in these treatments, the empirical model of evolution did not consistently yield more accurate results for inferred branch lengths (Fig. 3.12), topological accuracy (Fig. 3.10), or clade support (Fig. 3.11). When model choice made a difference, often the Poisson model outperformed the empirical model, as is especially evident for the internal branch length inferences of the neutral asexual reproduction treatment with the naïve ancestor (Fig. 3.12).

It is unclear why the highly accurate empirical models of evolution performed as poorly, if not worse, than the basic Poisson model. Differences in model performance are not attributable to overfitting, as both types of models have fixed rates and are therefore equally highly parameterized. While “all models are wrong, but some are useful” (Box, 1976; Sullivan and Joyce, 2005), it is unclear why the basic Poisson model functioned relatively well, or at least not substantially worse, than the empirical models. While the Poisson model is accurate with respect to the Avidian mutational model, it would not appear to at all characterize the substitution patterns occurring across experimental treatments (e.g., Table 3.03). Model choice

may especially have significant consequences over greater extents of evolutionary time, for example in the correction of multiple substitutions per site, and long branches should be affected more so than short branches (Felsenstein, 1978; Sullivan and Swofford, 2001). It seems plausible that the symmetrical, equidistant, and short branched topology evolved here did not offer difficulty that models of evolution may improve. Yet the inability for either model to closely resolve both branch lengths and topological accuracy for any taxon-character dataset should be highly concerning to systematists, because modern phylogenetics methods rely on the joint optimization of both aspects of a phylogeny.

Following the analysis presented in Chapter 2 and as first conducted by Hillis and Bull (1993) for their biological experimental phylogenetics research, the relationship between support values and clade accuracy is shown in Figure 3.11. The treatments analyzed in Chapter 2 produced conservative estimates of clade accuracy for support values greater than 30%, except for a single BI support value which was slightly conservative. The selection treatments analyzed here reproduced this result for only support values greater than 50%, and the neutral treatments were the opposite as in being liberal estimates for all such support values. This is a concerning result, as it indicates that the support-accuracy conclusions first shown by Hillis and Bull (1993) and still relied on by researchers (e.g., Sleator 2011) may not hold. The distinction between selection and neutral treatments may indicate that selection cannot just improve accuracy but also improve our estimate of it, but more research is warranted using other experimental designs and in other systems. An alternative hypothesis is that this result is not especially related to the experimental treatment conditions that produced the molecular data analyzed, but rather a factor of the absolute number of true clades insufficiently supported. For example, in the analyses shown in Chapter 2, and also by Hillis and Bull (1993), the maximum number of incorrect clades observed for any replicate was six, which is also the maximum possible for these eight-ingroup taxon evolutionary histories, while for these histories the maximum observed was 31 and maximum possible is 1022. Perhaps these larger trees, which are not exceptionally large compared to contemporary work (Li et al., 2015), have a greater rate

of false positive clade identification owing to their size. The implication of this would be that the greater number of alternative topologies necessitates an even greater bootstrap or posterior support sampling to alleviate the increased random error.

Conclusions

The work presented here demonstrates the potential of using digital evolution to conduct experimental phylogenetics. I have systematically evaluated a few of the complex processes that may affect phylogenetic accuracy. I did so using a much larger experimental design than any experimental phylogenetics research to date, and a much more biologically realistic experimental design than in Chapter 2, with the goal of demonstrating the biological realism, utility, generality, and overall potential of digital experimental phylogenetics. Compared to evaluating phylogenetic methods using simulations alone, this work has produced molecular evolutionary dynamics that closely resemble those observed in biological systems, and these processes led to curious phylogenetic inference results. The work presented here shows that complex population processes occur even in a digital system that is experiencing entirely neutral evolution, leading to differences from neutral theory expectations. When organisms reproduce sexually with recombination, as implemented in Avida, their evolutionary histories can be perfectly inferred, and under asexual reproduction natural selection restores some of the reduced accuracy, suggesting that both recombination and selection may aid phylogenetics. Digital evolution allows the construction of precise models of evolution, but phylogenetic methods were not improved with their use, producing similarly inaccurate branch lengths and clade relationships under very different models. Clade accuracy is not predictive of branch length accuracy for these evolutionary trees, and the clade support metrics of bootstrap support and posterior clade support are not close estimates of clade accuracy.

Analyses of the substitutions that occurred have begun to show the biological realism capable with this system. For example, natural selection's well-recognized effects of altering the distribution of mutational effects for adapted genotypes, reducing the total number of

substitutions that occur due to negative and stabilizing selection, and increasing the number of substitutions that co-occur in selective sweeps were each observed. Similarly, the combined effects of recombination and selection occurring together were observed, such as the fixation of greater numbers of beneficial alleles, greater rates of fixation in novel adaptive environments, and the combining of advantageous alleles onto the same linkage block. And yet, differences from neutral theory predictions indicate that the population genetic and genome evolution processes occurring in an evolving system are complex. For example, with neutral evolution treatments the number of substitutions and time to fixation were significantly different from expectations, and with selection treatments most substitutions exhibited sign-epistasis. Especially since the phylogenetically worst-performing treatments were those conducted under neutral evolution, these results suggest that such complex dynamics may not yet be adequately investigated using simulation approaches alone.

Surprisingly, and at least under the conditions evaluated here, selection and especially recombination restored reduced phylogenetic accuracy under asexual reproduction and neutral evolution. These results suggest an intriguing and novel mechanism by which phylogenetic patterns may be better inferred due to substitutions occurring throughout lineage evolution instead of in simultaneous bursts. This would be a similar dynamic to a molecular clock, but due to natural selection and recombination, and would seem to be relatively more beneficial to phylogenetic inference since neutral evolution treatments showed reduced accuracy. These results point to a novel benefit that recombination and natural selection could provide systematists, albeit these processes in even more complex contexts (i.e., even closer to biological reality) may have more costs than benefit (e.g., Lanier and Knowles 2012; Adams et al. 2018; Pang 2020). For example, recombination here could not disrupt positional homology since genomes had fixed length, so alignment was not required. Still, it would be nice to establish that recombination is not just a “nuisance” for phylogenetics, as it is perceived to be (Martin et al., 2011).

The unexpected phylogenetic benefits of recombination and selection observed here should be approached with a nod towards the generality of the experiment and system. An important caveat to experimental phylogenetics research is that an experiment shows exactly what happened under one set of conditions, but not every set of conditions. And with digital evolution especially, it's important to investigate results in multiple systems, including in biological contexts, to ensure the phenomena isn't system specific. For example, Avida currently implements two very different approaches to recombination, and it would be helpful to add a third setting to better approximate recombination rate as interpreted by biologists. Currently, one approach is to configure the genetical system for the number of equally sized modules, with each module having independent assortment from one another; this was employed in Chapter 2 with the number of modules equal to the genome length so that all loci exhibited independence. A second current approach is to have a single module but with two recombination breakpoints chosen randomly in the genome; this was the approach for the experiments in this chapter. A not-yet-implemented third approach would be to alter the number of randomly chosen paired breakpoints: zero (no recombination), one (as used here), and so on up to the size of the genome less one which would entail breakpoints between every locus and be equivalent to independent assortment among all loci. The benefit to this approach would be the ability to increase the recombination rate as biologists interpret it to have better control over how the evolving population approaches linkage equilibrium. For example, humans have much larger linkage blocks than do species like maize or *Arabidopsis thaliana* (Rafalski and Morgante, 2004; Wall and Pritchard, 2003), and with this new setting experimenters could better match such differences in molecular evolution otherwise produced by differences in life history or evolutionary history.

The influence of selection on phylogenetic accuracy might also be system dependent. In Avida, selection greatly improved phylogenetic accuracy under asexual reproduction conditions. However, biological experimental phylogenetics research tends to show that selection inhibits phylogenetic inference by producing homoplasy (Bull et al., 1997; Cunningham et al., 1997;

Fares et al., 1998), although see Leitner et al. (1996). The accuracy of inferences of digital evolution histories have been shown to be aided by selection via the production of more synapomorphies (Hagstrom et al., 2004; Hang et al., 2007, 2003). Unfortunately, the experiments presented in this chapter did not offer an ideal test of this, since these selective environments were explicitly designed to promote diversifying selection throughout the entire evolutionary history, so by design there should be greater synapomorphies. A hypothesis that remains to be tested is that the diversity in the potential genotype-to-phenotype map for digital organisms such as Avidians may make them especially unlikely to produce homoplasious evolution. For example, there are relatively few constraints to the genetic programming such that there are innumerable genotypes that code for the same phenotype. However, the impression that prompts this hypothesis may be biased since digital evolution experiments tend to start with a naïve ancestor, and such a "blank tape" genome would be especially unburdened by genotype-phenotype limitations due to historical contingency or other constraints. Further work exploring these trends of selection's impact on phylogenetic inference among both digital and biological systems is needed.

Digital evolution has great utility in allowing the tracking of data that would be impossible or incredibly burdensome to collect for biological systems. Digital evolution allowed the construction of models of evolution tailored to the substitution rates and state frequencies of the evolving populations under study with much greater precision than has been accomplished for biological systems. The sensitivity of model-based approaches was evaluated in this study with great accuracy while still maintaining the increased biological realism of experimental phylogenetics approaches. Although the model of evolution made little difference in the work presented, perhaps this would not be so under conditions dissimilar from the fully symmetrical and ultrametric evolutionary histories evolved here. Future work to this end could explore how model choice becomes increasingly important under varying adaptive evolution dynamics in addition to differently long branches, asymmetrical topologies, and other evolutionary divergence conditions. The construction of empirical models of evolution has

fulfilled the expectation of Hagstrom et al. (2004) who remarked that it "will be available in the near future" for Avidian systems. Although, the work presented here has not attempted to determine a single model widely applicable to differently parameterized Avida experiments. This may be a fruitful research direction that, as predicted by Hagstrom et al. (2004), would "further energize research."

The phylogenetic accuracy trends presented by these results are curious and their implications should be further investigated. Although clade accuracy was high across all treatments, it was comparatively low for the neutral evolution treatments. The treatments conducted under neutral evolution with asexual reproduction and inferred using the empirical model of evolution were predicted to have the greatest clade and branch length accuracy, since the evolutionary dynamics occurring should have been closest to the model of evolution and other assumptions used in the analysis. Yet these were among the lowest performing. Phylogenetic methods such as maximum likelihood and Bayesian inference iteratively optimize tree topology and branch lengths concurrently to produce the overall best tree. However, there was little relationship between clade accuracy and branch length accuracy in these analyses, perhaps suggesting room for improvement in the model of evolution or its use. And yet the use of very different models of evolution made little difference. Finally, the common metrics of clade support, bootstrap support values under maximum likelihood and posterior clade support probability under Bayesian inference, were not close estimates of clade accuracy, and their being under- or overestimations depended on whether natural selection was occurring. More research questions and hypotheses have been proposed than answered with these case-study experiments, and digital evolution will be a valuable tool among others in helping phylogeneticists to address them.

The T7 phage work of Hillis et al. (Hillis et al., 1992) presented the experimental approach in order to inspire work that "will fill an important void in the science of phylogenetic reconstruction." While the paucity of experimental phylogenetic research to date has left this void largely unfilled, the work presented here aims to spur greater research by presenting the

potential contributions digital evolution may bring. This potential rests on the approach's greater biological realism than simulations and greater utility and generality than biological experimental systems. By navigating these inherent tradeoffs in conducting phylogenetics research differently than past approaches, digital evolution just may "become the wave of the future in phylogenetics" as its trajectory was envisioned two decades ago. As we build better computational machinery and discover more about the underlying facets of molecular evolution, digital evolution has the potential to become more and more powerful and therefore relevant – for if reality is but a simulation (Bostrom, 2003) then what is digital evolution if not a proto-reality?

CHAPTER 4:

Shifting Student Understanding of the Importance of Variation for Evolution in a Course Featuring Digital Evolution

Introduction

Natural selection, and the basics of evolutionary change generally, are "staggeringly simple" (Coyne, 2010), "breathtakingly simple" (Chown, 2013), and so simple that Huxley remarked that it was stupid of him to not have thought of it (Huxley, 1887; Kalinowski et al., 2016). Evolution requires three concepts – variation, heredity, and differential reproduction (Godfrey-Smith, 2007): Mutation, heritable from parent to offspring, is the ultimate source of genetic variation that may affect phenotypic variation and variation in reproduction. Natural selection is the resulting process if phenotypic variation non-randomly changes in a population due to its effect on reproductive ability, and genetic drift is the resulting process if variation randomly changes due to random differences in reproduction (Gregory, 2009; Tibell and Harms, 2017). Yet evolution remains elusively difficult to teach, likely because "evolution, in a way, contradicts common sense" (Mayr, 1982) and "is probably one of the most counterintuitive ideas the human mind has encountered, so far" (Evans, 2008). Yet evolution is "the single best idea anyone has ever had" and is therefore worth the tremendous effort to teach well (Dennet, 1995; Gregory, 2009).

Among the many difficulties in understanding evolution, a large set of them are ultimately related to those borne from misapplying the intuitive reasoning strategy of essentialism to understand the concepts of species, inheritance, mechanisms of evolutionary change including natural selection, and the crux of the matter – the importance of individual-level variation for evolution. Essentialist thinking applied to a species entails individuals who are united by something that makes them unique, and this essence precludes evolutionary change. Taken in an absolute sense, essentialist thinking underlies creationist claims that a present-day species cannot be derived from any other species (Evans, 2008).

Essentialist reasoning can be incorporated into a conception of evolutionary change as transformationalism, which is distinguished from the scientifically valid explanation referred to as variationalism (Mayr, 2001). In transformationalism, evolution is defined as the change in the overall characteristics of one generation to the next. In variationalism, within-population variation is of utmost importance, so evolution is commonly defined as the change in the frequency of alleles, i.e., the genetic constituents of variation, over time (Freeman and Herron, 2014). This distinction of valuing individual-level differences in one generation *and* population-level differences between multiple generations in variationalism versus only valuing the latter in transformationalism may seem slight, but it is integral to understanding the mechanisms of modern evolutionary biology, especially natural selection. For example, differences between individuals are consequential enough to cause evolution under variationalism via differential reproduction, but under transformationalism, any difference between individuals is immaterial to evolution because it would be impossible for individuals to differ enough to produce such an effect.

I hypothesized that the direct exposure to and experimentation regarding variation and its importance using integrative thinking, statistical reasoning, and computer modeling via Avida-ED would produce transformational-to-variational shifts in student understanding. This hypothesis was tested within the whole-semester course context of a novel undergraduate introductory biology course, *Integrative Biology: From DNA to Populations*. Herein, I discuss a few of the difficulties in understanding, and therefore teaching, evolution, focusing on essentialism and its evolutionary formulation as transformationalism as contrasting the modern scientific formulation – variationalism. I then describe the features of the Avida-ED digital evolution platform that allow experiential learning of variationalism and outline further motivation for the curriculum development of the course. The course features addressed here, of which the digital evolution lab component is both key and one of many that address student understanding of variation, may collectively contribute to shifting student understanding of biological variation from transformationalism to variationalism. I describe the development,

use, and results of a modified transformationalism-variationalism assessment in a pre-/post-course design to measure this shift. And I conclude with suggestions regarding the incorporation of instructional material that engages students in developing a scientifically-modern understanding of variation and the evolutionary process.

A few of the difficulties in understanding and teaching evolution

Instruction in evolutionary concepts is not a matter of addition to students' existing set of knowledge, but rather fundamentally altering the way they see the world – an ontological shift (Sinatra et al., 2008; Tibell and Harms, 2017). Difficulty in understanding evolution and natural selection at least in part may rely on incorrect systems thinking, or misattribution at the appropriate level of organization (Chi, 2005). Humans are accustomed to direct, causal processes that typically have a central controlling agent that interacts in a series of sequential steps from a beginning to endpoint. Conversely, emergent process-type phenomena like evolution result in hierarchical systems "where patterns in a collective are generated by interactions between agents at the lower level" (Cooper, 2017). A failure to understand statistics and randomness can magnify the effects of misapplying systems thinking (Cooper, 2017; Fiedler et al., 2019; Speth et al., 2014; Tibell and Harms, 2017) – how could one understand that patterns such as distributions of traits can result from emergent processes based, at least in part, on random interactions among lower-level units? Additionally, Tibell and Harms (2017) point out that applying systems thinking in understanding evolution across large temporal and spatial scales is especially difficult.

A relatable inability in applying systems thinking is the vernacular of evolutionary biology, which includes many terms that have a different common folk application. The folk usages of "adapt" applies to individual organisms and occurs within a lifespan, whereas the scientific usage applies to a population across generations of evolution (Coley and Tanner, 2015; Shtulman, 2006). Biologists may seem to apply agency or intentionality to natural selection by using words like "force" or "pressure" even though they do not actually mean that natural selection is a direct causal process (Cooper, 2017; Gregory, 2009). Among yet other

misconceptions (also called naïve intuitions) are ideas related to organisms evolving in response to a need or the use/disuse of body parts (Coley and Tanner, 2012). These are just a few of the persistent folk ideas that differ from modern scientific thinking.

Increased undergraduate biology education, at least as has been implemented and assessed, can sometimes lead to the suppression but not supplantation of intuitive means of understanding (Shtulman and Valcarcel, 2012), or even strengthen misconceptions based on misapplications of fundamental modes of understanding (Coley and Tanner, 2015). So-called "intuitive conceptual systems," or mental shortcuts, are integral to helping us understand the world, especially as we develop this understanding as children, allowing us to simplify, categorize, and otherwise reduce the amount of information we process to make sense of the complexity that we encounter (Heddy and Sinatra, 2013; Sinatra et al., 2008). For example, knowing that an organism is a member of a certain species allows you to reliably predict (at least some of) its properties (Shtulman and Schulz, 2008). However these mental shortcuts cannot be applied in every application in which we might like to use them, and when we apply them incorrectly, they entail a serious cost as persistent and overarching misconceptions (Evans, 2008; Sinatra et al., 2008). Evolution is perhaps a quintessential example of this, with an array of common misconceptions (Gregory, 2009) stemming from multiple underlying cognitive construals (also called cognitive biases), including essentialism, teleology, intentionality, and anthropocentric thinking (Coley and Tanner, 2012; Evans, 2008; Sinatra et al., 2008). Within the field of evolutionary psychology, cognitive construals have been thought of as complex evolutionary adaptations for dealing with complex environments (Geary, 2007), and the stickiness or resistance in *not* applying these ideas by changing one's intuitive understanding may be an adaptation too (Sinatra et al., 2008).

