LEVERAGING LARGE DATASETS TO INVESTIGATE THE DISTRIBUTION OF FITNESS EFFECTS IN PLANTS

By

Miles David Roberts

A DISSERTATION

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

Genetics and Genome Sciences – Doctor of Philosophy

2025

ABSTRACT

Mutations—spontaneous changes in DNA sequences—are fundamental to evolution, and contribute to everything from disease susceptibility in people to agriculture. Population geneticists are tasked with understanding mutations - historically categorizing them as either beneficial, harmful, or neutral based on how they affect fitness. However, our understanding of these categories is often limited by incomplete data and an inability to travel back in time to track mutations. Over four chapters, I will explore how large datasets can enhance our understanding of these three fundamental mutation categories. First, I will use 27 terabases of gene expression data from 300 studies to investigate whether rarely expressed genes accumulate harmful mutations at higher rates than constitutively expressed genes. Next, I will leverage about 205 terabases of DNA-sequencing data and use k-mers (DNA subsequences) to measure neutral variation in 112 natural plant species. The results suggest that current methods for estimating diversity reliant on reference genomes underestimate genetic variation. The following chapter expands upon the idea of using k-mers to measure genetic diversity, numerically and analytically exploring the relationship between k-mer diversity and nucleotide diversity. Overall, k-mer diversity scales linearly with nucleotide diversity and we showcase the use of bloom filters to decrease the memory burden of the k-mer diversity calculations. Finally, I will use 250,000 evolutionary simulations to train machine learning models to infer the time it took for a fixed beneficial mutation to spread. Overall, large datasets like these hold many opportunities to revisit old biological questions and further our understanding of mutation trajectories.

To David, Cheree, Blake, and Paige Roberts

ACKNOWLEDGEMENTS

I am infinitely appreciative for the guidance and support of my advisor, Emily Josephs. Emily's patience, even in the face of my constant stubbornness, and her promotion of curiosity greatly shaped how I view research. This dissertation also greatly benefited from the input of my committee: Bob Vanburen, Shinhan Shiu, and Jeff Connor. I want to give a special thanks to Addie Thompson for being the outside evaluator for my comprehensive exam and for later stepping into my committee for my dissertation defense. I also want to thank Olivia Davis and Robert Williamson for being wonderful collaborators. We will see their awesome work in chapter 3.

I deeply appreciate the members of the Josephs lab, past and current, who have given me invaluable advice and have (usually) been willing to laugh at my puns. These folks are: Husain Agha, Sophie Buysse, Nate Catlin, Derek Denney, Asia Hightower, Daniela Palmer, Rebecca Panko, Gabbie Sandstedt, Mia Stevens, Chrissy Miller, Adrian Platts, Magie Williams, and Maya Wilson Brown. I was also lucky to have many amazing friends that helped keep my sprits high. In no particular order, these folks are: Laura Ford, Nick Johnson, Max Harman, Luke Strickland, Brandon Webster, Kim Fisher, Brandon Beal, John Botsford, Robin Waterman, and Carolyn Graham. Thank you all.

I also want to thank MSU's Council of Graduate Students and their dozens of representatives across MSU for their work in advocating for graduate and professional student needs and for giving me the opportunity to be their President. It was a great honor to represent my peers in that capacity and the experience shaped my understanding of leadership. Thank you especially to Deanne Arking and Allyn Shaw for helping me immensely with these efforts.

I want to acknowledge that it was my undergraduate genetics teacher William Cushwa that encouraged me to continue studying genetics and to engage in undergraduate research, which were both crucial to me pursuing a graduate degree. I did undergraduate research in Dr. Stephanie Porter's lab at Washington State University in Vancouver, Washington and I received incredibly valuable advice and mentorship that I continue to lean on to this day. I also want to acknowledge mentorship from the former post-docs Emily Helliwell and Camile Wendlandt, as well as former

graduate students Angeliqua Montoya, Zoie Lopez, and Niall Millar that were willing to train and mentor me during this time and motivated me to go to graduate school.

The work in this dissertation, as with all scientific works, leans heavily on the previous efforts of other scientists. I want to acknowledge that in chapters 1 and 2, I make extensive use of public databases with contributions from hundreds of scientists across the world. It is impossible to list all of the contributors here, but I am grateful for their work and for making their data public. Without their effort none of this would have been possible. I am also grateful for the anonymous peer reviewers that have commented on chapters 1-3.

I would like to thank the Plant Biology Department and the Ecology, Evolution, and Behavior Program at MSU for providing a welcoming environment where I could always participate even though I had technically no affiliation with either of them. My home program of Genetics and Genome Sciences was also always supportive and flexible in allowing me to pursue my personal educational goals.

I would like to thank the Biomolecular Sciences Program for supporting the 1st year of my PhD, the National Science Foundation and the National Institutes of Health for supporting the 2nd, 3rd, and 4th years of my PhD. I would also like to thank MSU's Institute for Cyber-Enabled Research for providing the computing power and a fellowship to learn cloud-based computing approaches.

To my family, David, Cheree, Blake, and Paige Roberts, your support is the reason why I got to pursue an education. I do not know where I would be without it. I love you all very much. To my partner, Riley Pizza, thank you for your endless support. You are the single greatest thing that's happened to me in graduate school and I cannot wait to see where our journey takes us.

- Miles

PREFACE

This dissertation is my original work. Chapters 1 and 3 are published in peer-reviewed journals under a CC-BY license, allowing unlimited sharing and reproduction as long as attribution is given. Chapter 2 is accepted for publication at Evolution Letters, but the preprint version is available at biorxiv under a CC-BY license. Chapter 4 is unpublished. I was responsible for all of data collection, data analysis, and writing in Chapters 1, 2, and 4 (with guidance from my advisor and committee). Chapter 3 was co-first authored by myself and Olivia Davis - a former undergraduate (now graduate) of Rose-Hulman Institute of Technology in Terre Haute, Indiana. Olivia carried out the simulations depicted in Figures 3.3 - 3.5, D1, and D2. I performed the simulations depicted in Figure 3.1, made Figure 3.2, and derived equations 3.7, 3.8, and 3.15. Olivia also wrote the text in the section **Testing the efficacy of k-mer measures of variation**, but all other text was written by me.

TABLE OF CONTENTS

INTRODUCTION

Population geneticists study genetic variation within and between populations. To understand variation, we need to understand the fate of new mutations. What determines whether a mutation spreads through a population or goes extinct? There are three frontiers that are currently revolutionizing our understanding of this question: (1) the collection of data across many environments, (2) the assembly of multiple genomes per population (i.e. pangenomics), and (3) machine learning (Hahn, 2018). In the course of this dissertation, we will familiarize ourselves with each of these areas, but to prepare for the journey, we will need a mental model as a guide.

Let's imagine the amount of genetic diversity in a population as the level of water in a pool. In this analogy, the process of mutation is the faucet or hose that adds water into the pool. If this were the only dynamic, then our pool would fill up forever. Thus, our pool must have some drains. One important drain is natural selection, which is a process where mutations that cause heritable, deleterious changes in survival or reproduction will not spread through a population and may eventually go extinct (i.e. purifying selection). On the other hand, mutations that cause heritable, beneficial changes to survival and reproduction will spread and may eventually be fixed (i.e. positive selection). If we imagine a world where all mutations are deleterious then we would expect natural selection to constantly remove mutations from our population. However, this expectation clearly does not match reality because essentially all natural populations have genetic variation (Leffler et al., 2012). One way to modify our expectation is to recognize that natural selection is not an omnipotent force - the pool drain only opens under certain circumstances.

One important limit to natural selection is that it can only act *when* a mutation is affecting fitness. If a mutation occurs in a non-expressed gene, for instance, then such a mutation may not affect fitness and will be invisible to natural selection. This masking of deleterious mutations can theoretically happen in any gene that is only expressed at certain times or in certain environments (Van Dyken and Wade, 2010). But on the other hand, natural selection often favors conditional expression (Snell-Rood et al., 2010). Many species of plants have evolved specific gene expression responses for specific stimuli, including mechanical stress (Braam and Davis, 1990), light stress

(Mishra et al., 2012), and heat stress (Zhang et al., 2017b). Thus, a trade-off arises between the added benefit of conditional expression vs the weakened ability to purify genes of deleterious mutations. Though the theory behind this relationship is well-understood (Van Dyken and Wade, 2010), there are no empirical investigations of the relationship between environment-specific expression and deleterious variation. This is in part because gene expression could not be measured across many environments easily until relatively recently. Chapter 1 leverages publicly available gene expression data across 200 different environments in *Arabidopsis thaliana* to test whether genes with more environment-specific expression have more deleterious variation.

What if we assume most mutations are neutral? This idea is the foundation of Neutral Theory (Kimura, 1983). Under this framework, the amount of variation in a population is simply a product of how many individuals are in the population and the rate of mutation. In other words, more individuals replicating DNA to produce offspring means more opportunities for mutations to occur in DNA. This is a straightforward expectation that is anchored in the foundational works of population genetics and evolutionary biology. Charles Darwin observed that "wide ranging, much diffused and common species vary most" (Darwin (1859), Chapter II) and R. A. Fisher supported this observation mathematically as "a numerous species, with the same frequency of mutation, will maintain a higher variability than will a less numerous species" (Fisher, 1923). However, when scientists empirically estimate the scaling between population size and genetic diversity, the result is much weaker than expected (Lewontin, 1974; Leffler et al., 2012; Buffalo, 2021). There are many biological mechanisms that contribute to this paradoxical result (Charlesworth and Jensen, 2022). However, one underexplored explanation is whether our methods of measuring genetic diversity simply do not capture all of the variation present. Chapters 2 and 3 will investigate whether the ubiquitous practice of using a reference genome to estimate diversity contributes to this paradox and how patterns in k-mers (DNA subsequences) provide a potential solution.

Finally, what about beneficial mutations? Positive selection can also be thought of as a drain in our pool analogy because fixation of beneficial alleles removes diversity around the site of the beneficial mutaiton - a signature called a selective sweep (Whitehouse and Schrider, 2023). The

amount of diversity that positive selection removes is proportional to the time it takes for the mutation to fix (Coop, 2020). Shorter times to fixation result in greater loss of diversity because there is less time for mutation and recombination to break up the sweep. After fixation, mutation and recombination will gradually reintroduce variation at the site of the sweep over time. This suggests that the times to fixation and age of a sweep are encoded in patterns of genomic diversity around the sweep site. However, times to fixation rarely estimated in real data, despite being a core quantity in studies of neutral theory and selective sweeps (Charlesworth, 2020; Zhao et al., 2013). In chapter 4, we will use machine learning to try to estimate times to fixation using a snapshot of a contemporary population.

CHAPTER 1: WEAKER SELECTION ON GENES WITH TREATMENT SPECIFIC EXPRESSION CONSISTENT WITH A LIMIT ON PLASTICITY EVOLUTION IN *ARABIDOPSIS THALIANA*[1]

Abstract

Differential gene expression between environments often underlies phenotypic plasticity. However, environment-specific expression patterns are hypothesized to relax selection on genes, and thus limit plasticity evolution. We collated over 27 terabases of RNA-sequencing data on *Arabidopsis thaliana* from over 300 peer-reviewed studies and 200 treatment conditions to investigate this hypothesis. Consistent with relaxed selection, genes with more treatment-specific expression have higher levels of nucleotide diversity and divergence at nonsynonymous sites but lack stronger signals of positive selection. This result persisted even after controlling for expression level, gene length, GC content, the tissue specificity of expression, and technical variation between studies. Overall, our investigation supports the existence of a hypothesized trade-off between the environment specificity of a gene's expression and the strength of selection on said gene in *A. thaliana*. Future studies should leverage multiple genome-scale datasets to tease apart the contributions of many variables in limiting plasticity evolution.

Introduction

Organisms must cope with ever-changing environmental conditions to survive and reproduce. If these changes in condition cannot be avoided or escaped, phenotypes that respond to environmental variation through phenotypic plasticity may be adaptive. For example, under low light, the same *Arabidopsis thaliana* genotype will produce more or larger leaves to capture more energy for photosynthesis (Pigliucci and Kolodynska, 2002). Plastic responses are partly controlled through differential gene expression between environments (Scheiner, 1993; Schlichting and Smith, 2002). Understanding the evolution of these condition-specific expression patterns could help reconcile the diversity of plastic responses observed in nature and engineer organisms to overcome environmental challenges.

However, not all organisms can respond plastically to environmental change, so it is crucial to understand the processes that constrain plasticity (Van Kleunen and Fischer, 2005). These constraints are usually characterized as either costs, where plasticity reduces fitness in some way, or limits to the evolution or maintenance of plasticity (DeWitt et al., 1998). Decades of research has attempted to measure the costs associated with plasticity (reviewed in Schneider 2022) but studies often fail to detect costs or find costs that are weak or restricted to certain environments (Van Kleunen and Fischer, 2005; Van Buskirk and Steiner, 2009; Auld et al., 2010). Theory also predicts that there will be strong selection to alleviate costs (Murren et al., 2015). Thus, limits may be more important than costs in shaping the evolution of plasticity.

Recent work suggests that relaxed selection can limit plasticity evolution (Snell-Rood et al., 2010; Murren et al., 2015). For instance, one hypothesis posits that genes are often under selection for environment-specific expression to minimize deleterious pleiotropy (Snell-Rood et al., 2010; McGuigan et al., 2014; Huber et al., 2017). However, narrowing the range of environments where a gene is expressed also reduces the opportunity for negative selection to act on deleterious mutations in the gene (Kawecki, 1994; Whitlock, 1996; Van Dyken and Wade, 2010). The accumulation of deleterious mutations could then cancel out any selective benefits of the environment-specific expression pattern. Thus, a trade-off arises between a gene's degree of environment-specific expression and the strength of negative selection acting on said gene. If we assume that environment-specific expression generally contributes to phenotypic plasticity, then this trade-off would potentially limit the maintenance of plasticity (Kawecki, 1994; Snell-Rood et al., 2010). Whether such a trade-off exists has not yet been tested, but the deposition of expression data from hundreds of experimental treatments across hundreds of labs into public repositories now enables approximating environment specificity as treatment specificity and linking treatment-specific expression to the rate of evolution.

One challenge in studying the relationship between treatment specificity and protein evolution is that many factors influence evolutionary rates (for review, see Rocha 2006; Gaut et al. 2011; Koonin 2011; Zhang and Yang 2015) and these factors are hard to disentangle. A protein's expres-

sion level is often considered the best predictor of its evolutionary rate (Rocha, 2006) - a result observed across all domains of life (Zhang and Yang, 2015) and sometimes considered a "law" of genome evolution (Koonin, 2011). Among multicellular organisms, the degree of tissue specificity in expression is also generally predictive of evolutionary rates (Duret and Mouchiroud, 2000; Larracuente et al., 2008; Winter et al., 2004; Zhang and Li, 2004; Slotte et al., 2011; Bush et al., 2015; Mukherjee et al., 2016; Groen et al., 2020; Huang, 2022). Additional factors that also influence evolutionary rates include exon edge conservation (Bush et al., 2015), mutational bias (Wang et al., 2004; Ossowski et al., 2010), gene length (Mukherjee et al., 2016), gene age (Moutinho et al., 2022), GC content (Zhang et al., 2002; Mukherjee et al., 2016), expression stochasticity (Groen et al., 2020), involvement in general vs specialized metabolism (Mukherjee et al., 2016), identity as a regulatory or structural gene (Wheeler et al., 2022), recombination rate (Langley et al., 2012), codon-bias (Betancourt and Presgraves, 2002), mating system (Wright et al., 2002; Glémin, 2007; Payne and Alvarez-Ponce, 2018), gene compactness (Larracuente et al., 2008; Mukherjee et al., 2016), co-expression or protein-protein interaction network connectivity (Alvarez-Ponce and Fares, 2012; Masalia et al., 2017; Mähler et al., 2017; Alvarez-Ponce et al., 2017; Josephs et al., 2017), gene body methylation (Takuno and Gaut, 2012), metabolic flux (Colombo et al., 2014), protein structure (Lin et al., 2007), essentiallity (Nembaware et al., 2002; Yang et al., 2003; Davis and Petrov, 2004), and even plant height (Lanfear et al., 2013). This overabundance of possible explanatory variables suggests that massive genome-scale datasets and careful statistical analysis are required to tease out the influence of treatment-specific expression on evolutionary rates.

To investigate the influence of treatment-specific expression on evolutionary rates, we compiled a dataset of gene expression data across over 200 treatments from over 300 peer-reviewed studies in *A. thaliana*. We annotated RNA-sequencing runs from these studies using standardized ontologies, then processed all of them with the same pipeline. Finally, we combined the resulting gene expression matrix with estimates of selection based on within-species polymorphism and between-species divergence to investigate whether genes with treatment-specific expression were under weaker negative selection.

Methods

RNA-seq run annotation

We amassed an initial set of RNA-seq runs from the Sustech Arabidopsis RNA-seq database V2 (Zhang et al., 2020) (http://ipf.sustech.edu.cn/pub/athrdb/) excluding any samples not associated with a publication or lacking a tissue type label. On May 24th, 2022 we also downloaded all run metadata from the Sequence Read Archive (SRA) returned by the following search term: ("Arabidopsis thaliana"[Organism] AND "RNA"[Source]) OR ("Arabidopsis thaliana"[Organism] AND "RNA-Seq"[Strategy]) OR ("Arabidopsis thaliana"[Organism] AND "TRANSCRIPTOMIC"[Source]). All SRA runs were linked to their associated publications, if possible, using Entrez. Any SRA run numbers that we could not link to a PUBMED ID or DOI were omitted. We then manually removed all SRA runs that originated from transgenic, mutant, hybrid, grafted, cell culture, polyploid, or aneuploid samples based on information in the SRA metadata and associated publications. Runs from any naturally-occurring *A. thaliana* accession were included. We also omitted SRA runs that focused on sequencing non-coding RNA (ncRNA-seq, miRNA-seq, lncRNA-seq, sRNA-seq, etc.). After applying these criteria, any bioprojects with 8 or fewer SRA run numbers remaining were also omitted.

All runs were labeled with treatment and tissue type descriptions using the Plant Experimental Conditions Ontology (PECO) and the Plant Ontology (PO) (Cooper et al., 2018), respectively, based on information in their associated publications and SRA metadata. In our analysis, control exposure was defined as long day conditions (12 hrs light exposure or longer, but not constant light) and growing temperatures in the range of 18° - 26°, inclusive, without explicit application of stress or nutrient limitation. Warm treatments were defined as 27° or higher, while cold treatments were defined as 17° or lower. Any studies that did not report both day length and growing temperature were omitted. Any runs that could not be linked to treatments based on their annotations in the SRA or Sustech databases were also omitted. Treatment with polyethylene glycol (PEG) was categorized as drought exposure. Samples from plants that were recovering from stress were categorized according to the growth conditions of the recovery state instead of the stressed state. When appro-

priate, we labeled samples with multiple PECO terms. For example, a sample that was subjected to both heat stress and high light stress would get two PECO terms (one for each stress) and be treated separately from samples subjected to only heat stress or only light stress. Tissue type labels were eventually collapsed to the following categories: whole plant, shoot, root, leaf, seed, and a combined category of flower and fruit tissues. The flower and fruit tissue categories were combined because of their developmental relationship and small size relative to the other categories. In the end, we had a dataset of 24,101 sequencing runs from 306 published studies.

RNA-seq run processing

All RNA-seq runs were processed using the same workflow to remove the effects of bioinformatic processing differences between studies on expression level. First, runs were downloaded using the SRA toolkit (v2.10.7), but 90 runs were not publicly available and thus failed to download. All successfully downloaded runs were trimmed using fastp v0.23.1 (Chen et al., 2018), requiring a minimum quality score of 20 and a minimum read length of at least 25 bp (-q 20 -l 25). Trimming results were compiled using multiqc v1.7 (Ewels et al., 2016). All trimmed runs were then aligned to a decoy-aware transcriptome index made by combining the primary transcripts of the Araport11 genome annotation (Cheng et al., 2017) with the *A. thaliana* genome in salmon v1.2.1 (Patro et al., 2017) using an index size of 25bp. The salmon outputs of each run were then combined with a custom R script to create an gene-by-run expression matrix. We omitted 423 runs with a mapping rate < 1 %, 215 runs with zero mapped transcripts, and 18 genes with zero mapped transcripts across all runs from further analysis. We note that although this cut-off does not exclude samples with more modest mapping rates (e.g. 20 - 60 %) the choice to include these samples was to avoid removing large chunks of data as "outliers" and analyzing only those samples that conform to our expectations.

Whole genome sequence data processing

We downloaded whole genome sequencing data for 1135 *A. thaliana* accessions from the 1001 genomes project panel (SRA project SRP056687) (Alonso-Blanco et al., 2016) using the SRA toolkit. All runs were trimmed using fastp (Chen et al., 2018), requiring a minimum quality score

of 20 and a read length of at least 30 bp (-q 20 -l 30). Trimmed reads were then aligned to the *A. thaliana* reference genome using BWA v0.7.17 (Li and Durbin, 2009). The alignments were sorted and converted to BAM format with SAMTOOLS v1.11 (Danecek et al., 2021), then optical duplicates were marked with picardtools v2.22.1. Haplotypes were called for each accession, then combined and jointly genotyped with GATK v4.1.4.1 assuming a sample ploidy of 2, heterozygosity of 0.001, indel-heterozyogsity of 0.001, and minimum base quality score of 20. Invariant sites were included in the genotype calls with the –include-non-variant-sites option. All calls were restricted to only coding sequence (CDS) regions based on the Araport11 annotation by supplying a BED file of CDS coordinates made with bedtools (v2.29.2). Following Korunes and Samuk (2021), variant and invariant sites were filtered separately using both GATK and vcftools v0.1.15 (Danecek et al., 2011). Variant sites were filtered if they met any of the following criteria: QD < 2, QUAL < 30, MQ < 40, FS > 60, HaplotypeScore > 13, MQRankSum < -12.5, ReadPosRankSum < -8.0, mean depth < 10, mean depth > 75, missing genotype calls > 20%, being an indel, or having more than 2 alleles. In the end, 1,915,859 variant sites across all coding sequences were retained for further analysis. Invariant sites were filtered if they met any of the following criteria: QUAL > 100, mean depth < 10, mean depth > 75, missing genotype calls > 20%. Finally, variant sites were annotated using snpEff (Java v15.0.2) (Cingolani et al., 2012b) and variants labeled as either missense or synonymous were separated into different files using SnpSift (Cingolani et al., 2012a).

Selection estimated from between-species divergence

We identified 1:1 orthologs between the primary transcripts of *A. thaliana* and *Arabidopsis lyrata* with Orthofinder v2.5.4 (Emms and Kelly, 2019). For each 1:1 ortholog, we aligned their protein sequences with MAFFT L-INS-I v7.475 (Katoh and Standley, 2013), then converted the protein alignments to gapless codon-based alignments using pal2nal v14 (Suyama et al., 2006). Using the gapless codon-based alignments, we estimated dN/dS using the Nei and Gojobori (1986) method implemented as a custom Biopython v1.79 script and implemented through the codeml program in the PAML package v4.9 (Yang, 2007). Unlike codeml, the custom Biopython script also returns counts of nonsynonymous (N) and synonymous sites (S) within each gene as described in

Nei and Gojobori (1986), which we later used to calculate nucleotide diversity per nonsynonymous site (π_N) and per synonymous site (π_S). Before proceeding with more analyses, we confirmed that our estimates of dN and dS were consistent between our Biopython script and codeml (Figure A5, Pearson correlations $dN : \rho = 0.9998$, $dS : \rho = 0.9809$). The outputs of the Biopython script were used in all subsequent analyses.

Selection estimated from within-species polymorphism

Nucleotide diversity (π) was calculated for each gene with pixy v1.2.3.beta1 (Korunes and Samuk, 2021) three times: once using all sites (both variant and invariant), once using missense sites plus invariant sites, and once using synonymous sites plus invariant sites. These estimates were then converted to π, π_N, and π_S, respectively, by first multiplying the per site estimate output from pixy by the number of sites included in the analysis. Then, to get π_N and π_S, the values from analyses of missense plus invariant, and synonymous plus invariant sites were divided by the N and S values for each gene, respectively, as determined by the method in Nei and Gojobori (1986).

We next calculated Tajima's D for each gene. First, we calculated π and Watterson's Theta (θ_W) for each variant site i within a gene (π_i and θ_{Wi} respectively). In this case, π_i was calculated as:

$$\pi_i = \left(\frac{n_i}{n_i - 1}\right)\left(1 - \sum_{j=1}^{2} p_{ij}^2\right) \tag{1.1}$$

Where n_i is the number of sequenced chromosomes with non-missing genotypes for variant i, p_{i1} is the frequency of the reference allele, and p_{i2} is the frequency of the alternative allele. Then, θ_{Wi} was calculated as:

$$\theta_{Wi} = \frac{1}{a_i} \tag{1.2}$$

Where a_i is:

$$a_i = \sum_{j=1}^{n_i - 1} \frac{1}{j} \tag{1.3}$$

This calculation of θ_{Wi} is equivalent to the usual calculation of θ_W with the number of segregating

sites set to one. Next, the variance in Tajima's D was calculated for each site as:

$$Var(\pi_i - \theta_{Wi}) = \frac{\frac{n_i+1}{3(n_i-1)} - \frac{1}{a_i}}{a_i} \tag{1.4}$$

This is equivalent to equation 38 in Tajima (1989) with the number of segregating sites set to one.

Finally, the results of the above calculations were combined in the following formula:

$$D_i = \frac{\pi_i - \theta_{Wi}}{\sqrt{Var(\pi_i - \theta_{Wi})}} \tag{1.5}$$

To get Tajima's D for each gene, we then averaged across the D_i values for all the variant sites within a gene.

We also calculated the direction of selection (DoS) statistic (Stoletzki and Eyre-Walker, 2011) for each gene. Counts of nonsynonymous and synonymous polymorphisms within each gene (P_N and P_S, respectively) were determined with bedtools (v2.29.2). The number of nonsynonymous and synonymous differences (D_N and D_S, respectively) between *A. thaliana* genes and their 1:1 *A. lyrata* orthologs, if present, were estimated during the process of calculating dN/dS in Biopython as described above. We then combined these components into the following equation:

$$DoS = \frac{D_N}{D_N + D_S} - \frac{P_N}{P_N + P_S} \tag{1.6}$$

We chose this metric, as opposed to the proportion of amino acid substitutions driven by positive selection (α), because it is less biased than α (Stoletzki and Eyre-Walker, 2011) and was successfully used in studies similar to ours (Paape et al., 2013). Furthermore, we found that α often returns uninterpretable negative values when applied to *A. thaliana*, perhaps because of an excess of slightly deleterious polymorphisms (Nordborg et al., 2005) due to their predominantly selfing mating system (Charlesworth, 1994).

Treatment specificity

Treatment specificity (τ) was estimated separately for runs from each tissue type using the following formula (Yanai et al., 2005):

$$\tau = \frac{\sum_{i=1}^{N} 1 - \frac{x_i}{\max x}}{N-1} \tag{1.7}$$

Where x is the vector of average expression values of a gene in each treatment category, measured in transcripts per million (TPM), and where N is the number of treatment categories. Dividing by N means that τ varies between zero and one, where zero indicates no specificity and one indicates exclusive specificity to a single treatment. We used this metric of specificity because it is consistently more robust than others (Kryuchkova-Mostacci and Robinson-Rechavi, 2017) and is normalized by the number of treatments included, making it comparable across data sets. We also applied the same formula to calculate tissue specificity in several different treatment conditions.

Simulating correlations between average expression and specificity index

Average expression level and measures of expression specificity are correlated by definition because genes with more treatment/tissue-specific expression will have lower average expression across all treatment/tissue categories. We ran two simulations to better illustrate the factors driving the correlation between average expression and the specificity index, τ. In both simulations, we generated 1000 random matrices, where each element x_{ij} represented the expression of gene i in experiment j, by sampling from a zero-inflated negative binomial distribution:

$$x_{ij} \sim ZINegBinom(N, p_1, p_2) \tag{1.8}$$

Where the size and probability parameters of the negative binomial component were $N = 100$ and $p_1 = 0.1$, respectively, while the probability of an expression value being non-zero was $p_2 = 0.4$. All matrices included 5 groups of columns, with 5 columns per group, representing replicates of tissue/treatment groups. For both simulations, we averaged across columns within each group to simulate the calculation of tissue/treatment-wide averages. We then applied the formula for τ across the rows of this averaged matrix to get expression specificity. In one simulation, we calculated expression level by averaging across the rows of the expression matrix. In a second simulation, we excluded experiments where a gene was not expressed ($x_{ij} = 0$) from the calculation of average expression.

Average expression, length, GC content, family size

Calculating the average expression of each gene was a three-step process. First, we averaged together runs with matching SRA experiment IDs because these runs represented technical replicates of the same biological sample and treatment conditions. Second, we partitioned our gene-by-experiment expression matrix by the tissue type each sample came from. Finally, for each tissue type's expression matrix, we averaged across all of the expression values of each gene across all experiments, excluding values < 5 transcripts per million (TPM). We excluded values < 5 TPM from the average expression calculation to avoid a high correlation between average expression and treatment-specificity, as has been reported in previous studies (Slotte et al., 2011). This high correlation occurs because an environment-specific gene will by definition also have low average expression across environments it is rarely expressed in. Furthermore, we excluded values < 5 TPM to avoid including small expression values that could be artifacts of alignment error.

The length and GC content of each gene was measured using the bedtools nuc command (v2.29.2) and included each gene's introns and untranslated regions when present. We included introns and untranslated regions in the estimate of gene length because they play important roles in determining rates of protein evolution (Castillo-Davis et al., 2002; Eisenberg and Levanon, 2003). Finally, the family size for each gene was estimated as the number of *A. thaliana* genes in their respective orthogroups output by OrthoFinder.

Partial correlation analysis

Not all treatment-tissue combinations were sampled in the overall RNA-seq dataset, causing confounding between the treatment and tissue labels. We resolved this in two ways. First, we subset the data to only the treatment conditions where all tissue types were represented. Second, we subset the data by tissue type and analyzed each subset separately. For each subset, we calculated partial spearman correlations between treatment specificity and our measures of selection (dN, π_N, Tajima's D, and DoS) after accounting for average expression (excluding values TPM < 5), gene length, and GC content using the ppcor R package (Kim, 2015). For partial correlation analyses involving π_N and Tajima's D, we also controlled for gene family size. We did not account for gene

family size in partial correlation analyses involving dN or DoS because these metrics apply to only genes with one family member in this study. When calculating partial correlations involving dN, we excluded any genes with saturating divergence ($dS > 1$). All statistical analyses and data visualizations used R v4.0.3 and used color palettes in the scico R package (Crameri, 2018; Pedersen and Crameri, 2023).

Surrogate variable analysis

We recalculated treatment specificity and repeated all partial correlation analyses after correcting each data subset for technical between-experiment variation (i.e. batch effects), following an approach from (Fukushima and Pollock, 2020). Batch effects include variables that influence gene expression measurements but are not of interest to this study, such as the sequencing platform and the library prep protocol used in each experiment. First, with our data already subset by tissue type, we further subset to only include treatments with RNA-seq runs from at least two studies. This minimizes confounding between-treatment variation with the technical between-experiment variation we aimed to account for. We then applied surrogate variable analysis (SVA) using the svaseq() function within the SVA package (Leek and Storey, 2007) to each of these subsets. Briefly, SVA models gene expression as:

$$x_{ij} = \mu_i + f(y_i) + e_{ij} \tag{1.9}$$

Where x_{ij} is the expression of gene i in experiment j, μ_i is the average expression of gene i across all experiments, and y_i is the value of a predictor variable of interest for gene i. Furthermore, $f(y_i)$ gives the deviation of gene i from its average expression based on the value of y_i and e_{ij} is the residual error. SVA takes this model and partitions the residual variance, e_{ij}, into:

$$x_{ij} = \mu_i + f(y_i) + \sum_{\ell=1}^{L} \gamma_{\ell i} g_{\ell j} + e^*_{ij} \tag{1.10}$$

Where $\sum_{\ell=1}^{L} \gamma_{\ell i} g_{\ell j}$ gives the summed effects of L unmodeled variables ($g_{\ell j}$) for each gene and e^*_{ij} gives the gene-specific noise in expression. SVA does not attempt to estimate what the unmodeled variables influencing expression are, but rather find a set of vectors (the surrogate variables) that

span the same space as $\boldsymbol{g}$:

$$x_{ij} = \mu_i + f(y_i) + \sum_{k=1}^{K} \lambda_{ki} h_{kj} + e^*_{ij} \tag{1.11}$$

Where each $\boldsymbol{h_k}$ is a surrogate variable and each $\boldsymbol{\lambda_k}$ gives the effects of each surrogate variable on gene expression. For our analyses, our predictor variable y_i was treatment type. To get a measure of expression where the effects of surrogate variables are removed, we then subtracted off the effects of surrogate variables from both sides of the above equation.

$$x_{ij} - \sum_{k=1}^{K} \lambda_{ki} h_{kj} = \mu_i + f(y_i) + e^*_{ij} \tag{1.12}$$

Where $x_{ij} - \sum_{k=1}^{K} \lambda_{ki} h_{kj}$ gives us our expression values accounting for the effects of surrogate variables. The net result here is a reduction in the amount of unexplained or seemingly stochastic variation in expression because sources of variation have been attributed to "surrogates" that span the same space as real batch variables. We also conducted principal component analysis in R before and after SVA to verify the removal of batch effects.

Results

Summary of tissue differentiation, treatment specificity, and selection in overall dataset

To understand how treatment specificity of gene expression affects evolutionary rates of proteins, we queried the Sequence Read Archive for all *A. thaliana* RNA-seq experiments published before May 2022. We then annotated these experiments with standardized tissue and treatment ontology terms, manually filtered the dataset, and then processed all RNA-seq runs with a standardized pipeline. The number of sequencing experiments associated with each combination of tissue and treatment labels is summarized in Tables S1. Overall, the most sampled tissue category was leaf (4,642 experiments) followed by root (3,348 experiments), whole plant (2,492 experiments), seed (1,866 experiments), shoot (1,106 experiments), then fruit and flower (266 experiments). The four most sampled treatment categories were control (5,701 experiments), cold air exposure (675 experiments), short day length (561 experiments), and short day length plus *Botrytis cinerea* exposure (407 experiments). Any sequencing runs that shared an SRA experiment ID were averaged to

produce individual gene expression values for each SRA experiment.

We first looked at the distribution of mapping rates across all RNA-seq runs. The median mapping rate was 72.39 % (Figure A1) and we excluded runs with a mapping rate < 1% from further analyses. We next performed a principal components analysis (PCA) on the expression matrix and observed strong differentiation between root and non-root tissues along PC2 (Figure 1.1). We also observed that nearly all genes had some degree of treatment specificity in their expression (Figures 1.2A, S3). Furthermore, only a small proportion of genes had strong signatures of selection based on dN/dS, π_N/π_S, DoS, or Tajima's D (Figure 1.2B-D, Figure A2). The treatment specificity of expression was lower on average in flower and fruit tissue compared to the other tissues (Figure A3). However, tissue specificity did not vary widely depending on the treatment condition (Figure A4).

