Evolutionary Consequences of Gen(om)e Duplications in Animals

by

Kyle T. David

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 6, 2022

Keywords: Polyploidy, Gene Duplication, Molecular Evolution, Bioinformatics

Approved by

Kenneth Halanych, Co-chair, Executive Director, Center for Marine Science
Jamie Oaks, Co-chair, Assistant Professor, Department of Biological Sciences
Laurie Stevison, Assistant Professor, Department of Biological Sciences
Nathan Whelan, Assistant Professor, Department of Fisheries

Abstract

My research interests can be broadly described as an attempt to explore the drivers of macroevolutionary processes and diversity patterns across taxa. One such driver is the contribution of gene and genome duplications. As the primary source of new genes in eukaryotes, duplication events are important drivers of molecular novelty and adaptation. In my first chapter, I wrote software to explore patterns of evolution following gene duplication across 77 whole vertebrate genomes. My coauthors and I discovered that duplicated genes evolve at a greater rate, more likely to diversify and acquire new functions than genes that originate from speciation events. This work demonstrates how molecular changes contribute to phenotypic evolution and the maintenance of lineages. Later in my PhD, I became further interested in the role environment has on such a dynamic. My second chapter, focusing on whole genome duplication events, explored the relationship between environment and genome copy number to test the hypothesis that species with whole genome duplications (polyploids) are more adaptable to extreme environments. Using comparisons between polyploids and diploids across five genera of South American frogs I discovered that polyploid frogs are more closely associated with environments affected by human agriculture and anthropogenic climate change. This work was expanded in my third chapter, which demonstrates the existence of a latitudinal polyploid gradient in amphibians, ray-finned fish, and insects. This pattern appears to be driven largely by glaciation cycles, further speaking to the increased capacity for polyploids in adapting to new, rapidly changing environments.

As mentioned above, I am primarily motivated by a desire to explore the relationship between evolution and diversity. To me this includes the empirical inquiries mentioned above but also methodological questions related to our imperfect appreciation of biodiversity and how it can lead to misconceptions about evolution. For my fourth dissertation chapter I explored how high-throughput sequencing technology is being distributed across the tree of life and found that species evenness in high-throughput sequencing experiments has been steadily decreasing over time. My fifth chapter uses unsupervised machine learning to attempt to characterize novel protein sequences in poorly understood groups.

Acknowledgments

First, I'd like to offer my sincerest, deepest, thanks to my primary advisor, Dr. Ken Halanych. Thank you for always indulging my curiosity and supporting me wherever it led. The workshops and meetings I attended on your recommendation and with your support have proven to be foundational experiences. I also seem to recall a certain 3-month research cruise that will remain with me for the rest of my life. I hope and believe this dissertation represents only the beginning of a long and rewarding collaboration. I'd also like to thank Dr. Jamie Oaks for his tenure as my interim advisor, both at the beginning and end of my PhD. Thank you for the hours spent in front of whiteboards walking through the finer points of molecular evolution modeling. Thank you also for the opportunities and mentorship you've given me with regard to prison education, an area I hope to continue developing throughout my career. Your teaching, advising, and mentorship style is a model I aspire to.

To the remainder of my committee I'd like to thank you first and foremost for remaining faculty at Auburn University. Thank you Dr. Laurie Stevison for your teaching service, I've taken more credits from you than anyone else, and it probably would have taken me at least another year to graduate if not for your course offerings. Thank you Dr. Nathan Whelan for your guidance over the years. Thank you also for being a gracious winner at pinball, and a gracious loser at *Magic: the Gathering*.

I'd also like to thank my friends and coworkers, who were more often than not one and the same. Thank you Dr. Yuannig Li, Dr. Viktoria Bogantes, Dr. Michael Tassia, Caitlin Redak, Oluchi Oyekwe, Yu Sun, Dr. James Townsend, Charlie Schaefer, Julia Bae, and far more colleagues, collaborators, and companions than I could possibly list here. Thank you to my erstwhile landlord, desk-partner, and bunkmate Damien Waits for your friendship and for your perspective both inside and outside academia, which I value greatly. I would also like to thank my friend and partner, Dr. Haley Hallowell, for her kindness, patience, and support, I love you.

Finally, I would like to thank my family, without whom none of this work would even have begun. Thank you to my sisters Laura, Claire, and K.C. Thank you to my father, Chris David, for all your generous support and advice over the years, I hope this work serves as a worthy addition to the David academic canon. Thank you to my mother, Karen David, for your love and support for my interest in the natural world since before I could walk. This dissertation is dedicated to you.

Table of Contents

List of Tables

List of Figures

Chapter I. Background

Though the phenomenon of gene duplication has been known for well over a century, its significance to evolutionary theory was largely ignored until Susumu Ohno's seminal work *Evolution by Gene Duplication*[1]. In his book, Ohno proposes three possible fates for retained duplicated genes: conservation, under which both genes maintain the same ancestral function; subfunctionalization[2], under which both genes share aspects of the ancestral function; and neofunctionalization[2], under which one gene maintains the ancestral function while the other acquires new function (Fig. 1.1). In the same year, Walter M. Fitch recognized the significance between gene copies that arise through speciation events and those that arise through duplication events, coining them orthologs and paralogs, respectively[3].

The mechanisms of gene duplication vary, ranging from single genes (in the case of retrotransposition), to long stretches of chromosome (in the case of non-allelic homologous recombination), to entire genomes (in the case of whole genome duplication, (WGD)). The expectations for each of these scenarios is similarly varied. For example, because retrogenes are duplicated in isolation without flanking regions they are much less likely to remain functional. Relocated genes may also lack the right genomic/epigenetic environment to replicate the ancestral function, becoming more prone to subfunctionalization or neofunctionalization as a result[4]. Smaller scale duplication events also create dosage effects, to the benefit or detriment of overall fitness. Conversely, WGD replicates the entire genome sequence, maintaining gene synteny. As a result, each gene has the same relative dosage as well as all of its flanking regions, increasing the likelihood the gene will be able to maintain its ancestral function. Genes retained after WGD are more likely to exhibit sensitivity to dosage imbalance, and are often more complex, associated with development, regulation, and cell to cell signaling [5–7]. However on the genomic level WGD results in the most dramatic changes of all, long stretches of DNA and even entire chromosomes often experience rapid rearrangement and loss, often over the course of just a few generations[8–10].

WGDs can have traumatic effects on phenotype as well as genotype. Organisms which experience WGD, called polyploids, often exhibit increased cell[11,12] and body size enlargement[13,14] (though this relationship is variable and still poorly understood[15–17]). Polyploids also often suffer from reduced fecundity and fitness, and are generally considered less fit than their diploid counterparts[15,18]. However despite the deleterious effects, there is a growing

appreciation for the significant role WGD plays in animal evolution. As predicted by Ohno in his book[1], there are two rounds of WGD near the base of vertebrates[19], and another specific to teleost-fish[20]. As genomic sampling improves there is growing evidence for more WGD across the animal tree, including near the base of Lepidoptera and Odonata[21], as well as within several arachnid[22,23], mollusc[24], and peracaridans (work currently in prep). WGD are often associated with large diversification events[25,26], persistence through mass extinctions[27], and evolutionary novelties[28,29]. Even lineages that are currently diploid likely benefited from WGD at some point in their evolutionary history. Approximately 30% of all vertebrate genes are the product of ancestral WGDs [30]. As humans we owe our color vision to WGD [31], as well as our full complement of Hox genes [32], and several gene families associated with immunity [33]. The apparent paradox between the prolific legacy of WGD in animal evolution contrasted with the dearth of extant animal polyploids is the space where much of the research presented in this dissertation occupies.

The remainder of this work covering the last two chapters is concerned with the narrowing scope of genetic investigations toward a small minority of model organisms, and possible methods for mediating this bias. Advances in sequencing technology have resulted in the expectation that genomic studies will become more representative of organismal diversity. To test this expectation, we explored species representation of nonhuman eukaryotes in the Sequence Read Archive. Interpreting protein function from sequence data is a fundamental goal of bioinformatics. However, our current understanding of protein diversity is bottlenecked by the small number of proteins which have been functionally validated in model organisms, limiting inferences especially in clades without model representatives. Unsupervised learning may help to ameliorate this bias by identifying highly complex patterns and structure from large datasets without external labels.

Chapter II. Patterns of Gene Evolution Following Duplication and Speciation in Vertebrates

Introduction

Homologous relationships between eukaryotic genes are typically categorized as either orthologous or paralogous. Orthologs are gene copies that arise through speciation events and paralogs are gene copies that arise through duplication events (Fig. 2.1). The distinction between orthologs and paralogs has important implications for molecular biology largely due to the assumption that orthologs maintain similarity over time as genes are expected to serve comparable roles in descendant species[34–36]. By contrast, a much wider range of fates are considered for paralogs[37–39]. In his seminal work, Ohno[1] hypothesized several possible fates for duplicated genes. Paralogs may maintain their ancestral function (conservation[1]), gain new function (neofunctionalization[40]), divide or specialize the ancestral function between copies (subfunctionalization[40]), or lose function entirely. Many additional models have since been proposed but which broadly fall under these main categories[38,39,41,42].

The expectation of conservation between orthologs and divergence between paralogs is sometimes referred to as the "Ortholog Conjecture"[3,43]; an assumption so pervasive throughout genomics that it is not always referenced explicitly. For example, the Ortholog Conjecture is often used indirectly to infer gene function[35,43]. Under this method, the known function of a gene from *Mus musculus* would be assumed as the function for orthologs in other species where the gene has not been functionally characterized. Many popular online databases such as KEGG[44], PANTHER[45], and eggNOG[46] rely on evidence from orthology to assign function. Additionally, as of October 2019, there are 386,841 proteins with orthology evidence in the Swiss-Prot sequence database representing 70% of all entries[47].

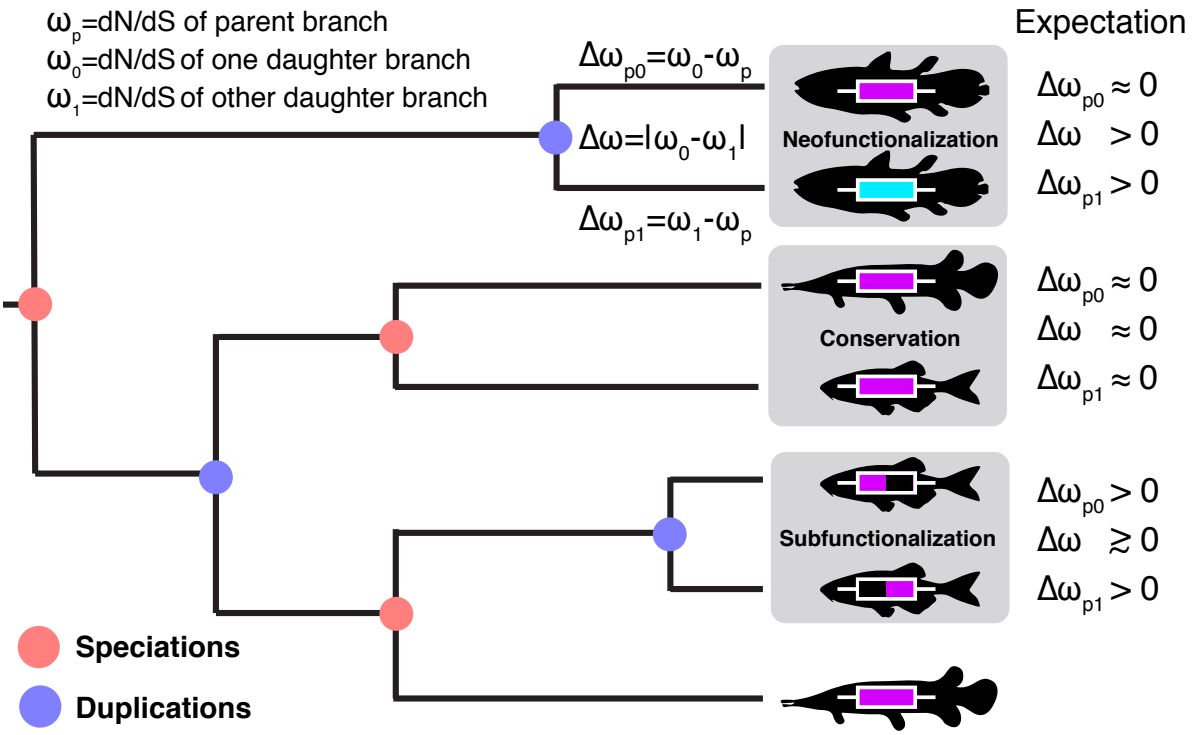

Fig. 2.1 A sample gene tree with duplication and speciation events. Genes that arose from a speciation event are orthologous to one another and genes that arose from a duplication event are paralogous. Under the Ortholog Conjecture, more changes are expected to occur along at least one of paralogous lineages following a duplication event. For each node we estimated the ratio of non-synonymous substitution rate (dN) over synonymous substitution rate (dS) for the two daughter lineages. We then calculated the absolute difference between the two ($\Delta\omega$) as well as the difference from the parent ($\Delta\omega_p$) for an estimate of divergence. Under neofunctionalization we would expect to see more nonsynonymous substitutions (positive $\Delta\omega_p$) in one lineage (whichever one is acquiring a new function) while the other lineage remains the same (small $\Delta\omega_p$), resulting in asymmetric selection between the two (large $\Delta\omega$). Under conservation, both lineages are expected to maintain similar levels of nonsynonymous substitutions as in the parent lineage (small $\Delta\omega_p$), resulting in symmetric selection (small $\Delta\omega$). Under subfunctionalization more nonsynonymous substitutions are predicted in both daughter lineages ($\Delta\omega_p$) as they partition aspects of the ancestral function. Subfunctionalization models generally do not make predictions regarding symmetry.

Given the importance of the Ortholog Conjecture in biology, the fact that only few studies test it explicitly (reviewed in[36,48] is surprising. A preliminary study using microarrays between human and mouse recovered significant differences in expression profiles between ortholog pairs, at levels comparable with gene pairs selected at random[49]. However, when these data were reanalyzed[50], orthologs were found to be significantly more similar to one another than paralogs. Nehrt et al. 2011[51] used Gene Ontology annotations and found no support for the Ortholog Conjecture between humans and mice, however other researchers have since noted several pitfalls associated with using Gene Ontology[52–55]. Recently, a study by Kryuchkova-Mostacci and Robinson-Rechavi[56] found strong support for the Ortholog Conjecture; however, reanalysis by Dunn, Zapata et al.[57] found the observed differences to be an artifact of node age,

not the divergence events themselves.

To address the validity of Ortholog Conjecture, we estimated the ratio of nonsynonymous to synonymous nucleotide substitution rates (ω=dN/dS) for daughter lineages descended from inferred speciation events (orthologous lineages) or duplication events (paralogous lineages). We then took the absolute difference of ω (Δω) between daughter lineages for an estimate of the difference in selective pressure experienced by daughter lineages (Fig. 2.1). If nonsynonymous substitutions are more likely between paralogs, we would expect to see greater Δω values following duplication events compared to speciation events (Fig. 2.1)[36]. We also measured the difference in ω of each daughter lineage from the parent ($\Delta\omega_p$) to see if one or both lineages diverged from the ancestral ratio. Δω was estimated for 234,066 speciation events and 16,978 duplication events in 6,244 gene families across 77 vertebrates.

Methods

We drew from 22,340 publicly available protein trees from the EnsemblCompara online database (release 90)[58,59]. To avoid issues with unknown calibration dates and branch lengths, we focused on patterns within vertebrates for 77 target taxa and 6 outgroup taxa with a well-established evolutionary history. Calibration dates taken from the most recent literature were assigned to 55 nodes in the species tree and implemented in a global clock model[60]. Under the global clock model one rate is used for each tree, represented in millions of years. Calibrating the data in this way enables us to compare nodes from different trees to one another, even though they likely have different rates of substitution. Trees that lacked at least one node with a calibration date (9,528) were discarded. Paralogs and orthologs were inferred using the species overlap algorithm[61]. Duplication labels were applied to nodes where the same species is represented in both daughter clades at least once. Speciation labels were then those nodes with no species overlap between the two daughter lineages. Put differently, speciation nodes gave rise to two discrete monophyletic groups, and duplication nodes did not (Fig. 2.1). Speciation and duplication labels generated from this method were compared with the annotations provided by EnsemblCompara[58], with which they were congruent in 97.5% of cases. This study ignores horizontal gene transfer events and their resultant xenologs, which are unlikely between vertebrates.

Trees with greater than approximately 170 tips (~5,000 trees) were found to be too

computationally intensive and excluded from our analyses. We also filtered 2,766 trees that did not contain at least one duplication and at least one speciation event. Of the remaining 6,303 trees, we filtered 31,039 nodes with an expected number of synonymous substitutions greater than 2 in either daughter lineage, which indicate saturation of substitutions, then 75,427 nodes with an expected number of synonymous substitutions less than 0.01 in either daughter lineage, which can lead to poor estimates of ω, and finally 61 nodes with ω >10 in one or both daughter lineages as outliers[62]. Our filtered dataset contains 6,244 trees with 16,978 duplication nodes and 234,066 speciation nodes.

We inferred selective pressure by estimating the rate of nonsynonymous substitutions relative to the rate of synonymous substitutions (ω)[63]. We estimated ω values with codeml, a maximum-likelihood method for codon-substitutions model within the PAML package[60]. To estimate Δω, we used a free-ratios model which allows separate ω values to be calculated for each branch in the tree[64,65]. The difference in selective pressure was then quantified by simply taking the absolute difference between the ω of the two daughter lineages (Δω) as well as the difference of one daughter lineage from the parent lineage ($\Delta\omega_P$) for each speciation and duplication node in our tree (Fig. 2.1). Per the PAML authors' recommendation we first ran a null model, which assumes uniform ω values to generate branch lengths and transition/transversion ratios in an effort to limit free parameters. All analyses were performed on the High Performance Computing Hopper Cluster at Auburn University.

To effectively test the Ortholog Conjecture, we compared empirical values to a null model, in which there was no difference between speciation and gene duplication events. For our null model, we assigned speciation and duplication labels randomly without replacement for each gene tree, removing any putative link between evolutionary events and Δω. To calculate p-values, we performed a two-tailed permutation test comparing our empirical estimates to those calculated from trees with permuted speciation/duplication node labels. Null distributions were approximated with 1,000 permutations. Under either model, the number of speciation and gene duplication nodes for each tree was kept the same as in the empirical trees. We also include Hedges' g[66] as a measure of effect size.

For our analysis, we were interested in testing not just differences between means but also similarity of distributions. Considering this, we calculated the overlap coefficient (OVL) between kernel density distributions of orthologous and paralogous lineages (Fig. 2.2A). The

OVL is the area in common under two probability density functions and represents the sum of the conditional misclassification probability as well as providing an intuitive measure of agreement between two similar distributions[67] . Under the most extreme interpretation of the Ortholog Conjecture, we would expect no overlap in Δω distributions between orthologous and paralogous lineages. OVLs were calculated by integrating the area under the intersection between the two density plots. We also perform a two-sample Kolmogorov-Smirnov test, which is a nonparametric test which estimates the likelihood of two samples (in this case Δω values between speciation and duplication events) being drawn from the same distribution[68].

All code required to update experiments and reproduce results/figures are available at https://github.com/KyleTDavid/OrthologConjecture2019. Original data files are available at https://figshare.com/projects/OrthologConjecture2019/63935.

Results & Discussion

On average, Δω was significantly smaller between orthologous lineages than paralogous lineages ($p<0.001$, K-S test $p<2.2E-16$, Hedges' $g=0.94$) (Fig. 2.2A). After a speciation event, resulting orthologs experience more similar patterns of molecular evolution (average Δω was 0.20 ± 0.38 standard deviations) to one another than do paralogs after a duplication event (0.60 ± 0.77). The OVL of kernel density distributions of Δω was 57.6% between orthologous and paralogous lineages. Small but significant ($p<0.001$, Hedges' $g=0.18$) differences in Δω were also recovered between different categories of duplication events, with greater Δω between duplication events leading to within-species paralogs (i.e., nodes in which all descendant leaves belong to the same species) (0.62 ± 0.77) than between-species paralogs (i.e., duplication nodes in which descendant leaves belong to more than one species) (0.48 ± 0.77).

We observe significant ($p<0.001$, Hedges' $g=0.88$) differences in the ω ratios themselves with an average ratio of 0.22 ± 0.36 for orthologous lineages compared to 0.56 ± 0.70 for paralogous lineages. High ω following gene duplications events is a well-documented phenomenon[4,41,69,70] . However whether or not relaxed selective pressures are experienced equally between copies remains controversial[71,72]. To address this, we also estimated the minimum and maximum difference between ω of each daughter lineage per pair and their parent lineage ($\Delta\omega_p$) (Fig. 2.1). ω for paralogous lineages increased by 0.21 ± 0.82 compared to their parent lineage, significantly ($p<0.001$, Hedges' $g=0.48$) higher than ω for orthologous lineages

which decrease by 0.003 ± 0.42. Although differences from the parent lineage were not recovered between the minimum orthologous and paralogous daughter lineages (p=1.00) they were recovered for the maximum lineages (p<0.01) (Fig. 2.2B). Additionally, effect sizes between the maximum and minimum lineages were >100x greater between paralogous (Hedges' g = 0.85) and orthologous lineages (Hedges' g = 6.7E-3).

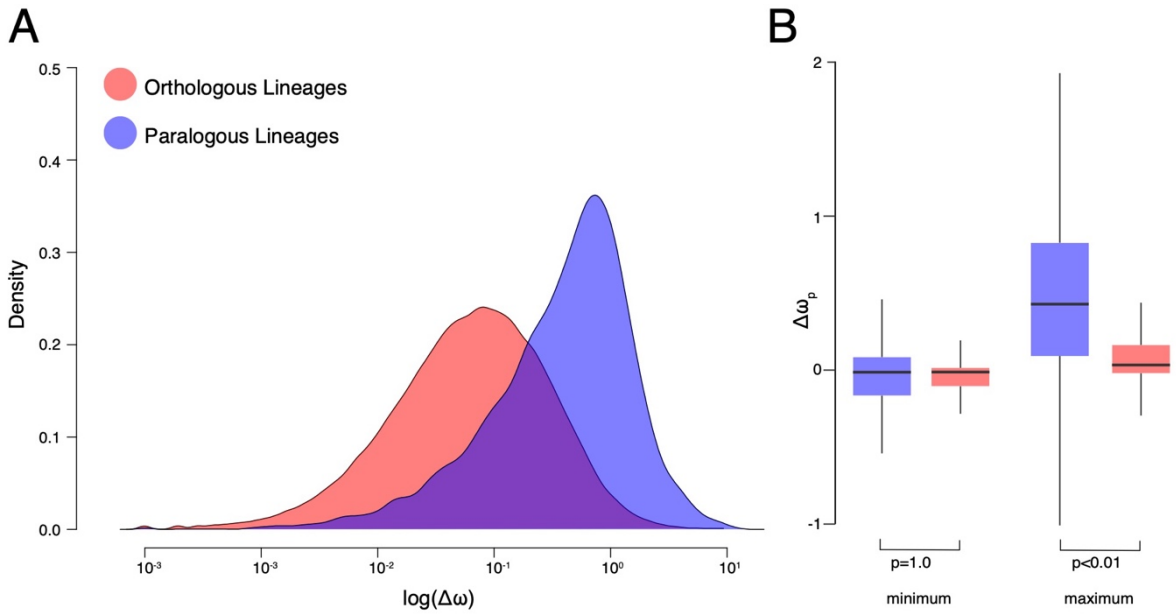

Fig. 2.2 A) kernel density plots of log transformed Δω of orthologous and paralogous lineages. B) $\Delta\omega_p$ of orthologous and paralogous lineages with each pair of daughter lineages categorized by maximum and minimum values.

