bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Why do species get a thin slice of π? Revisiting Lewontin's Paradox of Variation

Vince Buffalo

University of Oregon, Institute of Ecology and Evolution
Eugene, Oregon
vsbuffalo@gmail.com

February 3, 2021

Abstract

Under neutral theory, the level of polymorphism in an equilibrium population is expected to increase with population size. However, observed levels of diversity across metazoans vary only two orders of magnitude, while census population sizes (N_c) are expected to vary over several. This unexpectedly narrow range of diversity is a longstanding enigma in evolutionary genetics known as Lewontin's Paradox of Variation (1974). Since Lewontin's observation, it has been argued that selection constrains diversity across species, yet tests of this hypothesis seem to fall short of explaining the orders-of-magnitude reduction in diversity observed in nature. In this work, I revisit Lewontin's Paradox and assess whether current models of linked selection are likely to constrain diversity to this extent. To quantify the discrepancy between pairwise diversity and census population sizes across species, I combine genetic data from 172 metazoan taxa with estimates of census sizes from geographic occurrence data and population densities estimated from body mass. Next, I fit the relationship between previously-published estimates of genomic diversity and these approximate census sizes to quantify Lewontin's Paradox. While previous across-taxa population genetic studies have avoided accounting for phylogenetic non-independence, I use phylogenetic comparative methods to investigate the diversity census size relationship, estimate phylogenetic signal, and explore how diversity changes along the phylogeny. I consider whether the reduction in diversity predicted by models of recurrent hitchhiking and background selection could explain the observed pattern of diversity across species. Since the impact of linked selection is mediated by recombination map length, I also investigate how map lengths vary with census sizes. I find species with large census sizes have shorter map lengths, leading these species to experience greater reductions in diversity due to linked selection. Even after using high estimates of the strength of sweeps and background selection, I find linked selection likely cannot explain the shortfall between predicted and observed diversity levels across metazoan species. Furthermore, the predicted diversity under linked selection does not fit the observed diversity–census-size relationship, implying that processes other than background selection and recurrent hitchhiking must be limiting diversity.

A longstanding mystery in evolutionary genetics is that the observed levels of genetic variation across sexual species are confined to an unexpectedly narrow range. Under neutral theory, the average number of nucleotide differences between lineages (pairwise diversity, π) is determined by the balance of new mutations and their loss by genetic drift (Kimura and Crow 1964; Malécot 1948; Wright 1931). In particular, the expected diversity at neutral sites in a panmictic population of N_c

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

diploids is expected to be $\pi \approx 4N_c\mu$, where μ is per generation mutation rate. Given that metazoan germline mutation rates only differ 10-fold (10^{-8}–10^{-9}, Kondrashov and Kondrashov 2010; Lynch 2010), and census sizes vary over several orders of magnitude, one would expected under neutral theory that heterozygosity should also vary over several orders of magnitude. However, early allozyme surveys revealed that heterozygosity levels across a wide range of species varied just an order of magnitude (Lewontin 1974, p. 208); this anomaly is known as Lewontin's "Paradox of Variation". With modern sequencing-based estimates of π across taxa ranging over only three orders of magnitude (0.01–10%, Leffler et al. 2012), Lewontin's paradox has persisted unresolved through the genomics era.

From the beginning, explanations for Lewontin's Paradox have been framed in terms of the neutralist–selectionist controversy (Gillespie 1991, 2001; Kimura 1984; Lewontin 1974). The neutralist view is that beneficial alleles are sufficiently rare and deleterious alleles removed sufficiently quickly, that levels of genetic diversity are shaped predominantly by genetic drift and mutation (Kimura 1984). Specifically, *non-selective* processes decouple the effective population size implied by observed levels of diversity $\widehat{\pi}$, $\widetilde{N}_e = \widehat{\pi}/4\mu$, from the census size, N_c. By contrast, the selectionist view is that the direct and indirect effects of linked selection suppress diversity levels across taxa, specifically because the impact of linked selection is greater in large populations. Undoubtedly, these opposing views represent a false dichotomy, as population genomic studies have uncovered complex demographic histories that impact diversity within a species (e.g. Palkopoulou et al. 2015; Zhao et al. 2013), as well as evidence that selection depresses genome-wide diversity (e.g. Aguade et al. 1989; Begun and Aquadro 1992; Elyashiv et al. 2016; McVicker et al. 2009).

Possible Explanations of Lewontin's Paradox

A resolution of Lewontin's Paradox would involve a mechanistic description and quantification of the evolutionary processes that prevent diversity from scaling with census sizes across species. This would necessarily connect to the broader literature on the empirical relationship between diversity and population size (Frankham 1996; Nei and Graur 1984; Soulé 1976), and the ecological and life history correlates of genetic diversity (Nevo 1978; Nevo et al. 1984; Powell 1975). Three categories of processes stand out as potentially capable of decoupling census sizes from diversity: non-equilibrium demography, variance and skew in reproductive success, and selective processes.

It has long been appreciated that effective population sizes are typically less than census population sizes, tracing back to early debates between R.A. Fisher and Sewall Wright (Fisher and Ford 1947; Wright 1948). Possible causes of this divergence between effective and census population sizes include demographic history (e.g. population bottlenecks), extinction and recolonization dynamics, or the breeding structure of populations (e.g. the variance in reproductive success and population substructure). Early explanations for Lewontin's Paradox suggested bottlenecks during the last glacial maximum severely reduced population sizes (Kimura 1984; Nei and Graur 1984; Ohta and Kimura 1973), and emphasized that large populations recover to equilibrium diversity levels more slowly (Nei and Graur 1984, Kimura 1984 p. 203-204). Another explanation is that cosmopolitan species repeatedly endure extinction and recolonization events, which reduces effective population size (Maruyama and Kimura 1980; Slatkin 1977).

While chance demographic events like bottlenecks and recent expansions have long-term impacts on diversity (equilibrium is reached on the order of size of the population), characteristics of the breeding structure such as high variance (V_w) or skew in reproductive success also suppress diversity below the levels predicted by the census size (Wright 1938). In species like marine animals, females

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

are highly fecund, and dispersing larvae face extremely low survivorship, leading to high variance in reproductive success (Hauser and Carvalho 2008; Hedgecock and Pudovkin 2011; Waples et al. 2018, 2013). Such "sweepstakes" reproductive systems can lead to remarkably small ratios of effective population size to census population size (e.g. N_e/N_c can range from 10^{-6}–10^{-2}), since $N_e/N \approx 1/V_w$ (Hedgecock 1994; Nunney 1993, 1996; Wright 1938), and require multiple-merger coalescent processes to describe their genealogies (Eldon and Wakeley 2006). Overall, these reproductive systems diminish the diversity in many species, but seem unlikely to explain Lewontin's Paradox broadly across metazoans.

Alternatively, selective processes, and in particular the indirect effects of selection on linked neutral variation, could explain the observed narrow range of diversity. The earliest mathematical model of hitchhiking was proffered as a solution to Lewontin's Paradox (Maynard Smith and Haigh 1974). Since, empirical observations have demonstrated that linked selection shapes patterns of genome-wide diversity, as evidenced by the correlation between recombination and diversity in a variety of species (Aguade et al. 1989; Begun and Aquadro 1992; Cai et al. 2009; Cutter and Payseur 2003; Stephan and Langley 1998). Theoretic work to explain this pattern has considered diversity under a steady influx of new beneficial mutations (recurrent hitchhiking; Stephan 1995; Stephan et al. 1992), and purifying selection against new deleterious mutations (background selection, BGS; Charlesworth et al. 1993; Hudson and Kaplan 1995; Hudson and Kaplan 1994; Nordborg et al. 1996). Indeed, empirical work indicates background selection diminishes diversity around genic regions in a variety of species (Charlesworth 1996; Hernandez et al. 2011; McVicker et al. 2009), and now efforts have shifted towards teasing apart the effects of positive and negative selection on genomic diversity (Elyashiv et al. 2016).

An important class of theoretic selection models pertaining to Lewontin's Paradox are recurrent hitchhiking models that decouple diversity from the census population size. These models predict diversity when strongly selected beneficial mutations regularly enter and sweep through the population, trapping lineages and forcing them to coalesce (Gillespie 2000; Kaplan et al. 1989). In general, decoupling occurs under these hitchhiking models when the rate of coalescence due to selection is much greater than the rate of neutral coalescence (Coop and Ralph 2012, equation 22). Other selection models cannot alone decouple diversity from population size, *ceteris paribus*. For example, the reduction in diversity predicted under background selection and polygenic fitness variation is a proportion reduction in population size, mediated by the total recombination map length (Charlesworth et al. 1993; Nicolaisen and Desai 2012; Nordborg et al. 1996; Robertson 1961; Santiago and Caballero 1995, 1998).

Recent Approaches Towards Solving Lewontin's Paradox

Recently, Corbett-Detig et al. (2015) used population genomic data to estimate the reduction in diversity due to background selection and hitchhiking across 40 species, and showed the impact of selection increases with two proxies of census population size, species range and with body size. These authors argued this is evidence that selection could explain Lewontin's Paradox; however, in a re-analysis, Coop (2016) demonstrated that the observed scale of these reductions is insufficient to explain the orders-of-magnitude shortfall between observed and expected levels of diversity across species. Other recent work has found that certain life history characteristics related to parental investment, such as propagule size, are good predictors diversity in animals (Chen et al. 2017; Romiguier et al. 2014). Nevertheless, while these diversity correlates are important clues, they do not propose a mechanism by which these traits act to constrain diversity within a few orders of

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

magnitude.

Here, I revisit Lewontin's Paradox by integrating a variety of data sets and assessing the predicted reductions in diversity under different selection models. Prior surveys of genetic diversity either lacked census population size estimates, used allozyme-based measures of heterozygosity, or included fewer species. To address these shortcomings, I first estimate census sizes by combining predictions of population density based on body size with ranges estimated from geographic occurrence data. Using these estimates, I quantify the relationship between census size and previously-published genomic diversity estimates across 172 metazoan taxa within nine phyla, which provides a sense of the scale of the divergence between π and N_c that leads to Lewontin's Paradox.

Past work looking at the relationship between π and N_c has largely ignored phylogenetic non-independence across taxa (Felsenstein 1985). To address this shortcoming, I account for phylogenetic non-independence across taxa using a synthetic time-calibrated phylogeny and phylogenetic comparative methods (PCMs). Moreover, Lynch (2011) has argued that since coalescent times are much less than divergence times, considering phylogenetic non-independence is unnecessary for traits like effective population size. Using PCMs, I test this conjecture by estimating phylogenetic signal in the diversity census size relationship, and investigating how these traits evolve along the phylogeny.

Finally, I explore whether the predicted reductions of diversity under background selection and recurrent hitchhiking are sufficiently strong enough to resolve Lewontin's Paradox. These predicted reductions in diversity across species are generously estimated using selection strength parameters from *Drosophila melanogaster*, a species known to be strongly affected by linked selection. Given the effects of linked selection are mediated by recombination map length, I investigate how recombination map lengths vary with census population size using data from a previously-published survey (Stapley et al. 2017). I find map lengths are typically shorter in large–census-size species, increasing the effects of linked selection in these species, which might further decouple diversity from census size. Still, I find the combined impact of these selection models with available parameter estimates falls short in explaining Lewontin's Paradox, and discuss future avenues through which the Paradox of Variation could be fully resolved.

Results

Estimates of Census Population Size

A major impediment in quantifying Lewontin's Paradox has been estimating census population sizes across many taxa, especially for extremely abundant, cosmopolitan species that define the upper limit of ranges. Previous work has surveyed the literature for census size estimates (Frankham 1996; Nei and Graur 1984; Soulé 1976), or used range and body size or qualitative categories as proxies for census size (Corbett-Detig et al. 2015; Leffler et al. 2012). To quantify the relationship between genomic estimates of diversity and census population sizes, I first approximate census population sizes for 172 metazoan taxa (Figure 1). My approach predicts population densities from body sizes using a previously-observed linear relationship that holds across metazoans (Supplementary Figure S8; Damuth 1981, 1987). Then, from geographic occurrence data, I estimate range sizes. Finally, I estimate population size as the product of these predicted densities and range estimates (see Methods: *Macroecological Estimates of Population Size*). Note that the relationship between population density and body size is driven by energy budgets, and thus reflects macroecological equilibria

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

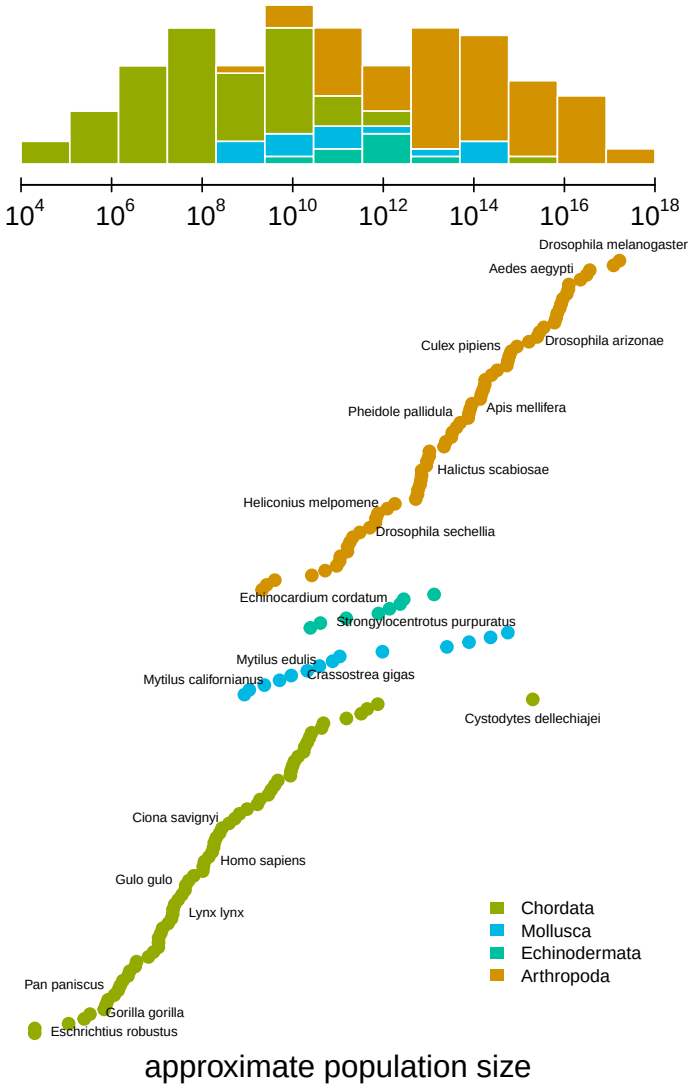

Figure 1: The distribution of approximate census population sizes estimated by this study. Some phyla containing few species were excluded for clarity.

(Damuth 1987); consequently, population sizes for taxa like humans and their domesticated species are underestimated. While these methods to estimate census size are crude and approximate, they can be efficiently calculated for numerous taxa and are sufficient to estimate the scale of Lewontin's Paradox.

Quantifying Lewontin's Paradox

To determine which ecological or evolutionary processes could decouple diversity from census population size, we first need to quantify this relationship across a wide variety of taxa. Previous work has found there is a significant relationship between heterozygosity and the logarithm of population size, but these studies relied on heterozygosity measured from allozyme data (Frankham 1996; Nei

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

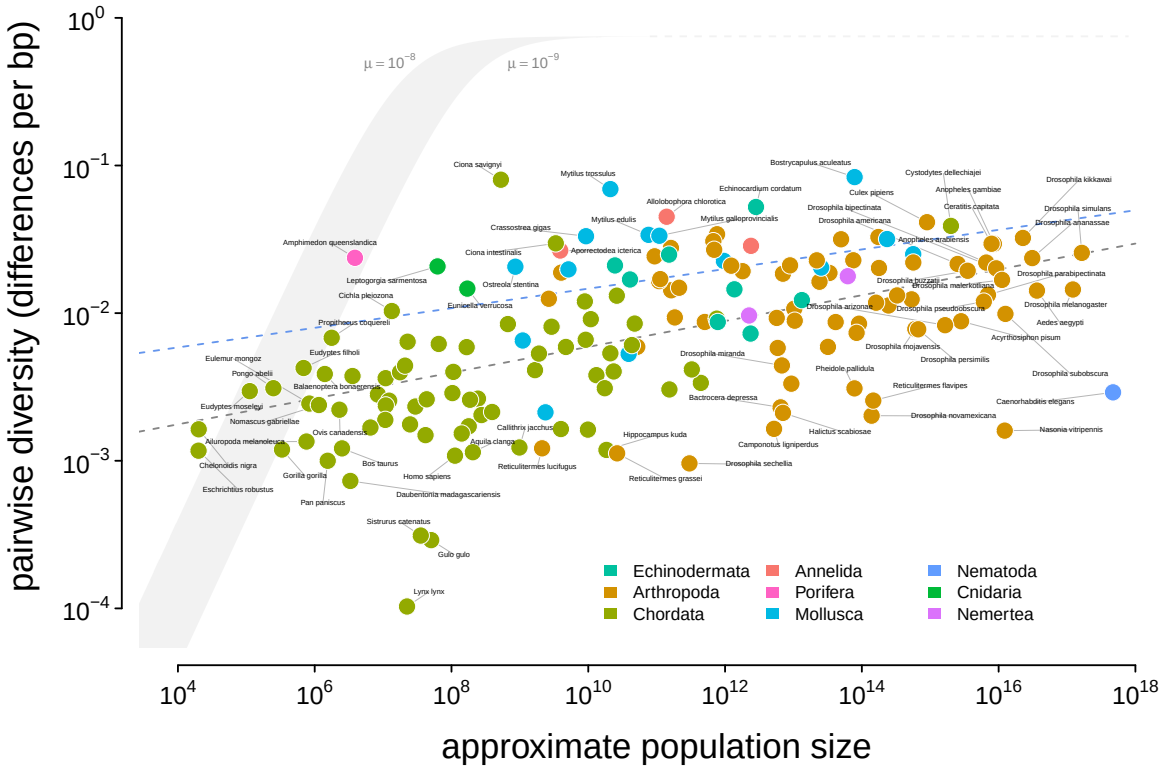

Figure 2: An illustration of Lewontin's Paradox of Variation. Pairwise diversity (data from Leffler et al. 2012, Corbett-Detig et al. 2015, and Romiguier et al. 2014), which varies around three orders of magnitude, shows a weak relationship with approximate population size, which varies over 12 orders of magnitude. The shaded curve shows the range of expected neutral diversity if N_e were to equal N_c under the four-alleles model, $\log_{10}(\pi) = \log_{10}(\theta) - \log_{10}(1 + {}^{4\theta}/_{3})$ where $\theta = 4N_c\mu$, for two mutation rates, $\mu = 10^{-8}$ and $\mu = 10^{-9}$, and the light gray dashed line represents the maximum pairwise diversity under the four alleles model. The dark gray dashed line is the OLS regression fit, and the blue dashed line is the regression fit using a phylogenetic mixed-effects model. Points are colored by phylum.

and Graur 1984; Soulé 1976). Here, I confirm these findings using pairwise diversity estimates from genomic sequence data and the estimated census sizes (Figure 2). The pairwise diversity estimates are from three sources: Leffler et al. (2012), Corbett-Detig et al. (2015), and Romiguier et al. (2014), and are predominantly from either synonymous or non-coding DNA (see Methods: *Diversity and Map Length Data*). Overall, an ordinary least squares (OLS) relationship on a log-log scale fits the data well (Figure 2). The OLS slope estimate is significant and implies an increase in diversity of 0.09 differences per basepair for every order of magnitude census size grows (95% confidence interval $[0.08, 0.12]$; see also the OLS fit per-phyla, Supplementary Figure S2).