Essentialism is the general idea that things can separately be placed into real and named categories (Gelman, 2003) that exhibit an underlying innateness (Knobe and Samuels, 2013), essence (Evans, 2008) or hidden causal power (Shtulman, 2006), whether or not we know, or can know, what that underlying nature is (Coley and Tanner, 2012). Crucially, this essence is

immutable or unchanging. Essentialist thinking can be applied to multiple levels of biology and is overall the idea that a core underlying facet of a level of organization (e.g., biological structure, species, or system) determines it's features and persistent identity (Coley and Tanner, 2015).

Essentialism is an example of natural history theory recapitulating the ontogeny of human cognitive development (Shtulman and Schulz, 2008). As recorded by Aristotle, in ancient Chinese thinking, and in folk ideas throughout other cultures, humans have long exhibited an essentialist understanding of living organisms (Evans, 2008; Shtulman, 2006). This was a formidable historical obstacle to modern evolutionary theory (Kalinowski et al., 2016), and Darwin spent great care dismantling it by writing at length on the importance of individual-level variation (Gregory, 2009). This understanding is also exhibited by children, especially until about 8-10 years old, after which they may be able to start to understand concepts like modern scientific notions of common descent and other variationalist concepts (Evans, 2008; Sinatra et al., 2008). This conceptual change is like a scientific revolution experienced within their own personal understanding (Sinatra et al., 2008).

Transformationalism and variationalism

In essentialism a species has an immutable essence that defines its nature; in transformationalism this essence is mutable and its change over time is evolution. Under transformationalism, individuals in a population change *all-together* over time because the underlying essence is evolving, for example all individual moths in a population becoming successively darker each generation. Under the scientifically-modern understanding of variationalism, genotypic (or phenotypic) individual variation changes *in frequency* within a population across generations, for example a black moth variant increasing in population frequency over generations (Shtulman, 2006). Note that someone with a transformational understanding is not blinded to variation within a population. They may admit that slight variation exists between individuals, but that this variation is inconsequential to the evolutionary process, since evolution must involve a change to the shared essence of the

species and therefore must act on all individuals (Coley and Tanner, 2015). Under both transformationalism and variationalism, individuals vary in terms of their genetic information (i.e., genotype) and/or expressed characteristics (i.e., phenotype), and over time generations of organisms change (i.e., evolve). Yet these conceptualizations vary in important respects.

In transformationalism, evolution acts in a single step where the environment directly affects the essence of all individuals of a species (Gregory, 2009; Shtulman, 2006). Modern molecular reductionism in biology may even provide a pseudoscientific species essence in the guise of DNA, genes, and chromosomes (Coley and Tanner, 2012; Fodor, 1998; Gelman, 2004). In my opinion, this could be at least one means by which increased biology education may reify essentialist reasoning for some students. In comparison, variationalism requires at least two steps: Mutation is the originator of variation by creating a variant allele in only one individual in the population. This is followed by an evolutionary process (e.g., natural selection or genetic drift) that operates to change the frequency of this variant in the population over successive generations (Gregory, 2009; Speth et al., 2014). Highlighting the importance of mutation as the source of variation, Gregory (2009) asks the rhetorical question, "How can an eliminative process like natural selection ever lead to creative outcomes?"

A byproduct of essentialist thinking is the persistent undervaluation of within-species variation; although individuals within a population may vary in some respects, this variation is limited and not important in an evolutionary context (Coley and Tanner, 2015; Shtulman and Schulz, 2008). Species are "manifestations of an underlying essence in which variability is irrelevant noise" (Coley and Muratore, 2012). Shtulman (2006) takes this further by claiming that transformationalists would intuit that any individual variation must be non-adaptive or maladaptive. In variationalism, differences between individuals are of critical importance (Coley and Muratore, 2012) and are characterized in terms of population frequencies, i.e., the number of individuals of the generation with a certain characteristic. In transformationalism, the population average (or something similar) is ontologically reified, i.e., made real (Gould and Duve, 1996; Shtulman, 2006). In variationalism, the population average is an abstraction and

not something real in and of itself (Coley and Muratore, 2012); it is a statistical property of a distribution of separate individual-level features, an emergent pattern (Cooper, 2017).

Variationalism requires the shift in thinking of natural selection not as an event but as an unbounded process (Gregory, 2009; Sinatra et al., 2008). Evolution is an emergent process at the population level resulting from random, undirected mutation occurring to individuals and inheritance of this genetic variation that may be expressed phenotypically as, ultimately, variation in reproduction among those individuals in the population (Gregory, 2009). In this way, natural selection results from a series of contingent events and continues unendingly (Cooper, 2017). Transformationalism is much simpler and intuitive because it uses folk understandings of causation – individual-level events like mutation and natural selection are neither necessary nor consequential, what matters is the direct inheritance among all members of the population of those traits which are suited to the environment (Sinatra et al., 2008). Further, transformationalism may lead to saltationist notions of complex traits arising suddenly in a single generation in response to a need (Gregory, 2009).

Trouble dislodging transformationalism

Biology education research has shown that novices tend to hold transformational views, and that it can be difficult to transform their thinking towards variationalism. As with other ideas that have a firmly intuitive basis, such as teleological reasoning, the transformationalist idea of a species having a shared essence is a sticky concept – one that is often impervious to dislodgement (Speth et al., 2014). Studies have shown that large proportions of children (Evans, 2008; Shtulman and Schulz, 2008), high school students (Furtak et al., 2014; Shtulman, 2006), and college students (Richard et al., 2017; Shtulman, 2006; Shtulman and Calabi, 2008) hold a transformational if not essentialist understanding of biological variation. Shtulman and Schulz (2008) found that only people with variationalist thinking could, across a majority of taxonomic and trait-class examples, affirm that within-species variation is both prevalent and probable. When expert and novice biologists have been asked to explain natural selection, a majority of experts exhibited variationist thinking by valuing this first step—the origin of heritable

variation—enough to include it in their explanations of natural selection while only 10% of undergraduates did so (Nehm and Ridgway, 2011; Speth et al., 2014). Coley and Tanner (2015) reported that a majority of non-biology major undergraduates agreed with an essentialist statement, while a minority of biology majors did, perhaps demonstrating a predilection towards opposing essentialist thinking with greater innate interest or curiosity if not increased expertise alone. While Shtulman and Schulz (2008) argue that an understanding of variationalism should be considered a necessary condition for understanding natural selection, other research has demonstrated that instruction in evolution, especially the variational processes of mutation and natural selection, is insufficient in altering these misconceptions (Gregory, 2009; Kalinowski et al., 2016). Instead, these authors report that it is most important to draw students' attention to variation between individuals and to provide examples of how important that variation is for evolution.

When helping students shift from transformationalism to variationalism, the incorporation of new information into students' mental models may result in mixed models of understanding (Evans, 2008). Shtulman (2006) even termed such models as "pre-variationalism" on a continuum between transformational and variational modes; although the learning progression of Furtak et al. (2014) had separate dimensions for "transformationist incorrect" and "variation," suggesting disagreement among researchers regarding the mutual incompatibility of these modes of understandings. The difficulty in fostering a shift in student thinking, and the likelihood of mixed model formation, is related to the cognitive "stickiness" of the essentialist intuitive reasoning strategy, which is innate and impervious to change (Gelman, 2004; Speth et al., 2014). Although difficult, Shtulman and Calabi (2008) describe a decrease in undergraduates reporting transformational understanding following instruction on Darwin's formulation of evolution by natural selection. However, Richard et al. (2017) discovered no significant differences between introductory and advanced biology majors' agreement with essentialist statements. Concerningly, the results of Coley and Tanner (2015) suggest that formal biology education can reify the relationship between essentialist intuitive reasoning and

essentialism among students who hold such views, which could strengthen the hold of transformational understanding among advanced biology students.

Experimentation in Avida-ED emphasizes individual variation

Evans (2008) describes the need of an instrument to observe evolution happening in action:

> "If we could speed up time, we would 'see' species as dynamic and biological change as contingent and non-directional; in effect, species would morph from one to another as environments change, or disappear entirely. Yet everyday cognition, mired as it is in a particular time and place, appears to obstruct this view of a dynamic world. What is needed is the equivalent of a microscope or telescope, such as a time-machine that transcends human cognitive and perceptual limitations."

Although we do not yet have such a machine to observe the evolution of biological organisms, we might have the next best thing in the Avida-ED model system, which allows students to readily observe the processes of evolution in action for digital organisms. By having a simplified genetic model and the color-coding of genotypic, phenotypic, and fitness differences, students observe that variation among individuals is created from parent to offspring and that this variation exists within a population context. And by engaging in what is occurring within rapidly changing populations, students have an opportunity to recognize that variation among individuals not only exists but that it is the raw material of evolution. Therefore, Avida-ED may be especially illustrative of variationalism.

Digital organisms have a greatly simplified genetic model compared to biological organisms. This may aid in highlighting genotypic differences between individuals as such variation may be obfuscated by complex genotype to phenotype relationships. In biological organisms, genetic information is transferred across multiple levels of organization and expression within an organism, for example, from DNA to RNA to protein and other molecular machinery to expressed phenotype. Avidian genetics is limited to the information transfer

between the instruction sequence to programmatic machinery (i.e., interactions between instructions' execution) and the expressed phenotype. This provides a simplified model that may lead to increased understanding of genetic differences by allowing students to focus on differences between organisms rather than focusing on the transfer of genetic information within an organism. Further, the myriad causes, types, and effects of mutations in biology add further complication to understanding genotype to phenotype relationships (Tibell and Harms, 2017). Avidians and Avida-ED provide a straightforward representation of how differences in genotypes are expressed as differences in phenotypes (Smith et al., 2016).

In Avida-ED, genotypic variation between individuals, as well as the cause of that variation, is readily observable. Using the organism viewer, students observe the life process of an Avidian, including reproduction. During reproduction random mutation may create variation in the genome of a single organism—the offspring—leaving the parent unchanged. Transformationalism would require a mutation to occur for both the parent and its offspring, or at least all of a population's progeny born at the same time or in the same generation. Differences between the parent and offspring are observed by evaluating changes in the labeled sequence of instructions, with mutations outlined in black. These genomic sequence differences are also often reinforced with a color difference, since sets of Avidian instructions are color-coded according to their approximate computational or programmatic function.

In Avida-ED, as in biological systems, single mutations may have an appreciable likelihood of resulting in consequential phenotypic and fitness effects in an evolving population. Using the population viewer, sets of organisms with discrete phenotypes (i.e., computational task completion) may be outlined in color, and the number of organisms performing each function is displayed among the population statistics. By default, organisms in the population are color-coded by fitness, and the average fitness of a population is displayed among the population statistics. Slight fitness differences often exist even for organisms with shared phenotypes, since genotypic differences may have consequences for reproduction efficiency (i.e., offspring cost). These differences are easily observable due to the high sensitivity of the

fitness color scale. Instead of a transformational depiction of all organisms having similar if not identical genotype, phenotype, fitness, and therefore color, a great deal of variation at each of these levels exists among contemporaneous individuals, producing a multitude of colors at any given timepoint.

When the environment is configured to reward certain tasks (in Avida-ED verbiage, when the resources associated with tasks are present in the environment), then the gain or loss of these phenotypes has significant fitness consequences. Avidians have short lifespans and rapidly reproduce to create large populations that experience hundreds of generations within minutes. Instead of a transformational picture of the entire population gradually changing color altogether over generations, differences among individuals are readily apparent as spatial, patchy changes in color-coded fitness over seconds or minutes. Slightly more efficient Avidians evolve over time, slowly altering the color across the population as their descendants with greater relative fitness increase in frequency; and Avidians with *de novo* mutations for task performance are observed to quickly produce similarly fitness-colored offspring, drastically changing the genotypic, phenotypic, and fitness makeup of the population. This provides a powerful visualization of variationalism as colors appear to rapidly evolve (i.e., change frequency) onscreen. Of course, it is not the color that is evolving, but rather the population of digital organisms.

This presentation of fitness driving the evolution of a population is key. However, fitness in biology can be extremely difficult to measure and track over time. In Avida-ED, fitness is automatically recorded for both individual organisms and the population average, and its change over time is graphed dynamically. These data are quantitative and readily exportable for further analyses. The lighter-colored genotypes produce more offspring and the population evolves accordingly – genotypic differences are driving fitness differences that drive the evolution of the population. Further, students observe this change quickly, being able to see how fitness at one time point affects a later time point by observing a series of intermediates.

Motivation for curriculum

Over the last couple decades, interest in modernizing science education in the United has reached a critical mass, resulting in the publication of several calls for instruction, curriculum, and assessment reform. With respect to biology content, calls for reform emphasize the curricular centrality of evolution as a major unifying theme. And they emphasize variational understanding of the evolutionary process. For example, as established in "Next Generation Science Standards" (2013), instruction in variationalism begins in first grade by establishing that phenotypic variation exists within a species. In third grade, the connection between genotypic and phenotypic variation is established, and, separately, differential fitness leading to adaptation. Middle school life sciences curriculum includes that mutation is the ultimate source of variation, that variation due to mutation may affect fitness, and that natural selection changes phenotypic frequency within a population. In high school, the requirements of natural selection, heritable variation causing differential fitness, are more greatly explored, connecting concepts established across primary and secondary education. Similar emphasis on individual-level within-population variation being required for evolution to occur is found within the life sciences core ideas in "A Framework for K-12 Science Education" (2012a) and is the first big idea in "AP Biology Curriculum Framework" (2011).

The most relevant reform product in the undergraduate biology curriculum space is "Vision and Change in Undergraduate Biology Education" (2011), which identifies five core foundational concepts that unify biological knowledge. Two of these, Evolution, and Information Flow, prominently feature variationalist understanding, as exhibited within the content-specific statements offered in the BioCore Guide (Brownell et al., 2014), which operationalized "Vision and Change" for a general biology curriculum.

Calls for reform also emphasize the importance of science and engineering practices in STEM education through the engagement of students in inquiry-based and research-based pedagogies (Auchincloss et al., 2014; Corwin et al., 2015; Linn et al., 2015; Weaver et al., 2008). For example, "A Framework for K-12 Science Education" (2012a) identified eight science and

engineering practices. These practices embody the means by which scientists discover and investigate the natural world and by which engineers evaluate and design the built world (National Research Council, 2012a), and are applicable in undergraduate education (Cooper, 2013).

Another national report, "Thinking Evolutionarily: Evolution Education Across the Life Sciences" (2012b), identified inquiry-based laboratory experiences as being especially key for helping students gain a better understanding of evolution. Furthermore, it has been shown that exercises and labs that directly address common misconceptions can be the most successful at overcoming the cognitive dissonance occurring when students' preexisting views are challenged (Grant, 2009; Robbins and Roy, 2007), a factor of special and significant concern for evolution education (Nelson, 2012). "Vision and Change" (2011) additionally highlighted that "themes of adaptation and genetic variation provide rich opportunities for students to work with relevant data and practice quantitative analysis and dynamic modeling."

Collectively, these documents call for a revolution in how students are taught, with best practices being both evidenced-based and backed by sound theoretical rationale. The Avida-ED lab curriculum has been designed to address the recommendations and best practices put forward by these documents.

Primarily four curricular components have been suggested to produce transformational-to-variational shifts in understanding. Since this is a profound conceptual change, Sinatra et al. (2008) proposed that direct exposure to the evolutionary phenomenon in class, and especially with a high degree of experimental engagement, has the greatest likelihood of success. Shtulman and Schulz (2008) and Shtulman and Calabi (2008) recommend that instructors make a point to highlight within-species variation across a multitude of traits. Cooper (2017) stresses the importance of introducing statistical reasoning and computer simulations or modeling as means to distinguish between the individual and population level understanding of emergent processes like natural selection. And Kalinowski et al. (2010) suggest that explicitly connecting molecular genetics concepts to evolutionary phenomenon, or further, connecting biological

ideas across fields of biology, especially molecular biology and genetics (Smith et al., 2009), may allow students to "make sense of the entire process from genes to populations" (Nehm et al., 2009; Speth et al., 2014). Engaging across biological fields in this manner is referred to as "integrative thinking" (Wake, 2008). For these reasons, I hypothesized that a course experience featuring direct exposure to and experimentation regarding variation and its importance (i.e., the introduction via mutation and change via population processes) using Avida-ED, and especially embedded within broader integrative biology contexts, would shift students from transformational to variational understanding.

Methods

University and course population context

Michigan State University is a large public research-intensive university located in the Midwestern United States, with an enrollment of 39,000 undergraduate students across fourteen colleges. While several introductory biology courses are taught, *Integrative Biology* was uniquely designed to be a one-semester course that explores many aspects of what is more commonly a two-semester introductory biology curriculum while maintaining the intellectual rigor required of STEM majors. We, principally Louise Mead and I, tailored the course to its target population of STEM students, consisting of engineering and non-life sciences majors needing only one semester of an introductory biology course and not requiring a separate laboratory course.

Fall 2016 through Summer 2018, a total of 354 students took the course, with one course section offered each of the six semesters (Table 4.01). A total of 95 student group research projects and poster presentations were produced at the culmination of these semesters. In-person classes met three times weekly for 50-minute learning sessions, one of which was reserved for the digital evolution lab, for 16 weeks; and online-only classes used an online video conferencing platform, Zoom, to facilitate individual student groups in meeting three times weekly for one-hour active learning sessions, two of which were reserved for digital

evolution labs, for seven weeks. Demographically, the students were primarily first-year (55%) or second-year (24%) students, and most were men (80%). Students were largely enrolled in the College of Engineering (60%) or College of Natural Sciences (29%), with the remainder enrolled across seven other colleges. Students majoring in 36 distinct major concentrations completed the course, with 85% of these being non-life sciences STEM majors, and 57% being either Mechanical Engineering, Computer Science, Mathematics, or Physics majors.

Table 4.01. Enrollment in *Integrative Biology: From DNA to Populations* in six consecutive semesters (N = 354).

Year	2016	2017			2018	
Semester	Fall	Spring	Summer	Fall	Spring	Summer
Enrollment	86	72	16	74	74	32

Course design

Avida-ED digital evolution lab

The accessible presentation of variationalism in Avida-ED is reinforced by the instructional circumstances of the digital evolution lab experience. The lab book is grounded by a core curriculum sequence of five activities culminating in an independent group-based research project (Smith et al., 2016; Kohn et al. 2018). The core curriculum has been classroom-tested in full and in part in undergraduate courses and high school Advanced Placement courses across the United States (Smith et al. 2016; Kohn et al. 2018; Lark et al. 2018). Avida-ED curriculum resources, distributed as ancillary to the lab book, include a popular science article on Avidians (Zimmer, 2005), the Avida-ED Quick Start Manual (accessible from the Help toolbar in Avida-ED), and curriculum activities exploring additional topics, such as genetic engineering and mutation rate evolution (Johnson et al., 2011a, 2011b; Lark et al., 2014).