Our results demonstrate support for the Ortholog Conjecture with regard to ω. The relative rates of nonsynonymous substitutions differs more between paralogous lineages than between orthologous lineages. Differences in amino acid sequence may further indicate functional/structural changes are more likely to occur between paralogs than orthologs as well, although the relationship between sequence, structure, and function is far from direct.

Our observation of higher Δω in paralogous lineages appears to be largely the result of a higher relative rate of nonsynonymous substitutions in just one daughter paralog (Fig. 2.2B). This pattern is commonly thought to be indicative of Ohno's neofunctionalization[1] or Francino's adaptive radiation[73] models of evolution[38,42], indicating that diversification may be the most common fate for at least one retained paralog copy. This interpretation is congruent with previous studies[4,74,75], such as Brunet et al. 2006[5] who also observed conserved vs. elevated selection between paralogous lineages resulting from the teleost specific whole genome

duplication event. It is worth noting; however, that a later study[76] using a more rigorous model found evidence for high levels of selection across lineages throughout vertebrates, regardless of homologous relationships.

Most gene duplication events in our dataset occur along lineages for which there is no evidence for whole genome duplication (98.3%), indicating that the Ortholog Conjecture is as, if not more, pronounced in single/several gene duplication events (such as those produced through unequal crossing-over) than whole genome duplication events. New gene copies generated through whole genome duplication are identical in terms of just not their sequence but location and context within the subgenome as well[36]. By contrast retrotransposed gene duplicates are relocated to an entirely new genomic environment in which it will be unlikely to reproduce the ancestral function and thus free to acquire new mutations. This scenario is supported by Han et al. 2009[4] who found that gene duplicates transferred to a new position are more likely to experience elevated ω ratios.

There were several clades in which the Ortholog Conjecture was not supported ($p>0.05$), namely: Sarcopterygii, Neopterygii, Amniota, and Mammalia. This suggests an unexpected trend of decreasing support for the Ortholog Conjecture as time since divergence increases (Fig. 2.3). This trend is likely the result of saturation in substitutions[77] over time as the expected number of synonymous substitutions begin to plateau at ~150 mya while the expected number of nonsynonymous substitutions continue to increase, obscuring differences between duplication and speciation events and increasing the overlap between the distributions. Additionally, estimates of ω have been demonstrated to exhibit slight bias when divergence levels are low[78]. As a result, more work may need to be done in order to more definitively resolve questions of ortholog and paralog evolution along especially old or young vertebrate lineages.

A similar pattern was recovered between different classifications of paralogs, with slightly smaller $\Delta\omega$ for between-species paralogs than within-species paralogs. This finding disagrees with previous results which suggest more conservation between within-species paralogs than between-species paralogs[51,53] . This discrepancy may be the result of saturation as noted above, as within-species duplication events are necessarily concentrated toward the tips of a phylogeny. However, when all nodes were filtered to those <1mya significant ($p<0.01$) differences were still recovered.

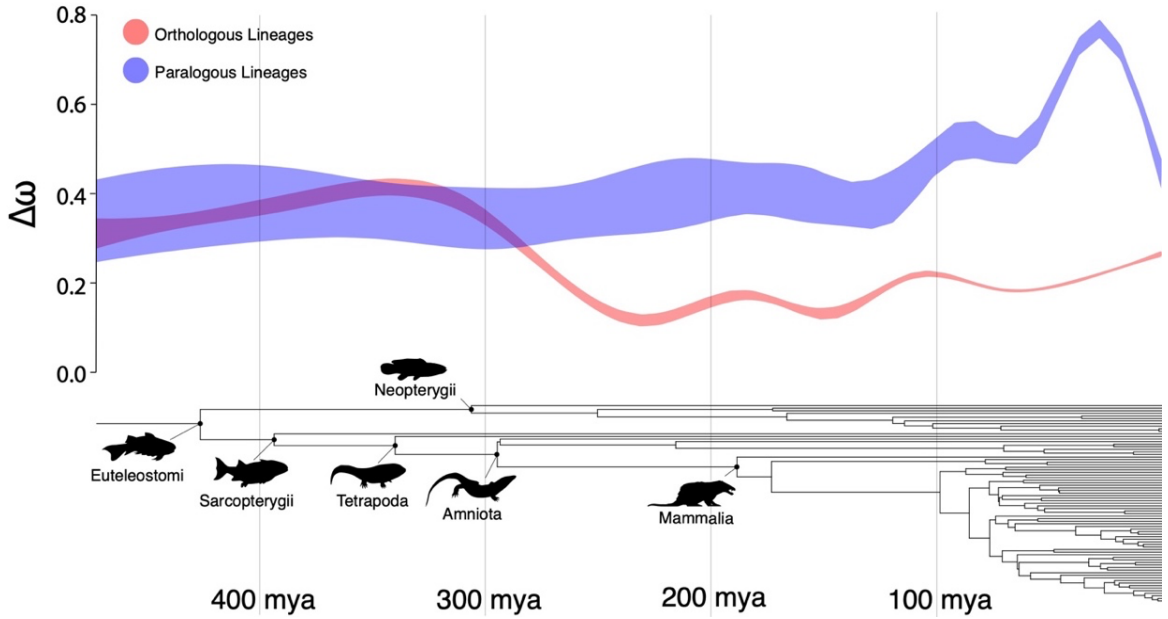

Fig. 2.3 General additive model 95% confidence intervals of Δω following speciation and duplication events over time, over the time-calibrated species tree of all taxa included in the study.

It should be noted that our approach only estimates the expected average ω for each branch. This could be a source of error, particularly if functionally relevant mutations occur early or late in the history of a lineage[4,38]. Similarly, our method is ignorant to models that involve multiple predictions at different stages of a paralog's history. For example, He and Zhang's subneofunctionalization model predicts paralogs undergo a short subfunctionalization period followed by sustained neofunctionalization[79].

In conclusion, our results reveal distinct differences in patterns of substitution between lineages descending from speciation and duplication events. ω ratios are on average more similar between daughter lineages following speciation events than duplication events. This greater amount of asymmetric evolution between paralogs appears to be driven by an increase in relaxed selection in just one of the lineages. The maximum change from the parent branch is 5.6x greater for paralogous lineages than orthologous lineages, whereas the minimum change from the parent branch is nearly the same (0.98x) (Fig. 2.2B). Taken together these trends seem to indicate that neofunctionalization, where one copy retains the ancestral function while the other acquires a new function, as the most dominant pattern for retained duplicated genes.

Chapter III. Spatial Proximity Between Polyploids Across South American Frog Genera

Introduction

Polyploids are organisms with more than two complete sets of chromosomes. Polyploids are created *via* whole genome duplication from diploids, either between separate lineages through hybridization (called allopolyploidy) or within a single lineage through the production of unreduced gametes (called autopolyploidy). The duplication of the genome is one of the most severe mutations found in nature, and polyploids are expected to experience unique evolutionary pressures and altered trajectories as a result[14,16,80,81]. In particular, polyploids are often found in extreme environments, which we define as those environments which exhibit statistically greater or lower values of a given environmental variable when compared to other occurrences within the clade. For example, polyploids can be found under conditions that are on average colder[82–85], drier[29,86–88], or more saline[89] than their diploid counterparts. This trend is best characterized in plants, but has also been observed in *Daphnia* water fleas[83] and *Neobatrachus* frogs[29]. Polyploids are also more common in disrupted environments that are subject to radical changes, such as those that have been previously glaciated[90–93].

Several nonexclusive hypotheses have been proposed to explain this correlation, which may be the result of selective or nonselective processes. First, the formation of polyploid gametes may be more likely within extreme or disrupted environments. Meiotic errors that lead to autopolyploids have been shown to occur more frequently under stressful conditions[94,95]. For example several mammalian cell types will undergo spontaneous polyploidization in response to metabolic or genotoxic stress[96]. Exposure to extreme temperatures is also known to induce autopolyploidy in a wide variety of organisms[97,98], including frogs[99,100]. In the case of allopolyploidy, environments that are frequently disrupted also provide potential hybrid zones, allowing secondary contact between species previously separated by glaciers[82,84] or sea-level rise[101]. As a result, we would expect a greater incidence of allopolyploids in these areas.

Extreme environments may exhibit conditions that are conducive to the formation of polyploids, but they are also less hospitable to diploids, which may in and of itself increase the success of polyploid populations. Although polyploids initially share their range with diploids, they are not predicted to be able to exist sympatrically for long. In a sexual polyploid population,

individuals increase fitness only by mating with other polyploids, not the ancestral diploid species[102]. The existence of a prezygotic reproduction barrier is therefore beneficial to polyploid populations[103]. Diploids and polyploids must also compete for resources, a scenario in which polyploids may be outmatched. Generally, polyploids are considered to be less fit when compared to closely related diploids for a variety of reasons. First, the diploid genome has been optimized for the environment under natural selection for far longer[18]. Despite evidence for hundreds of genome duplication events across the animal tree of life only a minority of extant species exhibit polyploidy, suggesting that the ultimate fate for most polyploid lineages is to either evolve into diploids or become extinct. Polyploid genomes, especially for allopolyploids, are often unstable and suffer from disruptions to gene expression and cellular processes which may incur fitness costs[15,16,80]. Indeed, fertility and survival has been shown to be reduced in polyploids[104]. Polyploids also experience greater mutation load due to their larger gene copy number[16]. Therefore, due to decreased competition from diploids and a reduced chance of cross-ploidy mating, polyploids are more likely to thrive in areas where diploids are less frequent[14,81,105]. Several studies across plants have identified niche differentiation as a significant factor in the formation and maintenance of polyploid populations[102,103,106–111], demonstrating the importance of ecological divergence between polyploids and diploids.

Polyploid species are not predicted to survive within the environment they originated in for long, however, they may especially well-suited to life outside it, as they are often thought to possess greater environmental resilience and adaptive potential compared to diploids. Polyploids' increased genome size results in more genetic variation and a higher average mutation rate per gene[16,112]. Polyploids also possess a greater capacity to mask deleterious mutations and safeguard against inbreeding depression *via* a lower incidence of homozygous genotypes[14,16]. Multiple gene copies also provide the opportunity for adaptation, as each gene now has a redundant copy that is free to specialize or evolve new function without compromising the ancestral role[1,80]. Additionally, as allopolyploids are also hybrids, heterosis may be involved as well[16,80]. There have been whole genome duplications detected across many different plant lineages near the Cretaceous-Paleogene boundary[27,29,113,114], speaking to polyploids' ability to adapt to or resist periods of severe ecological disruption. A link has also been discovered between polyploidy and invasiveness[115], polyploid plants are 20% more likely to be invasive than diploids[116], further indicating their ability to adapt to new conditions.

To test whether polyploids are more closely associated with particular environments we collected 13,556 occurrence records of 82 species across the 5 South American frog genera with verified polyploid (n=1,717) and diploid (n=11,839) members (*Ceratophrys, Chiasmocleis, Odontophrynus, Phyllomedusa,* and *Pleurodema*)[117].

Methods

Occurrence Records

Each of the five genera were queried in the Global Biodiversity Information Facility (GBIF) and downloaded on Jan. 16, 2020. GBIF is an international organization which organizes and curates biodiversity data provided by institutions from around the world. Although 78 institutions in total contributed occurrences to the dataset used in this study, the top 3 were the Bernardino Rivadavia Museum of Natural Science, Kansas University, and the Smithsonian Museum, which together contributed 45.3% of occurrences. The original dataset includes 15,339 occurrences from 113 published datasets (https://doi.org/10.15468/dl.akr84v). Synonyms were updated with the Open Tree Taxonomy[118] and AmphibiaWeb was used to resolve discrepancies between GBIF and Open Tree. 1,672 occurrences without species names were removed. Two species (three occurrences) that were no longer considered members of any of the five genera were also removed. 108 occurrences of the species *P. burmeisteri* were additionally removed as the species contains both diploid and tetraploid populations[119] and so individual occurrences could not be classified with any certainty. All species not verified as polyploid are presumed diploid. The final dataset contains 13,556 occurrences across 82 species. Range maps from the International Union for the Conservation of Nature[120] were also downloaded (version 2019-3) for all available species. The full record of each occurrence with all available metadata is available at https://doi.org/10.15468/dl.akr84v.

Range Overlap Estimation

Range estimates were performed using species distribution models (SDMs). Five climatic variables were collected from the WorldClim 2 dataset[121] at 30 second resolution ($\gtrsim$0.86km^2): mean monthly temperature, mean monthly precipitation, altitude, temperature seasonality, and precipitation seasonality. Seasonality describes the annual range of a variable, estimated from the standard deviation for temperature and the coefficient of variation for precipitation. SDMs were then constructed for each polyploid species using these climatic variables as predictors with a

maximum entropy approach[122]. From these SDMs, range overlap[123] was estimated using the age-range correlation method[124]. SDMs were executed using the R package dismo 1.1.5[125] and range overlap was estimated using the R package phyloclim 0.9.5[126].

Estimation of range overlap between >2 lineages requires a phylogenetic tree. To achieve this, a concatenated multiple species alignment of the 12S and 16S mitochondrial gene from 57 frog species (including all polyploids) was generated using T-Coffee[127]. A maximum likelihood tree was then constructed in IQ-TREE 2.0.3[128] under the GTR+F+I+G4 substitution model, the best-fitting model as determined by ModelFinder[129]. Neither genes, nor sites, were partitioned. The tree was then time-calibrated using a penalized likelihood method[130] using the chronos function as implemented in the R package ape 5.3[131]. Lower and upper age bounds for each genus were taken from the available literature and used to time-calibrate the maximum likelihood tree. All files required to evaluate and reproduce the tree building steps are available at https://github.com/KyleTDavid/FrogPloidy2020/tree/master/Tree.

Categorical Environmental Variables

Three main data sources were used to categorize environments across South America. The first was the Köppen-Geiger climate classification[132,133], which categorizes the earth's landmass into five groups on the basis of climate (Tropical, Arid, Temperate, Continental, Polar), and then further into 30 total subgroups on the basis of seasonal temperature and precipitation trends. Herein, we use an updated digital version of the original classification system[134]. Environmental regions were also categorized into one of 10 biomes, and further divided into a specific ecoregion (defined as a "large unit of land or water containing a geographically distinct assemblage of species, natural communities, and environmental conditions") as defined by the World Wildlife Fund [135]. Each individual occurrence was given a climate, biome, and ecoregion classification based on geographic location.

Quantitative Environmental Variables

In addition to categorical data, we collected several continuous environmental variables where we had some *a priori* expectation they would influence frog habitability, and where data was available throughout South America. As previously mentioned, five climatic variables (mean monthly temperature, mean monthly precipitation, altitude, temperature seasonality, and precipitation seasonality) were collected from the WorldClim 2 dataset[121]. Additionally, five anthropogenic variables (pasture usage, cropland usage, fertilizer application, manure

application, and pesticide application) were taken from the Socioeconomic Data and Applications Center. Inputs from fertilizer and manure were collected from Global Agricultural Inputs v1 dataset[136], land usage from pastures and croplands were collected from the Global Agricultural Lands dataset[137], and pesticide application was collected from the Global Pesticide Grids dataset[138].

Quantitative environmental variables were analyzed in several ways. Wilcoxon rank-sum tests were used to directly compare quantitative environmental variables between diploid and polyploid occurrences. As many occurrences were observed from the same species (and all species share some degree of common ancestry) occurrences cannot be said to be truly independent, and so violate an assumption of the test. We hope that by sampling polyploid and diploid species across several different genera the impact of this violation is reduced, however we cannot discou9nt the possibility that observed differences are the result of shared ancestry between occurrences rather than ploidy alone. To explicitly incorporate evolutionary history into our analyses we also perform a phylogenetic ANOVA[139] using the R package geiger 2.0.7[140], though with only eight polyploid species statistical power is severely limited. To explore polyploid and diploid distributions in higher-dimensional space, principal component analysis (PCA) employing singular value decomposition was conducted on all occurrences using every quantitative environmental variable. As all the same caveats apply as with the Wilcoxon rank-sum tests, a phylogenetic PCA[141] was also performed using the R package phytools 0.7.47[142].

Results & Discussion

Distribution of Diploid & Polyploid Frogs

For South American frogs, ploidy appears to be strongly correlated with species range. Range maps of diploid species collectively cover most of the continent (Fig. 3.1A). Noticeable exceptions to this are much of the Andes Mountains along the western coast and part of the Guiana Highlands in southeastern Venezuela. Diploids are also largely absent from a region in the southeast which encompasses southern Brazil, southern Paraguay, parts of Uruguay, and northeastern Argentina. Unlike the Andes or Guiana Highlands there are no clear topographical features that would prevent frogs from becoming established in this region. Indeed, this is where most of the polyploid species are located. The area south of 25°S and east of 65°W contains 64.8% of polyploid occurrences but only 7.1% of diploids (Fig. 3.1A). SDMs of polyploid and

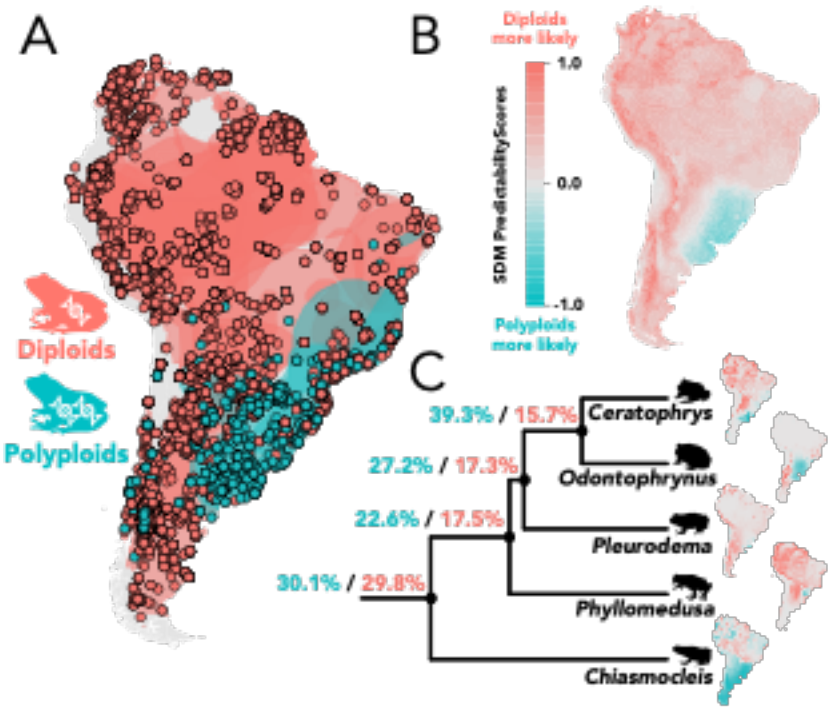

Figure 3.1. A) GBIF occurrences and IUCN range maps of the 82 frog species included in this study, colored by ploidy. B) Predictability scores generated from SDMs for diploids, subtracted by those for polyploids. Positive values (red) indicate areas where diploids are more likely to be present whereas negative values (blue) indicate areas where polyploids are more likely to be present. C) Range overlap between polyploid (blue) and diploid (red) species across genera, as estimated by age range correlation. Interpretation for genera specific maps is the same in Fig. 3.1B.

diploid taxa also reveal tighter spatial associations between polyploids. Diploids are more likely to be present throughout 88.4% of South America, with the exception of the southeastern region (Fig. 3.1B). Ancestral reconstructions of range overlap are also greater between polyploids than diploids across genera (Fig. 3.1C).

Characterizing the Polyploid Environment

In addition to geographic range, distributions across climate, biome, and ecoregion were also different between diploids and polyploids (Fig. 3.2). Polyploids are associated with temperate climates, with 84.7% of occurrences located within temperate regions, compared to 40.9% for diploids. This disparity is particularly true of the humid temperate climate, containing 49.5% of all polyploid occurrences but just 5.2% of diploids. Conversely diploid occurrences are most likely to be located within tropical climates (44.3%). The most common of these is the tropical rainforest climate (17.3%) for which only two polyploid occurrences have ever been reported (0.1%). Similarly, polyploids have less than half the relative frequency in forested biomes (31.1%) than diploids (63.2%). Instead, polyploid occurrences are more common in grasslands, savannas, and shrublands (58.7%).

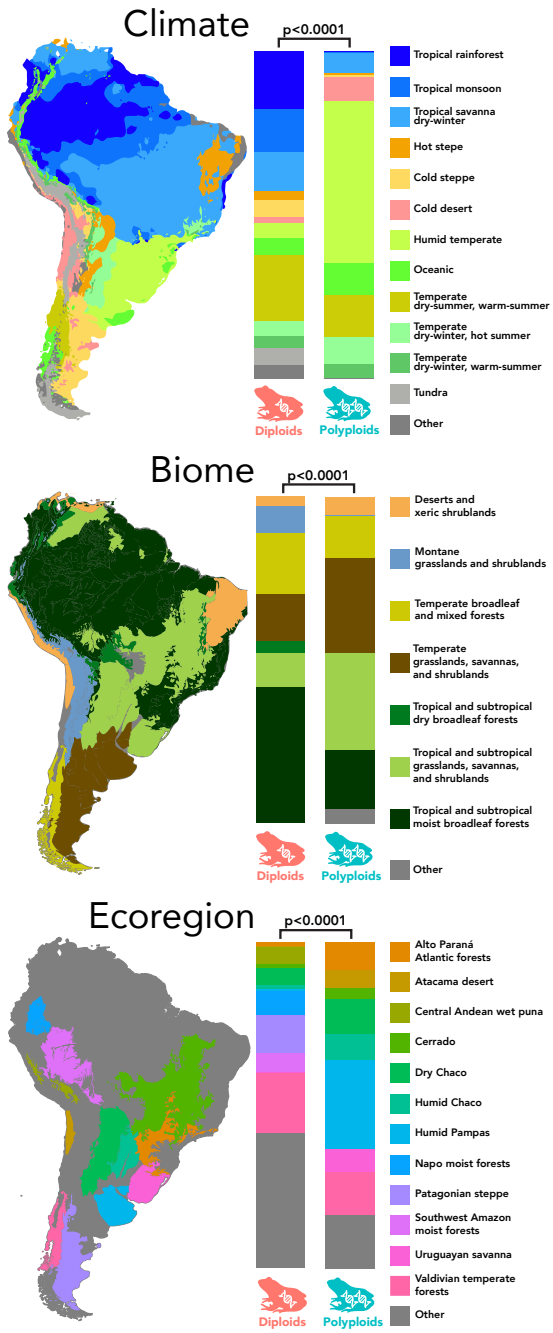

Figure 3.2. Categorical maps of South America accompanied with stacked bar plots showing the relative representation of diploid and polyploids on the bassis of climate, biome, and ecoregion.

Generally, polyploids appear more common throughout temperate grasslands and similar environments and are largely absent from the more tropical forested regions to the north. Occupying more temperate climates indicates that polyploids likely experience more seasonality, rather than the binary wet and dry seasons of the tropics. Indeed, temperature seasonality was the only variable with significant ($p<0.05$) differences between diploid and polyploid species across the phylogeny, though the strength of this trend varies across genera (Fig. 3.3).