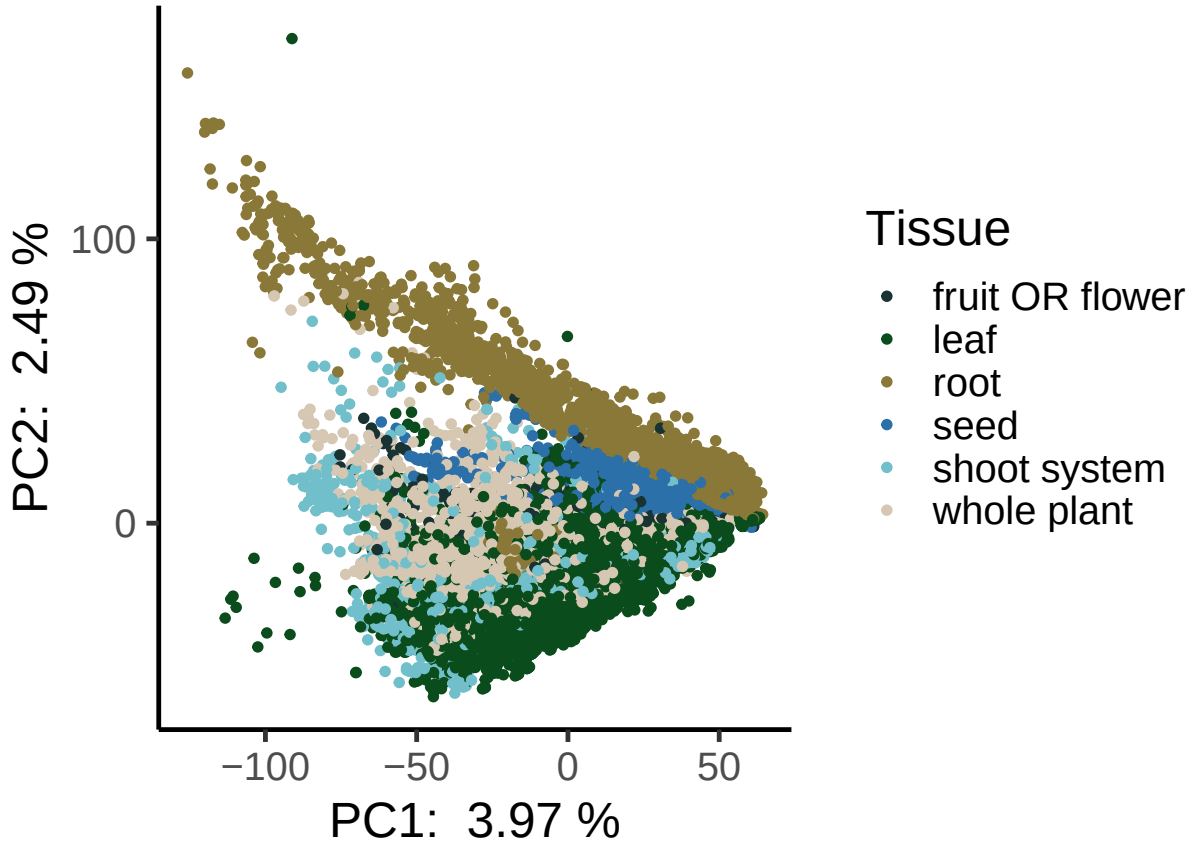

Figure 1.1: **Principal components analysis of all expression data.** Each point represents a different RNA-seq experiment and is colored by its associated tissue type. Experiments from all treatment conditions are included in this analysis. Plotting order was randomized to avoid overplotting.

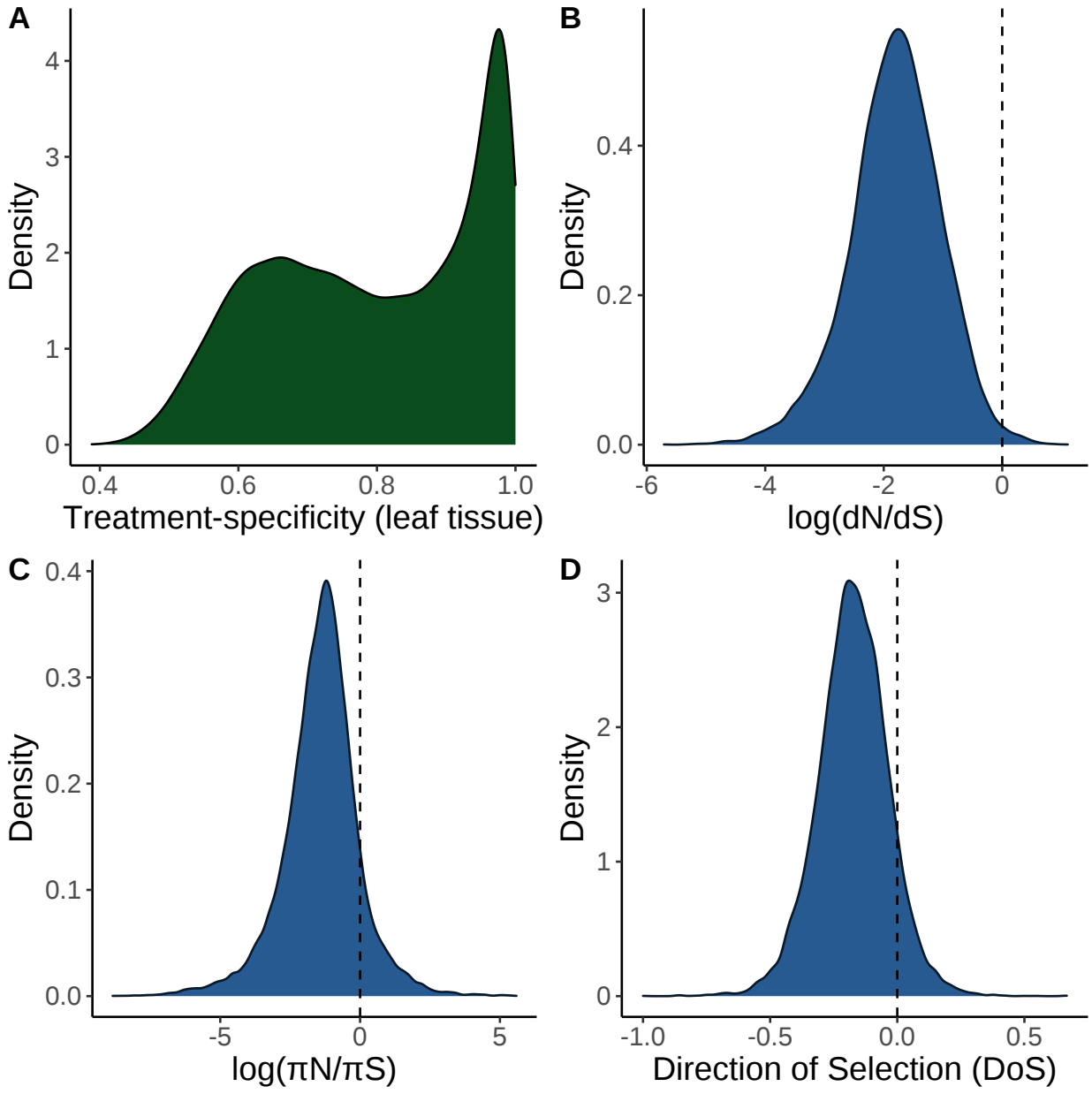

Figure 1.2: **Density plots of key variables measured in this study.** (**A**) Distribution of treatment specificity in leaf tissue expression across all genes included in this study. The area underneath the curve in a given interval of treatment specificity represents the proportion of genes in this study that fall within that range of treatment specificity. (**B**) Distribution of dN/dS across all genes included in this study. The area to the right of the dashed line represents the proportion of genes in this study with $dN/dS > 1$. (**C**) Distribution of π_N/π_S across all genes included in this study. The area to the right of the dashed line represents the proportion of genes in this study with $\pi_N/\pi_S > 1$. (**D**) Distribution of DoS across all genes in this study. Area to the right of the dashed line represents the propotion of genes with DoS > 0, which is interpreted as evidence of adaptive evolution.

Omitting samples with low expression disentangles expression level and specificity

Genes that are only expressed in one treatment or tissue will, by definition, have low mean expression across all environments or tissues (Wright et al., 2004). Thus, we sought a method of calculating expression level that was independent of treatment specificity. To better understand the relationship between average expression and treatment specificity, we calculated correlations between treatment-specificity and expression level while either including or excluding low expression values (TPM < 5) on our real RNA-seq dataset. We found that excluding low expression values decreased the correlation between average expression and treatment-specificity in leaf tissue samples (Figure 1.3) and other tissues (Figures A34 - A38) and replicated the result by simulating gene expression matrices (Figure A39). Thus, for all later partial correlation analyses (see next section) we quantified each gene's average expression after dropping experiments where the gene was not expressed (TPM < 5).

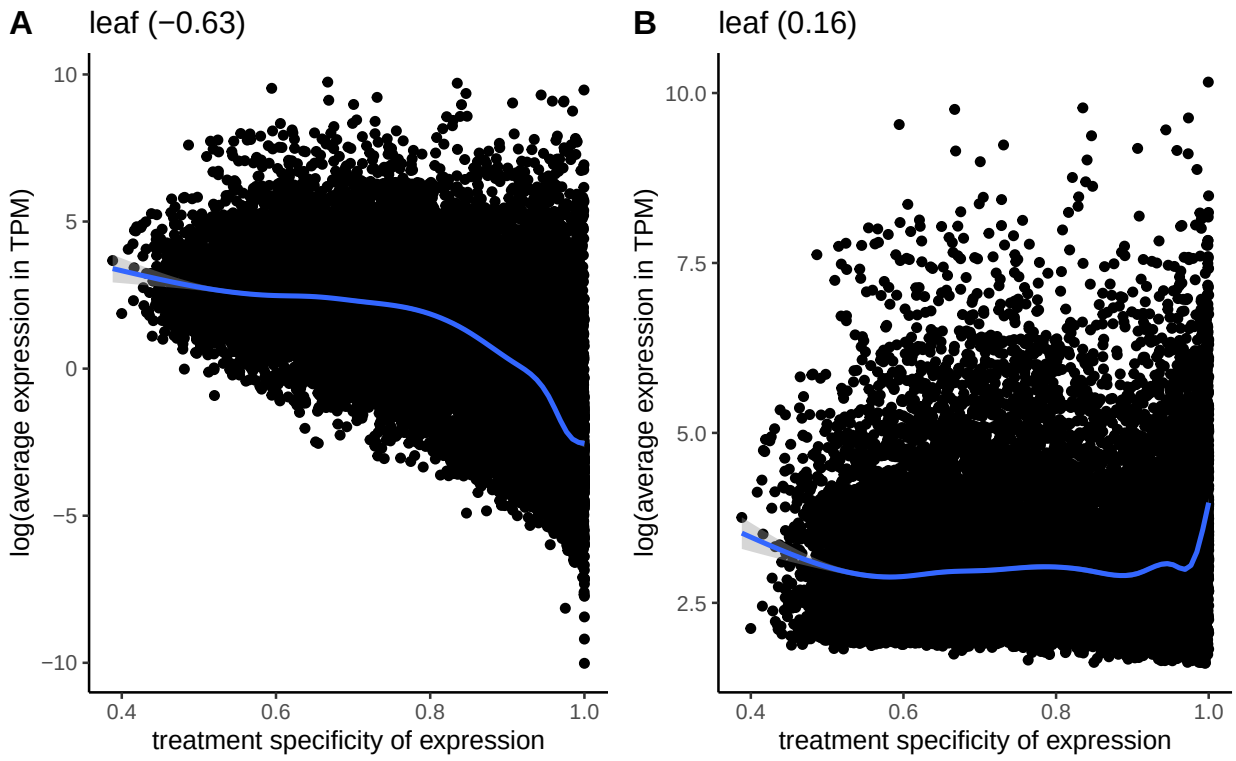

Figure 1.3: **Correlation between the average expression in transcripts per million (TPM) and treatment specificity of genes.** Figures show correlations when samples with low expression (< 5 TPM) are included (**A**) vs excluded (**B**). Expression level and treatment specificity were calculated using only data from leaf tissue samples. Line is a smoothing line with 95 % confidence intervals and values in parentheses give spearman correlation.

Treatment specificity correlates with levels of nonsynonymous diversity and divergence in genes

We next calculated partial correlations between treatment specificity and measures of selection after controlling for average expression, gene length, GC content, and tissue specificity in expression. These partial correlations were calculated separately for expression data on each tissue type and did not account for batch effects (see next section). Among leaf tissue samples, average expression had significant partial correlations with dN ($\rho = -0.19$, p-value = 2.1×10^{-122}) and π_N ($\rho = -0.17$, p-value = 2.8×10^{-175}) after controlling for other factors (Figures 1.4A,1.4B). Treatment specificity was more strongly correlated with dN ($\rho = 0.10$, p-value = 7.6×10^{-31}) and π_N ($\rho = 0.10$, p-value = 1.2×10^{-62}) than Tajima's D ($\rho = 0.01$, p-value = 3.1×10^{-7}) and DoS ($\rho = 0.04$, p-value = 2.3×10^{-06}, Figure 1.4C, 1.4D). Furthermore, the top 25% most

treatment-specific genes in leaf tissue for our dataset have average dN and π_N values nearly 2.5 times greater than the 25% least treatment-specific genes (dN = 0.025 vs 0.061; π_N = 0.0014 vs 0.0032). Meanwhile, the most and least treatment-specific genes have average Tajima's D values of are -0.44 and -0.43, respectively, and average DoS values of -0.19 and -0.14, respectively. The strongest partial correlation generally occurred between tissue specificity and treatment specificity (Spearman's $\rho = 0.53 - 0.60$, Figure 1.4). Gene family size had among the weakest partial correlations with π_N compared to other covariates, but strongly correlated with treatment specificity ($\rho = 0.12$, p-value = 6.3×10^{-84}, Figure 1.4B). All of these findings generally held when average expression and treatment specificity were calculated on data from other tissues (Table S2, Figures A6-A10).

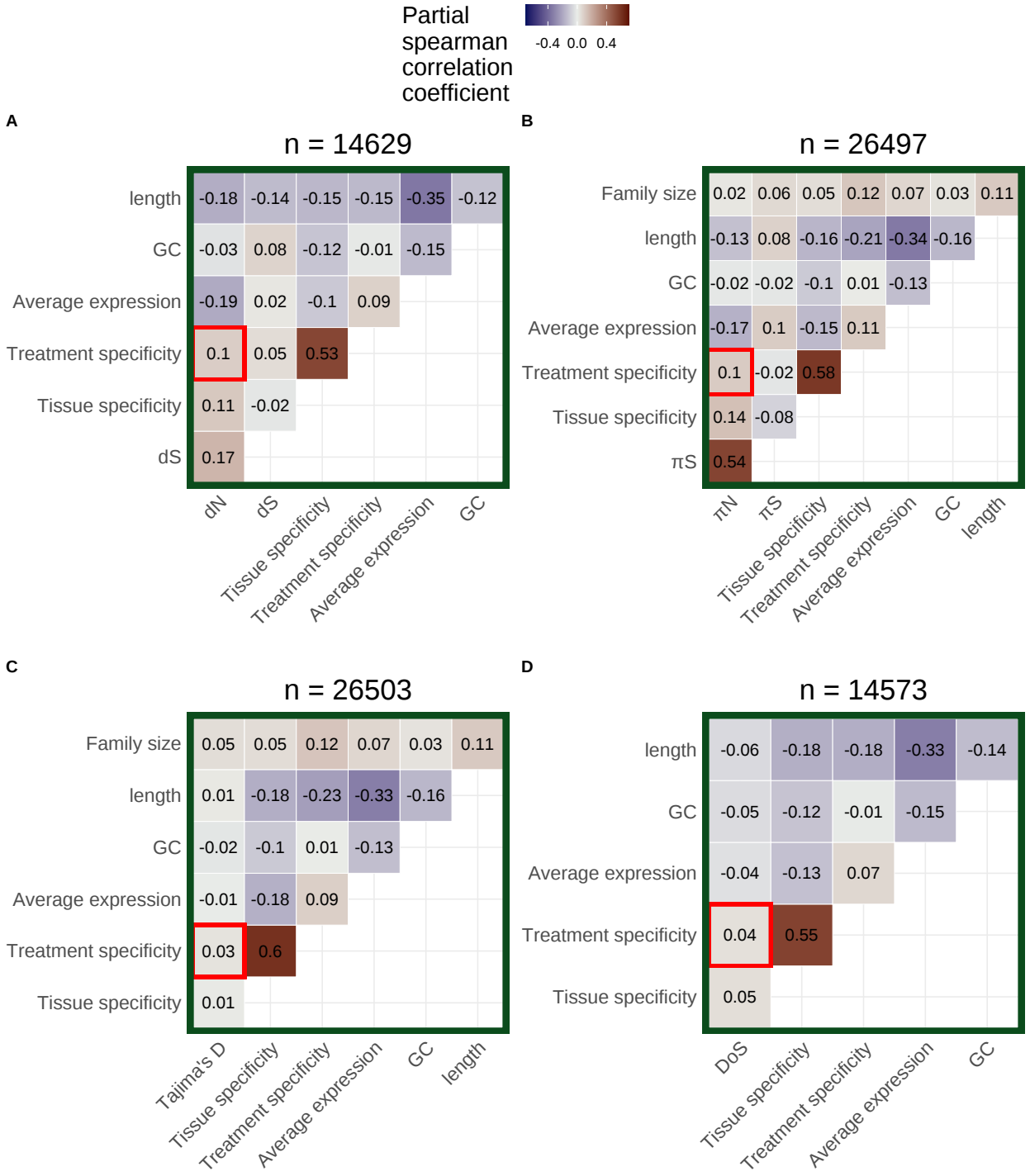

Figure 1.4: **Partial correlation analysis between specificity and selection.** Analyses include either (**A**) dN, (**B**) π_N, (**C**) Tajima's D, or (**D**) direction of selection (DoS) as a covariate. Average expression excludes values < 5 TPM and was calculated using only leaf tissue samples. Treatment specificity was also calculated using only leaf tissue samples. Tissue specificity was calculated using only control samples across all tissue categories. The number of genes included in each partial correlation analysis (n) is listed at the top of each heatmap.

Correlations between treatment specificity and nonsynonymous variation persist after controlling for batch effects and dataset imbalance

While combining gene expression data across multiple studies can increase the statistical power of an analysis, there are some potential concerns. First, if many tissue-treatment combinations are not sampled, the dataset will be unbalanced and the effects of tissue and treatment variation on expression could be confounded. Consistent with this expectation, there was a high correlation between tissue specificity and treatment specificity in our initial analyses (Figure 1.4, S6-S10). Furthermore, combining data from multiple laboratories could generate batch effects (Leek et al., 2010). To address the issues of imbalance and batch effects, we first subset our data to only include treatments where all tissue types were represented. This subset included the treatments of control, abscisic acid, continuous light, warm/hot air temperature, and cold air temperature. We then used SVA to correct for the influence of unknown batch effects on this data subset (Leek and Storey, 2007). After SVA, treatment specificity positively correlated with dN ($\rho = 0.10$, p-value = 1.6×10^{-32}) and π_N ($\rho = 0.07$, p-value = 1.5×10^{-23}) when average expression and treatment specificity were calculated on combined fruit and flower data (Figures A33). However, treatment specificity in other tissue types generally did not correlate with our measures of selection (Figures A28-S33, Table S4).

The inclusion of only five treatments in the above analysis could limit quantification of a gene's treatment specificity. Thus, in order to include data from a larger number of treatments, avoid dataset imbalance, and avoid batch effects, we split our expression matrix into six subsets by tissue category. We then further removed treatments that only had expression data from one study to avoid confounding treatment effects with study-specific batch effects. We applied SVA (Leek and Storey, 2007) to each of these tissue-specific subsets. After SVA, the expression profiles of most genes appear less treatment-specific (Figures A16-A21 panels A vs B). We also observed less separation in PCA space within treatment groups after SVA (for example, see Figures A16C and A16D). Average expression levels before SVA were generally correlated with expression levels after SVA (Figures A16-A21 panels A and B). In partial correlations on each SVA-corrected subset, treatment specificity significantly correlated with dN ($\rho = 0.13$, p-value = 6.9×10^{-50}) and π_N ($\rho = 0.16$, p-

value = 3.9×10^{-128}) but less strongly correlated with Tajima's D ($\rho = 0.04$, p-value = 6.6×10^{-10}) and DoS ($\rho = 0.05$, p-value = 2.0×10^{-8}) for the leaf tissue data subset (Table 1, Figures 1.5). These patterns were similar in other tissue types (Figures A11-A15, Table S3).

Pre/post-SVA	**Selection measure**	**Partial correlation: selection vs treatment-specificity**	**p-value**
Pre	dN	0.10	7.6×10^{-31}
Post	dN	0.13	6.9×10^{-50}
Pre	π_N	0.10	1.2×10^{-62}
Post	π_N	0.16	3.9×10^{-128}
Pre	Tajima's D	0.03	3.1×10^{-7}
Post	Tajima's D	0.04	6.6×10^{-10}
Pre	DoS	0.04	2.3×10^{-6}
Post	DoS	0.05	2.0×10^{-8}

Table 1: **Partial correlations between treatment-specificity and different measures of selection pre-SVA and post-SVA.** All correlation coefficients are spearman coefficients and are calculated only on leaf tissue samples. All p-values represent whether correlation coefficient significantly differs from 0.

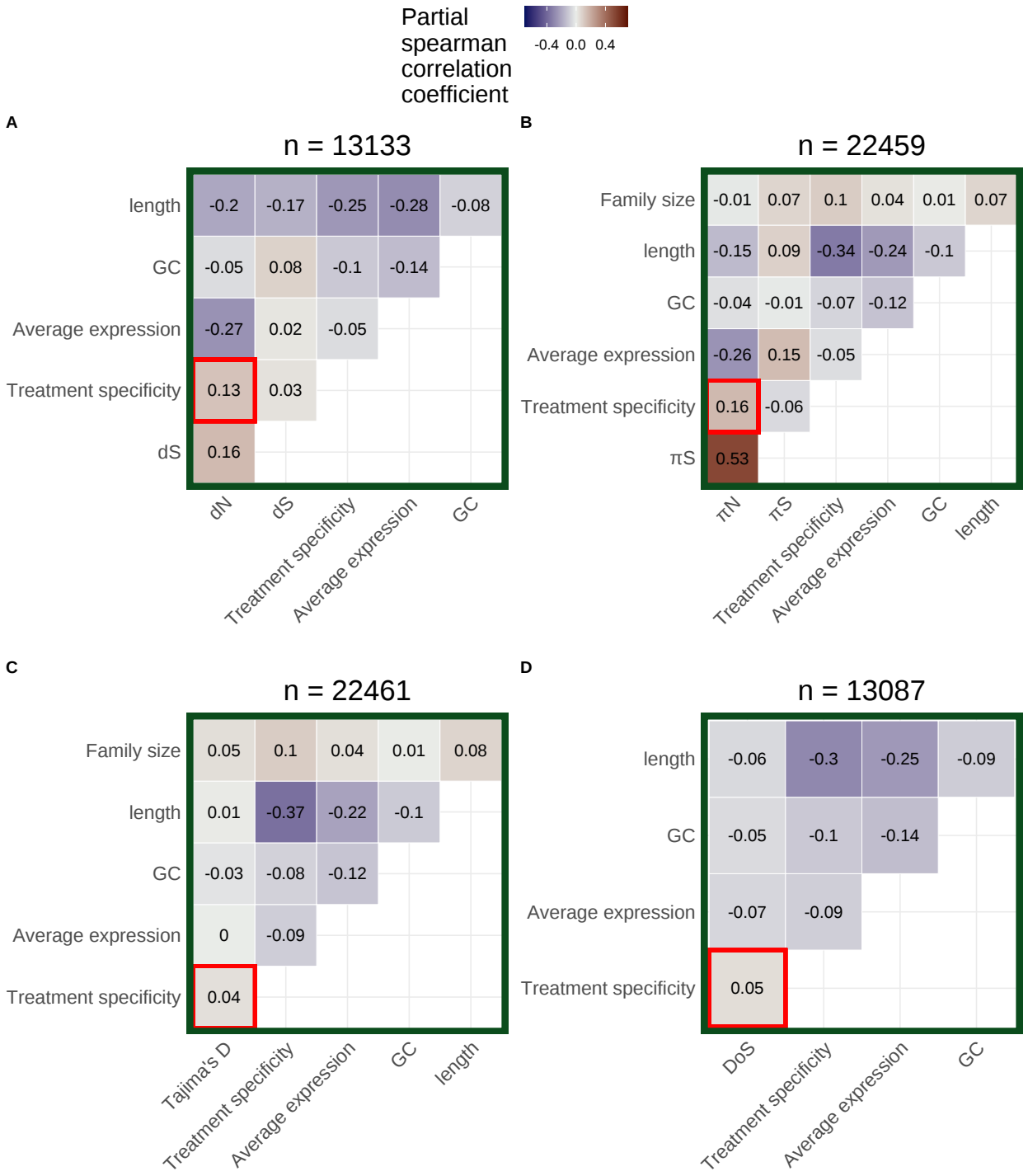

Figure 1.5: **Partial correlations between specificity and selection after SVA.** Analyses include (**A**) dN, (**B**) π_N, (**C**) Tajima's D, and (**D**) direction of selection (DoS) based on leaf tissue data subset after applying SVA. Data was further subset to include only treatment groups with data from more than one study before applying SVA. Average expression calculation excludes values < 5 TPM. The number of genes included in each partial correlation analysis (n) is listed at the top of each heatmap.

Discussion

Our main finding is that genes with more treatment-specific expression patterns are, on average, under weaker selective constraint in *A. thaliana*. This is evident by treatment-specific genes generally having higher values of π_N and dN, but not higher values of Tajima's D and DoS, compared to genes with more constitutive expression (Figures 1.4,1.5). Our result does not refute the possibility of strong positive selection on treatment-specific genes, as is the case for nucleotide binding site leucine rich repeat proteins (NBS-LRRs) in *A. thaliana* (Mondragón-Palomino et al., 2002). Rather, treatment-specific genes are simply under weaker selection on average compared to less treatment-specific genes. Altogether, this pattern is consistent with the hypothesis that a trade-off between the strength of selection and the treatment specificity of expression helps maintain variation in plasticity for *A. thaliana* (Snell-Rood et al., 2010; Van Dyken and Wade, 2010).

There are a few ways to think about the biological relevance of the correlations of treatment specificity with π_N and dN. First, the magnitude of treatment specificity's correlation with π_N and dN was generally half the magnitude of average expression's correlation with π_N and dN and similar to tissue specificity's correlation with π_N and dN. Both tissue specificity and average expression are thought to be important determinants of protein evolution (Bush et al., 2015; Wu et al., 2022), suggesting the comparable effects of treatment specificity may be important too. Second, the effect of treatment specificity on π_N and dN persisted even after simultaneously controlling for expression level, tissue specificity, gene length, GC content, and batch effects. Finally, the top 25% most treatment-specific genes in our dataset have average dN and π_N values nearly 2.5 times greater than the 25% least treatment-specific genes (dN = 0.025 vs 0.061; π_N = 0.0014 vs 0.0032), but relatively similar Tajima's D and DoS values (Tajima's D = -0.44 vs -0.43; DoS = -0.19 vs -0.14). These observations together suggest that treatment specificity is an important determinant of protein evolution.

This study disentangles several processes that were often difficult to resolve in previous research. First, many previous studies focus mainly on explaining trends in dN/dS (Slotte et al., 2011; Gaut et al., 2011; Bush et al., 2015), but both relaxed negative selection and increased posi-

tive selection can lead to increases in dN/dS. To tease apart these two processes, we additionally investigated treatment specificity's relationship with Tajima's D and DoS. Treatment specificity's weaker correlation with Tajima's D and DoS, compared to dN and π_N, suggests that relaxed negative selection plays a larger role than increased positive selection in explaining the high evolutionary rates of treatment-specific genes. Furthermore, measures of expression specificity are often highly correlated with expression level (Slotte et al., 2011; Alvarez-Ponce and Fares, 2012; Huang, 2022). When calculating a gene's expression level, we only included samples where said gene was expressed (TPM > 5) to get an estimate of expression level that was still correlated with dN and π_N, but was independent of expression specificity, allowing us to better disentangle these factors. Finally, previous studies have struggled to partition the factors that influence selection on genes in the presence of predictor variables with considerable error, such as expression level (Drummond et al., 2006; Plotkin and Fraser, 2007; Yang and Gaut, 2011). Error in expression measurements can often be attributed to unmeasured differences between RNA-sequencing experiments (Leek et al., 2010) and we accounted for these differences using SVA (Leek and Storey, 2007). Even after SVA, treatment specificity was strongly correlated with dN and π_N (Figures 1.5A-B), suggesting our results are not an artifact of errors in expression measurement or combining expression data across many studies.

Surprisingly, nearly all genes in *A. thaliana* have some degree of treatment specificity in their expression (Figures 1.2A, A3), reflecting results of previous studies on tissue specificity (Eisenberg and Levanon, 2003). The high prevalence of treatment specificity in our dataset is partly explained by batch effects because SVA significantly lowered the apparent treatment specificity of most genes (Figures A16B-A21B) and reduced within-treatment differentiation in PCA space (for example, see Figures A16C and A16D). This reduction in treatment-specificity likely happened because batch effects can include unrecorded between-treatment differences (e.g. the humidity of the growth chamber, light intensity, watering schedule, etc.). Controlling for these unrecorded between-treatment differences thus causes the expression of genes to be less treatment-specific. However, even after batch correction most genes still showed some degree of treatment specificity

(Figures A16B-A21B), suggesting it is rare for a gene to be expressed at the same level across many environments.

We also observed that genes with higher treatment specificity generally belonged to larger gene families. We expected gene family size to correlate with selection because singleton and duplicated genes often evolve at different rates (Jordan et al., 2004; Davis and Petrov, 2004). Theory also suggests that gene duplication leads to relaxation of selection on duplicates, allowing for neo- and sub-functionalization (Lynch and Conery, 2000; Aagaard et al., 2006). We could not investigate how gene family size correlates with dN or DoS because measuring these quantities requires identifying substitutions between orthologous genes. Thus, dN and DoS can only be reliably measured for 1:1 orthologs between *A. thaliana* and *A. lyrata*. However, π_N and Tajima's D can be calculated for genes in larger families and we did observe persistent correlations between family size and Tajima's D (For Figure 1.5C: $\rho = 0.05$, p-value = 3.1×10^{-12}; also see Figures A6C-A15C, A28C-A33C). Altogether, these correlations suggest that processes of gene duplication, neofunctionalization, and subfunctionalization could be connected to evolving some degree of treatment specificity.

Gene length was generally the second most correlated factor with dN and π_N in our study, just behind average expression. This is consistent with previous work suggesting that longer proteins require more energy to synthesize and are thus under stronger selective constraints (Urrutia and Hurst, 2001; Castillo-Davis et al., 2002; Eisenberg and Levanon, 2003; Urrutia and Hurst, 2003). However, while some previous studies in *A. thaliana* observe this same trend (Bush et al., 2015), others do not (Slotte et al., 2011). This discrepancy could be due to differences in how gene length is defined between studies. In this study, each gene's length included coding sequence as well as introns and untranslated regions, whereas other studies break down gene length into individual features (Bush et al., 2015). The goal of this study was not to understand differences in evolution between different gene features, so we included all gene features in our estimate of gene length. However, introns and untranslated regions experience different evolutionary patterns than coding sequences; for example, highly expressed genes being under selection for shorter introns (Castillo-

Davis et al., 2002; Eisenberg and Levanon, 2003). Therefore, future studies must clearly define even seemingly simple features like gene length to ensure that results are comparable across studies.

Although we focused on testing the idea that treatment specificity is responsible for relaxed negative selection in some genes, it is also possible that relaxed selection caused the evolution of treatment specificity. There is some evidence that relaxation of selection occurs before the evolution of expression specificity (Hunt et al., 2011) and may better explain cases of neo- and subfunctionalization (Lynch and Conery, 2000; Aagaard et al., 2006). Future experiments that look at the evolution of treatment specificity and sequence evolution across a broader phylogenetic scale may be helpful for determining the order of these processes.

In summary, this study investigates a trade-off between the treatment-specific expression of a gene and the strength of selection said gene experiences, which is hypothesized to limit plasticity evolution. Consistent with this hypothesis, genes in *A. thaliana* with more treatment-specific expression are under weaker selection compared to more evenly expressed genes. While we find that this trade-off exists, we could not dissect the direction of causality in the trade-off or determine how much this trade-off constrains plasticity evolution relative to other processes. However, these are exciting areas of future research. Future studies should ideally generate fully balanced datasets on gene expression acquired across natural environmental gradients. Taking these steps will contribute to a comprehensive understanding of the constraints on plasticity and protein evolution.

Data availability

Supplemental figures are available in Appendix A. All code for our bioinformatic workflows, data analysis, and figure creation can be found here: https://github.com/milesroberts-123/arabidopsis-conditional-expression. All supplementary tables are available online at: https://doi.org/10.1093/genetics/iyad074. The tissue type and treatment annotations for RNA-seq runs in our study can be found in Table S5. Genomic references, mapping rates, and a table of expression specificity; nucleotide diversity; and substitution rate values estimated for all *A. thaliana* genes included in this manuscript's analyses is available at: https://doi.org/10.5061/dryad.xd2547dnd. The genome assembly and annotation used in this study was originally downloaded from Phytozome: https:

//phytozome-next.jgi.doe.gov/.

CHAPTER 2: k-MER-BASED DIVERSITY SCALES WITH POPULATION SIZE PROXIES MORE THAN NUCLEOTIDE DIVERSITY IN A META-ANALYSIS OF 98 PLANT SPECIES[2]

Abstract

A key prediction of neutral theory 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. As a first step to testing this idea, we calculated nucleotide diversity (π) and k-mer-based metrics of genetic diversity across 112 plant species, amounting to over 205 terabases of DNA sequencing data from 27,488 individuals. After excluding 14 species with low coverage or no variant sites called, we compared how different diversity 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 are consistent with the possibility that variation not captured by common SNP-based analyses explains part of Lewontin's paradox in plants, but larger scale pangenomic studies are needed to definitively address this question.