The first five activities, constituting the core Avida-ED curriculum, are designed to support student understanding of evolution, the Avidian experimental system, and experimentation in science generally. First presented by Smith et al. (2016), the activities were expanded and modified by Kohn et al. (2018) to include an additional exercise on genetic drift,

further develop student expertise in the features of the Avida-ED system, incorporate assessment of science and engineering process skills, and otherwise expand the treatment of evolutionary concepts. Student familiarity with the system is jointly scaffolded with fundamental biology concepts. The biology content supports a conceptual progression from the initiation of differences between individuals in a population to the consequences of those differences: the origin of biological variation via mutation, the randomness by which mutations occur, the environment-specific concept of absolute fitness and population-specific concept of relative fitness, and finally the evolutionary processes of natural selection and genetic drift, highlighting the effects adaptive and non-adaptive change have on variation in an evolving population. The modern scientific understanding of variationalist evolutionary progression is emphasized, experimented on, and discussed throughout each of these core curriculum activities.

Being a true experimental study system, each of these activities involves the collection of novel data, and the results of any given Avida-ED lab experiment will likely not be as the instructor would predict for each individual student. Data sharing, statistical analysis, and classroom discussion uncovers the processes underlying the biological and experimental replicate variation. Since the aggregate results for each activity is highly predictable to the instructor although not to the average student, this portion of the curriculum fits the inquiry as opposed to traditional model of research authenticity (Fig. 4.01). Avida-ED activities include a ready means of online data collection, analysis, and presentation using established Google Sheets files. This allows students to compare their relatively limited set of data with that collected by a much larger set of students. By contributing their de-identified data, students can see their contributions towards an accumulating research dataset. These methods provide an excellent opportunity to discuss the importance of sample size when investigating phenomena fraught with experimental variation and, further, the potential for human error. For example, the accumulated results for the first lab exercise show two systematic differences from theoretically predicted outcomes. Such results are organic to the research experience, being

explainable by rare, though repeated student methodological and data transcriptional errors. The ensuing conversation highlights the reality of science being a human endeavor.

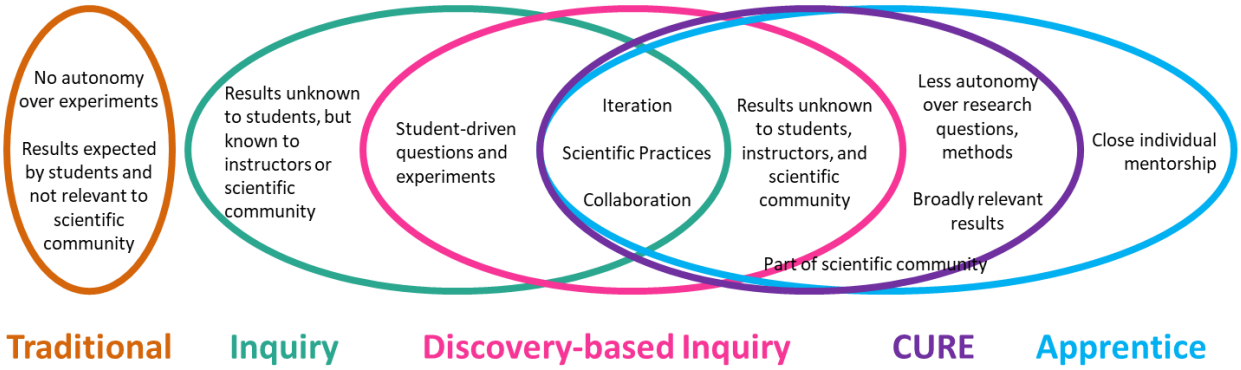


Figure 4.01. Models of research authenticity compiled from the literature on modes of investigation in laboratory environments (Ballen et al., 2017; Buck et al., 2008; Corwin et al., 2015; Domin, 1999; Goodwin et al., 2021; Seymour et al., 2004).

In Avida-ED curricula such as that used in the present study, after completing the core curriculum activities and gaining a familiarity with the model system specifically and experimentation in science generally, groups of students conduct independent research projects. Students are expected to draw connections across research systems by exploring biological phenomena within the digital experimental evolution system, and by testing evolutionary mechanisms that cannot typically be addressed during other types of biology lab experiences. This aspect of the Avida-ED curriculum highlights the full potential of the model system by encouraging students to conduct their own experiments, all while exploring the nature of scientific reasoning itself. For instance, students can change environmental and other variables and perform controlled experiments to test their own evolutionary hypotheses. Students can see for themselves how evolutionary hypotheses can be supported by empirical tests. Thus, students learn first-hand that scientists base conclusions upon repeatable empirical observations to construct arguments from evidence. By doing so, Avida-ED provides an environment for students to directly confront and correct their misconceptions about the scientific status of evolutionary theory and about the nature of scientific practices.

Student groups first brainstorm a research question and take ownership of the resulting research. Early in the process, each group prepares and presents a research proposal; this requirement is designed to prompt careful planning, and it allows the instructors an opportunity to provide formalized feedback early in the process when it can hopefully have the greatest impact. Feedback via critical yet encouraging guidance continues throughout the semester and, as is similar for other experimental evolution labs (Cotner and Hebert, 2016), is often crucial to helping students overcome common experimental pitfalls – for example, basic misunderstandings regarding Avidian biology, experimental design flaws, or failure to investigate phenomena left unexplored in the introductory exercises. Depending on the originality and experimental design creativity of the students, a proportion of experiments each semester investigate phenomena whose results are unknown to the instructor and potentially the scientific community broadly, and therefore clearly fit the discovery-based inquiry model of research authenticity (Fig. 4.01). The remaining experiments are inquiry-based, with results expected by the instructor with varying degrees of accuracy. Based on the sophistication of students' biology expertise, published research using biological systems may be sought by the students during their research process or provided by the instructors. Related publications may provide students with conceptual or experimental design inspiration for their research, in addition to reinforcing the analogy of Avidian experimentation with biology research in other living systems. Before completing the final draft of their research poster, students participate in a peer review process; this provides an opportunity for students to engage in the critical assessment of the content and practices involved in the work of others and, ideally, to metacognitively examine their own work. The conclusion of the digital evolution laboratory experience is a public poster presentation session in which student groups present their work to one another, the instructors, and members of the university community. This professional presentation opportunity is similar to that of a scientific conference poster session.

Other course features

The *Integrative Biology* curriculum is exemplified by its name – using integrative thinking (Wake, 2008) to unify the disciplines of biology across spatial, temporal, and taxonomic scales to holistically examine biological systems and processes. This provides context to and strengthens the variationalist patterns experimented upon in the digital evolution lab. A case-based approach was used with evolution as the organizing force across both the biological world and the course material. As Theodosius Dobzhansky (1973) famously remarked, "nothing in biology makes sense except in the light of evolution." The otherwise sundry biological minutiae and content is best understood and appreciated through the evolutionary lens. As such, evolution is an explicit theme of *Integrative Biology*, which consists of a sequence of seven content units. Each unit, except the first on tree-thinking, focuses on a specific biological system (for example, Caribbean anole lizards) and the scientists actively conducting that work. Integrative thinking allows the students to holistically investigate the biological phenomena of each unit. For example, students may examine the progression of a biological trait, exploring its genotypic basis, genetic expression, intracellular function, and resulting tissue- and organismal-evident phenotypic expression, and finally its population-level change due to an evolutionary process like natural selection (White et al., 2013). Other example progressions entail a different suite of biological subfields: organismal trait expression, population-level frequency variation, ecological mechanisms, and finally ecosystem effects. For each case, evolutionary origins, mechanisms, and/or consequences are investigated. This integrative approach emphasizes variationalism by highlighting the evolutionary consequences of variation among individuals in a population, and across a range of taxonomic diversity. Further, these biological examples reinforce the processes and patterns observed in the digital evolution lab using explicit analogies and discussions connecting the systems.

In addition to their research with Avida-ED, students interact with authentic biological data collected by the scientists researching the biology of the cases under study. Depending on the activity, students work with this data at multiple stages of the scientific process: selecting

which variable to research, collecting data, summarizing data statistically, graphing results, interpreting evidence, and/or presenting conclusions. One activity is meant to demonstrate how meticulous data collection in science can often be ("Lizard Evolution Virtual Lab," 2014). In this work, students measure several phenotypic characters of individuals among related species. Of course, it also prompts them to become familiar with individual-level variation – variation that is then discussed in terms of the macroevolution of several related species.

Fulfillment of calls for curriculum reform

Throughout the Avida-ED lab curriculum, students engage with each of the eight science and engineering practices identified in "A Framework for K-12 Science Education" (National Research Council, 2012a), including both the science and engineering-specific framings (Kohn et al., 2018). However, some practices incur lesser student engagement in the core activities as compared to the independent research project. For example, while the research topic, experimental methodology, and statistical analyses are provided as part of the core activities, students are prompted to engage with each of these through short answer assessment items and instructor-facilitated discussion. These early experiences prepare students to independently participate in the eight practices while conducting their research projects.

Working with authentic biological data, students further develop their science and engineering practices in contexts other than with Avidians. While most Avida-ED activities approach biological phenomena with a scientific lens by asking questions and constructing explanations, a few of the non-Avida-ED activities explicitly use the lens more associated with engineering, that of identifying problems and designing solutions. Proposing, evaluating, and iteratively adjusting a scientific model is explicitly practiced in some activities so that students actively engage in evaluating their own mental model of the phenomenon. For example, students iteratively create increasingly complex models of the interplay between communicative and morphological phenotypes with sexual and natural selection in the context of crickets and their parasites. Students also practice data management and basic statistical analyses and interpretations, including measures of experimental variation such as confidence

interval construction and statistical significance determination, emphasizing how important variation is for understanding biological mechanisms and interpreting hypotheses and experimental results.

Most recommended introductory undergraduate biology education content from "Vision and Change in Undergraduate Biology Education" (2011) and The BioCore Guide (Brownell et al., 2014) is addressed in the course. Across the course, over ninety percent of the BioCore Guide content statements are addressed and assessed at least once, with many revisited in multiple systems and using varied assessment styles. This fact highlights the pervasiveness of integrative thinking throughout the course along with the opportunity to show the connection between individual trait variation and its evolution. For example, the unit on selection and convergence in mouse coat color variation, adapted from White et al. (2013), revisits ten content statements covered earlier in the course and introduces ten new content statements from across four core concepts and all major subdisciplines of biology. Twenty percent of the content statements are specifically explored in the Avida-ED lab core activities. For example, one of the statements investigated in the third Avida-ED lab exercise is: "Mutations and epigenetic modifications can impact the regulation of gene expression and/or the structure and function of the gene product. If mutations affect phenotype and lead to increased reproductive success, the frequency of those alleles will increase in the population." This statement corresponds to the core concept of Evolution with connections to the biology subfields of Molecular, Cellular, and Developmental Biology.

Additional content statements may be revealed by student groups during their research process. Consider the core concept Evolution, sub-discipline Physiology statement, "Physiological systems are constrained by ancestral structures, physical limits, and the requirements of other physiological systems, leading to trade-offs that affect fitness." Some students discover that the evolution of Avidian functions can involve trade-offs wherein the gain of a selectively beneficial complex function requires, due to the Avidian's particular genetic or physiological machinery (Lenski et al. 1999), the loss of a simpler function. Another

statement commonly addressed is the Information flow core concept, Ecology/ Evolutionary Biology sub-discipline statement, "A genotype influences the range of possible phenotypes in an individual; the actual phenotype results from interactions between alleles and the environment." Some students come across this genotype-by-environment interaction in which an Avidian genotype inconsistently performs a function. This is because Avidian function performance requires encountering and utilizing random numbers in the digital environment, and some number combinations require fewer computational steps to output the product necessary to fulfill a function. Those students that encounter this phenomenon do not have the requisite Avidian taxonomic expertise to understand what is occurring. Thus, this can be an extenuating factor causing experiments to produce unexpected results. Once this genotype-by-environment interaction is elucidated by the instructor, this experience can provide a fantastic learning opportunity of “failing to succeed,” similar to that described by Linn et al. (2015) and Goodwin et al. (2021).

The course curriculum also addresses ten recommendations by Hillis (2007) for including evolution in introductory biology education. It demonstrates *that evolutionary research is ongoing* through units such as those on the *E. coli* Long Term Evolution Experiment (Lenski et al., 2015), in part adapted from the curriculum of White et al. (2013). The evolutionary processes of mutation, natural selection, genetic drift, and migration are explored and contrasted in multiple systems, *clarifying that evolution is not a synonym for natural selection* but rather change in individual variation over time. *Fresh examples* are used and continually updated, for example additional research on the evolution of quiet Hawaiian crickets (Balenger and Zuk, 2015) has recently been published (Schneider et al., 2018) and will be included in future iterations of the course. Examples such as the co-evolutionary significance of the human skin and gut microbiome are used to *show how evolution is relevant to human lives*. Numerous *examples of evolutionary biology from popular media* are incorporated, including many YouTube videos and NPR interviews of researchers conducting and explaining their work. The course includes *experimental evolution* both with a unit on biological experimentation in the *E.*

coli Long Term Evolution Experiment (Blount et al., 2008; Lenski and Travisano, 1994) and the digital evolution lab in which students conduct their own experimental research. Instead of being taught only in the context of organismal biology and ecology, Evolution is *integrated throughout* the course and is the unifying thread connecting the material, as has been called for by Smith et al. (2009), among others. *Tree-thinking* is explored in the first unit of the course to provide a conceptual basis and means of understanding later material. The *diversity of life* is highlighted using units featuring taxa from across biological diversity, including plants, mammals, bacteria, insects, mollusks, and reptiles. And finally, the *great magnitude of evolutionary time* is emphasized through discussions on the origin of sex and photosynthesis.

The Avida-ED laboratory experience shares features with other experiences that have been shown to have a multitude of benefits. Corwin et al. (2015) suggests that similar combinations of activities, especially in the context of course-based undergraduate research experiences, improves cognitive, psychosocial, and behavioral outcomes. For example: Avida-ED and other activities throughout the course require students to work in small, cooperative groups, a context with well-documented benefits (Smith, 1996; Smith et al., 2005). The research project, in particular, provides teamwork experience in addition to the greater student interaction facilitated by inquiry-based learning (Aditomo et al., 2013; Felder and Brent, 1996). Research experiences prompt students to more greatly identify themselves as being scientists or part of the science community (Carlone and Johnson, 2007), in addition to promoting interest in their degree and retention in STEM generally (Bangera and Brownell, 2014), especially for unrepresented persons in these fields (Eagan Jr. et al., 2013; Espinosa, 2011).

Furthermore, conducting research using Avidians necessitates an explicit interdisciplinary connection between biology and computer science; by solving conceptual problems in complex interdisciplinary systems students may improve learning and cognitive skill development (Betz, 1995). Computational tools in education have been shown to have well-documented benefits with regards to student learning and engagement. For example, such tools facilitate the connection of observed phenomena with underlying causal processes

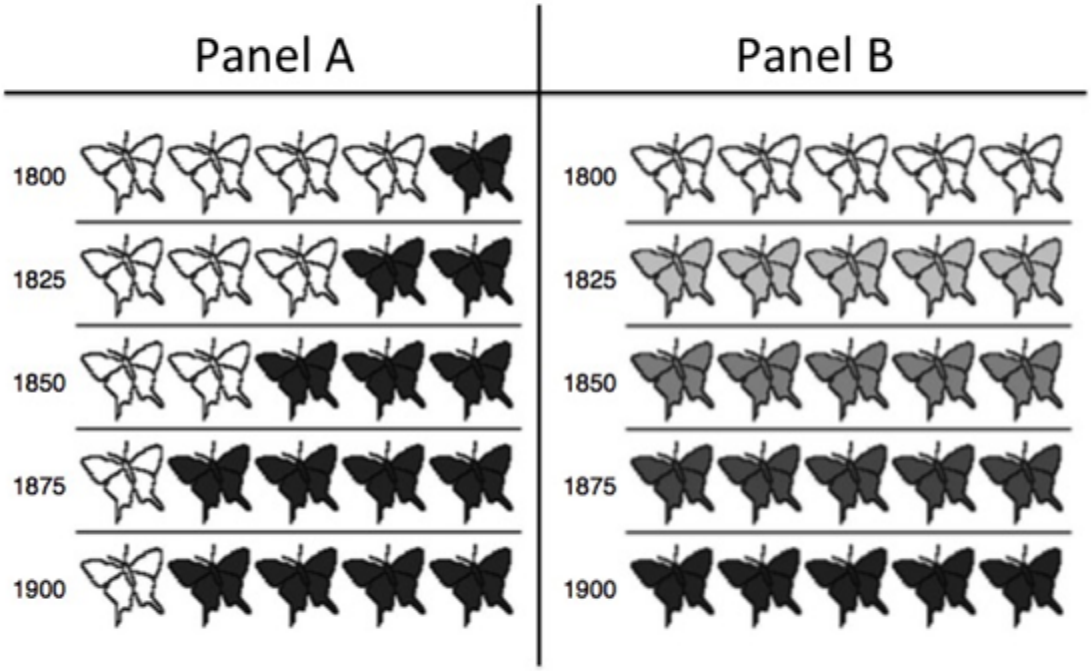
(Magana, 2017), and allow students to observe the unobservable (Trey and Khan, 2008), for example electrons moving in electrical circuits or, as with Avida-ED, variationalism as individual-level variation producing population-level evolutionary change within minutes. In some cases, students prefer computational tools over conventional tools and knowledge sources, reporting greater interest in the material (Akkoyun, 2017). Additionally, some students interact with Avida-ED in a manner akin to gamification, which can strongly encourage student engagement (Drace, 2013). Interest, in turn, influences affective response and persistence, which influences learning (Ainley and Ainley, 2011; Rotgans and Schmidt, 2009, 2014). Furthermore, by conducting authentic scientific investigation as part of a course, students tend to experience greater gains in content knowledge (Minner et al., 2010). While most of these benefits have not yet been shown with the use of Avida-ED specifically, the features of the digital evolution lab experience suggest they may indeed occur.