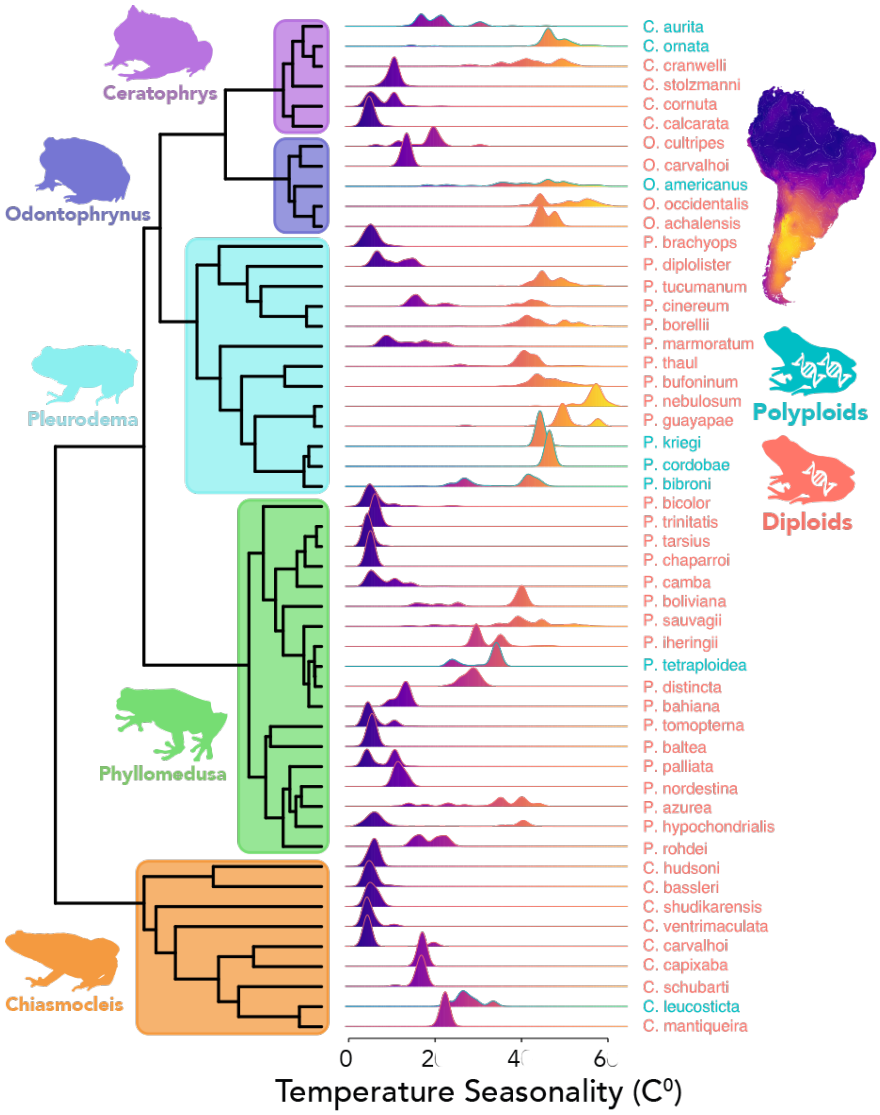

Figure 3.3. Density plots of temperature seasonality for each species with greater than two occurrences and available genetic data. Polyploid species on average occur in areas with greater temperature seasonality than diploid species ($p<0.05$).

In addition to being more temperate, seasonal, and less forested than the surrounding area, southeastern South America is subjected to unique anthropogenic inputs as well, having undergone an agricultural boom since the turn of the 20th century. Polyploid occurrences are more likely to be found in areas with large human impacts than diploids. In each of the sampled genera, polyploids were more frequent in areas with higher cropland usage, fertilizer application, and pesticide application compared to diploids between every comparison with statistical significance ($p<0.05$) (Fig. 3.4).

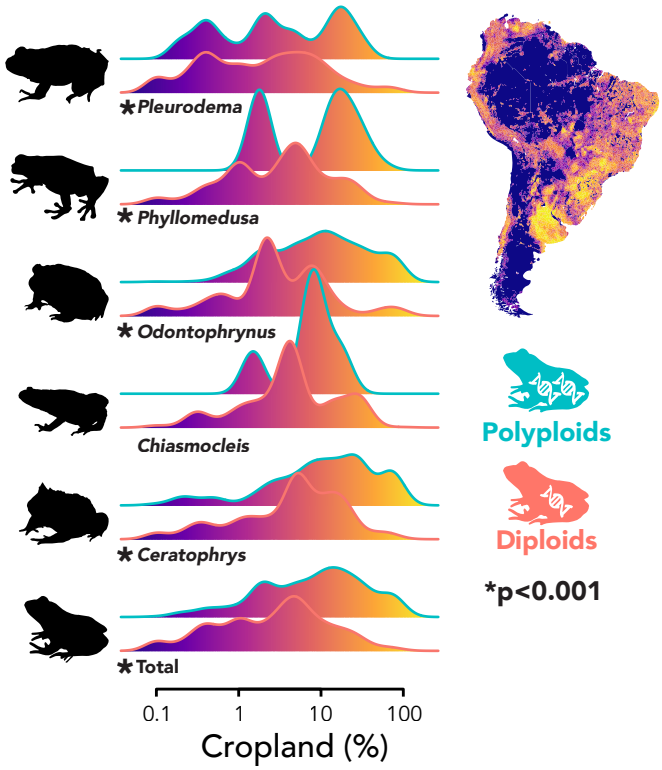

Figure 3.4 Cropland usage density plots for diploids and polyploids across and within genera. For all comparisons with statistical significance ($p<0.05$), polyploid occurrences are on average greater than diploid occurrences. Similar results were also recovered for fertilizer and pesticide application.

Discussion

As opposed to diploids, which are dispersed throughout most of South America, polyploids appear to occur almost exclusively within the southeastern region, regardless of genus. Occurrence data reveals a striking pattern between environment and ploidy. On average, polyploids are nearer to other polyploid species of different genera than they are to diploids of their own genus. This close spatial affinity between polyploids across genera suggests a particular region or environment that is conducive for polyploidy. We identify several unique features of southeastern South America that may have nonexclusively contributed to the exceptionally high occurrence of polyploids in the area.

First, the unique geological history of the region is worth briefly noting. The Paranese seaway was a marine incursion into the Paraná Basin (encompassing modern day west Brazil as well as northern Argentina and Uruguay) during the Mid to Late Miocene (~7-15 Ma)[143–145]. The subsequent retreat of the Paranese seaway would have provided new terrestrial environments for polyploid species to take advantage of. The retreat of the Paranese seaway as a geographic barrier may also have created new hybrid zones, leading to increased allopolyploid formation. According to phylogenetic analyses, all polyploid species diverged from their closest sampled relative during or after this period (<11.1 Ma). However, the Paranese seaway is not the only marine incursion to have occurred in South America, with the Tethys waterspout and Amazonian sea hypothesized to have existed at around the same time[146], so geological transformations alone are not sufficient to explain the occurrence of polyploids.

Another notable characteristic of southeastern South America is its seasonality, subject to fluctuations in temperature rather than precipitation compared to the surrounding area. Temperature seasonality was significant ($p<0.05$) between polyploid and diploid species across genera (Fig. 3.3). Temperature fluctuations have been previously identified as a condition common to fish and amphibian polyploids[97], possibly related to their resilience to disrupted environments. However temperature changes could also lead to an increase in polyploids through a nonselective process, as temperature shocks are a well-documented method for inducing autopolyploidy in many aquaculture species[98], as well as frogs[99,100].

Lastly, we consider relatively recent transformations to the region. Over the past century southeastern South America has experienced one of the largest upward precipitation trends in the world [147,148], linked to sea surface temperature anomalies[149–151] and ozone depletion[152]. This increased wetting has resulted in a farming boom[150–152], creating the most agriculturally productive region of South America[137]. Unfortunately this has also resulted in widespread habitat destruction and pollution, with southeastern South America suffering one of the largest biodiversity declines in the world[153]. Contamination from agrochemical inputs is of special concern to the conservation of amphibians, who can absorb deleterious chemicals through their permeable skin[154,155]. The Pampas ecoregion in particular has undergone two periods of ecological collapse over the last century as a result of unsustainable land management[156]. Notably, the Pampas is the most common ecoregion for polyploid occurrences (27.0%), where they have 54 times more relative frequency than diploids (0.5%) (Figure 3.2). In addition to possessing increased adaptive potential to extreme or novel environments, polyploids are also theorized to be more resilient to sudden and severe disruptions[29,157], such as those caused by anthropogenic intervention[158,159]. However empirical evidence remains controversial. A recent study comparing the effects of agrichemicals on the polyploid *O. americanus* and diploid *O. cordobae* found that although the polyploid species exhibited more micronucleases and cytoxic effects, possibly due to their increased genome or cell size, they showed significantly less nuclear abnormalities, indicating diploids may be more sensitive to pollutants [160]. This sensitivity may be why polyploid occurrences on average exhibit greater degrees of cropland usage as well as fertilizer and pesticide application than diploids (Fig. 3.4).

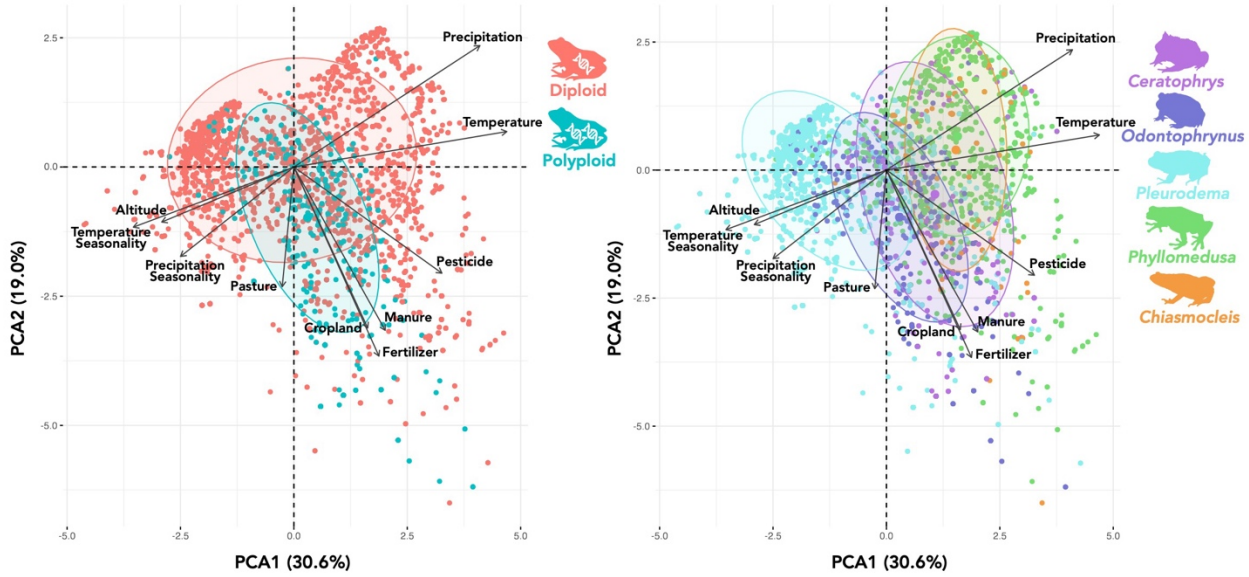

Figure 3.5. PCA biplot of all quantitative variables and occurrences across South America, colored by ploidy (left) and genera (right). Arrows indicate how each variable influences a principal component. Generally, arrows aligned to one another are correlated, positively if pointing in the same direction and negatively if pointing in opposite directions. Similarly arrows orthogonal to one another are uncorrelated. Each point represents an occurrence. Ellipses are drawn around points within one standard distribution.

PCA results show a similar trend (Fig. 3.5). Whereas diploid occurrences are distributed throughout feature-space, polyploids align more narrowly along the correlated anthropogenic variables. Furthermore, polyploid occurrences typically occur further along this "anthropogenic axis" than diploids. Phylogenetic PCAs reveal a similar pattern with regard to temperature seasonality. Taken together, these results seem to suggest that polyploids are more closely associated with areas with greater temperature fluctuations as well as those more highly impacted by human activities than diploids.

Conclusion

Occurrences of polyploid frog species reveal a pattern of significant clustering across genera. The tight spatial relationship between polyploids, and absence of diploids, suggests the environment of southeastern South America may be particularly conducive to polyploidy. This region is characterized by a temperate climate and dominated primarily by grassland, in contrast to the tropical rainforests in the north. Southeastern South America is also more seasonal than the surrounding region, experiencing higher temperature fluctuations as a result, which may lead to a higher incidence of polyploidization. The region has also experienced radical changes over just the past century, further influencing the occurrences and species ranges we observe today.

Southeastern South America has recently been subjected to large agricultural transformations, resulting in a steep decline in biodiversity. This radical shift may provide new opportunities to polyploid populations who are more resilient and/or better able to adapt to changing environments previously occupied by diploids. These nonexclusive scenarios are supported by previous studies which indicate polyploids are more resilient and adaptable to extreme conditions, as well as more likely to take advantage of new environments[16,29,80]. Southeastern South America is also noteworthy for its lack of diploid occurrences, possibly related to the environmental differences we described. This absence may also provide opportunities for polyploids they would have less competition and chance for cross-ploidy mating with diploids. Clearly, more work is required to distinguish between each of these hypotheses explicitly and may require sampling genomic to look for evidence of population expansions and/or migrations.

Chapter IV. Global Gradients in the Distribution of Animal Polyploids

Introduction

Whole genome duplication (WGD) is one of the most drastic mutations found in nature. Lineages which experience WGD are called polyploids, and can inherit their genome copies from either the same species (autopolyploidy) or between two different species (allopolyploidy). The effects of polyploidy vary dramatically but can have profound impacts on an organism's expression, cytotype, and phenotype, altering their evolutionary trajectory as a result[14,16,80,81]. The evolutionary consequences of polyploidy is best understood in plants, where approximately 35% of modern species are polyploid[161]. In animals, polyploidy is much less common and WGD has historically not thought to be a significant evolutionary driver[13,14]. However, WGD events have since been discovered at the base of some of the most successful animal clades[5,21,22], including vertebrates[1,19,162]. Despite the growing appreciation for the contribution of WGD in animal evolution, the significant factors influencing how polyploid animal lineages are established and maintained are still not well understood. One such factor is the possible presence of a latitudinal polyploid gradient. The tendency for polyploid species to increase with latitude has long been commented on in plants, as well as in several branchiopod species[163–168].

Several hypotheses have been proposed to explain this phenomenon, which may result from selective or non-selective processes. First, environmental conditions associated with higher latitudes may simply create more polyploids. Cold shocks are a common way to induce WGD a in a wide variety of species[97,98,169–171]. Additionally, heavily glaciated regions facilitate hybridization through secondary contact, a process commonly associated with polyploidy[82,84]. In addition to being conducive to the formation of polyploids, high-latitude environments may also be particularly favorable to the survival and maintenance of polyploid lineages. Many polyploids have broader ecological tolerances than their diploid counterparts and are predicted to be more adaptable to less hospitable or newer environments[107,172–174]. This is because the increased genome size following WGD affords polyploids more genetic variation and a higher average mutation rate per gene[16,112]. Multiple gene copies also creates opportunity for adaptation, as each gene now has a redundant copy that is free to specialize or acquire new function[1,80,175]. Additionally, since allopolyploids are also hybrids, factors associated with heterosis may also be

involved[16,80]. Polyploids are associated with areas that have recently been glaciated[92,93], and polyploid plants are 20% more likely to be invasive[115,116], speaking to their ability to acclimate to new environments.

Lastly, it is well-established that higher latitude environments exhibit reduced species richness, a near-ubiquitous pattern known as the latitudinal diversity gradient[176]. This reduced richness may in of itself provide a benefit to polyploids. Newly formed polyploids are often at risk of extinction from surrounding diploid species, either through direct competition (neopolyploids often suffer from decreased fitness[15,16,80,169]) or inviable cross-ploidy mating (a phenomenon known at the minority cytotype exclusion principle[102,103]). Polyploids may therefore be more likely to establish themselves in species poor areas with less competition and greater prezygotic reproductive barriers between themselves and surrounding diploids[14,81,105]. A recent study in plants confirmed the existence of a latitudinal polyploid gradient in that clade and attributed it in large part to these factors[85].

To explore the existence of a latitudinal polyploid gradient in animals and its potential drivers I collected global geographic, environmental, and climatic data coupled with evolutionary histories in three animal clades known to exhibit polyploidy: Amphibia (amphibians), Actinopterygii (ray-finned fishes), and Insecta (insects). A dataset of this size allows me to test previously unvalidated hypotheses on a global scale within and between taxa using a phylogenetically minded approach.

Methods

Species Occurrence Records

The Global Biodiversity Information Facility (GBIF) database was queried for occurrence data on May 23rd 2022. Three separate searches were performed, with each restricted to presence only data occurrences from one of the three clades under consideration by this study (Amphibia, Actinopterygii, Insecta). Occurrences with the GBIF 'basis of record' field *living specimen* were excluded, as these often describe animals from zoos or aquariums. Fossil specimens were similarly excluded. Occurrences flagged internally by GBIF for geospatial errors were also excluded from the search. After each dataset was downloaded, further filtering steps were performed. Coordinates with a reported uncertainty greater than 100km were removed, as well as those with a precision of less than one decimal place. Next, the R package *CoordinateCleaner*[177]

(v2.0.2) was employed to remove doubtful occurrences, such as those with equal latitude and longitude or which occur in country capitals, centroids, biodiversity institutions, or GBIF headquarters. Species names were reconciled with the Open Tree of Life reference taxonomy[178] (v3.3). Lastly, to avoid overrepresentation from specific sampling sites or sources, all coordinates were rounded to one decimal place and duplicate entries were removed. The final filtered dataset encompasses 32,880,023 occurrences.

Phylogenies

All major analyses reported in this study were performed using phylogenetic comparative methods. Phylogenetic trees were downloaded from previously published studies for each of the three focal clades. For Amphibia the tree of Jetz & Pyron (2018)[179] was selected, the reference amphibian phylogeny used by the VertLife project. The Actinopterygii tree was taken from Rabosky et al. (2018)[180] and is also programmatically available through the R package *fishtree*. In insects the tree of Chesters (2017)[181] was used, a species level phylogeny generated through a stepwise, hierarchical approach using multiple supermatrices. To ensure species labels were analogous with the GBIF occurrence data tip labels for each tree were also reconciled with Open Tree of Life reference taxonomy[178]. If species became synonymized after reconciliation one was retained pseudorandomly, with precedence given to those that shared a genus with their sisters, and the rest were pruned from the tree.

Ploidy Inference

Ploidy estimates were derived from literature, primarily from review articles or published databases[13,14,117,182,183]. When two sources disagreed the most recent was used. Species labelled "polyploid" for the purposes of this study are those in which polyploidy is known to commonly occur in natural populations, though not necessarily exclusively. In two special cases within the amphibian genera *Ambystoma* and *Pelophylax,* polyploid kleptons are maintained through hybridization. Due to the phenotypic similarity and range overlap within these complexes, as well as the challenges hybrids pose to standard phylogenetic methods, parent species are considered polyploid as well. I acknowledge this working definition likely includes many diploid occurrences; however, it accomplishes the primary goal of this study by encapsulating areas and environments where polyploid populations are established. This definition does not include species where polyploidy is considered rare or spontaneous or where polyploids are known only from experimental manipulations. Species without any ploidy data available are generally

presumed diploid, with the exception of a few genera and families where the majority of evaluated taxa are known polyploids.

Environmental Variables

Global rasters of environmental variables were drawn from a variety of publicly available sources. It has often been suggested that polyploids have greater ecological tolerances to extreme or challenging environmental conditions, such as the colder, drier climes associated with high latitudes. To explore this idea I collected data on five climate and environmental variables (mean annual temperature, temperature seasonality, annual precipitation, precipitation seasonality, and altitude) from the WorldClim 2.1[121] database at 5 minute resolution. Temperature variables are of particular interest, as extreme temperatures may be associated with a greater incidence of polyploidy not due to broader ecological tolerances, but because cold or heat shocks simply create more polyploid gametes. To further test this idea, I also calculated how climate variables have changed since the last glacial maximum (LDG), as estimated by the Community Climate System Model 4[184], at the same resolution. Similarly, I also collected data on estimated glacial cover at the LDG[185]. Estimates of glacial cover were taken from Hughes et al. (2011)[185]. Glacier shape files were rasterized, with glaciated cells receiving a value of 1, unglaciated cells receiving a value of 0, and cells located in the ocean receiving a value of NA. It has been suggested that polyploids are associated with new environments not due to their increased adaptive potential but due to the absence of other species, as interspecific competition and cross-ploidy matings are predicted to adversely impact polyploid populations[14,81,85,105]. For estimates of species richness I summed global rasters for amphibians, mammals, and birds species provided by BiodiversityMapping.org[186]. A limitation of this approach is the absence of any non-vertebrate clades, whose patterns of global richness may differ from vertebrates[187]. However, I am not aware of any global estimates of species richness in non-vertebrates that would be appropriate for this study at the time of writing. I also fail to include any marine clades, as including estimates of exclusively terrestrial or marine clades together within the same raster would introduce considerable bias. As all polyploid ray-finned fish identified by this study are freshwater, I elected to use the same terrestrial richness data for all three focal clades in order to better allow comparisons across them. While this dataset may not completely reflect true biodiversity it does capture the major global patterns (such as the latitudinal diversity gradient) of interest to this study. Finally, polyploids' expected ecological tolerance has been predicted to

extend to areas that have been transformed by human activities as well, with polyploids more resilient to anthropogenic impacts[158,159,188]. For this hypothesis I utilized the human footprint index provided by Sanderson et al.[189], as well as agricultural land use datasets from NASA's Socioeconomic Data and Applications Center[137]. All occurrence coordinates and rasters presented and discussed in this study us the WGS 84 reference system, however visualizations are provided as Robinson projections.

Hypothesis Testing

Values from each of the 14 environmental variables were paired with each species occurrence using the R package *raster*[190] (v3.5.9), and averaged across species. To get relative measures of the environmental differences between polyploids and diploids that are comparable across variables and clades, Cohen's *d* was calculated for all relevant comparisons using the R package *effsize*[191] (v0.8.1). Differences took the form: $(x_{polyploid} - x_{diploid})$, such that positive values indicate that the variable is higher for polyploids than for diploids, and negative values represent the inverse. A species' ploidy level and geographic range are expected to carry strong phylogenetic signal, and as a result efforts must be taken to distinguish meaningful relationships from covariance due to shared evolutionary history[192,193]. To achieve this, phylogenetic ANOVAs were performed[139], under which ANOVA test statistics are compared to a null distribution simulated from a Brownian-motion model. Additionally, phylogenetic ANOVAs were performed for each variable using the R package *geiger*[140] (v2.0.9), and tested against a null distribution generated from 1,000 phylogenetic simulations. As polyploidy is rare in animals, there is a concern that unbalanced sample sizes violate the homogenous variance assumption of ANOVA tests. To ensure tests were still able to distinguish between hypotheses, additional simulations were performed under which polyploid labels were assigned randomly while maintaining the same sample size. Under this null dataset no variable was found significant in any clade. Similarly, since the ranges for the majority of diploid species do not overlap with any deglaciated areas, the glaciation dataset is heavily zero-weighted and so violates the normality assumption. To address this, I also performed chi-squared tests between ploidy and glaciation where each species was labelled 'glaciated' (>=50% of occurrences within deglaciated area) or 'not glaciated' (<50% of occurrences within deglaciated area) which were significant ($p<0.05$) in each clade. After ANOVA testing, relative importance analysis was performed. For each clade, a series of phylogenetic multiplicative MANOVA models were performed. A model was

constructed for every combination of variables that were found significant under the ANOVA tests. AIC values were then compared within each clade to determine the best fitting models.