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, π) than expected (Frankham, 2012; Corbett-Detig et al., 2015; Buffalo, 2021). If we assume that

[2]This chapter is, as of writing, accepted for publication at *Evolution Letters*. The preprint version is currently available at the following DOI under a CC-BY License: https://doi.org/10.1101/2024.05.17.594778

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 μ is the mutation rate per base pair per generation. Mutations rates for SNPs and small indels vary relatively little across species (Cagan et al., 2022; Bergeron et al., 2023; 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 (i.e. demography, 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) and many studies have tested this hypothesis(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). These previous investigations often conclude that linked selection contributes to Lewontin's paradox, but not all report significant results (e.g. Schmid et al. (2005); Roselius et al. (2005); Flowers et al. (2012);

Wang et al. (2016)). It's been argued that studies focused on plant species especially tend to find weaker evidence for linked selection (Slotte, 2014). 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 too (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.

Two potential, but rarely explored, contributors to Lewontin's paradox are that (1) current methods for estimating genetic diversity systematically underestimate the true levels of genetic diversity in most populations and (2) changes or corrections to the calculation of diversity could yield stronger correlations with population size. 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. 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)). Sometimes alignment errors in regions of non-reference structural variation will also create false positive SNP calls, which will also affect diversity estimates (e.g. 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 deflated diversity estimates due to errors in the handling of missing genotype calls (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 and applying correct calculations of nucleotide diversity 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); Roberts et al. (2024)). 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 (Rahman et al., 2018; Voichek and Weigel, 2020; Wiersma et al., 2024).

We revisited Lewontin's paradox in plants using k-mer-based measures of genetic difference and corrected π calculations, aiming to test whether the inclusion of non-reference variation or modifications to diversity calculations could increase the scaling between population size proxies and diversity. 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.

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 8 species (*Aegilops tauschii*, *Hordeum vulgare*, *Lens culinaris*, *Pisum sativum*, *Thinopyrum intermedium*, *Trifolium pratense*, *Triticum aestivum*, *Triticum turgidum*) that had at least one chromosome longer than 2^{29} bp (about 537 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 correctly accounting for demography is a general limitation of Lewontin's paradox studies, as we can only estimate proxies of N_e for contemporary populations (Corbett-Detig et al., 2015; Buffalo, 2021). We include cultivation status in our downstream modeling to help account for systematic differences between cultivated and wild species (see **Statistical analysis**) and also repeat our analysis with multiple

types of population size proxies.

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 7.5×10^{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 7.5×10^{12} bp, we randomly dropped one individual at a time until 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.

SNP calling pipeline

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 and included all nuclear genome scaffolds, as identified by degenotate (v1.1.3, Mirchandani et al. (2024)), to focus solely on neutral diversity. Notably, we chose to not call SNPs in non-coding sequences in plants because plant genomes tend to be highly repetitive and are largely composed of transposable elements, making short read mapping difficult For example, even in the small-genome plant species *Capsella grandiflora*, SNPs can only be confidently called in 10% of intergenic sites (Williamson et al., 2014). 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. In short, these filters reflect the standard population genetics practices of removing non-biallelic SNPs and SNPs with a lot of missing genotype calls. The exact filter cutoffs also come from GATK Best Practices for hard filtering (Caetano-Anolles, 2023) that is widely used (e.g. (Yu et al., 2024; Ramirez-Ramirez et al., 2024; Glasenapp and Pogson, 2024; Ritter et al., 2024)). 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. We chose a minimum read depth of 5 to maintain comparability with our k-mer dissimilarity calculations which required minimum k-mer counts of 5 (see **k-mer counting**), but we also filtered out calls with very high read depths because these could represent read mismapping due to repetitive or paralogous regions. 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, 0.023 % of the genome)

and "scaffold04645" (2,838 bp, 0.000136 % of the genome), 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):

$$\pi_i = \left(\frac{n_i}{n_i - 1}\right)\left(1 - \sum_{j=1}^{a_i} p_{ij}^2\right) \tag{2.1}$$

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 jth 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 π_i across all M sites in the genome (including both variant and invariant sites):

$$\bar{\pi} = \frac{\sum_{i=1}^{M} \pi_i}{M} \tag{2.2}$$

Importantly, this calculation of nucleotide diversity does not assume that all missing genotype calls are homozygous for the reference genotype, meaning that our estimates of π will not be downwardly biased by missing data (Korunes and Samuk, 2021).

k-mer counting pipeline

For each species, we counted 30-mers in the coding sequences of their respective reference genome using KMC (v3.2.1, Kokot et al. (2017)). 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; Roberts et al., 2024). Next, we removed any 30-mers that matched 30-mers found in each species' 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 under selective constraints (Margulies et al., 2003; Guo et al., 2007). Thus, similar to the common analysis of 4-fold degenerate sites, our k-mer 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 found that non-coding regions and 4-fold degenerate sites have very similar levels of diversity (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. Our analysis choices here are thus in line with previous studies; however, larger scale pangenomic analyses will be helpful in relaxing this assumption.

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; Roberts et al., 2024). Thus, we randomly downsampled each species' 30-mer list to 10 million 30-mers with a frequency

≥ 5 in at least one sample in the species' 30-mer list. We chose to downsample to 10 million 30-mers to decrease disk space burden of storing k-mer counts and because several previous studies show that subsets of only 1 million k-mers or less can reliably estimate genetic dissimilarity in many systems (Ondov et al., 2016; Benoit et al., 2020; VanWallendael and Alvarez, 2022). The reason we also included a frequency cut-off of 5 is to omit k-mers containing sequencing errors, which predominantly occur as very low-frequency k-mers (Ranallo-Benavidez et al., 2020). We chose a frequency cutoff of 5 to to include more k-mers in some of our lower coverage datasets, but frequency cut-offs anywhere from 2-10 are used in plants (Voichek and Weigel, 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:

$$J_D(X, Y) = 1 - \frac{X \cap Y}{X \cup Y} \tag{2.3}$$

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 ≥ 5. 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:

$$B_D(X, Y) = 1 - \frac{2\sum_i^k \min(m_i^*(X), m_i^*(Y))}{\sum_i^k m_i^*(X) + m_i^*(Y)} \tag{2.4}$$

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):

$$m_i^*(X) = \frac{m_i(X)}{\sum_i m_i(X)} \tag{2.5}$$

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 proxies

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:

$$N = RD \tag{2.6}$$

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 Appendix C. 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 2.6 gives our main proxy for population size:

$$N \propto \frac{R}{h^2} \tag{2.7}$$

In our subsequent analyses, we refer to Equation 2.7 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 2.7 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 used the caper R package (Orme et al., 2018) to implement a statistical approach similar to Whitney et al. (2010). We performed partial phylogenetic least squares 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). We did not include sequencing coverage as a covariate at

this step because we would later exclude species with low coverage data from our analyses (see **Results**) and the relationship between coverage and diversity saturates at higher levels of coverage (Figures S3A, S3D, and S3G). We also did not include number of individuals as a covariate because this did not correlate with any of our diversity measures (Figures S3C, S3F, and S3I). 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 2.7 or it's components (range size and plant height), 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:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.8}$$

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 Ω. We first performed Cholesky decomposition on Ω to get matrix $\boldsymbol{C}$ such that:

$$\Omega = \boldsymbol{C}\boldsymbol{C}^T \tag{2.9}$$

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

$$\boldsymbol{C}^{-1}\boldsymbol{y} = \boldsymbol{C}^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{C}^{-1}\boldsymbol{\epsilon} \tag{2.10}$$

Which we will rewrite as:

$$\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^* \tag{2.11}$$

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:

$$\boldsymbol{y}^* = \beta_0\boldsymbol{x}_0^* + \beta_1\boldsymbol{x}_1^* + \beta_2\boldsymbol{x}_2^* + ... + \beta_{n-1}\boldsymbol{x}_{n-1}^* + \boldsymbol{\epsilon}^* \tag{2.12}$$

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:

$$\boldsymbol{y}^* - \sum_{i \neq k} \beta_i x_i^* = \beta_k \boldsymbol{x}_k^* + \boldsymbol{\epsilon}^* \tag{2.13}$$

We then plotted the values of $\boldsymbol{x}_k^*$ against $\boldsymbol{y}^* - \sum_{i \neq k} \beta_i x_i^*$, interpreting the slope (β_k) 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

There were 112 species in our initial dataset, each with estimates of population size proxies, nucleotide diversity, and k-mer diversity (Figure 2.1). 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 = 1.4×10^{-5}, Figure C1C). However, cultivation status was independent of both life cycle habit ($\chi^2 = 4.07 \times 10^{-31}$, p = 1, Figure C1A) and mating system (χ^2 = 0.53, p = 0.47, Figure C1B). 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 (Figure C2). 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 (Figure C2). There were no missing values for any of the variables investigated in this study, but there were three species with zero variant sites called (*Capsicum annuum*, *Heliosperma pusillum*, and *Papaver somniferum*) because they had very low coverage sequencing datasets (average coverages per individual of 0.035x, 0.169x, 0.043x, respectively, Table S1). We omitted these species 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. Mean coverage correlated with both nucleotide diversity (ρ = 0.33, p = 0.00033, Figure C3A) and k-mer diversity (Jaccard: ρ=

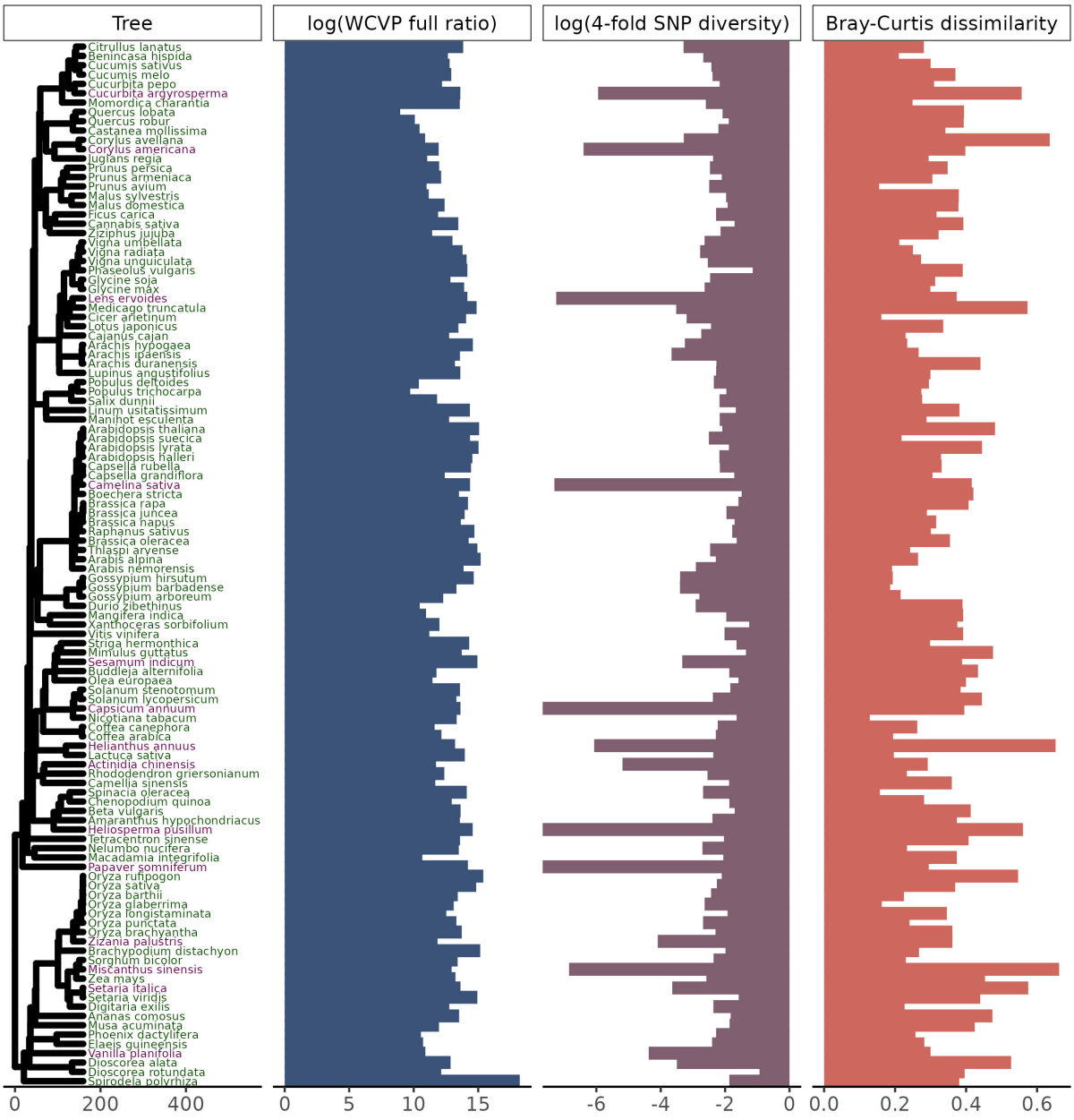

Figure 2.1: **Our study includes 112 plant species across a wide range of population sizes and diversity levels.** Species labeled in purple were considered outliers and omitted from downstream analyses (see Figure 2.2), but species labeled in green were retained. The phylogenetic tree is scaled in millions 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, see Equation 2.7). Nucleotide diversity is genome-wide average diversity at four-fold degenerate sites, log-transformed (base 10, see Equation 2.2). *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 (see Equation 2.4).

-0.53, p = 2.6×10^{-9}, Figure C3D; Bray-Curtis: ρ = -0.34, p = 0.00021, Figure C3G). Coefficient of variation in bp sequenced correlated strongly with k-mer diversity (Jaccard: ρ = 0.36, p = 0.00013, Figure C3E; Bray-Curtis: ρ = 0.42, p = 4.7×10^{-6}, Figure C3H) but not nucleotide diversity (ρ = -0.088, p = 0.36, Figure C3B). The number of individuals sequenced did not correlate with either nucleotide diversity or k-mer diversity (Figure C3C, C3F, C3I).

Given that mean coverage strongly correlated with both k-mer dissimilarity and nucleotide diversity, we decided to investigate this relationship further. We would expect that low coverage sequencing would produce artificially low nucleotide diversity and artificially high k-mer diversity. This is because our SNP-calling pipeline is generally tuned for high-coverage datasets (see **SNP calling pipelne**), so low coverage datasets produce few confident SNP calls. Conversely, low coverage sequencing only samples a small proportion of the total k-mer space in a set of reads. Thus, two independent low coverage samples are unlikely to share many k-mers in common and will have inflated k-mer diversity. Overall, we observed that species with very low mean coverage had both low nucleotide and high k-mer diversity (Figure 2.2A, Figure C4A). In contrast, there was no clear mapping of these abnormal diversity values with coefficient of variation in base pairs sequenced (Figure C5) or the number of individuals sequenced (Figure C6). Based on these results, we removed 10 species from our dataset with mean coverage per individual $\leq$ 0.5x. This included three species (*Capsicum annuum*, *Heliosperma pusillum*, and *Papaver somniferum*) with zero variant sites called because they had very low coverage (0.035x, 0.169x, 0.043x, respectively, Table S1). We also dropped 4 additional species with fewer than 1000 variant sites called: *Camelina sativa* , *Sesamum indicum*, *Setaria italica*, *Actinidia chinensis*, the first of which were just slightly above our coverage cutoff (0.63x, 0.92x, 0.63x, respecitively, Table S1). These filtering steps removed many of the species with abnormally low diversity and high dissimilarity values excluding (Figure 2.2B). 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

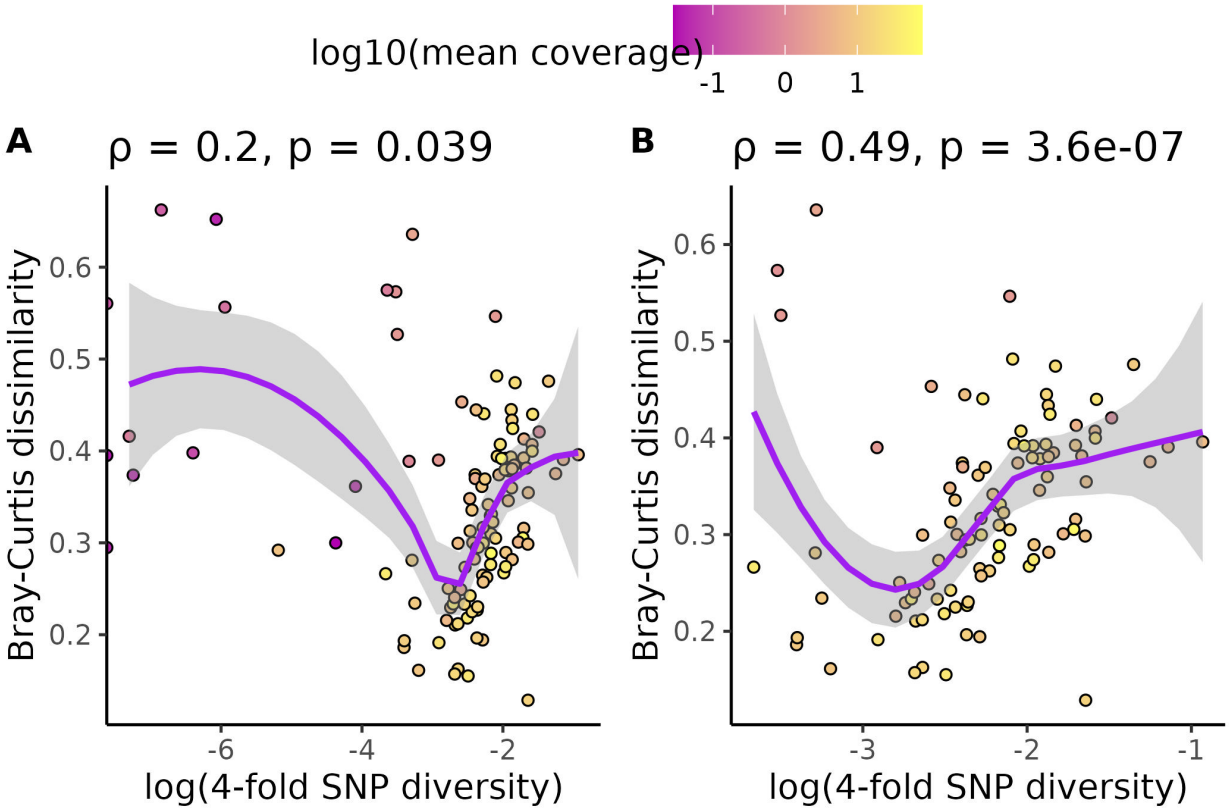

Figure 2.2: **Low coverage and low numbers of variant calls explains species with abnormally low diversity.** (A) shows the relationship between k-mer diversity and nucleotide diversity without omitting species with $\leq$ 0.5x coverage or $\leq$ 1000 SNP calls. (B) shows the same relationship, except 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 loess smoothing lines with 95% confidence intervals shaded in gray. Values across the top of each plot are Spearman correlation coefficients (ρ) and p-values that test whether each correlation coefficient differs from zero.

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 (ρ = 0.31, p = 0.00096, Figure C7A) or excluded (ρ = 0.48, p = 7.3×10^{-8}, Figure C7B). The omission of invaded ranges lowered the range size of several plant species based on WCVP range maps (Figure C7C) but had less effect on ranges estimated from GBIF occurrence data (Figure C7D).

We then calculated the ratio of range size to squared plant height (Equation 2.7) 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 (Figure 2.2), 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 8.9×10^8 to 1.7×10^{18}) and 13 (from 8.6×10^5 to 1.5×10^{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 $< 2.2 \times 10^{-16}$, Figure C8).

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 (Figure C9). 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, Figure C10A). 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 =

2.2×10^{-9}, Figure C10B; Bray-Curtis dissimilarity: β =0.79, SE = 0.11, p = 7.3×10^{-11}, Figure C10C). We observed the same qualitative trend when we included both native and invaded ranges in the range size-squared height ratio (Figure C10D-F), or used the GBIF-based range estimates instead of WCVP-based estimates (Figure C11). 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 Figure C10B-C), but models where Bray-Curtis dissimilarity was the response variable generally had lower adjusted R^2 (e.g. R^2 = 0.51 vs 0.41, models 2 and 3 in Table S4).

We also analyzed range size and plant height separately as population size proxies (Figure C12-C14). Overall, WCVP-estimated range size significantly affected nucleotide diversity ($\beta = 0.29$, SE = 0.072, p = 0.00011, Figure C12A) and k-mer diversity (Jaccard: $\beta = 0.92$, SE = 0.13, p = 9.9×10^{-11}, Figure C12B; Bray-Curtis: $\beta = 1.2$, SE = 0.13,p = 3.2×10^{-14}, Figure C12C), and this trend held when we estimated range size from GBIF occurrences (Figure C13A-C) or included invaded range area (Figure C12D-F and Figure C13D-F). On the other hand, nucleotide diversity did not scale with plant height (β = 0.13, SE = 0.19, p = 0.5, Figure C14A), but k-mer diversity marginally scaled downward with plant height (Jaccard: $\beta = -0.78$, SE =0.38, p = 0.046, Figure 14B; Bray-Curtis: $\beta = -0.77$, SE = 0.44, p = 0.088, Figure C14C).

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 (β = 0.035, SE = 0.063, p = 0.58, Figure 2.3A), but k-mer diversity did (Jaccard: $\beta = 0.54$, SE = 0.093, p = 8.8×10^{-8}, Figure C15; Bray-Curtis: β = 0.7, SE = 0.098, p = 2.2×10^{-10}, Figure 2.3B). Again, we got qualitatively similar results when we excluded invaded ranges in our range size estimates (Figure C16), used GBIF occurrences to estimate range size-squared height ratio (Figure C17) or used WCVP range size as the population size proxy (Figure C18). However, GBIF range size by itself did not scale with Jaccard dissimilarity (Figure C19B, C19E). Increased plant height associated with decreased k-mer diversity, but had no significant relationship with nucleotide diversity (Figure C20).

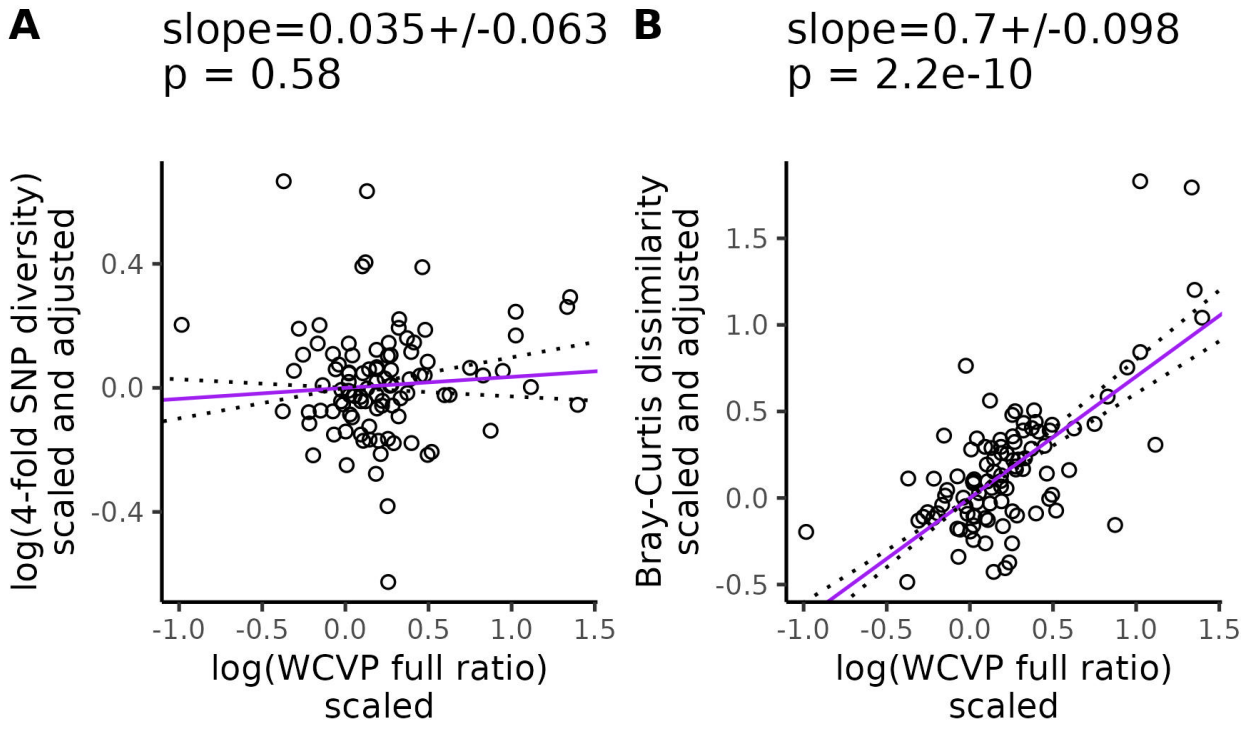

Figure 2.3: k**-mer diversity scales with population size proxies after controlling for genome size, life cycle habit, mating system, and cultivation status.** Purple lines are partial phylogenetic regression lines between diversity levels and the population size proxy. The y-axis (diversity) is scaled and adjusted according to Equation 2.13: scaling diversity levels to a standard normal distribution, followed by correcting for phylogenetic relatedness and 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 $\pm$ one standard error and p-values testing whether the slopes differ from zero. Dotted lines show the partial regression slope $\pm$ one standard error. 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.

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 = 8.4×10^{-14}, Figure C21; Bray-Curtis: $\beta = -4.2$, SE = 0.45, p = 4.5×10^{-15}, Figure 2.4B) and nucleotide diversity ($\beta - 1.8$, SE = 0.29, p = 1.4×10^{-8}, Figure 2.4A), 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 (Figure C22), or if our population size proxy was based on GBIF occurrences (Figure C23), or we used range size or plant height individually to control for population size variation (Figure C24-C26). Across all of these analyses, the partial regression relationship between genome size and diversity was always significantly negative.

Discussion

Our goal was to investigate whether genomic approaches that can capture more genetic variation than reference-based methods can improve the scaling between population size proxies and diversity. This was motivated by literature suggesting suggesting SNPs called against a single reference provide an incomplete picture of genome-wide polymorphism (Schmidt et al., 2021; Van-Wallendael and Alvarez, 2022; Jaegle et al., 2023; Sopniewski and Catullo, 2024). In total, we processed >205 terabases of publicly available sequencing data from the SRA over approximately

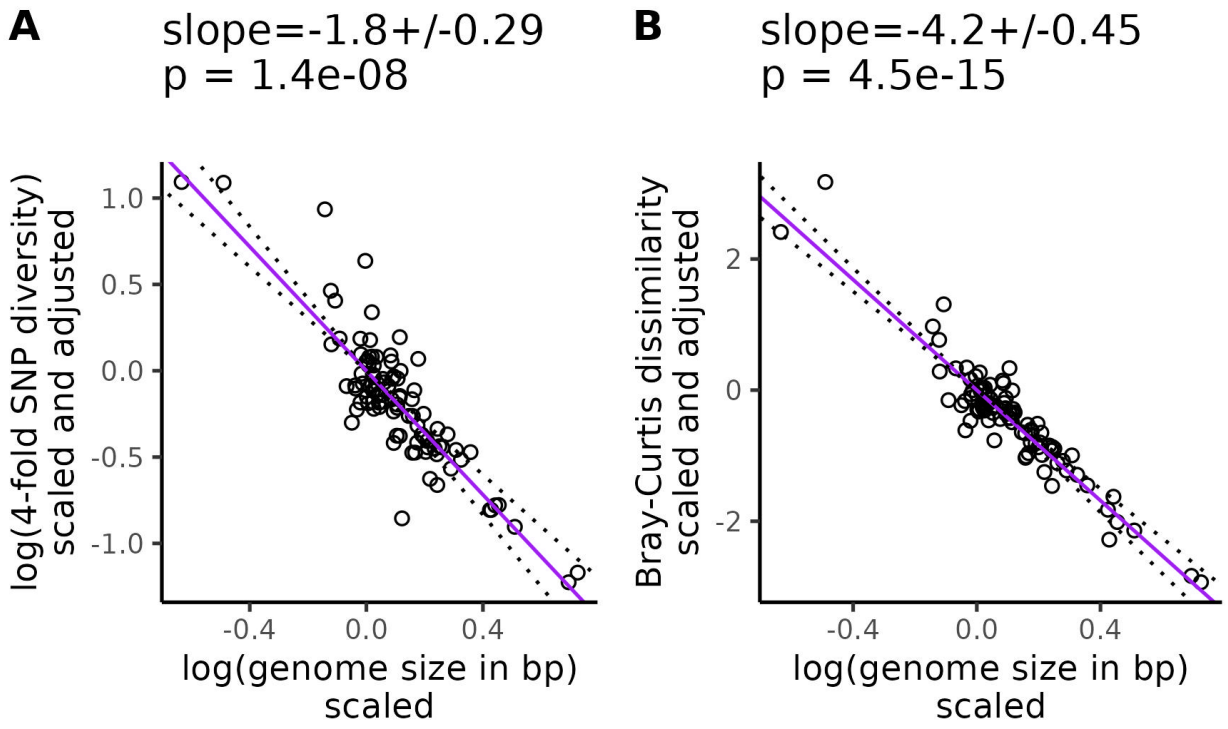

Figure 2.4: k**-mer diversity is more sensitive to genome size variation than nucleotide diversity.** Purple lines are partial phylogenetic regression lines between diversity levels and genome size (see Equation 2.13) after scaling diversity levels to a standard normal distribution (mean = 0, variance = 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 $\pm$ one standard error and p-values testing whether the slopes differ from zero. Dotted lines show the partial regression slope $\pm$ one standard error.

12 months of wall time, split between a maximum of 512 cores and 50 TB of disk space. After careful accounting for potential technical and phylogenetic confounding, the standardized slope between k-mer-based diversity and the range size-squared height ratio was up to 20 times larger than the same standardized slope for nucleotide diversity ($\beta = 0.035$ vs 0.7, Figure 2.3). We observed similar results across the two different measures of range size (Figure C17) and k-mer diversity (Figure C15). We also observed that k-mer-based diversity is more sensitive to variation in genome size compared to nucleotide diversity (Figure 2.4). Overall, these results are consistent with the possibility that diversity missed by reference-based analyses or changes in the diversity calculation itself can partly explain the weak scaling between population size and diversity. However, a larger pangenome scale analysis will be required to fully understand the genetic variants underlying this result.

Extending our findings to Lewontin's paradox directly is complicated by a few factors. First, it would be ideal to compare our diversity estimates to neutral expectations of diversity vs population size. For SNPs, the expected diversity under neutrality would be $4N_e\mu$ or a similar formula. For k-mers the exact value of μ is not clear, as k-mers also reflect non-SNP variation. Several k-mer-based measures of diversity are nearly perfectly correlated with π in neutral models where only SNPs are considered (Roberts et al., 2024), but the inclusion of non-SNP variation would probably change this relationship. Identifying the concrete variants underlying k-mers would help us better understand μ for k-mer, but this is currently not possible for mutations other than SNPs and small indels without pangenomic references (Uricaru et al., 2015; Gauthier et al., 2020). The eventual release of pangenomic references across many species will provide better context for increased scaling of k-mer diversity with population size. Similar to previous Lewontin's paradox studies, we also use both non-coding sequences (for k-mers) and 4-fold degenerate sites (for SNPs) as standards for estimating neutral diversity (Leffler et al., 2012; Buffalo, 2021), but there could be differences between coding and non-coding values of μ that contribute to results observed in our work and previous work. More estimates of μ are needed for non-SNP mutations (Quiroz et al., 2023), but we further caution that many estimates of μ will be limited in that they usually reflect

contemporary populations. Variation in μ over time, such as through TE bursts (Belyayev, 2014), could play an important role in diversity vs population size scaling and warrant further research.

In addition to limitations estimating μ and π, estimates of N_e also tend to be limited in investigations of Lewontin's paradox. Like previous analyses, our analysis assumes that contemporary population size estimates are good proxies for historic population sizes (Corbett-Detig et al., 2015; Buffalo and Coop, 2020). Common population size proxies such as range size and plant height only reflect the current census population size of a species. However, it is the long-term harmonic mean of the effective population size, not just the current size, that determines diversity levels within a population (Wright, 1940). There is some literature correlating aspects of species history with contemporary diversity levels (López□Delgado and Meirmans, 2022) and incorporating these metrics into Lewontin's paradox is an exciting avenue for future research. As a first step in this direction, we used plant range maps into native and invaded ranges to test the robustness of our results to invasion-related range size changes (Brown et al., 2023). Overall, our observations were remarkably similar no matter whether we included or excluded invaded ranges in our population size proxies (Figure C17A-C vs Figure C17D-F). Part of this apparent robustness was due to the insensitivity of our GBIF-based range size estimates to the inclusion of invaded ranges (Figure C7D). However, our WCVP-based range size estimates were drastically altered by the inclusion of invaded ranges (Figure C7C) and still yielded similar results (Figure 2.3, C15, C16). Although we cannot rule out the possibility that older historical events have affected contemporary diversity levels, our results appear to be robust to recent biological invasions.

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). For example, we did not filter out potentially contaminating sequences from our k-mer analysis, which could add noise to our k-mer dissimilarity values. However, this could only explain the increased scaling between k-mer diversity and population size if plants with larger population size proxies had a higher diversity of contaminants, which seems unlikely given that many of the sequencing datasets in our

study are from laboratory cultivated plants (see Table S1) and that we excluded low-frequency k-mers from our analyses. Our study is also unique in the multiple steps we took to limit the influence of systematic measurement errors on our coefficients, including reanalyzing all population-level sequencing data with one pipeline to limit the impact of bioinformatic parameter choices on our analysis (Mirchandani et al., 2024), filtering out low coverage datasets(Sandell et al., 2022), accounting for missing data in calculations of nucleotide diversity (Schmidt et al., 2021; Korunes and Samuk, 2021)and estimating range size with two different methods (WCVP range maps and GBIF occurrence records, Figure C6). 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.

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 Figure 2.3B vs Figure C15, 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 β = -3.7, SE = 0.42 in Figure 2.4). 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 (Figure 2.3A, C17-C19 panels A and D), but the relationship between k-mer diversity and population size proxies was often still highly significant (Figure 2.3B, C17-C19 panels B; C; E; F). The only exception was that Jaccard dissimilarity did not significantly scale with GBIF-based estimates of range size (Figure C19B, C19E). 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 may want to consider 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. 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, and understandings of each species' demographic history. Altogether, these methodological developments will hopefully reveal a more wholistic picture of variation across the tree of life.

Data availability

Supplemental figures are in appendix C. Our entire analysis is packaged as a snakemake workflow stored here: https://github.com/milesroberts-123/tajimasDacrossSpecies. All supplementary tables will be made available online at *Evolution Letters*. 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 covari-

ate 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.