Notably, this relationship has few outliers and is relatively homoscedastic. This is in part because of the log-log scale, in contrast to previous work (Nei and Graur 1984; Soulé 1976); see Supplementary Figure S1 for a version on a log-linear scale. However, it is noteworthy that few taxa have diversity estimates below $10^{-3.5}$ differences per basepair. Those that do, lynx (*Lynx Lynx*), wolverine (*Gulo gulo*), and Massasauga rattlesnake (*Sistrurus catenatus*) face habitat fragmentation

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

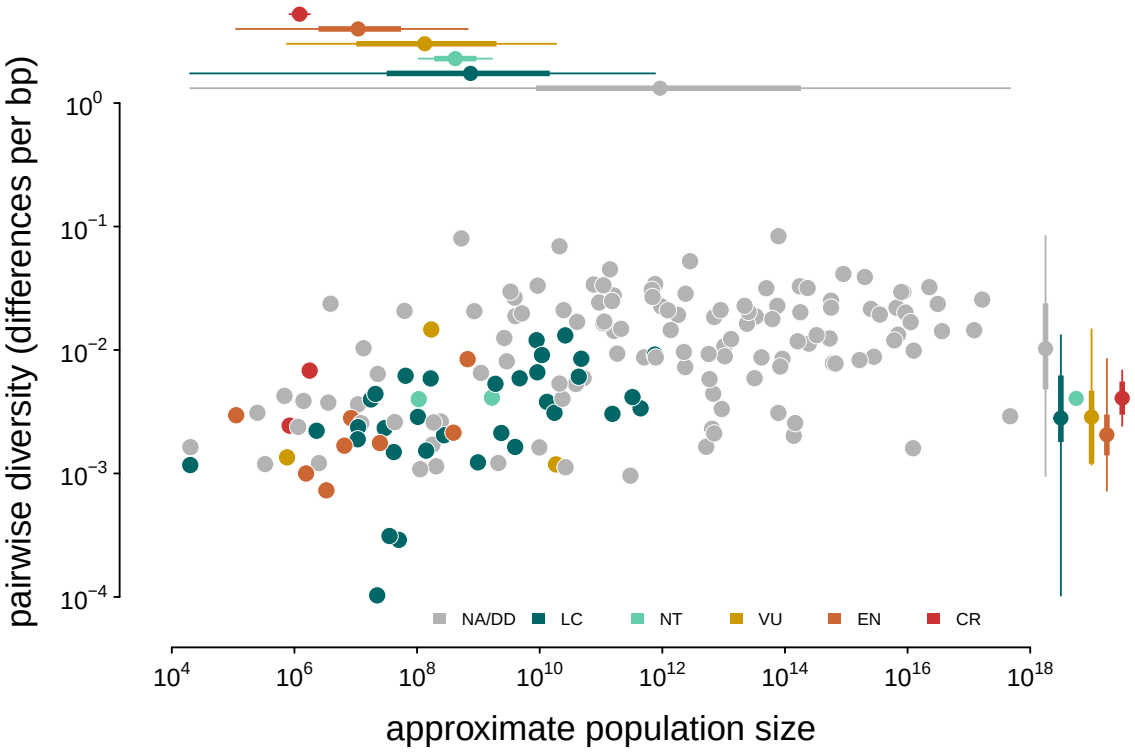

Figure 3: A version of Figure 2 with points colored by their IUCN Red List conservation status. Margin boxplots show the diversity and population size ranges (thin lines) and interquartile ranges (thick lines) for each category. NA/DD indicates no IUCN Red List entry, or Red List status Data Deficient; LC is Least Concern, NT is Near Threatened, VU is Vulnerable, EN is Endangered, and CR is Critically Endangered.

and declining population sizes. These three species are all in the IUCN Red List, but are listed as least concern (though their presence in the Red List indicates they are of conservation interest). I also investigated the relationship between species' IUCN Red List categories (an ordinal scale of how threatened a species is) and both diversity and population size, finding that species categorized as more threatened have both smaller population sizes and reduced diversity, compared to non-threatened species (Figure 3) consistent with past work (Spielman et al. 2004). A linear model of diversity regressed on population size has lower AIC when the IUCN Red List categories are included, and the estimates of the effect of IUCN status are all negative on diversity, though not all are significant in part because some categories have three or fewer species (Supplementary Table S1).

Phylogenetic Non-Independence and the Population Size Diversity Relationship

In quantifying Lewontin's Paradox, I have additionally fit some simple trait evolution models that account for phylogenetic non-independence, investigated whether there is a signal of phylogenetic non-independence, estimated the continuous trait values on the phylogeny, and assessed how diversity and population size evolve. Prior population genetic comparative studies have lacked

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

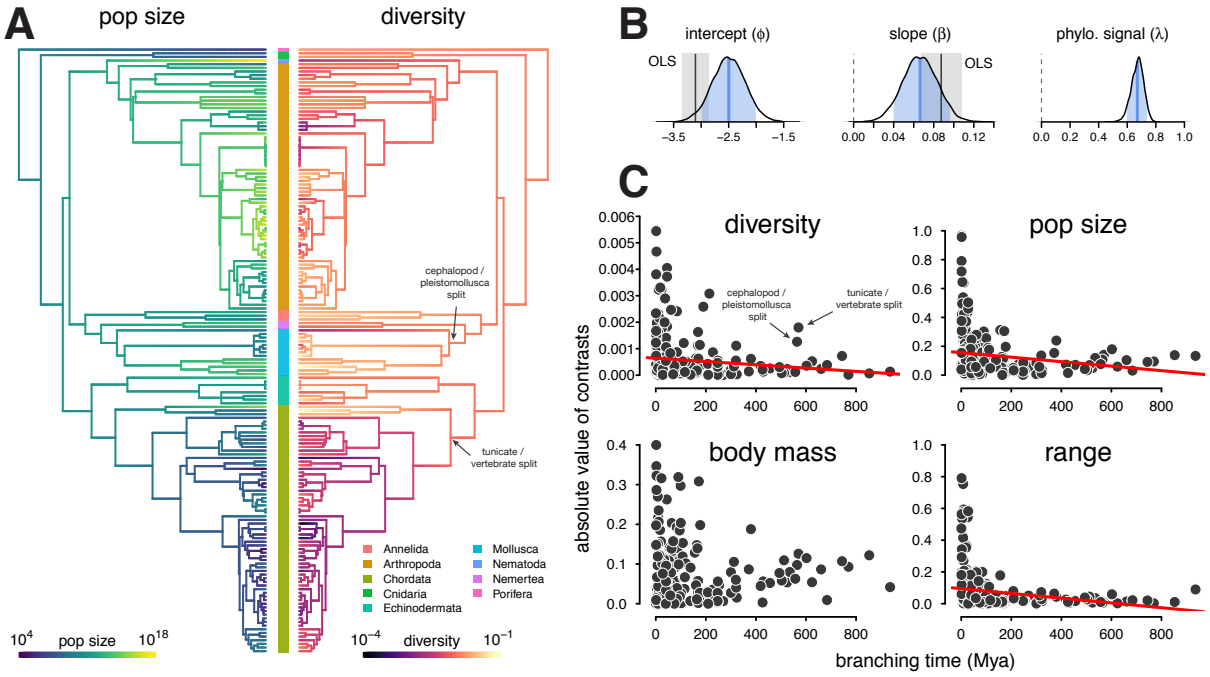

Figure 4: (A) The ancestral continuous trait estimates for the population size and diversity (differences per bp, log scaled) across the phylogeny of 166 taxa. The phyla of the tips are indicated by the color bar in the center. (B) The posterior distributions of the intercept, slope, and phylogenetic signal (λ, Villemereuil and Nakagawa 2014) of the phylogenetic mixed-effects model of diversity and population size (log scaled). Also shown are the 90% credible interval (light blue shading), posterior mean (blue line), OLS estimate (gray solid line), and bootstrap OLS confidence intervals (light gray shading). (C) The node-height tests of diversity, population size, and the two components of the population size estimates, body mass, and range (all traits on log scale before contrast was calculated). Each point shows the standardized phylogenetic independent contrast and branching time for a pair of lineages. Red lines are robust regression estimates (and are only shown for statistically significant relationships at the $\alpha = 0.05$ level). Note that some outlier pairs with very high phylogenetic independent contrasts were excluded (in all cases, these outliers were in the genus *Drosophila*).

time-calibrated phylogenies and assumed unit branch lengths (Whitney and Garland 2010), a shortcoming that has drawn criticism (Lynch 2011). Here, I use a synthetic time-calibrated phylogeny created from the DateLife project (O'Meara et al. 2020) to account for shared phylogenetic history (see Methods: *Phylogenetic Comparative Methods*).

Using a phylogenetic mixed-effects model (Hadfield and Nakagawa 2010; Lynch 1991; Villemereuil and Nakagawa 2014) implemented in Stan (Carpenter et al. 2017; Stan Development Team 2020), I estimated the linear relationship between diversity and population size (on a log-log scale) accounting for phylogeny, for the 166 taxa with non-missing data and present in the synthetic chronogram. As with the non-phylogenetic regression, this relationship was positive and significant (95% credible interval $0.04, 0.11$), though somewhat attenuated compared to the OLS estimates (Figure 2B). Since the population size estimates are based on range and body mass, they are essentially a composite trait; fitting phylogenetic mixed-effects models separately on body mass and range indicates these have significant negative and positive effects, respectively (Supplementary

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure S3).

With the phylogenetic mixed-effects model, I also estimated the variance of the phylogenetic effect (σ_p^2) and the residual variance (σ_r^2), which can be used to estimate a measure of the phylogenetic signal, $\lambda = \sigma_p^2/(\sigma_p^2+\sigma_r^2)$ (Lynch 1991; Villemereuil and Nakagawa 2014; see Freckleton et al. 2002 for a comparison to Pagel's λ). If the relationship between diversity and population size was free of shared phylogenetic history, $\lambda = 0$ and all the variance could be explained by evolution on the tips; this is analogous to Lynch's conjecture that coalescent times should be free of phylogenetic signal (2011). In the relationship between population size and diversity, the posterior mean of $\lambda = 0.67$ (90% credible interval $[0.59, 0.75]$) indicates that the majority of the variance perhaps might be due to shared phylogenetic history (Figure 4B).

A closer visual inspection of the estimated ancestral continuous values for diversity and population size on the phylogeny indicates the high phylogenetic signal seems to be driven in part by chordates having low diversity and small population sizes compared to non-chordates (Figure 4A). This suggests Gillespie's (1991) earlier critique that the π–N_c relationship was driven by chordate-arthropod differences may be valid. This problem resembles Felsenstein's worst-case scenario (Felsenstein 1985; Uyeda et al. 2018), where a singular event on a lineage separating two clades generates a spurious association between two traits. To further investigate whether clade-level differences dominated the relationship between diversity and population size, I fit phylogenetic mixed-effects models to phyla-level subsets of the data for clades with sufficient sample sizes (see Methods: *Phylogenetic Comparative Methods*). This analysis shows a significant positive relationship between diversity and population size in arthropods, and positive weak relationships in molluscs and chordates (Supplementary Figure S12). Each of the 90% credible intervals for slope overlap, indicating the relationship between π and N_c is similar across these clades.

One limitation of the phylogenetic mixed-effects models employed here is that they assume traits evolve under constant-rate Brownian motion. To test this assumption, I performed node-height tests (Freckleton and Harvey 2006). Node-height tests regress the absolute values of the standardized contrasts between lineages against the branching time (since present) of these lineages. Under Brownian Motion (BM), standardized contrasts are estimates of the rate of character evolution (Felsenstein 1985); if a trait evolves under constant rate BM, this relationship should be flat. For both diversity and population size, node-height tests indicate a significant increase in the rate of evolution towards the present (robust regression p-values 0.028 and 0.00070 respectively; Figure 4C). Considering the constituents of the population size estimate, range and body mass, separately, range but not body mass shows a significant increase (p-value 1.9×10^{-7}) in rate towards the present.

Interestingly, the diversity node-height test reveals two rate shifts at deeper splits (Figure 4C, top left) around 570 Mya. These nodes represent the branches between tunicates and vertebrates in chordates, and cephalopods and pleistomollusca (bivalves and gastropods) in molluscs. While the cephalopod-pleistomollusca split outlier may be an artifact of having a single cephalopod (*Sepia officinalis*) in the phylogeny, the tunicate-vertebrate split outlier is driven by the low diversity of vertebrates and the previously-documented exceptionally high diversity of tunicates (sea squirts; Nydam and Harrison 2010; Small et al. 2007). This deep node representing a rate shift in diversity could reflect a change in either effective population size or mutation rate, and there is some evidence of both in this genus *Ciona* (Small et al. 2007; Tsagkogeorga et al. 2012). Neither of these deep rate shifts in diversity is mirrored in the population size node-height test (Figure 4C, top right). Rather, it appears a trait impacting diversity but not census size (e.g. mutation rate or offspring distributions) has experienced a shift on the lineage separating tunicates and vertebrates. At nearly

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

600 Mya, these deep nodes illustrate a counterexample to Lynch's claim that expected coalescent times do not share phylogenetic history because they are less than divergence times.

Finally, an important caveat is the increase in rate towards the tips could be caused by measurement noise. Inspecting the lineage pairs that lead to this increase in rate towards the tips indicates these represent plausible rate shifts, e.g. between cosmopolitan and endemic sister species like *Drosophila simulans* and *Drosophila sechellia*; however, ruling out measurement noise entirely as an explanation would involve considering the uncertainty of diversity and population size estimates.

Assessing the Impact of Linked Selection on Diversity Across Taxa

The above analyses reemphasize the drastic shortfall of diversity levels as compared to census sizes. Linked selection has been proposed as the mechanism that acts to reduce diversity levels from what we would expect given census sizes (Corbett-Detig et al. 2015; Gillespie 2000; Maynard Smith and Haigh 1974). Here, I test this hypothesis by estimating the scale of diversity reductions expected under background selection and recurrent hitchhiking, and compare these to the observed relationship between π and N_c.

I quantify the effect of linked selection on diversity as the ratio of observed diversity (π) to the estimated diversity in the absence of linked selection (π_0), $R = {}^{\pi}/_{\pi_0}$. There are two difficulties in evaluating whether linked selection could resolve Lewontin's Paradox. The first difficulty is that π_0 is unobserved. Previous work has estimated π_0 using methods that exploit the spatial heterogeneity in recombination and functional density across the genome to fit linked selection models that incorporate both hitchhiking and background selection (Corbett-Detig et al. 2015; Elyashiv et al. 2016). The second difficulty is understanding how R varies across taxa, since we lack estimates of critical model parameters for most species. Still, I can address a key question: if diversity levels were determined by census sizes ($\pi_0 = 4N_c\mu$), are the combined effects of background selection and recurrent hitchhiking sufficient to reduce diversity to observed levels? Furthermore, does the relationship between census size and predicted diversity under linked selection across species, $\pi_{BGS+HH} = R\pi_0$, match the observed relationship in Figure 2?

