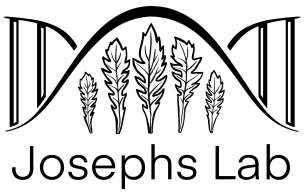


# Previously unmeasured genetic diversity explains part of Lewontin's paradox in a *k*-mer-based meta-analysis of 112 plant species

Miles Roberts[1*] & Emily B. Josephs[2,3,4]

**1 Genetics and Genome Sciences Program, Michigan State University, East Lansing MI**
**2 Department of Plant Biology, Michigan State University, East Lansing, MI**
**3 Ecology, Evolution, and Behavior Program, Michigan State University, East Lansing, MI**
**4 Plant Resilience Institute, Michigan State University, East Lansing, MI**

*** milesdroberts@gmail.com**

## Abstract

At the molecular level, most evolution is expected to be neutral. A key prediction of this expectation is that the level of genetic diversity in a population should scale with population size. However, as was noted by Richard Lewontin in 1974 and reaffirmed by later studies, the slope of the population size-diversity relationship in nature is much weaker than expected under neutral theory. We hypothesize that one contributor to this paradox is that current methods relying on single nucleotide polymorphisms (SNPs) called from aligning short reads to a reference genome underestimate levels of genetic diversity in many species. To test this idea, we calculated nucleotide diversity ($\pi$) and *k*-mer-based metrics of genetic diversity across 112 plant species, amounting to over 205 terabases of DNA sequencing data from 27,488 individual plants. We then compared how these different metrics correlated with proxies of population size that account for both range size and population density variation across species. We found that our population size proxies scaled anywhere from about 3 to over 20 times faster with *k*-mer diversity than nucleotide diversity after adjusting for evolutionary history, mating system, life cycle habit, cultivation status, and invasiveness. The relationship between *k*-mer diversity and population size proxies also remains significant after correcting for genome size, whereas the analogous relationship for nucleotide diversity does not. These results suggest that variation not captured by common SNP-based analyses explains part of Lewontin's paradox in plants.

## Lay Summary

Even after many revolutions in our ability to sequence and understand DNA, many important biological questions remain unsolved. One such problem is Lewontin's paradox, named after Richard Lewontin who first described it in 1974. The core of the paradox is a simple idea: species with more individuals should be more genetically diverse. The reasoning is that more individuals means more replication of DNA, and thus more opportunities for mutation to create new variation. However, species that differ massively in population size often have similar diversity levels. Lewontin's paradox has several potential, previously investigated mechanisms but what if one contributor is simply that our measurements of genetic diversity are off? Most studies estimate diversity by comparing sample genomes to a standard reference genome. While this approach is useful, it is impossible to measure variation in DNA that is not represented in the reference - a phenomenon known as reference bias. We estimate metrics of diversity that are free of reference-bias and re-investigate Lewontin's paradox in plants. Overall, we find that reference-free diversity metrics scale more with population size, compared to the reference-biased approach. While it is unlikely that reference-bias fully explains Lewontin's paradox, our analyses suggest that reference-bias plays an important role.

## Key words

Lewontin's paradox, k-mer, nucletide diversity, phylogenetic least squares, census size, neutral theory

# Introduction

Understanding the determinants of genetic diversity within populations is key to informing species conservation [Cole, 2003] and breeding efforts [Sanchez et al., 2023]. However, most species have far less genetic diversity (commonly estimated as pairwise nucleotide diversity, $\pi$) than expected [Frankham, 2012, Corbett-Detig et al., 2015, Buffalo, 2021]. If we assume that the vast majority of genetic variants are neutral, then the determinants of genetic diversity are encapsulated in neutral theory [Kimura, 1983]: $E[\pi] \approx 4N_e\mu$, where $E[\pi]$ is the expected level of genetic diversity, $N_e$ is the effective size of a population, and $\mu$ is the mutation rate per base pair per generation. Mutations rates vary relatively little across species (Cagan et al. [2022], Bergeron et al. [2023], reviewed in Quiroz et al. [2023]), while the total number of individuals in a species varies massively [Buffalo, 2021]. Thus, under neutral theory, population size should be a strong determinant of genetic diversity and species with larger population sizes should be more diverse. However, even some of the most abundant species studied to date have low genetic diversity compared to neutral theory expectations. For example, *Drosophila simulans* has an estimated population size $> 10^{14}$ and a diversity of $\pi \approx 0.01$, but an expected diversity of $\pi > 0.1$ [Buffalo, 2021]. This mismatch between expected and observed levels of neutral diversity across populations of varying size is known as Lewontin's paradox, named after Richard Lewontin who first described the phenomenon [Lewontin, 1974].

The potential mechanisms underlying Lewontin's paradox have been reviewed extensively [Leffler et al., 2012, Slotte, 2014, Ellegren and Galtier, 2016, Charlesworth and Jensen, 2022]. Multiple selective and demographic processes likely contribute to Lewontin's paradox; however, determining the relative importance of these processes remains a contentious area of research. The two most explored mechanisms are historic population size changes [Charlesworth and Jensen, 2022] and linked selection - whereby fixation or purging of selected alleles causes the loss of linked neutral alleles [Kojima and Schaffer, 1964, 1967, Smith and Haigh, 1974, Charlesworth et al., 1993, Charlesworth, 1994]. Linked selection is expected to reduce diversity more in regions of lower recombination and higher functional density [Slotte, 2014]. Thus, many studies have focused on measuring the correlations of recombination rate or functional density with either intraspecific or interspecific diversity, often observing significant correlations [Tenaillon et al., 2001, Hellmann et al., 2003, Nordborg et al., 2005, Roselius et al., 2005, Branca et al., 2011, Paape et al., 2012, Corbett-Detig et al., 2015, Silva-Junior and Grattapaglia, 2015, Wang et al., 2016, Phung et al., 2016, Mackintosh et al., 2019]. However, not all studies observe strong correlations between recombination and diversity [Schmid et al., 2005, Roselius et al., 2005, Flowers et al., 2012, Wang et al., 2016], especially studies focused on plant species (reviewed in Slotte [2014]), and such correlations could be explained by an association between recombination and mutation [Hellmann et al., 2003] (though the evidence for this is mixed, see Mackintosh et al. [2019]). There is also both empirical and theoretical evidence that linked selection is unlikely to explain the entirety of Lewontin's paradox, suggesting that demographic factors play an important role [Coop, 2016, Buffalo, 2021, Charlesworth and Jensen, 2022].

There are three main types of demographic changes proposed to contribute to Lewontin's paradox: contractions, expansions, and cyclical population size changes [Charlesworth and Jensen, 2022]. Population contractions cause loss of diversity. Thus, if many species' populations recently contracted (due to human activity, for example), then their contemporary diversity would be much lower than expected from their pre-contraction population sizes [Exposito-Alonso et al., 2022]. Recent population expansions could cause a similar mismatch. Because it takes many generations for populations to accumulate diversity compared to the timescale of typical expansions, contemporary diversity levels for an expanded population would be much smaller than expected from a post-expansion population size [Peart et al., 2020, Charlesworth and Jensen, 2022]. For a similar reason, species that have seasonal variation in their population sizes will also tend to have diversity levels closer to what one would expect based on their minimum size rather than their peak size [Wright, 1940]. Studies investigating Lewontin's paradox would ideally try to jointly infer these demographic histories alongside selective factors in natural populations. However, issues of model complexity and identifiability often prevent such joint estimation [Johri et al., 2020, 2022b,a], suggesting further explorations of Lewontin's paradox will require new approaches.

One potential, but rarely explored, contributor to Lewontin's paradox is that current methods for estimating genetic diversity systematically underestimate the true levels of genetic diversity in most populations. Lewontin's original observations and earlier studies on the population size-diversity relationship were based on allozymes, which detect variants in protein sequences [Lewontin, 1974, Nei and Graur, 1984]. More recent studies measure diversity using SNPs at more neutral four-fold degenerate sites (i.e. sites where mutations do not affect protein sequences) in DNA and generally observe greater within-species diversity and between-species divergence compared to allozymes [Li

and Sadler, 1991, Makałowski and Boguski, 1998, Bazin et al., 2006, Piganeau and Eyre-Walker, 2009]. However, current SNP-based methods are not perfect either and there is significant evidence that SNPs capture a biased and incomplete picture of genetic diversity. First, calling SNPs typically requires aligning reads to a reference genome, meaning any SNPs in regions that are not present or highly diverged from the reference genome will be excluded from analysis and thus downwardly bias diversity estimates [Golicz et al., 2020, Buffalo, 2021]. This downward bias is typically assumed to have little effect on the qualitative relationship between diversity and $N_e$ [Buffalo, 2021], but recent pangenomic studies have uncovered troves of non-reference variation across a variety of species (Ebler et al. [2022], Rice et al. [2023], reviewed in Bayer et al. [2020]). Second, many other classes of genetic variants contribute to genetic diversity besides SNPs, and SNPs can actually be a cryptic sign of larger-scale variation. For example, a large fraction of heterozygous SNP calls in *Arabidopsis thaliana* are actually the result of structural variation [Jaegle et al., 2023]. Finally, previous meta-analyses of population size and diversity data rely on scraping diversity estimates from previously published studies (Frankham [2012], Buffalo [2021], except see Corbett-Detig et al. [2015]). However, many studies report inaccurate SNP calls and diversity estimates due to errors in the handling of missing data [Korunes and Samuk, 2021, Schmidt et al., 2021, Sopniewski and Catullo, 2024] and may filter genotype calls differently, making comparisons across species difficult. Overall, errors in diversity calculations and omission of diversity in genomic regions that are either difficult or impossible to align to could partially explain Lewontin's paradox. Re-analyzing whole genome sequencing data with a common pipeline would make diversity estimates across species more comparable and easier to interpret [Buffalo, 2021, Mirchandani et al., 2024].

One useful pangenomics tool for measuring non-reference variation that is readily applicable to common short-read datasets is the $k$-mer. $k$-mers are subsequences of length $k$ derived from a larger sequence and they have a long history of use in computer science [Shannon, 1948], genome assembly [Turner et al., 2018], metagenomics [Benoit et al., 2016], and quantitative genetics [Rahman et al., 2018, Voichek and Weigel, 2020, Kim et al., 2020, Mehrab et al., 2021]. Recent studies have also demonstrated the utility of $k$-mers for measuring heterozygosity and genetic differences between individuals (commonly referred to as "dissimilarity" measures, Ondov et al. [2016], Vurture et al. [2017], Ranallo-Benavidez et al. [2020], VanWallendael and Alvarez [2022]). Typical analysis of $k$-mers involves only counting the presence/absence and/or frequencies of all $k$-mers in a set of reads, without aligning the reads to any reference, then deriving measures of genetic difference from such counts [Benoit et al., 2016]. Avoiding alignment allows one to incorporate sequences that would otherwise be omitted for lack of alignment to a reference genome.

We revisited Lewontin's paradox in plants using $k$-mer-based measures of genetic difference, aiming to test whether the inclusion of non-reference variation could partially resolve Lewontin's paradox. We compared how $k$-mer dissimilarity and typical SNP-based estimates of nucleotide diversity correlated with population size proxies across a large panel of plant species - all processed through the same bioinformatic pipeline. Our expectation was that if $k$-mers are better at capturing genomic variation than SNPs, $k$-mer dissimilarity would scale more rapidly with population size compared to nucleotide diversity.

# Materials and Methods

Our entire analysis is packaged as a snakemake workflow stored here: https://github.com/milesroberts-123/tajimasDacrossSpecies. This workflow includes the code to reproduce all of the steps individually explained below, along with instructions on how to run the code, and yaml files describing the exact configurations of software we used at each step. It also includes an example directed acyclic graph showing the order of steps a typical sample is processed through. The code detailing all initial, exploratory, and confirmatory data analyses as well as figure creation can be found as an R-markdown file in the github repository. The parameters for each software were kept constant across all datasets (except occasionally for the "–ploidy" parameter in GATK HaplotypeCaller) to ensure that variation in bioinformatic processing did not bias our results. All statistical analyses used R v4.2.2 [R Core Team, 2022] and all color palettes used in figure creation come from the scico R package [Pedersen and Crameri, 2023] to ensure color-blind accessibility.

## Population-level sequencing data collection

We started by building a list of species with high quality, publicly available reference genomes as well as population-level sequencing data. The source for the genome assembly and annotation used for each species in this

study is listed in Table S1. We first downloaded all genomes in Phytozome (https://phytozome-next.jgi.doe.gov/) with unrestricted data usage. We then downloaded all genomes for species from Ensembl plants (https://plants.ensembl.org/index.html) that were not already represented in Phytozome. Next, we downloaded genomes for additional species from the NCBI genome database (https://www.ncbi.nlm.nih.gov/genome/) that were not already present in either Phytozome or Ensembl and met all of the following criteria:

- matched filters: eukaryotic, plants, land plants, and exclude partial
- included assemblies of nuclear DNA (i.e. not just plastid genomes)
- included annotations of coding sequences

We also downloaded a genome for *Nicotiana tabaccum* from the Sol genomics network (https://solgenomics.net/). Finally, we omitted any species that had at least one chromosome longer than $2^{29}$ bp (about 512 Mb) from all downstream analyses because tabix indexing, which is often utilized for SNP-calling pipelines, does not support chromosomes exceeding this length. In the end, we were left with genome assemblies and annotations for 112 plant species (see Table S1).

Note that, similar to previous studies [Corbett-Detig et al., 2015, Buffalo, 2021], many of the plant species in this set of 112 are domesticated (see Table S1). This means that many of the species in our dataset have likely undergone recent demographic changes. However, we do not expect this to contribute to differences in the relationships between population size and nucleotide or $k$-mer diversity, because past demography affects both nucleotide and $k$-mer diversity. Furthermore, we include cultivation status in our downstream modeling to help account for systematic differences between cultivated and wild species (see **Statistical analysis**).

For each species with a reference genome, we searched for DNA-seq runs in the National Center for Biotechnology Information's Sequence Read Archive (SRA) with a name in the organism field that matched the species name (e.g. search for Arabidopsis lyrata[Organism] to get *Arabidopsis lyrata* runs). We downloaded the run info for each search and found the study with most sequenced individuals for inclusion in our analysis. Most datasets came from individual studies, with the exception of *Zea mays*, which included several studies described in Bukowski et al. [2017]. The datasets used for each species are listed in Table S1.

We limited the size of each species' dataset to no more than $7.5 \times 10^{12}$ bp and no more than 1200 individuals because this defined the amount of data our workflow could process without the peak memory limit exceeding 50 TB and the time limit for genotype calling exceeding 7 days. If a species' dataset exceeded either 1200 individuals or $7.5 \times 10^{12}$ bp, we randomly downsampled runs such that both of these limits were satisfied.