CHAPTER 3: k-MER-BASED APPROACHES TO BRIDGING PANGENOMICS AND POPULATION GENETICS[3]

Abstract

Many commonly studied species now have more than one chromosome-scale genome assembly, revealing a large amount of genetic diversity previously missed by approaches that map short reads to a single reference. However, many species still lack multiple reference genomes and correctly aligning references to build pangenomes can be challenging for many species, limiting our ability to study this missing genomic variation in population genetics. Here, we argue that k-mers are a very useful but underutilized tool for bridging the reference-focused paradigms of population genetics with the reference-free paradigms of pangenomics. We review current literature on the uses of k-mers for performing three core components of most population genetics analyses: identifying, measuring, and explaining patterns of genetic variation. We also demonstrate how different k-mer-based measures of genetic variation behave in population genetic simulations according to the choice of k, depth of sequencing coverage, and degree of data compression. Overall, we find that k-mer-based measures of genetic diversity scale consistently with pairwise nucleotide diversity (π) up to values of about $\pi = 0.025$ ($R^2 = 0.97$) for neutrally evolving populations. For populations with even more variation, using shorter k-mers will maintain the scalability up to at least $\pi = 0.1$. Furthermore, in our simulated populations, k-mer dissimilarity values can be reliably approximated from counting bloom filters, highlighting a potential avenue to decreasing the memory burden of k-mer based genomic dissimilarity analyses. For future studies, there is a great opportunity to further develop methods to identifying selected loci using k-mers.

Introduction

Two decades ago, assembling one reference genome for one eukaryotic species was an international, herculean effort (Venter et al., 2001). Now, individual laboratories can readily assemble and align multiple reference-quality genomes from the same species into pangenomes (Golicz et al., 2020). This shift toward pangenomes as the basis for genetic studies is already transforming our

understanding of genetic variation in populations. Analysis of pangenomes has uncovered vast quantities of genetic variation previously missed by the ubiquitous practice of aligning short reads to a single reference genome (Wong et al., 2020; Sirén et al., 2021; Ebler et al., 2022; Zhou et al., 2022; Rice et al., 2023; Liao et al., 2023), increased the power of trait mapping (Song et al., 2020; Chin et al., 2023), and resolved complex structural variations (Hickey et al., 2020; Song et al., 2024). Pangenomes have even revised our understanding of previously cataloged variation, because single nucleotide polymorphisms (SNPs) once identified by mapping reads against a single reference can sometimes be resolved as alignment errors due to structural variation (Jaegle et al., 2023). Altogether, pangenomes provide a more accurate representation of genetic variation, reducing the commonly observed phenomenon where one's choice of reference genome shapes the ultimate conclusions of a study (i.e. reference bias, Gage et al. (2019); Günther and Nettelblad (2019); Chen et al. (2021); Ebler et al. (2022)). This better ability to capture and explain patterns of genetic variation, combined with recent developments in pangenome assembly algorithms (Garrison et al., 2024; Hickey et al., 2020) and an explosion in pangenome sequencing for many non-model organisms (Lei et al., 2021), means that pangenomes will likely be the standard for population genetics analysis in the near future.

However, pangenomes can be difficult to assemble and tune in some contexts (Hahn, 2018; Song et al., 2024). First, pangenomes by definition require more sequencing data to assemble compared to a single reference genome, which can make building pangenomes expensive for large genome species or study systems with fewer resources. Furthermore, whole genome alignment (WGA) is key for pangenome assembly but also difficult to tune, affecting pangenome analysis. Modern WGA algorithms are impressive and are being used to produce pangenomes for a wide range of species (Garrison et al., 2024); however, current WGA algorithms are mainly developed and tuned to align human and model species genomes, and so do not always generalize well to genomes that are large, highly repetitive, highly diverse, polyploid, or containing high levels of structural variation (reviewed in Song et al. (2024)). For example, corn (*Zea mays*) has a famously repetitive and structurally variable genome and new WGA approaches needed to be developed just

to properly align corn genomes (Song et al., 2022). There are also still many open questions on how to best represent complex, nested variations in such alignments and tune alignment parameters (Song et al., 2024). A researcher's exact choice of alignment parameters and software can drastically affect the shape of a pangenome graph (Rice et al., 2023) and downstream genotype calls (O'Rawe et al., 2013; Li, 2014; Bush et al., 2020; Betschart et al., 2022; Sopniewski and Catullo, 2024). Given the challenges of computing and tuning alignments, approaches that skip alignment altogether could valuably complement or help guide pangenome assembly.

In this review we argue that the k-mers deserve more attention from population geneticists because they can complement the study of pangenomes. A k-mer is a sub-sequence of length k within a larger sequence. For example, the sequence "ATGCA", contains the unique 2-mers AT, TG, GC, and CA. The main benefit of k-mers is that they can be analyzed without alignment. Instead, one simply counts all of the unique k-mers present in a sample of reads (where k-mers that are reverse complements of each other are typically considered identical and are counted together Kokot et al. (2017)), then uses the resulting count matrix for downstream analysis (see Figure 3.2). Thus, k-mers derived from regions far diverged, absent, or otherwise unalignable in a given reference will not be automatically excluded from an analysis, allowing one to get a picture of pangenomic variation. k-mers have a long history of use in metagenomics (McClelland, 1985; Rosen et al., 2008; Dubinkina et al., 2016), phylogenetics (Kolekar et al., 2012; Haubold, 2014; Zielezinski et al., 2017, 2019; Bussi et al., 2021; Beichman et al., 2023; Jenike et al., 2024), computer science (Shannon, 1948), and quantitative genetics (Voichek and Weigel, 2020; Kim et al., 2020; Gupta, 2021; Onetto et al., 2022; Lemane et al., 2022). However, while applications of k-mers receive much attention in other disciplines, population genetic investigations using k-mers remain limited.

Our goal is to review k-mer-based approaches for identifying, measuring, and explaining patterns of genetic variation in populations. At the same time, we investigate the behavior of k-mer-based measures of variation to the choice of k, the depth of sequencing coverage, and the degree of data compression. We finally highlight some avenues to explore in k-mer-based (i.e. reference-free or alignment-free) population genetics. Overall, we advocate that k-mer-based approaches can be

valuable complements to common reference-based population genetics methods.

Box 1: What is the "best" value for k?

A common first question in k-mer based analyses is: What value(s) of k should be analyzed? The "best" k for an analysis is ultimately determined by a trade-off between the length of k and sequencing error: longer k-mers are more likely to represent unique genomic sequences, but are also more likely to contain a sequencing error (Rahman et al., 2018). In practice, many studies use a k of around 20-40 bp (Ponsero et al., 2023) because k-mers in this range can be reliably sequenced with short read data and often align uniquely to their source genome (Wu et al., 1991; Becher et al., 2022). For example, $k = 32$ captures 85.7 % of unique sequences in the human genome (Shajii et al., 2016), while $k = 21$ distinguishes many eukaryote, bacteria, and archea species (Bussi et al., 2021). However, there are many past studies that propose criteria for choosing specific values of k.

Choosing multiple values of k

One "brute force" approach to test the sensitivity of results to k is to simply repeat an analysis multiple times for different values of k. This approach is especially common among genome assembly algorithms (Chikhi and Medvedev, 2014; Durai and Schulz, 2016) but could be applied to almost any analysis in theory. The main drawback, however, is the high computational burden of performing the same analysis multiple times. It is also difficult to know without more information whether analyses performed for certain values of k produce more accurate results than analyses for other values of k, motivating the need for k selection criteria that can be either minimized or maximized.

Choosing k based on the number of unique non-erroneous k-mers

Higher values of k generally allow greater detection of unique sequences, but increase the probability of observing k-mers containing at least one sequencing error (Chikhi and Medvedev, 2014; Rahman et al., 2018). Each sequencing error can result in up to k erroneous k-mers, making it important to prevent errors from dominating one's analysis. Thus, choosing a k that maximizes the number of unique non-erroneous k-mers in a dataset is generally considered optimal for tasks like

genome assembly (Chikhi and Medvedev, 2014). This approach generally involves measuring the k-mer frequency spectrum and then fitting a model to the distribution to estimate which parts of the spectrum come from erroneous k-mers (Chikhi and Medvedev, 2014). Usually, the low-frequency end of the spectrum is dominated by erroneous k-mers because sequencing errors are unlikely to generate the same erroneous k-mers many times and usually convert real k-mers into k-mers not found in the source genome (Kelley et al., 2010). Although this criterion for choosing k could be applied to population genetic datasets, it is not if clear the resulting optimal k would vary significantly between genomes within the same species. Presumably, if genomes within the same species have considerable variation in repetitive content (Haberer et al., 2020), size (Schmuths et al., 2004), or ploidy (reviewed in Kolář et al. (2017)) then the optimal k could vary. Determining the value of k that maximizes the number of unique non-erroneous k-mers across all individuals in a population may be of interest for future population genetics studies.

Choosing k based on the probability of chance k-mer matches between samples

Another way to choose k is to think about the probability that a k-mer from one genome is also found in a second genome by chance alone. For example, it would be unsurprising for almost any pair of reasonably long, naturally-occurring DNA sequences to share the k-mer "AGC" because this k-mer is very short and could occur many times in a random sequence. How large must k be then before finding a k-mer in two different sequences is unlikely by chance alone? To discuss this question, we next present equations similar to ones in Ondov et al. (2016), except we generalize to account for variation in base composition between sequences being compared.

Let's begin by imagining we have two genome sequences we wish to compare: X_1 and X_2. First, to generate X_1, we sample with replacement the letters A, T, G, and C a total of L times with probabilities p_{A1}, p_{T1}, p_{G1}, and p_{C1}, respectively where $p_{A1} + p_{T1} + p_{G1} + p_{C1} = 1$. In other words, we sample our DNA from a multinomial distribution where we assume that each base is sampled independently of the preceding bases. This gives us one strand for X_1 and we can then generate the complementary strand by pairing A with T and G with C. We repeat this whole process once more to construct X_2 using probabilities p_{A2}, p_{T2}, p_{G2}, and p_{C2}, which may or may not match the

probabilities for X_1, except we only sample k bases. This gives us a k-mer, K, in the genome sequence of X_2. We now wish to find the probability that K occurs in X_1, assuming that the sampling of bases for K was independent of sampling bases for X_1.

If we imagine constructing X_1 and K at the same time, then the probability of drawing the same letter for both sequences at a given position is:

$$\Sigma_F = p_{A1}p_{A2} + p_{T1}p_{T2} + p_{G1}p_{G2} + p_{C1}p_{C2} \tag{3.1}$$

However, we also want to consider sequences that are reverse complements as identical, so the probability of drawing a pair of bases that match as reverse complements is:

$$\Sigma_R = p_{A1}p_{T2} + p_{T1}p_{A2} + p_{G1}p_{C2} + p_{C1}p_{G2} \tag{3.2}$$

The probability of drawing k pairs of bases in a row that are either identical or reverse complement matches is thus Σ_F^k and Σ_R^k, respectively, assuming that each base is sampled independently of all others. The complementary probability of not getting k matches in a row on the forward strand is $1 - \Sigma_F^k$ and, comparably, $1 - \Sigma_R^k$ for the reverse strand. Because X_1 contains L bases, there are a total of $L - k + 1$ k-mers in X_1, so the probability of a K not matching at any k-mers in X_1 is approximately $(1 - \Sigma_F^k)^{L-k+1}$ and $(1 - \Sigma_R^k)^{L-k+1}$. It should be noted that at this step we are assuming that the chance of K mismatching at a given position in X_1 is independent of the chance that K mismatches at other positions in X_1, which is not strictly true. For example, if K is the sequence "AAAAA" and we compare K to a subsequence in X_1 that is "AACAA" we would say that K and this subsequence do not match. However, with this information we would also know with certainty that adjacent k-mers in X_1 will also not match K because they will still contain a "C". However, we will continue with assuming that the mismatch between X_1 and K is independent at all positions because this makes our final equations more conservative (decreasing the Σ_F^k term in Equation 3.5 will decrease the overall value of Equation 3.5).

We can now say that the probability of finding the k-mer K at least once by chance alone at

any of the $L - k + 1$ positions in the sequence X_1 is:

$$P(K \in X_{1F}) = 1 - \left(1 - \Sigma_F^k\right)^{(L-k+1)} \tag{3.3}$$

$$P(K \in X_{1R}) = 1 - \left(1 - \Sigma_R^k\right)^{(L-k+1)} \tag{3.4}$$

Assuming $k << L$, Equations 3.3 and 3.4 approximate to:

$$P(K \in X_{1F}) \approx 1 - \left(1 - \Sigma_F^k\right)^{L} \tag{3.5}$$

$$P(K \in X_{1R}) \approx 1 - \left(1 - \Sigma_R^k\right)^{L} \tag{3.6}$$

We confirmed that equations 3.5 and 3.6 work as expected in simulations (Figure 3.1). Similar to Fofanov et al. (2004), we can then solve these equations to give the minimum k-mer length required to achieve a desired probability of chance k-mer matching of $q = P(K \in X_{1F}) = P(K \in X_{1R})$:

$$k_F = \left\lceil log_{\Sigma_F}\left(1 - (1 - q)^{1/L}\right)\right\rceil \tag{3.7}$$

$$k_R = \left\lceil log_{\Sigma_R}\left(1 - (1 - q)^{1/L}\right)\right\rceil \tag{3.8}$$

and now given a choice of q we are willing to tolerate, we can then use $\max(k_F, k_R)$ as a potential choice of k.

Equations 3.7 and 3.8 demonstrates that $k = 19$ reduces the probability of two 3 Gb genomes of random sequence sharing the same k-mer by chance to just 1 % (Ondov et al., 2016) (assuming

all bases are in equal proportion in both genomes, $p_{A1} = p_{T1} = p_{G1} = p_{C1} = p_{A2} = p_{T2} = p_{G2} = p_{C2} = 1/4$) while $k = 27$ gives $q = 1 \times 10^{-6}$ for a 10 Gb genome where both X_1 and X_2 have a GC conent of 42 % ($p_{A1} = p_{T1} = p_{A2} = p_{T2} = 0.29$ and $p_{G2} = p_{C2} = p_{G2} = p_{C2} = 0.21$). Slightly longer k-mers are required when the proportion of each base is not 25 % in both genomes because it is more likely then to have low-complexity strings of bases which make spurious matches more likely. For example, the required k-mer length increases to 34 bp if the GC content of both genomes is 21 % ($p_{A1} = p_{T1} = p_{A2} = p_{T2} = 0.395$ and $p_{G1} = p_{C1} = p_{G2} = p_{C2} = 0.105$), given a 10 Gb genome and $q = 1 \times 10^{-6}$.

In summary, if k is high enough, two genomes in a population are highly unlikely to share k-mers just through the accumulation of random sequences alone, suggesting that shared k-mers usually reflect shared ancestry. Developing additional k-mer selection criteria that explicitly model the influence of shared ancestry on the probability of k-mer sharing could improve approaches to choosing k.

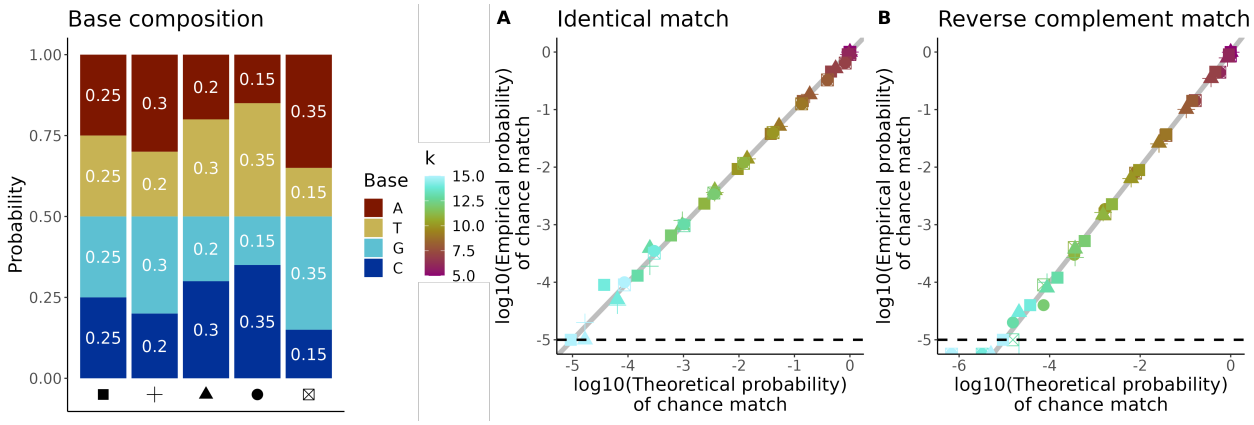

Figure 3.1: **The relationship of k-mer length and base composition to chance matching.** We verified (A) Equation 3.5 and (B) Equation 3.6 by randomly generating 5.5 million pairs of k-mers and genomes. Each simulation had 100,000 trials. For each trial we randomly generated a genome of length 10,000 bases using a given base composition (p_A, p_T, p_G, p_C in barchart) and also generated a random k-mer (color gradient shows k) using the same base composition. We then checked whether the random k-mer was in the random genome. In total, the 5 different base compositions $\times$ 11 different k-mer lengths $\times$ 100,000 trials per simulation gives 5.5 million trials total. Grey line is where the theoretical (i.e. Equations 3.5 and 3.6) and empirical probabilities (i.e. from simulations) of a chance match are equal. Dotted black line shows the resolution of the simulations; because we did 100,000 trials per simulation we cannot empirically estimate probabilities smaller than 1/100,000.

Choosing k based on the balance of shared vs differing k-mers

A final class of approaches for choosing k focuses on how if k is small, most k-mers are shared between most samples, but as k increases more sample-specific k-mers occur until no shared k-mers remain (Zhang et al., 2017a). Neither of these scenarios is usually desirable - the former makes all samples appear to be identical, while the latter makes all samples appear to be completely different. Thus, choosing k boils down to balancing the number of shared vs differing k-mers. Some statistics for informing this choice include cumulative relative entropy (Sims et al., 2009), relative sequence divergence (Sims et al., 2009), average number of common features (Zhang et al., 2017a), Shannon's diversity index (Zhang et al., 2017a), and χ^2 tests (Bai et al., 2017). However, there are two main drawbacks of these statistics. First, they were mainly developed for phylogenetic studies focused on short amino acid k-mers, so their utility for population genetics (which would probably focus on longer nucleotide k-mers) is untested. Furthermore, different values of k will appear

optimal for genomes of different sizes (Zhang et al., 2017a), making it hard to imagine that one optimal k exists for a population exhibiting substantial genome size variation. Extending these or similar approaches to optimize k for populations with genome size variation will be very useful for future studies.

Box 2: What's the expected value of k-mer diversity in a neutrally evolving population?

We argue that k-mers can be useful for initial assessments of genetic diversity, but two key questions are likely of interest to population geneticists: can k-mer diversity be related to nucleotide diversity (the common measure of genetic diversity) and what is the expected value of k-mer metrics for populations evolving under a given model? Here, we derive a simple bound on the expected number of k-mer differences between a pair of individuals in a neutrally evolving population. Our result is similar to (Shi et al., 2024), except we generalize beyond haploid organisms.

We start by assuming that (1) k is sufficiently long to capture all of the unique sequences in a genome, (2) each genome is sequenced at sufficient coverage to confidently identify all of the k-mers present, (3) k-mers are sequenced without error, and (4) only SNPs contribute to differences between genomes. Let i and j be individuals in a neutrally evolving population with ploidy level x and let y and z be two haploid genomes in the pool of i and j that differ by origin. Next, let K_i and K_j represent the set of k-mers identified in individual i and j's genomes respectively and let $|K_i|$ denote the size of the set K_i. If all of the genomes within i and j are identical, then all k-mers are shared between them. If we start by assuming i and j are haploid, this means a maximum of $2k$ k-mers are not shared between i and j for every pairwise difference between i and j (Iqbal et al., 2012; Younsi and MacLean, 2015). This relationship can be written as:

$$|K_i \cup K_j| - |K_i \cap K_j| \leq 2kD_{yz} \tag{3.9}$$

where D_{yz} represents the number of pairwise differences between haplotypes y and z and the left side represents the number of k-mers exclusive to either i or j. The reason for the using a "$\leq$" in Equation 3.9 is to account for three possibilities: (1) if two SNPs are less than k bases apart, there

will be fewer than $4k$ k-mers not shared between i and j, (2) it's possible for a SNP to turn one k-mer into a different k-mer that's already present elsewhere in a genome and (3) at higher ploidy levels, if a segregating site is heterozygous in i and j then i and j will share all k-mers between them.

To generalize beyond haploid organisms, we will replace the conversion factor of 2 in equation 3.9 above with a general function $a(x)$:

$$|K_i \cup K_j| - |K_i \cap K_j| \leq a(x) \sum_{y<z} D_{yz} \tag{3.10}$$

where $a(x)$ is a conversion factor that turns a number of pairwise differences into a number of sample-exclusive k-mers as a function of x and $\sum_{y<z} D_{yz}$ is all of the pairwise differences for any pair of haplotypes y and z in i and j. The values that maximize $a(x)$, maintaining the validity of using a "$\leq$" sign in Equation 3.10, are as follows:

$$a(x) = \begin{cases} \frac{2k}{x^2}, & \text{if } 1 \leq x \leq 3. \\ \frac{k}{2x-1}, & \text{if } x \geq 4. \end{cases} \tag{3.11}$$

When the population is haploid ($x = 1$), Equation 3.10 reduces to Equation 3.9. When $x = 2$ or 3, $a(x)$ is maximized when i and j are homozygous for different alleles, creating $2k$ sample-exclusive k-mers for every 4 or 9 pairwise differences, respectively. However, when $x \geq 4$, the situation that maximizes $a(x)$ is one where the SNP exists on only one haplotype in i or j, creating k sample-exclusive k-mers for every $2x - 1$ pairwise differences.

Now we will convert the right side of Equation 3.10 into nucleotide diversity (π). Summing across all pairs of individuals gives:

$$\sum_{i<j} |K_i \cup K_j| - |K_i \cap K_j| \leq a(x) \sum_{i<j} \sum_{y<z} D_{yz} \tag{3.12}$$

Next, we convert the right-hand side into genome-wide average π by dividing both sides by the

number of pairwise haplotype comparisons (Korunes and Samuk, 2021), which is the number of individuals sampled n times ploidy x, choose 2:

$$\frac{\sum_{i<j} |K_i \cup K_j| - |K_i \cap K_j|}{\binom{nx}{2}} \leq a(x) \left(\frac{\sum_{i<j} \sum_{y<z} D_{yz}}{\binom{nx}{2}} \right) \tag{3.13}$$

$$\frac{\sum_{i<j} |K_i \cup K_j| - |K_i \cap K_j|}{\binom{nx}{2}} \leq a(x)\pi \tag{3.14}$$

Next, isolating π on one side gives:

$$\frac{\sum_{i<j} |K_i \cup K_j| - |K_i \cap K_j|}{a(x)\binom{nx}{2}} \leq \pi \tag{3.15}$$

The intuition behind this formula is that, in a world where our assumptions are met, the number of k-mers that are not shared between a pair of samples is bounded by a multiple of π. The need to scale π by $a(x)$ to get a bound reflects the fact that a given SNP can be captured by multiple k-mers, but the exact relationship between pairwise differences and sample-exclusive k-mers depends on ploidy.

Finally, we can substitute π for the standard formula for the expected value of π in a neutrally evolving population at equilibrium (Tajima, 1996):

$$E\left[\frac{\sum_{i<j} |K_i \cup K_j| - |K_i \cap K_j|}{a(x)\binom{nx}{2}}\right] \leq \frac{2xN_e\mu}{1 + \frac{4}{3}2xN_e\mu} \tag{3.16}$$

where N_e is effective population size and μ is the mutation rate (probability of mutation per base pair per generation). The denominator on the right hand side of Equation 3.16 ensures that the expected value of π saturates as it approaches it's theoretical maximum of 0.75 (Tajima, 1996). Altogether, equations 3.15 and 3.16 provide simple bounds for the average number of k-mers that differentiate a pair of samples in terms of N_e and μ.

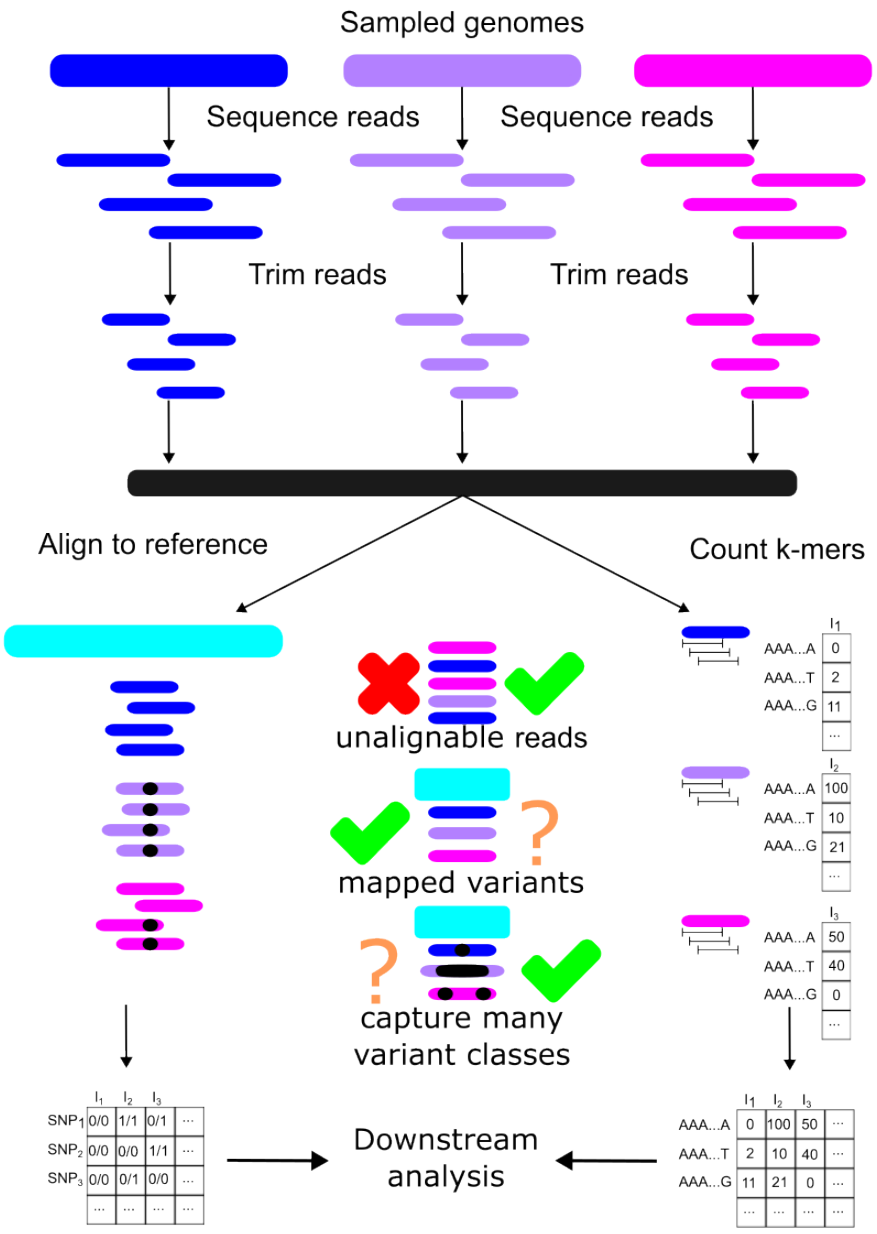

Figure 3.2: **Comparison between a typical SNP-calling workflow and a k-mer counting workflow.** A typical k-mer-based analysis begins the same as a SNP-based analysis: sequencing reads from sampled genomes followed by trimming and other quality control steps on the reads. The crucial difference comes down to whether the remaining reads are aligned to a reference sequence (discarding any unaligned reads) or are used for k-mer counting. Workflows may vary in terms of attempting to map k-mers to specific loci or calling variants other than SNPs (denoted with "?"). The end result for each workflow is typically a matrix where each row is a different genomic variant, each column is a sample, and the elements represent the variant states. For SNPs, the standard variant call format notation is to use 0/0 to represent homozygous reference genotypes, 0/1 or 1/0 to represent heterozygous genotypes, and 1/1 to represent homozygous alternate genotypes. In k-mer based analyses the variant states are instead counts of how many k-mers of a particular type were observed in each sample.

Identifying variation with k-mers

Many of the earlier studies using k-mers to identify variants *de novo* focus on microbes, where (despite small genome sizes) high levels of diversity make alignment and choosing a reference genome difficult (Gardner and Slezak, 2010). However, similar approaches have now been extended to other taxa. One class of approaches calls SNPs *de novo* simply by comparing k-mers between samples and searching for pairs of k-mers that differ at their central basepair (Gardner and Slezak, 2010; Gardner and Hall, 2013; Bedo et al., 2016; Li et al., 2022). Another, more popular, class of methods relies on de Bruijn graphs, which are graphs where k-mers are nodes and two nodes are connected if they share $k-1$ bases (Compeau et al., 2011). SNPs in a sample then appear as "bubbles" in these graphs and calling SNPs amounts to searching for these bubbles (Iqbal et al., 2012; Leggett et al., 2013; Younsi and MacLean, 2015; Uricaru et al., 2015; Standage et al., 2019; Gauthier et al., 2020). Finally, other approaches first identify k-mers that are present in all samples (i.e. "anchor" k-mers) and finds paths between anchor k-mers through either local alignment or traversing de Bruijn graphs (Audano et al., 2018; Kaplinski et al., 2021; Aylward et al., 2023). The main drawback to these approaches is that the relative positions of the resulting variant calls are usually unknown without further analysis (see Figure 3.2).

To call variants with known relative positions, it is possible to combine the strengths of k-mers with either pangenomes or databases of previously-identified variants. This is because unique variations in pangenomes will often be tagged by multiple unique k-mers. For example, one can compare k-mers with a reference set of SNPs (Shajii et al., 2016; Pajuste et al., 2017; Denti et al., 2019; Shi et al., 2023; Chu et al., 2024) or insertions (Puurand et al., 2019) to quickly genotype samples without alignment. Or if a reference pangenome assembly is available, it's possible to use k-mers to infer the path through the pangenome corresponding to a given sample genome (Iqbal et al., 2012; Ebler et al., 2022; Grytten et al., 2022; Häntze and Horton, 2023). However, similar to common alignment-based SNP calling practices, the called variants will be limited to variants present in the pangenome reference.

If references are not already available, k-mers can be especially helpful for identifying and as-

sessing variation in species with little prior knowledge. For example, the popular tools genomescope and smudgeplot use the distribution of k-mer counts in a sample (i.e. the k-mer frequency spectrum) to rapidly estimate important parameters, including genome size, heterozygosity, and ploidy structure all from unassembled read sets (Vurture et al., 2017; Ranallo-Benavidez et al., 2020). k-mer frequency spectra can also identify unwanted variation that is potentially due to contamination (reviewed in Cornet and Baurain (2022)) or sequencing error. k-mers have the useful property that sequencing errors mainly manifest as k-mers that occur just once or a few times in a sample of reads (Kelley et al., 2010). This is because random sequencing errors are unlikely to generate the same k-mer many times and (if k is long enough, see Box 1) usually produce k-mers not found in the target genome. Thus, excluding low-copy k-mers from an analysis can mitigate sequencing error and one can examine the k-mer frequency spectrum to choose an appropriate k-mer count cutoff (Zhao et al., 2018). Although k-mer frequency spectra can be usefully mined for genome parameters, their exact relationship to key population genetic parameters, such as measures of differentiation or diversity, remains underexplored.

Measuring variation with k-mers

After calling variants, quantifying levels of variation is a crucial step in many population genetics workflows. Because counting k-mers tends to be faster than alignment, k-mers could be especially helpful for rapid, initial assessments of diversity that complement or guide pangenome analysis. The standard approach to measure diversity in a population is to align sample sequences to a reference genome then calculate either (1) the average level of heterozygosity across sites or (2) the number of variants segregating in the sample. These are represented by Nei's (Equation 3.17; Nei and Li (1979); Nei and Tajima (1981)) and Watterson's (Equation 3.18; Watterson (1975)) estimators of diversity, respectively:

$$\pi = (1 - \sum_i p_i^2)(\frac{n}{n-1}) \tag{3.17}$$

$$\theta_w = \frac{S}{\sum_{a=1}^{a=n-1} \frac{1}{a}} \tag{3.18}$$

where n is the number of sequences in the sample, p_i is the frequency of the ith allele at a locus, and S is the number of segregating sites at a locus. While θ_w is based on a discrete count of variants (S), π is shaped by the allele frequencies of variants and will be higher if variants are common than if variants are rare (note that π is more commonly rewritten in terms of the average number of differences between sequences (Korunes and Samuk, 2021)).

Can analogous measures of variation be derived from k-mers? There are over 30 valid measures of genetic difference based on k-mer counts used in previous literature (Benoit et al., 2016; Zielezinski et al., 2017; Luczak et al., 2019; Zielezinski et al., 2019). However, the three most common k-mer dissimilarity measures are arguably Jaccard dissimilarity (Equation 3.19; Ondov et al. (2016)), Bray-Curtis dissimilarity (Equation 3.20; Dubinkina et al. (2016); Benoit et al. (2020)), and cosine dissimilarity (Equation 3.21, Choi et al. (2019)):

$$J(K_i, K_j) = 1 - \frac{K_i \cap K_j}{K_i \cup K_j} \tag{3.19}$$

$$B(C_i, C_j) = 1 - 2\sum_{b=1}^{4^k} \frac{\min(m_b(C_i), m_b(C_j))}{m_b(C_i) + m_b(C_j)} \tag{3.20}$$

$$C(C_i, C_j) = 1 - \frac{C_i \cdot C_j}{\|C_i\| \times \|C_j\|} \tag{3.21}$$

where k is the length of k-mers to be included in the comparison, K_i and K_j are the set of k-mers of length k present in a set of reads i and j, C_i and C_j are vectors of k-mer counts in a set of reads i and j, and m_i is a function that returns the relative frequency of the bth k-mer in a set (i.e. standardized such that $\sum_{b=1}^{4^k} m_b(C_i)$ and $\sum_{b=1}^{4^k} m_b(C_j)$ equal 1). These measures work similarly to the classical π and θ_w measures: the numerators are a measure of the number of sites that vary between individuals,

while the denominators are a measure of sample size (here the number of k-mers rather than number of haplotypes). The main difference, however, is that k-mer-based measures of genetic dissimilarity are not directly interpretable in terms of mutations, like π and θ_w can be with the assumption that each SNP represents one mutation (Haubold et al., 2011; Haubold and Pfaffelhuber, 2012). Any given k-mer may represent the combined presence of multiple mutations (Voichek and Weigel, 2020; Blanca et al., 2022) and any mutation can generate multiple new k-mers. Although this means that k-mer-based genetic dissimilarity measures are only proxies for the true mutational distance between individuals, they still effectively resolve relationships between lineages compared to alignment-based measures (VanWallendael and Alvarez, 2022).