Since we lack estimates of key linked selection parameters across species, I generously parameterize the hitchhiking and BGS models using estimates from *Drosophila melanogaster*, a species known to be strongly affected by linked selection (Sella et al. 2009). Under a generalized model of hitchhiking and background selection (Coop and Ralph 2012; Elyashiv et al. 2016) and assuming $N_e = N_c$, expected diversity is

$$\pi_{\text{BGS+HH}} \approx \frac{\theta}{{}^{1}/_{B(U,L)} + 2N_c S(\gamma, L, J)} \tag{1}$$

where $\theta = 4N_c\mu$, $B(U, L)$ is the effect of background selection, and $S(\gamma, L, J)$ is the rate of coalescence caused by sweeps (c.f. Elyashiv et al. 2016, equation 1, Coop and Ralph 2012 equation 20). Under background selection models with recombination, the reduction is $B(U, L) = \exp({}^{-U}/_{L})$ where U is the per diploid genome per generation deleterious mutation rate, and L is the recombination map length (Hudson and Kaplan 1995; Hudson and Kaplan 1994; Nordborg et al. 1996). This BGS model is similar to models of effective population size under polygenic fitness variation, and can account for other modes of linked selection (Robertson 1961; Santiago and Caballero 1995, 1998, see Appendix Section A2). The coalescent rate due to sweeps is $S(\gamma, L, J) = {}^{\gamma}/_{L}J$, where γ is

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

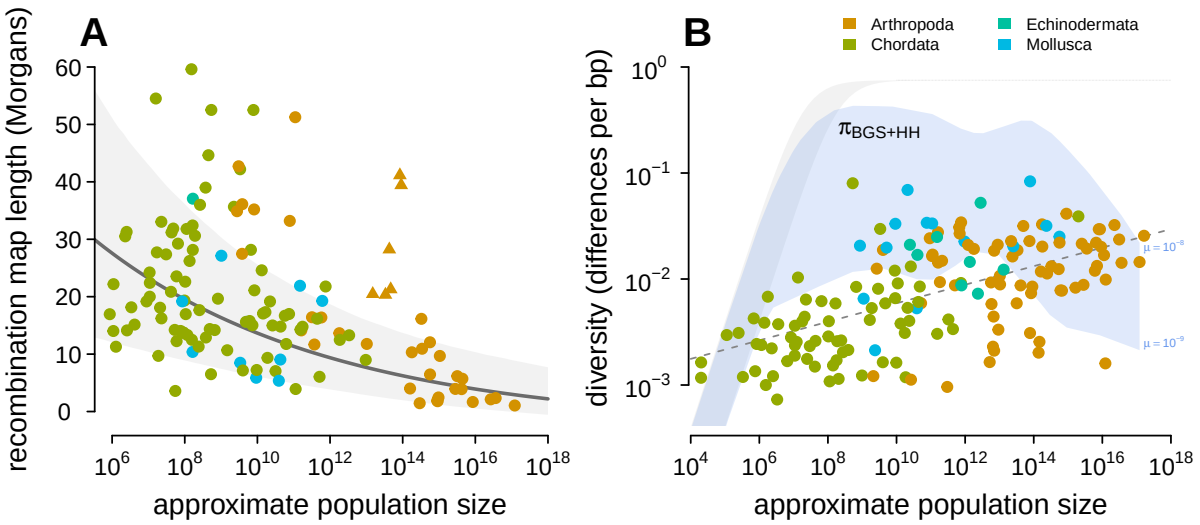

Figure 5: (A) The observed relationship between recombination map length (L) and census size (N_c) across 131 species with complete data and known phylogeny. Triangle points indicate social taxa excluded from the model fitting since these have adaptively higher recombination map lengths (Wilfert et al. 2007). The dark gray line is the estimated relationship under a phylogenetic mixed-effects model, and the gray interval is the 95% posterior average. (B) Points indicate the observed π–N_c relationship across taxa shown in Figure 2, and the blue ribbon is the range of predicted diversity were $N_e = N_c$ for $\mu = 10^{-8}$–10^{-9}, and after accounting for the expected reduction in diversity due to background selection and recurrent hitchhiking under *Drosophila melanogaster* parameters. In both plots, point color indicates phylum.

the number of adaptive substitutions per generation, and J is the probability a lineage is trapped by sweeps as they occur across the genome (c.f. $J_{2,2}$ in equation 15 of Coop and Ralph 2012).

Parameterizing the model this way, I then set the key parameters that determine the impact of recurrent hitchhiking and background selection (γ, J, and U) to high values estimated from *Drosophila melanogaster* by Elyashiv et al. (2016). My estimate of γ_{Dmel} based Elyashiv et al. implies $\nu_{\text{BP,Dmel}} \approx 2.34 \times 10^{-11}$, which is close to other estimates from *D. melanogaster* (see Supplementary Figure S11A). The rate of deleterious mutations per diploid genome, per generation is parameterized using the estimate from Elyashiv et al., $U_{\text{Dmel}} = 1.6$, which is a bit greater than previous estimates based on Bateman-Mukai approaches (Charlesworth 1987; Mukai 1988; Mukai 1985). Finally, the probability that a lineage is trapped in a sweep, J_{Dmel}, is calculated from the estimated genome-wide average coalescent rate due to sweeps from Elyashiv et al. (see Supplementary Figured S11B and Methods: *Predicted Reductions in Diversity* for more details on parameter estimates). Using these *Drosophila* parameters, I then explore how the predicted range of diversity levels under background selection and recurrent hitchhiking varies across species with recombination map length (L) and census population size (N_c).

Previous work has found that the impact of linked selection increases with N_c (Corbett-Detig et al. 2015; see also Supplementary Figure S10A), and it is often thought that this is driven by higher rates of adaptive substitutions in larger populations, despite equivocal evidence (Galtier 2016). However, there is another mechanism by which species with larger population sizes might experience a greater impact of linked selection: recombinational map length, L, is known to correlate with body mass (Burt and Bell 1987) and thus varies inversely with population size. As this is a critical

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

parameter that determines the genome-wide impact of both hitchhiking and background selection, I examine the relationship between recombination map length (L) and census population size (N_c) across taxa, using available estimates of map lengths across species (Corbett-Detig et al. 2015; Stapley et al. 2017). I find a significant non-linear relationship using phylogenetic mixed-effects models (Figure 5A; see Methods: *Phylogenetic Comparative Methods*). There is also a correlation between map length and genome size (Supplementary Figure S5) and genome size and population size (Supplementary Figure S4). Overall, the negative relationship between map length and census size indicates linked selection is expected to be stronger in short map length, high-N_c species.

Then, I predict the expected diversity (π_{BGS+HH}) under background selection and hitchhiking, were $N_e = N_c$, and assuming all species had the rate of sweeps and strength of BGS as *D. melanogaster*. Since neutral mutation rates μ are unknown and vary across species, I calculate the range of predicted π_{BGS+HH} estimates for $\mu = 10^{-8}$–10^{-9} (using the four-alleles model, Tajima 1996), and compare this to the observed relationship between π and N_c in Figured 5B. Under these parameters, linked selection begins to appreciably depress diversity around $N_c \approx 10^9$, since $S \approx 10^{-8}$–10^{-9} and linked selection dominates drift when $S > 1/2N$. Overall, this reveals two problems for the hypothesis that linked selection could solve Lewontin's Paradox. First, low to mid-N_c species (census sizes between 10^6–10^{14}) have sufficiently long map lengths that their diversity levels are only moderately reduced by linked selection, leading to a wide gap between predicted and observed diversity levels. For this not to be the case, the parameters that determine the strength of background selection and recurrent hitchhiking would need to be *higher* among these species than in *Drosophila melanogaster*. This would require that the rate of adaptive mutations or the deleterious mutation rate be orders of magnitude higher for species within this range than in *Drosophila*, which is incompatible with the rate of adaptive substitutions across species (Galtier 2016) and mutation rates (Lynch 2010). Furthermore, linked selection has been quantified in humans, which fall in this census size range, and has been found to be relatively weak (Boyko et al. 2008; Cai et al. 2009; Hellmann et al. 2008; Hernandez et al. 2011; McVicker et al. 2009). Second, while hitchhiking and BGS can reduce predicted diversity levels for high-N_c species ($N_c > 10^{14}$) to observed levels, this would imply available estimates of π_0 are underestimated by several orders of magnitude in *Drosophila* (Supplementary Figured S10B). Overall, while linked selection could decouple diversity from census size for high-N_c species, recurrent hitchhiking and background selection seem unlikely to explain the observed patterns of diversity across species under our understanding of the range of parameter estimates.

Discussion

Nearly fifty years after Lewontin's description of the Paradox of Variation, how evolutionary, life history, and ecological processes interact to constrain diversity across taxa to a narrow range remains a mystery. Since Wright (1931; 1938), population geneticists have appreciated that various demographic processes shrink effective population sizes compared to census sizes, yet it has remained unclear whether these neutral processes alone can explain Lewontin's Paradox and across-taxa diversity patterns. Alternatively, selective processes that act more strongly in larger populations could account for the observed narrow range of diversity. A critical first step to discerning the processes that act to transform census sizes to diversity levels across species is characterizing the observed π–N_c relationship.

Here, for 172 taxa, I estimated the relationship between genomic estimates of pairwise diversity

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

and approximate census population size. Previous surveys have used allozyme-based estimates, fewer taxa, or qualitative measures of population size. My estimates of census population sizes are quite approximate, since they use body size to predict density. An improved estimate might consider vagility (as Soulé 1976 did), though this is harder to do systematically across many taxa. Future work might also use other ecological information, such as total biomass and estimated numbers of species within phyla, to improve census size estimates (Bar-On et al. 2018; Mora et al. 2011). Still, it seems more accurate estimates would be unlikely to change the qualitative findings here, which resemble those of early surveys (Nei and Graur 1984; Soulé 1976).

One limitation of the dataset in this study is that diversity estimates are collated from a variety of sources rather than estimated with a single bioinformatic pipeline. This leads to technical noise across diversity estimates; perhaps the relationship between π and N_c found here could be tighter with a standardized bioinformatic pipeline. In addition to this technical variation, there might be systematic bioinformatic sources of bias in diversity estimates. For example high-diversity sequences may fail to align to the reference genome and end up unaccounted for, leading to a downward bias. Alternatively, high-diversity sequences might map to the reference genome, but adjacent mis-matching SNPs might be mistaken for a short insertion or deletion. While these issues might adversely affect the estimates in high-diversity species, it is unlikely they will qualitatively change the observed π–N_c relationship.

Macroevolution and Across-Taxa Population Genomics

Lewontin's Paradox arises from a comparison of diversity across species, yet surprisingly, previous work on this problem has not considered the impact of phylogenetic non-independence. I have addressed this limitation, showing that diversity does have a significant positive relationship with census size, after accounting for shared phylogenetic history among taxa. Additionally, I find a high degree of phylogenetic signal, and that arthropods and chordates form clusters, showing previous concern about phylogenetic non-independence was warranted (Gillespie 1991). Finally, this high degree of phylogenetic signal, as well as evidence of shifts in the rate of evolution of genetic diversity on deep timescales in molluscs and chordates, seem to contradict Lynch's (2011) claim that since coalescent times are much less than divergence times, they are not affected by shared phylogenetic history.

One can reconcile my findings with Lynch's claim by considering what evolutionary, ecological, life history, and demographic causal factors determine coalescent rates across species, and how these factors evolve across deep timescales. Lynch's conjecture that coalescent times should be free of phylogenetic signal may be true *were we to condition* on these causal factors that could be affected by shared phylogenetic history. In contrast, my estimates of phylogenetic signal in diversity are not conditioned on these factors. Importantly, even "correcting for" phylogeny implicitly favors certain causal interpretations over others (Uyeda et al. 2018; Westoby et al. 1995). Future work could try to untangle what causal factors determine coalescent times across species, as well as how these factors evolve across macroevolutionary timescales.

Furthermore, beyond just accounting for phylogenetic non-independence, macroevolution and phylogenetic comparative methods are a promising way to approach across-species population genomic questions. For example, one could imagine that diversification processes could contribute to Lewontin's Paradox. If large-N_c species were to have a rate of speciation that is greater than the rate at which mutation and drift reach equilibrium (which is indeed slower for large N_c species), this could act to decouple diversity from census population size. That is to say, even if the rate

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

of random demographic bottlenecks were constant across taxa, lineage-specific diversification processes could lead certain clades to be systematically further from demographic equilibrium, and thus have lower diversity than expected for their census population size.

Spatial and Demographic Processes

One limitation of this study is the inability to quantify the impact of spatial population genetic processes on the relationship between diversity and census population sizes across taxa. The genomic diversity estimates collated in this study unfortunately lack details about the sampling process and spatial data, which can have a profound impact on population genomic summary statistics (Battey et al. 2020). These issues could systematically bias species-wide diversity estimates; for example, if diversity estimates from a cosmopolitan species were primarily from a single subpopulation, diversity would be an underestimate relative to the entire population. However, biased spatial sampling alone seems incapable of explaining the π-N_c divergence in high-N_c taxa. In the extreme scenario in which only one subpopulation was sampled, F_{ST} would need to be close to one for population subdivision alone to sufficiently reduce the total population heterozygosity to explain the orders-of-magnitude shortfall between predicted and observed diversity levels. Yet, across-taxa surveys indicate that F_{ST} is almost never this high within species (Roux et al. 2016). Still, future work could quantify the extent to which spatial processes contribute to Lewontin's Paradox. For example, high-N_c taxa usually experience range expansions, likely with repeated founder effects and local extinction/recolonization dynamics that doubtlessly depress diversity. In particular, with the appropriate data, one could estimate the empirical relationship between dispersal distance, range size, and coalescent effective population size across taxa.

In this study, I have focused entirely on assessing the role of linked selection, rather than demography, in reducing diversity across taxa. In contrast to demographic models, models of linked selection have comparatively fewer parameters and more readily permit rough estimates of diversity reductions across taxa. Still, a full resolution of Lewontin's Paradox would require understanding how the demographic processes across taxa with incredibly heterogeneous ecologies and life histories transform N_c into N_e. With population genomic data becoming available for more species, this could involve systematically inferring the demographic histories of tens of species and looking for correlations in the frequency and size of bottlenecks with N_c across species.

How could selection still explain Lewontin's Paradox?

In this study, my goal was not to accurately estimate the levels in diversity across species, but rather to give linked selection the best possible chance to solve Lewontin's Paradox. Still, I find that even after parameterizing hitchhiking and background selection with strong selection parameter estimates from *Drosophila melanogaster*, the predicted patterns of diversity under linked selection poorly fit observed patterns of diversity across species. While this suggests these two common modes of linked selection seem unlikely to explain across-taxa patterns of diversity, there are three major potential limitations of my approach that need further evaluation.

First, I approximate the reduction in diversity using homogeneous background selection and recurrent hitchhiking models (Coop and Ralph 2012; Hudson and Kaplan 1995; Kaplan et al. 1989), when in reality, there is genome-wide heterogeneity in functional density, recombination rates, and the adaptive substitutions across species. Each of these factors mediate how strongly linked selection impacts diversity across the genome. Despite these model simplifications, my

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

predicted reduction in diversity in *Drosophila melanogaster* is 85%, which is reasonably close to the estimated 77% from the more realistic model of Elyashiv et al. that accounts for the actual position of substitutions, annotation features, and recombination rate heterogeneity (though it should be noted that these both use the same parameter estimates). Furthermore, even though my model fails to capture the heterogeneity of functionality density and recombination rate in real genomes, it is still extraordinary conservative, giving linked selection the best possible chance to decouple diversity from census size and explain Lewontin's Paradox. This is in part because the strong selection parameter estimates from *Drosophila melanogaster* used, but also because I assume that the effective population size is the census size. Even then, this decoupling only occurs in very high–census-size species, and implies that the diversity in the absence of linked selection, π_0, is currently underestimated by several orders of magnitude. Moreover, the study of Corbett-Detig et al. (2015) did consider recombination rate and functional density heterogeneity in estimating the reduction due to linked selection across species, yet their predicted reductions are orders of magnitude weaker than those considered here by assuming that $N_e = N_c$ (Supplementary Figure S10B). Overall, even with more realistic models of linked selection, current models of linked selection seem fundamentally unable to fit the diversity–census-size relationship.

Second, my model here only considers hard sweeps, and ignores the contribution of soft sweeps (e.g. from standing variation or recurrent mutations; Hermisson and Pennings 2005; Pennings and Hermisson 2006), partial sweeps (e.g those that do not reach fixation), and the interaction of sweeps and spatial processes. While future work exploring these alternative types of sweeps is needed, the predicted reductions in diversity found here under the simplified sweep model are likely relatively robust to these other modes of sweeps for a few reasons. First, the shape of the diversity–recombination curve is equivalent under models of partial sweeps and hard sweeps, though these imply different rates of sweeps (Coop and Ralph 2012). Second, in the limit where most fitness variation is due to weak soft sweeps from standing variation scattered across the genome (i.e. due to polygenic fitness variation), levels of diversity are well approximated by quantitative genetic linked selection models (QGLS Robertson 1961; Santiago and Caballero 1995, 1998). The reduction in diversity under these models is nearly identical to that under background selection models, in part because deleterious alleles at mutation-selection balance constitute a considerable component of fitness variation (see Appendix Section A2; Charlesworth 2015; Charlesworth and Hughes 2000). Finally, I also disregarded the interaction of sweeps and spatial processes. For populations spread over wide ranges, limited dispersal slows the spread of sweeps, allowing for new beneficial alleles to arise, spread, and compete against other segregating beneficial variants (Ralph and Coop 2010; Ralph and Coop 2015). Through limited dispersal should act to "soften sweeps" and not impact my findings for the reasons described above, future work could investigate how these processes impact diversity in ways not captured by hard sweep models.

Third, other selective processes, such as fluctuating selection or hard selective events, could reduce diversity in ways not captured by the background selection and hitchhiking model. Since frequency-independent fluctuating selection generally reduces diversity under most conditions (Novak and Barton 2017), this could lead seasonality and other sources of temporal heterogeneity to reduce diversity in large-N_c species with short generation times more than longer-lived species with smaller population sizes. Future work could consider the impact of fluctuating selection on diversity under simple models (Barton 2000) if estimates of key parameters governing the rate of such fluctuations were known across taxa. Additionally, another mode of selection that could severely reduce diversity across taxa, yet remains unaccounted for in this study, is periodic hard selective

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

events. These selective events could occur regularly in a species' history yet be indistinguishable from demographic bottlenecks from just population genomic data.