We downloaded the SRA runs associated with each individual using the SRA toolkit (v2.10.7), then trimmed low-quality base calls with fastp (v0.23.1, Chen et al. [2018]), requiring a minimum quality score of 20 and a minimum read length of 30 base pairs. For each species, we summarized the results of fastp trimming using multiqc (v1.18, Ewels et al. [2016]). After trimming, any fastq files that were technical replicates of the same individual were concatenated. Concatenated fastq files were then processed through two different workflows: SNP-calling and $k$-mer counting.

## Single-nucleotide polymorphism calling

We aligned sequencing reads for each individual to their respective reference genome using BWA MEM (v0.7.17, Li and Durbin [2009], Li [2013]), sorted the resulting BAM files with samtools (v1.11, Danecek et al. [2021]), and marked optical duplicates with picardtools (picard-slim v2.22.1, Institute [2019]). Next, we called SNPs with GATK HaplotypeCaller (v4.1.4.1, McKenna et al. [2010], Poplin et al. [2018]). We varied the –ploidy parameter for HaplotypeCaller between species depending on the actual ploidy recorded in the literature and whether individual subgenome assemblies were available. However, the vast majority of species in our dataset had a –ploidy paramter of 2. We restricted genotype calling to only 4-fold degenerate sites within the nuclear genome, as identified by degenotate (v1.1.3, Mirchandani et al. [2024]), to focus solely on neutral diversity. Runs for each species were then combined with GATK GenomicsDBImport, then genotyped with GATK GenotypeGVCFs, including invariant sites as done in Korunes and Samuk [2021]. Variant and invariant sites were separated with bcftools (v1.17, Danecek et al. [2021]) and then filtered separately, as recommended by Korunes and Samuk [2021]. Variant sites were removed from our analyses if they met at least one of the following criteria: number of alleles $> 2$, indel status = TRUE, fraction of

missing genotypes > 0.2, QD < 2.0, QUAL < 30.0, MQ < 40.00, FS > 60.0, HaplotypeScore > 13.0, MQRankSum < -12.5, and ReadPosRankSum < -8.0 [Caetano-Anolles, 2023]. For each species, we also required that each variant site have a minimum read depth of 5, but no more than 3 times the genome-wide average read depth at variant sites for that species. Meanwhile, invariant sites were removed from our analyses if they met at least one of the following criteria: QUAL > 100.0, read depth ≤ 5, or read depth ≥ 3 times the genome-wide average read depth at invariant sites for that species. Finally, invariant and variant sites were concatenated into a single VCF file per scaffold using bcftools. For *Brassica napus* and *Miscanthus sinensis*, scaffolds named "LK032656" (195,249 bp) and "scaffold04645" (2,838 bp), respectively, were omitted from our analyses because an error in SLURM job cancellation caused snakemake to prematurely delete intermediate files for these scaffolds. It is worth noting that different choices of genotype callers and filtering parameters could lead to different estimates of nucleotide diversity. However, our workflow is representative of SNP calling workflows used in many published population genetic analyses.

Using the SNP genotypes called from our pipeline, we then calculated genome-wide average nucleotide diversity at four-fold degenerate sites ($\bar{\pi}$) using the filtered set of variant and invariant sites. To do this, we first calculated heterozygosity at each four-fold degenerate site ($i$) according to Hahn [2018]:

$$\pi_i = \left(\frac{n_i}{n_i - 1}\right)\left(1 - \sum_{j=1}^{a_i} p_{ij}^2\right) \quad (1)$$

where $n_i$ is the number of sequenced chromosomes with non-missing genotypes for site $i$, $a_i$ is the number of alleles for site $i$, and $p_{ij}$ is the frequency of the $j$th allele at site $i$. For each invariant site, the equation reduces to $\pi_i = 0$ because $p_{i1} = 1$ and $a_i = 1$. To get $\bar{\pi}$, we then calculated the average value of $\pi_i$ across all $M$ sites in the genome (including both variant and invariant sites):

$$\bar{\pi} = \frac{\sum_{i=1}^{M} \pi_i}{M} \quad (2)$$

## $k$-mer counting

We chose to count $k$-mers of 30 base pairs (i.e. 30-mers) for all species in our dataset because previous $k$-mer-based analyses in plants typically analyzed $k$-mers in the range of 20 - 40 base pairs [Voichek and Weigel, 2020, Kim et al., 2020, VanWallendael and Alvarez, 2022, Ruperao et al., 2023] and because $k$-mers in this range can be reliably sequenced with short reads while capturing the majority of unique genomic sequences [Shajii et al., 2016, Ondov et al., 2016]. For each species, we built a database of the 30-mers that were present in the coding sequences of their reference genome using KMC (v3.2.1, Kokot et al. [2017]). Then, we counted 30-mers in each individuals' sequencing reads using KMC, removing any 30-mers that matched the database of 30-mers found in its corresponding set of coding sequences. This step intended to focus our $k$-mers down to a set that is evolving more neutrally on average, analogously to how we focused on only 4-fold degenerate SNPs in our SNP-calling pipeline. The justification for this approach is that non-coding sequences generally have weaker signals of interspecies conservation compared to coding sequences [Woolfe et al., 2005, Siepel et al., 2005, Johnsson et al., 2014]. Although, similarly to 4-fold degenerate sites, many studies have observed non-coding sequences that appear to be under selective constraints [Margulies et al., 2003, Guo et al., 2007]. Thus, similar to the common analysis of 4-fold degenerate sites, our analysis is limited by an inability to completely remove the effects of selection on sequence diversity.

Although comparing our $k$-mer and nucleotide diversity metrics will be affected by differences between coding and non-coding sequences, many previous studies have found that the average diversity of non-coding regions is often very similar to average diversity at 4-fold degenerate sites [Moriyama and Powell, 1996, Makałowski and Boguski, 1998, Halushka et al., 1999, Zwick et al., 2000, Tenaillon et al., 2001, Nordborg et al., 2005, Branca et al., 2011, Williamson et al., 2014, Wang et al., 2016, Phung et al., 2016, Mattila et al., 2017]. Previous investigations of Lewontin's paradox also found that diversity levels across species vary much more than diversity levels across different categories of putatively neutral sequences [Leffler et al., 2012, Buffalo, 2021] and subsequently pooled estimates of neutral diversity across different categories of sites. Similar to these previous studies, we thus assume that differences in linked selection between coding and non-coding sequences are negligible.

For most species in this study, we identified hundreds of millions of unique 30-mers. It would be computationally expensive to analyze all the 30-mers for every species. However, previous studies have shown that one can randomly

downsample $k$-mer sets with very minimal effects on measures of genomic dissimilarity [Fofanov et al., 2004, Benoit et al., 2020]. Thus, we randomly downsampled each species' 30-mer list to 10 million 30-mers with a frequency $\geq 5$ in at least one sample in the species' 30-mer list. The reason to include this frequency cut-off is to omit low frequency $k$-mers that result from sequencing errors [Ranallo-Benavidez et al., 2020]. We also chose to subset our 30-mer matrix to 10 million 30-mers to decrease disk space burden and because several previous studies show that subsets of only 1 million $k$-mers or less reliably estimate genetic dissimilarity in many systems [Ondov et al., 2016, Benoit et al., 2020, VanWallendael and Alvarez, 2022]. We then joined the subset $k$-mer counts for each individual into a single matrix for each species. We used this $k$-mer frequency matrix to measure genetic distance in two ways. First, we calculated Jaccard dissimilarity ($J_D$, Ondov et al. [2016]) between each pair of individuals in a species' dataset as:

$$J_D(X,Y) = 1 - \frac{X \cap Y}{X \cup Y} \tag{3}$$

where $X$ and $Y$ represent sets of unique $k$-mers identified as present in two different read sets. A $k$-mer is defined as present if it's frequency in a sample is $\geq 5$, but cutoffs anywhere from 2 - 10 are commonly used in the literature [Voichek and Weigel, 2020, VanWallendael and Alvarez, 2022]. We used a frequency cutoff of 5 to make our workflow amenable to lower mean coverage datasets. To get the genome-wide average Jaccard dissimilarity ($\bar{J_D}$), we took the average of all the pairwise Jaccard dissimilarities.

Jaccard dissimilarity is likely the most commonly used $k$-mer-based diversity measure [Ondov et al., 2016]. However, whether a $k$-mer reaches the frequency threshold needed to be identified as preset in a sample depends on the sequencing depth for the sample [VanWallendael and Alvarez, 2022]. Thus, we also calculated Bray-Curtis dissimilarity ($B_D$) between each pair of individuals in a species' dataset as:

$$B_D(X,Y) = 1 - \frac{2\sum_i^k \min(m_i^*(X), m_i^*(Y))}{\sum_i^k m_i^*(X) + m_i^*(Y)} \tag{4}$$

where $m_i^*(X)$ gives the normalized frequency of $k$-mer $i$ in genome $X$. The normalized frequencies are calculated by taking each frequency $m_i(X)$ and dividing it by the sum of the raw frequencies as in Dubinkina et al. [2016]:

$$m_i^*(X) = \frac{m_i(X)}{\sum_i m_i(X)} \tag{5}$$

This step accounts for variation in coverage between samples on $k$-mer frequency. To get the genome-wide average Bray-Curtis dissimilarity ($\bar{B_D}$), we again took the average of all the pairwise Bray-Curtis dissimilarities. Note that both Jaccard and Bray-Curtis dissimilarity are scaled in their denominators by either the total number of unique $k$-mers or total number of $k$-mers respectively, analogous to how nucleotide diversity is scaled by the number of sites included in the calculation.

## Population size estimation

Following similar methods to Corbett-Detig et al. [2015] and Buffalo [2021], we defined current census population size ($N$) as the product of species range size ($R$) in square kilometers and population density ($D$) in individuals per square kilometer:

$$N = RD \tag{6}$$

Estimation of both $R$ and $D$ are handled separately below. Importantly, these methods have the same drawback as described in Corbett-Detig et al. [2015] and Buffalo [2021]: contemporary estimates of $R$ and $D$ do not necessarily reflect the historical values of $R$ and $D$. However, since nearly all the species in this study lack long-term historical data on their population size, it is not currently possible to estimate long-term historical $N$ without making strong assumptions.

### Range size estimation from GBIF occurrence data

We first estimated range size based on Global Biodiversity Information Facility (GBIF) occurrence data from the rgbif package [Chamberlain and Boettiger, 2017]. For each species, we identified its GBIF taxon key(s). If the species

is domesticated, we used the taxon key(s) for a wild relative with an overlapping range when possible. We then downloaded all records associated with each taxon key that had an occurrence status of "PRESENT", had coordinates that mapped to land, had any basis of record other than "FOSSIL SPECIMEN", and recorded anywhere in a year $\geq$ 1943 and $\leq$ 2023. In addition, the records could not have any GBIF issue codes, except for the issue codes listed in Supplemental Methods. Similar to previous studies [Corbett-Detig et al., 2015, Buffalo, 2021], we estimated range size for domesticated species using GBIF occurrences from closely-related wild relatives because it is difficult to distinguish the native and introduced ranges of globally cultivated crop species with only occurrence data. Note, however, that we also used an additional method for estimating range size that is not burdened by this same assumption (see **Range size estimation from WCVP distribution maps**). The relatives used for each domesticated species is detailed in Table S1.

We followed methods of Buffalo [2021] to estimate range size from each species' set of GBIF occurrence data using the package alphahull [Pateiro-Lopez and Rodriguez-Casal, 2022]. We started with splitting the occurrence data by continent, in order to avoid estimating ranges that overlapped with oceans. We also only kept occurrences with unique latitude-longitude values to reduce the computational burden of alphahull's algorithms. We then added a small amount of random jitter (normally distributed with $\mu = 0$ and $\sigma = 1 \times 10^{-3}$) to the latitude-longitude coordinates of each unique occurrence to avoid errors in the triangulation algorithm of alphahull, which can break when there are lots of colinear points. Finally, we filtered out any continents which had fewer than 20 unique occurrences of a species. The only exceptions to this rule were *Solanum stenotomum*, *Dioscorea alata*, and *Rhododendron griersonianum*, for which we only required 8, 6, and 3 occurrences respectively due to the rarity of these species and thus a paucity of occurrence data. We then used alphahull to compute the alpha shape of each continent subset, which can be thought of as the smallest possible convex shape that encloses a set of points in a plane. We defined the alpha parameter for the alphahull package to be 200. We then used the R packages sf [Pebesma, 2018] and rworldmap [South, 2011] to measure the sizes of the alpha shapes in square kilometers after projecting them onto the Earth's surface. Finally, we took the estimated range polygons and filtered out ones that resided on continents in the introduced range of the species, as defined by the World Checklist of Vascular Plants (WCVP) [Govaerts et al., 2021]. The sum of the areas of the remaining polygons was our estimate of range size.

### Range size estimation from WCVP distribution maps

We also estimated range size from expert-drawn species distribution maps instead of species occurrence data. We used the rWCVP package [Brown et al., 2023] to download distribution maps from WCVP [Govaerts et al., 2021]. We then estimated range size for each species as either (1) the sum of the areas of all map elements labeled as "native" or "extinct" for that species or (2) the sum of the areas of all map elements labeled as "native", "invaded", or "extinct" for that species. Regions with an occurrence label of "dubious" were excluded from downstream analyses. In contrast to GBIF-derived ranges, we used distribution maps for domesticated species in this estimate of range size because the maps discriminate between the native and introduced ranges of species.

### Population density estimation from plant height

Similarly to previous studies, we use plant height as a proxy for plant population density [Corbett-Detig et al., 2015]. While it would be ideal to use actual population densities in our analyses, we could not find published estimates of population densities for many of the species in our dataset and all previous studies investigating Lewontin's paradox rely on population size proxies [Leffler et al., 2012, Corbett-Detig et al., 2015, Filatov, 2019, Buffalo, 2021]. We elaborate further on the limitations of using proxies in the Discussion, but at the time of writing this manuscript using proxies is the only way to achieve a sufficient sample size for investigating Lewontin's paradox.

We decided to use plant height rather than plant mass [Deng et al., 2012] as our measure of body size because plant height measurements are available for many more species in our dataset and also to make our results more comparable to previous studies that also use plant height [Corbett-Detig et al., 2015]. According to theory outlined in Deng et al. [2012], where $D$ is population density, $M$ is plant mass, and $h$ is plant height, $D \propto M^{-3/4}$ and $M \propto h^{8/3}$. Combining these two relationships gives $D \propto (h^{8/3})^{-3/4}$ which simplifies to $D \propto h^{-2}$. Adding this density-height relation to equation 6 gives our main proxy for population size:

$$N \propto \frac{R}{h^2} \tag{7}$$

In our subsequent analyses, we refer to Equation 7 as the range size-squared height ratio and we convert $R$ to square meters and $h$ to meters to make the ratio unitless. As Equation 7 suggests, we do not expect the range size-squared height ratio to exactly equal the true population size or be interpretable as a number of individuals. Rather, it is a quantity we expect to scale with population size. To calculate the range size-squared height ratio for each species, we downloaded plant height data from the EOL, which mainly comprised records summarized from the TRY database. If no height measurements were available for a species in the EOL, then we used estimates we found in published scientific literature. The only exceptions to this were *Vanilla planifolia* and *Rhododendron griersonianum*, where our height estimates came from the Kew Botanical Gardens' and the American Rhododendron Society's websites, respectively. The sources used for each height value are cited in Table S1.