While Equations 3.19, 3.20, 3.21 have all been successfully used to measure genetic dissimilarity between samples in past studies, the formulae highlight their benefits and drawbacks. First, Jaccard dissimilarity, perhaps the most commonly used k-mer dissimilarity metric (Ondov et al., 2016; Ruperao et al., 2023), requires only knowing k-mer presence/absence patterns in samples instead of k-mer counts and thus can take less memory to calculate than other k-mer-based dissimilarity measures. However, as a consequence Jaccard dissimilarity may not capture the effects of copy number variation and does not account for variation in coverage between samples, which affects whether a given k-mer is called as "present" in a sample (VanWallendael and Alvarez, 2022). Approaches that measure k-mer counts, like Bray-Curtis dissimilarity (Equation 3.20) and cosine dissimilarity (Equation 3.21), can better account for these influences, but may require more memory for storing counts (Liu et al., 2017; Choi et al., 2019). To alleviate this memory problem, many approaches calculate approximate k-mer dissimilarity measures with a small subset of k-mers (Ondov et al., 2016; Zhao, 2019; Benoit et al., 2020; Pellegrina et al., 2020). An alternative approach would be to instead compress the k-mer counts into a smaller array, keeping information from more k-mers while simultaneously alleviating memory burdens (Melsted and Pritchard, 2011), but such approaches have not been used to calculate genetic dissimilarity before. The relationship between k-mer-based dissimilarity measures and π is also rarely explored. While some studies have investigated the relationship between π and Jaccard dissimilarity (VanWallendael and Alvarez, 2022),

other k-mer dissimilarity measures are possible and no studies to our knowledge have compared these approaches for populations of varying levels of diversity - a key determinant of whether alignment-based genotype calls are accurate (Cornish and Guda, 2015; Bush et al., 2020). In the following sections, we investigate the efficacy of compressed and uncompressed k-mer-based measures of variation at capturing the true pairwise diversity of simulated populations.

Testing the efficacy of k-mer measures of variation.

We used simulations to investigate the relationship between the true value of π and genetic diversity measured from k-mer based approaches from simulated sequencing reads.

Simulations

We simulated a neutrally evolving 100kb segment of the *Arabidopsis thaliana* genome. We first forward simulated 300 neutrally evolving populations with SLiM 3 (Haller and Messer, 2019). Each population consisted of 100 individuals simulated for 1000 generations with a uniform recombination rate of 10^{-8}. To vary the diversity across simulations we varied the mutation rate between 10^{-6} and $2 * 10^{-4}$. For each simulation, we tracked the ancestry through tree-sequence recording which records the geneological history of all samples (Haller et al., 2019). From these trees, we generated sequences based on the *Arabidopsis thaliana* genome (chromosome 1 at positions 4,185,001-4,285,000) (Kent, 2002). We took a sub-sample of 10 individuals from the tips of the trees and randomly assigned nucleotides to each SNP in the sample using msprime version 1.2.0 and tskit version 0.5.6 (Baumdicker et al., 2022). From the sub-tree of the sampled genomes, we recorded the exact number of true average pair-wise differences (π_t) across the sample of 10 individuals (20 chromosomes) using msprime (Baumdicker et al., 2022).

Generating k-mers

To test the performance of the k-mer measures on unaligned reads, we simulated reads for each genome using an Illumina read simulator, InSilicoSeq 2.0.0 (Gourlé et al., 2019). We varied the read count to later investigate the effect of coverage as described below. We generated two sets of reads for each individual at coverages of 10x and 30x. After simulating reads, we counted k-mers within the reads using KMC3 (Deorowicz et al., 2015; Kokot et al., 2017). We generated k-mer

count vectors with $k = 10, 20, 30$, and 40 for each individual. We used a threshold-based approach to adjust the k-mer vectors to reduce the effects of sequencing errors; any k-mer count below the threshold value was set to zero for a particular sample before any dissimilarity calculations. For a threshold of 5, only k-mers with counts of 5 or more were considered when calculating the difference between two or more groups of k-mer counts. We present data with a threshold of 5, but note that a threshold of 0 is qualitatively similar with dissimilarity scores being slightly higher overall. This practice of filtering out low-coverage k-mers is analogous to the common practice of filtering out SNPs below a given minor allele frequency threshold (Asif et al., 2021). After k-mer counting, we calculated the genetic dissimilarity of populations with the Bray-Curtis (Equation 3.20) and cosine dissimilarity (Equation 3.21) measures.

The effect of k on k-mer similarity metrics

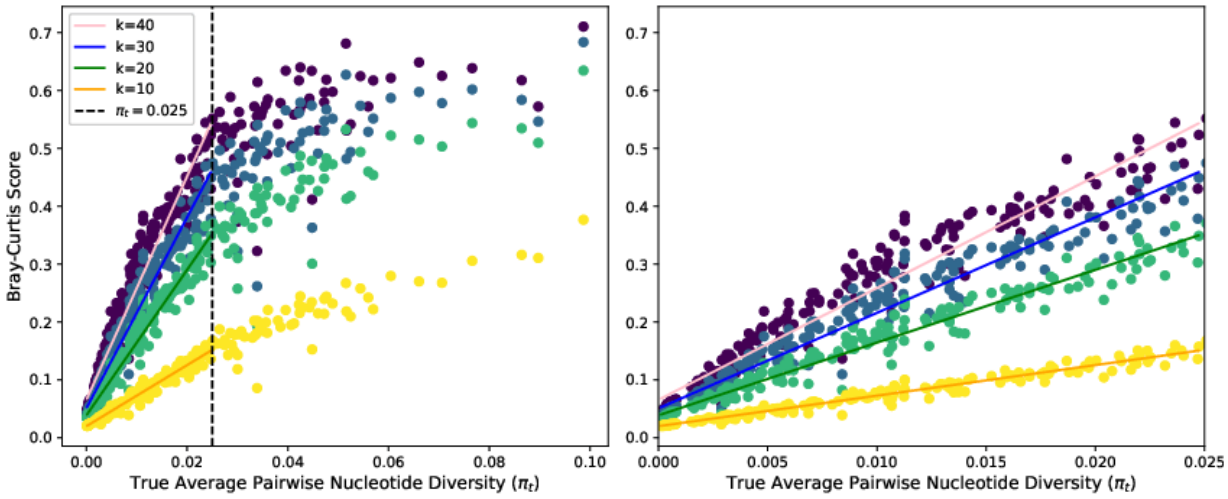

Figure 3.3: **Effect of k on Bray-Curtis dissimilarity.** Bray-Curtis dissimilarity calculated from simulated reads with a coverage of 30 where each point represents a sample of 20 chromosomes. Displayed are 10-mers (yellow), 20-mers (green), 30-mers (blue), and 40-mers (purple) each with a linear regression line for π_t between 0 and 0.025. The scores for populations with a π_t less than 0.025 is shown to the right.

Figure 3.3 shows the effect of k on the Bray-Curtis score with 30x coverage. We observe a plateau in the scores when diversity exceeds $\pi \approx 2.5\%$. This plateau occurs because, when diversity is high, SNPs cause most of the k-mers to be different between two samples. This effect

is especially true with high k values as one variant sampled in only one read can appear in up to $2k$ k-mers if it is sampled away from the edges of a read. The elevated number of k-mers at high diversity means that it is harder to interpret differences in dissimilarity measures between samples above this threshold, but this problem can be mitigated by using a lower k value, such as k=10 when the expected diversity in a sample is high. If π_t is expected to be below 0.025, larger k values can have higher precision and capture more unique k-mers in individual samples, better estimating true diversity.

While the data presented in Figure 3.3 were simulated with a coverage of 30x, Figure D1 shows that a coverage of 10x results in qualitatively similar scores when k=30. While coverage affects the precision of the measures, the overall trends and relative rankings between simulations remain consistent.

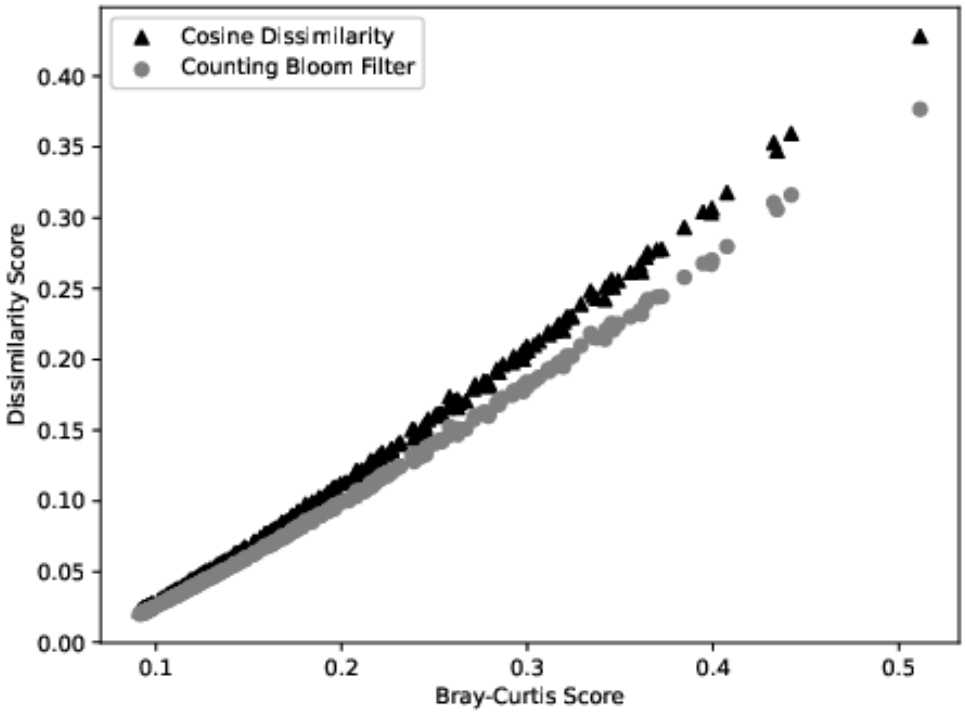

Figure 3.4: **Effect of bloom filter on cosine dissimilarity.** Cosine dissimilarity (triangle) and counting bloom filter scores (circle) from 10-mers of simulated reads with a coverage 30 compared directly against the Bray-Curtis score of each simulated set of 10 individuals (points).

The Counting Bloom filter approach to comparing genomes

Bloom filters are a data structure that can compress a k-mer count vector into a smaller array (Melsted and Pritchard, 2011). We can compute the dissimilarity of two compressed vectors using less memory and with increased efficiency as fewer entries are compared. Counting bloom filters (CBFs) are modifications of bloom filters (Fan et al., 2000), and are used in alternative k-mer count methods and error correction in sequencing data (Melsted and Pritchard, 2011; Roy et al., 2014; Shi et al., 2010). CBFs compress the k-mer vectors through several hash functions that map to a smaller array. Counting bloom filters will increase the k-mer count in the position as opposed to setting the count to one if a k-mer has been mapped there (see Figure 3.5). Because of the hash-functions and collisions, false positives are possible; here a "false positive" means that two k-mers of different sequences may map to the same locations in the vector (i.e. have a hash collision). However, identical mapping of two different k-mers is not a concern for relative comparison of diversity, since the same hash functions are used for each sample so collisions are consistent across samples. Collisions will cause diversity to be underestimated, but we can reduce collisions by increasing the number of hash-functions used in the process (Melsted and Pritchard, 2011).

A CBF gives a standardized way to compare across species/experiments if the vectors for k-mer counts are set to the same size and the same hash functions are used to generate the vectors across data sets. We can also perform the same cosine dissimilarity measure on the CBF vector that we use directly on the counts of k-mers. The cosine dissimilarity measure on the raw k-mer count vectors has the same R^2 to the compressed CBF vectors (0.97) when comparing the scores to π_t. Figure 3.4 shows that the cosine dissimilarity using the counting bloom filter is qualitatively similar to the cosine dissimilarity with the raw k-mer counts. This similarity and the high R^2 mean that the predictability of π_t is maintained while using the smaller, more manageable CBF data structure. Therefore, we can reliably use the CBF data structure as opposed to the raw k-mer counts to calculate the cosine dissimilarity of two samples.

The memory usage of a k-mers vector for a sample with 10x coverage and $k = 30$ under our simulations is around 4.5 MB per sample. The memory usage is reduced to 0.02 MB when

compressed to a 10,000 element array of unsigned 16-bit integers. Therefore, storing and using k-mer counts for measuring diversity scale much better under the CBF data structure.

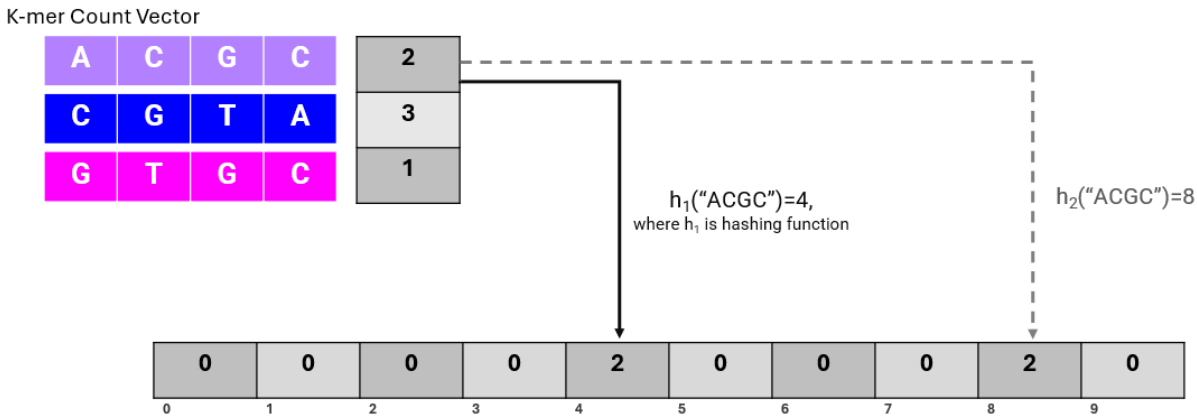

Figure 3.5: **A conceptual schematic of a Counting Bloom Filter.** When the first 4-mer, "ACGC", is hashed, the hash function h_1 outputs 4 and h_2 outputs 8. Both positions in the CBF array are incremented by 2 as that is the count of 'ACGC' in the given k-mer count vector.

The effect of array size on CBF measures of diversity

We found that the size of the CBF array scales the cosine dissimilarity score and keeps the rank of scores very similar as the array size changes. Since rank is maintained, we can take advantage of much smaller arrays without losing much information about the relative diversity scores between samples. Smaller arrays allow for less memory usage, and faster computing of cosine dissimilarity. Supplementary Figure D2 compares the scores of a CBF array of size 20 million and 10 thousand on the same simulations. We observe that the smaller array adjusts the scores down while maintaining the same rank of scores. For example, these results show that if a population has the lowest cosine score with the 20 million length array, it remains among the lowest scores with a 10 thousand length array. This result means that it may be possible to adjust the vector size such that the cosine dissimilarity scores are similar in magnitude and scale to the true pairwise nucleotide diversity which can be useful for prediction of the score without alignment, and can make interpretation simpler. However, the viability of this approach still needs to be investigated across different species and

genomic contexts, it is likely there is not a 'one size fits all' array size that would scale cosine dissimilarity to match π universally. Our simulations only consider SNPs, neutral evolution, and the *Arabidopsis thaliana* genome. It is still left to determine the effects of other mutations types such as insertions, deletions, and inversions as well as other evolution types. Therefore we should avoid the potential pitfalls of interpreting a scaled value directly as π unless future work shows that to be appropriate.

Explaining variation with k-mers

A common first step to understanding the evolutionary forces shaping variation in populations involves quantifying differentiation between populations. Patterns of differentiation in SNP genotypes are often summarized and plotted using dimension reduction techniques, such as principal component analysis (PCA) (Novembre and Stephens, 2008). k-mer genotypes are also amenable to PCA and recover the same differentiation patterns as SNP-based PCA (Liu et al., 2017; Murray et al., 2017; Rahman et al., 2018; Ho et al., 2019; Hrytsenko et al., 2022). Applying dimensional reduction to k-mers can even differentiate species (Rosen et al., 2008; Aflitos et al., 2015; Bernard et al., 2016; Linard et al., 2019; Boddé et al., 2022), something that can be difficult to do with SNPs without multiple sequence alignment. Interestingly, k-mers from repetitive sequences may not always differentiate populations that are clearly differentiated in terms of their SNP genotypes (Renny-Byfield and Baumgarten, 2020), suggesting that identifying the set of k-mers that best differentiate populations could be an important avenue for future research. Furthermore, while we are aware of one study in yeast that estimates admixture proportions from k-mers (Shi et al., 2024), more investigation is needed to see if this approach works for other species.

Besides polymorphism between populations, k-mers are also useful for explaining polymorphism patterns across genomes in terms of specific evolutionary forces. For instance, the k-mer profile of a given locus is often predictive of the local recombination rate (Liu et al., 2012; Haubold et al., 2013; Haubold, 2014; Frenkel et al., 2016; Al Maruf and Shatabda, 2019) and the local mutation rate (Aggarwala and Voight, 2016; Carlson et al., 2018; Bethune et al., 2022; Adams et al., 2023; Beichman et al., 2023; Liu and Samee, 2023). k-mers can capture information about re-

combination and mutation because these processes often associate with specific functional DNA motifs (Myers et al., 2008; Růžička et al., 2017). However, k-mers are also intrinsically sensitive to sequence changes and while there are some existing methods to identify new mutations (Nordström et al., 2013; Ho et al., 2019) and sites of recombination (Fletcher et al., 2021) solely from k-mers, further development is needed. Being able to predict fine scale variation in mutation and recombination rates solely from the k-mer profile of a reference sequence would be extremely beneficial because these processes frequently confound scans for sites of selection (Huber et al., 2016). However, this would require further study of the relationship between k-mers and mutation/recombination rates across wider ranges of species.

k-mers also offer opportunities to explore patterns of selection. One general assumption in k-mer literature is if a k-mer is shared across all individuals of a population or species then it is potentially under selection to be conserved (Bernard et al., 2016; Aylward et al., 2023). In contrast, a k-mer that is not present in any individual is possibly selected against (although this assumes that k is sufficiently small such that the total possible k-mer space is small, as might be true when analysing small amino acid-derived k-mers) (Georgakopoulos-Soares et al., 2021). However, it is unlikely that selection is solely responsible for patterns of k-mer sharing across individuals or species, so more rigorous approaches are needed. One potentially more rigorous approach would be to investigate k-mers that differentiate populations as candidates for loci underlying local adaptation, similar to F_{ST} (Kane and Rieseberg, 2007). Although, we are not aware of any papers that employ this specific approach, identifying group-specific k-mers is a very common and useful practice, with one example being the assembly of sex-specific sequences (Akagi et al., 2014; Ou et al., 2017; Liao et al., 2020; Neves et al., 2020; Mehrab et al., 2021; Wu et al., 2021; Behrens et al., 2022; Fong et al., 2023; Lichilín et al., 2023). There are some metrics of k-mer differentiation in published literature already (Rahman et al., 2018), but they do not have the same interpretation as F_{ST}. An alternative approach would be to march through a deBruijn graph (Aylward et al., 2023) to find strings of k-mers with selective sweep-like patterns, mainly low diversity, high linkage disequilibrium between k-mers, an excess of rare k-mers (Alachiotis and Pavlidis, 2018), high

differentiation between populations (Zhong et al., 2022), and high haplotype homozygosity (Klassmann and Gautier, 2022) which could be determined based on k-mer copy number (Vurture et al., 2017; Ranallo-Benavidez et al., 2020).

There are only a few model-oriented studies that compare the k-mer spectrum observed in a genome to a neutral expectation, mostly in the context of detecting selection on transcription factor binding motifs (Gerland and Hwa, 2002; Ke et al., 2008; Raijman et al., 2008; Yeang, 2010; Gyorgy, 2023). The basic idea behind these approaches is to first derive a neutral substitution model - which describes the probability of one nucleotide being substituted for another in a neutrally evolving sequence (Ke et al., 2008; Raijman et al., 2008) - then measure deviations from that neutral model according to the presence/absence of particular k-mers in a genome. Although these models are able to detect selection on k-mers that exist in multiple places across a genome, as is the case with binding motifs, it is unclear whether they could be applied to k-mers that represent unique genomic sequences.

Challenges in investigating k-mers

Despite the wide potential of k-mers to be useful for population genetic studies, there are three important limitations to keep in mind for most k-mer investigations. First, interpreting the biology of specific k-mers is often a challenge. One option is to take the candidate k-mers identified from an analysis and either align the k-mers themselves, the reads containing said k-mers, or assembled reads containing said k-mers to a reference genome (Voichek and Weigel, 2020) or a database of known sequences and motifs. While this is effective at pinpointing concrete type of variants at play in a system, it partially defeats the purpose of using k-mers in the first place because k-mers of interest may not be present in a reference genome. Second, since each sequencing error can generate up to k erroneous k-mers and general practice is to discard reads with many unique k-mers to reduce the effect of error rates (Zimin et al., 2013), analysis of k-mers typically requires datasets with high coverage ($>$10x) and low sequencing error rates. These criteria can potentially exclude long-read datasets (Vurture et al., 2017) or reduced-representation datasets where genomes are sequenced at lower coverage to cut costs. Approaches that apply k-mers in lower coverage or

higher error-rate datasets are needed. Third, k-mer-based analyses can frequently involve hundreds of millions or billions of unique k-mers. Storing and processing this many k-mers at once is often not feasible even on high performance computing systems. Two approaches to solving this issue are to subset one's k-mers (Ondov et al., 2016; Benoit et al., 2020; Zhao, 2019; Yi et al., 2021), or, as we discuss here, compress vectors of k-mer counts into an array of smaller size (Melsted and Pritchard, 2011). However, these solutions are usually developed for specific contexts, such as measuring genomic dissimilarity, and may not be applicable to every population genetic analysis. Future studies need to carefully consider common approaches to decreasing the disk burden of k-mer-based analyses and should explore new potential solutions.

These challenges to studying k-mers highlight how k-mer based methods should be viewed as complementary to, instead of better than, pangenomes in many situations. Among other benefits, pangenomes have ready biological interpretations and provide greater information on linkage between sequences. However, the fact that k-mers can be studied with fewer computing resources and simpler sequencing data sets makes them potentially useful for either preliminary pangenomic analyses or situations where constructing many reference quality genomes is not yet feasible.

Conclusions

Current literature demonstrates that k-mers are useful for identifying, measuring, and explaining variation within populations without performing alignment. Through our own simulations of neutrally evolving populations, we find that k-mer dissimilarity reliably scales with nucleotide diversity and k-mer matrices can be compressed with minimal loss of dissimilarity information. However, further development is needed to make k-mers amenable to a wider array of population genetic tasks, especially the identification of selected loci and population structure. Further developing alignment-free approaches to population genetics tasks will ultimately help guide and complement the analysis of pangenomes.

Future directions

- Can we develop criteria for choosing values of k that account for population-level variation in genome size or genome content?

- Can we develop k-mer-based estimates of population differentiation that are more similar to F_{ST}?
- Can we identify putative selective sweeps using k-mers?
- Can we develop applications of k-mers that are amenable to lower-coverage or pooled sequencing samples that are common in population genetics?
- k-mers are useful for getting preliminary estimates of important genome parameters that are useful for tuning pangenome assembly (Ranallo-Benavidez et al., 2020). What other types of preliminary pangenome analyses might k-mers be useful for?
- Can k-mers, either in combination with or without a pangenome reference, be useful for rapidly estimating allele frequencies and performing demographic inference?

Data availability

Supplemental figures can be found in Appendix D. Scripts for simulation and data analysis can be found at: https://github.com/williarj/kmers2024.

CHAPTER 4: SUMMARY STATISTICS COMPARABLE TO CONVOLUTIONAL NEURAL NETWORKS FOR INFERRING TIMES TO FIXATION[4]

Abstract

Times to fixation (t_f) are an essential ingredient for understanding neutral theory and positive selection. However, studies of t_f have largely subsided in recent years and few methods exist to predict the time to fixation for a fixed beneficial allele. A key difficulty in this prediction problem is disentangling t_f from the age of a sweep (t_a) because multiple different combinations of t_f and t_a can result in similar polymorphism patterns around the site of the sweep. We test whether convolutional neural networks (CNNs) can potentially perform better on this problem by simulating approximately 250,000 selective sweeps across five different demographic scenarios. The CNNs achieve comparable performance to approximate bayesian compuation methods used in previous approaches. The CNNs are better able to disentangle t_f from t_a for populations that are growing, but also give statistically worse predictions than ABC when trained on populations cycling between two different sizes. Altogether this suggests that perhaps few undiscovered signals remain in single timepoint unphased genotype data that can be used to disentangle t_f and t_a in singe population demographic scenarios.

Introduction

Adaptation requires increasing the frequency of beneficial alleles in a population, but these frequency changes take time. How long does it take for beneficial alleles to spread? This question was at the top of mind of naturalists in the early 20th century as there was doubt over how reasonable natural selection was as an evolutionary mechanism (see Charlesworth (2020) for a fuller historical treatment). The thinking was that if beneficial alleles change frequency only very slowly, then maybe natural selection is not important. However, early models outlined how selection can rapidly change allele frequencies, though most beneficial alleles are lost by drift Fisher (1923); Haldane (1927). Nonetheless the rate of new mutations is high enough such that new beneficial alleles are continually introduced and some eventually fix (Zhao et al., 2013). As it became appreciated that multiple fixations can build up between species, further investigations of fixation times were

[4] At the time of writing, this chapter is unpublished.

motivated by observations of molecular substitution rates (Kimura and Ohta, 1969) and trait degradation as a result of relaxed selection (Kimura, 1980). The theories from this foundational work have since been extended to understand how time to fixation is shaped by changing environments (Cui and Yuan, 2018; Kaushik and Jain, 2021) and population structure (Greven et al., 2016).

However, despite this foundation of knowledge, studies of times to fixation have subsided in recent years such that two different papers spaced 8 years apart both argue that times to fixation need more attention from modern methods (Zhao et al., 2013; Charlesworth, 2020). It is much more common for studies to estimate coalescence times (Y. C. Brandt et al., 2022), variant age (Bisschop et al., 2021), or allele frequency trajectories of currently segregating alleles (Stern et al., 2019). One reason for this focus is that coalescent times are foundational to important population genetic structures like ancestral-recombination graphs (Lewanski et al., 2024). Another reason is because t_f, in a simple model of a single additive *de novo* beneficial mutation, is tightly correlated to the selection coefficient. The mean time to fixation for a single additive beneficial mutation in a constant sized population is $2ln(2cN_e - 1)/s$ where c is ploidy level (Otto and Whitlock, 2013). This equation is sensitive to s more than N_e because s is outside of the logarithm. Thus, under this model one could study the selection coefficient and get essentially the same information as would be provided by t_f. However, the exact relationship between time to fixation and s is more complicated when arbitrary dominance, mating system variation, and demographic changes are considered (Glémin, 2012). Selection coefficients and coalescent times thus do not fully convey the timescale of fixation by themselves outside of simple scenarios.

How can we estimate t_f? Assuming a single *de novo* additive mutation that is sampled immediately after sweep completion, there are simple models describing how the dip in diversity around a selected site varies with t_f (Coop, 2020). There are more complex models requiring different sets of assumptions (He et al., 2020; Bisschop et al., 2021); however, not all study systems and situations will appropriately fit the assumptions in these models. As a result, there are several methods that forgo explicit models and instead simulate sweeps and use approximate bayesian computation (ABC) to estimate the timing of selection (Przeworski, 2003; Ormond et al., 2016; Nakagome et al.,

2019). Machine learning algorithms can also be trained on simulations and have proven useful for a variety of population genetics tasks (Kern and Schrider, 2018; Flagel et al., 2019; Torada et al., 2019; Sanchez et al., 2021; Whitehouse and Schrider, 2023), but have not yet been used to infer t_f.

Our goal is to use machine learning models that can infer t_f and are amenable to unphased genotype data. Our hypothesis is that machine learning models will outperform ABC models and could pick up on new signatures to differentiate between old, fast and young, slow sweeps (Hahn and Mishra, 2025).

Methods

Our entire analysis is available as a snakemake workflow, packaged here: https://github.com/milesroberts-123/selection-demography-cnn. This workflow includes all of the scripts used to simulate data, fit models, and generate figures as well as configuration files specifying the exact software versions used for each step.

Simulations

Our simulated selective sweeps are generated using SLiM (v4.0.1, Haller and Messer (2019)). The parameters for each simulation were generally drawn from uniform or log-uniform distributions. The exact distributions used for each parameter are listed in Table 4.1. Each simulation initializes a diploid population of size N_A individuals (i.e. the ancestral population size) and a chromosome of size L bp, then burns-in for $10N_A$ generations with mutation rate μ and recombination rate R. At κ generations post-burn-in, a sweep mutation is introduced at position $L/2$ bp (i.e. the middle of the simulated region). The beneficial mutation initially has a selection coefficient of 0, but once the frequency of the mutation reaches a value f_0 the selection coefficient is changed to a value s and given a dominance coefficient of h. After the frequency of the mutation reaches a value f_1, the sweep mutation is switched back to being neutral. After burn-in, the population either has a custom demography (specified as a vector of time steps and a corresponding vector of population sizes), or a demography determined by the following logistic map:

$$N_{t+1} = rN_t\left(1 - \frac{N_t}{K}\right) + N_t \tag{4.1}$$

where N_t gives the population size at generation t and $N_0 = N_A$. This equation has the following special properties (Phatak and Rao, 1995):

- $r = 0$, gives a constantly sized population.
- $0 < r < 2$ and $N_t < K$ gives a population that grows.
- $0 < r < 2$ and $N_t > K$ gives a decaying population.
- $2 < r < \sqrt{6}$ gives a population that cycles between two different values.
- $\sqrt{6} < r < 3$ gives a population that chaotically changes size. In practice, since our population size is a discrete number of individuals, the population can have a long cycle if happens to return to its initial population size of N_A.

From now on, we will refer to these respective demographies with the terms "constant", "growth", "decay", "cycling", or "chaotic", respectively. For growth or decay demographies, we restrict r to be $0 < r < 0.5$ to simulate gradual, rather than nearly instantaneous, population size changes. Because forward simulations slow down for larger populations, we also restrict K to be within 1 % - 50 % of N_A. For each simulation, we then track how long it takes for the beneficial mutation to fix. After fixation, we allow the simulation to run for an additional τ generations before taking a random sample of size n individuals from the population. If a sweep was lost from a population due to drift, we restarted the simulation to its post-burn-in state and initialized to a new random seed. If a simulation was restarted 1000 times and still did not result in a complete sweep, then we omitted that simulation from downstream analyses.

At the end of each simulation, we recorded the t_fs from the completed simulation and randomly sample n individuals (n = 128 for all analyses). We then downsampled the set of completed simulations to create a uniform distribution of $\log_{10}(t_f)$. Our downsampling procedure involved taking the initial $\log_{10}(t_f)$ distribution, dividing it up into bins of length 0.1, collecting the bin heights, and randomly sampling simulations from each bin (without replacement) until all bins were the same height. Our goal with this procedure was to include simulations with fixation times spanning approximately 50 - 20,000 generations while still retaining > 11000 simulations per demographic scenario. After downsampling, we randomly partitioned the remaining simulations to either the

training dataset (80 %), the validation dataset (10 %) or the testing dataset (10 %).

For our simulations, we kept the following parameters constant, but it is possible to configure different values in our workflow:

- L = 100 Kb
- κ = 1
- n = 128
- f_0 = 0
- f_1 = 1

In other words, we simulated only complete hard sweeps on chromosomes of size 100 Kb introduced 1 generation after burn-in and randomly sampled 128 individuals, which is a similar or larger sample size than most other sweep-related CNNs (Flagel et al., 2019; Torada et al., 2019; Whitehouse and Schrider, 2023).

Parameter	**constant**	**growth**	**decay**	**cycling**	**chaotic**
N_A	U(1000, 10000)	U(1000, 10000)	U(1000, 10000)	U(1000, 10000)	U(1000, 10000)
s	LU(log10(1/N_A), 0)	LU(log10(1/N_A), 0)	LU(log10(1/N_A), 0)	LU(log10(1/N_A), 0)	LU(log10(1/N_A), 0)
h	U(0,1)	U(0,1)	U(0,1)	U(0,1)	U(0,1)
μ	LU(-8.5, -7.5)	LU(-8.5, -7.5)	LU(-8.5, -7.5)	LU(-8.5, -7.5)	LU(-8.5, -7.5)
R	LU(-9, -7)	LU(-9, -7)	LU(-9, -7)	LU(-9, -7)	LU(-9, -7)
τ	LU(0, 4)	LU(0, 4)	LU(0, 4)	LU(0, 4)	LU(0, 4)
r	0	U(0,0.5)	U(0,0.5)	U(2, $\sqrt{6}$)	U($\sqrt{6}$, 3)
K	N_A	N_A×U(1.01, 2)	N_A×U(0.5, 0.99)	N_A×U(0.8, 1.2)	N_A×U(0.8, 1.2)

Table 4.1: **Sampling distributions for key simulation parameters for the five different demographic scenarios included in this study**. U is the uniform distribution and LU is the log uniform distribution. The first number for each distribution is the minimum and the second number is the maximum.

Selective sweep statistics

We calculated a suite of selective sweep summary statistics that were compatible with unphased genotype data and useful for selective sweep analysis (Kern and Schrider, 2018), including the number of segregating sites (S), nucleotide diversity (π), Watterson's theta (θ_W), Tajima's D (Tajima,

1989), variance in Tajima's D, number of unique genotypes (Kern and Schrider, 2018), unphased versions of h1; h2; h12; h123; h2h1 (Kern and Schrider, 2018), variance; skew; and kurtosis in the distribution of genotype mismatches (i.e. the g_{kl} statistic from Kern and Schrider (2018)), average genotypic correlation between pairs of SNPs (Rogers and Huff's R^2,Rogers and Huff (2009)), Kim's ω (Kim and Stephan, 2002), and unphased Messer's hscan (Schlamp et al., 2016).

Image-based sweep representation

We converted the VCF file output from each simulation into an image representation using a custom R script with ggplot (Wickham, 2016), ggnewscale (Campitelli, 2022), and cowplot (Wilke, 2020). This image representation is similar to (Flagel et al., 2019) but with some modifications for unphased genotype data. Each image was a gray-scale matrix with n rows corresponding to the sampled individuals and l columns corresponding to the l SNPs closest to the selective sweep site. We clustered the rows using according to their manhattan distance using the complete clustering algorithm in stats R package's hclust function. The color of each element in the matrix could be either black (homozygous genotype 0/0), grey (heterozygous genotype 0/1), or white (homozygous genotype 1/1). We allowed l to reach a maximum of 128 SNPs and padded each image with black columns if there were fewer than 128 SNPs in the sampled simulated population. Because each column of the image represents a SNP we also created a vector of SNP positions to go along with each image. These SNP positions were min-maxed normalized such that the SNP in the left-most column of an image was position 0 and the SNP at the right-most column of the image was position 1.

Convolutional Neural Networks

We used the image-based representations of genotypes in the regions of selective sweeps to train a convolutional neural network (CNN). Our CNN architecture was similar to Flagel et al. (2019), with two main branches: one to process a sweep image and another to process the vector of SNP positions associated with the columns of the sweep image. The image processing branch had three convolutional layers, each followed by a pooling and dropout layer. The final dropout layer was flattened and fed into a dense layer. The branch for processing the position information began with

an input layer with 128 neurons (one for each column in the image), followed by a dense layer and a dropout layer. The final dense layer for both the image-processing and position-processing branch were concatenated and fed into a final dense layer with dropout before being fed into a single output neuron.