Assessment

To assess transformational and variational reasoning, I used a three-item assessment featuring a graphical representation of color variation in the peppered moth *Biston betularia* during the industrial revolution in England (Fig. 4.02). The assessment stem established that moths darkened over time, that the coloration is heritable, and that the proposed samples are representative of the historical populations. Two potential collections of moths sampled across 100 years in 25-year intervals are proposed, Panel A and Panel B. Respondents are prompted to provide a separate explanation for each sample's phenotypic pattern and then choose and defend which collection would be more likely to have existed. Panel A presents sequentially darker populations that exhibit individual-level variation in white and dark moth variants, with dark moths increasing in frequency over time (Fig. 4.02). Note that in this panel, the population neither begins with all white moths (i.e., prior to a mutation conferring the dark coloration) nor ends with all black moths (i.e., fixation of the dark variant), but rather maintains contemporaneous phenotypic variation at each sampling interval. Panel B presents sequentially

darker populations that each exhibit a lack of variation among individuals, with each successive population being slighter darker than the previous sampling interval.

Previous observations have demonstrated that from 1800 to 1900, populations of moths darkened in England. It has also been established that moth coloration is inherited across generations. Imagine that scientists have just discovered a set of museum specimens that are representative of populations existing between 1800 and 1900. Panels A and B represent two possibilities of what these collections will look like.



1. If the collections match panel A, how would you explain the data?
2. If the collections match panel B, how would you explain the data?
3. Which of these patterns do you think is more likely to be observed in the collections, and why?

Figure 4.02. The assessment regarding melanic moth populations with idealized variational (Panel A) and transformational (Panel B) evolutionary trends, as modified from Shtulman (2006).

The assessment tool used in this study is modeled after a portion of an assessment by Shtulman (2006) that also addresses transformational and variational understanding of

biological variation. The entire assessment of Shtulman (2006) consisted of thirty items that aimed to distinguish variational and transformational interpretations for each of six evolutionary phenomena: variation, inheritance, adaptation, domestication, speciation, and extinction. The five-item subset on variation first presented the peppered moth scenario within an explicit adaptive evolution context and inquired "how might a change in the moths' environment brought about a change in the moths' color?" The remaining four items involved shading a set of 25 moth outlines, arrayed with five moths every 25 years as per Panel A or B in our assessment, but with no color variation other than that which the participant selected. Participants shaded each moth per row to reflect the color variation they would expect to have been observed historically, choosing from one of five grayscale values for each moth.

Our goals in modifying the assessment of Shtulman (2006) were to exclude the requirement of natural selection, to reduce the measurement error associated with the original scoring system, for which 31% of respondents produced uninterpretable shading patterns, and to facilitate digital survey collection by using standard item formatting. The moth figure used in our assessment (Fig. 4.02) reflects two competing scenarios – the ideal variational (panel A) and transformational (panel B) shading patterns sought by Shtulman (2006) in his Figure 2. While Shtulman (2006) offered some limited evidence on content validity for his assessment (i.e., three biology doctorates reviewed the items), it is unclear how this evidence relates to the entire instrument versus the specific portion used here with modifications.

Responses

Using an online survey platform, students completed the assessment twice, during the first week and then during the last two weeks of instruction for in-person course semesters and during the first and last week of online-only course semesters. Completion was incentivized with a small amount of extra course credit and was weighted such that completion of both pre- and post-course surveys awarded substantially greater extra credit. The transformationalism-variationalism items analyzed here (Fig. 4.02) were a subset of the survey, which additionally contained multiple choice and other open-ended items addressing other facets of evolution

education. A total of 649 surveys were returned, of which 611 were fully completed with respect to the transformationalism-variationalism portion, for an overall response rate of 84%. Of these, 103 students completed only the pre- or post-course survey, while 254 students completed both the pre- and post-course surveys for a joint response rate of 70%.

Scoring

Each survey submission was scored as transformational (T), variational (V), or other (O), with the latter including all responses that were not otherwise classifiable (Fig. 4.03). An initial scoring rubric was created based on the theoretical distinction between transformational and variational understanding (Coley and Muratore, 2012; Coley and Tanner, 2015; Gregory, 2009; Shtulman, 2006; Shtulman and Calabi, 2008; Shtulman and Schulz, 2008; Sinatra et al., 2008; Speth et al., 2014). Multiple scorers examined a set of responses for non-biology majors in a different course and revised the rubric following discussion. Two scorers then independently evaluated all *Integrative Biology* responses reported here, with discussion and rubric revision occurring throughout this process, before reaching finalized consensus scoring across all responses. Inter-rater reliability was calculated using Cohen's kappa between the two scorers prior to consensus being reached through discussion and was calculated for both all completed responses and the subset of paired pre-/post-course responses.

Scoring process determined from item 3 response:

If pattern A more likely → Evaluate items 1 and 3

If pattern B more likely → Evaluate items 2 and 3

O score – Any other response, including choosing both or neither

Scoring for items 1 and 3, explanation for panel A:

V – A variant within the population is changing or should change over time.

T – Explicitly refers to the essence of the species changing.

O – Any other response, including: The moth (population or species) is evolving, changing, adapting, developing, dominating, etc; natural selection or other evolutionary process without mention of phenotypic variation; uncertainty if moth variants can mate or if they are in the same population or species; variation among selective environments without mention of phenotypic variation.

Scoring for items 2 and 3, explanation for panel B:

V – A variant within the population is changing or should change over time and explicitly states that variation exists at a single time.

T – Explicitly refers to the essence of the species changing; or the moth (population, species, genotype, phenotype) is evolving, changing, adapting, developing, dominating, etc.

O – Any other response, including: Evolution is occurring rapidly without mention of individual variation; natural selection or other evolutionary process without mention of phenotypic variation; uncertainty if moth variants can mate or if they are in the same population or species; variation among selective environments without mention of individual phenotypic variation.

Figure 4.03. Melanic moth assessment rubric including the process for choosing which items to score and the criteria for awarding scores of variationalism (V), transformationalism (T), and other (O). Underlining indicates a scoring distinction between panels A and B, and colored font further indicates the same response as being scored differently depending on the chosen panel.

A single score was given for each participant per survey disbursement. The default score was O, with the response requiring explicit endorsement of an appropriate V or T response. The rationale for the rubric (Fig. 4.03) is that although both conceptualizations of variation may recognize that variation exists among individuals in a population, those with a variational understanding will deem it with consequential importance as contributing to evolutionary change over time. Note that only rejecting transformationalism by, for example, stating that the

population would not change in unison, was insufficient to score a V, as were references to genetics concepts alone, e.g., mutation, allele, dominance, or recessivity. The selection of items reviewed per participant depended on the item 3 response. If a student indicated that either panel A or panel B was most likely to be observed, then their score was based on their answer to item 3 and the item corresponding to their selection (item 1 or 2, respectively). The rationale for this differential scoring process is that a student may have difficulty justifying the panel that does not conform to their understanding of biological variation and evolutionary processes. The O ("other") score represents an inability to differentiate a response as transformational or variational (i.e., measurement error) and is not necessarily intended as a determination of a mixed model of understanding (Evans, 2008) or pre-variationalism (Shtulman, 2006).

An identically worded explanation for panel B with it being chosen as most likely may result in a different score than if it was the explanation for panel A by a student who chose panel A. Such instances are highlighted in Figure 4.03 with underlining and examples are provided in Table 4.02. Because the panels presented a distinction between whether variation among individuals exists within a population at a given time, we deemed that a student's choice of panel A provided some evidence of variational thinking and likewise for a choice of panel B and transformational thinking. Therefore, the rubric indicates a lower scoring barrier for T scores when choosing panel B and for V scores when choosing panel A. For example, a vague response such as "the moth population is evolving" was awarded an O if it was an explanation for panel A as being most likely, but a T for panel B (i.e., the blue text in Fig. 4.03). In describing panel B, a V score required additional explication that variation existed within the population at a given time, as opposed to simply being different between generations as represented in that panel. The additional O response in explanations of panel B, evolution occurring rapidly, was added because this may be a scientifically accurate explanation for a variational process, although this was insufficient by itself to warrant a V.

Table 4.02. Example scored responses of transformational (T), variational (V), and other (O) for panel chosen as most likely in item 3, with item 1 or 2 reviewed depending on panel chosen.

Panel, Score	Item 3	Item 1 or 2	Reason
A, V	Panel A is due to the fact you can see moths evolving in order to adapt to the area they are in.	With the predator sight being drawn to a darker colored moth, as the environment started to change, the ability to see the darker colored moths where difficult compared to seeing the lighter moths.	Variant should change over time
A, V	I would say A is more likely to be observed because populations don't all just change at once. Through adaption, one ancestor passes the trait on to offspring	One moth was randomly dark and passed this trait on to the offspring. This continued down the line	Variant changing over time
A, T	*No such responses provided by Integrative Biology students surveyed*		
A, O	A, because it would be gradual, not each generation being darker than the previous.	The moths gradually darkened over time.	Moth species is evolving
A, O	Panel A, as the evolutionary timeline in panel B is a bit fast. 100 years for a complete colour change.	The independent population of dark butterflies was more fit than the light ones.	Separate populations
B, V	B is more likely given that it is a more gradual process of color change.	There was a more gradual environmental change which selected moths with darker color to survive and reproduce over those without that phenotype.	Variation existing at a single time changes in the population
B, V	I think that panel B is more likely because the difference is too great in panel A between the moths for the two colors to coexist for that long.	Over time, a series of random mutations that made the moths slightly darker occurred. These mutations were beneficial enough that they outcompeted the other types and the entire population became gradually darker over time.	Variation existing at a single time changes in the population
B, T	Panel B because all members of the species darkened and had a chance at an increased survival rate.	The whole moth species gradually began to darken as opposed to a rapid darkening in a few members of the species.	Moth species is evolving
B, T	Panel B is definitely more likely to occur as all of the moths have many of the same survival and reproduction genes making it way more likely for them to darken as a population and gradually rather than individuals and suddenly.	Over the course of the century, the moth population as a whole was slowly developing a slightly darker shade of grey in order for survival. As the years progressed the shades became darker.	Moth species is evolving

Table 4.02 (continued)

B, O	Panel B. It seems unlikely that with 100 years of selective pressure that there would still be a light moth in 1900.	There was selective pressure on the light moths to darken.	Evolutionary process without mention of individual variation; also, rapid evolution
B, O	panel B because when an invasive species arrives, if its trait was more preferable then the other one it would take over extremely quick were panel A has a slow natural selection occurring.	over time the darker the moth became the higher chance of surviving the moth had allowing for more offspring, which slowly changed the color of the population.	Separate species; also, rapid evolution
Neither, O	Natural selection is more likely to occur because this is how most organisms adapt to their environment.	*Not reviewed*	Neither panel chosen
Both, O	I think they are both equally likely to occur. Both seem to be caused by the influence of a trait that wasn't there before.	*Not reviewed*	Both panels chosen

Statistics

As an initial control to evaluate whether respondents provided longer responses pre- and post-course, repeated-measures t-tests were conducted for the number of words and characters, and for both the complete response provided to all three items as well as just the scored portion of the response. A three-factor by two-factor Chi-square test was used to evaluate an overall difference in transformational (T), variational (V), and indeterminable ("other," O) responses among pre- and post-course scores among all completed responses and, separately, among the paired responses. Following this, each was partitioned into orthogonal two-by-two tables (Sharpe, 2015). This analysis independently evaluated the pair of factors of interest here, i.e., scores of transformational versus variational, and the pair of factors of lesser interest, i.e., scores classifiable as transformational or variational versus other. Since these sets of data were orthogonal, adjustment for multiple comparisons was not needed.

A McNemar-Bowker test was used to evaluate symmetry among pre-to-post-course score shifts among paired responses only; for example whether the number of students that

shifted from transformationalism to variationalism was significantly different from the number that shifted from variationalism to transformationalism (i.e., T-V vs. V-T). Individual McNemar tests were then used to evaluate each of the three pairs of comparisons (T-V vs. V-T, T-O vs. O-T, and V-O vs. O-V) and Bonferroni adjustments were conducted for the resulting p-values. All Chi-square and related analyses additionally included Cramér's V, which is a measure of association or effect size and has the same interpretation as a Pearson correlation coefficient. All statistical tests were completed using Microsoft Excel.

Results and Discussion

For students that responded to both the pre- and post-assessment, there was no significant difference in the length of responses provided for either assessment (Table 4.03) regardless of whether all three items or only the scored items were compared. To put these values into perspective, the first set of example responses in Table 4.02 for panel choice A & score O, choice B & score V, and choice A & score V are representative of the lower quartile, mean, and upper quartile word and character lengths. It was often extremely difficult to classify student understanding when sparsely detailed responses were provided. Since instances of insufficient evidence inflate the number of O scores, it is important that a lack of significant difference in response length pre- and post-course was found.

Table 4.03. Mean word and character counts for paired student responses with repeated-measures t-tests (N = 254, df = 253) showing no significant differences. Relevant word and character counts include only the subset of items used for scoring, see Figure 4.03.

	Mean Word Count	Mean Character Count	Mean Relevant Word Count	Mean Relevant Character Count
Pre-course	56.5	324.1	39.2	223.5
Post-course	54.9	319.8	37.7	217.3
p-value	0.4	0.69	0.33	0.49

The rubric appeared to function as expected. Inter-rater reliability was sufficiently high (Hallgren, 2012) between the two scorers. Percent agreement was 87.2% across all responses (N = 611), with a Cohen’s kappa of 0.783. For the paired responses only (N = 508), percent

agreement was 86.8% with a kappa of 0.774. The rubric was designed to differentiate among means of understanding depending on the panel chosen as most likely to have been historically observable (item 3), with a lower barrier for scoring variationalism when panel A was chosen and likewise for transformationalism and panel B. This result was observed. Of the 400 students that chose panel A, 76% of the responses were scored as variationalism and 0% as transformationalism; and of the 188 that chose panel B, 61% of the responses scored as transformational and 14% as variationalism. The remaining proportion of students whose understanding was indeterminable when choosing panel A or B, 24% and 26% respectively, was near identical, suggesting that scoring for this catch-all category was not biased depending on panel chosen.

There was a significant difference in the proportion of student understanding of variation pre- and post-course across all completed responses (Fig. 4.04). The difference in student understanding was statistically significant in the omnibus analysis of all responses ($\chi^2_{2,611} = 15.106$, $p < 0.001$) with an effect size of 0.157. This result was driven by the pre-/post-course differences in transformational and variational understanding, as evidenced by its partitioned analysis ($\chi^2_{1,445} = 13.401$, $p < 0.001$) with an effect size of 0.174. This contrasts with the orthogonal partitioning of students with transformational or variational thinking compared to students whose understanding was indeterminable, which was non-significant ($\chi^2_{1,611} = 1.822$, $p > 0.1$).

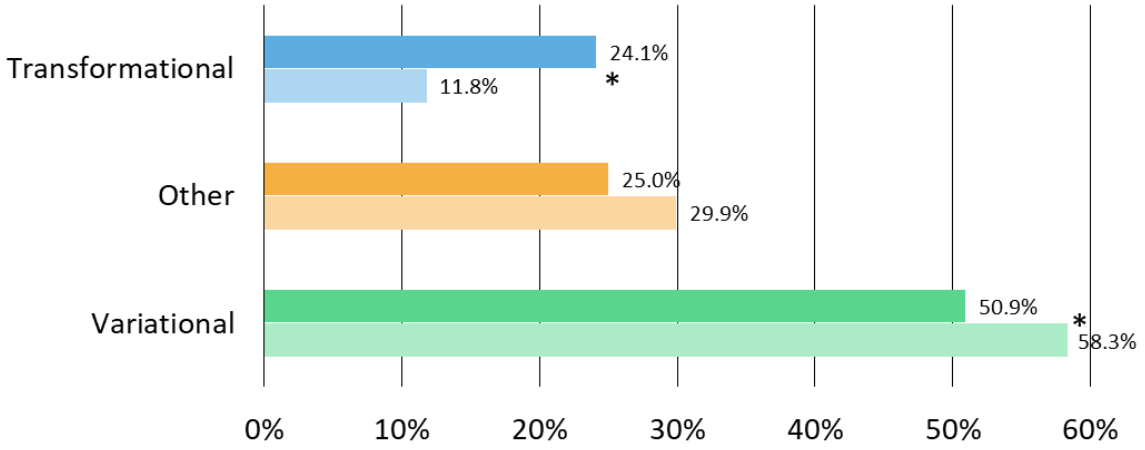


Figure 4.04. Percent of all students demonstrating understanding of biological variation pre-course (dark, N = 340) and post-course (light, N = 271), with asterisks indicating significance.

There was also a significant difference in student understanding for the subset of students that completed both pre- and post-course surveys (Fig. 4.05, overall states of understanding and black asterisks; Table 4.04, sums). The difference in student understanding was statistically significant in the omnibus analysis comparing changes in the overall states of understanding ($\chi^2_{2,508} = 12.711$, $p < 0.001$) with an effect size of 0.158. The post-hoc tests demonstrated that both orthogonal partitioned datasets were significant, comparing transformational and variational understanding ($\chi^2_{1,374} = 13.298$, $p < 0.001$) with an effect size of 0.189 and comparing transformational or variational thinking to students whose understanding was indeterminable ($\chi^2_{1,508} = 9.889$, $p < 0.005$). Therefore, the proportion of students with transformational understanding significantly decreased and the proportion with variational or indeterminable understanding significantly increased.