## Labeling species with genome size, mating system, ploidy, cultivation status, and life cycle habit

Table S1 contains citations for all studies that were used to label each species in our study with a genome size, mating system, ploidy level, cultivation status, and life-cycle habit. For determining genome size, we used estimates from flow cytometry and $k$-mer-spectra analyses whenever possible instead of using assembly size, since most assemblies do not contain the entire genome of the sequenced species. Most of our genome size estimates were 1C values acquired from publications cited in the Plant DNA C-values Database [Pellicer and Leitch, 2020]. Any estimates in terms of picograms (pg) of DNA were converted to base pairs using the following conversion factor: DNA in Mb = DNA in pg $\times 0.978 \times 10^9$ [Doležel et al., 2003]. If genome sizes in terms of pg were not available for a species, then we used the size of the species' genome assembly as the genome size.

We next labeled each species with a mating system (selfing, outcrossing, mixed, or clonal), cultivation status (wild or cultivated), and life cycle habit (annual, biennial, perennial, or mixed) because previous studies showed these factors to be important determinants of diversity in plants [Chen et al., 2017]. For classifying species into different mating systems, we used methods similar to a previous study [Opedal et al., 2023] and generally considered species with outcrossing rate $< 10$ % as "selfing", species with outcrossing rate between 10 - 90 % as "mixed", and species with outcrossing rate $> 90$ % as "outcrossing" when estimates of outcrossing rates were available. In the absence of outcrossing rate data, we also labeled species described as generally self-incompatible as "outcrossing" and we labeled species described as selfing as "selfing". The only exception to this was *Oryza brachyantha* for which we could not find mating system descriptions in peer-reviewed literature. Thus, we assumed that this species was most likely outcrossing because most of the other wild *Oryza* species in the dataset were classified as outcrossing. Because of the low number of mixed (14) and clonal (2) species in our dataset, we collapsed the selfing, mixed, and clonal species into a single "not outcrossing" category for later downstream analysis. Similarly, for life cycle habit, our dataset contained only 1 biennial species and 2 species that had a mixture of annual, biennial, and perennial forms. We combined these species with the perennial category to create a single "not annual" category. For cultivation status, we looked up each species in the EOL and classified species that had documented human uses (such as for food, fiber, fodder) or had some countries known to cultivate the species as "cultivated". All other species that did not meet these criteria were classified as "wild". The only exception to this was *Lactuca sativa*, which did not have any human uses listed in EOL at the time of writing this paper; however, it is commonly known as lettuce so we classified it as "cultivated". Finally, for ploidy levels, when more than one cytotype was described as present within a species we labeled the species with it's most common naturally-occurring cytotype. Citations to relevant literature used for each classification decision can be found in Table S1.

## Statistical analysis

The ultimate goal of our statistical analyses was to estimate the effect of our population size proxies on measures of diversity, comparing the effects of using $k$-mer-based or nucleotide diversity. To do this, we took an approach similar to Whitney et al. [2010] where we performed partial phylogenetic regressions controlling for evolutionary history (using a phylogeny obtained from timetree.org, Kumar et al. [2017, 2022]), mating system (outcrossing vs not outcrossing), cultivation status (wild vs cultivated), and life cycle habit (annual vs not annual). Similar to Whitney et al. [2010], we also scaled the dependent variables to be unitless with a mean of zero and unit variance across species (using the scale() function in R) before performing regression to make slopes more comparable across models

and account for the inherent differences in unit between nucleotide and $k$-mer diversity metrics. This approach can be summarized as follows:

$$\text{scale(diversity)} = \beta_0 + \beta_1 \times \log_{10}(\text{population size proxy}) + \beta_2 \times \text{mating system} + \beta_3 \times \text{cultivation status} + \beta_4 \times \text{life cycle habit} + \epsilon$$

where population size proxy refers to either Equation 7 or it's components (range size and plant height), covariance in the residuals is given by $Var[\epsilon] = \Omega$, diversity was estimated using either SNPs ($\log_{10}(\bar{\pi})$) or $k$-mers ($\bar{J_D}$ or $\bar{B_D}$), and the scale() function performs a z-transformation to make diversity unitless with mean of zero and unit variance. We also constructed a separate set of models where we included genome size as a covariate:

$$\text{scale(diversity)} = \beta_0 + \beta_1 \times \log_{10}(\text{population size proxy}) + \beta_2 \times \text{mating system} + \beta_3 \times \text{cultivation status} + \beta_4 \times \text{life cycle habit} + \beta_5 \times \log_{10}(\text{genome size}) + \epsilon$$

We controlled for genome size in a separate set of models because we had conflicting expectations on whether genome size would be a confounder or a mediator of the population size-diversity relationship. In other words, the effect of population size on diversity could act through genome size, since small populations may not experience strong enough selection to purge deleterious insertions [Lynch and Conery, 2003]. Including genome size as a covariate in this case would artificially diminish the estimated effect of population size on diversity. Alternatively, genome size could fundamentally alter the mode of adaptation in plant species [Mei et al., 2018], making genome size a confounder of the population size-diversity relationship.

After constructing our models, we visualized the relationship between population size and diversity or genome size and diversity with partial regression plots, following methods from Riddell [1977] and Blomberg et al. [2012]. Beginning with our initial phylogenetic least squares model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{8}$$

where $\boldsymbol{y}$ is a vector of diversity values, $X$ is the design matrix, $\boldsymbol{\beta}$ is a vector of regression coefficients, and $\boldsymbol{\epsilon}$ is a vector of residuals distributed normally about 0 with phylogenetic variance-covariance matrix $\Omega$. Using the variance-covariance matrix output from the caper R package [Orme et al., 2018], we first performed Cholesky decomposition to get matrix $\boldsymbol{C}$ such that:

$$\Omega = \boldsymbol{C}\boldsymbol{C}^T \tag{9}$$

We then took the inverse matrix $\boldsymbol{C}^{-1}$ and left-multiplied both sides of our regression equations to get:

$$\boldsymbol{C}^{-1}\boldsymbol{y} = \boldsymbol{C}^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{C}^{-1}\boldsymbol{\epsilon} \tag{10}$$

Which we will rewrite as:

$$\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^* \tag{11}$$

where $\boldsymbol{y}^* = \boldsymbol{C}^{-1}\boldsymbol{y}$, $\boldsymbol{X}^* = \boldsymbol{C}^{-1}\boldsymbol{X}$, and $\boldsymbol{\epsilon}^* = \boldsymbol{C}^{-1}\boldsymbol{\epsilon}$. In vector form, this equation is now:

$$\boldsymbol{y}^* = \beta_0\boldsymbol{x}_0^* + \beta_1\boldsymbol{x}_1^* + \beta_2\boldsymbol{x}_2^* + ... + \beta_{n-1}\boldsymbol{x}_{n-1}^* + \boldsymbol{\epsilon}^* \tag{12}$$

where $\beta_0\boldsymbol{x}_0^*$ is our intercept (Note that $\boldsymbol{x_0}$ was initially a column of 1's before being transformed by $\boldsymbol{C}^{-1}$). After fitting this model to our data with the standard lm() function in R, we collected all terms besides the primary variable of interest, $x_k^*$ (which would be a population size proxy or genome size in our case), and subtracted them from both sides of the equation to get:

$$\boldsymbol{y}^* - \sum_{i \neq k} \beta_i x_i^* = \beta_k \boldsymbol{x}_k^* + \boldsymbol{\epsilon}^* \tag{13}$$

We then plotted the values of $\boldsymbol{x}_k^*$ against $\boldsymbol{y}^* - \sum_{i \neq k} \beta_i x_i^*$, interpreting the slope as the effect of the primary variable on the response, scaled for phylogenetic relationships and adjusted for the effects of confounding factors.

# Results

## Low diversity species explained by low mean coverage

In total, we processed >205 terabases of publicly available sequencing data from the SRA over approximately 12 months of wall time, split between a maximum of 512 cores and 50 TB of disk space. There were 112 species in our initial dataset, each with estimates of population size proxies, nucleotide diversity, and $k$-mer diversity (Fig. 1). Out of these 112 species, 102 were diploids, 9 were tetraploids, and one was hexaploid, with haploid genome sizes ranging from 105 Mb to 5.06 Gb (Table S1). These species were further broken down into 57 annual species vs 55 not annual species (which were predominately perennial), 31 wild vs 81 cultivated species, and 55 outcrossing vs 57 not outcrossing species (which were predominantly selfing). Species classified as annual also tended to not be classified as outcrossing ($\chi^2 = 18.9$, p = $1.4 \times 10^{-5}$, Fig. S1C). However, cultivation status was independent of both life cycle habit ($\chi^2 = 4.07 \times 10^{-31}$, p = 1, Fig. S1A) and mating system ($\chi^2 = 0.53$, p = 0.47, Fig. S1B). There were no missing values for any of the variables investigated in this study, but there were three species with zero variant sites called that we omitted from all downstream analyses.

Before testing our central hypothesis, we investigated whether technical sequencing variables could explain any of the diversity values observed in our dataset. As would be expected for a meta-analysis of previously published data, sequencing parameters varied between species. The number of individuals sampled in each species varied from 3 to 1200 and the average depth of sequencing per individual varied from 0.028x to 79.7x (Fig. S2). Variation in the depth of sequencing between individuals, quantified as the coefficient of variation in base pairs sequenced, varied about 50-fold from 0.030 to 1.6 (Fig. S2). Mean coverage correlated with both nucleotide diversity ($\rho = 0.33$, p = 0.00033, Fig. S3A) and $k$-mer diversity (Jaccard: $\rho$= -0.53, p = $2.6 \times 10^{-9}$, Fig. S3D; Bray-Curtis: $\rho$ = -0.34, p = 0.00021, Fig. S3G). Coefficient of variation in bp sequenced correlated strongly with $k$-mer diversity (Jaccard: $\rho = 0.36$, p = 0.00013, Fig. S3E; Bray-Curtis: $\rho = 0.42$, p = $4.7 \times 10^{-6}$, Fig. S3H) but not nucleotide diversity ($\rho$ = -0.088, p = 0.36, Fig. S3B). The number of individuals sequenced did not correlate with either nucleotide diversity or $k$-mer diversity (Fig. S3C, S3F, S3I).

While screening the data for outliers, we expected that nucleotide diversity and $k$-mer-based diversity would be positively correlated across species and that deviations from this expectation might result from technical variation in how sequencing was performed. Overall, we observed that species with lower coverage did not follow the expected positive relationship between nucleotide and $k$-mer diversity (Fig. 2A, Fig. S4A). In contrast, there was no clear pattern in how the coefficient of variation in base pairs sequenced (Fig. S5) or the number of individuals sequenced (Fig. S6) affected the correlation between $k$-mer dissimilarity and nucleotide diversity. Based on these results, we removed 10 species from our dataset with mean coverage per individual $\leq$ 0.5x as well as 4 species with higher coverage but fewer than 1000 variant sites called. This included three species (*Capsicum annuum*, *Heliosperma pusillum*, and *Papaver somniferum*) with zero variant sites called. The correlation between nucleotide diversity and $k$-mer diversity was much more significant after excluding these species (Jaccard: $\rho = 0.34$, p = 0.00068, Fig. S4B; Bray-Curtis: $\rho = 0.49$, p = $3.6 \times 10^{-7}$, Fig. 2B). In total, we kept data for 98 species for downstream hypothesis testing.

## Range size-squared height ratio varies over more orders of magnitude than nucleotide diversity

We next investigated whether Lewontin's paradox applied to our dataset by comparing diversity estimates against population size proxies. For each species, we estimated range size using either GBIF occurrence data or WCVP range maps. Estimates from these two methods were significantly correlated no matter whether invaded ranges (as defined in the WCVP range maps) were included ($\rho = 0.31$, p = 0.00096, Fig. S7A) or excluded ($\rho = 0.48$, p = $7.3 \times 10^{-8}$, Fig S7B). The omission of invaded ranges lowered the range size of several plant species based on WCVP range maps (Fig. S7C) but had less effect on ranges estimated from GBIF occurrence data (Fig. S7D).