We performed 60 iterations of bayesian hyperparameter tuning implemented in keras (v2.11.0, Chollet and others (2015)) on each CNN. We chose 60 iterations because a random parameter search with 60 iterations will sample a model in the top 5 % of the performance range with probability 95 % (Bergstra and Bengio, 2012). Each iteration involved 2 epochs of model fitting to the training simulations and the final performance of the iteration was measured as mean squared error on the validation simulations. During tuning, each convolutional layer could have any multiple of 16 filters between 16 and 128 and any layer could have a dropout rate between 0 and 0.99. All dense layers could have any multiple of 32 neurons between 32, and 512. The first convolutional layer was fixed to have a kernal size of 7 and a stride of 2, while the second and third convolutional layers had a kernal size of 3 and stride of 1. All activation functions were rectified linear unit (ReLU) functions (Banerjee et al., 2019).

After hyperparameter tuning, we then trained a final CNN with the best performing hyperparameters. Each training epoch fit the CNN in batches of 32 images. We trained the CNN until validation performance did not improve for 20 epochs. After 20 epochs of no improvements, we reverted the CNN to the configuration with the best performance on the validation data. This early stopping criterion is meant to avoid overfitting the model on the training data. Final model performance was assessed on the testing data and we performed 100 iterations of Monte-Carlo sampling on each testing image to an estimate the uncertainty for each prediction, which we quantified as the standard deviation in predictions across the Monte-Carlo samples. The mean of the predictions across all the Monte-Carlo samples was used as the point estiamte for the outcome variable (t_f) from the CNN.

Dense Neural Networks

We also constructed dense neural networks (DNNs) that were trained on only summary statistics. Our DNNs had a input layer of 17 neurons (one for each summary statistic) followed by three dense layers with dropout. Similar to the training of CNNs, we did 60 iterations of hyperparmater tuning using bayesian hyperparameter optimization. During tuning, each of the dense layers could have a multiple of 8 neurons anywhere from 16 to 512 and a dropout rate anywhere from 0 to 0.99. Final model performance was assessed on the testing data and we performed 100 iterations of Monte-Carlo sampling on each testing image. The mean of the predictions across the Monte-Carlo samples was used as our point estimate for the outcome variable (t_f).

Approximate Bayesian Computation

We also performed ABC to estimate t_f using the rabc package (v2.2.1, Csilléry et al. (2012)) in R (R Core Team, 2022). Just like tuning hyperparameters of CNNs, performing ABC requires making several choices, including the choice of regression method, the tolerance level, and the type of point estimate used (i.e mean, median, or mode of the posterior distribution). Thus, similar to ML, it is typically desirable to try many hyperparameter choices and use the combination of hyperparameters with the best performance for final predictions. ABC approaches are typically tuned with a cross-validation procedure: randomly drop a simulation from the training set and then use the remaining training set to make a prediction for the dropped-out example, repeating many times Csilléry et al. (2012). However, since part of our hypothesis hinged on comparing the performance of ABC to CNNs and DNNs, we tuned ABC similarly to CNNs and DNNs to promote comparability.

For each set of simulations, we tried 60 different configurations of ABC (similar to doing 60 iterations of hyperparameter tuning for CNNs and DNNs), including 4 different methods (rejection, ridge regression, local linear regression, or neural network), 5 different tolerance values (0.025, 0.05, 0.1, 0.15, or 0.2), and three different point estimates (mean, median, or mode). For each model configuration, we predicted t_f for each simulation in the validation set using the training set as the baseline. We quantified the performance of the model as the Pearson correlation between

the model's predictions and the true value. For the best performing ABC model, we then measured final performance by making predictions for the testing dataset, using the training dataset as the background.

Partial R^2 calculations

Another aspect we were interested in was the benefit of adding more selective sweep statistics to explain variation in $t_f + t_a$ because many studies that scan for sweeps simply focus on a handful of site frequency spectrum (SFS) statistics like π and Tajima's D (Teshima et al., 2006). To investigate this, we calculated partial R^2 values for the following statistics: S, $log_{10}(\pi)$, $log_{10}(\theta_W)$, Tajima's D, Variance in Tajima's D, and the number of unique genotypes. These partial R^2 values can be interpreted as the proportion of variation not explain by these statistics that becomes explained by our remaining statistics.

The procedure to calculate partial R^2 is to first fit a ordinary least squares linear model, which we will call the full model:

$$log_{10}(t_f + t_a) \sim \text{S} + log_{10}(\pi) + log_{10}(\theta_W) + \text{D} + \text{Var[D]} + \text{number of unique genotypes} + \text{h1} + \text{h2} + \text{h12} + \text{h123} + \text{h2h1} + g_{kl} \text{ variance} + g_{kl} \text{ skew} + g_{kl} \text{ kurt} + \text{Messer's hscan} + \text{Rogers and Huff's } R^2 + log_{10}(\text{Kim's } \omega)$$

Then, fit a second model which we will call the reduced model:

$$log_{10}(t_f + t_a) \sim \text{S} + log_{10}(\pi) + log_{10}(\theta_W) + \text{D} + \text{Var[D]} + \text{number of unique genotypes}$$

Given the full and reduced model, the partial R^2 measuring the added predictive power of including more than just the SFS statistics is:

$$R^2_{partial} = \frac{SSE(reduced) - SSE(full)}{SSE(reduced)} \tag{4.2}$$

where SSE is the sum of squared errors.

Results

Differentiating old, fast sweeps from slow, young sweeps

We performed a total of 250,000 SLiM simulations across 5 different demographic scenarios (50,000 simulations per scenario), more than 79% of which successfully produced hard sweeps with valid values of all 17 selective sweep statistics. For each of these simulations, we recorded t_f and t_a for the sweep (Figure 4.1A). Most of these statistics produced clear, monotonic correlations with $t_f + t_a$ (Figure 4.1B, Figure E4-E7). However, for all of these statistics, there were divergent combinations of t_f and t_a that produced similar values of the selective sweep statistics. For example, sweeps that were old ($t_a > 1000$) and fast ($t_f < 100$) generally had similar π values to sweeps that were younger and slower ($t_f > 1000$, Figure 4.1C).

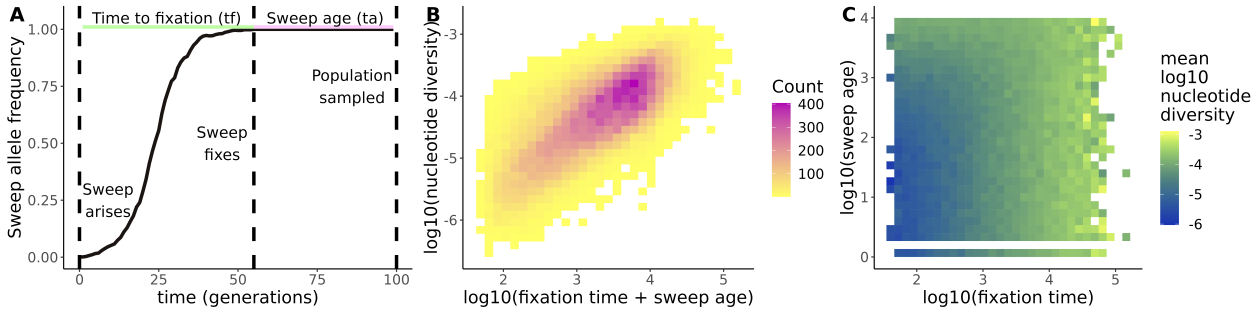

Figure 4.1: **Separately estimating t_f and t_a from summary statistics leads to unidentifiability.** (A) Definition of time to fixation and sweep age. Time to fixation is the time between when a sweep first arises and when it fixes. Sweep age is the time between the sweep fixes and when the population is sampled. Example allele frequency trajectory comes from simulating a beneficial allele with selection coefficient 0.5, dominance 0.5, and initial frequency of 1/1000 in a Wright-Fisher population of size 1000. (B) Nucleotide diversity in the sweep region linearly scales with the total lifetime of the sweep: $t_f + t_a$. (C) Young, slow sweeps (high t_f and low t_a) have similar nucleotide diversity as old, fast sweeps (low t_f and high t_a).

Partial R^2 of site frequency spectra statistics

Across the selective sweep statistics, the site frequency spectra statistics were generally positively correlated with each other and negatively correlated with the haplotype frequency statistics (Figure 4.2A). The two statistics related to linkage disequilibrium, Rogers and Huff's R^2 and Kim's ω, were generally not correlated with any of the other statistics, except for being slightly correlated to each other (Figure 4.2A). Across the 5 different demographic models, the partial R^2 (i.e. the

added predictive power) of including statistics beyond simple SFS statistics was significant, with values > 0.2 (Figure 4.2B). The growth demographic model had slightly higher partial R^2 (0.26) than the other scenarios and the chaotic demographic model had the lowest (partial $R^2 = 0.20$).

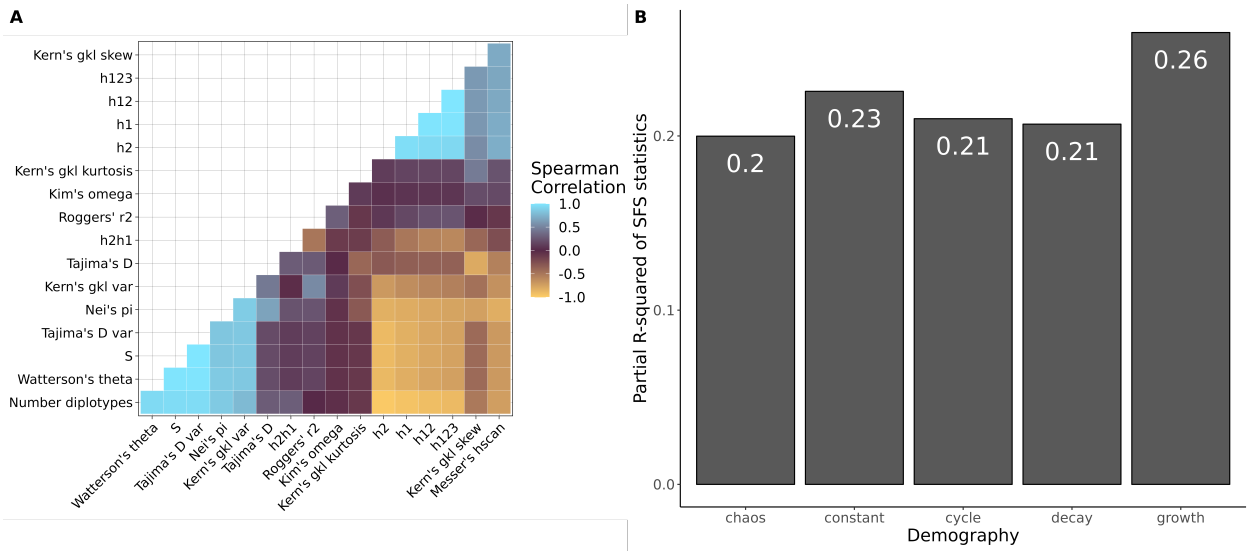

Figure 4.2: **Including more than site frequency spectrum statistics explains more variation in $t_f + t_a$.** (A) Pairwise Spearman correlations between unphased selective sweep statistics in the constant demography scenario. (B) Partial-R-squared values for a model with S, $log_{10}(\pi)$, $log_{10}(\theta_W)$, Tajima's D, Tajima's D variance, and number of diplotypes. In other words, these values describe the amount of additional variation in $t_f + t_a$ that explained when more than just these selective sweep statistics are added to the model.

Convolutional neural networks vs summary statistics for predicting t_f

For each of the 5 demographic scenarios, we built (1) Approximate Bayesian Computation, (2) dense neural network, and (3) convolutional neural network models to predict t_f. The pearson correlation between the true value of t_f and the predicted value of t_f for the best performing version of each model was generally > 0.7 (Figure 4.3, 4.4, E1-E3). By this measure of performance, the three types of models were not significantly different for most of the demographic scenarios tested (e.g. 95 % confidence intervals for constant demography: CNN = [0.705, 0.750], DNN = [0.719, 0.762], ABC = [0.731, 0.773]; Figure 4.3) with the exception of the cycling demography where the CNN performed worse (r = 0.656 (CNN) vs 0.728 (DNN) vs 0.74 (ABC), Figure E3). All three types of models also generally produced less accurate predictions for sweeps with a short t_f, but

$t_a > 1000$, incorrectly predicting that these sweeps had a much longer t_f (Figure 4.3). The one exception was the growth demography scenario where the CNN and DNN appeared to confuse old, fast and slow, young sweeps less often (Figure 4.4).

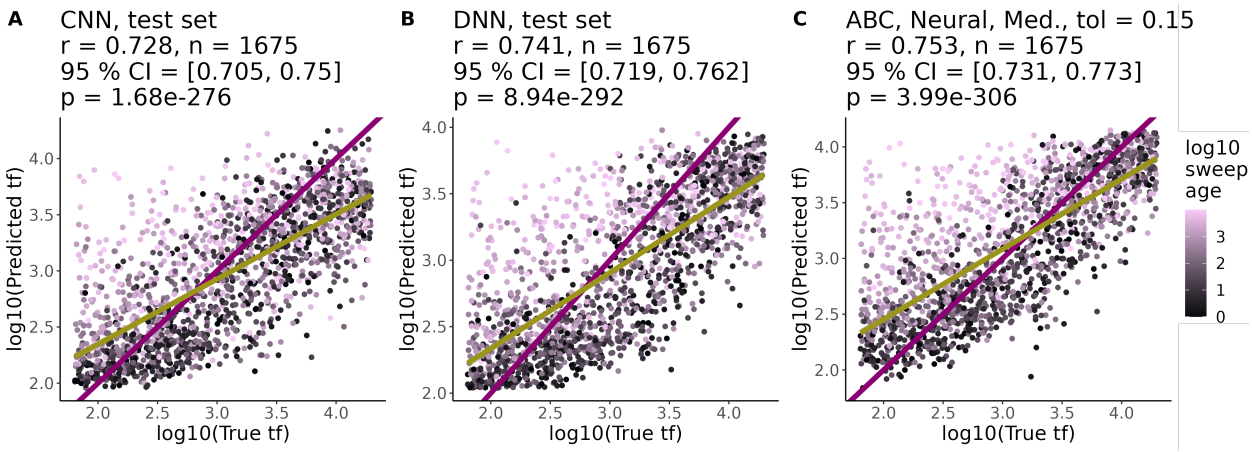

Figure 4.3: **CNNs perform similarly to DNNs and ABC in a constantly sized population.** Performance of best (A) CNN, (B) DNN, and (C) ABC models at predicting t_f under a constant demographic scenario. For each model, we list Pearson correlation coefficients between predicted and true values (r), the number of test set examples (n), the 95 % confidence intervals for r, and the p-value for testing if r is different from 0 (p). The purple line is the 1-1 reference line where predictions and truth are equal. The yellow line is a least squares regression line. All points are colored according to the age of the selective sweep in generations (log10 transformed).

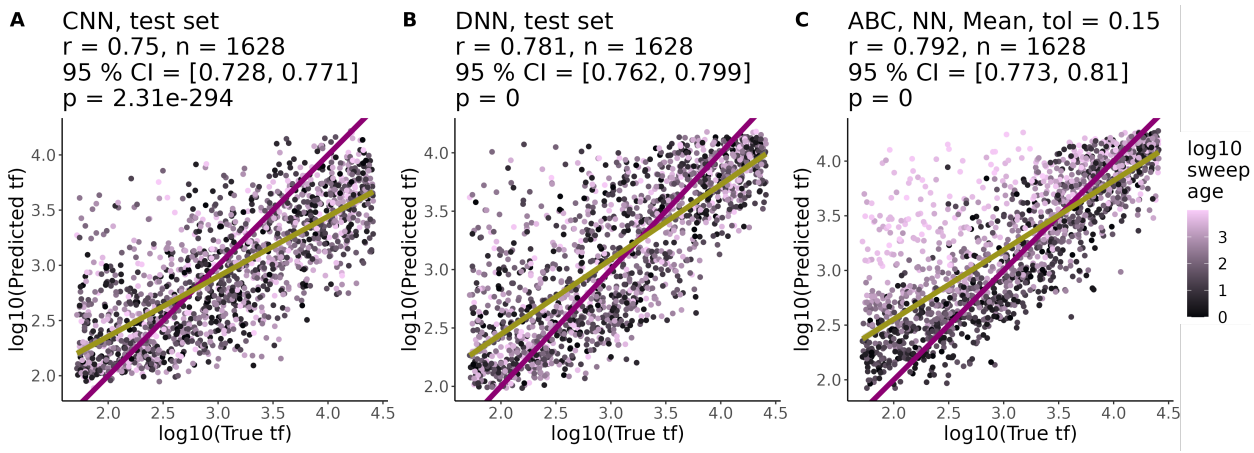

Figure 4.4: **CNNs and DNNs and ABC in a growing population.** Performance of best (A) CNN, (B) DNN, and (C) ABC models at predicting t_f under a constant demographic scenario. For each model, we list Pearson correlation coefficients between predicted and true values (r), the number of test set examples (n), the 95 % confidence intervals for r, and the p-value for testing if r is different from 0 (p). The purple line is the 1-1 reference line where predictions and truth are equal. The yellow line is a least squares regression line. All points are colored according to the age of the selective sweep in generations (log10 transformed).

Discussion

We simulated approximately 250,000 selective sweeps across a range of parameters and demographic scenarios, then used those simulations to train CNNs, DNNs, and ABC models to predict t_f. Our hypothesis was that CNNs could potentially better disentangle t_f from t_a and thus be better at predicting t_f. Our simulations reflected a core challenge in the problem of distinguishing t_f from t_a: sweeps with high t_f and low t_a (i.e. slow, young sweeps) leave similar signatures as sweeps with low t_f and high t_a (i.e. fast, old sweeps, Figure 4.1C). Each of the 17 selective sweep statistics we calculated explained additional variation in $t_f + t_a$ (Figure 4.2). In the end, our models trained on summary statistics (DNN and ABC) performed similarly to CNNs, which were trained on images of genotype matrices (Figure 4.3, 4.4, E1-E3).

Similar to previous studies on estimating coalescent times (Y. C. Brandt et al., 2022), we noticed a general bias where models overestimate low values of t_f and underestimate high values of t_f (Figure 4.3, 4.4, E1-E3). This is mostly likely due to a general identifiability problem where old, fast sweeps leave similar signatures as young, slow sweeps (Figure 4.1, E4-E7). However, in the

growth demography scenario CNNs and DNNs were apparently better able to disentangle t_f and t_a (Figure 4.4A and 4.4B) compared to ABC (Figure 4.4 C). This is possibly because two important signatures of sweeps, an excess of common alleles in site frequency spectra and excessive linkage, persist for longer in populations with higher N_e (Przeworski, 2002). Thus, the signatures of a short t_f persist for longer and prevent (or at least delay) confusion of old, fast sweeps with slow, young sweeps. The second case where ABC and CNNs diverged was the cyclic demography scenario, where the correlation between true and predicted t_f was lower for the CNN than the DNN and ABC, even though the DNN and ABC maintained a similar correlation compared to the constant demography scenario (Figure E2). This is potentially because repeated population bottlenecks have several consequences for the dynamics of sweeps (Wilson et al., 2014). In our case, if population bottlenecks are softening the sweeps, then it's possible that the CNN has trouble making predictions for this type of sweep compared to DNNs and ABC which are fed statistics specifically tuned for soft sweeps (H2H1, H123, etc.).

Summary statistics were overall effective at predicting t_f from unphased genotype data (Figure 4.3B-C, 4.4B-C). Our hypothesis was that CNNs could potentially identify new signatures besides known summary statistics and thus better predict t_f (Hahn and Mishra, 2025). However, the fact that ABC based on summary statistics performs the same or slightly better than a CNN suggests unphased genotype matrices do not contain signals for t_f that are not already captured by summary statistics. Many CNNs that classify selective sweeps use summary statistics partly to ease interpretability (Kern and Schrider, 2018; Lauterbur et al., 2023), but previous studies have also show CNNs trained on genotype matrices largely learn to reproduce summary statistics (Cecil and Sugden, 2023). It is still possible though that additional signatures of t_f not already described by summary statistics do arise in other demographies not explored in this study.

Two general directions forward for improving t_f predictions are to tweak the prediction algorithms or include additional sources of data. Inclusion of time series data (Whitehouse and Schrider, 2023) or phasing information could conceivably aid t_f prediction, but these are limited to cases where the researchers have ancient samples or an understanding of a recombination landscape.

Another option would be to include data from more than one population (Yair et al., 2021). Population differentiation provides information on the split time between populations and if a sweep is shared across populations, that would suggest that the sweep happened before the population split (Yair et al., 2021). The difficulty here is that one needs to know or derive a model of the population split and admixture beforehand so that it can be properly simulated. A combination of both including additional data types and tuning models will most likely be the fruitful path forward.

Data availability

Supplemental figures are in Appendix E. Our entire analysis is available as a snakemake workflow, packaged here: https://github.com/milesroberts-123/selection-demography-cnn.

BIBLIOGRAPHY

Aagaard, J. E., Willis, J. H., and Phillips, P. C. (2006). Relaxed selection among duplicate floral regulatory genes in Lamiales. *Journal of Molecular Evolution*, 63(4):493–503.

Adams, C. J., Conery, M., Auerbach, B. J., Jensen, S. T., Mathieson, I., and Voight, B. F. (2023). Regularized sequence-context mutational trees capture variation in mutation rates across the human genome. *PLOS Genetics*, 19(7):e1010807. Publisher: Public Library of Science.

Aflitos, S. A., Severing, E., Sanchez-Perez, G., Peters, S., de Jong, H., and de Ridder, D. (2015). Cnidaria: fast, reference-free clustering of raw and assembled genome and transcriptome NGS data. *BMC Bioinformatics*, 16(1):352.

Aggarwala, V. and Voight, B. F. (2016). An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature Genetics*, 48(4):349–355. Number: 4 Publisher: Nature Publishing Group.

Akagi, T., Henry, I. M., Tao, R., and Comai, L. (2014). A Y-chromosome–encoded small RNA acts as a sex determinant in persimmons. *Science*, 346(6209):646–650. Publisher: American Association for the Advancement of Science Section: Report.

Al Maruf, M. A. and Shatabda, S. (2019). iRSpot-SF: Prediction of recombination hotspots by incorporating sequence based features into Chou's Pseudo components. *Genomics*, 111(4):966–972.

Alachiotis, N. and Pavlidis, P. (2018). RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Communications Biology*, 1(1):1–11. Publisher: Nature Publishing Group.

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezwaan, T. M., Ding, W., Ecker, J. R., Exposito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D. G., Hancock, A. M., Henz, S. R., Holm, S., Horton, M., Jarsulic, M., Kerstetter, R. A., Korte, A., Korte, P., Lanz, C., Lee, C.-R., Meng, D., Michael, T. P., Mott, R., Muliyati, N. W., Nägele, T., Nagler, M., Nizhynska, V., Nordborg, M., Novikova, P. Y., Picó, F. X., Platzer, A., Rabanal, F. A., Rodriguez, A., Rowan, B. A., Salomé, P. A., Schmid, K. J., Schmitz, R. J., Seren, □., Sperone, F. G., Sudkamp, M., Svardal, H., Tanzer, M. M., Todd, D., Volchenboum, S. L., Wang, C., Wang, G., Wang, X., Weckwerth, W., Weigel, D., and Zhou, X. (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*, 166(2):481–491. Publisher: Elsevier.

Alvarez-Ponce, D. and Fares, M. A. (2012). Evolutionary rate and duplicability in the *Arabidopsis thaliana* protein–protein interaction network. *Genome Biology and Evolution*, 4(12):1263–1274.

Alvarez-Ponce, D., Feyertag, F., and Chakraborty, S. (2017). Position matters: network centrality considerably impacts rates of protein evolution in the human protein–protein interaction network. *Genome Biology and Evolution*, 9(6):1742–1756.

Asif, H., Alliey-Rodriguez, N., Keedy, S., Tamminga, C. A., Sweeney, J. A., Pearlson, G., Clementz, B. A., Keshavan, M. S., Buckley, P., Liu, C., Neale, B., and Gershon, E. S. (2021). GWAS significance thresholds for deep phenotyping studies can depend upon minor allele frequencies and sample size. *Molecular Psychiatry*, 26(6):2048–2055. Publisher: Nature Publishing Group.

Audano, P. A., Ravishankar, S., and Vannberg, F. O. (2018). Mapping-free variant calling using haplotype reconstruction from k-mer frequencies. *Bioinformatics*, 34(10):1659–1665.

Auld, J. R., Agrawal, A. A., and Relyea, R. A. (2010). Re-evaluating the costs and limits of adaptive phenotypic plasticity. *Proceedings of the Royal Society B: Biological Sciences*, 277(1681):503–511. Publisher: Royal Society.

Aylward, A. J., Petrus, S., Mamerto, A., Hartwick, N. T., and Michael, T. P. (2023). PanKmer: k-mer-based and reference-free pangenome analysis. *Bioinformatics*, 39(10):btad621.

Bai, X., Tang, K., Ren, J., Waterman, M., and Sun, F. (2017). Optimal choice of word length when comparing two Markov sequences using a chi-squared statistic. *BMC Genomics*, 18(6):732.

Banerjee, C., Mukherjee, T., and Pasiliao, E. (2019). An Empirical Study on Generalizations of the ReLU Activation Function. In *Proceedings of the 2019 ACM Southeast Conference*, pages 164–167, Kennesaw GA USA. ACM.

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschumar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., Quinto-Cortés, C. D., Rodrigues, M. F., Saunack, K., Sellinger, T., Thornton, K., van Kemenade, H., Wohns, A. W., Wong, Y., Gravel, S., Kern, A. D., Koskela, J., Ralph, P. L., and Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220(3):iyab229.

Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., and Edwards, D. (2020). Plant pan-genomes are the new reference. *Nature Plants*, 6(8):914–920. Publisher: Nature Publishing Group.

Bazin, E., Glémin, S., and Galtier, N. (2006). Population Size Does Not Influence Mitochondrial Genetic Diversity in Animals. *Science*, 312(5773):570–572. Publisher: American Association for the Advancement of Science.

Becher, H., Sampson, J., and Twyford, A. D. (2022). Measuring the invisible: The sequences causal of genome size differences in eyebrights (Euphrasia) revealed by *k*-mers. *Frontiers in Plant Science*, 13.

Bedo, J., Goudey, B., Wazny, J., and Zhou, Z. (2016). Information theoretic alignment free variant calling. *PeerJ Computer Science*, 2:e71. Publisher: PeerJ Inc.

Behrens, K. A., Koblmüller, S., and Kocher, T. D. (2022). Sex chromosomes in the tribe Cyprichromini (Teleostei: Cichlidae) of Lake Tanganyika. *Scientific Reports*, 12(1):17998. Number: 1 Publisher: Nature Publishing Group.

Beichman, A. C., Robinson, J., Lin, M., Moreno-Estrada, A., Nigenda-Morales, S., and Harris, K. (2023). Evolution of the mutation spectrum across a mammalian phylogeny. *Molecular Biology and Evolution*, 40(10):msad213.

Belyayev, A. (2014). Bursts of transposable elements as an evolutionary driving force. *Journal of Evolutionary Biology*, 27(12):2573–2584.

Benoit, G., Mariadassou, M., Robin, S., Schbath, S., Peterlongo, P., and Lemaitre, C. (2020). SimkaMin: fast and resource frugal de novo comparative metagenomics. *Bioinformatics*, 36(4):1275–1276.

Benoit, G., Peterlongo, P., Mariadassou, M., Drezen, E., Schbath, S., Lavenier, D., and Lemaitre, C. (2016). Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Computer Science*, 2:e94. Publisher: PeerJ Inc.

Bergeron, L. A., Besenbacher, S., Zheng, J., Li, P., Bertelsen, M. F., Quintard, B., Hoffman, J. I., Li, Z., St. Leger, J., Shao, C., Stiller, J., Gilbert, M. T. P., Schierup, M. H., and Zhang, G. (2023). Evolution of the germline mutation rate across vertebrates. *Nature*, 615(7951):285–291. Publisher: Nature Publishing Group.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13(null):281–305.

Bernard, G., Ragan, M. A., and Chan, C. X. (2016). Recapitulating phylogenies using k-mers: from trees to networks. *F1000Research*, 5:2789.

Betancourt, A. J. and Presgraves, D. C. (2002). Linkage limits the power of natural selection in *Drosophila*. *Proceedings of the National Academy of Sciences*, 99(21):13616–13620. Publisher: Proceedings of the National Academy of Sciences.

Bethune, J., Kleppe, A., and Besenbacher, S. (2022). A method to build extended sequence context models of point mutations and indels. *Nature Communications*, 13(1):7884. Number: 1 Publisher: Nature Publishing Group.

Betschart, R. O., Thiéry, A., Aguilera-Garcia, D., Zoche, M., Moch, H., Twerenbold, R., Zeller, T., Blankenberg, S., and Ziegler, A. (2022). Comparison of calling pipelines for whole genome sequencing: an empirical study demonstrating the importance of mapping and alignment. *Scientific Reports*, 12(1):21502. Publisher: Nature Publishing Group.

Bisschop, G., Lohse, K., and Setter, D. (2021). Sweeps in time: leveraging the joint distribution of branch lengths. *Genetics*, 219(2):iyab119.

Blanca, A., Harris, R. S., Koslicki, D., and Medvedev, P. (2022). The statistics of *k*-mers from a sequence undergoing a simple mutation process without spurious matches. *Journal of Computational Biology*, 29(2):155–168. Publisher: Mary Ann Liebert, Inc., publishers.

Blomberg, S. P., Lefevre, J. G., Wells, J. A., and Waterhouse, M. (2012). Independent contrasts and PGLS regression estimators are equivalent. *Systematic Biology*, 61(3):382–391.

Boddé, M., Makunin, A., Ayala, D., Bouafou, L., Diabaté, A., Ekpo, U. F., Kientega, M., Le Goff, G., Makanga, B. K., Ngangue, M. F., Omitola, O. O., Rahola, N., Tripet, F., Durbin, R., and Lawniczak, M. K. (2022). High-resolution species assignment of *Anopheles* mosquitoes using *k*-mer distances on targeted sequences. *eLife*, 11:e78775. Publisher: eLife Sciences Publications, Ltd.

Braam, J. and Davis, R. W. (1990). Rain-, wind-, and touch-induced expression of calmodulin and calmodulin-related genes in Arabidopsis. *Cell*, 60(3):357–364.

Branca, A., Paape, T. D., Zhou, P., Briskine, R., Farmer, A. D., Mudge, J., Bharti, A. K., Woodward, J. E., May, G. D., Gentzbittel, L., Ben, C., Denny, R., Sadowsky, M. J., Ronfort, J., Bataillon, T., Young, N. D., and Tiffin, P. (2011). 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. Publisher: Proceedings of the National Academy of Sciences.

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

Buffalo, V. (2021). Quantifying the relationship between genetic diversity and population size suggests natural selection cannot explain Lewontin's Paradox. *eLife*, 10:e67509. Publisher: eLife Sciences Publications, Ltd.

Buffalo, V. and Coop, G. (2020). Estimating the genome-wide contribution of selection to temporal allele frequency change. *Proceedings of the National Academy of Sciences*, 117(34):20672–20680. Publisher: Proceedings of the National Academy of Sciences.

Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., Rong, Z., Wang, B., Xu, D., Yang, B., Xie, C., Fan, L., Gao, S., Xu, X., Zhang, G., Li, Y., Jiao, Y., Doebley, J. F., Ross-Ibarra, J., Lorant, A., Buffalo, V., Romay, M. C., Buckler, E. S., Ware, D., Lai, J., Sun, Q., and Xu, Y. (2017). Construction of the third-generation *Zea mays* haplotype map. *GigaScience*, 7(4).

Bush, S. J., Foster, D., Eyre, D. W., Clark, E. L., De Maio, N., Shaw, L. P., Stoesser, N., Peto, T. E. A., Crook, D. W., and Walker, A. S. (2020). Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines. *GigaScience*, 9(2):giaa007.

Bush, S. J., Kover, P. X., and Urrutia, A. O. (2015). Lineage-specific sequence evolution and exon edge conservation partially explain the relationship between evolutionary rate and expression level in *A. thaliana*. *Molecular Ecology*, 24(12):3093–3106. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.13221.

Bussi, Y., Kapon, R., and Reich, Z. (2021). Large-scale k-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy. *PLOS ONE*, 16(10):e0258693. Publisher: Public Library of Science.

Caetano-Anolles, D. (2023). Hard-filtering germline short variants.

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

Campitelli, E. (2022). ggnewscale: Multiple Fill and Colour Scales in 'ggplot2'.

Carlson, J., Locke, A. E., Flickinger, M., Zawistowski, M., Levy, S., Myers, R. M., Boehnke, M., Kang, H. M., Scott, L. J., Li, J. Z., and Zöllner, S. (2018). Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nature Communications*, 9(1):3753. Number: 1 Publisher: Nature Publishing Group.

Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V., and Kondrashov, F. A. (2002). Selection for short introns in highly expressed genes. *Nature Genetics*, 31(4):415–418. Number: 4 Publisher: Nature Publishing Group.

Cecil, R. M. and Sugden, L. A. (2023). On convolutional neural networks for selection inference: revealing the lurking role of preprocessing, and the surprising effectiveness of summary statistics. Pages: 2023.02.26.530156 Section: New Results.

Chamberlain, S. A. and Boettiger, C. (2017). R Python, and Ruby clients for GBIF species occurrence data. Technical Report e3304v1, PeerJ Inc. ISSN: 2167-9843.

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

Charlesworth, B. (2020). How Long Does It Take to Fix a Favorable Mutation, and Why Should We Care? *The American Naturalist*, 195(5):753–771. Publisher: The University of Chicago Press.

Charlesworth, B. and Jensen, J. D. (2022). How can we resolve Lewontin's Paradox? *Genome Biology and Evolution*, 14(7):evac096.

Charlesworth, B., Morgan, M. T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303.

Chen, J., Glémin, S., and Lascoux, M. (2017). Genetic diversity and the efficacy of purifying selection across plant and animal species. *Molecular Biology and Evolution*, 34(6):1417–1428.

Chen, N.-C., Solomon, B., Mun, T., Iyer, S., and Langmead, B. (2021). Reference flow: reducing reference bias using multiple population genomes. *Genome Biology*, 22(1):8.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890.

Cheng, C.-Y., Krishnakumar, V., Chan, A. P., Thibaud□Nissen, F., Schobel, S., and Town, C. D. (2017). Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *The Plant Journal*, 89(4):789–804. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tpj.13415.

Chikhi, R. and Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, 30(1):31–37.