Measures of Effective Population Size, Timescales, and Lewontin's Paradox

Lewontin's Paradox shows the extent to which the effective population sizes implied by diversity, $\widetilde{N}_e$, diverge from census population sizes. However, there are a variety other effective population size estimates calculable from different data and summary statistics (Caballero 1994; Caballero 2020; Wang et al. 2016). These include estimators based on the observed decay in linkage disequilibrium or temporal estimators that use the variance in allele frequency change. These alternate estimators capture summaries of the effective population size on shorter timescales than coalescent-based estimators (Wang 2005), and thus could be used to tease out processes that impact the N_e-N_c relationship in the more recent past.

Temporal N_e estimators already play an important role in understanding another summary of the N_e-N_c relationship: the ratio N_e/N_c, which is an important quantity in conservation genetics (Frankham 1995; Mace and Lande 1991) and in understanding evolution in highly-fecund marine species. Surveys of the N_e/N_c relationship across taxa indicate mean N_e/N_c is on order of ≈ 0.1 (Frankham 1995; Palstra and Fraser 2012; Palstra and Ruzzante 2008), though the uncertainty in these estimates is high, and some species with sweepstakes reproduction systems like Pacific Oyster (*Crassostrea gigas*) can have $N_e/N_c \approx 10^{-6}$. Estimates of the N_e/N_c ratio are an important, and under appreciated piece of solving Lewontin's Paradox. For example, if N_e is estimated from the allele frequency change across a single generation (i.e. Waples 1989), N_e/N_c constrains the variance in reproductive success (Nunney 1993, 1996; Wright 1938). This implies that apart from species with sweepstakes reproductive systems, the variance in reproductives success each generation (whether heritable or non-heritable) is likely insufficient to significantly contribute to constraining $\widetilde{N}_e$ for most taxa. Still, further work is needed to characterize (1) how N_e/N_c varies with N_c across taxa (though see Palstra and Fraser 2012, Figure 2), and (2) the variance of N_e/N_c over longer time spans (i.e. how periodic sweepstakes reproductive events act to constrain N_e). Overall, characterizing how N_e/N_c varies across taxa and correlates with ecology and life history traits could provide clues into the mechanisms that leads propagule size and survivorship curves to be predictive of diversity levels across taxa (Barry et al. 2020; Hallatschek 2018; Romiguier et al. 2014).

Finally, short-term temporal N_e estimators may play an important role in resolving Lewontin's Paradox. These estimators, along with short-term estimates of the impact of linked selection (Buffalo and Coop 2019, 2020), can inform us how much diversity is depressed across shorter timescales, free from the rare strong selective events or severe bottlenecks that impact pairwise diversity. It could be that in any one generation, selection contributes more to the variance of allele frequency changes than drift, yet across-taxa patterns in diversity are better explained processes acting sporadically on longer timescales, such as colonization, founder effects, and bottlenecks. Thus, the pairwise diversity may not give us the best picture of the generation to generation evolutionary processes acting in a population to change allele frequencies. Furthermore, certain observed adaptations are inexplicable given implied long-term coalescent effective population sizes, and are only possible if short-term effective population sizes are orders of magnitude larger (Barton 2010; Karasov et al. 2010).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Conclusions

In *Building a Science of Population Biology* (2004), Lewontin laments the difficulty of uniting population genetics and population ecology into a cohesive discipline of population biology. Lewontin's Paradox of Variation remains a critical unsolved problem at the nexus of these two different disciplines: across species, we fail to understand the processes that connect a central parameter of population ecology, census size, to a central parameter of population genetics, effective population size. Given that selection seems to fall short in explaining Lewontin's Paradox, a full resolution will require a mechanistic understanding the ecological, life history, and macroevolutionary processes that connect N_c to N_e across taxa. While I have focused exclusively on metazoan taxa since their population densities are more readily approximated from body mass, a full resolution must also include plant species (with the added difficulties of variation in selfing rates, different dispersal strategies, pollination, etc.).

Looking at Lewontin's Paradox through an macroecological and macroevolutionary lens begets interesting questions outside of the realm of population genetics. Here, I have found that diversity and N_c have a surprisingly consistent relationship without many outliers, despite the wildly disparate ecologies, life histories, and evolutionary histories of the taxa included. Furthermore, taxa with very large census sizes have surprisingly low diversity. Is this explained by macroevolutionary processes, such as different rates of speciation for large-N_c taxa? Or, are the levels of diversity we observe today an artifact of our timing relative to the last glacial maximum, or the last major extinction? Did large-N_c prehistoric animal populations living in other geological eras have higher levels of diversity than our present taxa? Or, does ecological competition occur on shorter timescales such that strong population size contractions transpire and depress diversity, even if a species is undisturbed by climatic shifts or mass extinctions? Overall, patterns of diversity across taxa are determined by many overlaid evolutionary and ecological processes occurring on vastly different timescales. Lewontin's Paradox of Variation may persist unresolved for some time because the solution requires synthesis and model building at the intersection of all these disciplines.

Methods

Diversity and Map Length Data

The data used in this study are collated from a variety of previously published surveys. Of the 172 taxa with diversity estimates, 14 are from Corbett-Detig et al. (2015), 96 are from Leffler et al. (2012), and 62 are from Romiguier et al. (2014). The Corbett-Detig et al. data is estimated from four-fold degenerate sites, the Romiguier et al. data is synonymous sites, and the Leffler et al. data is estimated predominantly from silent, intronic, and non-coding sites. All types of diversity estimates from Leffler et al. (2012) were included to maximize the taxa in the study, since the variability of diversity across functional categories is much less than the diversity across taxa. Multiple diversity estimates per taxa were averaged. The total recombination map length data were from both Stapley et al. (2017; 30 taxa), and Corbett-Detig et al. (2015; 11 taxa). Both studies used sex-averaged recombination maps estimated with cross-based approaches; in some cases errors in the original data were found, documented, and corrected. These studies also included genome size estimates used to create Supplementary Figures S4 and S5.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Macroecological Estimates of Population Size

A rough approximation for total population size (census size) is $N_c = DR$, where D is the population density in individuals per km^2 and R is the range size in km^2. Since population density estimates are not available for many taxa included in this study, I used the macroecological abundance-body size relationship to predict population density from body size. Since body length measurements are more readily available than body mass, I collated body length data from various sources (see `https://github.com/vsbuffalo/paradox_variation/`); body lengths were averaged across sexes for sexually dimorphic species, and if only a range of lengths was available, the midpoint was used.

Then, I re-estimated the relationship between body mass and population density using the data in the appendix table of Damuth (1987), which includes 696 taxa with body mass and population density measurements across mammals, fish, reptiles, amphibians, aquatic invertebrates, and terrestrial arthropods. Though the abundance-body size relationship can be noisy at small spatial or phylogenetic scales (Chapter 5, Gaston and Blackburn 2008), across deeply diverged taxa such as those included in this study and Damuth (1987), the relationship is linear and homoscedastic (see Supplementary Figure S8). Using Stan (Stan Development Team 2020), I jointly estimated the relationship between body mass from body length using the Romiguier et al. (2014) taxa, and used this relationship to predict body mass for the taxa in this study. These body masses were then used to predict population density simultaneously, using the Damuth (1981) relationship. The code of this routine (`pred_popsize_missing_centered.stan`) is available in the GitHub repository (`https://github.com/vsbuffalo/paradox_variation/`).

To estimate range, I first downloaded occurrence records from Global Biodiversity Information Facility (*GBIF Occurrence Download* 2020) using the `rgbif` R package (Chamberlain and Boettiger 2017; Chamberlain et al. 2014). Using the occurrence locations, I inferred whether a species was marine or terrestrial, based on whether the majority of their recorded occurrences overlapped a continent using `rnaturalearth` and the `sf` packages (Pebesma 2018; South 2017). For each taxon, I estimated its range by finding the minimum α-shape containing these occurrences. The α parameters were set more permissive for marine species since occurrence data for marine taxa were sparser. Then, I intersected the inferred ranges for terrestrial taxa with continental polygons, so their ranges did not overrun landmasses (and likewise with marine taxa and oceans). I inspected diagnostic plots for each taxa for quality control (all of these plots are available in `paradox_variation` GitHub repository), and in some cases, I manually adjusted the α parameter or manually corrected the range based on known range maps (these changes are documented in the code `data/species_ranges.r` and `data/species_range_fixes.r`). The range of *C. elegans* was conservatively approximated as the area of the Western US and Western Europe based on the map in Frézal and Félix (2015). *Drosophila* species ranges are from the Drosophila Speciation Patterns website, (Yukilevich 2017; Yukilevich 2012). To further validate these range estimates, I have compared these to the qualitative range descriptions Leffler et al. (2012) (Supplementary Figure S7) and compared my α-shape method to a subset of taxa with range estimates from IUCN Red List (Chamberlain 2020; IUCN 2020; Supplementary Figure S6). Each census population size is then estimated as the product of range and density.

Phylogenetic Comparative Methods

Of the full dataset of 172 taxa with diversity and population size estimates, a synthetic calibrated phylogeny was created for 166 species that appear in phylogenies in DateLife project (O'Meara

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

et al. 2020; Sanchez-Reyes and O'Meara 2019). This calibrated synthetic phylogeny was then subset for the analyses based on what species had non-missing trait data. The diversity-population size relationship assessed by a linear phylogenetic mixed-effects model implemented in Stan (Stan Development Team 2020), according to the methods described in (Villemereuil and Nakagawa 2014, see `stan/phylo_mm_regression.stan` in the GitHub repository). This same Stan model was used to estimate the same relationship between arthropod, chordate, and mollusc subsets of the data (Supplementary Figure S12).

The relationship between recombination map length and the logarithm of population size is non-linear and heteroscedastic, and was fit using a lognormal phylogenetic mixed-effects model. Since social insects have longer recombination map lengths (Wilfert et al. 2007), social taxa were excluded when fitting this model. All `Rhat` (Vehtari et al. 2019) values were below 1.01 and the effective number of samples was over 1,000, consistent with good mixing; details about the model are available in the GitHub repository (`phylo_mm_lognormal.stan`). Continuous trait maps (Figure 4A and Supplementary Material Figure S14) were created using `phytools` (Revell 2012). Node-height tests were implemented based on the methods in `Geiger` (Harmon et al. 2008; Pennell et al. 2014), and use robust regression to fit a linear relationship between phylogenetic independent contrasts and branching times.

Predicted Reductions in Diversity

The predicted reductions in diversity due to linked selection are approximated using selection and deleterious mutation parameters from *Drosophila melanogaster*, and the recombination map length estimates from Stapley et al. (2017) and Corbett-Detig et al. (2015). The mathematical details of the simplified sweep model are explained in the Appendix Section A1. I use estimates of the number of substitutions, m, in genic regions between *D. melanogaster* and *D. simulans* from Hu et al. (2013). Following Elyashiv et al. (2016), only substitutions in UTRs and exons are included, since they found no evidence of sweeps in introns. Then, I average over annotation classes to estimate the mean proportion of substitutions that are beneficial, $\alpha_{\text{Dmel}} = 0.42$, which are consistent with the estimates of Elyashiv et al. and estimates from MacDonald–Kreitman test approaches (see Eyre-Walker 2006, Table 1). Then, I use divergence time estimates between *D. melanogaster* and *D. simulans* of 4.2×10^6 and estimate of ten generations per year (Obbard et al. 2012), calculating there are $\gamma_{\text{Dmel}} = {}^{\alpha m}/_{2T} = 2.26 \times 10^{-3}$ substitutions per generation. Given the length of the *Drosophila* autosomes, G, this implies that the rate of beneficial substitutions per basepair, per generation is $\nu_{BP,Dmel} = {}^{\gamma_{\text{Dmel}}}/_{G} = 2.34 \times 10^{-11}$. Finally, I estimate J_{Dmel} from the estimate of genome-wide average rate of sweeps from Elyashiv et al. (Supplementary Table S6) and assuming *Drosophila* $N_e = 10^6$. These *Drosophila melanogaster* hitchhiking parameter estimates are close to other previously-published estimates (Supplementary Figure S11). Finally, I use $U_{\text{Dmel}} = 1.6$, from Elyashiv et al. (2016). With these parameter estimates from *D. melanogaster*, the recombination map lengths across species, and Equation (1), I estimate $\pi_{\text{BGS+HH}}$ (assuming $N_c = N_c$) across all species. This leads to a range of predicted diversity ranges across species corresponding to $\mu = 10^{-8}$–10^{-9}; to visualize these, I take a convex hull of all diversity ranges and smooth this with R's `smooth.spline` function.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Acknowledgments

I would like to thank Andy Kern and Peter Ralph for helpful discussions and supporting me during this work, and Graham Coop for inspiration and helpful feedback during socially distanced nature walks at Yolo Basin. I thank Jessica Stapley for kindly providing the recombination map length data, and Yaniv Brandvain, Amy Collins, Doc Edge, Tyler Kent, Chuck Langley, Sally Otto, Jeff Ross-Ibarra, Aaron Stern, Anastasia Teterina, Michael Turelli, and my Kern-Ralph labmates for helpful discussions. Sarah Friedman, Katherine Corn, and Josef Uyeda provided very useful advice about phylogenetic comparative methods; yet I take full responsibility for any shortcomings of my analysis. I would like to also thank UO librarian Dean Walton for helping me track down some rather difficult to find older papers. This work was supported by an NIH Grant (1R01GM117241) awarded to Andrew Kern.

Appendix

A1 Simplified Sweep Effects Model

I use a simplified model of the effects of recurrent hitchhiking and background selection (BGS) occurring uniformly along a genome. Expected diversity is given by

$$\mathbb{E}(\pi) = \frac{\theta}{\theta + {}^{1}\!/_{B} + 2NS} \tag{A2}$$

$$\approx \frac{\theta}{{}^{1}\!/_{B} + 2NS} \tag{A3}$$

(cf. equation 1 Elyashiv et al. 2016, and equation 20 of Coop and Ralph 2012). The BGS component is given by Hudson and Kaplan (1995),

$$B(U, L) = N_e \exp\left(-\frac{U}{L}\right) \tag{A4}$$

and the hitchhiking component is

$$S = \frac{\nu_{\mathrm{BP}}}{r_{\mathrm{BP}}} J \tag{A5}$$

(cf. Coop and Ralph 2012 equation 20) where J is the probability that two lineages coalesce down to one, given sweeps occur uniformly along the genome. Under this homogeneous sweep model, J is

$$J = \int_0^L q_f(r)^2 dr \tag{A6}$$

where $q_f(r)$ is the approximate probability that a lineage is trapped by a sweep to frequency f when it is r recombination fraction away from this sweep (cf. Coop and Ralph 2012 equation 15).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Since I use *Drosophila melanogaster* parameter estimates from Elyashiv et al. (2016), I now reconcile their model's S term with the simple model above. They estimate S in *Drosophila melanogaster* using a composite likelihood model that considers hitchhiking and background selection simultaneously, using substitutions and stratifying by annotation. For a neutral position at site x, the coalescent rate due to sweeps is given by Elyashiv et al.'s equation 3,

$$S(x) = \frac{1}{T}\sum_{i_S}\alpha(i_S)\sum_{y\in a(i_S)}\int \exp(-r(x,y)\tau(s,N))g(s|i_S)ds \tag{A7}$$

where T is the number of generations that substitutions accrue, $i_S = 1,\ldots,I_S$ is the annotation class (e.g. exons, introns, UTRs), $\alpha(i_S)$ is the fraction of substitutions in annotation class i_S that are beneficial, $a(i_S)$ is the set of all substitutions in annotation class i_S, $\tau(s,N)$ is the fixation time of a site with additive effect s, and $g(s|i_S)$ is the distribution of selection coefficients for annotation class i_S.

Note, that we can recover the model of Coop and Ralph (2012) from this expression. Suppose there is only one annotation class, and α fraction of substitutions are beneficial, and one selection coefficient $\bar{s}$, (i.e. $g(s) = \delta_0(s-\bar{s})$), then

$$S(x) = \frac{\alpha}{T}\sum_{y\in a}\exp(-r(x,y)\tau(\bar{s},N)). \tag{A8}$$

Let the number of substitutions be $m := |a|$, and imagine their positions are uniformly distributed on a segment of length G basepairs with the focal site is the middle at position $x = 0$. Then, each substitution y is a random distance $l_y \sim U(-G/2, G/2)$ away from the focal site. Assuming the recombination rate is a constant r_{BP} per basepair, and approximating the sum with an integral, we have,

$$S = \frac{\alpha}{T}\sum_{i=1}^{m}\mathbb{E}_{l_i}\left(\exp(-r_{\mathrm{BP}}l_i\tau(\bar{s},N))\right) \tag{A9}$$

$$= \frac{\alpha}{TG}\sum_{i=1}^{m}\int_0^G \exp(-r_{\mathrm{BP}}\ell\tau(\bar{s},N))d\ell \tag{A10}$$

$$= \frac{\alpha m}{TG}\int_0^G \exp(-r_{\mathrm{BP}}\ell\tau(\bar{s},N))d\ell \tag{A11}$$

Using u-substitution with $r = \ell r_{\mathrm{BP}}$ this simplifies to

$$S = \frac{\alpha m}{TGr_{\mathrm{BP}}}\int_0^L \exp(-r\tau(\bar{s},N))dr \tag{A12}$$

where $L = Gr_{\mathrm{BP}}$.