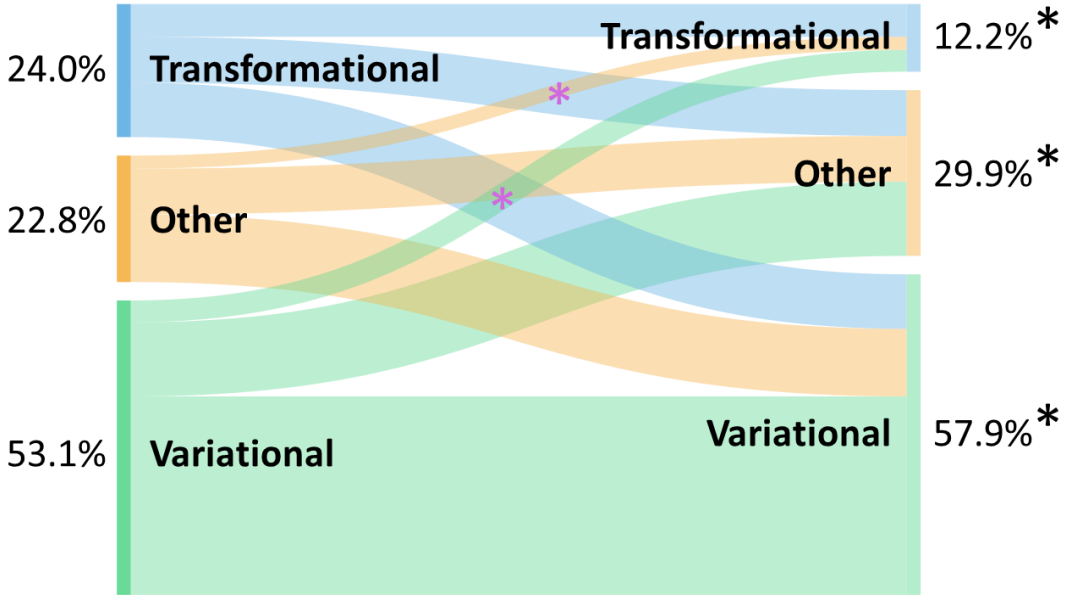


Figure 4.05. Percent of paired-response students (N = 254) demonstrating understanding of biological variation pre- and post-course, with asterisks indicating significance between pre/post states of understanding (black) and directionality of shifts between states (pink).

Table 4.04. Detailed percentage breakdowns for paired-response students (N = 254) demonstrating understanding of biological variation pre- and post-course.

		Pre-course			
		Transformational	Other	Variational	Sum
Post-	Transformational	5.9%	2.4%	3.9%	12.2%
	Other	8.3%	8.3%	13.4%	29.9%
	Variational	9.8%	12.2%	35.8%	57.9%
	Sum	24.0%	22.8%	53.1%	100%

There was asymmetry in the directionality of shifts between states of understanding for the paired-response students (Fig. 4.05, colors crossing and pink asterisks; Table 4.04, off-diagonal percentages). The directionality of shifts in student understanding was statistically significant in the omnibus analysis ($\chi^2_{3,254} = 14.900$, $p < 0.005$) with an effect size of 0.242. This result was driven by two significant factors: greater transitions from transformational to variational understanding rather than in the reverse ($\chi^2_{1,35} = 6.429$, adj. $p < 0.05$) and greater transitions from transformational to indeterminable understanding than in the reverse ($\chi^2_{1,27} = 8.333$, adj. $p < 0.005$). In comparison, there was a lack of significance in transitions between variational and indeterminable understanding ($\chi^2_{1,65} = 0.138$, adj. $p > 0.1$). These results

provide the means to explain the significant overall shifts in student understanding among paired-response students (Fig. 4.05, black asterisks) and potentially, albeit untestable, among all students (Fig. 4.04).

These results are encouraging, especially in comparison to prior work. The pre-course proportion of students identified here as transformationalists is similar to that identified for nonbiology majors and biology majors by Richard et al. (2017), and for the biology majors by Coley and Tanner (2015), although less so than their nonmajors. Unlike Richard et al. (2017), who saw no change in transformational understanding between entering and advanced biology majors (i.e., after multiple semesters of instruction), the proportion of students here was halved at the conclusion of their one, and likely only, college biology course (Fig. 4.04). At least half of this change is attributable to students shifting from transformational to variational understanding, with the remainder shifting to the catch-all "other" classification (Fig. 4.05). Shtulman and Calabi (2008), using the assessment of Shtulman (2006), reported a slightly lower prevalence of transformational thinking among biology students than reported here. Unfortunately, their results are not comparable to those reported here because their assessment measured variational and transformational understanding on a single continuous interval scale and with respect to inheritance, adaptation, domestication, speciation, and extinction in addition to variation alone. For example, Shtulman (2006) classified students as pre-variationalists if they held variational views of adaptation and inheritance but transformational views of variation and the other factors assessed. Additionally, it would be ideal if the effect size results observed here could be compared to those in prior work (Sun et al., 2010); unfortunately, they were not reported.

The proportion of students that could not be classified as having either variational or transformational understanding (i.e., "other") increased pre-/post-course for paired-response students (Fig. 4.05), although not significantly so for all students (Fig. 4.04). Perhaps the increased proportion of paired-students coded as "other" may be attributed to their incorporation of new information into mixed models of understanding as students shift from

transformational to variationalism (Evans, 2008) or pre-variationalism (Shtulman, 2006). This might especially explain the 8.3% of paired-response students that shifted from transformational to "other," which was statistically significant from the 2.4% that switched from "other" to transformational (Table 4.04). Research will be necessary to test this hypothesis, especially using methods with reduced measurement error. While slightly more students shifted from variational to "other" than in the reverse direction (13.4% versus 12.2%, Table 4.04), this difference was not significant; combined, these groups constitute more than a quarter of all paired-response students and are the clearest indicators of measurement error for this assessment. Overall, the proportion of "other" is slightly lower than the 31% of ambiguous responses in the original assessment (Shtulman, 2006), indicating that our assessment had limited success in reducing measurement error.

It is evident from reviewing student responses that there are a variety of means by which a student may preserve a transformationalist perspective. For example, students may misunderstand the mutational process as one that acts on all individuals in the population simultaneously. In this way, the species' genomic identity can be construed as its essence and the mutation as its means of change (Coley and Tanner, 2012; Fodor, 1998; Gelman, 2004). Seemingly based on a reductionist, molecular means, this transformational model might appear to have a modern scientific basis; although of course it entails a gross misunderstanding about how mutation occurs. It is unclear if these observations support mixed model creation (Evans, 2008; Shtulman, 2006), the strengthening of transformationalism following biological education (Coley and Tanner, 2015), or both. If such student understanding is a conflation of the concepts of mutation occurring to an individual and substitution occurring within a population then it may be an instance of confusion over levels of the system, i.e., a failure in population thinking (Cooper, 2017; Mayr, 1994). Or it could instead be a case of students improperly applying terminology for the mechanism of change (e.g., mutation instead of substitution) or the product of such change (e.g., mutation instead of allele) and therefore being incorrectly classified non-variationalists.

Other students seem to propose ecological or physiological processes that were not intended. For example, many students were classified as “other” due to their indication that multiple populations or species of moth existed in panel A, and others proposed phenotypic plasticity to explain panel B. These biologically plausible responses may be due to either creative thinking or may instead be a cover for an innate essentialist reasoning strategy by suggesting that evolution is not occurring at all.

Conclusions

This work demonstrated a significant change in student understanding of variation after a single semester of instruction. The cognitive stickiness of the essentialist intuitive reasoning strategy as expressed through a transformationalist understanding of the importance of individual variation (Gelman, 2004; Speth et al., 2014) was not shown here, as 9.8% of all students switched from transformational to variationalism, with the potential improvement of up to 18% whose “other” responses may be indicative of mixed-model reasoning (Table 4.04). Stated in a different way, relative to the proportion of students that began as transformationalists, 40% of them switched to variational understanding and a total of up to 75% potentially improved their understanding. This result might be attributable to the direct exposure to and experimentation regarding variation and its importance using integrative thinking, statistical reasoning, and computer modeling via Avida-ED, as hypothesized here and suggested generally by Sinatra et al. (2008), Kalinowski et al. (2010), Speth et al. (2014), and Cooper (2017). However, the Avida-ED curriculum was one component among many in a course that emphasized within-species variation and population thinking in considering evolution across multiple levels of biological organization, as also hypothesized here and suggested generally by others (Kalinowski et al., 2010; Nehm et al., 2009; Shtulman and Calabi, 2008; Shtulman and Schulz, 2008; Smith et al., 2009; Speth et al., 2014). Further, throughout the course, observations and results made using Avida-ED were directly compared with those gathered from biological sources such that an understanding of the importance of variation in

one context may be tied to an understanding in others. Due to the thorough integration of Avida-ED in the course and the fact that the assessment was administered not immediately before and after specific Avida-ED curricula, it is impossible to attribute the results shown here to anything other than the course experience as a whole. Additionally, while the results presented here are substantial and intriguing, further assessment development for measuring student understanding of variation is welcomed.

The Avida-ED digital evolution lab experience was a hallmark of the *Integrative Biology* course. This course should serve as an exemplar for how to best incorporate integrative thinking using case-based content exploration (see White et al., 2013) and a digital evolution lab using Avida-ED into an undergraduate introductory biology curriculum to engage students in a variational understanding (Kohn et al., 2018). The goal in designing *Integrative Biology* was to create a rigorous introductory biology course for non-life sciences STEM majors that integrates a case-based approach, introducing students to the levels of biological organization while using evolution as a central organizing framework. As such, it was designed according to the specific circumstances of its institution and student body; yet the curricular niche it fills is likely applicable to other universities catering to non-life sciences STEM majors, especially engineers, mathematicians, and physicists, and its approach to variational understanding is applicable more broadly still.

While Avida-ED has been used in other courses across a range of geographically and educationally diverse contexts (Lark et al., 2018), the effect size of the curricular interventions on student understanding of variation reported here may not be equivalent. Lark (2014) has established that Avida-ED implementation success is positively correlated with instructor familiarity and comfort, presentation and exploration during in-person classes and especially laboratory-type learning opportunities and student completion of introductory activities followed by guided inquiry investigation. *Integrative Biology* was instructed by Avida-ED curriculum developers whom in reduced-class-size computer lab environments implemented the full set of recommended curricula, including the student driven research project.

Additionally, Avidians were explicitly analogized with asexual study systems and experiments of each were discussed side-by-side, numerous biological concepts initially introduced with biological systems were contemporaneously explored in Avida-ED and vice versa, and students were given the opportunity to learn through research success and, as importantly, failure with instructor guidance. Further, this is, to my knowledge, the first application of a digital evolution lab within a larger course context in which the importance of variation among individuals is repeatedly explored in multiple biological systems and across biological fields using integrative thinking. Instructors interested in similarly shifting student understanding of variation may, therefore, find success by incorporating a digital evolution lab using Avida-ED into their classroom and integrating student discoveries therein throughout the curriculum.

Care was taken in this study to measure student understanding of a difficult concept, and further improvements to the measurement tool would be useful contributions to the field of biology education research. Especially considering the relatively high proportion of measurement error observed here, which was marginally improved over Shtulman (2006), further efforts to refine our assessment instrument are welcomed. Biology education assessments are numerous, difficult to compare, and routinely in a state of development (Mead et al., 2019). Since the assessment used here was closely based on that offered by Shtulman (2006), the validity of that instrument at least somewhat lends evidence for the validity of this instrument. Even so, one or more validity analyses could be conducted to confirm that it is measuring the intended distinction between variational and transformational understanding of biological variation. Further refinement of the rubric may also be warranted, especially if it reduces the classification of indeterminable ("other") understanding while maintaining its validity. For example, the explicit rejection of transformationalism or variationalism alone should, I think, necessitate evidence of the other mode, although others may disagree, e.g., Furtak et al. (2014).

The original assessment offered evidence of content validity via agreement among experts (Shtulman, 2006), although their expert pool of three biology doctorates was small. The

modifications made here, while expected to maintain this validity, could be similarly tested and with a larger pool of experts. Additionally, evidence of substantive validity would be most convincing (Campbell and Nehm, 2013). Substantive validity is often shown through think-aloud interviews demonstrating that the cognitive processes used to answer the assessment are as intended. This could confirm, for example, that a student response of "the moth population is evolving" is most likely intended as a transformational response in the context of panel B and perhaps clarify what the same statement may mean in the panel A context (i.e., the blue text in Fig. 4.03). Student interviews may also provide further insight other than with respect to the rubric. For example, in our understanding of what students mean when they state or imply that the moth variants in panel A are separate species or populations, or in understanding their nuanced transformational or perhaps mixed-model explanations, especially because expert variationalist biologists may struggle to understand the now-alien transformational form of understanding, despite essentialist thinking in biological contexts being universal during childhood (Evans, 2008; Sinatra et al., 2008). As an alternative to major revisions to the instrument itself, future uses could routinely include structured interviews, especially with students otherwise scoring as "other," with the aim of reducing measurement error by classifying a greater proportion of students as holding either transformational or variational understanding.

Follow-up work should explore how changes in variational thinking relates to other biology topics. Most notably, Shtulman and Schulz (2008) argue that natural selection can only be understood once variational thinking has been established. However, the work of Kalinowski et al. (2016) for the CANS instrument showed that student understanding of variation was largely independent of their understanding of evolution generally. While it might not be possible to tease apart a causal relationship, analyses comparing increased variationalism with the understanding of natural selection, genetic drift, mutation, and other factors would be interesting. It might also be worthwhile to investigate the correlation between variational thinking and acceptance of evolution, although admittedly transformationalism is still an

evolutionary process, albeit closely related and perhaps indistinguishable on the present assessment from the non-evolutionary essentialism.

REFERENCES

REFERENCES

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinforma. Oxf. Engl. 21, 2104–2105. https://doi.org/10.1093/bioinformatics/bti263

Adami, C., 2006. Digital genetics: unravelling the genetic basis of evolution. Nat. Rev. Genet. 7, 109–118. https://doi.org/10.1038/nrg1771

Adami, C., Ofria, C., Collier, T.C., 2000. Evolution of biological complexity. Proc. Natl. Acad. Sci. 97, 4463–4468.

Adams, R.H., Schield, D.R., Card, D.C., Castoe, T.A., 2018. Assessing the Impacts of Positive Selection on Coalescent-Based Species Tree Estimation and Species Delimitation. Syst. Biol. 67, 1076–1090. https://doi.org/10.1093/sysbio/syy034

Aditomo, A., Goodyear, P., Bliuc, A.-M., Ellis, R.A., 2013. Inquiry-based learning in higher education: principal forms, educational objectives, and disciplinary variations. Stud. High. Educ. 38, 1239–1258.

Agren, J.A., Williamson, R.J., Campitelli, B.E., Wheeler, J., 2017. Greenbeards in yeast: an undergraduate laboratory exercise to teach the genetics of cooperation. J. Biol. Educ. 51, 228–236.

Ainley, M., Ainley, J., 2011. Student engagement with science in early adolescence: The contribution of enjoyment to students' continuing interest in learning about science. Contemp. Educ. Psychol. 36, 4–12.

Akkoyun, O., 2017. New simulation tool for teaching–learning processes in engineering education. Comput. Appl. Eng. Educ. 25, 404–410.

Altekar, G., Dwarkadas, S., Huelsenbeck, J.P., Ronquist, F., 2004. Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20, 407–415.

Alters, B.J., Nelson, C.E., 2002. Perspective: Teaching evolution in higher education. Evolution 56, 1891–1901.

American Association for the Advancement of Science, 2011. Vision and change in undergraduate biology education: a call to action. American Association for the Advancement of Science, Washington, DC.

Arenas, M., 2012. Simulation of Molecular Data under Diverse Evolutionary Scenarios. PLoS Comput. Biol. 8, e1002495. https://doi.org/10.1371/journal.pcbi.1002495

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., Grosdidier, A., Hernandez, C., Ioannidis, V., Kuznetsov, D., Liechti, R., Moretti, S., Mostaguir, K., Redaschi, N., Rossier, G., Xenarios, I., Stockinger, H., 2012. ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res. 40, W597–W603. https://doi.org/10.1093/nar/gks400

Auchincloss, L.C., Laursen, S.L., Branchaw, J.L., Eagan, K., Graham, M., Hanauer, D.I., Lawrie, G., McLinn, C.M., Pelaez, N., Rowland, S., Towns, M., Trautmann, N.M., Varma-Nelson, P., Weston, T.J., Dolan, E.L., 2014. Assessment of Course-Based Undergraduate Research Experiences: A Meeting Report. CBE-Life Sci. Educ. 13, 29–40. https://doi.org/10.1187/cbe.14-01-0004

Balenger, S.L., Zuk, M., 2015. Roaming Romeos: male crickets evolving in silence show increased locomotor behaviours. Anim. Behav. 101, 213–219. https://doi.org/10.1016/j.anbehav.2014.12.023

Ballen, C.J., Blum, J.E., Brownell, S., Hebert, S., Hewlett, J., Klein, J.R., McDonald, E.A., Monti, D.L., Nold, S.C., Slemmons, K.E., Soneral, P.A.G., Cotner, S., 2017. A Call to Develop Course-Based Undergraduate Research Experiences (CUREs) for Nonmajors Courses. CBE Life Sci. Educ. 16. https://doi.org/10.1187/cbe.16-12-0352

Bangera, G., Brownell, S.E., 2014. Course-Based Undergraduate Research Experiences Can Make Scientific Research More Inclusive. CBE-Life Sci. Educ. 13, 602–606. https://doi.org/10.1187/cbe.14-06-0099

Barbançon, F., Evans, S.N., Nakhleh, L., Ringe, D., Warnow, T., 2013. An experimental study comparing linguistic phylogenetic reconstruction methods. Diachronica 30, 143–170. https://doi.org/10.1075/dia.30.2.01bar

Baum, B.R., Duncan, T., Stuessy, T., 1984. Application of compatibility and parsimony methods at the infraspecific, specific, and generic levels in Poaceae, in: Cladistics: Perspectives on the Reconstruction of Evolutionary History. Columbia Univ. Press, New York, pp. 192–220.

Baum, D.A., Smith, S.D., 2013. Tree Thinking An Introduction to Phylogenetic Biology. Roberts and Company Publishers Inc, the United States.

Beckmann, B.E., McKinley, P.K., Ofria, C., 2008. Selection for group-level efficiency leads to self-regulation of population size, in: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation. ACM, pp. 185–192.

Beerenwinkel, N., Siebourg, J., 2012. Probability, statistics, and computational science, in: Evolutionary Genomics. Springer, pp. 77–110.

Betz, J.A., 1995. Computer games: Increase learning in an interactive multidisciplinary environment. J. Educ. Technol. Syst. 24, 195–205.

Bishop, B.A., Anderson, C.W., 1990. Student conceptions of natural selection and its role in evolution. J. Res. Sci. Teach. 27, 415–427.

Bishop, M.J., Friday, A.E., 1987. Tetrapod relationships: the molecular evidence, in: Molecules and Morphology in Evolution: Conflict or Compromise. Cambridge University Press, Cambridge, England, pp. 123–139.

Blount, Z.D., Borland, C.Z., Lenski, R.E., 2008. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. Proc. Natl. Acad. Sci. 105, 7899–7906. https://doi.org/10.1073/pnas.0803151105

Bostrom, N., 2003. Are You Living in a Computer Simulation? Philos. Q. 53, 243–255.