We then calculated the ratio of range size to squared plant height (Equation 7) using height values from the EOL. We used this ratio as our primary population size proxy in downstream analyses. After excluding species with $<$ 0.5x coverage and $<$ 1000 variant sites called (Fig. 2), nucleotide diversity varied over about 4 orders of magnitude for the species in our dataset (from 0.00021 to 0.117, Table S2), while the ratio of range size to squared plant height based on WCVP and GBIF range estimation methods (including both native and invaded ranges) varied over 10 (from

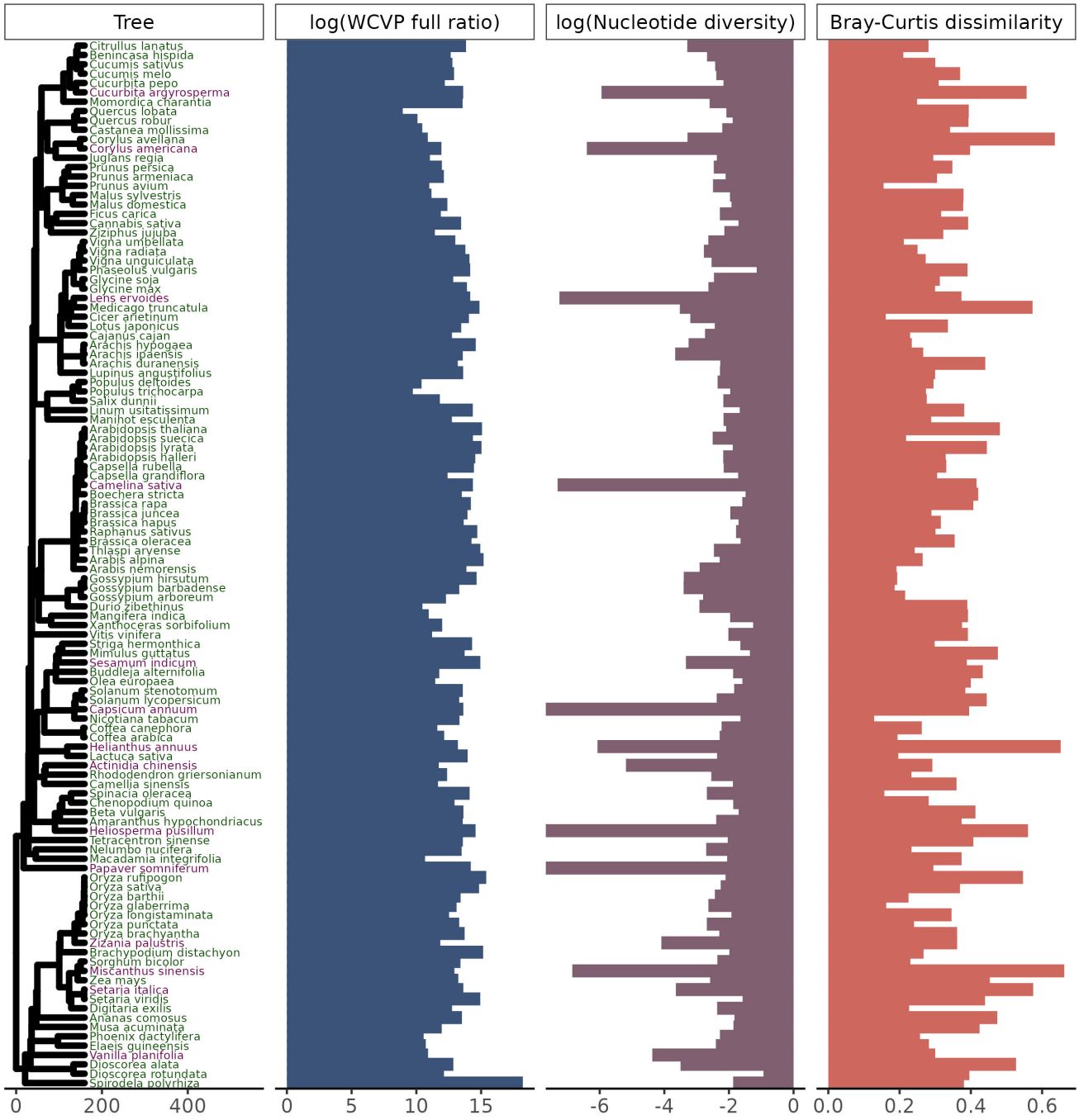


**Figure 1. Our study includes 112 plant species across a wide range of population sizes and diversity levels.** Species labeled in purple were considered outliers and omitted from downstream analyses (see Fig. 2), but species labeled in green were retained. The phylogenetic tree is scaled in millions of years. The WCVP full ratio is a unitless population size proxy equal to the ratio of range area, estimated using WCVP range maps, to squared plant height and is log-transformed (base 10). Nucleotide diversity is genome-wide average diversity at four-fold degenerate sites, log-transformed (base 10). *Capsicum annuum*, *Heliosperma pusillum*, and *Papaver somniferum* had nucleotide diversity values of zero and so have bars at the plotting limit ($\log(0) = -\infty$). Bray-Curtis dissimilarity is average pairwise Bray-Curtis dissimilarity across all pairs of individuals in a species' sample.

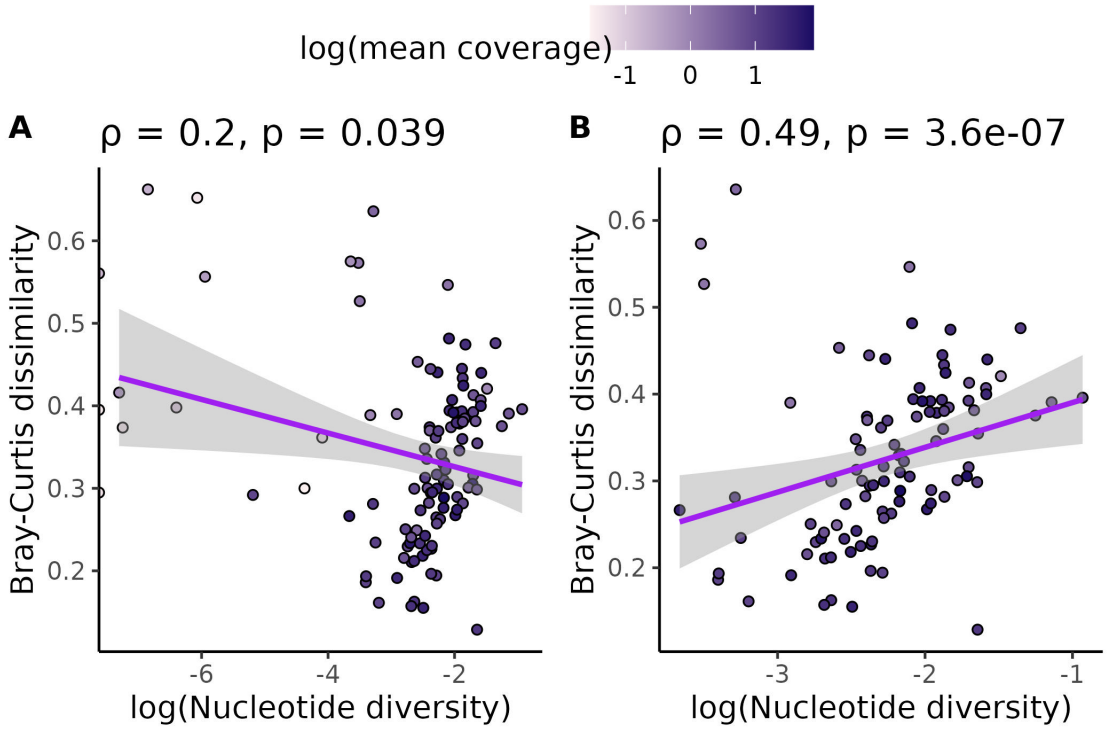


**Figure 2. Omitting species with low coverage and low numbers of variant calls increased the positive correlation between nucleotide and $k$-mer diversity.** (A) shows the relationship between $k$-mer diversity and nucleotide diversity without omitting species with $\leq$ 0.5x coverage or $\leq$ 1000 SNP calls. (B) shows the same relationship, except species with $\leq$ 0.5x coverage or $\leq$ 1000 SNP calls are omitted. Each data point is a species. All species' points are colored by the log (base 10) of average genome-wide coverage per individual for that species. Purple lines are linear regressions with 95% confidence intervals shaded in gray. Values across the top of each plot are Spearman correlation coefficients ($\rho$) and p-values that test whether each correlation coefficient differs from zero.

$8.9 \times 10^8$ to $1.7 \times 10^{18}$) and 13 (from $8.6 \times 10^5$ to $1.5 \times 10^{18}$) orders of magnitude, respectively (Table S2). Mean pairwise Bray-Curtis dissimilarity values varied about 4.9-fold across species, from 0.13 to 0.64, while mean pairwise Jaccard dissimilarity varied about 22-fold, from 0.040 to 0.87 (Table S2). Bray-Curtis dissimilarity values correlated with Jaccard dissimilarity values across species ($\rho = 0.76$, $p < 2.2 \times 10^{-16}$, Fig. S8).

## $k$-mer diversity scales with population size proxies more than nucleotide diversity

The core of Lewontin's paradox is that a population's diversity does not scale much with population size. If $k$-mers capture a wider range of genetic variation compared to SNPs, population size will scale more with $k$-mer diversity than nucleotide diversity. If we did not control for shared evolutionary history or any confounding variables (mating system, life cycle habit, cultivation status, or genome size), then none of our diversity measures significantly correlated with the range size-squared height ratio (Fig. S9). After controlling for confounding variables, nucleotide diversity marginally scaled with the range size-squared height ratio ($\beta = 0.14$, SE = 0.056, p = 0.017, Fig. S10A). However, the relationship between $k$-mer diversity and the range size-squared height ratio was highly significant, with generally a greater slope (Jaccard: $\beta = 0.64$, SE = 0.096 $p = 2.2 \times 10^{-9}$, Fig. S10B; Bray-Curtis dissimilarity: $\beta$ =0.79, SE = 0.11, $p = 7.3 \times 10^{-11}$, Fig. S10C). We observed the same qualitative trend when we included both native and invaded ranges in the range size-squared height ratio (Fig. S10D-F), or used the GBIF-based range estimates instead of WCVP-based estimates (Fig. S11). Interestingly, we often observed Bray-Curtis dissimilarity having a larger slope with the range size-squared height ratio compared to Jaccard dissimilarity ($\beta = 0.64$ vs 0.79 Fig. S10B-C), but models where Bray-Curtis dissimilarity was the response variable generally had lower adjusted $R^2$ (Table S4).

We also analyzed range size and plant height separately as population size proxies (Fig. S12-S14). Overall, WCVP-estimated range size significantly affected nucleotide diversity ($\beta = 0.29$, SE = 0.072, p = 0.00011, Fig. S12A) and $k$-mer diversity (Jaccard: $\beta = 0.92$, SE = 0.13, $p = 9.9 \times 10^{-11}$, Fig. S12B; Bray-Curtis: $\beta = 1.2$, SE = 0.13,$p = 3.2 \times 10^{-14}$, Fig. S12C), and this trend held when we estimated range size from GBIF occurrences (Fig. S13A-C) or included invaded range area (Fig. S12D-F and Fig. S13D-F). On the other hand, plant height did not scale with nucleotide diversity ($\beta = 0.13$, SE = 0.19, p = 0.5, Fig. S14A), but marginally scaled downward with increasing $k$-mer diversity (Jaccard: $\beta = -0.78$, SE =0.38, p = 0.046, Fig. 14B; Bray-Curtis: $\beta = -0.77$, SE = 0.44, p = 0.088, Fig. S14C).

Finally, we repeated our partial phylogenetic regressions controlling for genome size as an additional covariate. In this case, nucleotide diversity did not scale with the range size-squared height ratio ($\beta = 0.035$, SE = 0.063, p = 0.58, Fig. 3A), but $k$-mer diversity did (Jaccard: $\beta = 0.54$, SE = 0.093, $p = 8.8 \times 10^{-8}$, Fig. S15; Bray-Curtis: $\beta = 0.7$, SE = 0.098, $p = 2.2 \times 10^{-10}$, Fig. 3B). Again, we got qualitatively similar results when we excluded invaded ranges in our range size estimates (Fig. S16), used GBIF occurrences to estimate range size-squared height ratio (Fig. S17) or used WCVP range size as the population size proxy (Fig. S18). However, GBIF range size by itself did not scale with Jaccard dissimilarity (Fig. S19B, S19E). Increased plant height associated with decreased $k$-mer diversity, but had no significant relationship with nucleotide diversity (Fig. S20).

## $k$-mer diversity scales with genome size more than nucleotide diversity

We also investigated the relationship between diversity and genome size because we expected genome size to potentially play a role in the mechanism underlying the greater scaling of $k$-mer diversity with population size. Genome size is often a strong predictor of diversity [Lynch and Conery, 2003]. Among eukaryotes, variation in genome size is largely explained by variation in transposable element abundance [Flavell et al., 1974, Kidwell, 2002, Lynch and Conery, 2003, Muñoz-Diez et al., 2012, Tenaillon et al., 2011, Nystedt et al., 2013, Ibarra-Laclette et al., 2013], which contribute substantially to the repetitive sequence content of genomes and increase the difficulty of aligning short reads to a reference genome (reviewed in Goerner-Potvin and Bourque [2018]). Thus, our expectation was that $k$-mer-based diversity measures are more sensitive to genome size variation compared to nucleotide diversity.

Increasing genome size was associated with decreasing $k$-mer diversity (Jaccard: $\beta = -3.7$, SE = 0.42, $p = 8.4 \times 10^{-14}$, Fig. S21; Bray-Curtis: $\beta = -4.2$, SE = 0.45, $p = 4.5 \times 10^{-15}$, Fig. 4B) and nucleotide diversity ($\beta - 1.8$, SE = 0.29, $p = 1.4 \times 10^{-8}$, Fig. 4A), after controlling for variation in the range size-squared height ratio, mating system, life cycle habit, cultivation status, and evolutionary history. We got qualitatively similar results when the population size proxy we corrected for excluded invaded ranges (Fig. S22), or if our population size proxy was based on GBIF occurrences (Fig. S23), or we used range size or plant height individually to control for population size

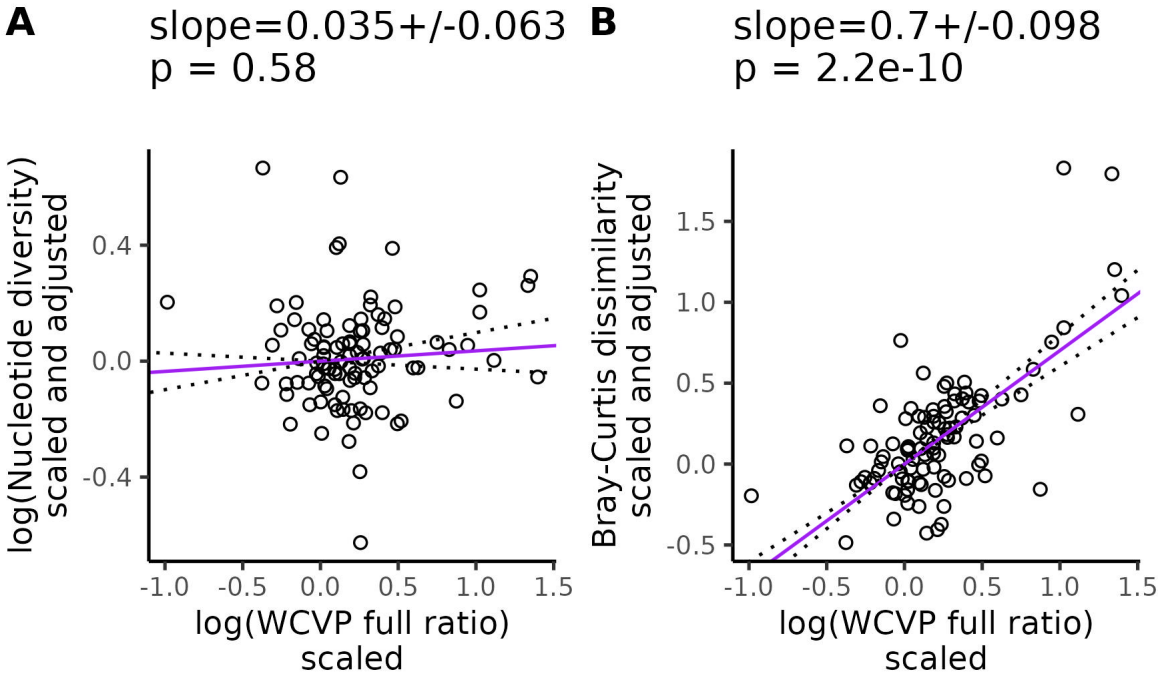


**Figure 3. $k$-mer diversity scales with population size proxies after controlling for genome size, life cycle habit, mating system, and cultivation status.** WCVP full ratio is a population size proxy estimated as the ratio of range size recorded in WCVP range maps (including invaded ranges) to squared plant height. Purple lines are partial phylogenetic regression lines between diversity levels and the population size proxy (see Equation 13) after scaling diversity levels to a standard normal distribution (mean = 0, variance = 1), followed by scaling diversity levels and population sizes according to their phylogenetic relatedness, and finally adjusting for the confounding variables (genome size, life cycle habit, mating system, and cultivation status). The values at the top of each plot give the slope of the partial regression ± one standard error and p-values testing whether the slopes differ from zero. Dotted lines show the partial regression slope ± one standard error.