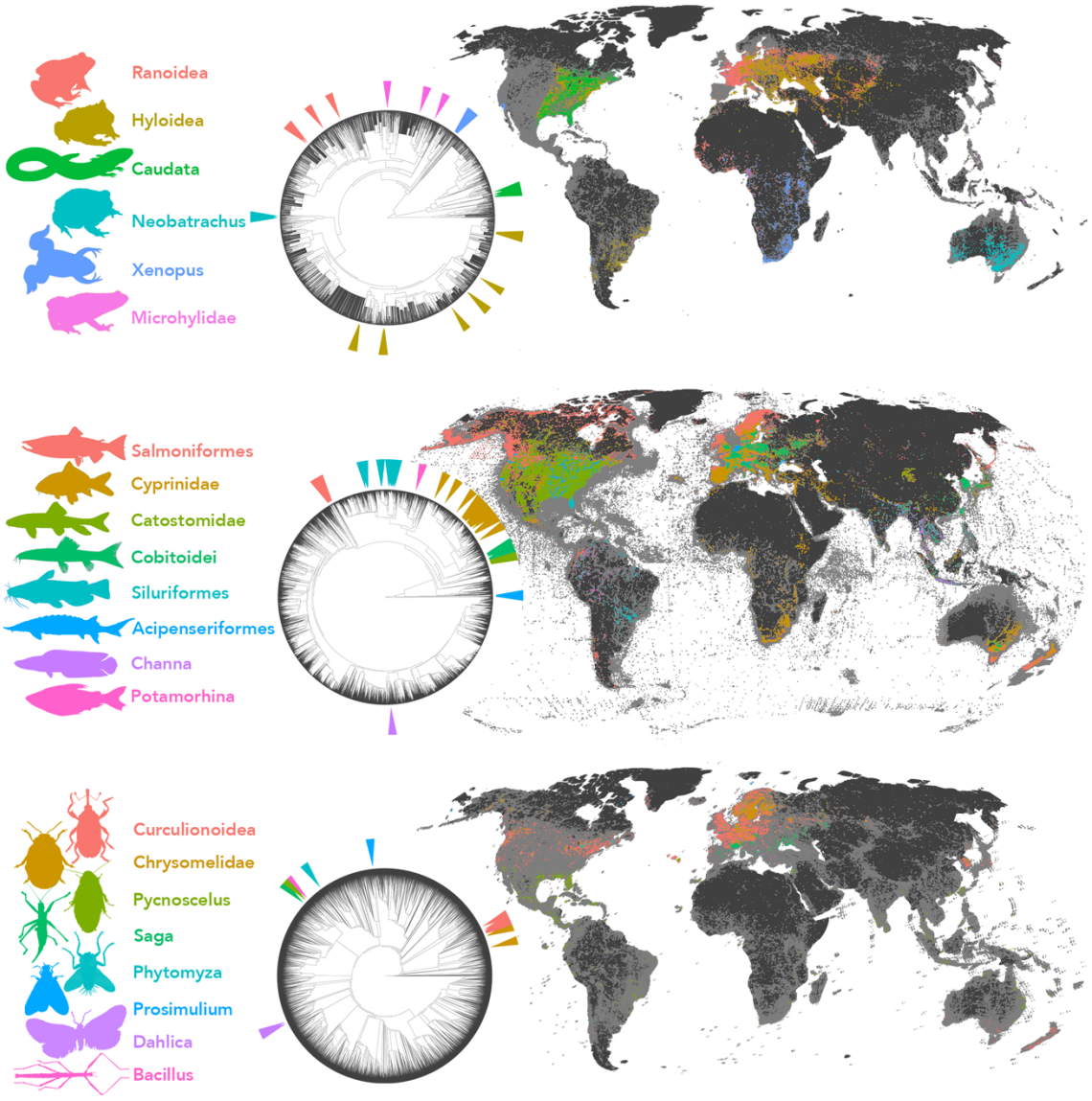

Figure 4.1. Phylogenetic and geographic sampling of this study. Polyploids are highlighted and colored by clade. Grey dots and unlabeled tips represent diploids.

Results & Discussion

Herein I present a global dataset of diploid and polyploid species' occurrences in three major animal clades. Each occurrence is associated with a variety of metadata describing the environment, climate, anthropogenic influence, and history of the locality. I have used these variables to explore potential drivers of polyploid occurrence and distribution, testing several

longstanding hypotheses. Additionally, previously published phylogenies for each clade were employed to ensure results were not the result of shared evolutionary history[193]. The assembled data encompasses 57,905 species including 471 polyploids across 2,223 terrestrial, freshwater, and marine ecoregions (Fig. 4.1, Fig. 4.2).

A Latitudinal Polyploid Gradient in Animals

My results reveal the clear existence of a latitudinal polyploid gradient in animals. Of the 471 polyploid species identified in this study only 19.5% have a mean latitudinal range between the Tropics of Cancer and Capricorn. This disparity is most evident in hexapods where 97% of polyploid species occur outside of the tropics but is also apparent in amphibians and ray-finned fishes where the frequency of temperate species is far greater in polyploids than diploids (50.9% vs. 21.0% in Amphibia and 78.7% vs. 42.2% in Actinopterygii) (Fig. 4.2, Fig. 4.3). Additionally, absolute latitude was a significant ($p<0.05$) variable in phylogenetic ANOVAs across all three clades (Fig. 4.4).

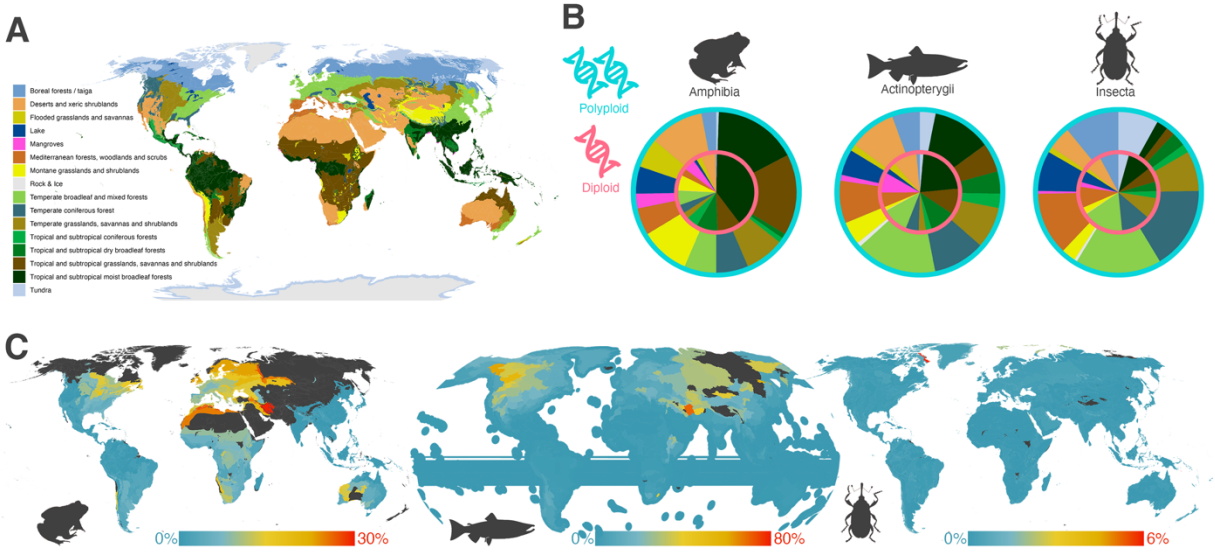

Figure 4.2. A) Biomes of the world as classified by the world wildlife foundation. B) Pie charts comparing the representation of biomes for polyploid (blue) and diploid (pink) species across clades. C) Each ecoregion (terrestrial for Amphibia and Insecta, freshwater and marine for Actinopterygii) with at least 10 species, colored by proportion of polyploid species.

Analysis of the world's biomes reveals similar patterns. Temperate and polar biomes including tundra, taiga, temperate broadleaf, and temperate coniferous forests all had greater relative frequencies of polyploids than diploids in each clade. By contrast, biomes such as tropical and subtropical dry or moist broadleaf forests and mangroves had reduced frequencies

(Fig. 4.2). For example, diploid amphibian species occur in tropical and subtropical dry broadleaf forests at a frequency 9 times higher than that for polyploid species. At finer scale resolution, on the level of ecoregions, fewer similarities exist between clades however the general pattern remains. Ecoregions within the greatest 1% of polyploid frequency include the Scandinavian and Russian taiga for amphibians, upper Yukon for ray-finned fish, and Arctic desert for hexapods. As with tropical biomes, species-rich ecoregions have relatively few polyploids. In amphibians, the three ecoregions with the highest species richness (all Andean montane forests) have no recorded polyploid occurrences whatsoever.

Environmental Conditions and Changing Climate Dynamics Differ Significantly between Diploids and Polyploids

Geographic patterns of animal polyploids reveal a tendency for polyploidy incidence to increase with latitude. There are several hypotheses which may explain this trend, as many variables that may contribute to conditions that are favorable for the formation and/or survival of polyploid lineages also covary with latitude. To assess and distinguish between these hypotheses, I collected environmental data from each occurrence in the dataset, averaged across species. Comparisons between diploid and polyploid species revealed significant trends shared across clades. Of the 14 potential explanatory variables, 13 were found to be significant ($p<0.05$ following Bonferroni correction) in at least one clade and 6 were significant in all 3 clades (Fig. 4.4). These six include all variables pertaining to temperature as well as glaciation and absolute latitude. Temperature seasonality had the largest absolute effect size across clades, followed by change in temperature and mean annual temperature (Fig. 4.3). Polyploids generally appear to occupy areas that are colder and have greater annual temperature fluctuations than diploids. This result suggests a selection-independent driver of polyploidy. Cold temperatures are known to contribute to unreduced gamete formation, which can lead to autotetraploidy. Indeed, cold shocks are an established method to produce artificial autotetraploids in aquaculture and has been demonstrated in ray-finned fish, frogs, and branchiopods[98–100,170]. However, polyploids are also particularly common in those areas which have undergone large temperature shifts since the LDG (positive for mean annual temperature, negative for temperature seasonality), which suggests an adaptive response. In addition to the adaptive advantages outlined above, polyploids may be particularly well suited to cold climates. The effects of polyploidy are varied, however one consequence that appears relatively consistent in animals is an increase in cell and/or body

size, both of which are predicted to be advantageous in colder environments[194,195].

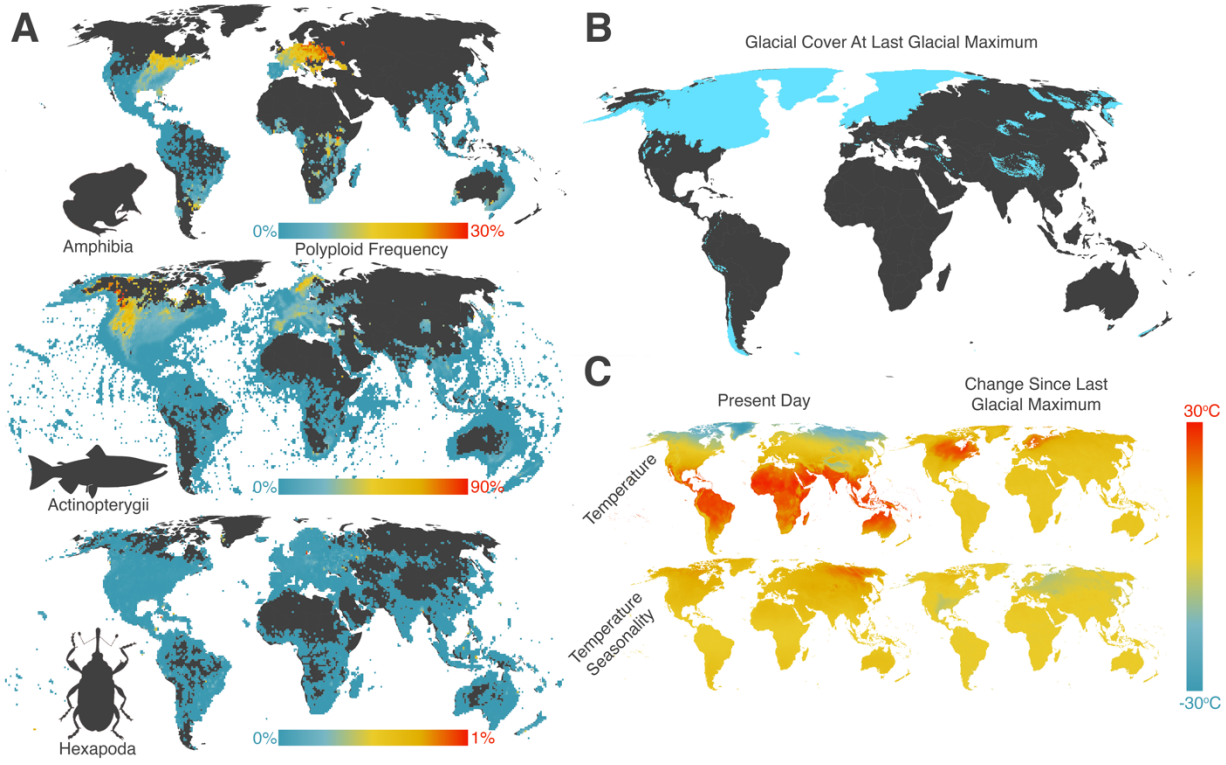

Figure 4.3. A) Global representation of polyploid species in each major clade, excluding cells with fewer than 10 total species. B) Estimated glacial cover at LGM. C) Mean annual temperature and temperature seasonality (standard deviation), estimated for modern day as well as Δ from LGM.

Overall, temperature dynamics and changing climates appear to be the most dramatic differences between polyploid and diploid distributions. However, my analysis also recovered several significant clade-specific effects that are worth discussing. Precipitation was a significant factor shared between amphibians and ray-finned fish, with polyploid species more common in drier environments. This trend is most evident in amphibians, where polyploid species consistently occur in drier conditions than diploids within their genus, including *Neobatrachus* in Australia, *Phyllomedusa* in South America, *Xenopus* in Africa, *Dryophytes* in North America, and *Pelophylax* throughout Europe and parts of Asia. It has previously been noted that polyploidy facilitates niche shifts into drier climates in *Neobatrachus*, and coupled with the result here suggests that polyploidy is a common mechanism for arid climate adaptation in frogs[196]. Actinopterygian polyploids were similarly linked to areas of reduced precipitation. Precipitation is known to covary with latitude, so it is possible this observed trend is simply a correlation artifact. However, there are specific arid regions where both amphibian and

actinopterygian polyploids appear with disproportionate frequency. In particular, xeric shrublands, grasslands, and steppes throughout Western, Central, and Eastern Asia exhibit some of the highest relative polyploid richness. Of the seven aquatic ecoregions with majority actinopterygian polyploid species two occur in Iran and one encompasses Western Mongolia. In Amphibia, ecoregions such as the Persian Gulf desert, East Afghan montane conifer forests, and three separate ecoregions stretching across the northern Himalayan foothills have only polyploid species. Unfortunately sampling in this area is extremely sparse, often with just one species identified per ecoregion, so my ability to make conclusions about this pattern is limited. In a related patten, ray-finned fish polyploids are also significantly associated with high-altitude environments. The now polyphyletic schizothoracine fishes (known as "mountain carps") are known polyploids[182] distributed throughout the Tibetan plateau, and show molecular signatures of high altitude adaptation[197]. Additionally, the majority-polyploid Salmonidae are common throughout mountainous regions in northwest North America. As with aridity in amphibians, the repeated association of high-altitude environments and polyploidy in Actinopterygii across clades and continents may indicate a common adaptive pathway. Finally, species richness is often cited as having a negative relationship with polyploid incidence due to competition from diploids and cross-ploidy mating. I found limited evidence to support this idea, species richness was not significant between amphibian diploids and polyploids. There was a significant negative association in Actinopterygii, albeit with relatively low Cohen's *d*. However, large and significant differences were found in Insecta. One possible element that may help explain this discrepancy is the relationship between parthenogenesis and polyploidy. Polyploidy is tightly linked with thelytokous parthenogenesis in several animal clades including insects, but not amphibians or ray-finned fish[13]. Parthenogenic organisms are also predicted to inhabit species-

poor areas, for similar reasons as polyploids[198].

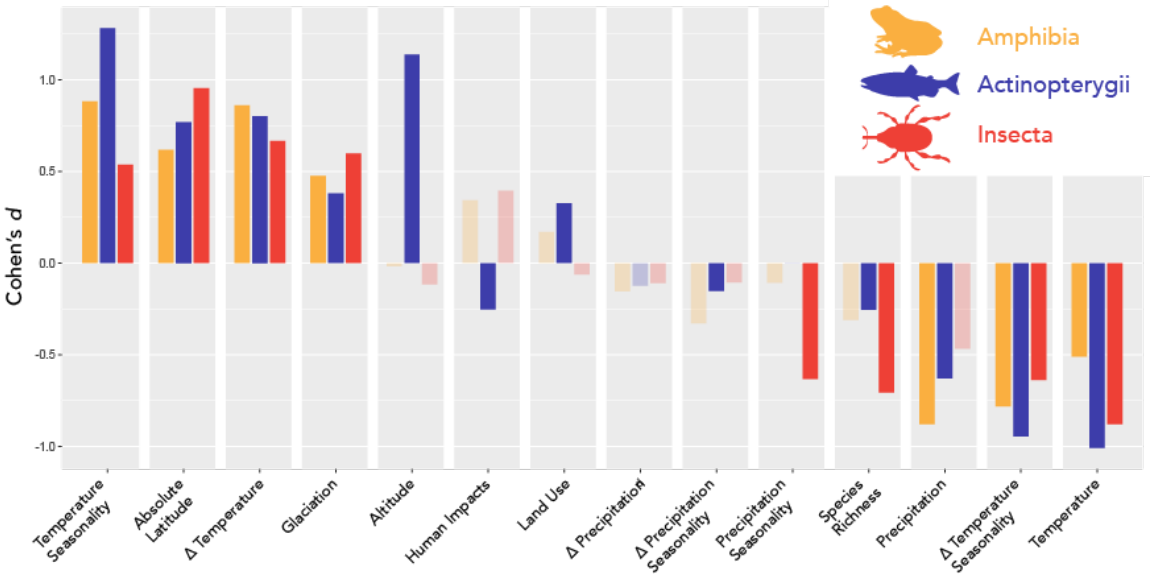

Figure 4.4. Cohen's *d* between polyploids and diploids in each major clade for every environmental variable considered by this study. Positive values indicate that the variable is higher for polyploids than for diploids, and negative values represent the inverse. Transparent bars indicate that the difference was not significant for phylogenetic ANOVAs following Bonferroni correction ($p<0.05$).

Glaciation Cycles Drive Polyploidy in Animals

Comparisons between polyploid and diploid species revealed several large, significant differences between environmental variables across clades. To see on which of these variables ploidy had the largest impact I performed a relative importance analysis. Briefly, a series of multiplicative phylogenetic MANOVAs[139] were constructed from every combination of each significant variable for each clade. This resulted in 127 models for Amphibia, 4,095 models for Actinopterygii, and 255 models for Insecta. Within each clade AICs for each model were then compared to explore which variables contribute most to the differences between polyploids and diploids. Models that only included glaciation were the best fit in Amphibia and Insecta, whereas a multiplicative model including glaciation and land use was the best fit for the Actinopterygii dataset (Table 4.1). Each of these models far outperformed any other, with Akaike weights[199] approaching 1. The glaciation variable also has the largest relative importance in each clade.

Across clades, glaciation best explains the discrepancy between the polypoid and diploid environment. Polyploids are more likely to be found in areas that were glaciated at the LGM. This finding is supported by the effect size analysis, and also serves to explain the latitudinal polyploid gradient. This result contributes to a model of evolution under which recently de-

glaciated areas provide environments that are conducive to polyploidy. The expanded genomic toolkit of polyploids may enable them to more rapidly adapt to new environments, as has been shown recently in frogs[196] and switchgrass[200]. Migrating into new environments also allows polyploids to "escape" the range of their diploid counterparts, allowing them to avoid cross-ploidy competition and mating which is predicted to favor diploids. While de-glaciation is the largest and most obvious mechanism through which species are exposed to new environments, human introduction is another example. Invasiveness is commonly associated with plants, polyploid plants are 20% more likely to be invasive than diploids[115]. I recovered this finding in part; polyploids had more invasive species than diploids in all three clades though chi-squared tests were not significant ($p>0.05$) in insects, perhaps due to the low sample size of polyploid invasives (1 species). Anthropogenic transformations such as land clearance and fertilizer application might similarly represent "new" environments more favorable to polyploids. In Actinopterygii polyploids were more common than diploids in areas used for agriculture, and the agricultural land use variable appeared alongside glaciation in the highest ranking model. However more work is required to disentangle this possible effect from potential covariates.

Table 4.1. Top five highest ranking phylogenetic M/ANOVA models in each clade, sorted by ΔAIC

Amphibia	Actinopterygii	Hexapoda
Glaciation **(0)**	Glaciation+Land Use **(0)**	Glaciation **(0)**
ΔTemperature **(31329)**	Glaciation **(9920)**	ΔTemperature **(250507)**
Temperature **(40716)**	Land Use **(17108)**	Temperature **(256386)**
Glaciation+ΔTemperature **(41557)**	ΔTemperature **(71181)**	Absolute Latitude **(320606)**
Absolute Latitude **(49092)**	Temperature **(83424)**	Precipitation Seasonality **(350120)**

Conclusions

There are several elements not explored by this study that may be of interest to future research. First, this work does not distinguish between various levels of polyploid. Triploidy, tetraploidy, hexaploidy, etc. were all treated equally for the purposes of this study. In theory octoploids are as different from tetraploids as tetraploids are from diploids, and future work may benefit from distinguishing between these cases explicitly. Similarly, I make no attempt to classify polyploids as either allopolyploids or autopolyploids, as evidence for either case is missing or contradictory in the majority of species. Glaciation may be particularly relevant in the case of allopolyploids,

as deglaciated areas represent potential secondary contact zones for the formation of polyploid hybrids. However, such a scenario would also need to explain the presence of the many autopolyploid species also associated with recently deglaciated areas observed by this study.

A strong latitudinal polyploid gradient was observed in each of the three major clades investigated by this study, with frequency of polyploid species increasing with distance from the equator. This trend was apparent across clades at the scale of climate zones, biomes, and ecoregions. This pattern is no doubt the result of many processes and may indicate adaptation to cold environments, dry environments in the case of amphibians, or high-altitude environments in the case of ray-finned fish. However, relative importance analysis indicates that differences in ploidy level are best explained by the extent to which species' ranges were glaciated in the LGM. This finding supports the idea that polyploids are more adaptable to new environments brought about through rapid change, which may also apply to introduced species and environments transformed through anthropogenic intervention. Furthermore, evidence for the increased adaptability and environmental plasticity of polyploids presented here and elsewhere may also serve to explain why polyploidy often precedes diversification events, evolutionary novelties, and persistence in the face of mass extinction.

Chapter V. Sequencing Disparity in the Genomic Era

Who to Sequence?