To simplify this notation, note that the rate of adaptive substitutions per basepair per generation is $\nu_{\mathrm{BP}} = {}^{\alpha m}/_{GT}$, so

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

$$S = \frac{\nu_{\mathrm{BP}}}{r_{\mathrm{BP}}} \int_0^L \exp(-r\tau(\bar{s}, N))dr \tag{A13}$$

This is analogous to the second term of Coop and Ralph (2012) equation 17, with $k = i = 2$ and $x = 1$ (e.g. conditioning on a sweep to fixation). Note that there appears to be a factor of two error in Elyashiv et al. (2016) compared to Coop and Ralph (2012); here I include the factor of two. Then,

$$S = \frac{\nu_{\mathrm{BP}}}{r_{\mathrm{BP}}} \underbrace{\int_0^L \exp(-2r\tau(\bar{s}, N))dr}_{J} \tag{A14}$$

where the integral is equal to J (c.f. $J_{2,2}$ of equation 15 in Coop and Ralph 2012) since a simple model of $q_f(r) = f\exp(-2r\tau(s, N))$ and if we condition on fixation, $f = 1$. This expression is useful to generalize across species, since we know N and L. Additionally, we have estimates of α and m/T in *Drosophila* and other species. In Elyashiv et al, they consider the number of substitutions per generation in genic regions only; it should be noted that the number of coding basepairs varies little across species. For convenience, I define $\gamma = \alpha m/T$ as the number of adaptive substitutions per generation per entire genome, such that $S(\gamma, L, J) = \gamma/L\ J$ used in the main text. Using the estimates of $m \approx 4.5 \times 10^5$, $\alpha \approx 0.42$, and $T \approx 8.4 \times 10^7$ from the Supplementary Material of Elyashiv et al., I arrive at $\gamma \approx 0.00226$ adaptive substitutions per generation, per genome. For a ≈ 100 megabase genome, this translates to a $\nu_{\mathrm{BP}} \approx 2.34 \times 10^{-11}$, which is close to previous estimates (Supplementary Figure S11). For J, I use an empirical estimate calculated from the genome-wide average of the rate of coalescent events due to sweeps, from Supplementary Table S6 of Elyashiv et al. ($r_s = 2NS \approx 0.92$). This implies $J \approx 4.46 \times 10^{-4}$. Alternatively, I have tried using the estimated distribution of selection coefficients from Elyashiv et al., but this led to a weaker estimate of J, since the adaptive substitutions considered tend to cluster around genic regions. Note that these *Drosophila* sweep parameters I have used are close to previous estimates (Supplementary Figures S11 A and B).

A2 Background Selection and Polygenic Fitness Models

Throughout the main text, I use recurrent hitchhiking and background selection models to estimate the reduction in diversity due to linked selection. Another class of linked selection models, which I refer to as quantitative genetic linked selection models (QGLS; Robertson 1961; Santiago and Caballero 1995, 1998), can also depress genome-wide diversity. Furthermore, these models may depress diversity at neutral sites unlinked to the regions containing fitness variation. While I did not explicitly incorporate these models into my estimates of the diversity reductions, their effect is implicit in background selection models because they are analytically nearly identical. Here, I briefly sketch out the connection between BGS and QGLS models.

Under the Santiago and Caballero (1998) model, the effective population size is $N_e^{\mathrm{SC98}} = N\exp(-C^2/(1-Z)L)$, where C^2 is the standardized heritable fitness variation, $1 - Z$ is the decay of genetic variance through time, and L is the recombination map length. This model can accommodate a variety of modes of selection such as selection on an infinitesimal trait (Santiago and

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Caballero 1995, p. 1016), and the flux of either weakly advantageous or deleterious alleles (Santiago and Caballero 1998, p. 2109). If the source of fitness variation is entirely the input of new deleterious mutations with heterozygous effect sh at rate U per diploid genome per generation, then under mutation-selection balance, the equilibrium relative variance in reproductive success $C^2 = Ush$ (Crow and Kimura 1970; Caballero 2020, p. 167), and $Z = 1 - sh - {}^1/_{2N_c}$ (Santiago and Caballero 1998). Thus, if ${}^1/_{2N_c} << sh << 1$, then ${}^{C^2}/_{(1-Z)} \approx U$ and $N_e^{\mathrm{SC98}} \approx N \exp({}^{-U}/_{L})$, which is the BGS model used in the main text and is a result of many background selection models with similar assumptions (Hudson and Kaplan 1994 eqn. 15; Hudson and Kaplan 1995 eqn. 9; Nordborg et al. 1996 eqn. 4; Barton 1995 eqn. 22b). Intuitively, the similarity of these models reflects the fact that a substantial proportion of heritable fitness variation is caused by the continual flux of deleterious alleles across the genome under mutation-selection balance (Charlesworth 2015; Charlesworth and Hughes 2000).

References

Aguade, M, N Miyashita, and C H Langley (1989). "Reduced variation in the yellow-achaete-scute region in natural populations of Drosophila melanogaster". *Genetics* 122.3, pp. 607–615.

Andolfatto, Peter (2007). "Hitchhiking effects of recurrent beneficial amino acid substitutions in the Drosophila melanogaster genome". *Genome Res.* 17.12, pp. 1755–1762.

Bar-On, Yinon M, Rob Phillips, and Ron Milo (2018). "The biomass distribution on Earth". *Proc. Natl. Acad. Sci. U. S. A.* 115.25, pp. 6506–6511.

Barry, Pierre, Thomas Broquet, and Pierre-Alexandre Gagnaire (2020). "Life tables shape genetic diversity in marine fishes".

Barton, N H (1995). "Linkage and the limits to natural selection". *Genetics* 140.2, pp. 821–841.

— (2000). "Genetic hitchhiking". *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 355.1403, pp. 1553–1562.

Barton, Nick (2010). "Understanding adaptation in large populations". *PLoS Genet.* 6.6, e1000987.

Battey, C J, Peter L Ralph, and Andrew D Kern (2020). "Space is the Place: Effects of Continuous Spatial Structure on Analysis of Population Genetic Data". *Genetics* 215.1, pp. 193–214.

Begun, D J and C F Aquadro (1992). "Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster". *Nature* 356.6369, pp. 519–520.

Boyko, Adam R et al. (2008). "Assessing the evolutionary impact of amino acid mutations in the human genome". *PLoS Genet.* 4.5, e1000083.

Buffalo, Vince and Graham Coop (2019). "The Linked Selection Signature of Rapid Adaptation in Temporal Genomic Data". *Genetics* 213.3, pp. 1007–1045.

— (2020). "Estimating the genome-wide contribution of selection to temporal allele frequency change". *Proceedings of the National Academy of Sciences* 117.34, pp. 20672–20680.

Burt, A and G Bell (1987). "Mammalian chiasma frequencies as a test of two theories of recombination". *Nature* 326.6115, pp. 803–805.

Caballero, A (1994). "Developments in the prediction of effective population size". *Heredity* 73 (Pt 6), pp. 657–679.

Caballero, Armando (2020). *Quantitative Genetics*. Cambridge University Press.

Cai, James J, J Michael Macpherson, Guy Sella, and Dmitri A Petrov (2009). "Pervasive Hitchhiking at Coding and Regulatory Sites in Humans". *PLoS Genet.* 5.1, e1000336–13.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Carpenter, Bob et al. (2017). "Stan: A Probabilistic Programming Language". *Journal of Statistical Software, Articles* 76.1, pp. 1–32.

Chamberlain, Scott (2020). *rredlist: 'IUCN' Red List Client.*

Chamberlain, Scott and Carl Boettiger (2017). *R Python, and Ruby clients for GBIF species occurrence data.*

Chamberlain, Scott, K Ram, V Barve, and D Mcglinn (2014). "rgbif: interface to the global biodiversity information facility API". *R package version 0. 7* 7.

Charlesworth, B (1987). *The heritability of fitness.* Sexual selection: testing the alternatives.

— (1996). "Background selection and patterns of genetic diversity in Drosophila melanogaster". *Genet. Res.* 68.2, pp. 131–149.

Charlesworth, B, M T Morgan, and D Charlesworth (1993). "The effect of deleterious mutations on neutral molecular variation". *Genetics* 134.4, pp. 1289–1303.

Charlesworth, Brian (2015). "Causes of natural variation in fitness: evidence from studies of Drosophila populations". *Proc. Natl. Acad. Sci. U. S. A.* 112.6, pp. 1662–1669.

Charlesworth, Brian and Kimberly A Hughes (2000). "The Maintenance of Genetic Variation in Life-History Traits". *Evolutionary Genetics: From Molecules to Morphology, Vol 1.* Ed. by R S Singh and C Krimbas. Vol. 1. Cambridge University Press, pp. 369–392.

Chen, Jun, Sylvain Glémin, and Martin Lascoux (2017). "Genetic Diversity and the Efficacy of Purifying Selection across Plant and Animal Species". *Mol. Biol. Evol.* 34.6, pp. 1417–1428.

Coop, Graham (2016). *Does linked selection explain the narrow range of genetic diversity across species?* Tech. rep. Cold Spring Harbor Labs Journals.

Coop, Graham and Peter Ralph (2012). "Patterns of neutral diversity under general models of selective sweeps". *Genetics* 192.1, pp. 205–224.

Corbett-Detig, Russell B, Daniel L Hartl, and Timothy B Sackton (2015). "Natural selection constrains neutral diversity across a wide range of species". *PLoS Biol.* 13.4, e1002112.

Crow, James Franklin and Motoo Kimura (1970). *An Introduction to Population Genetics Theory.* New York, Evanston and London: Harper & Row, Publishers.

Cutter, Asher D and Bret A Payseur (2003). "Selection at linked sites in the partial selfer Caenorhabditis elegans". *Mol. Biol. Evol.* 20.5, pp. 665–673.

Damuth, John (1981). "Population density and body size in mammals". *Nature* 290.5808, pp. 699–700.

— (1987). "Interspecific allometry of population density in mammals and other animals: the independence of body mass and population energy-use". *Biol. J. Linn. Soc. Lond.* 31.3, pp. 193–246.

Eldon, Bjarki and John Wakeley (2006). "Coalescent processes when the distribution of offspring number among individuals is highly skewed". *Genetics* 172.4, pp. 2621–2633.

Elyashiv, Eyal et al. (2016). "A Genomic Map of the Effects of Linked Selection in Drosophila". *PLoS Genet.* 12.8, e1006130.

Eyre-Walker, Adam (2006). "The genomic rate of adaptive evolution". *Trends Ecol. Evol.* 21.10, pp. 569–575.

Felsenstein, Joseph (1985). "Phylogenies and the Comparative Method". *Am. Nat.* 125.1, pp. 1–15.

Fisher, R A and E B Ford (1947). "The spread of a gene in natural conditions in a colony of the moth Panaxia dominula". *Heredity* 1, pp. 143–174.

Frankham, Richard (1995). "Effective population size/adult population size ratios in wildlife: a review". *Genet. Res.* 66.2, pp. 95–107.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Frankham, Richard (1996). "Relationship of Genetic Variation to Population Size in Wildlife". *Conserv. Biol.* 10.6, pp. 1500–1508.

Freckleton, R P, P H Harvey, and M Pagel (2002). "Phylogenetic analysis and comparative data: a test and review of evidence". *Am. Nat.* 160.6, pp. 712–726.

Freckleton, Robert P and Paul H Harvey (2006). "Detecting non-Brownian trait evolution in adaptive radiations". *PLoS Biol.* 4.11, e373.

Frézal, Lise and Marie-Anne Félix (2015). "C. elegans outside the Petri dish". *Elife* 4.

Galtier, Nicolas (2016). "Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis". *PLoS Genet.* 12.1, e1005774.

Gaston, Kevin and Tim Blackburn (2008). *Pattern and Process in Macroecology*. John Wiley & Sons.

GBIF Occurrence Download (2020).

Gillespie, J H (1991). *The causes of molecular evolution*. Oxford: Oxford University Press Google Scholar.

— (2000). "Genetic drift in an infinite population. The pseudohitchhiking model". *Genetics* 155.2, pp. 909–919.

— (2001). "Is the population size of a species relevant to its evolution?" *Evolution* 55.11, pp. 2161–2169.

Hadfield, J D and S Nakagawa (2010). "General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters". *J. Evol. Biol.* 23.3, pp. 494–508.

Hallatschek, Oskar (2018). "Selection-Like Biases Emerge in Population Models with Recurrent Jackpot Events". *Genetics* 210.3, pp. 1053–1073.

Harmon, L J, J T Weir, C D Brock, R E Glor, and W Challenger (2008). *GEIGER: investigating evolutionary radiations*.

Hauser, Lorenz and Gary R Carvalho (2008). "Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts". *Fish Fish* 9.4, pp. 333–362.

Hedgecock, Dennis (1994). "Does variance in reproductive success limit effective population sizes of marine organisms". *Genetics and evolution of aquatic organisms* 122, pp. 122–134.

Hedgecock, Dennis and Alexander I Pudovkin (2011). "Sweepstakes Reproductive Success in Highly Fecund Marine Fish and Shellfish: A Review and Commentary". *Bull. Mar. Sci.* 87.4, pp. 971–1002.

Hellmann, Ines et al. (2008). "Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals". *Genome Res.* 18.7, pp. 1020–1029.

Hermisson, Joachim and Pleuni S Pennings (2005). "Soft sweeps: molecular population genetics of adaptation from standing genetic variation". *Genetics* 169.4, pp. 2335–2352.

Hernandez, Ryan D et al. (2011). "Classic selective sweeps were rare in recent human evolution". *Science* 331.6019, pp. 920–924.

Hu, Tina T, Michael B Eisen, Kevin R Thornton, and Peter Andolfatto (2013). "A second-generation assembly of the Drosophila simulans genome provides new insights into patterns of lineage-specific divergence". *Genome Res.* 23.1, pp. 89–98.

Hudson, R R and N L Kaplan (1995). "Deleterious background selection with recombination". *Genetics* 141.4, pp. 1605–1617.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Hudson, Richard R and Norman L Kaplan (1994). "Gene Trees with Background Selection". *Non-Neutral Evolution: Theories and Molecular Data*. Ed. by Brian Golding. Boston, MA: Springer US, pp. 140–153.

IUCN (2020). *The IUCN Red List of Threatened Species*. `https://www.iucnredlist.org`. Accessed: 2020-10-31.

Jensen, Jeffrey D, Kevin R Thornton, and Peter Andolfatto (2008). "An approximate bayesian estimator suggests strong, recurrent selective sweeps in Drosophila". *PLoS Genet.* 4.9, e1000198.

Kaplan, N L, R R Hudson, and C H Langley (1989). "The "hitchhiking effect" revisited". *Genetics* 123.4, pp. 887–899.

Karasov, Talia, Philipp W Messer, and Dmitri A Petrov (2010). "Evidence that adaptation in Drosophila is not limited by mutation at single sites". *PLoS Genet.* 6.6, e1000924.

Kimura, M and J F Crow (1964). "The number of alleles that can be maintained in a finite population". *Genetics* 49, pp. 725–738.

Kimura, Motoo (1984). *The neutral theory of molecular evolution*. Cambridge University Press.

Kondrashov, Fyodor A and Alexey S Kondrashov (2010). "Measurements of spontaneous rates of mutations in the recent past and the near future". *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365.1544, pp. 1169–1176.

Leffler, Ellen M et al. (2012). "Revisiting an Old Riddle: What Determines Genetic Diversity Levels within Species?" *PLoS Biol.* 10.9, e1001388–9.

Lewontin, Richard C (1974). *The genetic basis of evolutionary change*. Vol. 560. Columbia University Press New York.

Lewontin, Richard C, Rama S Singh, and Marcy K Uyenoyama (2004). "Building a science of population biology". *The Evolution of Population Biology*. Cambridge University Press, pp. 7–20.

Li, Haipeng and Wolfgang Stephan (2006). "Inferring the demographic history and rate of adaptive substitution in Drosophila". *PLoS Genet.* 2.10, e166.

Lynch, Michael (1991). "Methods for the analysis of comparative data in evolutionary biology". *Evolution* 45.5, pp. 1065–1080.

— (2010). "Evolution of the mutation rate". *Trends Genet.* 26.8, pp. 345–352.

— (2011). "Statistical inference on the mechanisms of genome evolution". *PLoS Genet.* 7.6, e1001389.

Mace, Georgina M and Russell Lande (1991). "Assessing Extinction Threats: Toward a Reevaluation of IUCN Threatened Species Categories". *Conserv. Biol.* 5.2, pp. 148–157.

Macpherson, J Michael, Guy Sella, Jerel C Davis, and Dmitri A Petrov (2007). "Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in Drosophila". *Genetics* 177.4, pp. 2083–2099.

Malécot, Gustave (1948). *Mathématiques de l'hérédité*. Paris: Masson.

Maruyama, T and M Kimura (1980). "Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent". *Proc. Natl. Acad. Sci. U. S. A.* 77.11, pp. 6710–6714.

Maynard Smith, John and John Haigh (1974). "The hitch-hiking effect of a favourable gene". *Genet. Res.* 23.1, pp. 23–35.

McVicker, Graham, David Gordon, Colleen Davis, and Phil Green (2009). "Widespread genomic signatures of natural selection in hominid evolution". *PLoS Genet.* 5.5, e1000471.

Mora, Camilo, Derek P Tittensor, Sina Adl, Alastair G B Simpson, and Boris Worm (2011). "How many species are there on Earth and in the ocean?" *PLoS Biol.* 9.8, e1001127.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Mukai, T (1988). *Genotype-environment interaction in relation to the maintenance of genetic variability in populations of Drosophila melanogaster*. Proceedings of the Second International Conference On Quantitative Genetics.

Mukai, Terumi (1985). "Experimental Verification of the Neutral Theory". *Population Genetics and Molecular Evolution*. Ed. by Tomoko Ohta and Kenichi Aoki. Berlin: Springer-Verlag, pp. 125–145.

Nei, Masatoshi and Dan Graur (1984). "Extent of protein polymosphism and the neutral mutation theory". *Evol. Biol.* 17, pp. 73–118.

Nevo, E (1978). "Genetic variation in natural populations: patterns and theory". *Theor. Popul. Biol.* 13.1, pp. 121–177.