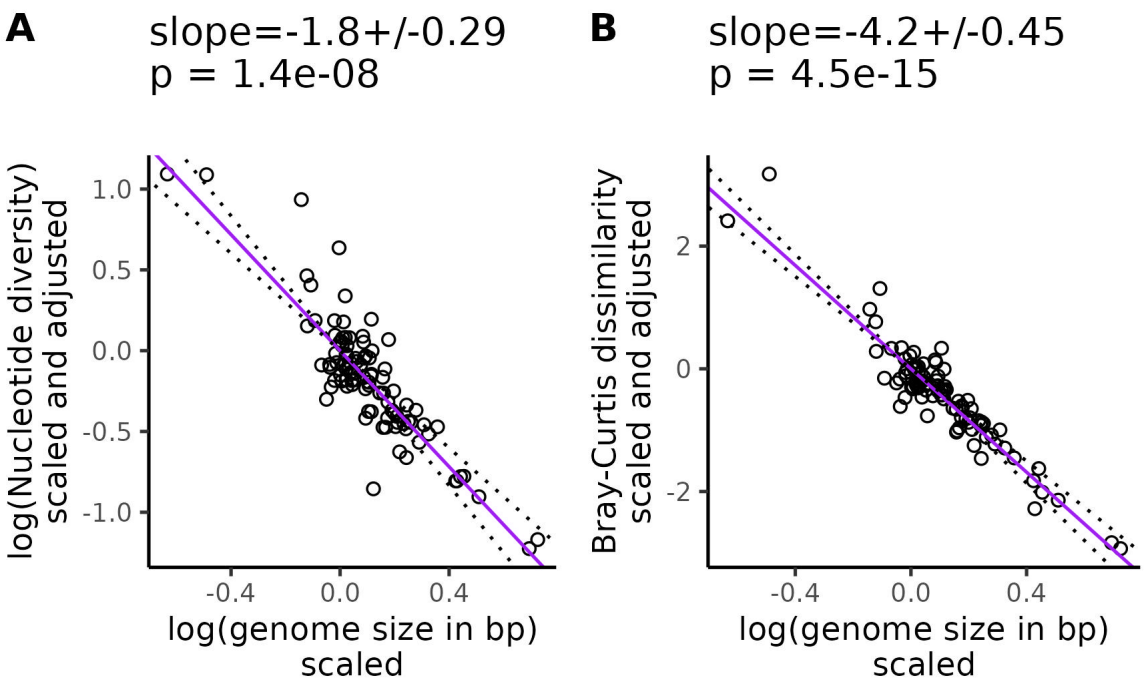


**Figure 4.** *k*-**mer diversity is more sensitive to genome size variation than nucleotide diversity.** Purple lines are partial phylogenetic regression lines between diversity levels and genome size (see Equation 13) after scaling diversity levels to a standard normal distribution (mean = 0, variance = 1), followed by scaling diversity levels and population sizes according to their phylogenetic relatedness, and finally adjusting for the confounding effects of mating system, cultivation status, life cycle habit, and population size. Here we used the ratio of range size to squared plant height, where range size was estimated from ranges in WCVP range maps (including invaded ranges). The values at the top of each plot give the slope of the partial regression ± one standard error and p-values testing whether the slopes differ from zero. Dotted lines show the partial regression slope ± one standard error.

variation (Fig. S24-S26). Across all of these analyses, the partial regression relationship between genome size and diversity was always significantly negative.

## Discussion

Our primary goal was to investigate whether genomic approaches that can capture more genetic variation than standard SNP-based methods can explain the longstanding observation that species with large population sizes have less genetic variation than expected. After careful accounting for potential technical and phylogenetic confounding, the slope between *k*-mer-based diversity and the range size-squared height ratio was up to 20 times larger than the same slope for nucleotide diversity ($\beta = 0.035$ vs 0.7, Fig. 3). We observed similar results across the two different measures of range size (Fig. S17) and *k*-mer diversity (Fig. S15). We also observed that *k*-mer-based diversity is more sensitive to variation in genome size compared to nucleotide diversity (Fig. 4). Overall, these results suggests that diversity missed by SNPs explains part of Lewontin's paradox in plants, consistent with literature suggesting that SNPs provide an incomplete picture of genome-wide polymorphism [Schmidt et al., 2021, VanWallendael and Alvarez, 2022, Jaegle et al., 2023, Sopniewski and Catullo, 2024].

One limitation of our investigation was that we were not able to compare our *k*-mer diversity scales to a neutral expectation of how *k*-mer diversity scales with $N_e$. Doing so would have allowed us to estimate what proportion of Lewontin's paradox is explained by using *k*-mer diversity instead of nucleotide diversity measures. Instead we can only compare the slopes of how *k*-mer and nucleotide diversity scale with population size proxies. We deliberately avoided comparing our data to a neutral expectation for two main reasons. First, we can only estimate proxies of

census size that are not interpretable as numbers of individuals, which is what a theoretical expectation would most likely be based on. Furthermore, robustly estimating the diversity-census size relationship across species requires controlling for evolutionary history and other confounding variables. This will transform the axes of a diversity-population size partial regression plot into a scale that's not interpretable in the units of the original measures (see Equation 10). Thus, we must restrict our conclusions to whether $k$-mer diversity scales with census size proxies faster than nucleotide diversity. This observation is consistent with the hypothesis that the exclusion of non-reference variation explains a part of Lewontin's paradox. However, exactly what proportion of the paradox is explained by our results remains unknown.

As with all regression-based analyses, our results are also ultimately sensitive to error in the measurement of both covariates (population size proxies, genome size, mating system, life cycle habit, or cultivation status) and outcome variables (nucleotide or $k$-mer diversity). Random covariate measurement errors (i.e. error that is not systematically higher/lower for different values of the covariate) bias regression coefficients toward zero [Hutcheon et al., 2010, Nab et al., 2021]. Similarly, random measurement error in the outcome variables increases the standard errors of the covariates, weakening the statistical significance of detected relationships [Hutcheon et al., 2010]. However, our results remain statistically significant despite the potential for error. Our study is also unique in the multiple steps we took to limit the influence of systematic measurement errors on our coefficients. First, we reanalyzed all population-level sequencing data with a single pipeline to limit between-study variation and the impact of bioinformatic parameter choices on our analysis [Mirchandani et al., 2024]. Second, to minimize error in both nucleotide and $k$-mer diversity measures, we omitted species with coverage below 0.5x from our study, because having low coverage strongly correlated with having low diversity (Fig. 2). This threshold is consistent with previous $k$-mer-based phylogenetic studies that found dropping coverage to 0.5x changes tree topologies compared to coverage levels $\geq$ 1x [Sandell et al., 2022]. Third, we accounted for the presence of missing data in calculations of nucleotide diversity [Schmidt et al., 2021, Korunes and Samuk, 2021]. And finally, we estimated range size with two different methods (WCVP range maps and GBIF occurrence records, Fig. S6). Although we could not control for some covariates [Willis, 1922, Romiguier et al., 2014, Guo et al., 2024] due to a dearth of data, our study is still the largest reanalysis of population-level sequencing data in plants that we know of to date. The availability of our workflow also makes it easy for our study to be extended as more population-level sequencing data is released.

Another limitation of most investigations into Lewontin's paradox is the assumption that contemporary population size estimates are good proxies for historic population sizes [Corbett-Detig et al., 2015, Buffalo and Coop, 2020]. While the long-term harmonic mean of the effective population size determines diversity levels within a population [Wright, 1940], population size proxies such as range size and plant height only reflect the current census population size of a species. The separation of plant range maps into native and invaded ranges [Brown et al., 2023] offered an opportunity to test the robustness of our results to invasion-related range size changes. Overall, our observations were remarkably similar no matter whether we included or excluded invaded ranges in our population size proxies (Fig. S17A-C vs Fig. S17D-F). Part of this apparent robustness was due to the insensitivity of our GBIF-based range size estimates to the inclusion of invaded ranges (Fig. S7D). However, our WCVP-based range size estimates were drastically altered by the inclusion of invaded ranges (Fig. S7C) and still yielded similar results (Fig. 3, S15, S16). Although we cannot rule out the possibility that older historical events have affected contemporary diversity levels, our results appear to be robust to some recent human-caused population size changes.

Interestingly, the estimated effect of our population size proxies on diversity was often slightly larger for Bray-Curtis dissimilarity than Jaccard dissimilarity (for example, $\beta = 0.7$ vs 0.54 from Fig. 3B vs Fig. S15, Table S4). In contrast, the range size-squared height ratio was often slightly more predictive of Jaccard dissimilarity than Bray-Curtis dissimilarity (Table S4). We could not test whether these trends were statistically significant, but the benefits of different $k$-mer metrics in predicting measures of population size warrant further study. Our expectation is that $k$-mer diversity measures based on frequency, such as Bray-Curtis dissimilarity, better capture diversity compared to measures based on purely $k$-mer presence/absence, such as Jaccard dissimilarity, because they explicitly measure copy number variation. However, accurately measuring $k$-mer frequencies likely requires higher sequencing coverage than calling presence/absence, which could explain why Bray-Curtis dissimilarity generally scaled more with population size but had a lower $R^2$ compared to Jaccard dissimilarity (Table S4). Future studies using higher coverage population level sequencing data could help test this hypothesis.

$k$-mer frequencies are known to be highly informative of genomic structure, with one common application of $k$-mers being the estimation of genome size [Vurture et al., 2017, Pflug et al., 2020]. Similar to previous studies, we

observed that nucleotide diversity was negatively correlated with genome size [Lynch and Conery, 2003, Chen et al., 2017], but we observed an even stronger negative correlation for $k$-mer diversity ($\beta = -1.8$, SE = 0.29 vs $\beta$ = -3.7, SE = 0.42 in Fig. 4). $k$-mers also appeared to explain diversity patterns that scaled with population size beyond those explained by genome size, while nucleotide diversity did not. After controlling for genome size, the relationship between our population size proxies and nucleotide diversity was not significant (Fig. 3A, S17-S19 panels A and D), but the relationship between $k$-mer diversity and population size proxies was often still highly significant (Fig. 3B, S17-S19 panels B; C; E; F). The only exception was that Jaccard dissimilarity did not significantly scale with GBIF-based estimates of range size (Fig. S19B, S19E). This additional scaling of $k$-mer diversity with population size beyond just the effects of genome size and confounding variables suggests that $k$-mers capture some element of the population size-diversity relationship that is absent from nucleotide diversity.

Our results do not negate the fact that other important factors also underlie Lewontin's paradox, such as past demographic fluctuations and linked selection. However, our results do suggest that future studies of Lewontin's paradox would benefit from considering diversity outside one reference genome. The increasing availability of pangenomes across species [Göktay et al., 2021, Zhou et al., 2022, Rice et al., 2023, Wang et al., 2023] offers many opportunities to revisit this classic population genetics question. While our results suggest that including non-reference variation may partially satisfy Lewontin's paradox, exactly how much of the paradox is explained by non-reference variation, whether our findings apply outside of plants, and the relative importance of non-reference variation to other factors in explaining Lewontin's paradox is still unknown. Ideal future studies would use pangenomic genotyping methods across a wide range of species with a standardized pipeline, combined with multiple proxies of population size. Altogether, these methodological developments will hopefully reveal a more wholistic picture of variation across the tree of life.

## Data availability

Our entire analysis is packaged as a snakemake workflow stored here: https://github.com/milesroberts-123/tajimasDacrossSpecies. Table S1 contains the metadata for all of the datasets used in this study, including sources for genome assemblies, genome annotations, population-level sequencing datasets, and GBIF observations. Table S2 contains all of the covariate and response variable values used for fitting our phylogenetic least squares models. Table S3 contains the estimated coefficients of all of our phylogenetic least squares models and their related statistics, including p-values and standard errors. Table S4 contains the model-level statistics for each phylogenetic least squares model, including $R^2$ values and F-test results. If necessary, we are also prepared to publish the following datasets upon acceptance of this manuscript in the accepting journal's preferred repository: matrices of $k$-mer counts (93 G), VCF files of filtered variants (202 G), multiqc reports of fastp read trimming (244 M), species range maps (87M, downloaded from Plants of the World Online), and plant height values (downloaded from Encyclopedia of Life), and our species tree (downloaded from timetree.org).

## Conflicts of interest

The authors declare no conflicts of interest.

## Acknowledgments

We would like to thank Jeff Conner, Husain Agha, Sophie Buysse, Nathan Catlin, Adrian Platts, Gabrielle Sandstedt and the rest of the Josephs lab for informal comments on early drafts of this manuscript. We would also like to thank the Institute for Cyber-Enabled Research at Michigan State University for providing the computing power used for this research. Finally, we would like to thank the hundreds of scientists throughout the world that sequenced plant populations and publicly released their raw data. Without them this work would not have been possible.

## Author contributions

Both M.D.R and E.B.J contributed to the initial conceptualization and planning of this research. M.D.R wrote all of the code and conducted all statistical analyses. Both M.D.R and E.B.J contributed to drafting, editing, and reviewing the manuscript.

## Funding

This work was funded by a National Institutes of Health grant (R35 GM142829) to E.B.J., an Integrated Training Model in Plant And Computational Sciences Fellowship (National Science Foundation: DGE-1828149) to M.D.R., a Plant Biotechnology for Health and Sustainability Fellowship (National Institute Of General Medical Sciences of the National Institutes of Health : T32-GM110523) to M.D.R., and a Michigan State University Institute for Cyber-Enabled Research Cloud Computing Fellowship to M.D.R. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## References

P. E. Bayer, A. A. Golicz, A. Scheben, J. Batley, and D. Edwards. Plant pan-genomes are the new reference. *Nature Plants*, 6(8):914–920, Aug. 2020. ISSN 2055-0278. doi: 10.1038/s41477-020-0733-0. URL https://www.nature.com/articles/s41477-020-0733-0. Publisher: Nature Publishing Group.

E. Bazin, S. Glémin, and N. Galtier. Population Size Does Not Influence Mitochondrial Genetic Diversity in Animals. *Science*, 312(5773):570–572, Apr. 2006. doi: 10.1126/science.1122033. URL https://www.science.org/doi/full/10.1126/science.1122033. Publisher: American Association for the Advancement of Science.