Chin, C.-S., Behera, S., Khalak, A., Sedlazeck, F. J., Sudmant, P. H., Wagner, J., and Zook, J. M. (2023). Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes. *Nature Methods*, 20(8):1213–1221. Number: 8 Publisher: Nature Publishing Group.

Choi, I., Ponsero, A. J., Bomhoff, M., Youens-Clark, K., Hartman, J. H., and Hurwitz, B. L. (2019). Libra: scalable k-mer–based tool for massive all-vs-all metagenome comparisons. *GigaScience*, 8(2):giy165.

Chollet, F. and others (2015). Keras. Published: urlhttpskeras.io.

Chu, J., Rong, J., Feng, X., and Li, H. (2024). ntsm: an alignment-free, ultra-low-coverage, sequencing technology agnostic, intraspecies sample comparison tool for sample swap detection. *GigaScience*, 13:giae024.

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., and Lu, X. (2012a). Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Frontiers in Genetics*, 3. Publisher: Frontiers Media SA.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012b). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, 6(2):80–92.

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

Colombo, M., Laayouni, H., Invergo, B. M., Bertranpetit, J., and Montanucci, L. (2014). Metabolic flux is a determinant of the evolutionary rates of enzyme-encoding genes. *Evolution*, 68(2):605–613. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/evo.12262.

Compeau, P. E. C., Pevzner, P. A., and Tesler, G. (2011). Why are de Bruijn graphs useful for genome assembly? *Nature biotechnology*, 29(11):987–991.

Coop, G. (2016). Does linked selection explain the narrow range of genetic diversity across species? Pages: 042598 Section: Contradictory Results.

Coop, G. (2020). *Population and Quantitative Genetics*. github, 3 edition.

Cooper, L., Meier, A., Laporte, M.-A., Elser, J. L., Mungall, C., Sinn, B. T., Cavaliere, D., Carbon, S., Dunn, N. A., Smith, B., Qu, B., Preece, J., Zhang, E., Todorovic, S., Gkoutos, G., Doonan, J. H., Stevenson, D. W., Arnaud, E., and Jaiswal, P. (2018). The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Research*, 46(D1):D1168–D1180.

Corbett-Detig, R. B., Hartl, D. L., and Sackton, T. B. (2015). Natural selection constrains neutral diversity across a wide range of species. *PLOS Biology*, 13(4):e1002112. Publisher: Public Library of Science.

Cornet, L. and Baurain, D. (2022). Contamination detection in genomic data: more is not enough. *Genome Biology*, 23(1):60.

Cornish, A. and Guda, C. (2015). A comparison of variant calling pipelines using genome in a bottle as a reference. *BioMed Research International*, 2015(1):456479. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2015/456479.

Crameri, F. (2018). Geodynamic diagnostics, scientific visualisation and StagLab 3.0. *Geoscientific Model Development*, 11(6):2541–2562. Publisher: Copernicus GmbH.

Csilléry, K., François, O., and Blum, M. G. B. (2012). abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, 3(3):475–479.

Cui, F. and Yuan, B. (2018). Fixation probability of a beneficial mutation conferring decreased generation time in changing environments. *BMC Systems Biology*, 12(S4):48.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and Group, . G. P. A. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158. Publisher: Oxford Academic.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):giab008.

Darwin, C. (1859). *On the Origin of Species*. Simon & Schuster. Num Pages: 9.

Davis, J. C. and Petrov, D. A. (2004). Preferential duplication of conserved proteins in eukaryotic genomes. *PLOS Biology*, 2(3):e55. Publisher: Public Library of Science.

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

Denti, L., Previtali, M., Bernardini, G., Schönhuth, A., and Bonizzoni, P. (2019). MALVA: Genotyping by Mapping-free ALlele detection of known VAriants. *iScience*, 18:20–27.

Deorowicz, S., Kokot, M., Grabowski, S., and Debudaj-Grabysz, A. (2015). KMC 2: fast and resource-frugal k-mer counting. *Bioinformatics*, 31(10):1569–1576. Publisher: Oxford Academic.

DeWitt, T. J., Sih, A., and Wilson, D. S. (1998). Costs and limits of phenotypic plasticity. *Trends in Ecology & Evolution*, 13(2):77–81.

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

Drummond, D. A., Raval, A., and Wilke, C. O. (2006). A single determinant dominates the rate of yeast protein evolution. *Molecular Biology and Evolution*, 23(2):327–337.

Dubinkina, V. B., Ischenko, D. S., Ulyantsev, V. I., Tyakht, A. V., and Alexeev, D. G. (2016). Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC Bioinformatics*, 17(1):38.

Durai, D. A. and Schulz, M. H. (2016). Informed kmer selection for de novo transcriptome assembly. *Bioinformatics*, 32(11):1670–1677.

Duret, L. and Mouchiroud, D. (2000). Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Molecular Biology and Evolution*, 17(1):68–070.

Ebler, J., Ebert, P., Clarke, W. E., Rausch, T., Audano, P. A., Houwaart, T., Mao, Y., Korbel, J. O., Eichler, E. E., Zody, M. C., Dilthey, A. T., and Marschall, T. (2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nature Genetics*, 54(4):518–525. Number: 4 Publisher: Nature Publishing Group.

Eisenberg, E. and Levanon, E. Y. (2003). Human housekeeping genes are compact. *Trends in Genetics*, 19(7):362–365.

Ellegren, H. and Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews Genetics*, 17(7):422–433. Number: 7 Publisher: Nature Publishing Group.

Emms, D. M. and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1):238.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048.

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

Fan, L., Cao, P., Almeida, J., and Broder, A. (2000). Summary cache: a scalable wide-area web cache sharing protocol. *IEEE/ACM Transactions on Networking*, 8(3):281–293.

Filatov, D. A. (2019). Extreme Lewontin's Paradox in Ubiquitous Marine Phytoplankton Species. *Molecular Biology and Evolution*, 36(1):4–14.

Fisher, R. A. (1923). XXI.—On the Dominance Ratio. *Proceedings of the Royal Society of Edinburgh*, 42:321–341.

Flagel, L., Brandvain, Y., and Schrider, D. R. (2019). The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference. *Molecular Biology and Evolution*, 36(2):220–238.

Flavell, R. B., Bennett, M. D., Smith, J. B., and Smith, D. B. (1974). Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochemical Genetics*, 12(4):257–269.

Fletcher, K., Zhang, L., Gil, J., Han, R., Cavanaugh, K., and Michelmore, R. (2021). AFLAP: assembly-free linkage analysis pipeline using k-mers from genome sequencing data. *Genome Biology*, 22(1):115.

Flowers, J. M., Molina, J., Rubinstein, S., Huang, P., Schaal, B. A., and Purugganan, M. D. (2012). 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.

Fofanov, Y., Luo, Y., Katili, C., Wang, J., Belosludtsev, Y., Powdrill, T., Belapurkar, C., Fofanov, V., Li, T.-B., Chumakov, S., and Pettitt, B. M. (2004). How independent are the appearances of n-mers in different genomes? *Bioinformatics*, 20(15):2421–2428.

Fong, L. J. M., Darolti, I., Metzger, D. C. H., Morris, J., Lin, Y., Sandkam, B. A., and Mank, J. E. (2023). Evolutionary history of the *Poecilia picta* sex chromosomes. *Genome Biology and Evolution*, 15(3):evad030.

Frankham, R. (2012). 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. Number: 3 Publisher: Nature Publishing Group.

Frenkel, S., Kirzhner, V., Frenkel, Z., and Korol, A. B. (2016). Organizational heterogeneity of the human genome: significant variation of recombination rate of 100 kbp sequences within GC ranges. In *2016 Second International Symposium on Stochastic Models in Reliability Engineering, Life Science and Operations Management (SMRLO)*, pages 414–420.

Fukushima, K. and Pollock, D. D. (2020). Amalgamated cross-species transcriptomes reveal organ-specific propensity in gene expression evolution. *Nature Communications*, 11(1):4459. Number: 1 Publisher: Nature Publishing Group.

Gage, J. L., Vaillancourt, B., Hamilton, J. P., Manrique-Carpintero, N. C., Gustafson, T. J., Barry, K., Lipzen, A., Tracy, W. F., Mikel, M. A., Kaeppler, S. M., Buell, C. R., and de Leon, N. (2019). Multiple Maize Reference Genomes Impact the Identification of Variants by Genome-Wide Association Study in a Diverse Inbred Panel. *The Plant Genome*, 12(2):180069. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3835/plantgenome2018.09.0069.

Gardner, S. N. and Hall, B. G. (2013). When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. *PLOS ONE*, 8(12):e81760. Publisher: Public Library of Science.

Gardner, S. N. and Slezak, T. (2010). Scalable SNP analyses of 100+ vacterial or viral genomes. *Journal of Forensic Research*, 01(03).

Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., Hagmann, J., Vorbrugg, S., Marco-Sola, S., Kubica, C., Ashbrook, D. G., Thorell, K., Rusholme-Pilcher, R. L., Liti, G., Rudbeck, E., Golicz, A. A., Nahnsen, S., Yang, Z., Mwaniki, M. N., Nobrega, F. L., Wu, Y., Chen, H., de Ligt, J., Sudmant, P. H., Huang, S., Weigel, D., Soranzo, N., Colonna, V., Williams, R. W., and Prins, P. (2024). Building pangenome graphs. *Nature Methods*, 21(11):2008–2012. Publisher: Nature Publishing Group.

Gaut, B., Yang, L., Takuno, S., and Eguiarte, L. E. (2011). The patterns and causes of variation in plant nucleotide substitution rates. *Annual Review of Ecology, Evolution, and Systematics*, 42(1):245–266.

Gauthier, J., Mouden, C., Suchan, T., Alvarez, N., Arrigo, N., Riou, C., Lemaitre, C., and Peterlongo, P. (2020). DiscoSnp-RAD: de novo detection of small variants for RAD-Seq population genomics. *PeerJ*, 8:e9291. Publisher: PeerJ Inc.

Georgakopoulos-Soares, I., Yizhar-Barnea, O., Mouratidis, I., Hemberg, M., and Ahituv, N. (2021). Absent from DNA and protein: genomic characterization of nullomers and nullpeptides across functional categories and evolution. *Genome Biology*, 22(1):245.

Gerland, U. and Hwa, T. (2002). On the selection and evolution of regulatory DNA motifs. *Journal of Molecular Evolution*, 55(4):386–400.

Glasenapp, M. R. and Pogson, G. H. (2024). Selection Shapes the Genomic Landscape of Introgressed Ancestry in a Pair of Sympatric Sea Urchin Species. *Genome Biology and Evolution*, 16(6):evae124.

Glémin, S. (2007). Mating systems and the efficacy of selection at the molecular level. *Genetics*, 177(2):905–916.

Glémin, S. (2012). Extinction and fixation times with dominance and inbreeding. *Theoretical Population Biology*, 81(4):310–316.

Goerner-Potvin, P. and Bourque, G. (2018). Computational tools to unmask transposable elements. *Nature Reviews Genetics*, 19(11):688–704. Publisher: Nature Publishing Group.

Golicz, A. A., Bayer, P. E., Bhalla, P. L., Batley, J., and Edwards, D. (2020). Pangenomics comes of age: from bacteria to plant and animal applications. *Trends in Genetics*, 36(2):132–145.

Gourlé, H., Karlsson-Lindsjö, O., Hayer, J., and Bongcam-Rudloff, E. (2019). Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics*, 35(3):521–522.

Govaerts, R., Nic Lughadha, E., Black, N., Turner, R., and Paton, A. (2021). The World Checklist of Vascular Plants, a continuously updated resource for exploring global plant diversity. *Scientific Data*, 8(1):215. Number: 1 Publisher: Nature Publishing Group.

Greven, A., Pfaffelhuber, P., Pokalyuk, C., and Wakolbinger, A. (2016). The fixation time of a strongly beneficial allele in a structured population. *Electronic Journal of Probability*, 21(none):1–42. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

Groen, S. C., Ćalić, I., Joly-Lopez, Z., Platts, A. E., Choi, J. Y., Natividad, M., Dorph, K., Mauck, W. M., Bracken, B., Cabral, C. L. U., Kumar, A., Torres, R. O., Satija, R., Vergara, G., Henry, A., Franks, S. J., and Purugganan, M. D. (2020). The strength and pattern of natural selection on gene expression in rice. *Nature*, 578(7796):572–576. Number: 7796 Publisher: Nature Publishing Group.

Grytten, I., Dagestad Rand, K., and Sandve, G. K. (2022). KAGE: fast alignment-free graph-based genotyping of SNPs and short indels. *Genome Biology*, 23(1):209.

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

Guo, X., Wang, Y., Keightley, P. D., and Fan, L. (2007). Patterns of selective constraints in non-coding DNA of rice. *BMC Evolutionary Biology*, 7(1):208.

Gupta, P. K. (2021). GWAS for genetics of complex quantitative traits: Genome to pangenome and SNPs to SVs and k-mers. *BioEssays*, 43(11):2100109. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.202100109.

Gyorgy, A. (2023). Competition and evolutionary selection among core regulatory motifs in gene expression control. *Nature Communications*, 14(1):8266. Number: 1 Publisher: Nature Publishing Group.

Göktay, M., Fulgione, A., and Hancock, A. M. (2021). 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.

Günther, T. and Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLOS Genetics*, 15(7):e1008302. Publisher: Public Library of Science.

Haberer, G., Kamal, N., Bauer, E., Gundlach, H., Fischer, I., Seidel, M. A., Spannagl, M., Marcon, C., Ruban, A., Urbany, C., Nemri, A., Hochholdinger, F., Ouzunova, M., Houben, A., Schön, C.-C., and Mayer, K. F. X. (2020). European maize genomes highlight intraspecies variation in repeat and gene content. *Nature Genetics*, 52(9):950–957. Publisher: Nature Publishing Group.

Hahn, M. W. (2018). *Molecular Population Genetics*. Oxford University Press. Google-Books-ID: 3BDkswEACAAJ.

Hahn, M. W. and Mishra, S. R. (2025). Estimating recombination using only the allele frequency spectrum. Pages: 2025.02.01.635998 Section: New Results.

Haldane, J. B. S. (1927). A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 23(7):838–844.

Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W., and Ralph, P. L. (2019). Tree-sequence recording in slim opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, 19(2):552–566.

Haller, B. C. and Messer, P. W. (2019). SLiM 3: forward genetic simulations beyond the Wright–Fisher model. *Molecular Biology and Evolution*, 36(3):632–637.

Halushka, M. K., Fan, J.-B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R., and Chakravarti, A. (1999). Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genetics*, 22(3):239–247. Publisher: Nature Publishing Group.

Haubold, B. (2014). Alignment-free phylogenetics and population genetics. *Briefings in Bioinformatics*, 15(3):407–418.

Haubold, B., Krause, L., Horn, T., and Pfaffelhuber, P. (2013). An alignment-free test for recombination. *Bioinformatics*, 29(24):3121–3127.

Haubold, B. and Pfaffelhuber, P. (2012). Alignment-free population genomics: an efficient estimator of sequence diversity. *G3 Genes|Genomes|Genetics*, 2(8):883–889.

Haubold, B., Reed, F. A., and Pfaffelhuber, P. (2011). Alignment-free estimation of nucleotide diversity. *Bioinformatics*, 27(4):449–455.

He, Z., Dai, X., Beaumont, M., and Yu, F. (2020). Estimation of Natural Selection and Allele Age from Time Series Allele Frequency Data Using a Novel Likelihood-Based Approach. *Genetics*, 216(2):463–480.

Hellmann, I., Ebersberger, I., Ptak, S. E., Pääbo, S., and Przeworski, M. (2003). A neutral explanation for the correlation of diversity with recombination rates in humans. *The American Journal of Human Genetics*, 72(6):1527–1535.

Hickey, G., Heller, D., Monlong, J., Sibbesen, J. A., Sirén, J., Eizenga, J., Dawson, E. T., Garrison, E., Novak, A. M., and Paten, B. (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biology*, 21(1):35.

Ho, E. K. H., Macrae, F., Latta, IV, L. C., Benner, M. J., Sun, C., Ebert, D., and Schaack, S. (2019). Intraspecific variation in microsatellite mutation profiles in *Daphnia magna*. *Molecular Biology and Evolution*, 36(9):1942–1954.

Hrytsenko, Y., Daniels, N. M., and Schwartz, R. S. (2022). Determining population structure from k-mer frequencies. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '22, page 1, New York, NY, USA. Association for Computing Machinery.

Huang, Y.-F. (2022). Dissecting genomic determinants of positive selection with an evolution-guided regression model. *Molecular Biology and Evolution*, 39(1):msab291.

Huber, C. D., DeGiorgio, M., Hellmann, I., and Nielsen, R. (2016). Detecting recent selective sweeps while controlling for mutation rate and background selection. *Molecular Ecology*, 25(1):142–156.

Huber, C. D., Kim, B. Y., Marsden, C. D., and Lohmueller, K. E. (2017). Determining the factors driving selective effects of new nonsynonymous mutations. *Proceedings of the National Academy of Sciences*, 114(17):4465–4470. Publisher: Proceedings of the National Academy of Sciences.

Hunt, B. G., Ometto, L., Wurm, Y., Shoemaker, D., Yi, S. V., Keller, L., and Goodisman, M. A. D. (2011). Relaxed selection is a precursor to the evolution of phenotypic plasticity. *Proceedings of the National Academy of Sciences*, 108(38):15936–15941. Publisher: National Academy of Sciences Section: Biological Sciences.

Häntze, H. and Horton, P. (2023). Effects of spaced k-mers on alignment-free genotyping. *Bioinformatics*, 39(Supplement_1):i213–i221.

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

Institute, B. (2019). Picard toolkit. *Broad Institute, GitHub repository*.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, 44(2):226–232. Number: 2 Publisher: Nature Publishing Group.

Jaegle, B., Pisupati, R., Soto-Jiménez, L. M., Burns, R., Rabanal, F. A., and Nordborg, M. (2023). Extensive sequence duplication in *Arabidopsis* revealed by pseudo-heterozygosity. *Genome Biology*, 24(1):44.

Jenike, K. M., Campos-Domínguez, L., Boddé, M., Cerca, J., Hodson, C. N., Schatz, M. C., and Jaron, K. S. (2024). Guide to k-mer approaches for genomics across the tree of life. arXiv:2404.01519.

Johnsson, P., Lipovich, L., Grandér, D., and Morris, K. V. (2014). Evolutionary conservation of long noncoding RNAs; sequence, structure, function. *Biochimica et biophysica acta*, 1840(3):1063–1071.

Johri, P., Aquadro, C. F., Beaumont, M., Charlesworth, B., Excoffier, L., Eyre-Walker, A., Keightley, P. D., Lynch, M., McVean, G., Payseur, B. A., Pfeifer, S. P., Stephan, W., and Jensen, J. D. (2022a). Recommendations for improving statistical inference in population genomics. *PLOS Biology*, 20(5):e3001669. Publisher: Public Library of Science.

Johri, P., Charlesworth, B., and Jensen, J. D. (2020). Toward an evolutionarily appropriate null model: jointly inferring demography and purifying selection. *Genetics*, 215(1):173–192.

Johri, P., Eyre-Walker, A., Gutenkunst, R. N., Lohmueller, K. E., and Jensen, J. D. (2022b). On the prospect of achieving accurate joint estimation of selection with population history. *Genome Biology and Evolution*, 14(7):evac088.

Jordan, I. K., Wolf, Y. I., and Koonin, E. V. (2004). Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evolutionary Biology*, 4(1):22.

Josephs, E. B., Wright, S. I., Stinchcombe, J. R., and Schoen, D. J. (2017). The relationship between selection, network connectivity, and regulatory variation within a population of *Capsella grandiflora*. *Genome Biology and Evolution*, 9(4):1099–1109.

Kane, N. C. and Rieseberg, L. H. (2007). Selective Sweeps Reveal Candidate Genes for Adaptation to Drought and Salt Tolerance in Common Sunflower, Helianthus annuus. *Genetics*, 175(4):1823–1834.

Kaplinski, L., Möls, M., Puurand, T., Pajuste, F.-D., and Remm, M. (2021). KATK: Fast genotyping of rare variants directly from unmapped sequencing reads. *Human Mutation*, 42(6):777–786. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.24197.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780. Publisher: Oxford Academic.

Kaushik, S. and Jain, K. (2021). Time to fixation in changing environments. *Genetics*, 219(3):iyab148.

Kawecki, T. J. (1994). Accumulation of Deleterious Mutations and the Evolutionary Cost of Being a Generalist. *The American Naturalist*, 144(5):833–838. Publisher: The University of Chicago Press.

Ke, S., Zhang, X. H.-F., and Chasin, L. A. (2008). Positive selection acting on splicing motifs reflects compensatory evolution. *Genome Research*, 18(4):533–543. 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.

Kelley, D. R., Schatz, M. C., and Salzberg, S. L. (2010). Quake: quality-aware detection and correction of sequencing errors. *Genome Biology*, 11(11):R116.

Kent, W. J. (2002). The human genome browser at UCSC. *Genome Res.*, 12(6):996–1006.

Kern, A. D. and Schrider, D. R. (2018). diploS/HIC: An Updated Approach to Classifying Selective Sweeps. *G3 Genes|Genomes|Genetics*, 8(6):1959–1970.

Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, 115(1):49–63.

Kim, J.-H., Park, J.-S., Lee, C.-Y., Jeong, M.-G., Xu, J. L., Choi, Y., Jung, H.-W., and Choi, H.-K. (2020). 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. Publisher: Public Library of Science.

Kim, S. (2015). ppcor: an R package for a fast calculation to semi-partial correlation coefficients. *Communications for statistical applications and methods*, 22(6):665–674.

Kim, Y. and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160(2):765–777. WOS:000174097600036.

Kimura, M. (1980). Average time until fixation of a mutant allele in a finite population under continued mutation pressure: Studies by analytical, numerical, and pseudo-sampling methods. *Proceedings of the National Academy of Sciences*, 77(1):522–526. Publisher: Proceedings of the National Academy of Sciences.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press. Google-Books-ID: olIoSumPevYC.

Kimura, M. and Ohta, T. (1969). The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population. *Genetics*, 61(3):763–771.

Klassmann, A. and Gautier, M. (2022). Detecting selection using extended haplotype homozygosity (EHH)-based statistics in unphased or unpolarized data. *PLoS ONE*, 17(1):e0262024.

Kojima, K. and Schaffer, H. E. (1964). Accumulation of epistatic gene complexes. *Evolution*, 18(1):127–129.

Kojima, K.-i. and Schaffer, H. E. (1967). Survival Process of Linked Mutant Genes. *Evolution*, 21(3):518–531. Publisher: [Society for the Study of Evolution, Wiley].

Kokot, M., Długosz, M., and Deorowicz, S. (2017). KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*, 33(17):2759–2761. Publisher: Oxford Academic.

Kolekar, P., Kale, M., and Kulkarni-Kale, U. (2012). Alignment-free distance measure based on return time distribution for sequence analysis: Applications to clustering, molecular phylogeny and subtyping. *Molecular Phylogenetics and Evolution*, 65(2):510–522.

Kolář, F., Čertner, M., Suda, J., Schönswetter, P., and Husband, B. C. (2017). Mixed-ploidy species: progress and opportunities in polyploid research. *Trends in Plant Science*, 22(12):1041–1055. Publisher: Elsevier.

Koonin, E. V. (2011). Are there laws of genome evolution? *PLOS Computational Biology*, 7(8):e1002173. Publisher: Public Library of Science.

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

Kryuchkova-Mostacci, N. and Robinson-Rechavi, M. (2017). A benchmark of gene expression tissue-specificity metrics. *Briefings in Bioinformatics*, 18(2):205–214.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. *Molecular Biology and Evolution*, 34(7):1812–1819.

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., Stecher, G., and Hedges, S. B. (2022). TimeTree 5: an expanded resource for species divergence times. *Molecular Biology and Evolution*, 39(8):msac174.

Lanfear, R., Ho, S. Y. W., Jonathan Davies, T., Moles, A. T., Aarssen, L., Swenson, N. G., Warman, L., Zanne, A. E., and Allen, A. P. (2013). Taller plants have lower rates of molecular evolution. *Nature Communications*, 4(1):1879. Number: 1 Publisher: Nature Publishing Group.

Langley, C. H., Stevens, K., Cardeno, C., Lee, Y. C. G., Schrider, D. R., Pool, J. E., Langley, S. A., Suarez, C., Corbett-Detig, R. B., Kolaczkowski, B., Fang, S., Nista, P. M., Holloway, A. K., Kern, A. D., Dewey, C. N., Song, Y. S., Hahn, M. W., and Begun, D. J. (2012). Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics*, 192(2):533–598.

Larracuente, A. M., Sackton, T. B., Greenberg, A. J., Wong, A., Singh, N. D., Sturgill, D., Zhang, Y., Oliver, B., and Clark, A. G. (2008). Evolution of protein-coding genes in *Drosophila*. *Trends in Genetics*, 24(3):114–123.

Lauterbur, M. E., Munch, K., and Enard, D. (2023). Versatile detection of diverse selective sweeps with Flex-sweep. Pages: 2022.11.15.516494 Section: New Results.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739. Number: 10 Publisher: Nature Publishing Group.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLOS Genetics*, 3(9):e161. Publisher: Public Library of Science.

Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Ségurel, L., Venkat, A., Andolfatto, P., and Przeworski, M. (2012). Revisiting an old riddle: what determines genetic diversity levels within species? *PLOS Biology*, 10(9):e1001388. Publisher: Public Library of Science.

Leggett, R. M., Ramirez-Gonzalez, R. H., Verweij, W., Kawashima, C. G., Iqbal, Z., Jones, J. D. G., Caccamo, M., and MacLean, D. (2013). Identifying and classifying trait linked polymorphisms in non-reference species by walking coloured de Bruijn graphs. *PLOS ONE*, 8(3):e60058. Publisher: Public Library of Science.

Lei, L., Goltsman, E., Goodstein, D., Wu, G. A., Rokhsar, D. S., and Vogel, J. P. (2021). Plant pan-genomics comes of age. *Annual Review of Plant Biology*, 72(1):411–435. _eprint: https://doi.org/10.1146/annurev-arplant-080720-105454.

Lemane, T., Chikhi, R., and Peterlongo, P. (2022). kmdiff, large-scale and user-friendly differential k-mer analyses. *Bioinformatics*, 38(24):5443–5445.

Lewanski, A. L., Grundler, M. C., and Bradburd, G. S. (2024). The era of the ARG: An introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics. *PLOS Genetics*, 20(1):e1011110.

Lewontin, R. C. (1974). *The genetic basis of evolutionary change*. Columbia University Press.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio].

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20):2843–2851.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, W. H. and Sadler, L. A. (1991). Low Nucleotide Diversity in Man. *Genetics*, 129(2):513–523.

Li, Y., Patel, H., and Lin, Y. (2022). Kmer2SNP: Reference-free heterozygous SNP calling using k-mer frequency distributions. In Ng, C. and Piscuoglio, S., editors, *Variant Calling: Methods and Protocols*, pages 257–265. Springer US, New York, NY.

Liao, Q., Du, R., Gou, J., Guo, L., Shen, H., Liu, H., Nguyen, J. K., Ming, R., Yin, T., Huang, S., and Yan, J. (2020). The genomic architecture of the sex-determining region and sex-related metabolic variation in Ginkgobiloba. *The Plant Journal*, 104(5):1399–1409. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tpj.15009.

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., Garg, S., Groza, C., Guarracino, A., Harvey, W. T., Heumos, S., Howe, K., Jain, M., Lu, T.-Y., Markello, C., Martin, F. J., Mitchell, M. W., Munson, K. M., Mwaniki, M. N., Novak, A. M., Olsen, H. E., Pesout, T., Porubsky, D., Prins, P., Sibbesen, J. A., Sirén, J., Tomlinson, C., Villani, F., Vollger, M. R., Antonacci-Fulton, L. L., Baid, G., Baker, C. A., Belyaeva, A., Billis, K., Carroll, A., Chang, P.-C., Cody, S., Cook, D. E., Cook-Deegan, R. M., Cornejo, O. E., Diekhans, M., Ebert, P., Fairley, S., Fedrigo, O., Felsenfeld, A. L., Formenti, G., Frankish, A., Gao, Y., Garrison, N. A., Giron, C. G., Green, R. E., Haggerty, L., Hoekzema, K., Hourlier, T., Ji, H. P., Kenny, E. E., Koenig, B. A., Kolesnikov, A., Korbel, J. O., Kordosky, J., Koren, S., Lee, H., Lewis, A. P., Magalhães, H., Marco-Sola, S., Marijon, P.,

McCartney, A., McDaniel, J., Mountcastle, J., Nattestad, M., Nurk, S., Olson, N. D., Popejoy, A. B., Puiu, D., Rautiainen, M., Regier, A. A., Rhie, A., Sacco, S., Sanders, A. D., Schneider, V. A., Schultz, B. I., Shafin, K., Smith, M. W., Sofia, H. J., Abou Tayoun, A. N., Thibaud-Nissen, F., Tricomi, F. F., Wagner, J., Walenz, B., Wood, J. M. D., Zimin, A. V., Bourque, G., Chaisson, M. J. P., Flicek, P., Phillippy, A. M., Zook, J. M., Eichler, E. E., Haussler, D., Wang, T., Jarvis, E. D., Miga, K. H., Garrison, E., Marschall, T., Hall, I. M., Li, H., and Paten, B. (2023). A draft human pangenome reference. *Nature*, 617(7960):312–324. Publisher: Nature Publishing Group.

Lichilín, N., Salzburger, W., and Böhne, A. (2023). No evidence for sex chromosomes in natural populations of the cichlid fish *Astatotilapia burtoni*. *G3 Genes|Genomes|Genetics*, 13(3):jkad011.

Lin, Y.-S., Hsu, W.-L., Hwang, J.-K., and Li, W.-H. (2007). Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Molecular Biology and Evolution*, 24(4):1005–1011.

Linard, B., Swenson, K., and Pardi, F. (2019). Rapid alignment-free phylogenetic identification of metagenomic sequences. *Bioinformatics*, 35(18):3303–3312.

Liu, G., Liu, J., Cui, X., and Cai, L. (2012). Sequence-dependent prediction of recombination hotspots in *Saccharomyces cerevisiae*. *Journal of Theoretical Biology*, 293:49–54.

Liu, S., Zheng, J., Migeon, P., Ren, J., Hu, Y., He, C., Liu, H., Fu, J., White, F. F., Toomajian, C., and Wang, G. (2017). Unbiased *k*-mer analysis reveals changes in copy number of highly repetitive sequences during maize domestication and improvement. *Scientific Reports*, 7(1):42444. Number: 1 Publisher: Nature Publishing Group.

Liu, Z. and Samee, M. (2023). Structural underpinnings of mutation rate variations in the human genome. *Nucleic Acids Research*, 51(14):7184–7197.

Luczak, B. B., James, B. T., and Girgis, H. Z. (2019). A survey and evaluations of histogram-based statistics in alignment-free sequence comparison. *Briefings in Bioinformatics*, 20(4):1222–1237.

Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155. Publisher: American Association for the Advancement of Science.

Lynch, M. and Conery, J. S. (2003). The origins of genome complexity. *Science*, 302(5649):1401–1404. Publisher: American Association for the Advancement of Science.

López□Delgado, J. and Meirmans, P. G. (2022). History or demography? Determining the drivers of genetic variation in North American plants. *Molecular Ecology*, 31(7):1951–1962.

Mackintosh, A., Laetsch, D. R., Hayward, A., Charlesworth, B., Waterfall, M., Vila, R., and Lohse, K. (2019). The determinants of genetic diversity in butterflies. *Nature Communications*, 10(1):3466. Number: 1 Publisher: Nature Publishing Group.

Makałowski, W. and Boguski, M. S. (1998). 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. Publisher: Proceedings of the National Academy of Sciences.

Margulies, E. H., Blanchette, M., Program, N. C. S., Haussler, D., and Green, E. D. (2003). Identification and characterization of multi-species conserved sequences. *Genome Research*, 13(12):2507–2518. 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.

Masalia, R. R., Bewick, A. J., and Burke, J. M. (2017). Connectivity in gene coexpression networks negatively correlates with rates of molecular evolution in flowering plants. *PLOS ONE*, 12(7):e0182289. Publisher: Public Library of Science.

Mattila, T. M., Tyrmi, J., Pyhäjärvi, T., and Savolainen, O. (2017). Genome-Wide Analysis of Colonization History and Concomitant Selection in Arabidopsis lyrata. *Molecular Biology and Evolution*, 34(10):2665–2677.

McClelland, M. (1985). Selection againstdam methylation sites in the genomes of DNA of enterobacteriophages. *Journal of Molecular Evolution*, 21(4):317–322.

McGuigan, K., Collet, J. M., Allen, S. L., Chenoweth, S. F., and Blows, M. W. (2014). Pleiotropic mutations are subject to strong stabilizing selection. *Genetics*, 197(3):1051–1062.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303. 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.

Mehrab, Z., Mobin, J., Tahmid, I. A., and Rahman, A. (2021). Efficient association mapping from k-mers—An application in finding sex-specific sequences. *PLOS ONE*, 16(1):e0245058. Publisher: Public Library of Science.

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

Melsted, P. and Pritchard, J. K. (2011). Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC Bioinformatics*, 12(1):333.

Mirchandani, C. D., Shultz, A. J., Thomas, G. W. C., Smith, S. J., Baylis, M., Arnold, B., Corbett-Detig, R., Enbody, E., and Sackton, T. B. (2024). A fast, reproducible, high-throughput variant calling workflow for population genomics. *Molecular Biology and Evolution*, 41(1):msad270.

Mishra, Y., Johansson Jänkänpää, H., Kiss, A. Z., Funk, C., Schröder, W. P., and Jansson, S. (2012). *Arabidopsis* plants grown in the field and climate chambers significantly differ in leaf morphology and photosystem components. *BMC Plant Biology*, 12(1):6.

Mondragón-Palomino, M., Meyers, B. C., Michelmore, R. W., and Gaut, B. S. (2002). Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. *Genome Research*, 12(9):1305–1315. 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.

Moriyama, E. N. and Powell, J. R. (1996). Intraspecific nuclear DNA variation in Drosophila. *Molecular Biology and Evolution*, 13(1):261–277.

Moutinho, A. F., Eyre-Walker, A., and Dutheil, J. Y. (2022). Strong evidence for the adaptive walk model of gene evolution in *Drosophila* and *Arabidopsis*. *PLOS Biology*, 20(9):e3001775. Publisher: Public Library of Science.

Mukherjee, D., Mukherjee, A., and Ghosh, T. C. (2016). Evolutionary rate heterogeneity of primary and secondary metabolic pathway genes in *Arabidopsis thaliana*. *Genome Biology and Evolution*, 8(1):17–28.

Murray, K. D., Webers, C., Ong, C. S., Borevitz, J., and Warthmann, N. (2017). kWIP: The k-mer weighted inner product, a de novo estimator of genetic similarity. *PLOS Computational Biology*, 13(9):e1005727. Publisher: Public Library of Science.