The use of model organisms, such as maize (*Zea mays*), the mouse (*Mus musculus*), and the fruit fly (*Drosophila melanogaster*), have contributed greatly to our understanding of biology via their tractability and large research communities[201]. Thus, when whole genome sequencing came of age, focusing on organisms that have been widely used as models was sensible. With cost-effective, high-throughput sequencing, however, many barriers that limited the use of non-model organisms have been removed. Advances in non-model and reduced representation genome sampling approaches have enabled researchers to sequence virtually any organism more cheaply and easily than ever before[202]. These advances enable comparative studies with broad sampling from across the tree of life that can elucidate the origins and variation in cellular mechanisms thereby ushering in a new era of discovery[203]. In light of this new approach, many researchers have predicted that the lines between model and non-model organisms will blur or disappear entirely[201,202,204,205]

Trends in High-Throughput Sequencing Biodiversity

To explore how high-throughput sequencing efforts are distributed across the diversity of eukaryotic life, we accessed all nonhuman eukaryotic sequencing experiments in the Sequence Read Archive (SRA) using the Entrez Direct suite of UNIX commands. The SRA is a high-throughput sequence database administered by the DNA Data Bank of Japan, European Nucleotide Archive, and the National Center for Biotechnology Information (NCBI). Note that experiments are defined in the SRA as "a unique sequencing result for a specific sample" and can be from experimental or descriptive research. Experiments may use one of many different sequencing strategies, though RNA-Seq (37.6% of experiments) and whole genome sequencing (22.6% of experiments) are the most common. The search was executed on 20 January 2019 and returned 1,874,638 experiments. Of those, 29,578 (1.6%) experiments were removed either because they had pooled samples from multiple species or missing data. Experiments were restricted from those published between 2010, the first year with a sufficient (>100) number of species for our analysis, and December 2018, the most recent month with complete data at the time of the search. The final dataset includes 1,808,136 high-throughput sequencing experiments from 24,288 unique species.

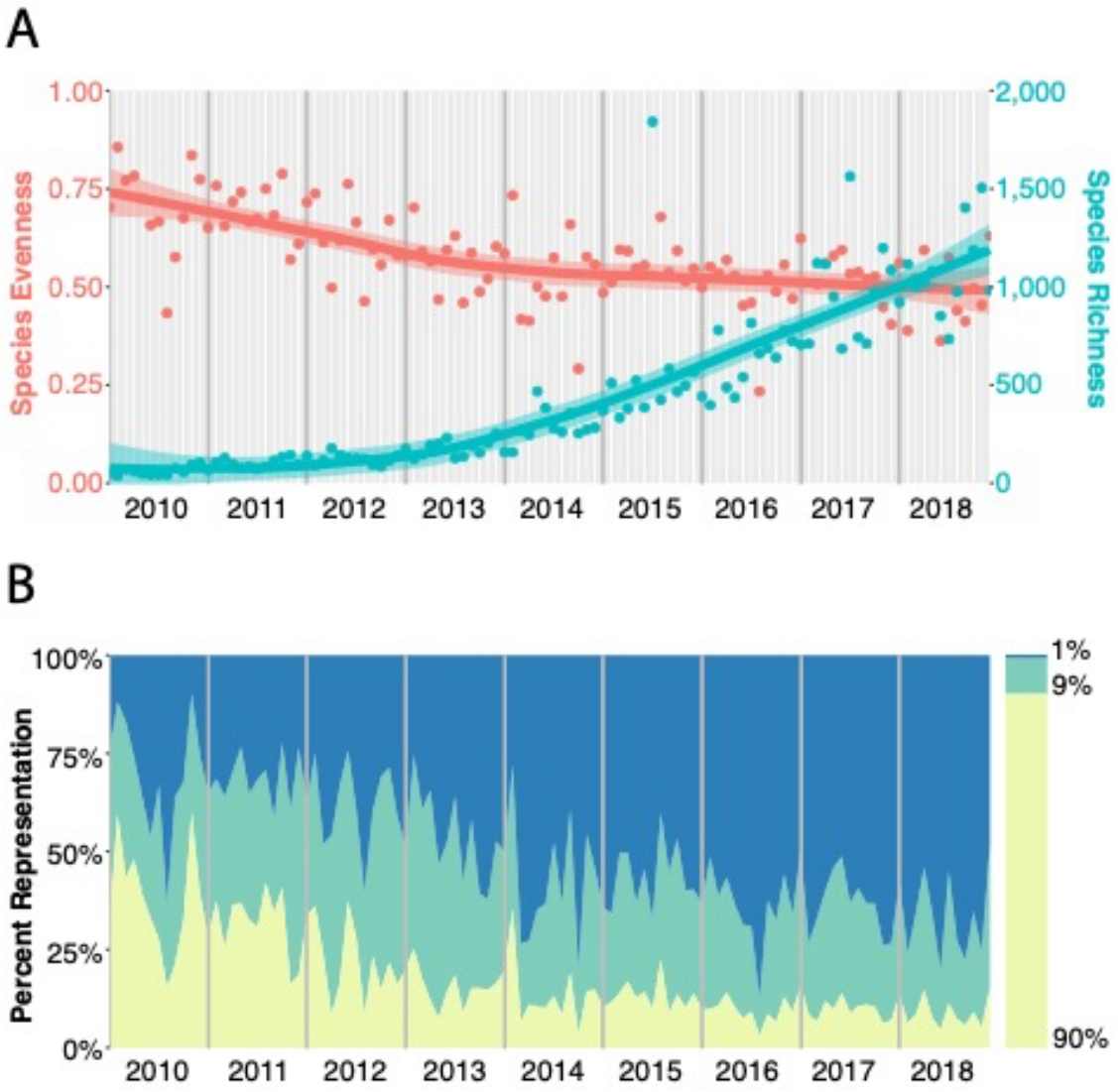

Figure 5.1. A) Pielou's species evenness (H/ln(S) where H is Shannon's diversity index and S is species richness) and species richness of all nonhuman eukaryotic sequencing experiments in the SRA calculated for each month between January 2010 and December 2018. Richness has increased over time ($p<1.1E-32$) while evenness has decreased ($p<2.5E-13$). B) Relative representation of the top 1%, top 2-10%, and bottom 90% of species in the SRA by number of experiments for each month between January 2010 and July 2018.

The top 1% of species with the most experiments represent 85.3% of all experiments accessed. The top 1% includes 120 animals, 99 plants, 14 fungi, and 9 protists. Fifteen phyla are represented, although 83.1% of species are either streptophytes (green plants; n=98), chordates (n=73), or arthropods (n=30). At the species level, the mouse *Mus musculus* is the most represented by a wide margin with 523,192 experiments, 4.4x more than any other species, and representing 28.9% of all experiments. All 13 model organisms officially recognized by the

National Institute of Health (NIH) are featured in the top 1%, which together represent 44.7% of accessed experiments.

As expected, species richness has increased over time from 453 unique species sequenced in 2010 to 9,696 in 2018. However, despite an increase in the number of unique species sequenced over time, species evenness has decreased significantly (Fig. 5.1A; $p<2.5E-13$). This trend appears to be mediated at least in part by a growing preference toward relatively fewer study species. The top 1% represented 49.7% of experiments in 2010; however, in 2017 the top 1% represented 80.4%. The top 1% of species for each month has been increasing at a rate of about 5.0% $year^{-1}$ (Fig. 5.1B; $p<2.2E-16$).

Of the 24,288 species we accessed, 1,146 have significant increases in the number of experiments over time. Given that high-throughput sequencing has increased exponentially, seeing large increases in number of experiments over time for highly studied organisms is not surprising. Indeed, with the exception of *Plasmodium falciparum,* the top five species with the greatest increases were all NIH recognized models (in descending order: *M. musculus*, *Saccharomyces cerevisiae*, *P. falciparum*, *Arabidopsis thalina*, and *D. melanogaster*) (Fig. 5.2).

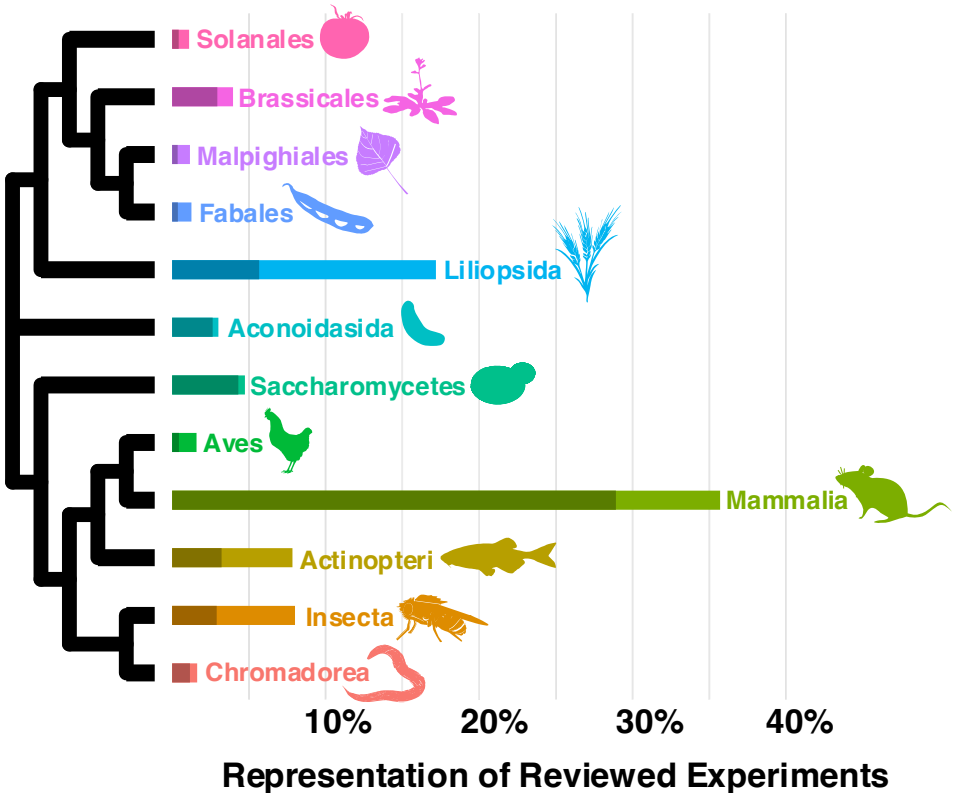

Figure 5.2. Relative representation of all clades (classes or orders if class does not exist) with more than 1% of total experiments in the dataset. Together these clades represent 87% of all reviewed experiments. Dark bars represent relative representation of the most studied species in each clade (pictured).

We also explored change in relative

frequency of experiments for each organism over time. In terms of relative frequency, the only NIH model that maintained a significant increase over time was the mouse. The two species with the largest decreases in relative frequency were *D. melanogaster* and *Caenorhabditis elegans*. While longstanding models are becoming less dominant, other organisms, such as the olive baboon (*Papio anubis*) and mummichog (*Fundulus heteroclitus*), appear to be receiving more attention.

As anticipated, species richness in high-throughput sequencing experiments increased significantly over time (Fig. 5.1A; $p<1.1E-32$). This finding demonstrates a trend toward sequencing greater taxonomic diversity, driven by recent initiatives such as those undertaken by Genome 10K and the Global Invertebrate Genomics Alliance (GIGA). Notably, in November 2018, the Earth BioGenome Project was launched, which aims to sequence all eukaryotic species genomes in 10 years.

Although more unique species are being sequenced over time, the disparity of sequencing efforts is widening, suggesting more focus is being put on relatively fewer species. In particular, the mouse has 1.8x more experiments than all other NIH models combined, and the number of experiments is growing linearly at a rate of 150.8 per month, 4.3x faster than any other organism. Mouse is also the only NIH model whose relative representation has increased, taking up a larger proportion of total sequencing experiments over time (increasing at a rate 3.5x that of any other organism).

"All Models are Wrong"

Model organisms have been a fundamental aspect of biology for at least 150 years[201]; however, they are not without problems. As statistician George Box famously articulated, "All models are wrong, but some are useful." Many researchers have previously commented on pitfalls of model-centric research[204,205], which can result in disastrous and even lethal consequences as with the infamous fialuridine and thalidomide drug trials during the 20th century[206,207].

In particular, there are many questions for which mouse models are not well suited, in spite of their popularity. In recent years, the mouse has been shown to be an imperfect representative of human disorders, particularly with regard to neurological and immune disease as well as cancer, resulting in frustratingly few real-world applications relative to the investment[208–212]. Additionally, early approximations of human gene count based on homology with mouse were overestimated by ~60,000, a number which was reduced to ~10,000 by a study

that used the more compact genome of the pufferfish *Tetraodon nigroviridis*[213]. Current estimates of human gene count rely on comparisons across many different species.

Although the model organism philosophy is instrumental to mechanistically understand intraspecific biological processes, it is singularly inappropriate to address these questions in an evolutionary context and therefore not capable of answering questions on the origins and variation of such mechanisms. Broad taxonomic sampling is a prerequisite to assess evolutionary processes and patterns. For example, a centralized nervous system was thought to be an ancestral character of bilaterians with a single evolutionary origin based on similar expression patterns in *D. melanogaster*, the annelid worm *Platynereis dumerilii*, and vertebrates. However a study in 2018, based on novel sequence data from additional groups within *Bilateria*, found different nervous-system architectures even between closely related taxa, suggesting nerve cords evolved within *Bilateria* multiple times[214]. Similar patterns remain to be explored for many other biological characters, such as gastrulation and segmentation[215].

The Sangerian Shortfall

Just as the Linnaean Shortfall describes how few species have been formally described, we define the Sangerian Shortfall as the lack of knowledge regarding most species' genomes. The 24,288 species represented in the SRA represent 2.0% of the 1,186,221 described eukaryotic species in NCBI's taxonomy database and only 0.0027% of the 8.7 million eukaryotic species thought to exist on earth. For species with whole genome sequences available, representation is even lower, 9,613 species as of December 2018. Of the 14,927 species currently listed as endangered or critically endangered by the IUCN, only 2.6% had high-throughput sequence data.

Concluding Remarks

Advances in high-throughput sequencing technology have enabled researchers to sample from more species than ever before. In spite of this, genomic sampling is becoming more model-focused as relatively more attention is being paid to fewer species. Negative effects of this trend extend beyond biodiversity and evolution studies to many fields including pharmacology, development, genetics, and neurobiology. To improve taxonomic diversity in genomic studies we recommend 1) developing and improving funding resources to increase our understanding of biodiversity (such as the Dimensions of Biodiversity and PurSUit programs offered by the U.S. National Science Foundation) and 2) taking steps to ensure reviewer panels are represented by

researchers working on a variety of study systems to reduce bias in funding decisions. We predict that these recommendations, if acted upon, will not only improve our understanding of variation and diversity in nature but will also foster collaborations across different research fields and systems. We conclude that molecular researchers should attempt to select models based on their relevance to biological questions over ease of use and use comparative approaches to address questions in an evolutionary framework whenever possible.

All code required to update experiments and reproduce results/figures are available at https://github.com/KyleTDavid/SRA2019. Original data files are available at https://figshare.com/projects/SRA2019/39296.

Chapter VI. Unsupervised Deep Learning Can Identify Protein Functional Groups From Unaligned Sequences

Introduction

As sequencing technology continues to improve, there is an ever-increasing need to adequately annotate and characterize novel protein sequences and their predicted functions. With thousands of new sequences being uploaded every day, predicting the function of every protein directly with conventional experimental studies such as gene knockouts or assays is not possible. Thus, it becomes necessary to attempt to infer protein function automatically from the sequence alone. Many such methods exist but fundamentally operate the same way: by matching the sequence of a protein with unknown function to a reference sequence of a protein with known function, and then assuming that the functions are the same. These matches are generally identified by sequence similarity, as inferred through methods such as the BLAST algorithm[216], closely related homology, through hidden Markov models such as HMMER[217], or by orthology using tree inference methods such as OrthoFinder[218,219]. Of the 219,740,215 proteins listed in the UniProtKB database[47,220] (release 2021_03) only 0.77% had experimental evidence, compared to 31% with homology-based evidence. This reference-based approach is clearly a powerful tool to assign putative identity to unknown protein sequences. However, a consequence of this approach is that our assessment of protein function and diversity is restricted to only those sequences that with sufficient pairwise site similarity with reference sequences which have been experimentally validated in model taxa. For example, in the Gene Ontology database[221,222] (release 2021-07) 73% of annotations with experimental evidence belong to just five species[223]. Three of these five are human, rat, and mouse, which suggests that unless a protein is similar to one found Euarchontoglirine mammals, accurate annotation may be difficult. Reference-based approaches may cause researchers to miss or underestimate protein functional diversity, particularly in non-model taxa. For example, clades like cnidarians which are more phylogenetically distant from model species, have larger “dark” regions of the proteome about which little is known (Fig. 6.1). This trend can create bias, leading researchers to conflate organismal complexity with research attention[203,224].

One obvious solution would be to expand the taxonomic scope of our experimental investigations and explore new model systems which better reflect the diversity of life. However,

the field of molecular biology appears to be moving in the opposite direction, with more sequencing effort being disproportionately funneled into fewer taxa each year[225]. We also acknowledge that developing new model systems is resource intensive and intractable for many organisms. However, a more complete understanding of molecular function and diversity across the tree of life is of paramount importance to many questions throughout the field of biology. Failure to account for molecular diversity of non-model organisms will lead to a systemic underestimation of gene family size in these groups[226], which can compromise functional studies such as those dependent on reporter constructs[227] or CRISPR-Cas9[228]. Phylogenetically-informed gene sampling is also necessary for assessing genotype-to-phenotype relationships, as assessing most phenotypic diversity is impossible with classical genetics[229]. Finally, broad sampling is a prerequisite to address evolutionary questions that are comparative by nature. For example, a 2018 study drawing on data from across Bilateria identified multiple independent origins of nerve cords within the group, upending the previous consensus of a single origin based on just a few taxa[214]. As non-model organisms become increasingly relevant in answering biological questions, there is a clear need to develop *ab initio* methods for characterizing novel protein sequences and their functions.

One possible avenue for exploring protein function without directly relying on model systems is through unsupervised deep learning. Deep learning refers to a class of machine learning methods characterized by multilayer neural networks designed to identify progressively complex patterns from large datasets[230]. Unsupervised deep learning (in contrast to supervised learning) does not learn from externally-provided examples or labels, but rather attempts to learn patterns inherent to the data in order to construct a generalized model[231]. Several supervised deep learning methods are already being developed for protein classification which may outperform conventional techniques[232–235]. However, as these supervised methods are reliant on high confidence training examples, they are not able to escape the fundamental biases associated with reference-based methods.

One common implementation of unsupervised deep learning is through the use of variational autoencoders[236] (VAEs). A VAE consists of two neural networks, an encoder and a decoder. The encoder q compresses the input data x into a latent representation z from parameters θ: $q_\theta(z|x)$. The decoder ρ then samples the latent representation and attempts a reconstruction of data x' from parameters ϕ: $\rho_\phi(x'|z)$. The loss function can be approximated as $x' \cong \rho_\phi(q_\theta(x))$. The

VAE tries to optimize θ and ϕ to produce the highest fidelity reconstruction of x after it has been mapped to the lower-dimensional z. Once trained, the learned distribution can then be sampled to generate its own data. For example, a VAE trained on images of objects[237], animals[238], or faces[238] may be able to generate realistic original images of its own. In biology, one recent study used a VAE to simulate realistic human genotypes[239]. For the purposes of this study, however, we are less interested in the generative power of these models than the latent representation itself. The VAE is forced to map input efficiently into a reduced latent representation with the expectation that it will learn key, identifiable features from training data and distribute latent variables in a way that is meaningful to researchers, (e.g., by placing proteins of the same family or with the same function close together). In this way a VAE can be thought of as an analog to principal component analysis (PCA), although VAEs allow for nonlinear relationships which may enable them to capture greater variation in data[240–242].

VAEs have become increasingly popular in recent years[243], and have been developed in biology for problems such as predicting the effects of mutations[241], visualizing population structure[240], and species delimitation[244]. These and most other VAEs use a multivariate normal distribution to represent latent space, with the output of the encoder providing the means and standard deviations to parameterize the distribution. However, in theory almost any type of distribution should work. One variant that has recently shown promising results, known as a vector quantized variational autoencoder (VQ-VAE), uses a discrete distribution[245]. The intuition behind such an approach posits that a discrete distribution may be more appropriate for categorical data. For example, with image datasets like CIFAR-10 each discrete latent variable could ideally represent one of the images classes (dog, frog, horse, etc.).

To explore the applications unsupervised deep learning may have for molecular biology we created DeepSeqProt, a VQ-VAE which accepts unaligned protein sequence data as input. We wrote DeepSeqProt in part to explore how the latent embeddings produced by VQ-VAEs reflect protein functional diversity as it is currently understood (e.g. canonical protein families). We are further interested in the utility unsupervised deep learning may have for protein

clustering and annotation, and whether it offers any advantages over conventional reference-based methods, especially with regard to function.

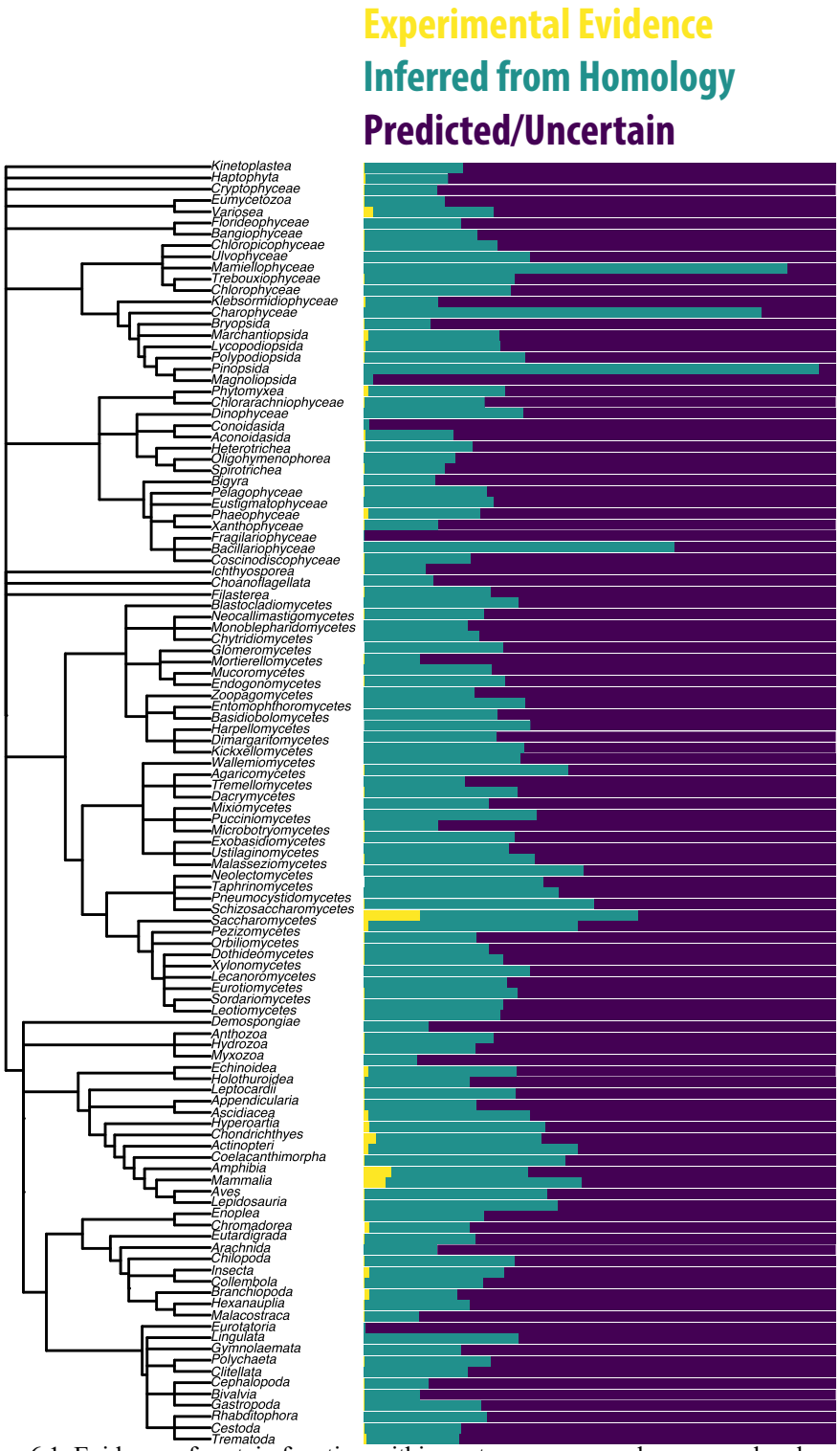

Figure 6.1. Evidence of protein function within proteomes summed across each eukaryotic class in UniProt (release 2021_03).