G. Benoit, P. Peterlongo, M. Mariadassou, E. Drezen, S. Schbath, D. Lavenier, and C. Lemaitre. Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Computer Science*, 2:e94, Nov. 2016. ISSN 2376-5992. doi: 10.7717/peerj-cs.94. URL https://peerj.com/articles/cs-94. Publisher: PeerJ Inc.

G. Benoit, M. Mariadassou, S. Robin, S. Schbath, P. Peterlongo, and C. Lemaitre. SimkaMin: fast and resource frugal de novo comparative metagenomics. *Bioinformatics*, 36(4):1275–1276, Feb. 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz685. URL https://doi.org/10.1093/bioinformatics/btz685.

L. A. Bergeron, S. Besenbacher, J. Zheng, P. Li, M. F. Bertelsen, B. Quintard, J. I. Hoffman, Z. Li, J. St. Leger, C. Shao, J. Stiller, M. T. P. Gilbert, M. H. Schierup, and G. Zhang. Evolution of the germline mutation rate across vertebrates. *Nature*, 615(7951):285–291, Mar. 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-05752-y. URL https://www.nature.com/articles/s41586-023-05752-y. Publisher: Nature Publishing Group.

S. P. Blomberg, J. G. Lefevre, J. A. Wells, and M. Waterhouse. Independent contrasts and PGLS regression estimators are equivalent. *Systematic Biology*, 61(3):382–391, May 2012. ISSN 1063-5157. doi: 10.1093/sysbio/syr118. URL https://doi.org/10.1093/sysbio/syr118.

A. Branca, T. D. Paape, P. Zhou, R. Briskine, A. D. Farmer, J. Mudge, A. K. Bharti, J. E. Woodward, G. D. May, L. Gentzbittel, C. Ben, R. Denny, M. J. Sadowsky, J. Ronfort, T. Bataillon, N. D. Young, and P. Tiffin. Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. *Proceedings of the National Academy of Sciences*, 108(42):E864–E870, Oct. 2011. doi: 10.1073/pnas.1104032108. URL https://www.pnas.org/doi/abs/10.1073/pnas.1104032108. Publisher: Proceedings of the National Academy of Sciences.

M. J. M. Brown, B. E. Walker, N. Black, R. H. A. Govaerts, I. Ondo, R. Turner, and E. Nic Lughadha. rWCVP: a companion R package for the World Checklist of Vascular Plants. *New Phytologist*, 240(4):1355–1365, 2023. ISSN 1469-8137. doi: 10.1111/nph.18919. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/nph.18919. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/nph.18919.

V. Buffalo. Quantifying the relationship between genetic diversity and population size suggests natural selection cannot explain Lewontin's Paradox. *eLife*, 10:e67509, Aug. 2021. ISSN 2050-084X. doi: 10.7554/eLife.67509. URL https://doi.org/10.7554/eLife.67509. Publisher: eLife Sciences Publications, Ltd.

V. Buffalo and G. Coop. Estimating the genome-wide contribution of selection to temporal allele frequency change. *Proceedings of the National Academy of Sciences*, 117(34):20672–20680, Aug. 2020. doi: 10.1073/pnas.1919039117. URL https://www.pnas.org/doi/10.1073/pnas.1919039117. Publisher: Proceedings of the National Academy of Sciences.

R. Bukowski, X. Guo, Y. Lu, C. Zou, B. He, Z. Rong, B. Wang, D. Xu, B. Yang, C. Xie, L. Fan, S. Gao, X. Xu, G. Zhang, Y. Li, Y. Jiao, J. F. Doebley, J. Ross-Ibarra, A. Lorant, V. Buffalo, M. C. Romay, E. S. Buckler, D. Ware, J. Lai, Q. Sun, and Y. Xu. Construction of the third-generation *Zea mays* haplotype map. *GigaScience*, 7 (4), Dec. 2017. ISSN 2047-217X. doi: 10.1093/gigascience/gix134. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5890452/.

D. Caetano-Anolles. Hard-filtering germline short variants, Oct. 2023. URL https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants.

A. Cagan, A. Baez-Ortega, N. Brzozowska, F. Abascal, T. H. H. Coorens, M. A. Sanders, A. R. J. Lawson, L. M. R. Harvey, S. Bhosle, D. Jones, R. E. Alcantara, T. M. Butler, Y. Hooks, K. Roberts, E. Anderson, S. Lunn, E. Flach, S. Spiro, I. Januszczak, E. Wrigglesworth, H. Jenkins, T. Dallas, N. Masters, M. W. Perkins, R. Deaville, M. Druce, R. Bogeska, M. D. Milsom, B. Neumann, F. Gorman, F. Constantino-Casas, L. Peachey, D. Bochynska, E. S. J. Smith, M. Gerstung, P. J. Campbell, E. P. Murchison, M. R. Stratton, and I. Martincorena. Somatic mutation rates scale with lifespan across mammals. *Nature*, 604(7906):517–524, Apr. 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04618-z. URL https://www.nature.com/articles/s41586-022-04618-z. Publisher: Nature Publishing Group.

S. A. Chamberlain and C. Boettiger. R Python, and Ruby clients for GBIF species occurrence data. Technical Report e3304v1, PeerJ Inc., Sept. 2017. URL https://peerj.com/preprints/3304. ISSN: 2167-9843.

B. Charlesworth. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genetical Research*, 63(3):213–227, June 1994. doi: 10.1017/s0016672300032365.

B. Charlesworth and J. D. Jensen. How can we resolve Lewontin's Paradox? *Genome Biology and Evolution*, 14(7): evac096, July 2022. ISSN 1759-6653. doi: 10.1093/gbe/evac096. URL https://doi.org/10.1093/gbe/evac096.

B. Charlesworth, M. T. Morgan, and D. Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303, Aug. 1993. ISSN 1943-2631. doi: 10.1093/genetics/134.4.1289. URL https://doi.org/10.1093/genetics/134.4.1289.

J. Chen, S. Glémin, and M. Lascoux. Genetic diversity and the efficacy of purifying selection across plant and animal species. *Molecular Biology and Evolution*, 34(6):1417–1428, June 2017. ISSN 0737-4038. doi: 10.1093/molbev/msx088. URL https://doi.org/10.1093/molbev/msx088.

S. Chen, Y. Zhou, Y. Chen, and J. Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17): i884–i890, Sept. 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty560. URL https://doi.org/10.1093/bioinformatics/bty560.

C. T. Cole. Genetic variation in rare and common plants. *Annual Review of Ecology, Evolution, and Systematics*, 34 (1):213–237, 2003. doi: 10.1146/annurev.ecolsys.34.030102.151717. URL https://doi.org/10.1146/annurev.ecolsys.34.030102.151717. _eprint: https://doi.org/10.1146/annurev.ecolsys.34.030102.151717.

G. Coop. Does linked selection explain the narrow range of genetic diversity across species?, Mar. 2016. URL https://www.biorxiv.org/content/10.1101/042598v1. Pages: 042598 Section: Contradictory Results.

R. B. Corbett-Detig, D. L. Hartl, and T. B. Sackton. Natural selection constrains neutral diversity across a wide range of species. *PLOS Biology*, 13(4):e1002112, Apr. 2015. ISSN 1545-7885. doi: 10.1371/journal.pbio.1002112. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002112. Publisher: Public Library of Science.

P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):giab008, Feb. 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab008. URL https://doi.org/10.1093/gigascience/giab008.

J. Deng, W. Zuo, Z. Wang, Z. Fan, M. Ji, G. Wang, J. Ran, C. Zhao, J. Liu, K. J. Niklas, S. T. Hammond, and J. H. Brown. Insights into plant size-density relationships from models and agricultural crops. *Proceedings of the National Academy of Sciences*, 109(22):8600–8605, May 2012. doi: 10.1073/pnas.1205663109. URL https://www.pnas.org/doi/10.1073/pnas.1205663109. Publisher: Proceedings of the National Academy of Sciences.

J. Doležel, J. Bartoš, H. Voglmayr, and J. Greilhuber. Nuclear DNA content and genome size of trout and human. *Cytometry Part A*, 51A(2):127–128, 2003. ISSN 1552-4930. doi: 10.1002/cyto.a.10013. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.a.10013. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cyto.a.10013.

V. B. Dubinkina, D. S. Ischenko, V. I. Ulyantsev, A. V. Tyakht, and D. G. Alexeev. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC Bioinformatics*, 17(1):38, Jan. 2016. ISSN 1471-2105. doi: 10.1186/s12859-015-0875-7. URL https://doi.org/10.1186/s12859-015-0875-7.

J. Ebler, P. Ebert, W. E. Clarke, T. Rausch, P. A. Audano, T. Houwaart, Y. Mao, J. O. Korbel, E. E. Eichler, M. C. Zody, A. T. Dilthey, and T. Marschall. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nature Genetics*, 54(4):518–525, Apr. 2022. ISSN 1546-1718. doi: 10.1038/s41588-022-01043-w. URL https://www.nature.com/articles/s41588-022-01043-w. Number: 4 Publisher: Nature Publishing Group.

H. Ellegren and N. Galtier. Determinants of genetic diversity. *Nature Reviews Genetics*, 17(7):422–433, July 2016. ISSN 1471-0064. doi: 10.1038/nrg.2016.58. URL https://www.nature.com/articles/nrg.2016.58. Number: 7 Publisher: Nature Publishing Group.

P. Ewels, M. Magnusson, S. Lundin, and M. Käller. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, Oct. 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw354. URL https://doi.org/10.1093/bioinformatics/btw354.

M. Exposito-Alonso, T. R. Booker, L. Czech, L. Gillespie, S. Hateley, C. C. Kyriazis, P. L. M. Lang, L. Leventhal, D. Nogues-Bravo, V. Pagowski, M. Ruffley, J. P. Spence, S. E. Toro Arana, C. L. Weiß, and E. Zess. Genetic diversity loss in the Anthropocene. *Science*, 377(6613):1431–1435, Sept. 2022. doi: 10.1126/science.abn5642. URL https://www-science-org.proxy2.cl.msu.edu/doi/10.1126/science.abn5642. Publisher: American Association for the Advancement of Science.

D. A. Filatov. Extreme Lewontin's Paradox in Ubiquitous Marine Phytoplankton Species. *Molecular Biology and Evolution*, 36(1):4–14, Jan. 2019. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msy195. URL https://academic.oup.com/mbe/article/36/1/4/5142658.

R. B. Flavell, M. D. Bennett, J. B. Smith, and D. B. Smith. Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochemical Genetics*, 12(4):257–269, Oct. 1974. ISSN 1573-4927. doi: 10.1007/BF00485947. URL https://doi.org/10.1007/BF00485947.

J. M. Flowers, J. Molina, S. Rubinstein, P. Huang, B. A. Schaal, and M. D. Purugganan. Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice. *Molecular Biology and Evolution*, 29(2):675–687, Feb. 2012. ISSN 0737-4038. doi: 10.1093/molbev/msr225. URL https://doi.org/10.1093/molbev/msr225.

Y. Fofanov, Y. Luo, C. Katili, J. Wang, Y. Belosludtsev, T. Powdrill, C. Belapurkar, V. Fofanov, T.-B. Li, S. Chumakov, and B. M. Pettitt. How independent are the appearances of n-mers in different genomes? *Bioinformatics*, 20(15):2421–2428, Oct. 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth266. URL https://doi.org/10.1093/bioinformatics/bth266.

R. Frankham. How closely does genetic diversity in finite populations conform to predictions of neutral theory? Large deficits in regions of low recombination. *Heredity*, 108(3):167–178, Mar. 2012. ISSN 1365-2540. doi: 10.1038/hdy.2011.66. URL https://www.nature.com/articles/hdy201166. Number: 3 Publisher: Nature Publishing Group.

P. Goerner-Potvin and G. Bourque. Computational tools to unmask transposable elements. *Nature Reviews Genetics*, 19(11):688–704, Nov. 2018. ISSN 1471-0064. doi: 10.1038/s41576-018-0050-x. URL https://www.nature.com/articles/s41576-018-0050-x. Publisher: Nature Publishing Group.

A. A. Golicz, P. E. Bayer, P. L. Bhalla, J. Batley, and D. Edwards. Pangenomics comes of age: from bacteria to plant and animal applications. *Trends in Genetics*, 36(2):132–145, Feb. 2020. ISSN 0168-9525. doi: 10.1016/j.tig.2019.11.006. URL https://www.sciencedirect.com/science/article/pii/S016895251930246X.

R. Govaerts, E. Nic Lughadha, N. Black, R. Turner, and A. Paton. The World Checklist of Vascular Plants, a continuously updated resource for exploring global plant diversity. *Scientific Data*, 8(1):215, Aug. 2021. ISSN 2052-4463. doi: 10.1038/s41597-021-00997-6. URL https://www.nature.com/articles/s41597-021-00997-6. Number: 1 Publisher: Nature Publishing Group.

Q. Guo, H. Qian, J. Zhang, and P. Liu. The relationships between species age and range size. *Journal of Biogeography*, 00(n/a):1–9, Feb. 2024. ISSN 1365-2699. doi: 10.1111/jbi.14809. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jbi.14809. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jbi.14809.

X. Guo, Y. Wang, P. D. Keightley, and L. Fan. Patterns of selective constraints in noncoding DNA of rice. *BMC Evolutionary Biology*, 7(1):208, Nov. 2007. ISSN 1471-2148. doi: 10.1186/1471-2148-7-208. URL https://doi.org/10.1186/1471-2148-7-208.

M. Göktay, A. Fulgione, and A. M. Hancock. A new catalog of structural variants in 1,301 *A. thaliana* lines from Africa, Eurasia, and North America reveals a signature of balancing selection at defense response genes. *Molecular Biology and Evolution*, 38(4):1498–1511, Apr. 2021. ISSN 1537-1719. doi: 10.1093/molbev/msaa309. URL https://doi.org/10.1093/molbev/msaa309.

M. W. Hahn. *Molecular Population Genetics*. Oxford University Press, 2018. ISBN 978-0-87893-965-7. Google-Books-ID: 3BDkswEACAAJ.

M. K. Halushka, J.-B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz, and A. Chakravarti. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genetics*, 22(3):239–247, July 1999. ISSN 1546-1718. doi: 10.1038/10297. URL https://www.nature.com/articles/ng0799_239. Publisher: Nature Publishing Group.

I. Hellmann, I. Ebersberger, S. E. Ptak, S. Pääbo, and M. Przeworski. A neutral explanation for the correlation of diversity with recombination rates in humans. *The American Journal of Human Genetics*, 72(6):1527–1535, June 2003. ISSN 0002-9297. doi: 10.1086/375657. URL https://www.sciencedirect.com/science/article/pii/S0002929707604510.