Box, G.E., 1976. Statistics and Science. J Am Stat Assoc 71, 791–799.

Brooks, D.J., Fresco, J.R., Lesk, A.M., Singh, M., 2002. Evolution of Amino Acid Frequencies in Proteins Over Deep Time: Inferred Order of Introduction of Amino Acids into the Genetic Code. Mol. Biol. Evol. 19, 1645–1655. https://doi.org/10.1093/oxfordjournals.molbev.a003988

Brown, J.M., Hedtke, S.M., Lemmon, A.R., Lemmon, E.M., 2010. When Trees Grow Too Long: Investigating the Causes of Highly Inaccurate Bayesian Branch-Length Estimates. Syst. Biol. 59, 145–161. https://doi.org/10.1093/sysbio/syp081

Brownell, S.E., Freeman, S., Wenderoth, M.P., Crowe, A.J., 2014. BioCore Guide: a tool for interpreting the core concepts of Vision and Change for biology majors. CBE-Life Sci. Educ. 13, 200–211.

Bryant, D., 2003. A classification of consensus methods for phylogenetics. DIMACS Ser. Discrete Math. Theor. Comput. Sci. 61, 163–184.

Buck, L.B., Bretz, S.L., Towns, M.H., 2008. Characterizing the level of inquiry in the undergraduate laboratory. J. Coll. Sci. Teach. 38, 52–58.

Bull, J.J., Badgett, M.R., Wichman, H.A., Huelsenbeck, J.P., Hillis, D.M., Gulati, A., Ho, C., Molineux, I.J., 1997. Exceptional convergent evolution in a virus. Genetics 147, 1497–1507.

Bull, J.J., Cunningham, C.W., Molineux, I.J., Badgett, M.R., Hillis, D.M., 1993. Experimental Molecular Evolution of Bacteriophage T7. Evolution 47, 993–1007. https://doi.org/10.2307/2409971

Burke, M.K., Dunham, J.P., Shahrestani, P., Thornton, K.R., Rose, M.R., Long, A.D., 2010. Genome-wide analysis of a long-term evolution experiment with Drosophila. Nature 467, 587–590. https://doi.org/10.1038/nature09352

Burmeister, A.R., Smith, J.J., 2016. Evolution across the Curriculum: Microbiology. J. Microbiol. Biol. Educ. 17, 252–260. https://doi.org/10.1128/jmbe.v17i2.988

Campbell, C.E., Nehm, R.H., 2013. A Critical Analysis of Assessment Quality in Genomics and Bioinformatics Education Research. CBE Life Sci. Educ. 12, 530–541. https://doi.org/10.1187/cbe.12-06-0073

Carlone, H.B., Johnson, A., 2007. Understanding the science experiences of successful women of color: Science identity as an analytic lens. J. Res. Sci. Teach. 44, 1187–1218. https://doi.org/10.1002/tea.20237

Cavalli-Sforza, L.L., Edwards, A.W., 1967. Phylogenetic analysis: models and estimation procedures. Evolution 21, 550–570.

CG, N., LaBar, T., Hintze, A., Adami, C., 2017. Origin of life in a digital microcosm. Phil Trans R Soc A 375, 20160350. https://doi.org/10.1098/rsta.2016.0350

Chi, M.T.H., 2005. Commonsense Conceptions of Emergent Processes: Why Some Misconceptions Are Robust. J. Learn. Sci. 14, 161–199. https://doi.org/10.1207/s15327809jls1402_1

Chow, S.S., Wilke, C.O., Ofria, C., Lenski, R.E., Adami, C., 2004. Adaptive Radiation from Resource Competition in Digital Organisms. Science 305, 84–86. https://doi.org/10.1126/science.1096307

Chown, M., 2013. What a Wonderful World: One Man's Attempt to Explain the Big Stuff. Faber & Faber.

Clune, J., Goldsby, H.J., Ofria, C., Pennock, R.T., 2011. Selective pressures for accurate altruism targeting: evidence from digital evolution for difficult-to-test aspects of inclusive fitness theory. Proc. R. Soc. Lond. B Biol. Sci. 278, 666–674.

Clune, J., Misevic, D., Ofria, C., Lenski, R.E., Elena, S.F., Sanjuán, R., 2008. Natural Selection Fails to Optimize Mutation Rates for Long-Term Adaptation on Rugged Fitness Landscapes. PLoS Comput. Biol. 4, e1000187. https://doi.org/10.1371/journal.pcbi.1000187

Clune, J., Ofria, C., Pennock, R.T., 2007. Investigating the emergence of phenotypic plasticity in evolving digital organisms, in: European Conference on Artificial Life. Springer, pp. 74–83.

Coley, J., Muratore, T., 2012. Trees, Fish, and Other Fictions: Folk Biological Thought and its Implications for Understanding Evolutionary Biology, in: Evolution Challenges: Integrating Research and Practice In Teaching and Learning About Evolution. pp. 22–46. https://doi.org/10.1093/acprof:oso/9780199730421.003.0002

Coley, J.D., Tanner, K., 2015. Relations between Intuitive Biological Thinking and Biological Misconceptions in Biology Majors and Nonmajors. CBE—Life Sci. Educ. 14, ar8. https://doi.org/10.1187/cbe.14-06-0094

Coley, J.D., Tanner, K.D., 2012. Common Origins of Diverse Misconceptions: Cognitive Principles and the Development of Biology Thinking. CBE—Life Sci. Educ. 11, 209–215. https://doi.org/10.1187/cbe.12-06-0074

College Board, 2011. AP Biology Curriculum Framework 2012–2013. College Board, New York.

Connelly, B.D., Zaman, L., Ofria, C., McKinley, P.K., 2010. Social Structure and the Maintenance of Biodiversity., in: ALIFE. pp. 461–468.

Cooper, M.M., 2013. Chemistry and the Next Generation Science Standards. J. Chem. Educ. 90, 679–680. https://doi.org/10.1021/ed400284c

Cooper, R.A., 2017. Natural selection as an emergent process: instructional implications. J. Biol. Educ. 51, 247–260. https://doi.org/10.1080/00219266.2016.1217905

Cooper, T.F., Ofria, C., 2003. Evolution of stable ecosystems in populations of digital organisms, in: Artificial Life VIII: Proceedings of the Eighth International Conference on Artificial Life. pp. 227–232.

Corwin, L.A., Graham, M.J., Dolan, E.L., 2015. Modeling Course-Based Undergraduate Research Experiences: An Agenda for Future Research and Evaluation. CBE-Life Sci. Educ. 14, es1. https://doi.org/10.1187/cbe.14-10-0167

Cotner, S., Hebert, S., 2016. Bean Beetles Make Biology Research Sexy. Am. Biol. Teach. 78, 233–240. https://doi.org/10.1525/abt.2016.78.3.233

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. Proc. Natl. Acad. Sci. 110, E3171–E3178.

Covert III, A.W., Smith, L., Derrberry, D.Z., Wilke, C.O., 2012. What does sex have to do with it: tracking the fate of deleterious mutations in sexual populations. Artif. Life 13, 32–36. https://doi.org/10.7551/978-0-262-31050-5-ch005

Coyne, J.A., 2010. Why evolution is true. Viking, New York.

Cummings, M.P., Handley, S.A., Myers, D.S., Reed, D.L., Rokas, A., Winka, K., 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst. Biol. 52, 477–487.

Cunningham, C.W., Jeng, K., Husti, J., Badgett, M., Molineux, I.J., Hillis, D.M., Bull, J.J., 1997. Parallel molecular evolution of deletions and nonsense mutations in bacteriophage T7. Mol. Biol. Evol. 14, 113–116.

Cunningham, C.W., Zhu, H., Hillis, D.M., 1998. Best-fit maximum-likelihood models for phylogenetic inference: Empirical tests with known phylogenies. Evolution 52, 978–987.

Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C., 1978. 22 a model of evolutionary change in proteins, in: Atlas of Protein Sequence and Structure. National Biomedical Research Foundation Silver Spring, MD, pp. 345–352.

de Varigny, H., 1892. Experimental Evolution. MacMillan.

de Visser, J.A.G.M., Hermisson, J., Wagner, G.P., Meyers, L.A., Bagheri-Chaichian, H., Blanchard, J.L., Chao, L., Cheverud, J.M., Elena, S.F., Fontana, W., 2003. Perspective: evolution and detection of genetic robustness. Evolution 57, 1959–1972.

Degnan, J.H., DeGiorgio, M., Bryant, D., Rosenberg, N.A., 2009. Properties of consensus methods for inferring species trees from gene trees. Syst. Biol. 58, 35–54.

Dennet, D.C., 1995. Darwin's Dangerous Idea: Evolution and the meanings of life. Simon & Schuster.

Dobzhansky, T., 1973. Nothing in Biology Makes Sense except in the Light of Evolution. Am. Biol. Teach. 35, 125–129. https://doi.org/10.2307/4444260

Domin, D.S., 1999. A Review of Laboratory Instruction Styles. J. Chem. Educ. 76, 543. https://doi.org/10.1021/ed076p543

Drace, K., 2013. Gamification of the Laboratory Experience to Encourage Student Engagement †. J. Microbiol. Biol. Educ. 14, 273–274. https://doi.org/10.1128/jmbe.v14i2.632

Eagan Jr., M.K., Hurtado, S., Chang, M.J., Garcia, G.A., Herrera, F.A., Garibay, J.C., 2013. Making a Difference in Science Education: The Impact of Undergraduate Research Programs. Am. Educ. Res. J. 50, 683–713. https://doi.org/10.3102/0002831213482038

Elena, S.F., Wilke, C.O., Ofria, C., Lenski, R.E., 2007. Effects of population size and mutation rate on the evolution of mutational robustness. Evolution 61, 666–674. https://doi.org/10.1111/j.1558-5646.2007.00064.x

Elsberry, W.R., Grabowski, L.M., Ofria, C., Pennock, R.T., 2009. Cockroaches, drunkards, and climbers: Modeling the evolution of simple movement strategies using digital organisms, in: Artificial Life, 2009. ALife'09. IEEE Symposium On. IEEE, pp. 92–99.

Espinosa, L., 2011. Pipelines and Pathways: Women of Color in Undergraduate STEM Majors and the College Experiences That Contribute to Persistence. Harv. Educ. Rev. 81, 209–241. https://doi.org/10.17763/haer.81.2.92315ww157656k3u

Evans, E.M., 2008. Conceptual Change and Evolutionary Biology: A Developmental Analysis, in: Vosniadou, S. (Ed.), International Handbook of Research on Conceptual Change. Routledge, New York, pp. 263–294.

Eyre-Walker, A., Keightley, P.D., 2007. The distribution of fitness effects of new mutations. Nat. Rev. Genet. 8, 610–618. https://doi.org/10.1038/nrg2146

Fares, M.A., Barrio, E., Becerra, N., Escarmís, C., Domingo, E., Moya, A., 1998. The foot-and-mouth disease RNA virus as a model in experimental phylogenetics. Int. Microbiol. 1, 311–318.

Felder, R.M., Brent, R., 1996. Navigating the bumpy road to student-centered instruction. Coll. Teach. 44, 43–47.

Felsenstein, J., 2005. PHYLIP (Phylogeny Inference Package). Department of Genome Sciences, University of Washington, Seattle.

Felsenstein, J., 2004. Inferring phylogenies. Sinauer associates, Sunderland, MA.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.

Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27, 401–410.

Felsenstein, J., 1974. The Evolutionary Advantage of Recombination. Genetics 78, 737–756.

Fiedler, D., Sbeglia, G.C., Nehm, R.H., Harms, U., 2019. How strongly does statistical reasoning influence knowledge and acceptance of evolution? J. Res. Sci. Teach. 56, 1183–1206.

Fitch, W.M., Atchley, W.R., 1985. Evolution in inbred strains of mice appears rapid. Science 228, 1169–1175. https://doi.org/10.1126/science.4001935

Fitch, W.M., Margoliash, E., 1967. Construction of phylogenetic trees. Science 155, 279–284.

Fodor, J.A., 1998. Concepts: Where cognitive science went wrong. Oxford University Press.

Fortuna, M.A., Zaman, L., Ofria, C., Wagner, A., 2017. The genotype-phenotype map of an evolving digital organism. PLOS Comput. Biol. 13, e1005414. https://doi.org/10.1371/journal.pcbi.1005414

Fortuna, M.A., Zaman, L., Wagner, A.P., Ofria, C., 2013. Evolving digital ecological networks. PLoS Comput. Biol. 9, e1002928.

Freeman, S., Herron, J.C., 2014. Evolutionary analysis, Fifth. ed. Pearson Prentice Hall, Upper Saddle River, NJ.

Furtak, E.M., Morrison, D., Kroog, H., 2014. Investigating the Link Between Learning Progressions and Classroom Assessment. Sci. Educ. 98, 640–673. https://doi.org/10.1002/sce.21122

Garland, T., Rose, M.R., 2009. Darwin's Other Mistake, in: Garland, T., Rose, M.R. (Eds.), Experimental Evolution: Concepts, Methods, and Applications of Selection Experiments. University of California Press, Berkeley, pp. 3–13.

Geary, D., 2007. Educating the Evolved Mind: Conceptual Foundations for an Evolutionary Educational Psychology, in: Carlson, J., Levin, J.R. (Eds.), Educating the Evolved Mind: Conceptual Foundations for an Evolutionary Educational Psychology. IAP, pp. 1–99.

Gelman, S., 2004. Psychological essentialism in children. Trends Cogn. Sci. 8, 404–409. https://doi.org/10.1016/j.tics.2004.07.001

Gelman, S.A., 2003. The Essential Child: Origins of Essentialism in Everyday Thought. Oxford University Press, New York.

Gerrish, P.J., Lenski, R.E., 1998. The fate of competing beneficial mutations in an asexual population. Genetica 102, 127.

Godfrey-Smith, P., 2007. Conditions for evolution by natural selection. J. Philos. 104, 489–516.

Goings, S., Clune, J., Ofria, C., Pennock, R.T., 2004. Kin selection: The rise and fall of kin-cheaters, in: Proceedings of the Ninth International Conference on Artificial Life. pp. 303–308.

Goldsby, H.J., Cheng, B.H., 2008. Avida-MDE: a digital evolution approach to generating models of adaptive software behavior, in: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation. ACM, pp. 1751–1758.

Goldsby, H.J., Cheng, B.H., McKinley, P.K., Knoester, D.B., Ofria, C.A., 2008. Digital evolution of behavioral models for autonomic systems, in: Autonomic Computing, 2008. ICAC'08. International Conference On. IEEE, pp. 87–96.

Goldsby, H.J., Dornhaus, A., Kerr, B., Ofria, C., 2012. Task-switching costs promote the evolution of division of labor and shifts in individuality. Proc. Natl. Acad. Sci. 109, 13686–13691.

Goldsby, H.J., Knoester, D.B., Kerr, B., Ofria, C., 2014a. The effect of conflicting pressures on the evolution of division of labor. PloS One 9, e102713.

Goldsby, H.J., Knoester, D.B., Ofria, C., Kerr, B., 2014b. The Evolutionary Origin of Somatic Cells under the Dirty Work Hypothesis. PLoS Biol. 12, e1001858. https://doi.org/10.1371/journal.pbio.1001858

Goodwin, E.C., Anokhin, V., Gray, M.J., Zajic, D.E., Podrabsky, J.E., Shortlidge, E.E., 2021. Is This Science? Students' Experiences of Failure Make a Research-Based Course Feel Authentic. CBE—Life Sci. Educ. 20, ar10. https://doi.org/10.1187/cbe.20-07-0149

Gould, S.J., Duve, C. de, 1996. Full house: The spread of excellence from Plato to Darwin. Nature 383, 771–771.

Grabowski, L.M., Bryson, D.M., Dyer, F.C., Pennock, R.T., Ofria, C., 2013. A Case Study of the De Novo Evolution of a Complex Odometric Behavior in Digital Organisms. PLoS ONE 8, e60466. https://doi.org/10.1371/journal.pone.0060466

Grant, B.W., 2009. Practitioner research improved my students' understanding of evolution by natural selection in an introductory biology course. Teach. Issues Exp. Ecol. 6.

Graybeal, A., 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. 47, 9–17.

Graybeal, A., 1994. Evaluating the Phylogenetic Utility of Genes: A Search for Genes Informative About Deep Divergences Among Vertebrates. Syst. Biol. 43, 174–193. https://doi.org/10.2307/2413460

Green, D.S., Bozzone, D.M., 2001. A test of hypotheses about random mutation: using classic experiments to teach experimental design. Am. Biol. Teach. 63, 54–58.

Green, J.H., Koza, A., Moshynets, O., Pajor, R., Ritchie, M.R., Spiers, A.J., 2011. Evolution in a test tube: rise of the Wrinkly Spreaders. J. Biol. Educ. 45, 54–59. https://doi.org/10.1080/00219266.2011.537842

Gregory, T.R., 2009. Understanding Natural Selection: Essential Concepts and Common Misconceptions. Evol. Educ. Outreach 2, 156–175. https://doi.org/10.1007/s12052-009-0128-1

Gupta, A., LaBar, T., Miyagi, M., Adami, C., 2016. Evolution of Genome Size in Asexual Digital Organisms. Sci. Rep. 6, 25786. https://doi.org/10.1038/srep25786

Hagstrom, G.I., Hang, D.H., Ofria, C., Torng, E., 2004. Using Avida to test the effects of natural selection on phylogenetic reconstruction methods. Artif. Life 10, 157–166.

Hall, B.G., 2005. Comparison of the Accuracies of Several Phylogenetic Methods Using Protein and DNA Sequences. Mol. Biol. Evol. 22, 792–802. https://doi.org/10.1093/molbev/msi066

Hallgren, K.A., 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutor. Quant. Methods Psychol. 8, 23–34.

Handelsman, J., Houser, B., Kriegel, H., 1997. Biology brought to life: a guidebook to teaching students to think like scientists. McGraw-Hill Primis.

Hang, D., Ofria, C., Schmidt, T.M., Torng, E., 2003. The effect of natural selection on phylogeny reconstruction algorithms, in: Genetic and Evolutionary Computation Conference. Springer, pp. 13–24.

Hang, D., Torng, E., Ofria, C., Schmidt, T.M., 2007. The effect of natural selection on the performance of maximum parsimony. BMC Evol. Biol. 7, 94. https://doi.org/10.1186/1471-2148-7-94

Hartl, D.L., Clark, A.G., 2007. Principles of population genetics, Fourth. ed. Sinauer associates, Sunderland, MA.

Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22, 160–174.

Heddy, B.C., Sinatra, G.M., 2013. Transforming misconceptions: Using transformative experience to promote positive affect and conceptual change in students learning about biological evolution. Sci. Educ. 97, 723–744.

Hill, W.G., Robertson, A., 1966. The effect of linkage on limits to artificial selection. Genet. Res. 8, 269–294.

Hillis, D.M., 2007. Making evolution relevant and exciting to biology students. Evolution 61, 1261–1264.

Hillis, D.M., 1995. Approaches for assessing phylogenetic accuracy. Syst. Biol. 44, 3–16.

Hillis, D.M., Bull, J.J., 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42, 182–192.

Hillis, D.M., Bull, J.J., White, M.E., Badgett, M.R., Molineux, I.J., 1993. Experimental Approaches to Phylogenetic Analysis. Syst. Biol. 42, 90–92.

Hillis, D.M., Bull, J.J., White, M.E., Badgett, M.R., Molineux, I.J., 1992. Experimental Phylogenetics: Generation of a Known Phylogeny. Science 255, 589–592.

Hillis, D.M., Huelsenbeck, J.P., Cunningham, C.W., 1994. Application and accuracy of molecular phylogenies. Science 264, 671–677.

Hoang, D.T., Vinh, L.S., Flouri, T., Stamatakis, A., von Haeseler, A., Minh, B.Q., 2018. MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. BMC Evol. Biol. 18. https://doi.org/10.1186/s12862-018-1131-3

Hoang, D.T., Vinh, L.S., Flouri, T., Stamatakis, A., von Haeseler, A., Minh, B.Q., 2017. MPBoot version 1.1.0 User Manual.

Hormoz, S., 2013. Amino acid composition of proteins reduces deleterious impact of mutations. Sci. Rep. 3, 2919. https://doi.org/10.1038/srep02919

Howe, K., Bateman, A., Durbin, R., 2002. QuickTree: building huge Neighbour-Joining trees of protein sequences. Bioinformatics 18, 1546–1547.

Huang, H., He, Q., Kubatko, L.S., Knowles, L.L., 2010. Sources of Error Inherent in Species-Tree Estimation: Impact of Mutational and Coalescent Effects on Accuracy and Implications for Choosing among Different Methods. Syst. Biol. 59, 573–583. https://doi.org/10.1093/sysbio/syq047

Huang, H., Sukumaran, J., Smith, S.A., Knowles, Ll., 2017. Cause of gene tree discord? Distinguishing incomplete lineage sorting and lateral gene transfer in phylogenetics. PeerJ Prepr. https://doi.org/10.7287/peerj.preprints.3489v1

Huelsenbeck, J.P., 1995. Performance of phylogenetic methods in simulation. Syst. Biol. 44, 17–48.

Huelsenbeck, J.P., Nielsen, R., 1999. Effect of nonindependent substitution on phylogenetic accuracy. Syst. Biol. 48, 317–328.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R., Bollback, J.P., 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. science 294, 2310–2314.

Huerta-Cepas, J., Serra, F., Bork, P., 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol. Biol. Evol. 33, 1635–1638. https://doi.org/10.1093/molbev/msw046

Hunter, J.D., 2007. Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. 9, 90–95. https://doi.org/10.1109/MCSE.2007.55

Huxley, T.H., 1887. On the Reception of the "Origin of Species," in: Darwin, F. (Ed.), The Life and Letters of Charles Darwin, Including an Autobiographical Chapter. Kpjm <irrau, pp. 179–204.

Jin, L., Nei, M., 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. 7, 82–102.

Johnson, W., Lark, A., Pennock, R., Mead, L., 2011a. Evolving TCE Biodegraders - Activity [WWW Document]. www.teachengineering.org. URL https://www.teachengineering.org/activities/view/mis_avida_lesson01_activity2 (accessed 3.1.18).

Johnson, W., Pennock, R., Mead, L., 2011b. Studying Evolution with Digital Organisms - Activity [WWW Document]. www.teachengineering.org. URL https://www.teachengineering.org/activities/view/mis_avida_lesson01_activity1 (accessed 3.1.18).

Johnson, W.R., Lark, A., 2018. Evolution in Action in the Classroom: Engaging Students in Science Practices to Investigate and Explain Evolution by Natural Selection. Am. Biol. Teach. 80, 92–99.

Jones, D.T., Taylor, W.R., Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. Bioinformatics 8, 275–282.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules, in: Munro, H.N. (Ed.), Mammalian Protein Metabolism, III. Academic Press, New York, pp. 21–132.

Junier, T., 2011. Newick Utilities Tutorial Version 1.6.0.

Junier, T., Zdobnov, E.M., 2010. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics 26, 1669–1670. https://doi.org/10.1093/bioinformatics/btq243

Kalinowski, S.T., Leonard, M.J., Andrews, T.M., 2010. Nothing in evolution makes sense except in the light of DNA. CBE—Life Sci. Educ. 9, 87–97.

Kalinowski, S.T., Leonard, M.J., Taper, M.L., 2016. Development and Validation of the Conceptual Assessment of Natural Selection (CANS). CBE—Life Sci. Educ. 15, ar64. https://doi.org/10.1187/cbe.15-06-0134

Kawecki, T.J., Lenski, R.E., Ebert, D., Hollis, B., Olivieri, I., Whitlock, M.C., 2012. Experimental evolution. Trends Ecol. Evol. 27, 547–560. https://doi.org/10.1016/j.tree.2012.06.001

Kim, S., de Medeiros, B.A.S., Byun, B.-K., Lee, S., Kang, J.-H., Lee, B., Farrell, B.D., 2018. West meets East: How do rainforest beetles become circum-Pacific? Evolutionary origin of Callipogon relictus and allied species (Cerambycidae: Prioninae) in the New and Old Worlds. Mol. Phylogenet. Evol. 125, 163–176. https://doi.org/10.1016/j.ympev.2018.02.019

Kimura, M., 1983. The neutral theory of molecular evolution. Cambridge University Press.

Kimura, M., 1968. Evolutionary Rate at the Molecular Level. Nature 217, 624–626. https://doi.org/10.1038/217624a0

Kimura, M., Ohta, T., 1969. The average number of generations until fixation of a mutant gene in a finite population. Genetics 61, 763.

King, J.L., Jukes, T.H., 1969. Non-Darwinian Evolution. Science 164, 788–798.

Klopfstein, S., Massingham, T., Goldman, N., 2017. More on the Best Evolutionary Rate for Phylogenetic Analysis. Syst. Biol. 66, 769–785. https://doi.org/10.1093/sysbio/syx051

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows, in: Loizides, F., Schmidt, B. (Eds.), Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press, pp. 87–90.

Knobe, J., Samuels, R., 2013. Thinking like a scientist: Innateness as a case study. Cognition 126, 72–86.

Knoester, D.B., Goldsby, H.J., McKinley, P.K., 2013. Genetic Variation and the Evolution of Consensus in Digital Organisms. IEEE Trans. Evol. Comput. 17, 403–417. https://doi.org/10.1109/TEVC.2012.2201725

Knoester, D.B., McKinley, P.K., Beckmann, B., Ofria, C., 2007a. Directed evolution of communication and cooperation in digital organisms, in: European Conference on Artificial Life. Springer, Berlin, Heidelberg, pp. 384–394.

Knoester, D.B., McKinley, P.K., Ofria, C.A., 2007b. Using group selection to evolve leadership in populations of self-replicating digital organisms, in: Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation. ACM, pp. 293–300.

Kohn, C., Wiser, M.J., Pennock, R.T., Smith, J.J., Mead, L.S., 2018. A Digital Technology-based Introductory Biology Course Designed for Engineering and Other Non-life Science STEM Majors. Comput. Appl. Eng. Educ. https://doi.org/10.1002/cae.21986

Kolaczkowski, B., Thornton, J.W., 2008. A mixed branch length model of heterotachy improves phylogenetic accuracy. Mol. Biol. Evol. 25, 1054–1066.

Krist, A.C., Showsh, S.A., 2007. Experimental evolution of antibiotic resistance in bacteria. Am. Biol. Teach. 69, 94–97.

Kuhner, M.K., Felsenstein, J., 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11, 459–468.

Kvitek, D.J., Sherlock, G., 2011. Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. PLoS Genet. 7, e1002056.

LaBar, T., Adami, C., 2017. Evolution of drift robustness in small populations. Nat. Commun. 8, 1012. https://doi.org/10.1038/s41467-017-01003-7

LaBar, T., Adami, C., 2016. Different Evolutionary Paths to Complexity for Small and Large Populations of Digital Organisms. PLOS Comput. Biol. 12, e1005066. https://doi.org/10.1371/journal.pcbi.1005066

Lanier, H.C., Knowles, L.L., 2012. Is Recombination a Problem for Species-Tree Analyses? Syst. Biol. 61, 691–701. https://doi.org/10.1093/sysbio/syr128

Lark, A., Richmond, G., Mead, L.S., Smith, J.J., Pennock, R.T., 2018. Exploring the Relationship between Experiences with Digital Evolution and Students' Scientific Understanding and Acceptance of Evolution. Am. Biol. Teach. 80, 74–86.

Lark, A., Richmond, G., Pennock, R.T., 2014. Modeling evolution in the classroom: The case of Fukushima's mutant butterflies. Am. Biol. Teach. 76, 450–454.

Lark, A.M., 2014. Teaching and learning with digital evolution: Factors influencing implementation and student outcomes (Ph.D.). Michigan State University, United States -- Michigan.

Le, S.Q., Gascuel, O., 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25, 1307–1320. https://doi.org/10.1093/molbev/msn067

Lehmer, L.M., Ragsdale, B.D., Daniel, J., Hayashi, E., Kvalstad, R., 2011. Plastic bag clip discovered in partial colectomy accompanying proposal for phylogenic plastic bag clip classification. BMJ Case Rep. 2011, bcr0220113869.

Leitner, T., Escanilla, D., Franzen, C., Uhlen, M., Albert, J., 1996. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc. Natl. Acad. Sci. 93, 10864–10869.

Lenski, R.E., 2001. Twice as natural. Nature 414, 255.

Lenski, R.E., Barrick, J.E., Ofria, C., 2006. Balancing robustness and evolvability. PLoS Biol. 4, e428.

Lenski, R.E., Ofria, C., Collier, T.C., Adami, C., 1999a. Lenski, Richard E., et al. "Genome complexity, robustness and genetic interactions in digital organisms. Nature 400, 661–664.

Lenski, R.E., Ofria, C., Collier, T.C., Adami, C., 1999b. Genome complexity, robustness and genetic interactions in digital organisms. Nature 400, 661–664. https://doi.org/10.1038/23245

Lenski, R.E., Ofria, C., Pennock, R.T., Adami, C., 2003. The evolutionary origin of complex features. Nature 423, 139.

Lenski, R.E., Travisano, M., 1994. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. Proc. Natl. Acad. Sci. 91, 6808–6814. https://doi.org/10.1073/pnas.91.15.6808

Lenski, R.E., Wiser, M.J., Ribeck, N., Blount, Z.D., Nahum, J.R., Morris, J.J., Zaman, L., Turner, C.B., Wade, B.D., Maddamsetti, R., Burmeister, A.R., Baird, E.J., Bundy, J., Grant, N.A., Card, K.J., Rowles, M., Weatherspoon, K., Papoulis, S.E., Sullivan, R., Clark, C., Mulka, J.S., Hajela, N., 2015. Sustained fitness gains and variability in fitness trajectories in the long-term evolution experiment with Escherichia coli. Proc R Soc B 282. https://doi.org/10.1098/rspb.2015.2292

Li, H.-L., Wang, W., Mortimer, P.E., Li, R.-Q., Li, D.-Z., Hyde, K.D., Xu, J.-C., Soltis, D.E., Chen, Z.-D., 2015. Large-scale phylogenetic analyses reveal multiple gains of actinorhizal nitrogen-fixing symbioses in angiosperms associated with climate change. Sci. Rep. 5, 14023. https://doi.org/10.1038/srep14023

Li, W.-H., 1997. Molecular Evolution. Sinauer Associates Inc, Sunderland, Mass.

Linn, M.C., Palmer, E., Baranger, A., Gerard, E., Stone, E., 2015. Undergraduate research experiences: Impacts and opportunities. Science 347, 1261757. https://doi.org/10.1126/science.1261757

Liu, S.-Y.V., Frédérich, B., Lavoué, S., Chang, J., Erdmann, M.V., Mahardika, G.N., Barber, P.H., 2018. Buccal venom gland associates with increased of diversification rate in the fang blenny fish Meiacanthus (Blenniidae; Teleostei). Mol. Phylogenet. Evol. 125, 138–146. https://doi.org/10.1016/j.ympev.2018.03.027

Lizard Evolution Virtual Lab [WWW Document], 2014. . HHMI BioInteractive. URL https://www.biointeractive.org/classroom-resources/lizard-evolution-virtual-lab (accessed 4.8.21).

Magana, A.J., 2017. Modeling and Simulation in Engineering Education: A Learning Progression. J. Prof. Issues Eng. Educ. Pract. 143, 04017008.

Martin, D.P., Lemey, P., Posada, D., 2011. Analysing recombination in nucleotide sequences. Mol. Ecol. Resour. 11, 943–955. https://doi.org/10.1111/j.1755-0998.2011.03026.x

Mayr, E., 2001. What evolution is. Basic Book, New York.

Mayr, E., 1994. Typological Versus Population Thinking, in: Sober, E. (Ed.), Conceptual Issues in Evolutionary Biology. The Mit Press. Bradford Books, pp. 157–160.

Mayr, E., 1982. The growth of biological thought: Diversity, evolution, and inheritance. Harvard University Press.

McCaldon, P., Argos, P., 1988. Oligopeptide biases in protein sequences and their use in predicting protein coding regions in nucleotide sequences. Proteins Struct. Funct. Bioinforma. 4, 99–122.

McKinley, P., Cheng, B.H., Ofria, C., Knoester, D., Beckmann, B., Goldsby, H., 2008. Harnessing digital evolution. Computer 41.

Mead, L.S., Kohn, C., Warwick, A., Schwartz, K., 2019. Applying measurement standards to evolution education assessment instruments. Evol. Educ. Outreach 12, 5. https://doi.org/10.1186/s12052-019-0097-y

Mead, L.S., Mates, A., 2009. Why science standards are important to a strong science curriculum and how states measure up. Evol. Educ. Outreach 2, 359–371.

Messer, P.W., 2013. SLiM: Simulating Evolution with Selection and Linkage. Genetics 194, 1037–1039. https://doi.org/10.1534/genetics.113.152181

Minh, B.Q., Trifinopoulos, J., Schrempf, D., Schmidt, H.A., 2017. IQ-TREE version 1.6.0: Tutorials and Manual.

Minner, D.D., Levy, A.J., Century, J., 2010. Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. J. Res. Sci. Teach. 47, 474–496.

Misevic, D., Lenski, R.E., Ofria, C., 2004. Sexual reproduction and Muller's ratchet in digital organisms, in: Ninth International Conference on Artificial Life. pp. 340–345.

Misevic, D., Ofria, C., Lenski, R.E., 2010. Experiments with Digital Organisms on the Origin and Maintenance of Sex in Changing Environments. J. Hered. 101, S46–S54. https://doi.org/10.1093/jhered/esq017

Misevic, D., Ofria, C., Lenski, R.E., 2006. Sexual reproduction reshapes the genetic architecture of digital organisms. Proc. R. Soc. B Biol. Sci. 273, 457–464. https://doi.org/10.1098/rspb.2005.3338

Miyamoto, M.M., Cracraft, J., 1991. Phylogenetic inference, DNA sequence analysis, and the future of molecular systematics, in: Phylogenetic Analysis of DNA Sequences. Oxford Univ. Press New York, pp. 3–17.

Moore, R.M., Harrison, A.O., McAllister, S.M., Polson, S.W., Wommack, K.E., 2020. Iroki: automatic customization and visualization of phylogenetic trees. PeerJ 8, e8584. https://doi.org/10.7717/peerj.8584

National Research Council, 2012a. A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. National Academies Press.

National Research Council, 2012b. Thinking evolutionarily: Evolution education across the life sciences: National Academies Press, Washington, DC.

Nehm, R.H., Poole, T.M., Lyford, M.E., Hoskins, S.G., Carruth, L., Ewers, B.E., Colberg, P.J.S., 2009. Does the Segregation of Evolution in Biology Textbooks and Introductory Courses Reinforce Students' Faulty Mental Models of Biology and Evolution? Evol. Educ. Outreach 2, 527–532. https://doi.org/10.1007/s12052-008-0100-5

Nehm, R.H., Reilly, L., 2007. Biology majors' knowledge and misconceptions of natural selection. BioScience 57, 263–272.

Nehm, R.H., Ridgway, J., 2011. What do experts and novices "see" in evolutionary problems? Evol. Educ. Outreach 4, 666–679.

Nei, M., 1987. Molecular evolutionary genetics. Columbia university press, New York.

Nei, M., Kumar, S., 2000. Molecular evolution and phylogenetics. Oxford university press.

Nelson, C.E., 2012. Why don't undergraduates really "get" evolution? What can faculty do?, in: Karl, S.K.B., Rosengren, S., Evans, E.M., Sinatra, G.M. (Eds.), Evolution Challenges: Integrating Research and Practice in Teaching and Learning about Evolution. Oxford University Press, New York, pp. 311–347.

NGSS Lead States, 2013. Next generation science standards: For states, by states. National Academies Press, Washington, DC.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol. 32, 268–274. https://doi.org/10.1093/molbev/msu300

Oakley, T.H., 2009. A Critique of Experimental Phylogenetics, in: Garland, T., Rose, M.R. (Eds.), Experimental Evolution: Concepts, Methods, and Applications of Selection Experiments. University of California Press, Berkeley, pp. 659–669.

Oakley, T.H., Cunningham, C.W., 2000. Independent contrasts succeed where ancestor reconstruction fails in a known bacteriophage phylogeny. Evolution 54, 397–405.

Ofria, C., 2015. Why use Artificial Life to Study Evolutionary Biology? Devolab Digit. Evol. Lab. URL http://devosoft.org/using-artificial-life-to-test-evolutionary-hypotheses-part-1-why-artificial-life/ (accessed 3.2.18).

Ofria, C., Adami, C., Collier, T.C., 2003. Selective pressures on genomes in molecular evolution. J. Theor. Biol. 222, 477–483.