Methods

To investigate the viability of unsupervised deep learning for protein annotation we wrote a VQ-VAE[245] program, called DeepSeqProt, whose general architecture is detailed below (Fig. 6.2). Our primary motivation was to develop a simple, but useful, deep learning framework for unaligned, unannotated, protein sequence data. To this end we limited hyperparameters wherever possible while also allowing flexibility on behalf of the user. DeepSeqProt is written in Python3 and uses the

PyTorch[246] machine learning library. More details are available at DeepSeqProt's GitHub repository (https://github.com/KyleTDavid/DeepSeqProt). DeepSeqProt can also be run interactively in the public Google Colaboratory Notebook available at https://colab.research.google.com/drive/1FcOCECzhUg35PcXjb-LfdsSYDQoZurV3?usp=sharing.

Data Preprocessing

Before applying neural networks to categorical amino acid data, sequences must first be converted into numeric form. DeepSeqProt accomplishes this through one-hot encoding, where each of the 20 canonical amino acids is translated to a unique binary array 20 elements long. Alanine, for example, is represented as [1, 0…0, 0] and Valine is represented as [0, 0…0, 1]. Any character not associated with one of the canonical amino acids (e.g., unknowns, ambiguity codes, nonstandard amino acids) is represented by an array of all zeroes. One necessary restriction of deep learning is that most operations require the input data to have the same shape, which presents a problem for amino acid sequence data which can be of variable length. Some methods overcome this limitation either by padding sequences with additional zeroes up to a standard length[233] (while discarding or truncating longer sequences), and/or by grouping sequences of the same/similar length together[13,16]. However, in the latter case we worry that artificially ordering the training data *a-priori* may introduce unintended bias during training. We also experimented with zero-padding but this caused the model to become fixated on sequence length, distributing embeddings along an axis correlating with sequence size alone. DeepSeqProt instead uses a linear interpolation to 'stretch' (or shrink) the one-hot encodings for each amino acid along an arbitrarily defined standard sequence length. In this way each input sequence can be thought of as 20 one-dimensional arrays, or 'channels', of a standardized arbitrary length, where each channel represents the relative contribution of a given amino acid along the length of the sequence.

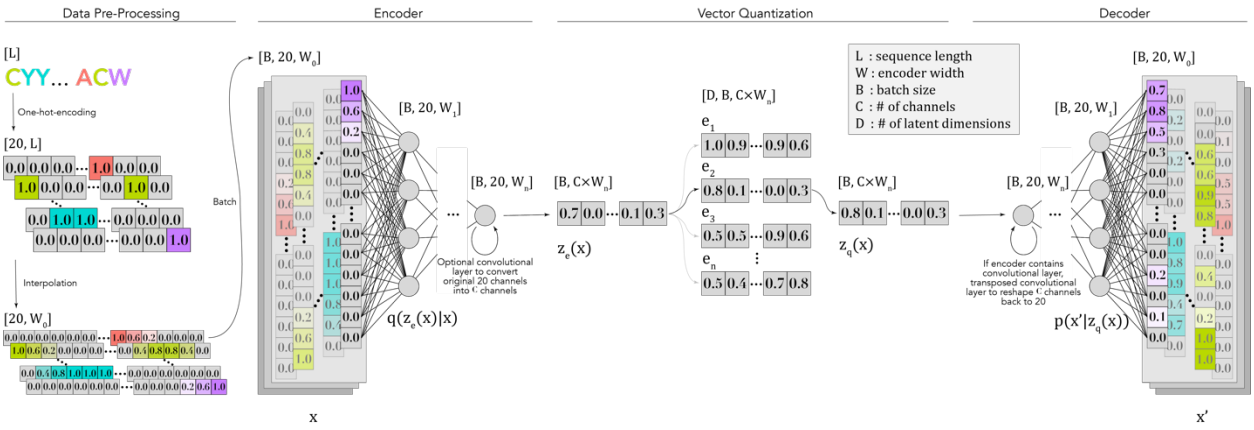

Figure 6.2. Overview of DeepSeqProt. Bracketed values represent the data shape at each step.

Encoder/Decoder

All VAEs contain an encoder and a decoder. Encoders and decoders are both neural networks, encoders are designed to compress the input data into a reduced latent representation, whereas decoders are designed to reconstruct the input data by sampling from the reduced latent space. Several different encoder/decoder schema were tested, with multiple convolutional layers and residual neural network architectures. Upon release, DeepSeqProt uses a fully connected feedforward network, which performed best during development (see also[240]). The depth and width of the network is provided by the user as a vector. For example, the default architecture of [2000, 1500, 1000, 500, 1] will consists of 4 hidden layers. The first value in the vector represents the standard length to which all input sequences will be interpolated to, and each successive value represents the 'width' of each hidden layer in order. In our example, the first hidden layer is 1500 'neurons' wide and is fed by fully connected linear transformations from the encoded input sequence, which have all been interpolated to a standard length of 2000. These values will then be passed through to the next hidden layer which is 1000 neurons wide, and so on. Note that each layer should have the same or fewer neurons to the one before it to accommodate the formulation behind autoencoders, which expect a progressive compression of the input data. The decoder architecture is always the inverse of the encoder. To continue our example the first hidden layer of the decoder will be a single neuron and ultimately terminate in a vector the same shape as the encoder input, as is necessary to estimate loss.

After traversing the fully connected layers but prior to vector quantization, DeepSeqProt also provides the option for a single convolutional layer. This layer allows data to be compressed

further along the 20 amino acid channels, which is necessary to visualize the latent space in 2 or 3 dimensions. The default kernel size, stride, and dilation of this layer are all equal to one, to preserve the data shape. Naturally if this option is selected a complementary transposed convolutional layer, or 'deconvolutional' layer, is automatically applied at the start of the decoder.

Vector Quantization

After passing through the encoder, the data is in shape [B, C, W], where B is the batch size, C is the number of channels (20, unless the optional convolution was applied), and W is the width of the final layer of the decoder. The first step of the vector quantization layer is to flatten the data to N vectors of length D, where D is the number of dimensions in the embedding space. Recall that the motivation for the vector quantization step is to provide a discrete distribution in latent space, in hopes that the model will use this distribution as a type of classifier for the categorical input data. As we want each sequence to be associated with a single embedding, our encoded data must be flattened to shape [N=B, D=C×W]. The number of embeddings is arbitrary and specified by the user. The default number is 1,000 though in development the model rarely used more than 100. If users wished to force the model to use a binary classifier, for example, they could use two embeddings instead.

After the encoded data has been flattened into the same dimensionality as the embedding vectors, distances are calculated from the encoded data to each of the embedding vectors. The closest vector is then reshaped back to [B, C, W] and passed to the decoder.

Loss and Learning

A noteworthy feature of VQ-VAEs compared to conventional VAEs is that the closest embedding lookup step has no gradient, and therefore cannot be trained through backpropagation. VQ-VAEs can circumvent this by simply copying the gradient from decoder input to encoder output[245], similar to the straight-through estimator for stochastic neurons[247]. The loss function itself is made up of three components, reconstruction loss, codebook loss, and commitment loss. Reconstruction loss optimizes the encoder and decoder by encouraging the decoder to compute a faithful reconstruction of the data x' from the original input data x. Since reconstruction loss receives no gradient from the embeddings, codebook loss is used to move the embedding e toward the encoder output $\boxed{z_e(x)}$. Finally, commitment loss prevents the latent space from growing arbitrarily large by encouraging the encoder to commit to an embedding.

During training, reconstruction loss is measured by the mean squared error of the input data from the reconstructed data, divided by the data variance. Likewise, codebook loss and commitment loss are measured by the mean squared error of the embeddings from the encoder output. Commitment loss also contains a weight hyperparameter β to allow the user some control over the size of latent space. The complete loss function can therefore be represented as

$$L = \log p\left(x' \middle| z_q(x)\right) + \|\mathrm{sg}[z_e(x)] - e\|_2^2 + \beta\|z_e(x) - \mathrm{sg}[e]\|_2^2$$

where sg indicates the stop gradient operator to indicate that these values are not directly updated by the gradient during training. Prior to loss estimation, the embeddings themselves are further updated by exponential moving averages (EMA). EMA optimization was suggested (though not implemented) by the original authors of VQ-VAE and we found that it performed better for our dataset than the original method[245]. Training stops automatically if the average loss of a sliding 1,000 update window fails to improve, or if the maximum number of training updates set by the user is reached. DeepSeqProt uses the Adam optimizer[248].

Benchmarking

The aim of DeepSeqProt is to learn latent embeddings which are each associated with specific groups of protein sequences. In this sense DeepSeqProt can be thought of as a clustering tool, albeit with added information from the structure of latent space. We benchmarked DeepSeqProt alongside several clustering algorithms designed for unaligned protein sequences. CD-HIT[249] was one of the first[250] clustering algorithms designed for large databases and uses short word filtering followed by greedy incremental clustering by sequence similarity[249,251]. CD-HIT uses a clustering threshold parameter which controls the short word filtering step, for which we tested two values, 50% and 75%. MMseqs2[252] is a similar software suite designed to quickly and sensitively search and cluster large sequence datasets[252] . MMseqs2 also has a minimum sequence identity parameter similar to CD-HIT's clustering threshold for which we tested values 25%, 50%, and 75%. Finally, we performed Markov clustering on a similarity matrix generated from an all vs. all DIAMOND[253] search. Each clustering algorithm was run on three proteomes, *Saccharomyces cerevisiae* (UP000002311), *Mus musculus* (UP000000589), and *Arabidopsis thaliana* (UP000006548) from the Uniprot protein database[47]. We selected several summary statistics to measure the accuracy and precision of each method.

First, we wanted to see how well DeepSeqProt embeddings represent entire protein families, as defined by UniProt. To achieve this, we calculated how many representatives from

each family were captured within the same embedding. We call this metric 'family completeness', formally defined as $\frac{\sum_{i=1}^{i=I} \text{Arg max}(\{y_1 \ldots y_n\})}{T}$ where I is the total number of protein families, y is the number of representatives from a given protein family across n embeddings, and T is the total number of protein sequences in the dataset. In other words, family completeness is the greatest fraction of sequences that are categorized together with other members of the same family. We also calculated the adjusted mutual information[254] (AMI) for protein families. Mutual information[255,256], or information gain, is a measure of mutual dependence between two random variables. AMI adds a correction to account for agreement due to random chance. Importantly, AMI is agnostic to the number of clusters and can be used to make fair comparisons across different methods. An AMI of 0 means that assigned labels (clusters) are totally incomplete and an AMI of 1 means that labels are totally complete and homogeneous.

Finally, we are interested in how well DeepSeqProt captures protein function. To this end, we performed gene ontology enrichment analysis[257] on each embedding/cluster produced by DeepSeqProt and our benchmark clustering algorithms (Table S1). From these we calculate the fraction of proteins that possess at least one of the gene ontology terms found significant ($p<0.01$ after Benjamini-Hochberg[258] correction) for that protein's assigned embedding. This metric, referred to hereafter as 'GO accuracy' is formally represented by the equation $\frac{\sum_{i=1}^{i=J} n(G)}{T}$ where $n(G)$ is the number of proteins in a set G with at least one significant gene ontology term in a given embedding across J embeddings. Notably, each of these summary statistics has a lower bound at or approaching 0 and an upper bound of 1. In addition to information gained from which proteins are associated with which embeddings, the position of the embeddings themselves may reflect local and global structure within functional space. To explore how deep learning may be used to visualize high dimensional protein function we reran DeepSeqProt with an additional convolutional layer, restricting the embedding vectors two dimensions to better visualize the latent space.

Results & Discussion

Both DeepSeqProt models (default and 2D convolutional) were trained on the 1,612 eukaryotic reference proteomes currently available in the UniProt database (release 2021_03). Training was performed on a single NVIDIA T4 Tensor Core GPU. The default model converged after ~4 hours and the 2D convolutional model converged after ~1.5. Once trained, models processed training proteomes on a standard Intel Xeon Gold 6248R CPU. Testing the model took ~2 minutes for all three proteomes for both models, or ~2-3 seconds per every 1,000 sequences.

Clustering

Embeddings produced from DeepSeqProt provide a distinct view of protein organization compared to conventional clustering methods based on pairwise similarity between sites (Fig. 6.3). DeepSeqProt had the highest family completeness statistic across taxa, from 0.68 in *M. musculus* to 0.76 in *S. cerevisiae*, offering an average 0.31 increase over conventional clustering. By contrast AMI values for DeepSeqProt were consistently lower compared to other methods, with an average decrease of 0.26. These results indicate that while DeepSeqProt embeddings were more inclusive to members of the same protein family, they were less likely to resolve protein families as homogeneous sets. With regard to GO accuracy, DeepSeqProt outperformed other methods, with an average increase of 0.35. This is likely because clustering through pairwise site similarity can produce very granular results that are too small to provide statistical power to enrichment analysis. This effect is especially apparent in the *S. cerevisiae* proteome due to its smaller size (Fig. 6.3), where as few as 2% of clusters had significant enrichment.

DeepSeqProt consistently outperformed all other benchmarks with regard to protein family completeness and gene ontology accuracy but underperformed on adjusted mutual information. Overall, these results indicate that DeepSeqProt generates fewer, larger protein sets that generally contain members with shared gene ontologies, but do not adequately reflect protein families. One strength of neural networks for analyzing biological sequences is their ability to analyze latent factors between sites across the entire length of the sequence, compared to the sitewise factors used by conventional clustering[241]. Therefore, it is perhaps unsurprising that DeepSeqProt is more “lumper” than “splitter” compared to conventional clustering, as single site differences are less likely to dissuade the model from using the same embedding for similar proteins. Despite the lack of specificity, groups assigned by DeepSeqProt embeddings may still be of use for researchers interested in capturing the entire diversity of a protein family with less regard to how many other, similar proteins may be included as well, or for capturing broad

functional classes. For example, DeepSeqProt may serve as a useful precursor to homology inference or tree inference steps in programs like OrthoFinder[219]. DeepSeqProt will also likely capture more diverse homologs from distantly related or underrepresented taxa than conventional clustering, as it is less dissuaded by single site sequence dissimilarity.

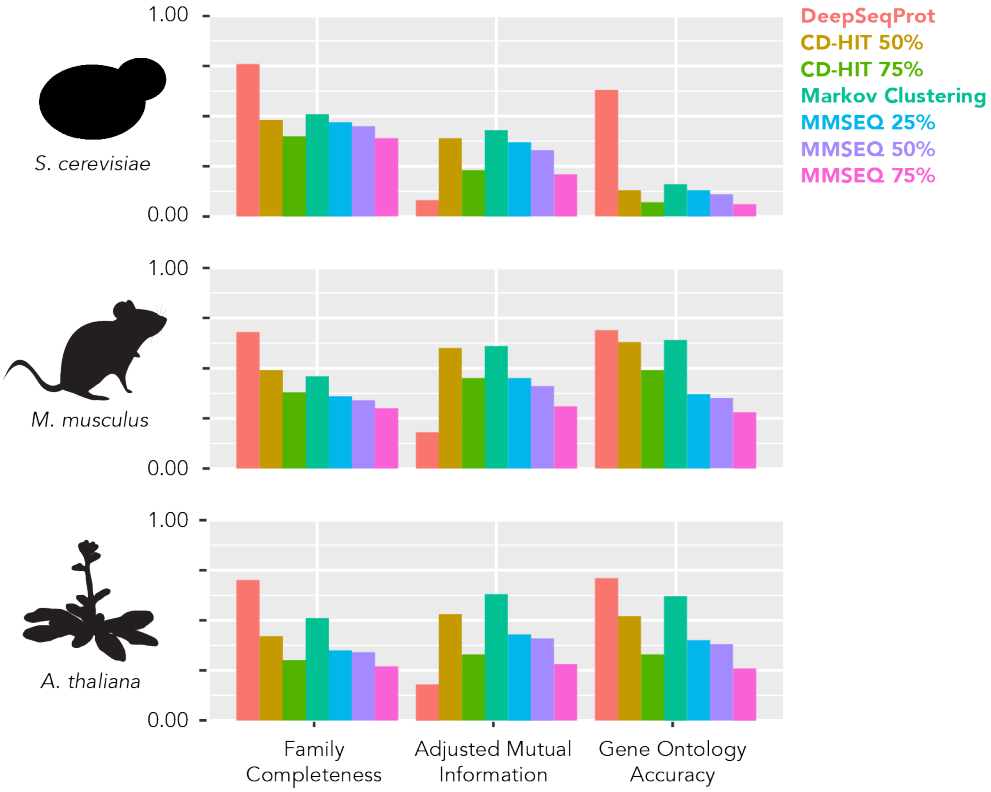

Figure 6.3. Benchmarking results.

Visualization

Embeddings trained in a two- or three-dimensional latent space offer a way to visualize protein sets along axes that contain biologically relevant information. In this way, DeepSeqProt performs similarly to other dimensionality reduction techniques such as UMAP[259] or t-SNE[260] (though a recent study suggests the visualizations produced by such methods are more arbitrary than previously suspected[261]). We find that DeepSeqProt embeddings preserve the general local and global structure of proteome functional diversity (Fig. 6.4). For example, embeddings whose members are most associated with DNA-templated transcription are located closely together within latent space, while also coinciding with other embeddings associated with RNA polymerase II transcription. This “transcription space” can be contrasted with “signaling space” around quadrant IV of Figure 6.4 where cell membrane signaling associated embeddings are

found, home to families like G proteins or defensin-like proteins in the case of *Arabidopsis*. This type of structure is apparent across all three aspects of gene ontology and shared across eukaryotic kingdoms. These visualizations demonstrate the generality and versatility of DeepSeqProt in "mapping" protein sequence data in functional space and may be a useful tool in categorizing proteins from diverse taxonomic groups distantly related to model organisms.

What Can Unsupervised Learning Teach Us About Molecular Diversity?

Traditional methods of assessing protein function are reliant on connecting new sequences to a small number of high confidence references. Although this approach is powerful, it is also necessarily limited. Such methods will at best underestimate the full breadth of molecular diversity and at worst create systematic bias, leading to a warped perception of biology[203,262]. Unsupervised learning provides an alternative approach. Information is not tied to specific references but instead can be produced from the data itself in aggregate. Given enough data an adequately specified model will learn complex features and relationships of novel proteins on its own, providing a model-agnostic view of protein functional diversity. One potential advantage to this approach is reduced bias, training datasets are limited only by the availability of sequence data and can otherwise be as taxonomically balanced as the user wants, rather than relying on similarity to a handful of model organisms. The tradeoff for this reduced bias is a lack of specificity. For example, DeepSeqProt is unable to report with confidence whether a protein belongs to a particular family, and reference information is still necessary in order to interpret its results. However, reference data does not contribute to the operation of the model itself.

We predict that both unsupervised approaches and traditional reference-based methods will be important to the future of protein annotation. As we demonstrate here, autoencoders are more likely to generate fewer, larger, groups compared to traditional clustering. This offers a "greedy" approach for identifying diverse proteins which largely fall under the same functional classes. These groups can then be examined themselves using a more conventional, conservative approach to increase resolution and provide specific annotations. Unsupervised learning may also serve as a way to benchmark traditional methods in order to "keep them honest". For example, if proteins from a non-model organism are routinely associated with the same embeddings and regions of latent space as known cell-signaling proteins, but lack any cell-signaling domains themselves, it may indicate a lack of information in reference databases or an undiscovered protein family.

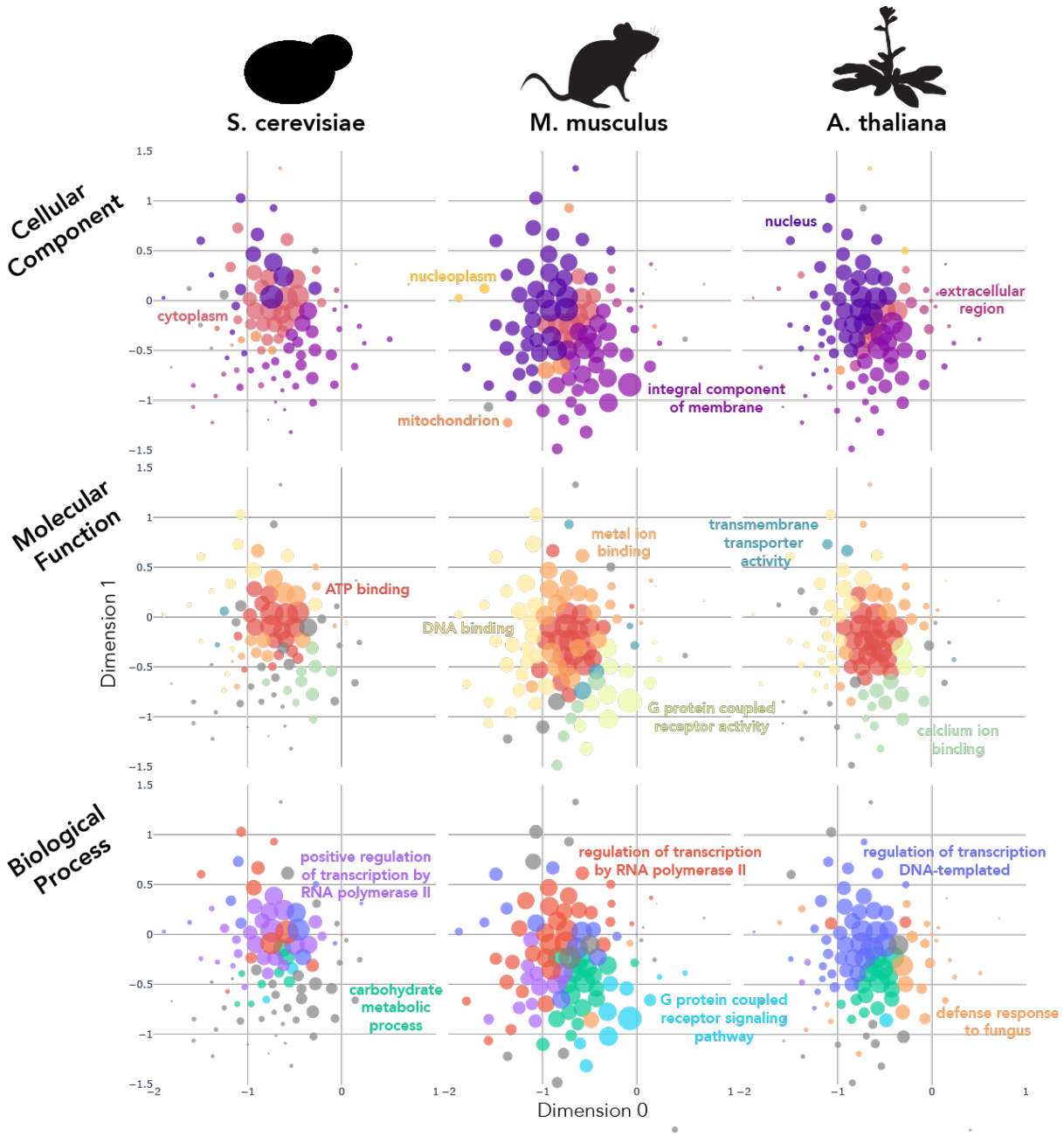

Figure 6.4. Visualizations in two-dimensional latent space. Circles represent embeddings, size of the circle represents the relative number of proteins associated with the embedding, and colors represent the most popular ontology for that embedding of the top six most common ontologies total. Embeddings that do not contain any proteins with one of these common ontologies are grey.