J. A. Hutcheon, A. Chiolero, and J. A. Hanley. Random measurement error and regression dilution bias. *BMJ*, 340: c2289, June 2010. ISSN 0959-8138, 1468-5833. doi: 10.1136/bmj.c2289. URL https://www.bmj.com/content/340/bmj.c2289. Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting.

E. Ibarra-Laclette, E. Lyons, G. Hernández-Guzmán, C. A. Pérez-Torres, L. Carretero-Paulet, T.-H. Chang, T. Lan, A. J. Welch, M. J. A. Juárez, J. Simpson, A. Fernández-Cortés, M. Arteaga-Vázquez, E. Góngora-Castillo, G. Acevedo-Hernández, S. C. Schuster, H. Himmelbauer, A. E. Minoche, S. Xu, M. Lynch, A. Oropeza-Aburto, S. A. Cervantes-Pérez, M. de Jesús Ortega-Estrada, J. I. Cervantes-Luevano, T. P. Michael, T. Mockler, D. Bryant, A. Herrera-Estrella, V. A. Albert, and L. Herrera-Estrella. Architecture and evolution of a minute plant genome. *Nature*, 498(7452):94–98, June 2013. ISSN 1476-4687. doi: 10.1038/nature12132. URL `https://www.nature.com/articles/nature12132`. Publisher: Nature Publishing Group.

B. Institute. Picard toolkit. *Broad Institute, GitHub repository*, 2019. URL `https://broadinstitute.github.io/picard/`.

B. Jaegle, R. Pisupati, L. M. Soto-Jiménez, R. Burns, F. A. Rabanal, and M. Nordborg. Extensive sequence duplication in *Arabidopsis* revealed by pseudo-heterozygosity. *Genome Biology*, 24(1):44, Mar. 2023. ISSN 1474-760X. doi: 10.1186/s13059-023-02875-3. URL `https://doi.org/10.1186/s13059-023-02875-3`.

P. Johnsson, L. Lipovich, D. Grandér, and K. V. Morris. Evolutionary conservation of long noncoding RNAs; sequence, structure, function. *Biochimica et biophysica acta*, 1840(3):1063–1071, Mar. 2014. ISSN 0006-3002. doi: 10.1016/j.bbagen.2013.10.035. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3909678/`.

P. Johri, B. Charlesworth, and J. D. Jensen. Toward an evolutionarily appropriate null model: jointly inferring demography and purifying selection. *Genetics*, 215(1):173–192, May 2020. ISSN 1943-2631. doi: 10.1534/genetics.119.303002. URL `https://doi.org/10.1534/genetics.119.303002`.

P. Johri, C. F. Aquadro, M. Beaumont, B. Charlesworth, L. Excoffier, A. Eyre-Walker, P. D. Keightley, M. Lynch, G. McVean, B. A. Payseur, S. P. Pfeifer, W. Stephan, and J. D. Jensen. Recommendations for improving statistical inference in population genomics. *PLOS Biology*, 20(5):e3001669, May 2022a. ISSN 1545-7885. doi: 10.1371/journal.pbio.3001669. URL `https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3001669`. Publisher: Public Library of Science.

P. Johri, A. Eyre-Walker, R. N. Gutenkunst, K. E. Lohmueller, and J. D. Jensen. On the prospect of achieving accurate joint estimation of selection with population history. *Genome Biology and Evolution*, 14(7):evac088, July 2022b. ISSN 1759-6653. doi: 10.1093/gbe/evac088. URL `https://doi.org/10.1093/gbe/evac088`.

M. G. Kidwell. Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, 115(1):49–63, May 2002. ISSN 1573-6857. doi: 10.1023/A:1016072014259. URL `https://doi.org/10.1023/A:1016072014259`.

J.-H. Kim, J.-S. Park, C.-Y. Lee, M.-G. Jeong, J. L. Xu, Y. Choi, H.-W. Jung, and H.-K. Choi. Dissecting seed pigmentation-associated genomic loci and genes by employing dual approaches of reference-based and k-mer-based GWAS with 438 *Glycine* accessions. *PLOS ONE*, 15(12):e0243085, Dec. 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0243085. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0243085`. Publisher: Public Library of Science.

M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, 1983. ISBN 978-0-521-31793-1. Google-Books-ID: olIoSumPevYC.

K. Kojima and H. E. Schaffer. Accumulation of epistatic gene complexes. *Evolution*, 18(1):127–129, Mar. 1964. ISSN 0014-3820, 1558-5646. doi: 10.1111/j.1558-5646.1964.tb01577.x. URL `https://academic.oup.com/evolut/article/18/1/127/6868468`.

K.-i. Kojima and H. E. Schaffer. Survival Process of Linked Mutant Genes. *Evolution*, 21(3):518–531, 1967. ISSN 0014-3820. doi: 10.2307/2406613. URL `https://www.jstor.org/stable/2406613`. Publisher: [Society for the Study of Evolution, Wiley].

M. Kokot, M. Długosz, and S. Deorowicz. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*, 33 (17):2759–2761, Sept. 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx304. URL `https://academic.oup.com/bioinformatics/article/33/17/2759/3796399`. Publisher: Oxford Academic.

K. L. Korunes and K. Samuk. pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources*, 21(4):1359–1368, 2021. ISSN 1755-0998. doi: 10.1111/1755-0998.13326. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13326. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13326.

S. Kumar, G. Stecher, M. Suleski, and S. B. Hedges. TimeTree: a resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution*, 34(7):1812–1819, July 2017. ISSN 0737-4038. doi: 10.1093/molbev/msx116. URL https://doi.org/10.1093/molbev/msx116.

S. Kumar, M. Suleski, J. M. Craig, A. E. Kasprowicz, M. Sanderford, M. Li, G. Stecher, and S. B. Hedges. TimeTree 5: an expanded resource for species divergence times. *Molecular Biology and Evolution*, 39(8):msac174, Aug. 2022. ISSN 1537-1719. doi: 10.1093/molbev/msac174. URL https://doi.org/10.1093/molbev/msac174.

E. M. Leffler, K. Bullaughey, D. R. Matute, W. K. Meyer, L. Ségurel, A. Venkat, P. Andolfatto, and M. Przeworski. Revisiting an old riddle: what determines genetic diversity levels within species? *PLOS Biology*, 10(9):e1001388, Sept. 2012. ISSN 1545-7885. doi: 10.1371/journal.pbio.1001388. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001388. Publisher: Public Library of Science.

R. C. Lewontin. *The genetic basis of evolutionary change*. Columbia University Press, 1974. ISBN 0-231-03392-3.

H. Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, May 2013. URL http://arxiv.org/abs/1303.3997. arXiv:1303.3997 [q-bio].

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25 (14):1754–1760, July 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp324. URL https://doi.org/10.1093/bioinformatics/btp324.

W. H. Li and L. A. Sadler. Low Nucleotide Diversity in Man. *Genetics*, 129(2):513–523, Oct. 1991. ISSN 0016-6731. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1204640/.

M. Lynch and J. S. Conery. The origins of genome complexity. *Science*, 302(5649):1401–1404, Nov. 2003. doi: 10.1126/science.1089370. URL https://www.science.org/doi/full/10.1126/science.1089370. Publisher: American Association for the Advancement of Science.

A. Mackintosh, D. R. Laetsch, A. Hayward, B. Charlesworth, M. Waterfall, R. Vila, and K. Lohse. The determinants of genetic diversity in butterflies. *Nature Communications*, 10(1):3466, Aug. 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11308-4. URL https://www.nature.com/articles/s41467-019-11308-4. Number: 1 Publisher: Nature Publishing Group.

W. Makałowski and M. S. Boguski. Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. *Proceedings of the National Academy of Sciences*, 95(16):9407–9412, Aug. 1998. doi: 10.1073/pnas.95.16.9407. URL https://www.pnas.org/doi/10.1073/pnas.95.16.9407. Publisher: Proceedings of the National Academy of Sciences.

E. H. Margulies, M. Blanchette, N. C. S. Program, D. Haussler, and E. D. Green. Identification and characterization of multi-species conserved sequences. *Genome Research*, 13(12):2507–2518, Dec. 2003. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.1602203. URL https://genome.cshlp.org/content/13/12/2507. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

T. M. Mattila, J. Tyrmi, T. Pyhäjärvi, and O. Savolainen. Genome-Wide Analysis of Colonization History and Concomitant Selection in Arabidopsis lyrata. *Molecular Biology and Evolution*, 34(10):2665–2677, Oct. 2017. ISSN 0737-4038. doi: 10.1093/molbev/msx193. URL https://doi.org/10.1093/molbev/msx193.

A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, Sept. 2010. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.107524.110. URL `https://genome.cshlp.org/content/20/9/1297`. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Z. Mehrab, J. Mobin, I. A. Tahmid, and A. Rahman. Efficient association mapping from k-mers—An application in finding sex-specific sequences. *PLOS ONE*, 16(1):e0245058, Jan. 2021. ISSN 1932-6203. doi: 10.1371/journal.pone.0245058. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0245058`. Publisher: Public Library of Science.

W. Mei, M. G. Stetter, D. J. Gates, M. C. Stitzer, and J. Ross-Ibarra. Adaptation in plant genomes: Bigger is different. *American Journal of Botany*, 105(1):16–19, 2018. ISSN 1537-2197. doi: 10.1002/ajb2.1002. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/ajb2.1002`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajb2.1002.

C. D. Mirchandani, A. J. Shultz, G. W. C. Thomas, S. J. Smith, M. Baylis, B. Arnold, R. Corbett-Detig, E. Enbody, and T. B. Sackton. A fast, reproducible, high-throughput variant calling workflow for population genomics. *Molecular Biology and Evolution*, 41(1):msad270, Jan. 2024. ISSN 1537-1719. doi: 10.1093/molbev/msad270. URL `https://doi.org/10.1093/molbev/msad270`.

E. N. Moriyama and J. R. Powell. Intraspecific nuclear DNA variation in Drosophila. *Molecular Biology and Evolution*, 13(1):261–277, Jan. 1996. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a025563. URL `https://doi.org/10.1093/oxfordjournals.molbev.a025563`.

C. Muñoz-Diez, C. Vitte, J. Ross-Ibarra, B. S. Gaut, and M. I. Tenaillon. Using nextgen sequencing to investigate genome size variation and transposable element content. In M.-A. Grandbastien and J. M. Casacuberta, editors, *Plant Transposable Elements: Impact on Genome Structure and Function*, pages 41–58. Springer, Berlin, Heidelberg, 2012. ISBN 978-3-642-31842-9. doi: 10.1007/978-3-642-31842-9_3. URL `https://doi.org/10.1007/978-3-642-31842-9_3`.

L. Nab, M. van Smeden, R. H. Keogh, and R. H. H. Groenwold. Mecor: An R package for measurement error correction in linear regression models with a continuous outcome. *Computer Methods and Programs in Biomedicine*, 208:106238, Sept. 2021. ISSN 0169-2607. doi: 10.1016/j.cmpb.2021.106238. URL `https://www.sciencedirect.com/science/article/pii/S0169260721003126`.

M. Nei and D. Graur. Extent of protein polymorphism and the neutral mutation theory. *Evolutionary Biology*, 17: 73–118, 1984.

M. Nordborg, T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian, H. Zheng, E. Bakker, P. Calabrese, J. Gladstone, R. Goyal, M. Jakobsson, S. Kim, Y. Morozov, B. Padhukasahasram, V. Plagnol, N. A. Rosenberg, C. Shah, J. D. Wall, J. Wang, K. Zhao, T. Kalbfleisch, V. Schulz, M. Kreitman, and J. Bergelson. The pattern of polymorphism in *Arabidopsis thaliana*. *PLOS Biology*, 3(7):e196, May 2005. ISSN 1545-7885. doi: 10.1371/journal.pbio.0030196. URL `https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0030196`. Publisher: Public Library of Science.

B. Nystedt, N. R. Street, A. Wetterbom, A. Zuccolo, Y.-C. Lin, D. G. Scofield, F. Vezzi, N. Delhomme, S. Giacomello, A. Alexeyenko, R. Vicedomini, K. Sahlin, E. Sherwood, M. Elfstrand, L. Gramzow, K. Holmberg, J. Hällman, O. Keech, L. Klasson, M. Koriabine, M. Kucukoglu, M. Käller, J. Luthman, F. Lysholm, T. Niittylä, Olson, N. Rilakovic, C. Ritland, J. A. Rosselló, J. Sena, T. Svensson, C. Talavera-López, G. Theißen, H. Tuominen, K. Vanneste, Z.-Q. Wu, B. Zhang, P. Zerbe, L. Arvestad, R. Bhalerao, J. Bohlmann, J. Bousquet, R. Garcia Gil, T. R. Hvidsten, P. de Jong, J. MacKay, M. Morgante, K. Ritland, B. Sundberg, S. Lee Thompson, Y. Van de Peer, B. Andersson, O. Nilsson, P. K. Ingvarsson, J. Lundeberg, and S. Jansson. The Norway spruce genome sequence and conifer genome evolution. *Nature*, 497(7451):579–584, May 2013. ISSN 1476-4687. doi: 10.1038/nature12211. URL `https://www.nature.com/articles/nature12211`. Publisher: Nature Publishing Group.

B. D. Ondov, T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren, and A. M. Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1):132, June 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0997-x. URL https://doi.org/10.1186/s13059-016-0997-x.

H. Opedal, W. S. Armbruster, T. F. Hansen, A. Holstad, C. Pélabon, S. Andersson, D. R. Campbell, C. M. Caruso, L. F. Delph, C. G. Eckert, Lankinen, G. M. Walter, J. Ågren, and G. H. Bolstad. Evolvability and trait function predict phenotypic divergence of plant populations. *Proceedings of the National Academy of Sciences*, 120(1): e2203228120, Jan. 2023. doi: 10.1073/pnas.2203228120. URL https://www.pnas.org/doi/10.1073/pnas.2203228120. Publisher: Proceedings of the National Academy of Sciences.

D. Orme, R. Freckleton, G. Thomas, T. Petzoldt, S. Fritz, N. Isaac, and W. Pearse. caper: Comparative Analyses of Phylogenetics and Evolution in R, Apr. 2018. URL https://CRAN.R-project.org/package=caper.

T. Paape, P. Zhou, A. Branca, R. Briskine, N. Young, and P. Tiffin. Fine-scale population recombination rates, hotspots, and correlates of recombination in the *Medicago truncatula* genome. *Genome Biology and Evolution*, 4(5): 726–737, Jan. 2012. ISSN 1759-6653. doi: 10.1093/gbe/evs046. URL https://doi.org/10.1093/gbe/evs046.

B. Pateiro-Lopez and A. Rodriguez-Casal. alphahull: Generalization of the convex hull of a sample of points in the plane, 2022. URL https://CRAN.R-project.org/package=alphahull.