Nevo, Eviatar, Avigdor Beiles, and Rachel Ben-Shlomo (1984). "The Evolutionary Significance of Genetic Diversity: Ecological, Demographic and Life History Correlates". *Evolutionary Dynamics of Genetic Diversity*. Springer Berlin Heidelberg, pp. 13–213.

Nicolaisen, Lauren E and Michael M Desai (2012). "Distortions in genealogies due to purifying selection". *Mol. Biol. Evol.* 29.11, pp. 3589–3600.

Nordborg, Magnus, Brian Charlesworth, and Deborah Charlesworth (1996). "The effect of recombination on background selection*". *Genet. Res.* 67.02, pp. 159–174.

Novak, Sebastian and Nicholas H Barton (2017). "When Does Frequency-Independent Selection Maintain Genetic Variation?" *Genetics* 207.2, pp. 653–668.

Nunney, Leonard (1993). "The influence of mating system and overlapping generations on effective population size". *Evolution* 47.5, pp. 1329–1341.

— (1996). "The influence of variation in female fecundity on effective population size". *Biol. J. Linn. Soc. Lond.* 59.4, pp. 411–425.

Nydam, Marie L and Richard G Harrison (2010). "Polymorphism and divergence within the ascidian genus Ciona". *Mol. Phylogenet. Evol.* 56.2, pp. 718–726.

O'Meara, Brian et al. (2020). *datelife: Go from a List of Taxa or a Tree to a Chronogram using Open Scientific Data*.

Obbard, Darren J et al. (2012). "Estimating divergence dates and substitution rates in the Drosophila phylogeny". *Mol. Biol. Evol.* 29.11, pp. 3459–3473.

Ohta, T and M Kimura (1973). "A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population". *Genet. Res.* 22.2, pp. 201–204.

Palkopoulou, Eleftheria et al. (2015). "Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth". *Curr. Biol.* 25.10, pp. 1395–1400.

Palstra, Friso P and Dylan J Fraser (2012). "Effective/census population size ratio estimation: a compendium and appraisal". *Ecol. Evol.* 2.9, pp. 2357–2365.

Palstra, Friso P and Daniel E Ruzzante (2008). "Genetic estimates of contemporary effective population size: what can they tell us about the importance of genetic stochasticity for wild population persistence?" *Mol. Ecol.* 17.15, pp. 3428–3447.

Pebesma, Edzer J (2018). "Simple features for R: Standardized support for spatial vector data". *R J.* 10.1, p. 439.

Pennell, M W et al. (2014). *geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to ph ylogenetic trees*.

Pennings, Pleuni S and Joachim Hermisson (2006). "Soft sweeps II–molecular population genetics of adaptation from recurrent mutation or migration". *Mol. Biol. Evol.* 23.5, pp. 1076–1084.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Powell, Jeffrey R (1975). "Protein variation in natural populations of animals". *Evolutionary Biology: Volume 8*. Ed. by Dobzhansky, Theodosius, and Hecht, Max K. and Steere, William C. New York: Plenum Press, pp. 79–199.

Ralph, Peter and Graham Coop (2010). "Parallel Adaptation: One or Many Waves of Advance of an Advantageous Allele?" *Genetics* 186.2, pp. 647–668.

Ralph, Peter L and Graham Coop (2015). "The Role of Standing Variation in Geographic Convergent Adaptation*". *Am. Nat.* 186.S1, S5–S23.

Revell, Liam J (2012). "phytools: an R package for phylogenetic comparative biology (and other things)". *Methods Ecol. Evol.* 3.2, pp. 217–223.

Robertson, Alan (1961). "Inbreeding in artificial selection programmes". *Genet. Res.* 2.2, pp. 189–194.

Romiguier, J et al. (2014). "Comparative population genomics in animals uncovers the determinants of genetic diversity". *Nature* 515.7526, pp. 261–263.

Roux, Camille et al. (2016). "Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence". *PLoS Biol.* 14.12, e2000234.

Sanchez-Reyes, Luna L and Brian O'Meara (2019). *datelife: Leveraging databases and analytical tools to reveal the dated Tree of Life*.

Santiago, E and A Caballero (1995). "Effective size of populations under selection". *Genetics* 139.2, pp. 1013–1030.

— (1998). "Effective size and polymorphism of linked neutral loci in populations under directional selection". *Genetics* 149.4, pp. 2105–2117.

Sella, Guy, Dmitri A Petrov, Molly Przeworski, and Peter Andolfatto (2009). "Pervasive natural selection in the Drosophila genome?" *PLoS Genet.* 5.6, e1000495.

Shapiro, Joshua A et al. (2007). "Adaptive genic evolution in the Drosophila genomes". *Proceedings of the National Academy of Sciences* 104.7, pp. 2271–2276.

Slatkin, M (1977). "Gene flow and genetic drift in a species subject to frequent local extinctions". *Theor. Popul. Biol.* 12.3, pp. 253–262.

Small, Kerrin S, Michael Brudno, Matthew M Hill, and Arend Sidow (2007). "Extreme genomic variation in a natural population". *Proc. Natl. Acad. Sci. U. S. A.* 104.13, pp. 5698–5703.

Soulé, M E (1976). "Allozyme variation, its determinants in space and time". *Molecular evolution*. Ed. by F J Ayala. Sunderland, Massachusetts: Sinauer Associates, pp. 60–77.

South, Andy (2017). *rnaturalearth: World Map Data from Natural Earth*.

Spielman, Derek, Barry W Brook, and Richard Frankham (2004). "Most species are not driven to extinction before genetic factors impact them". *Proc. Natl. Acad. Sci. U. S. A.* 101.42, pp. 15261–15264.

Stan Development Team (2020). *Stan Modeling Language Users Guide and Reference Manual*.

Stapley, Jessica, Philine G D Feulner, Susan E Johnston, Anna W Santure, and Carole M Smadja (2017). "Variation in recombination frequency and distribution across eukaryotes: patterns and processes". *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 372.1736.

Stephan, W (1995). "An improved method for estimating the rate of fixation of favorable mutations based on DNA polymorphism data". *Mol. Biol. Evol.* 12.5, pp. 959–962.

Stephan, W and C H Langley (1998). "DNA polymorphism in lycopersicon and crossing-over per physical length". *Genetics* 150.4, pp. 1585–1593.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Stephan, Wolfgang, Thomas H E Wiehe, and Marcus W Lenz (1992). "The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory". *Theor. Popul. Biol.* 41.2, pp. 237–254.

Tajima, F (1996). "The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites". *Genetics* 143.3, pp. 1457–1465.

Tsagkogeorga, Georgia, Vincent Cahais, and Nicolas Galtier (2012). "The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate Ciona intestinalis". *Genome Biol. Evol.* 4.8, pp. 740–749.

Uyeda, Josef C, Rosana Zenil-Ferguson, and Matthew W Pennell (2018). "Rethinking phylogenetic comparative methods". *Syst. Biol.* 67.6, pp. 1091–1109.

Vehtari, Aki, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner (2019). "Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC". arXiv: `1903.08008 [stat.CO]`.

Villemereuil, Pierre de and Shinichi Nakagawa (2014). "General Quantitative Genetic Methods for Comparative Biology". *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice*. Ed. by László Zsolt Garamszegi. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 287–303.

Wang, J (2005). "Estimation of effective population sizes from data on genetic markers". *Proceedings of the Royal Society of London B: Biological Sciences* 360.1459, pp. 1395–1409.

Wang, J, E Santiago, and A Caballero (2016). "Prediction and estimation of effective population size". *Heredity* 117.4, pp. 193–206.

Waples, R S (1989). "A generalized approach for estimating effective population size from temporal changes in allele frequency". *Genetics* 121.2, pp. 379–391.

Waples, Robin S, Peter M Grewe, Mark W Bravington, Richard Hillary, and Pierre Feutry (2018). "Robust estimates of a high Ne/N ratio in a top marine predator, southern bluefin tuna". *Sci Adv* 4.7, eaar7759.

Waples, Robin S, Gordon Luikart, James R Faulkner, and David A Tallmon (2013). "Simple life-history traits explain key effective population size ratios across diverse taxa". *Proc. Biol. Sci.* 280.1768, p. 20131339.

Westoby, Mark, Michelle R Leishman, and Janice M Lord (1995). "On misinterpreting the 'phylogenetic correction'". *Journal of Ecology* 83, pp. 531–534.

Whitney, Kenneth D and Theodore Garland (2010). "Did Genetic Drift Drive Increases in Genome Complexity?" *PLoS Genet.* 6.8, e1001080–6.

Wilfert, L, J Gadau, and P Schmid-Hempel (2007). "Variation in genomic recombination rates among animal taxa and the case of social insects". *Heredity* 98.4, pp. 189–197.

Wright, S (1948). "On the roles of directed and random changes in gene frequency in the genetics of populations". *Evolution* 2.4, pp. 279–294.

Wright, Sewall (1931). "Evolution in Mendelian populations". *Genetics* 16.2, p. 97.

— (1938). "Size of population and breeding structure in relation to evolution". *Science* 87.2263, pp. 430–431.

Yukilevich, R (2017). *Drosophila Speciation Patterns*. `www.Drosophila-speciation-patterns.com`. Accessed: 2020-5-27.

Yukilevich, Roman (2012). "Asymmetrical patterns of speciation uniquely support reinforcement in Drosophila". *Evolution* 66.5, pp. 1430–1446.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Zhao, Shancen et al. (2013). "Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation". *Nat. Genet.* 45.1, pp. 67–71.

Supplementary Table and Figures

S1 Diversity and IUCN Red List Status

	mean	2.5 %	97.5 %
β_0	-2.80	-3.10	-2.50
β_{LC}	-0.37	-0.55	-0.19
β_{NT}	-0.20	-0.80	0.41
β_{VU}	-0.31	-0.81	0.18
β_{EN}	-0.38	-0.71	-0.05
β_{CR}	-0.01	-0.63	0.60
β_{N_c}	0.07	0.05	0.10

Table S1: The regression estimates of full IUCN Red List population size model for diversity, $\log_{10}(\pi) = \beta_0 + \beta_{LC}LC + \beta_{NT}NT + \beta_{VU}VU + \beta_{EN}EN + \beta_{CR}CR + \beta_{N_c}\log_{10}(N_c)$; $df = 165$. Using AIC to compare this full model to a reduced model of $\log_{10}(\pi) = \beta_0 + \beta_{N_c}\log_{10}(N_c)$, $AIC_{\text{full}} = 203.18$, $AIC_{\text{reduced}} = 212.41$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

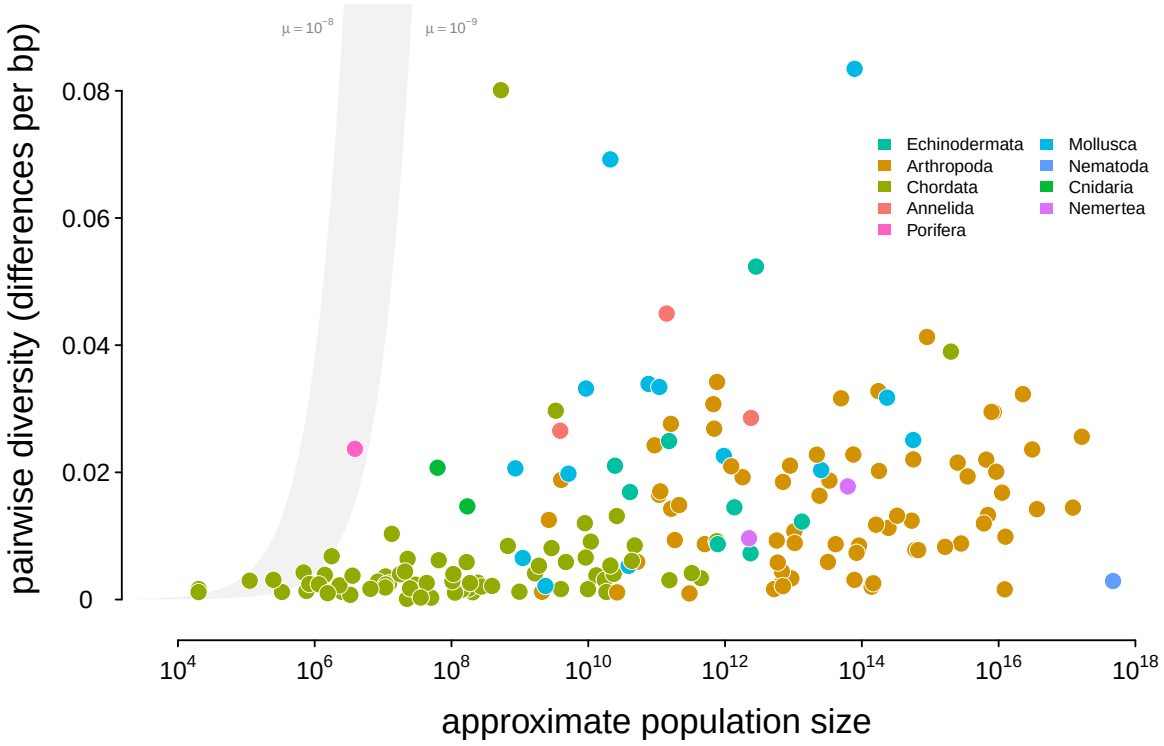

Figure S1: A version of Figure 2 with diversity on a linear, rather than log, scale. Points are colored by phylum, and the shaded region is the predicted neutral level of diversity assuming $N_e = N_c$ with mutation range ranging between $10^{-10} \leq \mu \leq 10^{-8}$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

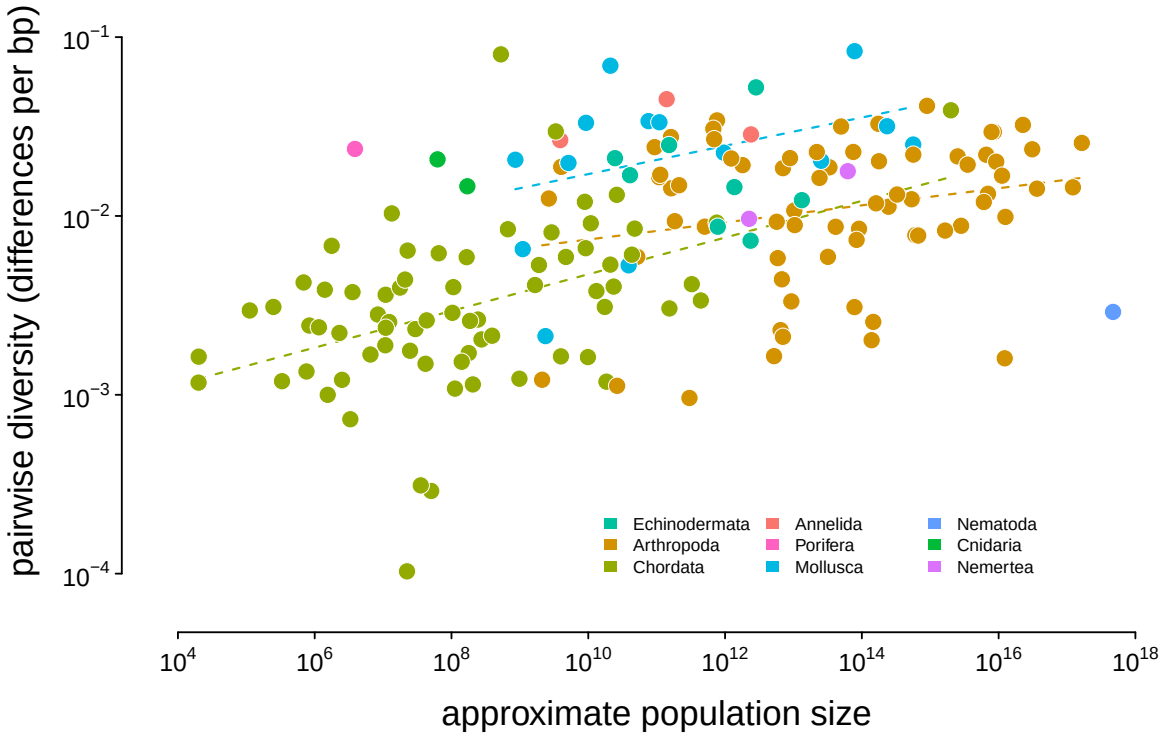

Figure S2: Diversity and approximate population size for 172 taxa, colored by phylum; the dashed lines indicate the non-phylogenetic OLS estimates of the relationship between population size and diversity grouped by phyla.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

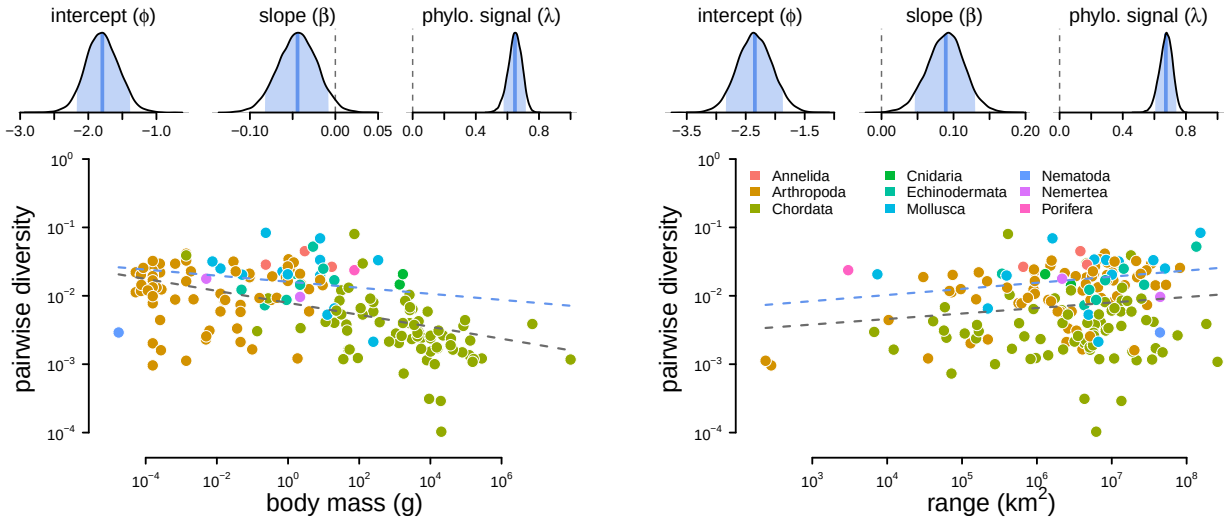

Figure S3: The relationship between diversity (differences per basepair) and body mass (left) and range (right) across 172 species. The top row are posterior distributions of parameters estimated using the phylogenetic mixed-effects model using 166 taxa in the synthetic phylogeny for the intercept, slope, and phylogenetic signal from the mixed-effects model. The bottom row contain each species as a point, colored by phyla. The gray dashed line is the non-phylogenetic standard regression estimate, and the blue dashed line is the relationship fit by the phylogenetic mixed-effects model.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