Ofria, C., Bryson, D.M., Wilke, C.O., 2009. Avida: A Software Platform for Research in Computational Evolutionary Biology, in: Komosinski, M. (Ed.), Artificial Life Models in Software. Springer, London, pp. 3–36.

Ofria, C., Huang, W., Torng, E., 2008. On the gradual evolution of complexity and the sudden emergence of complex features. Artif. Life 14, 255–263.

Ofria, C., Wilke, C.O., 2004. Avida: A software platform for research in computational evolutionary biology. Artif. Life 10, 191–229.

Ohta, T., Kimura, M., 1971. On the constancy of the evolutionary rate of cistrons. J. Mol. Evol. 1, 18–25.

O'Neill, B., 2003. Digital evolution. PLoS Biol. 1, e18.

Ostrowski, E.A., Ofria, C., Lenski, R.E., 2015. Genetically integrated traits and rugged adaptive landscapes in digital organisms. BMC Evol. Biol. 15. https://doi.org/10.1186/s12862-015-0361-x

Ostrowski, E.A., Ofria, C., Lenski, R.E., 2007. Ecological Specialization and Adaptive Decay in Digital Organisms. Am. Nat. 169, E1–E20. https://doi.org/10.1086/510211

Pang, T.Y., 2020. A coarse-graining, ultrametric approach to resolve the phylogeny of prokaryotic strains with frequent homologous recombination. Bmc Evol. Biol. 20, 52. https://doi.org/10.1186/s12862-020-01616-5

Pennock, R.T., 2007a. Learning evolution and the nature of science using evolutionary computing and artificial life. McGill J. Educ. 42, 211–224.

Pennock, R.T., 2007b. Models, simulations, instantiations, and evidence: the case of digital evolution. J. Exp. Theor. Artif. Intell. 19, 29–42.

Pentz, J.T., Limberg, T., Beermann, N., Ratcliff, W.C., 2015. Predator Escape: An Ecologically Realistic Scenario for the Evolutionary Origins of Multicellularity. Evol. Educ. Outreach 8. https://doi.org/10.1186/s12052-015-0041-8

Perez, F., Granger, B.E., 2007. IPython: A System for Interactive Scientific Computing. Comput. Sci. Eng. 9, 21–29. https://doi.org/10.1109/MCSE.2007.53

Petrie, A., Finkel, S.E., Erbe, J., 2005. Use of long-term E. coli cultures to study generation of genetic diversity & teach general microbiology laboratory skills. Am. Biol. Teach. 67, 87–92.

Pin, L.C., Teen, L.P., Ahmad, A., Usup, G., 2001. Genetic diversity of Ostreopsis ovata (Dinophyceae) from Malaysia. Mar. Biotechnol. 3, 246–255.

Plunkett, A.D., Yampolsky, L.Y., 2010. When a Fly Has to Fly to Reproduce: Selection against Conditional Recessive Lethals in Drosophila. Am. Biol. Teach. 72, 12–15.

Psonis, N., Antoniou, A., Karameta, E., Leaché, A.D., Kotsakiozi, P., Darriba, D., Kozlov, A., Stamatakis, A., Poursanidis, D., Kukushkin, O., Jablonski, D., Crnobrnja–Isailović, J., Gherghel, I., Lymberakis, P., Poulakakis, N., 2018. Resolving complex phylogeographic patterns in the Balkan Peninsula using closely related wall-lizard species as a model system. Mol. Phylogenet. Evol. 125, 100–115. https://doi.org/10.1016/j.ympev.2018.03.021

Rabosky, D.L., Lovette, I.J., 2008. Explosive Evolutionary Radiations: Decreasing Speciation or Increasing Extinction Through Time? Evolution 62, 1866–1875. https://doi.org/10.1111/j.1558-5646.2008.00409.x

Rafalski, A., Morgante, M., 2004. Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet. 20, 103–111. https://doi.org/10.1016/j.tig.2003.12.002

Rambaut, A., 2018. FigTree. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh.

Rambaut, A., Grass, N.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics 13, 235–238. https://doi.org/10.1093/bioinformatics/13.3.235

Randall, R.N., Radford, C.E., Roof, K.A., Natarajan, D.K., Gaucher, E.A., 2016. An experimental phylogeny to benchmark ancestral sequence reconstruction. Nat. Commun. 7, 12847. https://doi.org/10.1038/ncomms12847

Ratcliff, W.C., Raney, A., Westreich, S., Cotner, S., 2014. A novel laboratory activity for teaching about the evolution of multicellularity. Am. Biol. Teach. 76, 81–87.

Richard, M., Coley, J.D., Tanner, K.D., 2017. Investigating Undergraduate Students' Use of Intuitive Reasoning and Evolutionary Knowledge in Explanations of Antibiotic Resistance. CBE—Life Sci. Educ. 16, ar55. https://doi.org/10.1187/cbe.16-11-0317

Robbins, J.R., Roy, P., 2007. The natural selection: identifying & correcting non-science student preconceptions through an inquiry-based, critical approach to evolution. Am. Biol. Teach. 69, 460–466.

Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147. https://doi.org/10.1016/0025-5564(81)90043-2

Robson, R.L., Burns, S., 2011. Gain in student understanding of the role of random variation in evolution following teaching intervention based on Luria-Delbruck experiment. J. Microbiol. Biol. Educ. JMBE 12, 3.

Rokas, A., Carroll, S.B., 2006. Bushes in the Tree of Life. PLOS Biol. 4, e352. https://doi.org/10.1371/journal.pbio.0040352

Ronquist, F., Huelsenbeck, J.P., Teslenko, M., 2011. MrBayes version 3.2 Manual: Tutorials and Model Summaries.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539–542.

Rose, M.R., Passananti, H.B., Margarida, M. (Eds.), 2004. Methuselah flies: a case study in the evolution of aging. World Scientific.

Rotgans, J., Schmidt, H., 2009. Examination of the context-specific nature of self-regulated learning. Educ. Stud. 35, 239–253.

Rotgans, J.I., Schmidt, H.G., 2014. Situational interest and learning: Thirst for knowledge. Learn. Instr. 32, 37–50.

Ruiz-Mirazo, K., Peretó, J., Moreno, A., 2004. A Universal Definition of Life: Autonomy and Open-Ended Evolution. Orig. Life Evol. Biosph. 34, 323–346. https://doi.org/10.1023/B:ORIG.0000016440.53346.dc

Sage, R.D., Atchley, W.R., Capanna, E., 1993. House Mice as Models in Systematic Biology. Syst. Biol. 42, 523–561. https://doi.org/10.2307/2992487

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Sanson, G.F., Kawashita, S.Y., Brunstein, A., Briones, M.R., 2002. Experimental phylogeny of neutrally evolving DNA sequences generated by a bifurcate series of nested polymerase chain reactions. Mol. Biol. Evol. 19, 170–178.

Scarcelli, N., Kover, P.X., 2009. Standing genetic variation in FRIGIDA mediates experimental evolution of flowering time in Arabidopsis. Mol. Ecol. 18, 2039–2049. https://doi.org/10.1111/j.1365-294X.2009.04145.x

Schneider, W.T., Rutz, C., Hedwig, B., Bailey, N.W., 2018. Vestigial singing behaviour persists after the evolutionary loss of song in crickets. Biol. Lett. 14, 20170654. https://doi.org/10.1098/rsbl.2017.0654

Seymour, E., Hunter, A.-B., Laursen, S.L., DeAntoni, T., 2004. Establishing the benefits of research experiences for undergraduates in the sciences: First findings from a three-year study. Sci. Educ. 88, 493–534. https://doi.org/10.1002/sce.10131

Sharpe, D., 2015. Your Chi-Square Test is Statistically Significant: Now What? Pract. Assess. Res. Eval. 20, 10.

Shtulman, A., 2006. Qualitative diverences between naïve and scientific theories of evolution. Cognit. Psychol. 52, 170–194.

Shtulman, A., Calabi, P., 2008. Learning, Understanding, and Acceptance: The Case of Evolution. Proc. Annu. Meet. Cogn. Sci. Soc. 30, 7.

Shtulman, A., Schulz, L., 2008. The Relation Between Essentialist Beliefs and Evolutionary Reasoning. Cogn. Sci. Multidiscip. J. 32, 1049–1062. https://doi.org/10.1080/03640210801897864

Shtulman, A., Valcarcel, J., 2012. Scientific knowledge suppresses but does not supplant earlier intuitions. Cognition 124, 209–215.

Simmons, M., 2012. Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. Cladistics 28, 208–222.

Sinatra, G.M., Brem, S.K., Evans, E.M., 2008. Changing Minds? Implications of Conceptual Change for Teaching and Learning about Biological Evolution. Evol. Educ. Outreach 1, 189–195. https://doi.org/10.1007/s12052-008-0037-8

Sleator, R.D., 2011. Phylogenetics. Arch. Microbiol. 193, 235–239. https://doi.org/10.1007/s00203-011-0677-x

Smith, G.P., Golomb, M., Billstein, S.K., Smith, S.M., 2015. The Luria-Delbrück fluctuation test as a classroom investigation in Darwinian evolution. Am. Biol. Teach. 77, 614–619.

Smith, J.J., Baum, D.A., Moore, A., 2009. The need for molecular genetic perspectives in evolutionary education (and vice versa). Trends Genet. 25, 427–429. https://doi.org/10.1016/j.tig.2009.09.001

Smith, J.J., Johnson, W.R., Lark, A.M., Mead, L.S., Wiser, M.J., Pennock, R.T., 2016. An Avida-ED digital evolution curriculum for undergraduate biology. Evol. Educ. Outreach 9, 9.

Smith, K.A., 1996. Cooperative learning: Making “groupwork” work. New Dir. Teach. Learn. 1996, 71–82.

Smith, K.A., Sheppard, S.D., Johnson, D.W., Johnson, R.T., 2005. Pedagogies of engagement: Classroom-based practices. J. Eng. Educ. 94, 87–101.

Sober, E., 1993. Experimental Tests of Phylogenetic Inference Methods. Syst. Biol. 42, 85–89. https://doi.org/10.2307/2992558

Sokal, R.R., Michener, C.D., 1958. A statistical method for evaluating systematic relationship. Univ. Kans. Sci. Bull. 28, 1409–1438.

Sousa, A., Zé-Zé, L., Silva, P., Tenreiro, R., 2008. Exploring tree-building methods and distinct molecular data to recover a known asymmetric phage phylogeny. Mol. Phylogenet. Evol. 48, 563–573. https://doi.org/10.1016/j.ympev.2008.04.030

Speth, E.B., Long, T.M., Pennock, R.T., Ebert-May, D., 2009. Using Avida-ED for teaching and learning about evolution in undergraduate introductory biology courses. Evol. Educ. Outreach 2, 415–428.

Speth, E.B., Shaw, N., Momsen, J., Reinagel, A., Le, P., Taqieddin, R., Long, T., 2014. Introductory Biology Students' Conceptual Models and Explanations of the Origin of Variation. CBE—Life Sci. Educ. 13, 529–539. https://doi.org/10.1187/cbe.14-02-0020

Stamatakis, A., 2016. The RAxML v8.2.X Manual.

Steel, M., 2013. Consistency of Bayesian inference of resolved phylogenetic trees. J. Theor. Biol. 336, 246–249.

Stone, A.C., Griffiths, R.C., Zegura, S.L., Hammer, M.F., 2002. High levels of Y-chromosome nucleotide diversity in the genus Pan. Proc. Natl. Acad. Sci. 99, 43–48.

Strelioff, C.C., Lenski, R.E., Ofria, C., 2010. Evolutionary dynamics, epistatic interactions, and biological information. J. Theor. Biol. 266, 584–594. https://doi.org/10.1016/j.jtbi.2010.07.025

Sukumaran, J., Holder, M.T., 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26, 1569–1571.

Sullivan, J., Joyce, P., 2005. Model Selection in Phylogenetics. Annu. Rev. Ecol. Evol. Syst. 36, 445–466. https://doi.org/10.1146/annurev.ecolsys.36.102003.152633

Sullivan, J., Swofford, D.L., 2001. Should We Use Model-Based Methods for Phylogenetic Inference When We Know That Assumptions About Among-Site Rate Variation and Nucleotide Substitution Pattern Are Violated? Syst. Biol. 50, 723–729. https://doi.org/10.1080/106351501753328848

Sun, S., Pan, W., Wang, L.L., 2010. A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. J. Educ. Psychol. 102, 989–1004. https://doi.org/10.1037/a0019507

Swenson, M.S., Barbançon, F., Warnow, T., Linder, C.R., 2010. A simulation study comparing supertree and combined analysis methods using SMIDGen. Algorithms Mol. Biol. 5, 8.

Swofford, D.L., 1998. PAUP*. Phylogenetic analysis using parsimony. Sinauer Associates, Sunderland, MA.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. 30, 2725–2729. https://doi.org/10.1093/molbev/mst197

Tateno, Y., Takezaki, N., Nei, M., 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol. Biol. Evol. 11, 261–277.

Tavaré, S., 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect. Math. Life Sci. 17, 57–86.

Taylor, T., Bedau, M., Channon, A., Ackley, D., Banzhaf, W., Beslon, G., Dolson, E., Froese, T., Hickinbotham, S., Ikegami, T., McMullin, B., Packard, N., Rasmussen, S., Virgo, N., Agmon, E., Clark, E., McGregor, S., Ofria, C., Ropella, G., Spector, L., Stanley, K.O., Stanton, A., Timperley, C., Vostinar, A., Wiser, M., 2016. Open-Ended Evolution: Perspectives from the OEE Workshop in York. Artif. Life 22, 408–423. https://doi.org/10.1162/ARTL_a_00210

Tehrani, J.J., 2013. The Phylogeny of Little Red Riding Hood. PLoS ONE 8, e78871. https://doi.org/10.1371/journal.pone.0078871

The International Society for Artificial Life Awards: Winners [WWW Document], 2017. . Int. Soc. Artif. Life. URL http://alife.org/news/2017-isal-awards-winners (accessed 2.23.18).

The UniProt Consortium, 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45, D158–D169. https://doi.org/10.1093/nar/gkw1099

Tibell, L.A.E., Harms, U., 2017. Biological Principles and Threshold Concepts for Understanding Natural Selection: Implications for Developing Visualizations as a Pedagogic Tool. Sci. Educ. 26, 953–973. https://doi.org/10.1007/s11191-017-9935-x

Trey, L., Khan, S., 2008. How science students can learn about unobservable phenomena using computer-based analogies. Comput. Educ. 51, 519–529.

Valverde, S., Solé, R.V., Elena, S.F., 2012. Evolved Modular Epistasis in Artificial Organisms. Artif. Life 13, 111–115.

Vartanian, J.-P., Henry, M., Wain-Hobson, S., 2001. Simulating pseudogene evolution in vitro: Determining the true number of mutations in a lineage. Proc. Natl. Acad. Sci. 98, 13172–13176. https://doi.org/10.1073/pnas.221334898

Wake, M.H., 2008. Integrative Biology: Science for the 21st Century. BioScience 58, 349–353. https://doi.org/10.1641/B580410

Wall, J.D., Pritchard, J.K., 2003. Haplotype blocks and linkage disequilibrium in the human genome. Nat. Rev. Genet. 4, 587.

Wang, L.-S., Leebens-Mack, J., Wall, P.K., Beckmann, K., de Pamphilis, C.W., Warnow, T., 2011. The Impact of Multiple Protein Sequence Alignment on Phylogenetic Estimation. IEEE/ACM Trans. Comput. Biol. Bioinform. 8, 1108–1119. https://doi.org/10.1109/TCBB.2009.68

Weaver, G.C., Russell, C.B., Wink, D.J., 2008. Inquiry-based and research-based laboratory pedagogies in undergraduate science. Nat. Chem. Biol. 4, 577.

Whelan, S., Goldman, N., 2001. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Mol. Biol. Evol. 18, 691–699. https://doi.org/10.1093/oxfordjournals.molbev.a003851

White, P.J.T., Heidemann, M.K., Smith, J.J., 2013. A New Integrative Approach to Evolution Education. BioScience 63, 586–594. https://doi.org/10.1525/bio.2013.63.7.11

Wiens, J.J., Cannatella, D., 1998. Does Adding Characters with Missing Data Increase or Decrease Phylogenetic Accuracy? Syst. Biol. 47, 625–640. https://doi.org/10.1080/106351598260635

Wiens, J.J., Soltis, P., 2005. Can Incomplete Taxa Rescue Phylogenetic Analyses from Long-Branch Attraction? Syst. Biol. 54, 731–742. https://doi.org/10.1080/10635150500234583

Wilcox, T.P., Zwickl, D.J., Heath, T.A., Hillis, D.M., 2002. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. Mol. Phylogenet. Evol. 25, 361–371. https://doi.org/10.1016/S1055-7903(02)00244-0

Wilke, C.O., Adami, C., 2002. The biology of digital organisms. Trends Ecol. Evol. 17, 528–532.

Wilke, C.O., Wang, J.L., Ofria, C., Lenski, R.E., Adami, C., 2001. Evolution of digital organisms at high mutation rates leads to survival of the flattest. Nature 412, 331.

Yang, Z., 2006. Computational molecular evolution. Oxford University Press, New York.

Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39, 306–314.

Yang, Z., Rannala, B., 2012. Molecular phylogenetics: principles and practice. Nat. Rev. Genet. 13, 303.

Yedid, G., Ofria, C.A., Lenski, R.E., 2009. Selective Press Extinctions, but Not Random Pulse Extinctions, Cause Delayed Ecological Recovery in Communities of Digital Organisms. Am. Nat. 173, E139–E154. https://doi.org/10.1086/597228

Yedid, G., Ofria, C.A., Lenski, R.E., 2008. Historical and contingent factors affect re-evolution of a complex feature lost during mass extinction in communities of digital organisms. J. Evol. Biol. 21, 1335–1357. https://doi.org/10.1111/j.1420-9101.2008.01564.x

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 Biol. 12, e1002023.

Zhang, Y., Sun, J., Rouse, G.W., Wiklund, H., Pleijel, F., Watanabe, H.K., Chen, C., Qian, P.-Y., Qiu, J.-W., 2018. Phylogeny, evolution and mitochondrial gene order rearrangement in scale worms (Aphroditiformia, Annelida). Mol. Phylogenet. Evol. 125, 220–231. https://doi.org/10.1016/j.ympev.2018.04.002

Zimmer, C., 2005. Testing Darwin. Discov. Mag.