C. R. Peart, S. Tusso, S. D. Pophaly, F. Botero-Castro, C.-C. Wu, D. Aurioles-Gamboa, A. B. Baird, J. W. Bickham, J. Forcada, F. Galimberti, N. J. Gemmell, J. I. Hoffman, K. M. Kovacs, M. Kunnasranta, C. Lydersen, T. Nyman, L. R. de Oliveira, A. J. Orr, S. Sanvito, M. Valtonen, A. B. A. Shafer, and J. B. W. Wolf. Determinants of genetic variation across eco-evolutionary scales in pinnipeds. *Nature Ecology & Evolution*, 4(8):1095–1104, Aug. 2020. ISSN 2397-334X. doi: 10.1038/s41559-020-1215-5. URL https://www.nature.com/articles/s41559-020-1215-5. Publisher: Nature Publishing Group.

E. Pebesma. Simple Features for R: Standardized support for spatial vector data. *The R Journal*, 10(1):439–446, 2018. URL https://doi.org/10.32614/RJ-2018-009.

T. Pedersen and F. Crameri. scico, Aug. 2023. URL https://github.com/thomasp85/scico.

J. Pellicer and I. J. Leitch. The plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytologist*, 226(2):301–305, 2020. ISSN 1469-8137. doi: 10.1111/nph.16261. URL https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/nph.16261. _eprint: https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/nph.16261.

J. M. Pflug, V. R. Holmes, C. Burrus, J. S. Johnston, and D. R. Maddison. Measuring genome sizes using read-depth, k-mers, and flow cytometry: Methodological comparisons in beetles (Coleoptera). *G3 Genes|Genomes|Genetics*, 10 (9):3047–3060, Sept. 2020. ISSN 2160-1836. doi: 10.1534/g3.120.401028. URL https://doi.org/10.1534/g3.120.401028.

T. N. Phung, C. D. Huber, and K. E. Lohmueller. Determining the effect of natural selection on linked neutral divergence across species. *PLOS Genetics*, 12(8):e1006199, Aug. 2016. ISSN 1553-7404. doi: 10.1371/journal.pgen.1006199. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006199. Publisher: Public Library of Science.

G. Piganeau and A. Eyre-Walker. Evidence for Variation in the Effective Population Size of Animal Mitochondrial DNA. *PLOS ONE*, 4(2):e4396, Feb. 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0004396. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0004396. Publisher: Public Library of Science.

R. Poplin, V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, G. A. V. d. Auwera, D. E. Kling, L. D. Gauthier, A. Levy-Moonshine, D. Roazen, K. Shakir, J. Thibault, S. Chandran, C. Whelan, M. Lek, S. Gabriel, M. J. Daly, B. Neale, D. G. MacArthur, and E. Banks. Scaling accurate genetic variant discovery to tens of

thousands of samples, July 2018. URL https://www.biorxiv.org/content/10.1101/201178v3. Pages: 201178 Section: New Results.

D. Quiroz, M. Lensink, D. J. Kliebenstein, and J. G. Monroe. Causes of mutation rate variability in plant genomes. *Annual Review of Plant Biology*, 74(1):751–775, 2023. doi: 10.1146/annurev-arplant-070522-054109. URL https://doi.org/10.1146/annurev-arplant-070522-054109. _eprint: https://doi.org/10.1146/annurev-arplant-070522-054109.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

A. Rahman, I. Hallgrímsdóttir, M. Eisen, and L. Pachter. Association mapping from sequencing reads using k-mers. *eLife*, 7:e32920, June 2018. ISSN 2050-084X. doi: 10.7554/eLife.32920. URL https://doi.org/10.7554/eLife.32920. Publisher: eLife Sciences Publications, Ltd.

T. R. Ranallo-Benavidez, K. S. Jaron, and M. C. Schatz. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, 11(1):1432, Mar. 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-14998-3. URL https://www.nature.com/articles/s41467-020-14998-3. Publisher: Nature Publishing Group.

E. S. Rice, A. Alberdi, J. Alfieri, G. Athrey, J. R. Balacco, P. Bardou, H. Blackmon, M. Charles, H. H. Cheng, O. Fedrigo, S. R. Fiddaman, G. Formenti, L. A. F. Frantz, M. T. P. Gilbert, C. J. Hearn, E. D. Jarvis, C. Klopp, S. Marcos, A. S. Mason, D. Velez-Irizarry, L. Xu, and W. C. Warren. A pangenome graph reference of 30 chicken genomes allows genotyping of large and complex structural variants. *BMC Biology*, 21(1):267, Nov. 2023. ISSN 1741-7007. doi: 10.1186/s12915-023-01758-0. URL https://doi.org/10.1186/s12915-023-01758-0.

W. C. Riddell. Prediction in Generalized Least Squares. *The American Statistician*, 31(2):88–90, 1977. ISSN 0003-1305. doi: 10.2307/2683049. URL https://www.jstor.org/stable/2683049. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

J. Romiguier, P. Gayral, M. Ballenghien, A. Bernard, V. Cahais, A. Chenuil, Y. Chiari, R. Dernat, L. Duret, N. Faivre, E. Loire, J. M. Lourenco, B. Nabholz, C. Roux, G. Tsagkogeorga, A. a.-T. Weber, L. A. Weinert, K. Belkhir, N. Bierne, S. Glémin, and N. Galtier. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, 515(7526):261–263, Nov. 2014. ISSN 1476-4687. doi: 10.1038/nature13685. URL https://www.nature.com/articles/nature13685. Number: 7526 Publisher: Nature Publishing Group.

K. Roselius, W. Stephan, and T. Städler. The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics*, 171(2):753–763, Oct. 2005. ISSN 0016-6731. doi: 10.1534/genetics.105.043877. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1456785/.

P. Ruperao, P. Gandham, D. A. Odeny, S. Mayes, S. Selvanayagam, N. Thirunavukkarasu, R. R. Das, M. Srikanda, H. Gandhi, E. Habyarimana, E. Manyasa, B. Nebie, S. P. Deshpande, and A. Rathore. Exploring the sorghum race level diversity utilizing 272 sorghum accessions genomic resources. *Frontiers in Plant Science*, 14, 2023. ISSN 1664-462X. URL https://www.frontiersin.org/articles/10.3389/fpls.2023.1143512.

D. Sanchez, S. B. Sadoun, T. Mary-Huard, A. Allier, L. Moreau, and A. Charcosset. Improving the use of plant genetic resources to sustain breeding programs' efficiency. *Proceedings of the National Academy of Sciences*, 120 (14):e2205780119, Apr. 2023. doi: 10.1073/pnas.2205780119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2205780119. Publisher: Proceedings of the National Academy of Sciences.

F. L. Sandell, N. Stralis-Pavese, J. M. McGrath, B. Schulz, H. Himmelbauer, and J. C. Dohm. Genomic distances reveal relationships of wild and cultivated beets. *Nature Communications*, 13(1):2021, Apr. 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29676-9. URL https://www.nature.com/articles/s41467-022-29676-9. Number: 1 Publisher: Nature Publishing Group.

K. J. Schmid, S. Ramos-Onsins, H. Ringys-Beckstein, B. Weisshaar, and T. Mitchell-Olds. A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. *Genetics*, 169(3):1601–1615, Mar. 2005. ISSN 1943-2631. doi: 10.1534/genetics.104.033795. URL https://doi.org/10.1534/genetics.104.033795.

T. L. Schmidt, M.-E. Jasper, A. R. Weeks, and A. A. Hoffmann. Unbiased population heterozygosity estimates from genome-wide sequence data. *Methods in Ecology and Evolution*, 12(10):1888–1898, 2021. ISSN 2041-210X. doi: 10.1111/2041-210X.13659. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13659. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13659.

A. Shajii, D. Yorukoglu, Y. William Yu, and B. Berger. Fast genotyping of known SNPs through approximate k-mer matching. *Bioinformatics*, 32(17):i538–i544, Sept. 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw460. URL https://doi.org/10.1093/bioinformatics/btw460.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL https://ieeexplore.ieee.org/abstract/document/6773024. Conference Name: The Bell System Technical Journal.

A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, Aug. 2005. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.3715005. URL https://genome.cshlp.org/content/15/8/1034. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

O. B. Silva-Junior and D. Grattapaglia. Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of *Eucalyptus grandis*. *New Phytologist*, 208(3):830–845, 2015. ISSN 1469-8137. doi: 10.1111/nph.13505. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/nph.13505. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/nph.13505.

T. Slotte. The impact of linked selection on plant genomic variation. *Briefings in Functional Genomics*, 13(4): 268–275, July 2014. ISSN 2041-2649. doi: 10.1093/bfgp/elu009. URL https://doi.org/10.1093/bfgp/elu009.

J. M. Smith and J. Haigh. The hitch-hiking effect of a favourable gene. *Genetics Research*, 23(1):23–35, Feb. 1974. ISSN 1469-5073, 0016-6723. doi: 10.1017/S0016672300014634. URL https://www.cambridge.org/core/journals/genetics-research/article/hitchhiking-effect-of-a-favourable-gene/918291A3B62BD50E1AE5C1F22165EF1B.

J. Sopniewski and R. A. Catullo. Estimates of heterozygosity from single nucleotide polymorphism markers are context-dependent and often wrong. *Molecular Ecology Resources*, n/a(n/a):e13947, Mar. 2024. ISSN 1755-0998. doi: 10.1111/1755-0998.13947. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13947. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13947.

A. South. rworldmap: A new R package for mapping global data. *The R Journal*, 3(1):35–43, June 2011. URL http://journal.r-project.org/archive/2011-1/RJournal_2011-1_South.pdf.

M. I. Tenaillon, M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley, and B. S. Gaut. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. mays L.). *Proceedings of the National Academy of Sciences*, 98(16):9161–9166, July 2001. doi: 10.1073/pnas.151244298. URL https://www.pnas.org/doi/abs/10.1073/pnas.151244298. Publisher: Proceedings of the National Academy of Sciences.

M. I. Tenaillon, M. B. Hufford, B. S. Gaut, and J. Ross-Ibarra. Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. *Genome Biology and Evolution*, 3:219–229, Jan. 2011. ISSN 1759-6653. doi: 10.1093/gbe/evr008. URL https://doi.org/10.1093/gbe/evr008.

I. Turner, K. V. Garimella, Z. Iqbal, and G. McVean. Integrating long-range connectivity information into de Bruijn graphs. *Bioinformatics*, 34(15):2556–2565, Aug. 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty157. URL https://doi.org/10.1093/bioinformatics/bty157.

A. VanWallendael and M. Alvarez. Alignment-free methods for polyploid genomes: Quick and reliable genetic distance estimation. *Molecular Ecology Resources*, 22(2):612–622, 2022. ISSN 1755-0998. doi: 10.1111/1755-0998.13499. URL http://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13499. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13499.

Y. Voichek and D. Weigel. Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nature Genetics*, 52(5):534–540, May 2020. ISSN 1546-1718. doi: 10.1038/s41588-020-0612-7. URL https://www.nature.com/articles/s41588-020-0612-7. Number: 5 Publisher: Nature Publishing Group.

G. W. Vurture, F. J. Sedlazeck, M. Nattestad, C. J. Underwood, H. Fang, J. Gurtowski, and M. C. Schatz. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, 33(14):2202–2204, July 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx153. URL https://doi.org/10.1093/bioinformatics/btx153.

J. Wang, N. R. Street, D. G. Scofield, and P. K. Ingvarsson. Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related *Populus* species. *Genetics*, 202(3):1185–1200, Mar. 2016. ISSN 1943-2631. doi: 10.1534/genetics.115.183152. URL https://doi.org/10.1534/genetics.115.183152.

J. Wang, W. Yang, S. Zhang, H. Hu, Y. Yuan, J. Dong, L. Chen, Y. Ma, T. Yang, L. Zhou, J. Chen, B. Liu, C. Li, D. Edwards, and J. Zhao. A pangenome analysis pipeline provides insights into functional gene identification in rice. *Genome Biology*, 24(1):19, Jan. 2023. ISSN 1474-760X. doi: 10.1186/s13059-023-02861-9. URL https://doi.org/10.1186/s13059-023-02861-9.

K. D. Whitney, E. J. Baack, J. L. Hamrick, M. J. W. Godt, B. C. Barringer, M. D. Bennett, C. G. Eckert, C. Goodwillie, S. Kalisz, I. J. Leitch, and J. Ross-Ibarra. A role for nonadaptive processes in plant genome size evolution? *Evolution*, 64(7):2097–2109, July 2010. ISSN 0014-3820. doi: 10.1111/j.1558-5646.2010.00967.x. URL https://doi.org/10.1111/j.1558-5646.2010.00967.x.

R. J. Williamson, E. B. Josephs, A. E. Platts, K. M. Hazzouri, A. Haudry, M. Blanchette, and S. I. Wright. Evidence for Widespread Positive and Negative Selection in Coding and Conserved Noncoding Regions of Capsella grandiflora. *PLOS Genetics*, 10(9):e1004622, Sept. 2014. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004622. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004622. Publisher: Public Library of Science.

J. C. Willis. *Age and Area: A Study in Geographical Distribution and Origin of Species*. The University Press, 1922. Google-Books-ID: yBs4AAAAMAAJ.

A. Woolfe, M. Goodson, D. K. Goode, P. Snell, G. K. McEwen, T. Vavouri, S. F. Smith, P. North, H. Callaway, K. Kelly, K. Walter, I. Abnizova, W. Gilks, Y. J. K. Edwards, J. E. Cooke, and G. Elgar. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biology*, 3(1):e7, Jan. 2005. ISSN 1544-9173. doi: 10.1371/journal.pbio.0030007. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC526512/.

S. Wright. Breeding structure of populations in relation to speciation. *The American Naturalist*, 74(752):232–248, 1940. ISSN 0003-0147. URL https://www.jstor.org/stable/2457575. Publisher: [University of Chicago Press, American Society of Naturalists].

Y. Zhou, Z. Zhang, Z. Bao, H. Li, Y. Lyu, Y. Zan, Y. Wu, L. Cheng, Y. Fang, K. Wu, J. Zhang, H. Lyu, T. Lin, Q. Gao, S. Saha, L. Mueller, Z. Fei, T. Städler, S. Xu, Z. Zhang, D. Speed, and S. Huang. Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*, 606(7914):527–534, June 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04808-9. URL `https://www.nature.com/articles/s41586-022-04808-9`. Number: 7914 Publisher: Nature Publishing Group.

M. E. Zwick, D. J. Cutler, and A. Chakravarti. Patterns of Genetic Variation in Mendelian and Complex Traits. *Annual Review of Genomics and Human Genetics*, 1(1):387–407, Sept. 2000. ISSN 1527-8204, 1545-293X. doi: 10.1146/annurev.genom.1.1.387. URL `https://www.annualreviews.org/doi/10.1146/annurev.genom.1.1.387`.