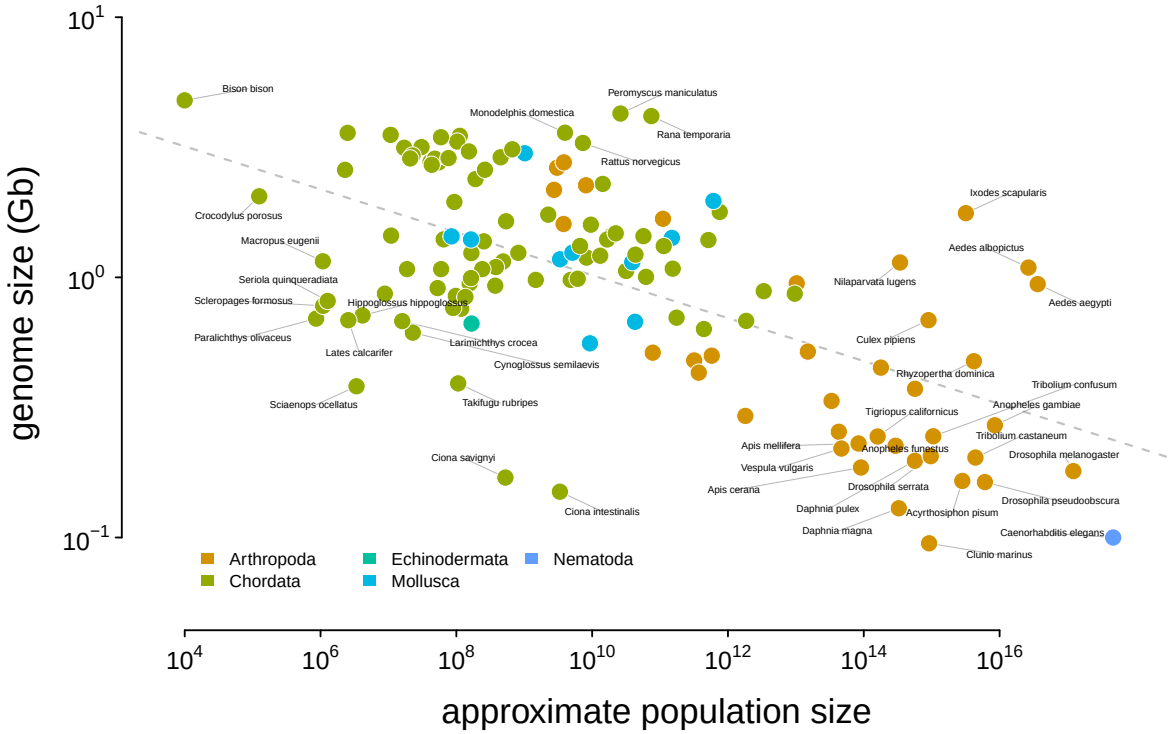

Figure S4: The relationship between genome size and approximate census population size. The dashed gray line indicates the OLS fit. Tiger salamander (*Ambystoma tigrinum*) was excluded because of its exceptionally large genome size (30Gbp).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

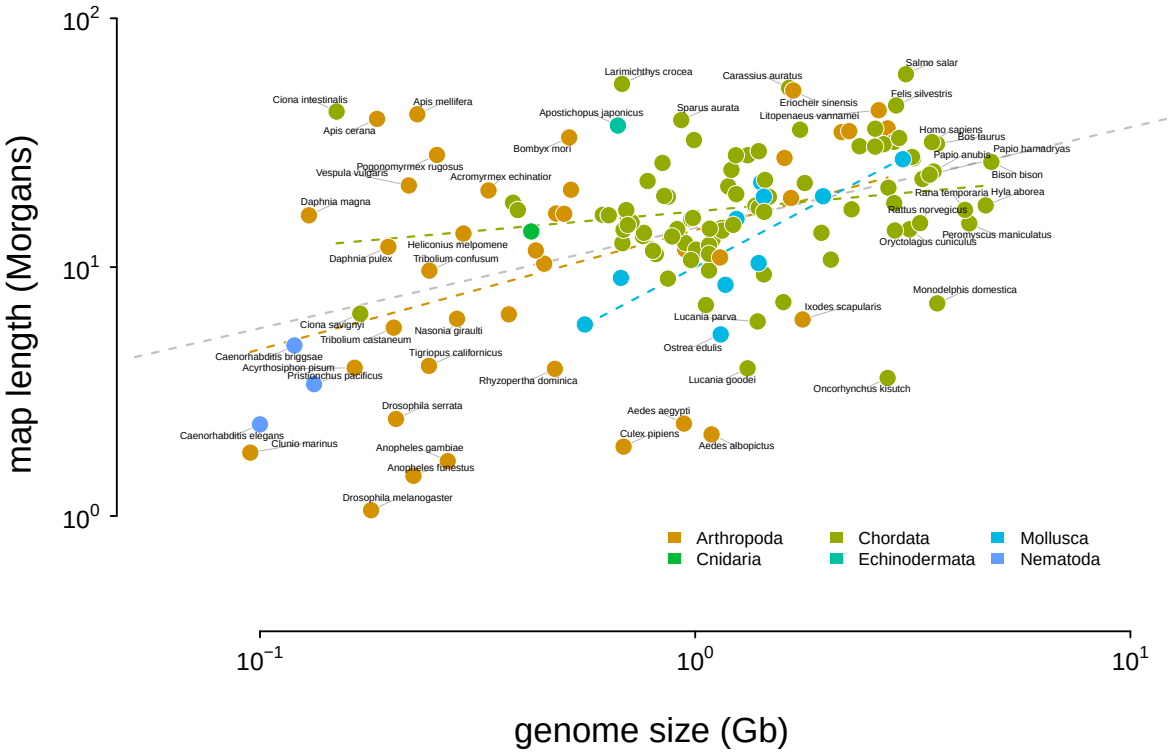

Figure S5: The relationship between genome size and recombination map length. The dashed gray line indicates the OLS fit for all taxa, and the dashed colored dashed lines indicate the linear relationship fit by phyla. Tiger salamander (*Ambystoma tigrinum*) was excluded because of its exceptionally large genome size (30Gbp).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

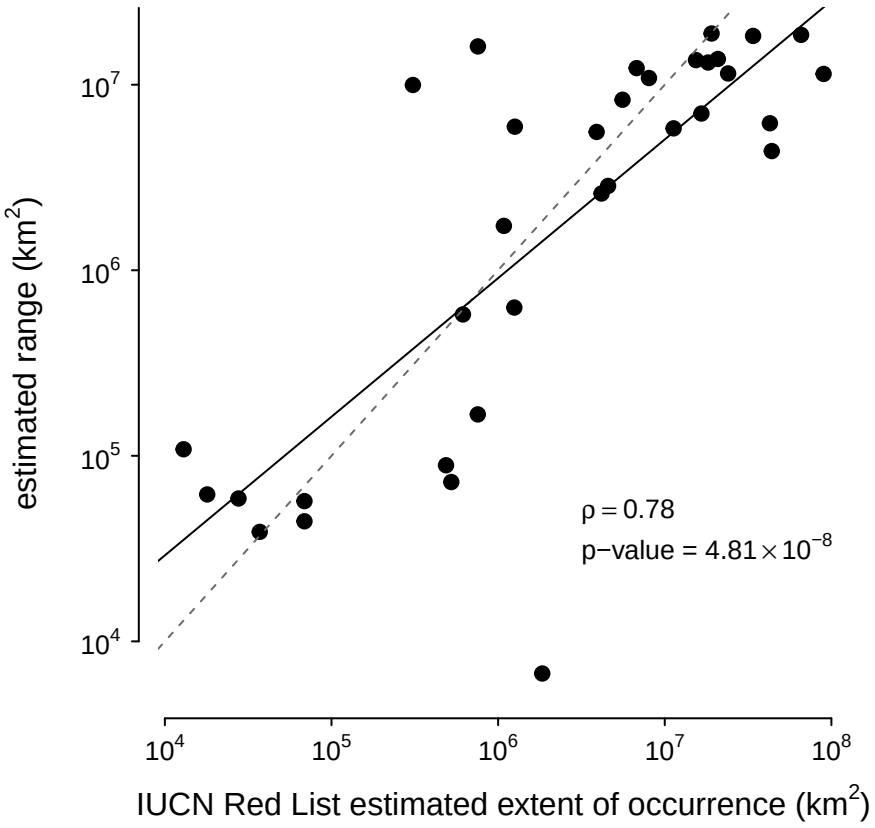

Figure S6: The correspondence between the ranges estimated with the alpha hull method applied to GBIF data used in this paper and IUCN Red List's Extent of Occurrence for the subset of species in both datasets. Note that the IUCN Red List contains predominantly endangered species, which leads to ascertainment bias; still, the high correlation between the estimated ranges shows the alpha hull method works well.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

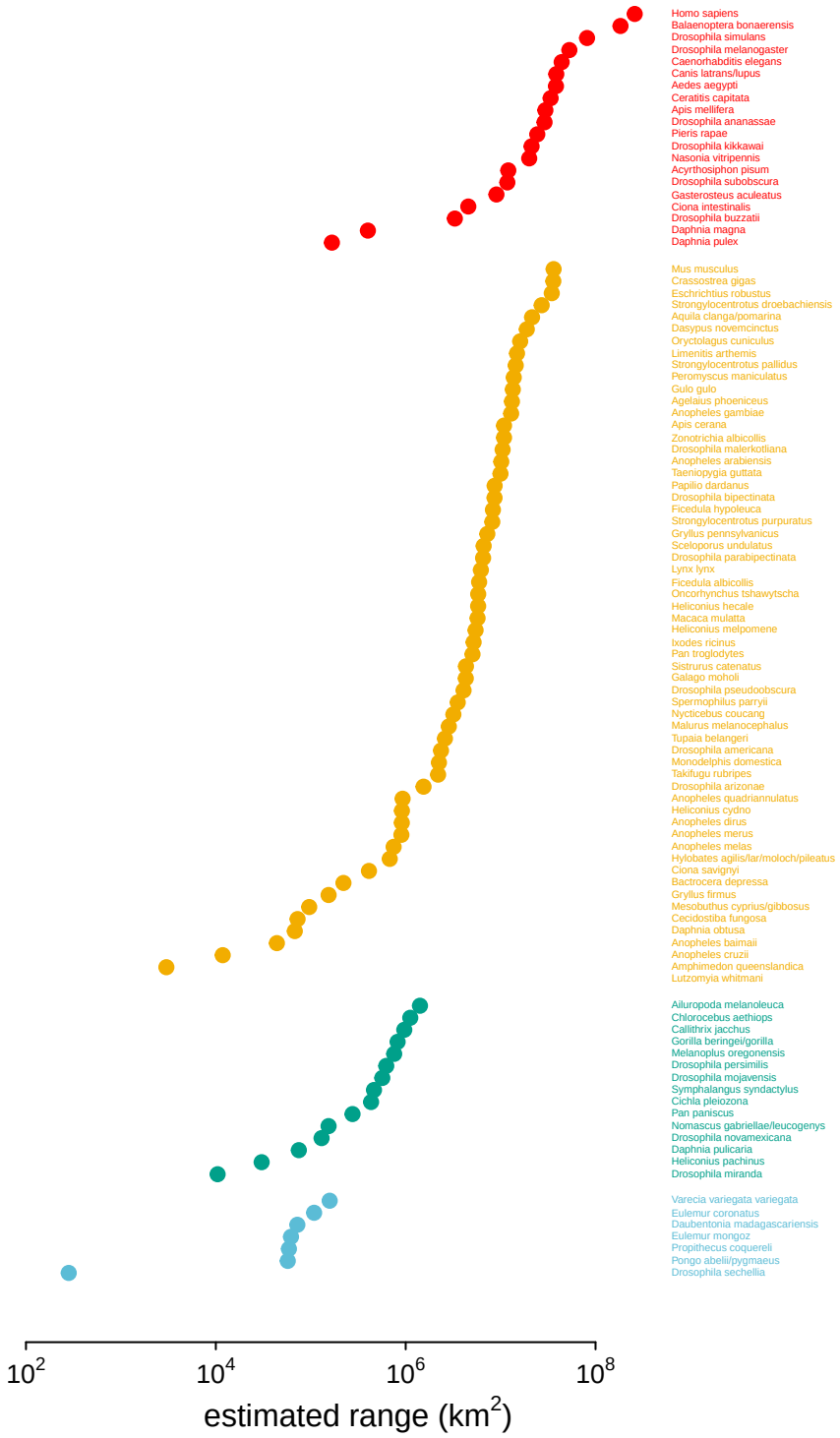

Figure S7: The estimated ranges using GBIF occurrence data, ordered within and colored by the original range category labels assigned in Leffler et al. (2012).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

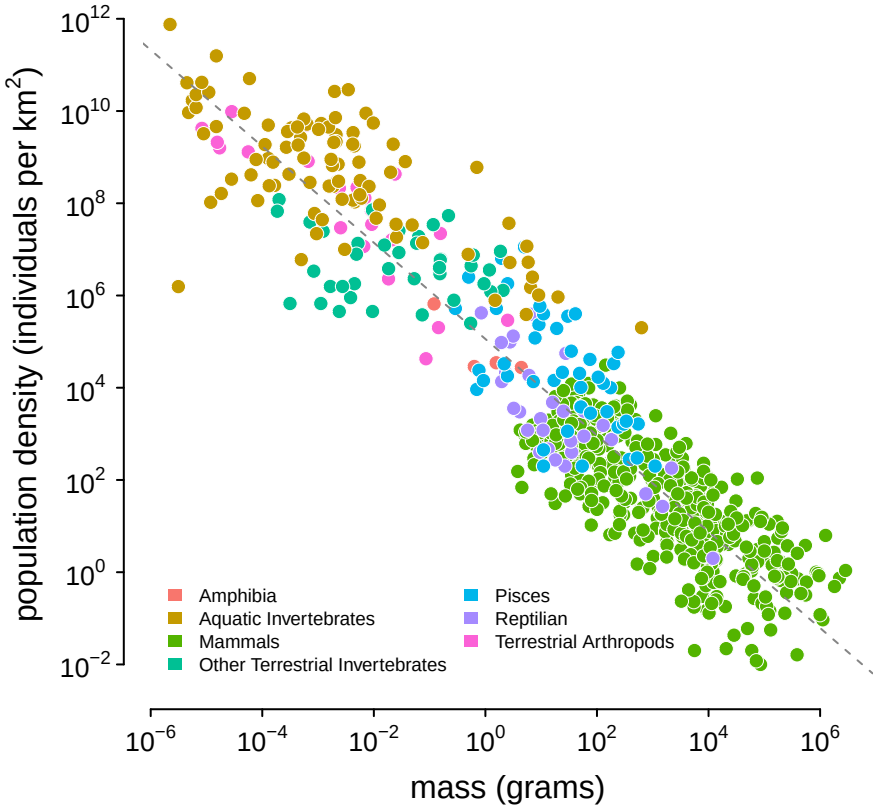

Figure S8: The appendix table of Damuth (1987); the color indicates Damuth's original group labels. The dashed line was estimated using a lognormal regression model in Stan. References to each measurement are available in Damuth (1987).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

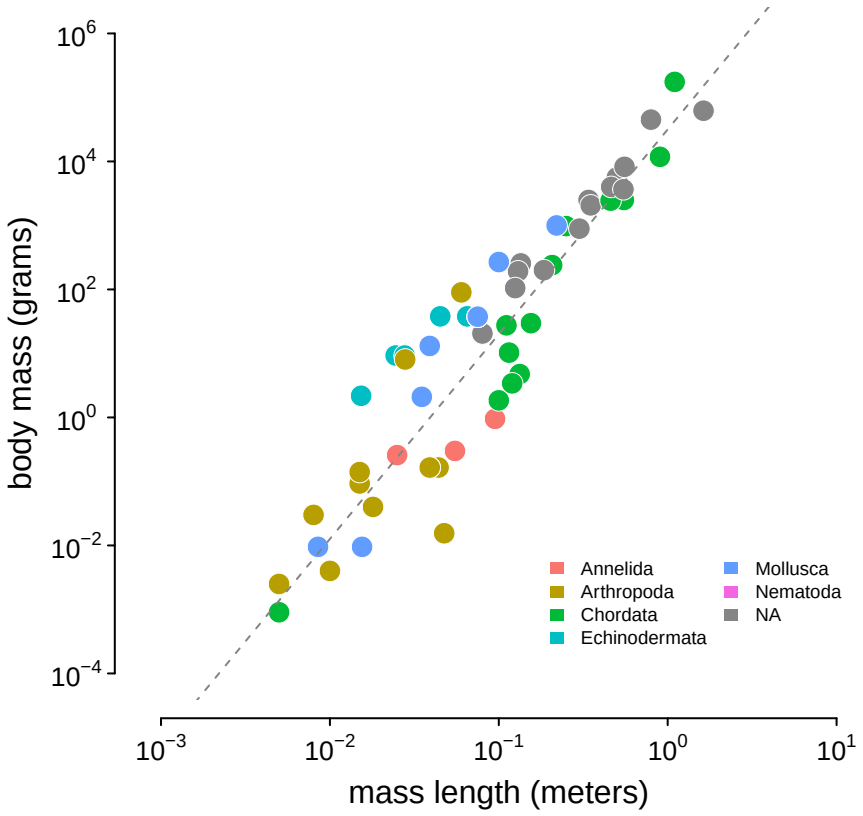

Figure S9: The relationship between body length (meters) and body mass (grams) in the Romiguier et al. (2014) data set, used to infer body masses for taxa. The gray dashed line is the line of best fit inferred using Stan.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