Use Cases & Caveats

The primary motivation behind DeepSeqProt was to develop a method for exploring protein function that was not reference dependent. However, DeepSeqProt is not a replacement for experimental inquiry, as all machine learning models ultimately need a ground truth to validate their results against. The inclusion of high-confidence reference data from model organisms will

not contribute to the performance of the model itself, however, they are still necessary in order to interpret results.

For the purposes of this paper, we focus on analyzing proteome-scale data, though note that DeepSeqProt can be trained and run on any number of protein sequences in any context. Proteins found to be differentially expressed, for example, can be analyzed on a model trained on background proteins to observe whether up/downregulated proteins cluster together, and where such cluster(s) are distributed in latent space. Similarly, while we discuss a very general model trained across eukaryotes here, researchers may train on any taxon or clade they prefer. Indeed, we may expect to see greater resolution and increased model performance on datasets tailored toward more exclusive clades.

Conclusions

Here we present an unsupervised deep learn model, called DeepSeqProt, that is capable of learning salient biological features from unaligned, unannotated sequences. DeepSeqProt is capable of recognizing broad functional classes across eukaryotic life, as evidenced by its ability to cluster sequences of similar function together and distributing such clusters coherently throughout latent space. Such a method has uses for analyzing protein sequences, particularly for non-model groups. More importantly, however, DeepSeqProt provides a starting point for unsupervised deep learning and its applications in molecular biology.

Data Availability

DeepSeqProt and all scripts used for benchmarking and validation are available at https://github.com/KyleTDavid/DeepSeqProt. This repository also contains pickled python objects of both trained models discussed in this paper. An interactive version of DeepSeqProt is also available at https://colab.research.google.com/drive/1FcOCECzhUg35PcXjb-LfdsSYDQoZurV3?usp=sharing.

References

1. Ohno, S. *Evolution by gene duplication*. (Springer Science & Business Media, 1970).
2. Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. & Postlethwait, J. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151,** 1531–1545 (1999).
3. Fitch, W. M. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19,** 99–113 (1970).
4. Han, M. V., Demuth, J. P., McGrath, C. L., Casola, C. & Hahn, M. W. Adaptive evolution of young gene duplicates in mammals. *Genome Res.* **19,** 859–867 (2009).
5. Brunet, F. G., Crollius, H. R., Paris, M., Aury, J.-M., Gibert, P., Jaillon, O., Laudet, V. & Robinson-Rechavi, M. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol. Biol. Evol.* **23,** 1808–1816 (2006).
6. Li, Z., Defoort, J., Tasdighian, S., Maere, S., Van de Peer, Y. & De Smet, R. Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* **28,** 326–344 (2016).
7. Singh, P. P., Arora, J. & Isambert, H. Identification of ohnolog genes originating from whole genome duplication in early vertebrates, based on synteny comparison across multiple genomes. *PLoS Comput. Biol.* **11,** (2015).
8. Wendel, J. F. in *Plant Mol. Evol.* 225–249 (Springer, 2000).
9. Song, K., Lu, P., Tang, K. & Osborn, T. C. Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. *Proc. Natl. Acad. Sci.* **92,** 7719–7723 (1995).
10. Adams, K. L., Cronn, R., Percifield, R. & Wendel, J. F. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci.* **100,** 4649–4654 (2003).
11. Cavalier-Smith, T. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *J. Cell Sci.* **34,** 247–278 (1978).
12. Gregory, T. R. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol. Rev.* **76,** 65–101 (2001).
13. Gregory, T. R. & Mable, B. K. in *Evol. Genome* 427–517 (Elsevier, 2005).
14. Otto, S. P. & Whitton, J. Polyploid incidence and evolution. *Annu. Rev. Genet.* **34,** 401–437 (2000).
15. Doyle, J. J. & Coate, J. E. Polyploidy, the nucleotype, and novelty: the impact of genome doubling on the biology of the cell. *Int. J. Plant Sci.* **180,** 1–52 (2019).
16. Otto, S. P. The evolutionary consequences of polyploidy. *Cell* **131,** 452–462 (2007).
17. Itgen, M. W., Natalie, G. R., Siegel, D. S., Sessions, S. K. & Mueller, R. L. Genome size drives morphological evolution in organ-specific ways. *Evolution* (2022).
18. Levin, D. A. The ecological transition in speciation. *New Phytol.* **161,** 91–96 (2004).
19. Simakov, O., Bredeson, J., Berkoff, K., Marletaz, F., Mitros, T., Schultz, D. T., O'Connell, B. L., Dear, P., Martinez, D. E. & Steele, R. E. Deeply conserved synteny and the evolution of metazoan chromosomes. *Sci. Adv.* **8,** eabi5884 (2022).
20. Glasauer, S. M. & Neuhauss, S. C. Whole-genome duplication in teleost fishes and its evolutionary consequences. *Mol. Genet. Genomics* **289,** 1045–1060 (2014).
21. Li, Z., Tiley, G. P., Galuska, S. R., Reardon, C. R., Kidder, T. I., Rundell, R. J. & Barker, M. S. Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc. Natl. Acad. Sci.* **115,** 4713–4718 (2018).

22. Clarke, T. H., Garb, J. E., Hayashi, C. Y., Arensburger, P. & Ayoub, N. A. Spider transcriptomes identify ancient large-scale gene duplication event potentially important in silk gland evolution. *Genome Biol. Evol.* **7,** 1856–1870 (2015).
23. Nossa, C. W., Havlak, P., Yue, J.-X., Lv, J., Vincent, K. Y., Brockmann, H. J. & Putnam, N. H. Joint assembly and genetic mapping of the Atlantic horseshoe crab genome reveals ancient whole genome duplication. *GigaScience* **3,** 2047–217X (2014).
24. Liu, C., Ren, Y., Li, Z., Hu, Q., Yin, L., Wang, H., Qiao, X., Zhang, Y., Xing, L. & Xi, Y. Giant African snail genomes provide insights into molluscan whole-genome duplication and aquatic–terrestrial transition. *Mol. Ecol. Resour.* **21,** 478–494 (2021).
25. Hoegg, S., Brinkmann, H., Taylor, J. S. & Meyer, A. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J. Mol. Evol.* **59,** 190–203 (2004).
26. Tank, D. C., Eastman, J. M., Pennell, M. W., Soltis, P. S., Soltis, D. E., Hinchliff, C. E., Brown, J. W., Sessa, E. B. & Harmon, L. J. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytol.* **207,** 454–467 (2015).
27. Fawcett, J. A., Maere, S. & Van De Peer, Y. Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proc. Natl. Acad. Sci.* **106,** 5737–5742 (2009).
28. Moriyama, Y. & Koshiba-Takeuchi, K. Significance of whole-genome duplications on the emergence of evolutionary novelties. *Brief. Funct. Genomics* **17,** 329–338 (2018).
29. Van de Peer, Y., Mizrachi, E. & Marchal, K. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18,** 411 (2017).
30. Nakatani, Y., Takeda, H., Kohara, Y. & Morishita, S. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.* **17,** 1254–1265 (2007).
31. Dulai, K. S., von Dornum, M., Mollon, J. D. & Hunt, D. M. The evolution of trichromatic color vision by opsin gene duplication in New World and Old World primates. *Genome Res.* **9,** 629–638 (1999).
32. Beeman, R. W. A homoeotic gene cluster in the red flour beetle. *Nature* **327,** 247 (1987).
33. Nei, M., Gu, X. & Sitnikova, T. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc. Natl. Acad. Sci.* **94,** 7799–7806 (1997).
34. Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* **39,** 309–338 (2005).
35. Dolinski, K. & Botstein, D. Orthology and functional conservation in eukaryotes. *Annu Rev Genet* **41,** 465–507 (2007).
36. Studer, R. A. & Robinson-Rechavi, M. How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet.* **25,** 210–216 (2009).
37. Otto, S. P. & Yong, P. in *Adv. Genet.* **46,** 451–483 (Elsevier, 2002).
38. Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* **11,** 97 (2010).
39. Hahn, M. W. Distinguishing among evolutionary models for the maintenance of gene duplicates. *J. Hered.* **100,** 605–617 (2009).
40. Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. & Postlethwait, J. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151,** 1531–1545 (1999).

41. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290,** 1151–1155 (2000).
42. Lynch, M. & Katju, V. The altered evolutionary trajectories of gene duplicates. *TRENDS Genet.* **20,** 544–549 (2004).
43. Gabaldón, T. & Koonin, E. V. Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.* **14,** 360–366 (2013).
44. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44,** D457–D462 (2015).
45. Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A. & Narechania, A. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13,** 2129–2141 (2003).
46. Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S. & Kuhn, M. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44,** D286–D293 (2015).
47. Consortium, U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47,** D506–D515 (2018).
48. Bleidorn, C. *Phylogenomics*. (Springer International Publishing, 2017).
49. Yanai, I., Graur, D. & Ophir, R. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *Omics J. Integr. Biol.* **8,** 15–24 (2004).
50. Liao, B.-Y. & Zhang, J. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol. Biol. Evol.* **23,** 530–540 (2005).
51. Nehrt, N. L., Clark, W. T., Radivojac, P. & Hahn, M. W. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput. Biol.* **7,** e1002073 (2011).
52. Chen, X. & Zhang, J. The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. *PLoS Comput. Biol.* **8,** e1002784 (2012).
53. Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M. & Dessimoz, C. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput. Biol.* **8,** e1002514 (2012).
54. Thomas, P. D., Wood, V., Mungall, C. J., Lewis, S. E., Blake, J. A. & Consortium, G. O. On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: a short report. *PLoS Comput. Biol.* **8,** e1002386 (2012).
55. Rogozin, I. B., Managadze, D., Shabalina, S. A. & Koonin, E. V. Gene family level comparative analysis of gene expression in mammals validates the ortholog conjecture. *Genome Biol. Evol.* **6,** 754–762 (2014).
56. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. Tissue-specificity of gene expression diverges slowly between orthologs, and rapidly between paralogs. *PLoS Comput. Biol.* **12,** e1005274 (2016).
57. Dunn, C. W., Zapata, F., Munro, C., Siebert, S. & Hejnol, A. Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proc. Natl. Acad. Sci.* **115,** E409–E417 (2018).

58. Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R. & Birney, E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19,** 327–335 (2009).
59. Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M., Amode, R. & Brent, S. Ensembl comparative genomics resources. *Database* **2016,** (2016).
60. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).
61. Huerta-Cepas, J., Dopazo, H., Dopazo, J. & Gabaldón, T. The human phylome. *Genome Biol.* **8,** R109 (2007).
62. Villanueva-Cañas, J. L., Laurie, S. & Albà, M. M. Improving genome-wide scans of positive selection by using protein isoforms of similar length. *Genome Biol. Evol.* **5,** 457–467 (2013).
63. Yang, Z. & Bielawski, J. P. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15,** 496–503 (2000).
64. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15,** 568–573 (1998).
65. Yang, Z. On the best evolutionary rate for phylogenetic analysis. *Syst. Biol.* **47,** 125–133 (1998).
66. Hedges, L. V. Distribution theory for Glass's estimator of effect size and related estimators. *J. Educ. Stat.* **6,** 107–128 (1981).
67. Inman, H. F. & Bradley Jr, E. L. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Commun. Stat.-Theory Methods* **18,** 3851–3874 (1989).
68. Massey Jr, F. J. The Kolmogorov-Smirnov test for goodness of fit. *J. Am. Stat. Assoc.* **46,** 68–78 (1951).
69. Seoighe, C., Johnston, C. R. & Shields, D. C. Significantly different patterns of amino acid replacement after gene duplication as compared to after speciation. *Mol. Biol. Evol.* **20,** 484–490 (2003).
70. Johnston, C. R., O'dushlaine, C., Fitzpatrick, D. A., Edwards, R. J. & Shields, D. C. Evaluation of whether accelerated protein evolution in chordates has occurred before, after, or simultaneously with gene duplication. *Mol. Biol. Evol.* **24,** 315–323 (2006).
71. Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Selection in the evolution of gene duplications. *Genome Biol.* **3,** research0008. 1 (2002).
72. Scannell, D. R. & Wolfe, K. H. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Res.* **18,** 137–147 (2008).
73. Francino, M. P. An adaptive radiation model for the origin of new gene functions. *Nat. Genet.* **37,** 573 (2005).
74. Conant, G. C. & Wagner, A. Asymmetric sequence divergence of duplicate genes. *Genome Res.* **13,** 2052–2058 (2003).
75. Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428,** 617 (2004).
76. Studer, R. A., Penel, S., Duret, L. & Robinson-Rechavi, M. Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome Res.* **18,** 1393–1402 (2008).

77. Cannarozzi, G. M. & Schneider, A. *Codon evolution: mechanisms and models*. (Oxford University Press, 2012).
78. Stoletzki, N. & Eyre-Walker, A. The Positive Correlation between d N/d S and d S in Mammals Is Due to Runs of Adjacent Substitutions. *Mol. Biol. Evol.* **28,** 1371–1380 (2010).
79. He, X. & Zhang, J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169,** 1157–1164 (2005).
80. Comai, L. The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* **6,** 836 (2005).
81. Parisod, C., Holderegger, R. & Brochmann, C. Evolutionary consequences of autopolyploidy. *New Phytol.* **186,** 5–17 (2010).
82. Stebbins, G. L. Polyploidy and the distribution of the arctic-alpine flora: new evidence and a new approach. *Bot Helvetica* **94,** 1–13 (1984).
83. Adamowicz, S. J., Gregory, T. R., Marinone, M. C. & Hebert, P. D. New insights into the distribution of polyploid Daphnia: the Holarctic revisited and Argentina explored. *Mol. Ecol.* **11,** 1209–1217 (2002).
84. Brochmann, C., Brysting, A. K., Alsos, I. G., Borgen, L., Grundt, H. H., Scheen, A.-C. & Elven, R. Polyploidy in arctic plants. *Biol. J. Linn. Soc.* **82,** 521–536 (2004).
85. Rice, A., Šmarda, P., Novosolov, M., Drori, M., Glick, L., Sabath, N., Meiri, S., Belmaker, J. & Mayrose, I. The global biogeography of polyploid plants. *Nat. Ecol. Evol.* **3,** 265–273 (2019).
86. Thompson, K. A., Husband, B. C. & Maherali, H. Climatic niche differences between diploid and tetraploid cytotypes of Chamerion angustifolium (Onagraceae). *Am. J. Bot.* **101,** 1868–1875 (2014).
87. Manzaneda, A. J., Rey, P. J., Anderson, J. T., Raskin, E., Weiss-Lehman, C. & Mitchell-Olds, T. Natural variation, differentiation, and genetic trade-offs of ecophysiological traits in response to water limitation in Brachypodium distachyon and its descendent allotetraploid B. hybridum (Poaceae). *Evolution* **69,** 2689–2704 (2015).
88. Diallo, A. M., Nielsen, L. R., Kjær, E. D., Petersen, K. K. & Ræbild, A. Polyploidy can confer superiority to West African Acacia senegal (L.) Willd. trees. *Front. Plant Sci.* **7,** 821 (2016).
89. Chao, D.-Y., Dilkes, B., Luo, H., Douglas, A., Yakubova, E., Lahner, B. & Salt, D. E. Polyploids exhibit higher potassium uptake and salinity tolerance in Arabidopsis. *Science* **341,** 658–659 (2013).
90. Borgen, L. I. V. & Hultfgård, U.-M. Parnassia palustris: a genetically diverse species in Scandinavia. *Bot. J. Linn. Soc.* **142,** 347–372 (2003).
91. Jørgensen, M. H., Carlsen, T., Skrede, I. & Elven, R. Microsatellites resolve the taxonomy of the polyploid Cardamine digitata aggregate (Brassicaceae). *Taxon* **57,** 882–892 (2008).
92. Oberlander, K. C., Dreyer, L. L., Goldblatt, P., Suda, J. & Linder, H. P. Species-rich and polyploid-poor: Insights into the evolutionary role of whole-genome duplication from the Cape flora biodiversity hotspot. *Am. J. Bot.* **103,** 1336–1347 (2016).
93. Novikova, P. Y., Hohmann, N. & Van de Peer, Y. Polyploid Arabidopsis species originated around recent glaciation maxima. *Curr. Opin. Plant Biol.* **42,** 8–15 (2018).
94. Kreiner, J. M., Kron, P. & Husband, B. C. Frequency and maintenance of unreduced gametes in natural plant populations: associations with reproductive mode, life history and genome size. *New Phytol.* **214,** 879–889 (2017).

95. Mason, A. S., Nelson, M. N., Yan, G. & Cowling, W. A. Production of viable male unreduced gametes in Brassica interspecific hybrids is genotype specific and stimulated by cold temperatures. *BMC Plant Biol.* **11,** 103 (2011).
96. Pandit, S. K., Westendorp, B. & de Bruin, A. Physiological significance of polyploidization in mammalian cells. *Trends Cell Biol.* **23,** 556–566 (2013).
97. Mable, B. K., Alexandrou, M. A. & Taylor, M. I. Genome duplication in amphibians and fish: an extended synthesis. *J. Zool.* **284,** 151–182 (2011).
98. Zhou, L. & Gui, J. Natural and artificial polyploids in aquaculture. *Aquac. Fish.* **2,** 103–111 (2017).
99. Ueda, H. Mating calls of autotriploid and autotetraploid males in Hyla japonica. *Sci. Rep. Lab. Amphib. Biol.* **12,** 177–189 (1993).
100. Keller, M. J. & Carl Gerhardt, H. Polyploidy alters advertisement call structure in gray treefrogs. *Proc. R. Soc. Lond. B Biol. Sci.* **268,** 341–345 (2001).
101. Grismer, J. L., Bauer, A. M., Grismer, L. L., Thirakhupt, K., Aowphol, A., Oaks, J. R., Wood Jr, P. L., Onn, C. K., Thy, N. & Cota, M. Multiple origins of parthenogenesis, and a revised species phylogeny for the Southeast Asian butterfly lizards, Leiolepis. *Biol. J. Linn. Soc.* **113,** 1080–1093 (2014).
102. Levin, D. A. Minority cytotype exclusion in local plant populations. *Taxon* **24,** 35–43 (1975).
103. Husband, B. C. & Schemske, D. W. Ecological Mechanisms of Reproductive Isolation between Diploid and Tetraploid Chamerion angustifolium. *J. Ecol.* **88,** 689–701 (2000).
104. Ramsey, J. & Schemske, D. W. Neopolyploidy in flowering plants. *Annu. Rev. Ecol. Syst.* **33,** 589–639 (2002).
105. Madlung, A. Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity* **110,** 99 (2013).
106. Baniaga, A. E., Marx, H. E., Arrigo, N. & Barker, M. S. Polyploid plants have faster rates of multivariate niche differentiation than their diploid relatives. *Ecol. Lett.* **23,** 68–78 (2020).
107. Levin, D. A. Polyploidy and novelty in flowering plants. *Am. Nat.* **122,** 1–25 (1983).
108. Fowler, N. L. & Levin, D. A. Ecological constraints on the establishment of a novel polyploid in competition with its diploid progenitor. *Am. Nat.* **124,** 703–711 (1984).
109. Felber, F. Establishment of a tetraploid cytotype in a diploid population: effect of relative fitness of the cytotypes. *J. Evol. Biol.* **4,** 195–207 (1991).
110. Rodriguez, D. J. A model for the establishment of polyploidy in plants. *Am. Nat.* **147,** 33–46 (1996).
111. Fowler, N. L. & Levin, D. A. Critical factors in the establishment of allopolyploids. *Am. J. Bot.* **103,** 1236–1251 (2016).
112. Selmecki, A. M., Maruvka, Y. E., Richmond, P. A., Guillet, M., Shoresh, N., Sorenson, A. L., De, S., Kishony, R., Michor, F. & Dowell, R. Polyploidy can drive rapid adaptation in yeast. *Nature* **519,** 349 (2015).
113. Vanneste, K., Baele, G., Maere, S. & Van de Peer, Y. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Res.* **24,** 1334–1347 (2014).
114. Vanneste, K., Maere, S. & Van de Peer, Y. Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Philos. Trans. R. Soc. B Biol. Sci.* **369,** 20130353 (2014).

115. Te Beest, M., Le Roux, J. J., Richardson, D. M., Brysting, A. K., Suda, J., Kubešová, M. & Pyšek, P. The more the better? The role of polyploidy in facilitating plant invasions. *Ann. Bot.* **109,** 19–45 (2011).
116. Pandit, M. K., Pocock, M. J. & Kunin, W. E. Ploidy influences rarity and invasiveness in plants. *J. Ecol.* **99,** 1108–1115 (2011).
117. Schmid, M., Evans, B. J. & Bogart, J. P. Polyploidy in Amphibia. *Cytogenet. Genome Res.* **145,** 315–330 (2015).
118. Michonneau, F., Brown, J. W. & Winter, D. J. rotl: an R package to interact with the Open Tree of Life data. *Methods Ecol. Evol.* **7,** 1476–1481 (2016).
119. Batistic, R. F., Soma, M., Beçak, M. L. & Beçak, W. Further studies on polyploid amphibians: a diploid population of Phyllomedusa burmeisteri. *J. Hered.* **66,** 160–162 (1975).
120. IUCN. The IUCN Red List of Threatened Species. Version 2019-3. (2019). at <http://www.iucnredlist.org>
121. Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37,** 4302–4315 (2017).
122. Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E. & Yates, C. J. A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.* **17,** 43–57 (2011).
123. Warren, D. L., Glor, R. E. & Turelli, M. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evol. Int. J. Org. Evol.* **62,** 2868–2883 (2008).
124. Fitzpatrick, B. M. & Turelli, M. The geography of mammalian speciation: mixed signals from phylogenies and range maps. *Evolution* **60,** 601–615 (2006).
125. Hijmans, R. J., Phillips, S., Leathwick, J., Elith, J. & Hijmans, M. R. J. Package 'dismo'. *Circles* **9,** 1–68 (2017).
126. Heibl, C. & Calenge, C. Phyloclim: integrating phylogenetics and climatic niche modelling. *R Package Version 09-4 Softw.* (2013).
127. Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J.-M., Taly, J.-F. & Notredame, C. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.* **39,** W13–W17 (2011).
128. Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32,** 268–274 (2015).
129. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14,** 587 (2017).
130. Kim, J. & Sanderson, M. J. Penalized likelihood phylogenetic inference: bridging the parsimony-likelihood gap. *Syst. Biol.* **57,** 665–674 (2008).
131. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20,** 289–290 (2004).
132. Köppen, W. Versuch einer Klassifikation der Klimate, vorzugsweise nach ihren Beziehungen zur Pflanzenwelt. *Geogr. Z.* **6,** 593–611 (1900).
133. Geiger, R. Klassifikation der klimate nach W. Köppen. *Landolt-Börnstein–Zahlenwerte Funkt. Aus Phys. Chem. Astron. Geophys. Tech.* **3,** 603–607 (1954).
134. Kottek, M., Grieser, J., Beck, C., Rudolf, B. & Rubel, F. World map of the Köppen-Geiger climate classification updated. *Meteorol. Z.* **15,** 259–263 (2006).

135. Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E. & Morrison, J. C. Terrestrial Ecoregions of the World: A New Map of Life on EarthA new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience* **51,** 933–938 (2001).
136. Potter, P., Ramankutty, N., Bennett, E. M. & Donner, S. D. Characterizing the Spatial Patterns of Global Fertilizer Application and Manure Production. *Earth Interact.* **14,** 1–22 (2010).
137. Ramankutty, N., Evan, A. T., Monfreda, C. & Foley, J. A. Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000. *Glob. Biogeochem. Cycles* **22,** (2008).
138. Maggi, F., Tang, F. H., la Cecilia, D. & McBratney, A. PEST-CHEMGRIDS, global gridded maps of the top 20 crop-specific pesticide application rates from 2015 to 2025. *Sci. Data* **6,** 1–20 (2019).
139. Garland Jr, T., Dickerman, A. W., Janis, C. M. & Jones, J. A. Phylogenetic analysis of covariance by computer simulation. *Syst. Biol.* **42,** 265–292 (1993).
140. Pennell, M. W., Eastman, J. M., Slater, G. J., Brown, J. W., Uyeda, J. C., FitzJohn, R. G., Alfaro, M. E. & Harmon, L. J. geiger v2. 0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics* **30,** 2216–2218 (2014).
141. Revell, L. J. Size-correction and principal components for interspecific comparative studies. *Evol. Int. J. Org. Evol.* **63,** 3258–3268 (2009).
142. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3,** 217–223 (2012).
143. Candela, A. M., Bonini, R. A. & Noriega, J. I. First continental vertebrates from the marine Paraná Formation (Late Miocene, Mesopotamia, Argentina): Chronology, biogeography, and paleoenvironments. *Geobios* **45,** 515–526 (2012).
144. Gross, M., Ramos, M. I. F. & Piller, W. E. A minute ostracod (Crustacea: Cytheromatidae) from the Miocene Solimões Formation (western Amazonia, Brazil): evidence for marine incursions? *J. Syst. Palaeontol.* **14,** 581–602 (2016).
145. Ruskin, B. G., Dávila, F. M., Hoke, G. D., Jordan, T. E., Astini, R. A. & Alonso, R. Stable isotope composition of middle Miocene carbonates of the Frontal Cordillera and Sierras Pampeanas: Did the Paranaense seaway flood western and central Argentina? *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **308,** 293–303 (2011).
146. Ortiz-Jaureguizar, E. & Cladera, G. A. Paleoenvironmental evolution of southern South America during the Cenozoic. *J. Arid Environ.* **66,** 498–532 (2006).
147. Liebmann, B., Vera, C. S., Carvalho, L. M. V., Camilloni, I. A., Hoerling, M. P., Allured, D., Barros, V. R., Báez, J. & Bidegain, M. An Observed Trend in Central South American Precipitation. *J. Clim.* **17,** 4357–4367 (2004).
148. Gonzalez, P., Neilson, R. P., Lenihan, J. M. & Drapek, R. J. Global patterns in the vulnerability of ecosystems to vegetation shifts due to climate change. *Glob. Ecol. Biogeogr.* **19,** 755–768 (2010).
149. Haylock, M. R., Peterson, T. C., Alves, L. M., Ambrizzi, T., Anunciação, Y. M. T., Baez, J., Barros, V. R., Berlato, M. A., Bidegain, M., Coronel, G., Corradi, V., Garcia, V. J., Grimm, A. M., Karoly, D., Marengo, J. A., Marino, M. B., Moncunill, D. F., Nechet, D., Quintana, J., Rebello, E., Rusticucci, M., Santos, J. L., Trebejo, I. & Vincent, L. A. Trends in Total and

Extreme South American Rainfall in 1960–2000 and Links with Sea Surface Temperature. *J. Clim.* **19,** 1490–1512 (2006).
150. Barros, V. R., Doyle, M. E. & Camilloni, I. A. Precipitation trends in southeastern South America: relationship with ENSO phases and with low-level circulation. *Theor. Appl. Climatol.* **93,** 19–33 (2008).
151. Seager, R., Naik, N., Baethgen, W., Robertson, A., Kushnir, Y., Nakamura, J. & Jurburg, S. Tropical Oceanic Causes of Interannual to Multidecadal Precipitation Variability in Southeast South America over the Past Century. *J. Clim.* **23,** 5517–5539 (2010).
152. Gonzalez, P. L. M., Polvani, L. M., Seager, R. & Correa, G. J. P. Stratospheric ozone depletion: a key driver of recent precipitation trends in South Eastern South America. *Clim. Dyn.* **42,** 1775–1792 (2014).
153. Newbold, T., Hudson, L. N., Arnell, A. P., Contu, S., De Palma, A., Ferrier, S., Hill, S. L., Hoskins, A. J., Lysenko, I. & Phillips, H. R. Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. *Science* **353,** 288–291 (2016).
154. Baker, N. J., Bancroft, B. A. & Garcia, T. S. A meta-analysis of the effects of pesticides and fertilizers on survival and growth of amphibians. *Sci. Total Environ.* **449,** 150–156 (2013).
155. Ortiz, M. E., Marco, A., Saiz, N. & Lizana, M. Impact of Ammonium Nitrate on Growth and Survival of Six European Amphibians. *Arch. Environ. Contam. Toxicol.* **47,** 234–239 (2004).
156. Viglizzo, E. F. & Frank, F. C. Ecological interactions, feedbacks, thresholds and collapses in the Argentine Pampas in response to climate and farming during the last century. *Quat. Int.* **158,** 122–126 (2006).
157. Schoenfelder, K. P. & Fox, D. T. The expanding implications of polyploidy. *J Cell Biol* **209,** 485–491 (2015).
158. Levin, D. A. Has the Polyploid Wave Ebbed? *Front. Plant Sci.* **11,** (2020).
159. Otto, S. P. Adaptation, speciation and extinction in the Anthropocene. *Proc. R. Soc. B* **285,** 20182047 (2018).
160. Pollo, F. E., Grenat, P. R., Otero, M. A., Babini, S., Salas, N. E. & Martino, A. L. Evaluation in situ of genotoxic and cytotoxic response in the diploid/polyploid complex Odontophrynus (Anura: Odontophrynidae) inhabiting agroecosystems. *Chemosphere* **216,** 306–312 (2019).
161. Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B. & Rieseberg, L. H. The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci.* **106,** 13875–13879 (2009).
162. Dehal, P. & Boore, J. L. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3,** e314 (2005).
163. Zhang, L. & Lefcort, H. The effects of ploidy level on the thermal distributions of brine shrimp Artemia parthenogenetica and its ecological implications. *Heredity* **66,** 445–452 (1991).
164. Beaton, M. J. & Hebert, P. D. Geographical parthenogenesis and polyploidy in Daphnia pulex. *Am. Nat.* **132,** 837–845 (1988).
165. Ward, R. D., Bickerton, M. A., Finston, T. & Hebert, P. D. Geographical cline in breeding systems and ploidy levels in European populations of Daphnia pulex. *Heredity* **73,** 532–543 (1994).

166. Little, T. J. & Hebert, P. D. N. Clonal diversity in high arctic ostracodes. *J. Evol. Biol.* **10,** 233–252 (1997).
167. Dufresne, F. & Hebert, P. D. Pleistocene glaciations and polyphyletic origins of polyploidy in an arctic cladoceran. *Proc. R. Soc. Lond. B Biol. Sci.* **264,** 201–206 (1997).
168. Barata, C., Hontoria, F., Amat, F. & Browne, R. Competition between sexual and parthenogenetic Artemia: temperature and strain effects. *J. Exp. Mar. Biol. Ecol.* **196,** 313–328 (1996).
169. Ramsey, J. & Schemske, D. W. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu. Rev. Ecol. Syst.* **29,** 467–501 (1998).
170. Gross, F. Untersuchungen über die Polyploidie und die Variabilität bei Artemia salina. *Naturwissenschaften* **20,** 962–967 (1932).
171. Bogart, J. P., Elinson, R. P. & Licht, L. E. Temperature and sperm incorporation in polyploid salamanders. *Science* **246,** 1032–1034 (1989).
172. Stebbins Jr, C. L. Variation and evolution in plants. *Var. Evol. Plants* (1950).
173. Lokki, J. & Saura, A. in *Polyploidy* 277–312 (Springer, 1980).
174. Uyeno, T. & Smith, G. R. Tetraploid origin of the karyotype of catostomid fishes. *Science* **175,** 644–646 (1972).
175. David, K. T., Oaks, J. R. & Halanych, K. M. Patterns of gene evolution following duplications and speciations in vertebrates. *PeerJ* **8,** e8813 (2020).
176. Hillebrand, H. On the Generality of the Latitudinal Diversity Gradient. *Am. Nat.* **163,** 192–211 (2004).
177. Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza, M. & Scharn, R. CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods Ecol. Evol.* **10,** 744–751 (2019).
178. Rees, J. A. & Cranston, K. Automated assembly of a reference taxonomy for phylogenetic data synthesis. *Biodivers. Data J.* (2017).
179. Jetz, W. & Pyron, R. A. The interplay of past diversification and evolutionary isolation with present imperilment across the amphibian tree of life. *Nat. Ecol. Evol.* **2,** 850–858 (2018).
180. Rabosky, D. L., Chang, J., Title, P. O., Cowman, P. F., Sallan, L., Friedman, M., Kaschner, K., Garilao, C., Near, T. J., Coll, M. & Alfaro, M. E. An inverse latitudinal gradient in speciation rate for marine fishes. *Nature* **559,** 392–395 (2018).
181. Chesters, D. Construction of a species-level tree of life for the insects and utility in taxonomic profiling. *Syst. Biol.* **66,** 426–439 (2017).
182. Arai, R. *Fish karyotypes: a check list*. (Springer Science & Business Media, 2011).
183. Leggatt, R. A. & Iwama, G. K. Occurrence of polyploidy in the fishes. *Rev. Fish Biol. Fish.* **13,** 237–246 (2003).
184. Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., Lawrence, D. M., Neale, R. B., Rasch, P. J., Vertenstein, M., Worley, P. H., Yang, Z.-L. & Zhang, M. The Community Climate System Model Version 4. *J. Clim.* **24,** 4973–4991 (2011).
185. Hughes, P. D., Ehlers, J. & Gibbard, P. L. in *Quat. Glaciat.-Extent Chronol. Part IV-Closer Look* 1–14 (Elsevier BV, 2011).
186. Jenkins, C. N., Pimm, S. L. & Joppa, L. N. Global patterns of terrestrial vertebrate diversity and conservation. *Proc. Natl. Acad. Sci.* **110,** E2602–E2610 (2013).

187. Jenkins, C. N., Guénard, B., Diamond, S. E., Weiser, M. D. & Dunn, R. R. Conservation implications of divergent global patterns of ant and vertebrate diversity. *Divers. Distrib.* **19,** 1084–1092 (2013).
188. David, K. T. & Halanych, K. M. Spatial proximity between polyploids across South American frog genera. *J. Biogeogr.* **48,** 991–1000 (2021).
189. Sanderson, E. W., Jaiteh, M., Levy, M. A., Redford, K. H., Wannebo, A. V. & Woolmer, G. The Human Footprint and the Last of the Wild: The human footprint is a global map of human influence on the land surface, which suggests that human beings are stewards of nature, whether we like it or not. *BioScience* **52,** 891–904 (2002).
190. Hijmans, R. J. & van Etten, J. raster: Geographic analysis and modeling with raster data. R package version 2.0-12. (2012).
191. Torchiano, M. & Torchiano, M. M. Package 'effsize'. *Package "Effsize* (2020).
192. Morlon, H., Schwilk, D. W., Bryant, J. A., Marquet, P. A., Rebelo, A. G., Tauss, C., Bohannan, B. J. & Green, J. L. Spatial patterns of phylogenetic diversity. *Ecol. Lett.* **14,** 141–149 (2011).
193. Felsenstein, J. Phylogenies and the comparative method. *Am. Nat.* **125,** 1–15 (1985).
194. Dufresne, F. & Hebert, P. D. N. Temperature-related differences in life-history characteristics between diploid and polyploid clones of the Daphnia pulex complex. *Écoscience* **5,** 433–437 (1998).
195. Timofeev, S. F. Bergmann's Principle and Deep-Water Gigantism in Marine Crustaceans. *Biol. Bull. Russ. Acad. Sci.* **28,** 646–650 (2001).
196. Novikova, P. Y., Brennan, I. G., Booker, W., Mahony, M., Doughty, P., Lemmon, A. R., Moriarty Lemmon, E., Roberts, J. D., Yant, L. & Van de Peer, Y. Polyploidy breaks speciation barriers in Australian burrowing frogs Neobatrachus. *PLoS Genet.* **16,** e1008769 (2020).
197. Li, Y., Ren, Z., Shedlock, A. M., Wu, J., Sang, L., Tersing, T., Hasegawa, M., Yonezawa, T. & Zhong, Y. High altitude adaptation of the schizothoracine fishes (Cyprinidae) revealed by the mitochondrial genome analyses. *Gene* **517,** 169–178 (2013).
198. Fujita, M. K., Singhal, S., Brunes, T. O. & Maldonado, J. A. Evolutionary dynamics and consequences of parthenogenesis in vertebrates. *Annu. Rev. Ecol. Evol. Syst.* **51,** 191–214 (2020).
199. Anderson, D. & Burnham, K. Model selection and multi-model inference. *Second NY Springer-Verl.* **63,** 10 (2004).
200. Napier, J. D., Grabowski, P. P., Lovell, J. T., Bonnette, J., Mamidi, S., Gomez-Hughes, M. J., VanWallendael, A., Weng, X., Handley, L. H., Kim, M. K., Boe, A. R., Fay, P. A., Fritschi, F. B., Jastrow, J. D., Lloyd-Reilley, J., Lowry, D. B., Matamala, R., Mitchell, R. B., Rouquette, F. M., Wu, Y., Webber, J., Jones, T., Barry, K., Grimwood, J., Schmutz, J. & Juenger, T. E. A generalist–specialist trade-off between switchgrass cytotypes impacts climate adaptation and geographic range. *Proc. Natl. Acad. Sci.* **119,** e2118879119 (2022).
201. Müller, B. & Grossniklaus, U. Model organisms—a historical perspective. *J. Proteomics* **73,** 2054–2063 (2010).
202. Goldstein, B. & King, N. The future of cell biology: emerging model organisms. *Trends Cell Biol.* **26,** 818–824 (2016).
203. Dunn, C. W. & Ryan, J. F. The evolution of animal genomes. *Curr. Opin. Genet. Dev.* **35,** 25–32 (2015).
204. Davis, R. H. The age of model organisms. *Nat. Rev. Genet.* **5,** 69 (2004).

205. Bolker, J. Model organisms: There's more to life than rats and flies. *Nature* **491,** 31 (2012).
206. Warkany, J. Why I doubted that thalidomide was the cause of the epidemic of limb defects of 1959 to 1961. *Teratology* **38,** 217–219 (1988).
207. Xu, D., Nishimura, T., Nishimura, S., Zhang, H., Zheng, M., Guo, Y.-Y., Masek, M., Michie, S. A., Glenn, J. & Peltz, G. Fialuridine induces acute liver failure in chimeric TK-NOG mice: a model for detecting hepatic drug toxicity prior to human testing. *PLoS Med.* **11,** e1001628 (2014).
208. Schnabel, J. Neuroscience: standard model. *Nat. News* **454,** 682–685 (2008).
209. Geerts, H. Of mice and men. *CNS Drugs* **23,** 915–926 (2009).
210. Seok, J., Warren, H. S., Cuenca, A. G., Mindrinos, M. N., Baker, H. V., Xu, W., Richards, D. R., McDonald-Smith, G. P., Gao, H. & Hennessy, L. Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc. Natl. Acad. Sci.* **110,** 3507–3512 (2013).
211. Baker, D. & Amor, S. Mouse models of multiple sclerosis: lost in translation? *Curr. Pharm. Des.* **21,** 2440–2452 (2015).
212. Perlman, R. L. Mouse models of human diseaseAn evolutionary perspective. *Evol. Med. Public Health* **2016,** 170–176 (2016).
213. Crollius, H. R., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P. & Quétier, F. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nat. Genet.* **25,** 235 (2000).
214. Martín-Durán, J. M., Pang, K., Børve, A., Lê, H. S., Furu, A., Cannon, J. T., Jondelius, U. & Hejnol, A. Convergent evolution of bilaterian nerve cords. *Nature* **553,** 45 (2018).
215. Dunn, C. W., Giribet, G., Edgecombe, G. D. & Hejnol, A. Animal phylogeny and its evolutionary implications. *Annu. Rev. Ecol. Evol. Syst.* **45,** 371–395 (2014).
216. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
217. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39,** W29–W37 (2011).
218. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16,** 157 (2015).
219. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20,** 1–14 (2019).
220. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49,** D480–D489 (2021).
221. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S. & Eppig, J. T. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).
222. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49,** D325–D334 (2021).
223. Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., Hub, A. & Group, W. P. W. AmiGO: online access to ontology and annotation data. *Bioinformatics* **25,** 288–289 (2009).
224. Dunn, C. W., Leys, S. P. & Haddock, S. H. The hidden biology of sponges and ctenophores. *Trends Ecol. Evol.* **30,** 282–291 (2015).

225. David, K. T., Wilson, A. E. & Halanych, K. M. Sequencing disparity in the genomic era. *Mol. Biol. Evol.* **36,** 1624–1627 (2019).
226. Tassia, M. G., Whelan, N. V. & Halanych, K. M. Toll-like receptor pathway evolution in deuterostomes. *Proc. Natl. Acad. Sci.* 201617722 (2017).
227. Cavalieri, V. & Spinelli, G. Early asymmetric cues triggering the dorsal/ventral gene regulatory network of the sea urchin embryo. *eLife* **3,** e04664 (2014).
228. Connahs, H., Tlili, S., van Creij, J., Loo, T. Y. J., Banerjee, T. D., Saunders, T. E. & Monteiro, A. Activation of butterfly eyespots by Distal-less is consistent with a reaction-diffusion process. *Development* **146,** dev169367 (2019).
229. Smith, S. D., Pennell, M. W., Dunn, C. W. & Edwards, S. V. Phylogenetics is the New Genetics (for Most of Biodiversity). *Trends Ecol. Evol.* **35,** 415–425 (2020).
230. Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A. & Telenti, A. A primer on deep learning in genomics. *Nat. Genet.* **51,** 12–18 (2019).
231. Bengio, Y., Courville, A. C. & Vincent, P. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR Abs12065538* **1,** 2012 (2012).
232. Feldbauer, R., Gosch, L., Lüftinger, L., Hyden, P., Flexer, A. & Rattei, T. DeepNOG: fast and accurate protein orthologous group assignment. *Bioinformatics* **36,** 5304–5312 (2020).
233. Seo, S., Oh, M., Park, Y. & Kim, S. DeepFam: deep learning based alignment-free method for protein family modeling and prediction. *Bioinformatics* **34,** i254–i262 (2018).
234. Sureyya Rifaioglu, A., Doğan, T., Jesus Martin, M., Cetin-Atalay, R. & Atalay, V. DEEPred: Automated Protein Function Prediction with Multi-task Feed-forward Deep Neural Networks. *Sci. Rep.* **9,** 7344 (2019).
235. Bileschi, M. L., Belanger, D., Bryant, D., Sanderson, T., Carter, B., Sculley, D., DePristo, M. A. & Colwell, L. J. Using Deep Learning to Annotate the Protein Universe. *bioRxiv* 626507 (2019). doi:10.1101/626507
236. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *ArXiv13126114 Cs Stat* (2014). at <http://arxiv.org/abs/1312.6114>
237. Dosovitskiy, A., Springenberg, J. T., Tatarchenko, M. & Brox, T. Learning to Generate Chairs, Tables and Cars with Convolutional Networks. *ArXiv14115928 Cs* (2017). at <http://arxiv.org/abs/1411.5928>
238. Razavi, A., Oord, A. van den & Vinyals, O. Generating Diverse High-Fidelity Images with VQ-VAE-2. *ArXiv190600446 Cs Stat* (2019). at <http://arxiv.org/abs/1906.00446>
239. Montserrat, D. M., Bustamante, C. & Ioannidis, A. Class-conditional vae-gan for local-ancestry simulation. *ArXiv Prepr. ArXiv191113220* (2019).
240. Battey, C. J., Coffing, G. C. & Kern, A. D. Visualizing population structure with variational autoencoders. *G3 GenesGenomesGenetics* **11,** (2021).
241. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15,** 816–822 (2018).
242. Kramer, M. A. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.* **37,** 233–243 (1991).
243. Doersch, C. Tutorial on variational autoencoders. *ArXiv Prepr. ArXiv160605908* (2016).
244. Derkarabetian, S., Castillo, S., Koo, P. K., Ovchinnikov, S. & Hedin, M. A demonstration of unsupervised machine learning in species delimitation. *Mol. Phylogenet. Evol.* **139,** 106562 (2019).
245. Oord, A. van den, Vinyals, O. & Kavukcuoglu, K. Neural discrete representation learning. *ArXiv Prepr. ArXiv171100937* (2017).

246. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N. & Antiga, L. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32,** 8026–8037 (2019).
247. Bengio, Y., Léonard, N. & Courville, A. Estimating or propagating gradients through stochastic neurons for conditional computation. *ArXiv Prepr. ArXiv13083432* (2013).
248. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *ArXiv Prepr. ArXiv14126980* (2014).
249. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).
250. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17,** 282–283 (2001).
251. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28,** 3150–3152 (2012).
252. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35,** 1026–1028 (2017).
253. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18,** 366–368 (2021).
254. Vinh, N. X., Epps, J. & Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **11,** 2837–2854 (2010).
255. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27,** 379–423 (1948).
256. Kreer, J. A question of terminology. *IRE Trans. Inf. Theory* **3,** 208–208 (1957).
257. Klopfenstein, D. V., Zhang, L., Pedersen, B. S., Ramírez, F., Vesztrocy, A. W., Naldi, A., Mungall, C. J., Yunes, J. M., Botvinnik, O. & Weigel, M. GOATOOLS: A Python library for Gene Ontology analyses. *Sci. Rep.* **8,** 1–17 (2018).
258. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57,** 289–300 (1995).
259. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv Prepr. ArXiv180203426* (2018).
260. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9,** (2008).
261. Chari, T., Banerjee, J. & Pachter, L. The specious art of single-cell genomics. *bioRxiv* (2021).
262. Tassia, M. G., David, K. T., Townsend, J. P. & Halanych, K. M. TIAMMAt: Leveraging biodiversity to revise protein domain models, evidence from innate immunity. *Mol. Biol. Evol.* (2021).