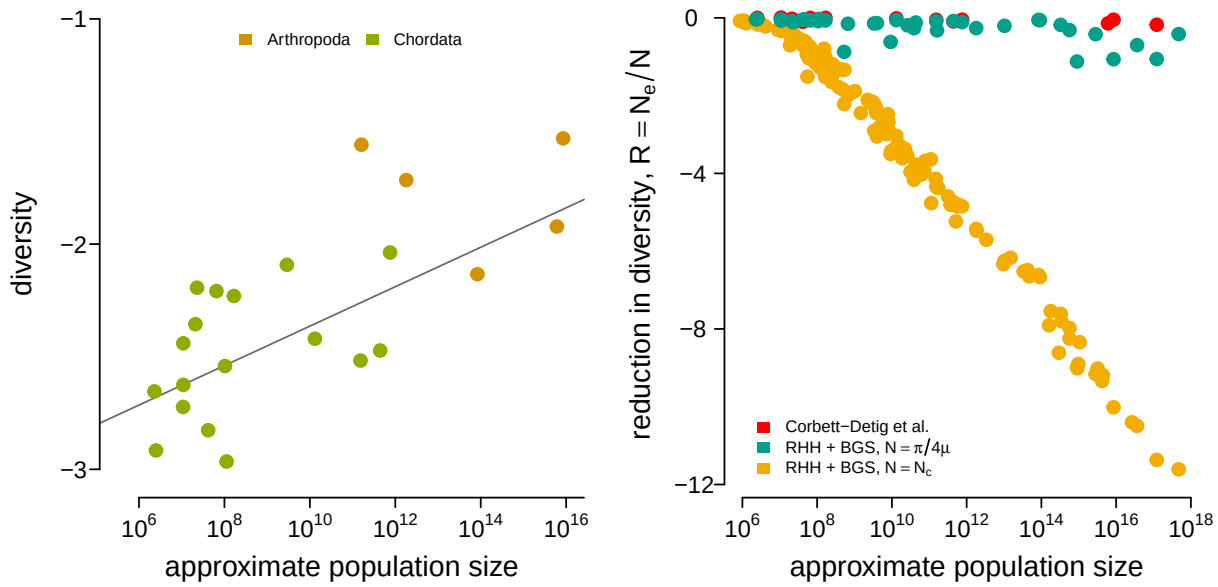

Figure S10: (A) The diversity data from Corbett-Detig et al. (2015) and the census population size estimated here for metazoan taxa. The blue point is selfing *C. elegans* and was excluded from the OLS fit. (B) The reductions in diversity, $R = {}^{N_e}/_{N}$, plotted against census size across species. The red points are the reductions estimated by Corbett-Detig et al. (2015). This confirms Corbett-Detig et al.'s (2015) finding that the impact of selection ($I = 1 - R$) increases with census population size (though, in the original paper size body size and range were used as separate proxy variables for census population size). The green and red points are the predicted reduction in diversity under the recurrent hitchhiking (RHH) and background selection (BGS) model using the *Drosophila melanogaster* parameters as described in the main text. The reduction in the diversity due to sweeps, from Equation (1), is determined by the term $2NS$. Green points treat N as the implied effective population size from diversity $\widetilde{N}_e = {}^{\widehat{\pi}}/_{4\mu}$, assuming $\mu = 10^{-9}$. Yellow points treat N as the census size, $N = N_c$. Overall, using the census size, e.g. $2N_cS$, leads to reductions in diversity that far exceed the empirical estimates of Corbett-Detig et al. and reasonable model-based predictions from $\widetilde{N}_e$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

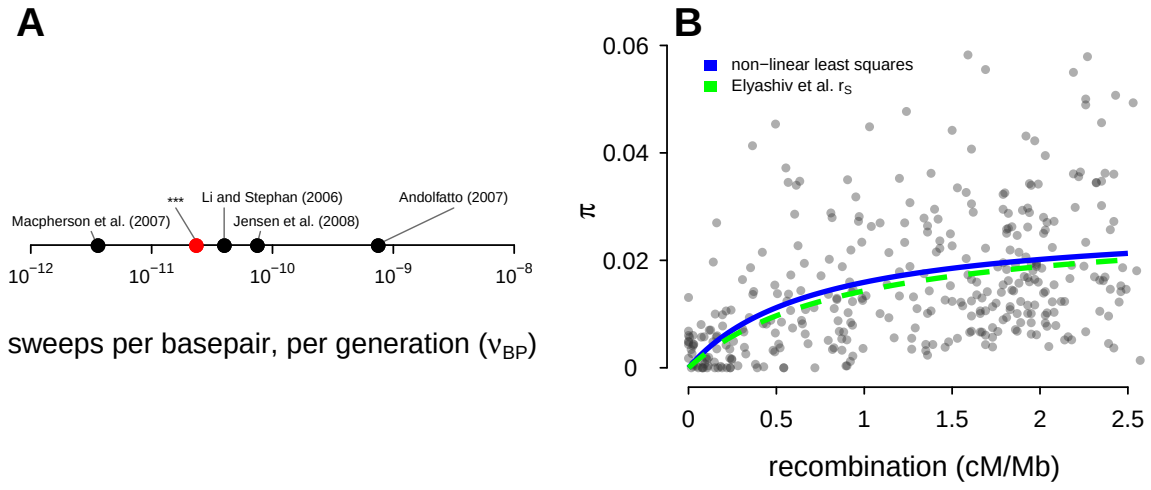

Figure S11: (A) The estimate of the number of sweeps per basepair, per genome (ν_{BP}) from Table 2 of Elyashiv et al. (2016) (the studies included are Andolfatto 2007; Li and Stephan 2006; Macpherson et al. 2007 and Jensen et al. 2008). (B) Points are the data from Shapiro et al. (2007). The blue line is the non-linear least squares fit to the data, and the green dashed line is the sweep model parameterized by the genome-wide average sweep coalescent rate $2NS \approx 0.92$ from the classic sweep and background selection model of Elyashiv et al. (2016) (r_s in Supplementary Table S6).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

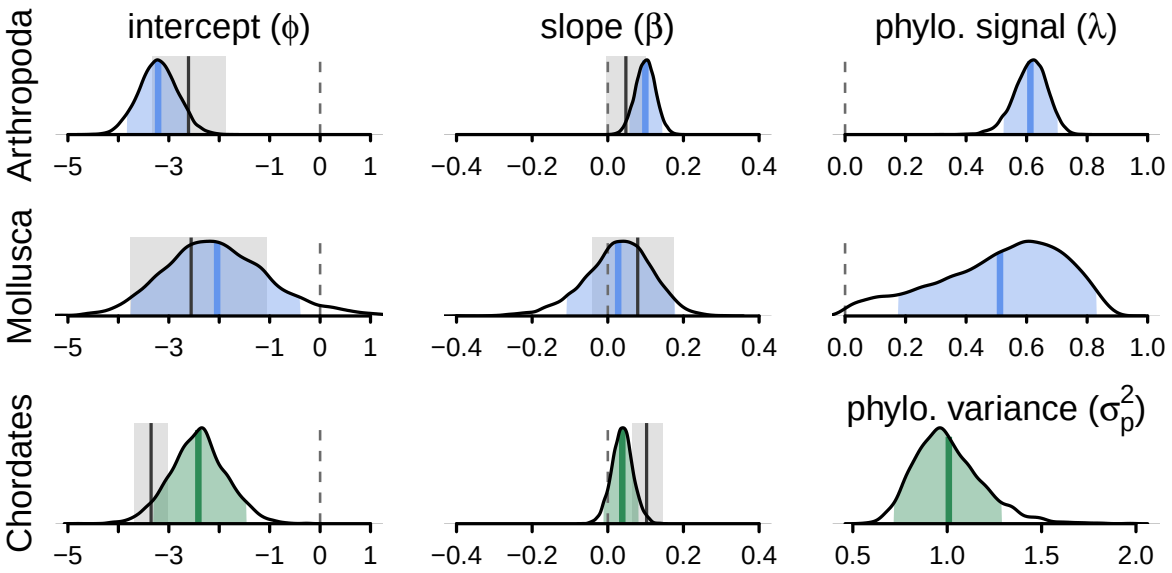

Figure S12: The posterior distributions for the parameters of the phylogenetic mixed-effects model of diversity and population size (this is analogous to Figure 4B) fit separately on chordates ($n = 68$), molluscs ($n = 13$), and arthropods ($n = 68$). The phylogenetic mixed-effects model for chordates indicated the best-fitting model had no residual variance ($\sigma_r^2 = 0$), so an alternate model without this variance component was used to ensure proper convergence; this model is shown in green. The light blue (green) shaded regions are the 90% credible intervals, the blue (green) lines the posterior averages, the gray shaded regions the OLS bootstrap 95% confidence intervals, and the gray lines the OLS estimate. Note that unlike Figure 4, the OLS estimate uses all taxa, not just those present in the phylogeny, since splitting the data by phyla reduces sample sizes (OLS with just the subset of taxa in the phylogeny is not significant for either chordates and arthropods). The vertical dashed gray line indicates zero.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

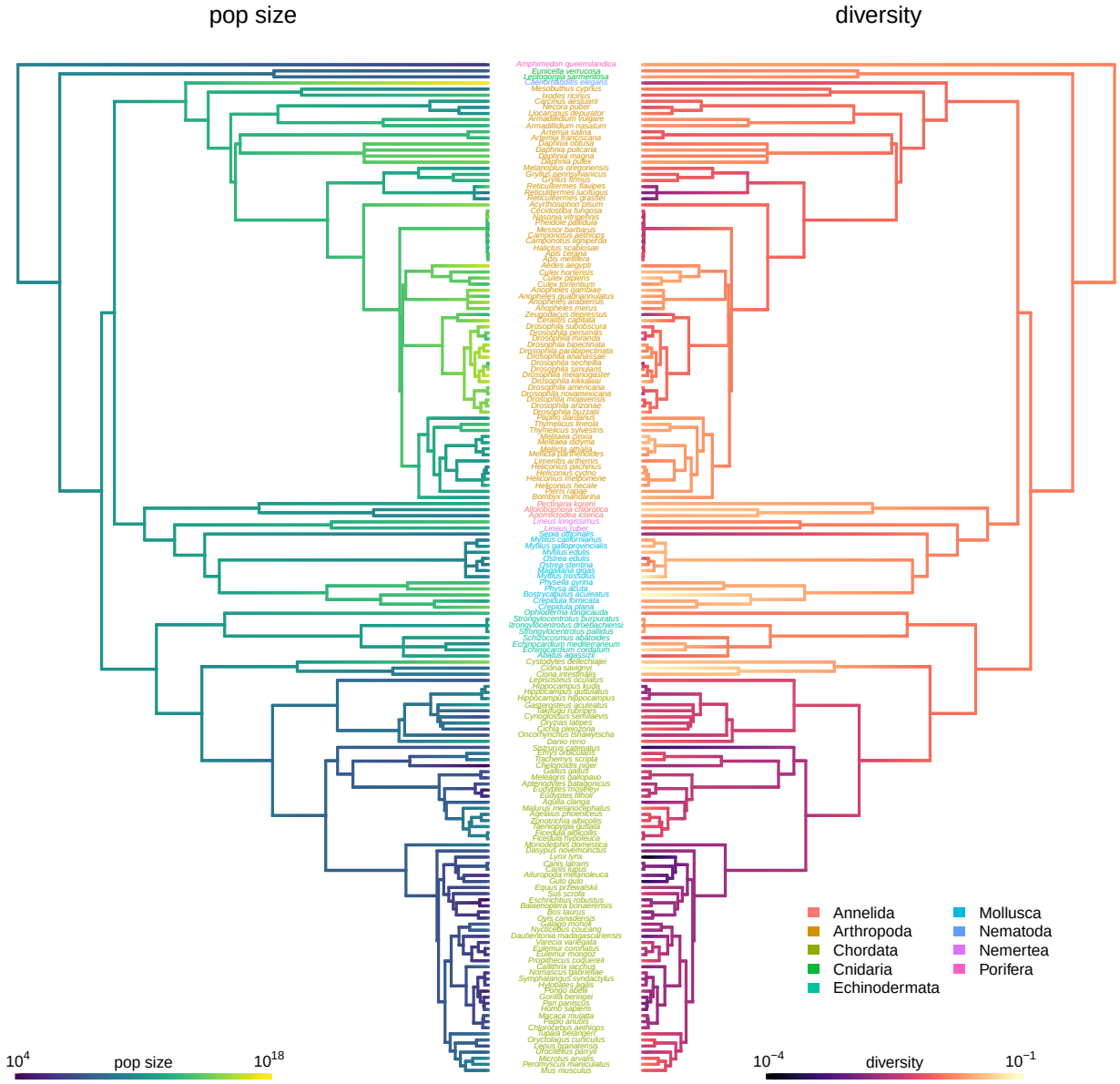

Figure S13

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.03.429633; this version posted February 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

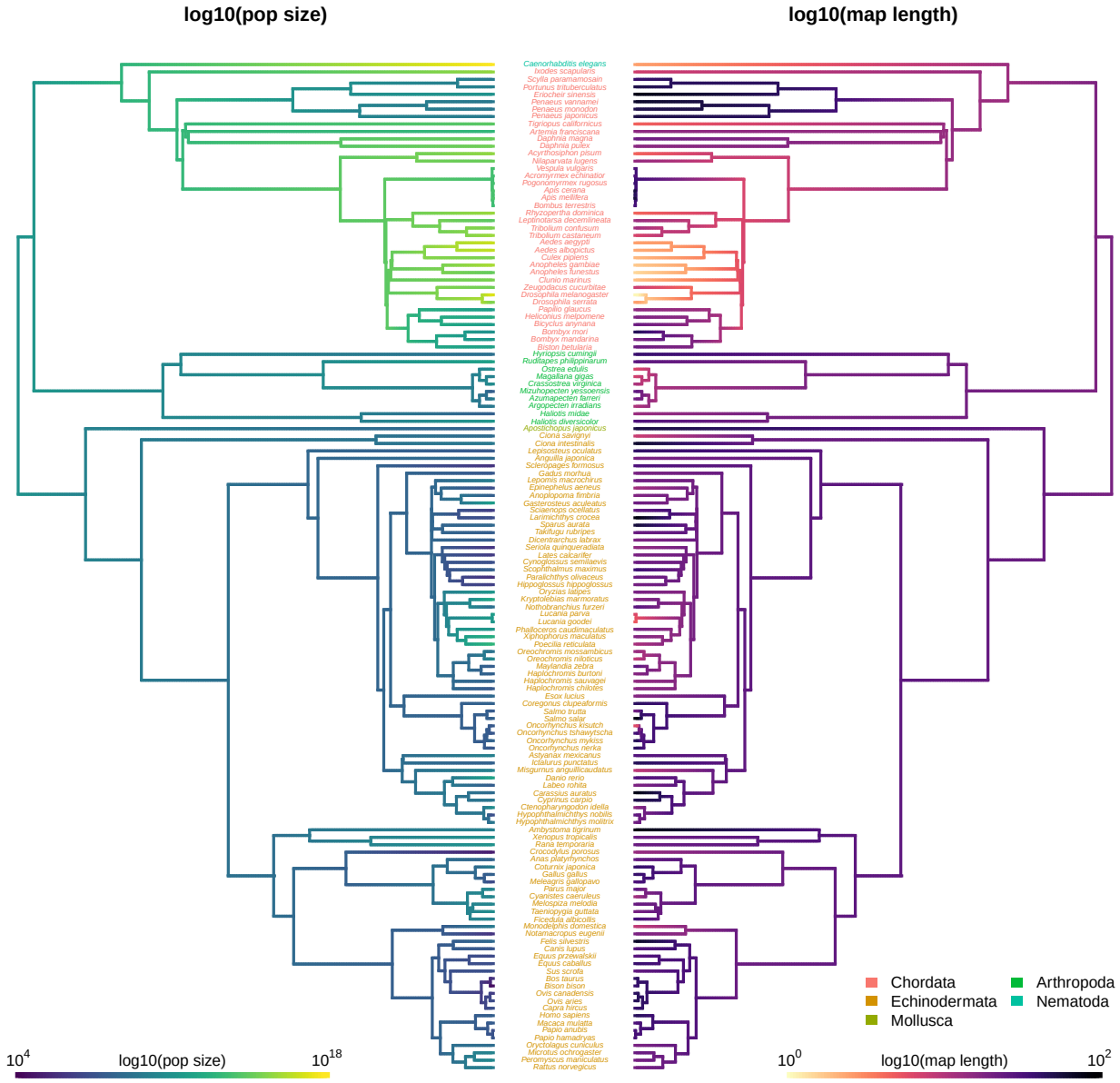

Figure S14