Murren, C. J., Auld, J. R., Callahan, H., Ghalambor, C. K., Handelsman, C. A., Heskel, M. A., Kingsolver, J. G., Maclean, H. J., Masel, J., Maughan, H., Pfennig, D. W., Relyea, R. A., Seiter, S., Snell-Rood, E., Steiner, U. K., and Schlichting, C. D. (2015). Constraints on the evolution of phenotypic plasticity: limits and costs of phenotype and plasticity. *Heredity*, 115(4):293–301. Bandiera_abtest: a Cc_license_type: cc_y Cg_type: Nature Research Journals Number: 4 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Evolutionary theory Subject_term_id: evolutionary-theory.

Muñoz-Diez, C., Vitte, C., Ross-Ibarra, J., Gaut, B. S., and Tenaillon, M. I. (2012). Using nextgen sequencing to investigate genome size variation and transposable element content. In Grandbastien, M.-A. and Casacuberta, J. M., editors, *Plant Transposable Elements: Impact on Genome Structure and Function*, pages 41–58. Springer, Berlin, Heidelberg.

Myers, S., Freeman, C., Auton, A., Donnelly, P., and McVean, G. (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature Genetics*, 40(9):1124–1129. Number: 9 Publisher: Nature Publishing Group.

Mähler, N., Wang, J., Terebieniec, B. K., Ingvarsson, P. K., Street, N. R., and Hvidsten, T. R. (2017). Gene co-expression network connectivity is an important determinant of selective constraint. *PLoS Genetics*, 13(4):e1006402.

Nakagome, S., Hudson, R. R., and Di Rienzo, A. (2019). Inferring the model and onset of natural selection under varying population size from the site frequency spectrum and haplotype structure.

Proceedings of the Royal Society B: Biological Sciences, 286(1896):20182541. Publisher: Royal Society.

Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, 3(5):418–426.

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

Nei, M. and Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, 76(10):5269–5273. Publisher: Proceedings of the National Academy of Sciences.

Nei, M. and Tajima, F. (1981). DNA polymorphism detectable by restriction endonucleases. *Genetics*, 97(1):145–163.

Nembaware, V., Crum, K., Kelso, J., and Seoighe, C. (2002). Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. *Genome Research*, 12(9):1370–1376. 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.

Neves, C. J., Matzrafi, M., Thiele, M., Lorant, A., Mesgaran, M. B., and Stetter, M. G. (2020). Male linked genomic region determines sex in dioecious *Amaranthus palmeri*. *Journal of Heredity*, 111(7):606–612.

Nordborg, M., Hu, T. T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobsson, M., Kim, S., Morozov, Y., Padhukasahasram, B., Plagnol, V., Rosenberg, N. A., Shah, C., Wall, J. D., Wang, J., Zhao, K., Kalbfleisch, T., Schulz, V., Kreitman, M., and Bergelson, J. (2005). The pattern of polymorphism in *Arabidopsis thaliana*. *PLOS Biology*, 3(7):e196. Publisher: Public Library of Science.

Nordström, K. J. V., Albani, M. C., James, G. V., Gutjahr, C., Hartwig, B., Turck, F., Paszkowski, U., Coupland, G., and Schneeberger, K. (2013). Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. *Nature Biotechnology*, 31(4):325–330. Number: 4 Publisher: Nature Publishing Group.

Novembre, J. and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40(5):646–649. Publisher: Nature Publishing Group.

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., Vicedomini, R., Sahlin, K., Sherwood, E., Elfstrand, M., Gramzow, L., Holmberg, K., Hällman, J., Keech, O., Klasson, L., Koriabine, M., Kucukoglu, M., Käller, M., Luthman, J., Lysholm, F., Niittylä, T., Olson, □., Rilakovic, N., Ritland, C., Rosselló, J. A., Sena, J., Svensson, T., Talavera-López, C., Theißen, G., Tuominen, H., Vanneste, K., Wu, Z.-Q., Zhang, B., Zerbe, P., Arvestad, L., Bhalerao, R., Bohlmann, J., Bousquet, J., Garcia Gil, R., Hvidsten, T. R., de Jong, P., MacKay, J., Morgante, M., Ritland, K., Sundberg, B., Lee Thompson, S., Van de Peer, Y., Andersson, B., Nilsson, O., Ingvarsson,

P. K., Lundeberg, J., and Jansson, S. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature*, 497(7451):579–584. Publisher: Nature Publishing Group.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., and Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1):132.

Onetto, C. A., Sosnowski, M. R., Heuvel, S. V. D., and Borneman, A. R. (2022). Population genomics of the grapevine pathogen *Eutypa lata* reveals evidence for population expansion and intraspecific differences in secondary metabolite gene clusters. *PLOS Genetics*, 18(4):e1010153. Publisher: Public Library of Science.

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

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W. E., Wei, Z., Wang, K., and Lyon, G. J. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Medicine*, 5(3):28.

Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N., and Pearse, W. (2018). caper: Comparative Analyses of Phylogenetics and Evolution in R.

Ormond, L., Foll, M., Ewing, G. B., Pfeifer, S. P., and Jensen, J. D. (2016). Inferring the age of a fixed beneficial allele. *Molecular Ecology*, 25(1):157–169. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.13478.

Ossowski, S., Schneeberger, K., Lucas-Lledó, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., Weigel, D., and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*, 327(5961):92–94. Publisher: American Association for the Advancement of Science.

Otto, S. P. and Whitlock, M. C. (2013). Fixation Probabilities and Times. In *eLS*. John Wiley & Sons, Ltd. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470015902.a0005464.pub3.

Ou, M., Yang, C., Luo, Q., Huang, R., Zhang, A.-D., Liao, L.-J., Li, Y.-M., He, L.-B., Zhu, Z.-Y., Chen, K.-C., and Wang, Y.-P. (2017). An NGS-based approach for the identification of sex-specific markers in snakehead (*Channa argus*). *Oncotarget*, 8(58):98733–98744.

Paape, T., Bataillon, T., Zhou, P., J. Y. Kono, T., Briskine, R., Young, N. D., and Tiffin, P. (2013). Selection, genome-wide fitness effects and evolutionary rates in the model legume *Medicago truncatula*. *Molecular Ecology*, 22(13):3525–3538. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.12329.

Paape, T., Zhou, P., Branca, A., Briskine, R., Young, N., and Tiffin, P. (2012). Fine-scale population recombination rates, hotspots, and correlates of recombination in the *Medicago truncatula* genome. *Genome Biology and Evolution*, 4(5):726–737.

Pajuste, F.-D., Kaplinski, L., Möls, M., Puurand, T., Lepamets, M., and Remm, M. (2017). FastGT: an alignment-free method for calling common SNVs directly from raw sequencing reads. *Scientific Reports*, 7(1):2537. Number: 1 Publisher: Nature Publishing Group.

Pateiro-Lopez, B. and Rodriguez-Casal, A. (2022). alphahull: Generalization of the convex hull of a sample of points in the plane.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419. Number: 4 Publisher: Nature Publishing Group.

Payne, B. L. and Alvarez-Ponce, D. (2018). Higher rates of protein evolution in the self-fertilizing plant *Arabidopsis thaliana* than in the out-crossers *Arabidopsis lyrata* and *Arabidopsis halleri*. *Genome Biology and Evolution*, 10(3):895–900.

Peart, C. R., Tusso, S., Pophaly, S. D., Botero-Castro, F., Wu, C.-C., Aurioles-Gamboa, D., Baird, A. B., Bickham, J. W., Forcada, J., Galimberti, F., Gemmell, N. J., Hoffman, J. I., Kovacs, K. M., Kunnasranta, M., Lydersen, C., Nyman, T., de Oliveira, L. R., Orr, A. J., Sanvito, S., Valtonen, M., Shafer, A. B. A., and Wolf, J. B. W. (2020). Determinants of genetic variation across eco-evolutionary scales in pinnipeds. *Nature Ecology & Evolution*, 4(8):1095–1104. Publisher: Nature Publishing Group.

Pebesma, E. (2018). Simple Features for R: Standardized support for spatial vector data. *The R Journal*, 10(1):439–446.

Pedersen, T. and Crameri, F. (2023). scico.

Pellegrina, L., Pizzi, C., and Vandin, F. (2020). Fast approximation of frequent *k*-mers and applications to metagenomics. *Journal of Computational Biology*, 27(4):534–549. Publisher: Mary Ann Liebert, Inc., publishers.

Pellicer, J. and Leitch, I. J. (2020). 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. _eprint: https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/nph.16261.

Pflug, J. M., Holmes, V. R., Burrus, C., Johnston, J. S., and Maddison, D. R. (2020). Measuring genome sizes using read-depth, k-mers, and flow cytometry: Methodological comparisons in beetles (Coleoptera). *G3 Genes|Genomes|Genetics*, 10(9):3047–3060.

Phatak, S. C. and Rao, S. S. (1995). Logistic map: A possible random-number generator. *Physical Review E*, 51(4):3670–3678.

Phung, T. N., Huber, C. D., and Lohmueller, K. E. (2016). Determining the effect of natural selection on linked neutral divergence across species. *PLOS Genetics*, 12(8):e1006199. Publisher: Public Library of Science.

Piganeau, G. and Eyre-Walker, A. (2009). Evidence for Variation in the Effective Population Size of Animal Mitochondrial DNA. *PLOS ONE*, 4(2):e4396. Publisher: Public Library of Science.

Pigliucci, M. and Kolodynska, A. (2002). Phenotypic plasticity to light intensity in *Arabidopsis thaliana*: invariance of reaction norms and phenotypic integration. *Evolutionary Ecology*, 16(1):27–47.

Plotkin, J. B. and Fraser, H. B. (2007). Assessing the determinants of evolutionary rates in the presence of noise. *Molecular Biology and Evolution*, 24(5):1113–1121.

Ponsero, A. J., Miller, M., and Hurwitz, B. L. (2023). Comparison of *k*-mer-based *de novo* comparative metagenomic tools and approaches. *Microbiome Research Reports*, 2(4):null–null. Publisher: OAE Publishing Inc.

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Auwera, G. A. V. d., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., and Banks, E. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. Pages: 201178 Section: New Results.

Przeworski, M. (2002). The Signature of Positive Selection at Randomly Chosen Loci. *Genetics*, 160(3):1179–1189.

Przeworski, M. (2003). Estimating the Time Since the Fixation of a Beneficial Allele. *Genetics*, 164(4):1667–1676.

Puurand, T., Kukuškina, V., Pajuste, F.-D., and Remm, M. (2019). AluMine: alignment-free method for the discovery of polymorphic Alu element insertions. *Mobile DNA*, 10(1):31.

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

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rahman, A., Hallgrímsdóttir, I., Eisen, M., and Pachter, L. (2018). Association mapping from sequencing reads using k-mers. *eLife*, 7:e32920. Publisher: eLife Sciences Publications, Ltd.

Raijman, D., Shamir, R., and Tanay, A. (2008). Evolution and selection in yeast promoters: Analyzing the combined effect of diverse transcription factor binding sites. *PLOS Computational Biology*, 4(1):e7. Publisher: Public Library of Science.

Ramirez-Ramirez, A. R., Bidot-Martínez, I., Mirzaei, K., Rasoamanalina Rivo, O. L., Menéndez-Grenot, M., Clapé-Borges, P., Espinosa-Lopez, G., and Bertin, P. (2024). Comparing the performances of SSR and SNP markers for population analysis in Theobroma cacao L., as alternative approach to validate a new ddRADseq protocol for cacao genotyping. *PLOS ONE*, 19(5):e0304753.

Ranallo-Benavidez, T. R., Jaron, K. S., and Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, 11(1):1432. Publisher: Nature Publishing Group.

Renny-Byfield, S. and Baumgarten, A. (2020). Repetitive DNA content in the maize genome is uncoupled from population stratification at SNP loci. *BMC Genomics*, 21(1):1–10. Number: 1 Publisher: BioMed Central.

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

Riddell, W. C. (1977). Prediction in Generalized Least Squares. *The American Statistician*, 31(2):88–90. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Ritter, E. J., Cousins, P., Quigley, M., Kile, A., Kenchanmane Raju, S. K., Chitwood, D. H., and Niederhuth, C. (2024). From buds to shoots: insights into grapevine development from the Witch's Broom bud sport. *BMC Plant Biology*, 24(1):283.

Roberts, M. D., Davis, O., Josephs, E. B., and Williamson, R. J. (2024). k-mer-based approaches to bridging pangenomics and population genetics. Version Number: 1.

Rocha, E. P. C. (2006). The quest for the universals of protein evolution. *Trends in Genetics*, 22(8):412–416.

Rogers, A. R. and Huff, C. (2009). Linkage Disequilibrium Between Loci With Unknown Phase. *Genetics*, 182(3):839–844.

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

Roselius, K., Stephan, W., and Städler, T. (2005). The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics*, 171(2):753–763.

Rosen, G., Garbarine, E., Caseiro, D., Polikar, R., and Sokhansanj, B. (2008). Metagenome fragment classification using n-mer frequency profiles. *Advances in Bioinformatics*, 2008:1–12.

Roy, R. S., Bhattacharya, D., and Schliep, A. (2014). Turtle: Identifying frequent k -mers with cache-efficient algorithms . *Bioinformatics*, 30(14):1950–1957.

Ruperao, P., Gandham, P., Odeny, D. A., Mayes, S., Selvanayagam, S., Thirunavukkarasu, N., Das, R. R., Srikanda, M., Gandhi, H., Habyarimana, E., Manyasa, E., Nebie, B., Deshpande, S. P., and Rathore, A. (2023). Exploring the sorghum race level diversity utilizing 272 sorghum accessions genomic resources. *Frontiers in Plant Science*, 14.

Růžička, M., Kulhánek, P., Radová, L., Čechová, A., Špačková, N., Fajkusová, L., and Réblová, K. (2017). DNA mutation motifs in the genes associated with inherited diseases. *PLOS ONE*, 12(8):e0182377. Publisher: Public Library of Science.

Sanchez, D., Sadoun, S. B., Mary-Huard, T., Allier, A., Moreau, L., and Charcosset, A. (2023). Improving the use of plant genetic resources to sustain breeding programs' efficiency. *Proceedings of the National Academy of Sciences*, 120(14):e2205780119. Publisher: Proceedings of the National Academy of Sciences.

Sanchez, T., Cury, J., Charpiat, G., and Jay, F. (2021). Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. *Molecular Ecology Resources*, 21(8):2645–2660. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13224.

Sandell, F. L., Stralis-Pavese, N., McGrath, J. M., Schulz, B., Himmelbauer, H., and Dohm, J. C. (2022). Genomic distances reveal relationships of wild and cultivated beets. *Nature Communications*, 13(1):2021. Number: 1 Publisher: Nature Publishing Group.

Scheiner, S. M. (1993). Genetics and evolution of phenotypic plasticity. *Annual Review of Ecology and Systematics*, 24(1):35–68. _eprint: https://doi.org/10.1146/annurev.es.24.110193.000343.

Schlamp, F., Van Der Made, J., Stambler, R., Chesebrough, L., Boyko, A. R., and Messer, P. W. (2016). Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. *Molecular Ecology*, 25(1):342–356.

Schlichting, C. D. and Smith, H. (2002). Phenotypic plasticity: linking molecular mechanisms with evolutionary outcomes. *Evolutionary Ecology*, 16(3):189–211.

Schmid, K. J., Ramos-Onsins, S., Ringys-Beckstein, H., Weisshaar, B., and Mitchell-Olds, T. (2005). 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.

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

Schmuths, H., Meister, A., Horres, R., and Bachmann, K. (2004). Genome size variation among accessions of *Arabidopsis thaliana*. *Annals of Botany*, 93(3):317–321.

Schneider, H. (2022). Characterization, costs, cues, and future perspectives of phenotypic plasticity. *Annals of botany*.

Shajii, A., Yorukoglu, D., William Yu, Y., and Berger, B. (2016). Fast genotyping of known SNPs through approximate k-mer matching. *Bioinformatics*, 32(17):i538–i544.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423. Conference Name: The Bell System Technical Journal.

Shi, G., Dai, Y., Zhou, D., Chen, M., Zhang, J., Bi, Y., Liu, S., and Wu, Q. (2024). An alignment- and reference-free strategy using k-mer present pattern for population genomic analyses. *Mycology*, 0(0):1–15. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/21501203.2024.2358868.

Shi, H., Schmidt, B., Liu, W., and Müller-Wittig, W. (2010). Quality-score guided error correction for short-read sequencing data using cuda. *Procedia Computer Science*, 1(1):1129–1138. ICCS 2010.

Shi, Z. J., Nayfach, S., and Pollard, K. S. (2023). Identifying species-specific k-mers for fast and accurate metagenotyping with Maast and GT-Pro. *STAR Protocols*, 4(1):101964.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050. 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.

Silva-Junior, O. B. and Grattapaglia, D. (2015). 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. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/nph.13505.

Sims, G. E., Jun, S.-R., Wu, G. A., and Kim, S.-H. (2009). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences*, 106(8):2677–2682. Publisher: Proceedings of the National Academy of Sciences.

Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., Sibbesen, J. A., Hickey, G., Chang, P.-C., Carroll, A., Gupta, N., Gabriel, S., Blackwell, T. W., Ratan, A., Taylor, K. D., Rich, S. S., Rotter, J. I., Haussler, D., Garrison, E., and Paten, B. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, 374(6574):abg8871. Publisher: American Association for the Advancement of Science.

Slotte, T. (2014). The impact of linked selection on plant genomic variation. *Briefings in Functional Genomics*, 13(4):268–275.

Slotte, T., Bataillon, T., Hansen, T. T., St. Onge, K., Wright, S. I., and Schierup, M. H. (2011). Genomic determinants of protein evolution and polymorphism in *Arabidopsis*. *Genome Biology and Evolution*, 3:1210–1219.

Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetics Research*, 23(1):23–35.

Snell-Rood, E. C., Van Dyken, J. D., Cruickshank, T., Wade, M. J., and Moczek, A. P. (2010). Toward a population genetic framework of developmental evolution: the costs, limits, and consequences of phenotypic plasticity. *BioEssays*, 32(1):71–81. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.200900132.

Song, B., Buckler, E. S., and Stitzer, M. C. (2024). New whole-genome alignment tools are needed for tapping into plant diversity. *Trends in Plant Science*, 29(3):355–369. Publisher: Elsevier.

Song, B., Marco-Sola, S., Moreto, M., Johnson, L., Buckler, E. S., and Stitzer, M. C. (2022). AnchorWave: Sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. *Proceedings of the National Academy of Sciences*, 119(1):e2113075119. Publisher: Proceedings of the National Academy of Sciences.

Song, J.-M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., Liu, D., Wang, B., Lu, S., Zhou, R., Xie, W.-Z., Cheng, Y., Zhang, Y., Liu, K., Yang, Q.-Y., Chen, L.-L., and Guo, L. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nature Plants*, 6(1):34–+. Place: Berlin Publisher: Nature Portfolio WOS:000508165000002.

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

South, A. (2011). rworldmap: A new R package for mapping global data. *The R Journal*, 3(1):35–43.

Standage, D. S., Brown, C. T., and Hormozdiari, F. (2019). Kevlar: A mapping-free framework for accurate discovery of *de novo* variants. *iScience*, 18:28–36.

Stern, A. J., Wilton, P. R., and Nielsen, R. (2019). An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLOS Genetics*, 15(9):e1008384. Publisher: Public Library of Science.

Stoletzki, N. and Eyre-Walker, A. (2011). Estimation of the neutrality index. *Molecular Biology and Evolution*, 28(1):63–70.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 34(suppl_-2):W609–W612.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595.

Tajima, F. (1996). The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics*, 143(3):1457–1465.

Takuno, S. and Gaut, B. S. (2012). Body-methylated genes in *Arabidopsis thaliana* are functionally important and evolve slowly. *Molecular Biology and Evolution*, 29(1):219–227.

Tenaillon, M. I., Hufford, M. B., Gaut, B. S., and Ross-Ibarra, J. (2011). Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. *Genome Biology and Evolution*, 3:219–229.

Tenaillon, M. I., Sawkins, M. C., Long, A. D., Gaut, R. L., Doebley, J. F., and Gaut, B. S. (2001). 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. Publisher: Proceedings of the National Academy of Sciences.

Teshima, K. M., Coop, G., and Przeworski, M. (2006). How reliable are empirical genomic scans for selective sweeps? *Genome Research*, 16(6):702–712.

Torada, L., Lorenzon, L., Beddis, A., Isildak, U., Pattini, L., Mathieson, S., and Fumagalli, M. (2019). ImaGene: a convolutional neural network to quantify natural selection from genomic data. *BMC Bioinformatics*, 20(9):337.

Turner, I., Garimella, K. V., Iqbal, Z., and McVean, G. (2018). Integrating long-range connectivity information into de Bruijn graphs. *Bioinformatics*, 34(15):2556–2565.

Uricaru, R., Rizk, G., Lacroix, V., Quillery, E., Plantard, O., Chikhi, R., Lemaitre, C., and Peterlongo, P. (2015). Reference-free detection of isolated SNPs. *Nucleic Acids Research*, 43(2):e11.

Urrutia, A. O. and Hurst, L. D. (2001). Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics*, 159(3):1191–1199.

Urrutia, A. O. and Hurst, L. D. (2003). The signature of selection mediated by expression on human genes. *Genome Research*, 13(10):2260–2264. 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.

Van Buskirk, J. and Steiner, U. K. (2009). The fitness costs of developmental canalization and plasticity. *Journal of Evolutionary Biology*, 22(4):852–860. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1420-9101.2009.01685.x.

Van Dyken, J. D. and Wade, M. J. (2010). The genetic signature of conditional expression. *Genetics*, 184(2):557–570.

Van Kleunen, M. and Fischer, M. (2005). Constraints on the evolution of adaptive phenotypic plasticity in plants. *New Phytologist*, 166(1):49–60. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8137.2004.01296.x.

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

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington,

K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351. Publisher: American Association for the Advancement of Science.

Voichek, Y. and Weigel, D. (2020). Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nature Genetics*, 52(5):534–540. Number: 5 Publisher: Nature Publishing Group.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., and Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, 33(14):2202–2204.

Wang, H.-c., Singer, G. A. C., and Hickey, D. A. (2004). Mutational bias affects protein evolution in flowering plants. *Molecular Biology and Evolution*, 21(1):90–96.

Wang, J., Street, N. R., Scofield, D. G., and Ingvarsson, P. K. (2016). Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related *Populus* species. *Genetics*, 202(3):1185–1200.

Wang, J., Yang, W., Zhang, S., Hu, H., Yuan, Y., Dong, J., Chen, L., Ma, Y., Yang, T., Zhou, L., Chen, J., Liu, B., Li, C., Edwards, D., and Zhao, J. (2023). A pangenome analysis pipeline provides insights into functional gene identification in rice. *Genome Biology*, 24(1):19.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256–276.

Wheeler, L. C., Walker, J. F., Ng, J., Deanna, R., Dunbar-Wallis, A., Backes, A., Pezzi, P. H., Palchetti, M. V., Robertson, H. M., Monaghan, A., de Freitas, L. B., Barboza, G. E., Moyroud, E., and Smith, S. D. (2022). Transcription factors evolve faster than their structural gene targets in the flavonoid pigment pathway. *Molecular Biology and Evolution*, 39(3):msac044.

Whitehouse, L. S. and Schrider, D. R. (2023). Timesweeper: accurately identifying selective sweeps using population genomic time series. *Genetics*, 224(3):iyad084.

Whitlock, M. C. (1996). The Red Queen Beats the Jack-Of-All-Trades: The Limitations on the Evolution of Phenotypic Plasticity and Niche Breadth. *The American Naturalist*, 148:S65–S77. Publisher: [University of Chicago Press, American Society of Naturalists].

Whitney, K. D., Baack, E. J., Hamrick, J. L., Godt, M. J. W., Barringer, B. C., Bennett, M. D., Eckert, C. G., Goodwillie, C., Kalisz, S., Leitch, I. J., and Ross-Ibarra, J. (2010). A role for nonadaptive processes in plant genome size evolution? *Evolution*, 64(7):2097–2109.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wiersma, A. T., Hamilton, J. P., Vaillancourt, B., Brose, J., Awale, H. E., Wright, E. M., Kelly, J. D., and Buell, C. R. (2024). k-mer genome-wide association study for anthracnose and BCMV resistance in a Phaseolus vulgaris Andean Diversity Panel. *The Plant Genome*, n/a(n/a):1–17. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/tpg2.20523.

Wilke, C. O. (2020). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'.

Williamson, R. J., Josephs, E. B., Platts, A. E., Hazzouri, K. M., Haudry, A., Blanchette, M., and Wright, S. I. (2014). Evidence for Widespread Positive and Negative Selection in Coding and Conserved Noncoding Regions of Capsella grandiflora. *PLOS Genetics*, 10(9):e1004622. Publisher: Public Library of Science.

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

Wilson, B. A., Petrov, D. A., and Messer, P. W. (2014). Soft Selective Sweeps in Complex Demographic Scenarios. *Genetics*, 198(2):669–684.

Winter, E. E., Goodstadt, L., and Ponting, C. P. (2004). Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Research*, 14(1):54–61. 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.

Wong, K. H. Y., Ma, W., Wei, C.-Y., Yeh, E.-C., Lin, W.-J., Wang, E. H. F., Su, J.-P., Hsieh, F.-J., Kao, H.-J., Chen, H.-H., Chow, S. K., Young, E., Chu, C., Poon, A., Yang, C.-F., Lin, D.-S., Hu, Y.-F., Wu, J.-Y., Lee, N.-C., Hwu, W.-L., Boffelli, D., Martin, D., Xiao, M., and Kwok, P.-Y. (2020). Towards a reference genome that captures global genetic diversity. *Nature Communications*, 11(1):5482. Number: 1 Publisher: Nature Publishing Group.

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K., Walter, K., Abnizova, I., Gilks, W., Edwards, Y. J. K., Cooke, J. E., and Elgar, G. (2005). Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biology*, 3(1):e7.

Wright, S. (1940). Breeding structure of populations in relation to speciation. *The American Naturalist*, 74(752):232–248. Publisher: [University of Chicago Press, American Society of Naturalists].

Wright, S. I., Lauga, B., and Charlesworth, D. (2002). Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. *Molecular Biology and Evolution*, 19(9):1407–1420.

Wright, S. I., Yau, C. B. K., Looseley, M., and Meyers, B. C. (2004). Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Molecular Biology and Evolution*, 21(9):1719–1726.

Wu, D. Y., Ugozzoli, L., Pal, B. K., Qian, J., and Wallace, R. B. (1991). The effect of temperature and oligonucleotide primer length on the specificity and efficiency of amplification by the polymerase chain reaction. *DNA and Cell Biology*, 10(3):233–238. Publisher: Mary Ann Liebert, Inc., publishers.

Wu, M., Haak, D. C., Anderson, G. J., Hahn, M. W., Moyle, L. C., and Guerrero, R. F. (2021). Inferring the genetic basis of sex determination from the genome of a dioecious nightshade. *Molecular Biology and Evolution*, 38(7):2946–2957.

Wu, Z., Cai, X., Zhang, X., Liu, Y., Tian, G.-b., Yang, J.-R., and Chen, X. (2022). Expression level is a major modifier of the fitness landscape of a protein coding gene. *Nature Ecology & Evolution*, 6(1):103–115. Number: 1 Publisher: Nature Publishing Group.

Y. C. Brandt, D., Wei, X., Deng, Y., Vaughn, A. H., and Nielsen, R. (2022). Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics*, 221(1):iyac044.

Yair, S., Lee, K. M., and Coop, G. (2021). The timing of human adaptation from Neanderthal introgression. *Genetics*, 218(1):iyab052.

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., Lancet, D., and Shmueli, O. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, 21(5):650–659.

Yang, J., Gu, Z., and Li, W.-H. (2003). Rate of protein evolution versus fitness effect of gene deletion. *Molecular Biology and Evolution*, 20(5):772–774.

Yang, L. and Gaut, B. S. (2011). Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. *Molecular Biology and Evolution*, 28(8):2359–2369.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.

Yeang, C.-H. (2010). Quantifying the strength of natural selection of a motif sequence. In Moulton, V. and Singh, M., editors, *Algorithms in Bioinformatics*, Lecture Notes in Computer Science, pages 362–373, Berlin, Heidelberg. Springer.

Yi, H., Lin, Y., Lin, C., and Jin, W. (2021). Kssd: sequence dimensionality reduction by k-mer substring space sampling enables real-time large-scale datasets analysis. *Genome Biology*, 22(1):84.

Younsi, R. and MacLean, D. (2015). Using 2*k*+2 bubble searches to find single nucleotide polymorphisms in *k*-mer graphs. *Bioinformatics*, 31(5):642–646.

Yu, H., Zhang, K., Cheng, G., Mei, C., Wang, H., and Zan, L. (2024). Genome-wide analysis reveals genomic diversity and signatures of selection in Qinchuan beef cattle. *BMC Genomics*, 25(1):558.

Zhang, H., Zhang, F., Yu, Y., Feng, L., Jia, J., Liu, B., Li, B., Guo, H., and Zhai, J. (2020). A comprehensive online database for exploring 20,000 public *Arabidopsis* RNA-seq libraries. *Molecular Plant*, 13(9):1231–1233.

Zhang, J. and Yang, J.-R. (2015). Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics*, 16(7):409–420. Number: 7 Publisher: Nature Publishing Group.

Zhang, L. and Li, W.-H. (2004). Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Molecular Biology and Evolution*, 21(2):236–239.

Zhang, L., Vision, T. J., and Gaut, B. S. (2002). Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. *Molecular Biology and Evolution*, 19(9):1464–1473.

Zhang, Q., Jun, S.-R., Leuze, M., Ussery, D., and Nookaew, I. (2017a). Viral phylogenomics using an alignment-free method: A three-step approach to determine optimal length of *k*-mer. *Scientific Reports*, 7(1):40712. Publisher: Nature Publishing Group.

Zhang, S.-S., Yang, H., Ding, L., Song, Z.-T., Ma, H., Chang, F., and Liu, J.-X. (2017b). Tissue-Specific Transcriptomics Reveals an Important Role of the Unfolded Protein Response in Maintaining Fertility upon Heat Stress in Arabidopsis. *The Plant Cell*, 29(5):1007–1023.

Zhao, L., Lascoux, M., Overall, A. D. J., and Waxman, D. (2013). The Characteristic Trajectory of a Fixing Allele: A Consequence of Fictitious Selection That Arises from Conditioning. *Genetics*, 195(3):993–1006.

Zhao, L., Xie, J., Bai, L., Chen, W., Wang, M., Zhang, Z., Wang, Y., Zhao, Z., and Li, J. (2018). Mining statistically-solid k-mers for accurate NGS error correction. *BMC Genomics*, 19(10):912.

Zhao, X. (2019). BinDash, software for fast genome distance estimation on a typical personal laptop. *Bioinformatics*, 35(4):671–673.

Zhong, L., Zhu, Y., and Olsen, K. M. (2022). Hard versus soft selective sweeps during domestication and improvement in soybean. *Molecular Ecology*, 31(11):3137–3153. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.16454.

Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K., Zhang, J., Lyu, H., Lin, T., Gao, Q., Saha, S., Mueller, L., Fei, Z., Städler, T., Xu, S., Zhang, Z., Speed, D., and Huang, S. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*, 606(7914):527–534. Number: 7914 Publisher: Nature Publishing Group.

Zielezinski, A., Girgis, H. Z., Bernard, G., Leimeister, C.-A., Tang, K., Dencker, T., Lau, A. K., Röhling, S., Choi, J. J., Waterman, M. S., Comin, M., Kim, S.-H., Vinga, S., Almeida, J. S., Chan, C. X., James, B. T., Sun, F., Morgenstern, B., and Karlowski, W. M. (2019). Benchmarking of alignment-free sequence comparison methods. *Genome Biology*, 20(1):144.

Zielezinski, A., Vinga, S., Almeida, J., and Karlowski, W. M. (2017). Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, 18(1):186.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, 29(21):2669–2677.

Zwick, M. E., Cutler, D. J., and Chakravarti, A. (2000). Patterns of genetic variation in Mendelian and complex traits. *Annual Review of Genomics and Human Genetics*, 1:387–407.

APPENDIX A: SUPPLEMENTAL FIGURES FOR CHAPTER 1

General data curation

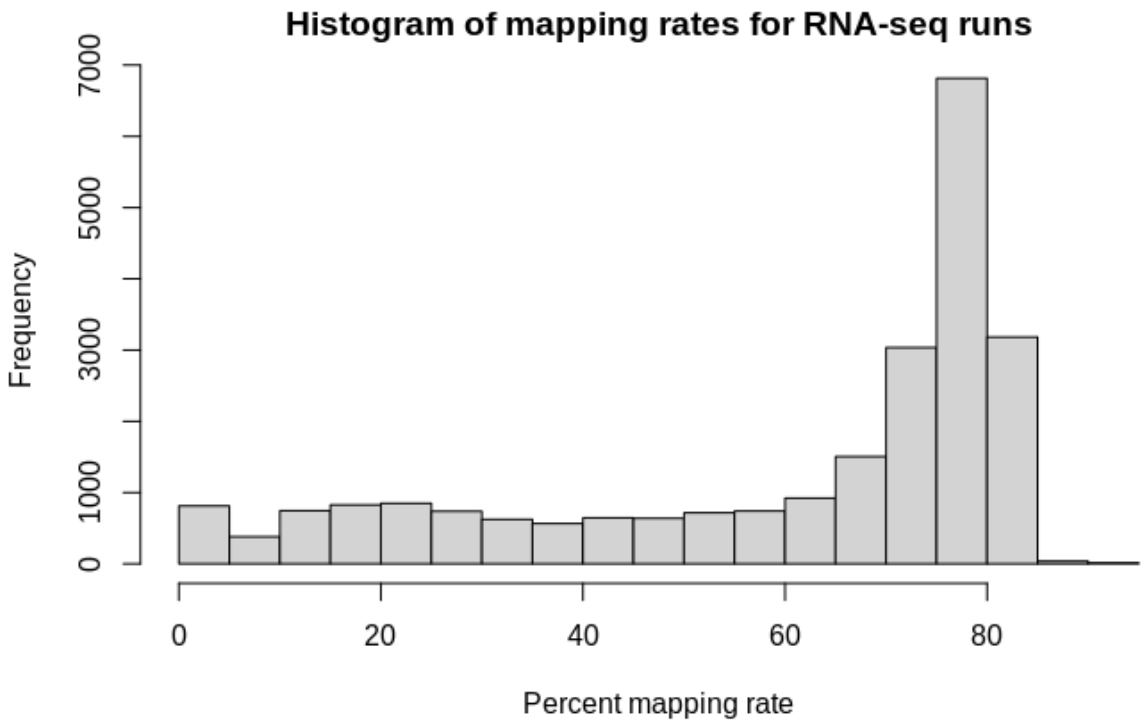

Figure A1: **Histogram of mapping rates for RNA-seq runs.** Bin width is 5 %.

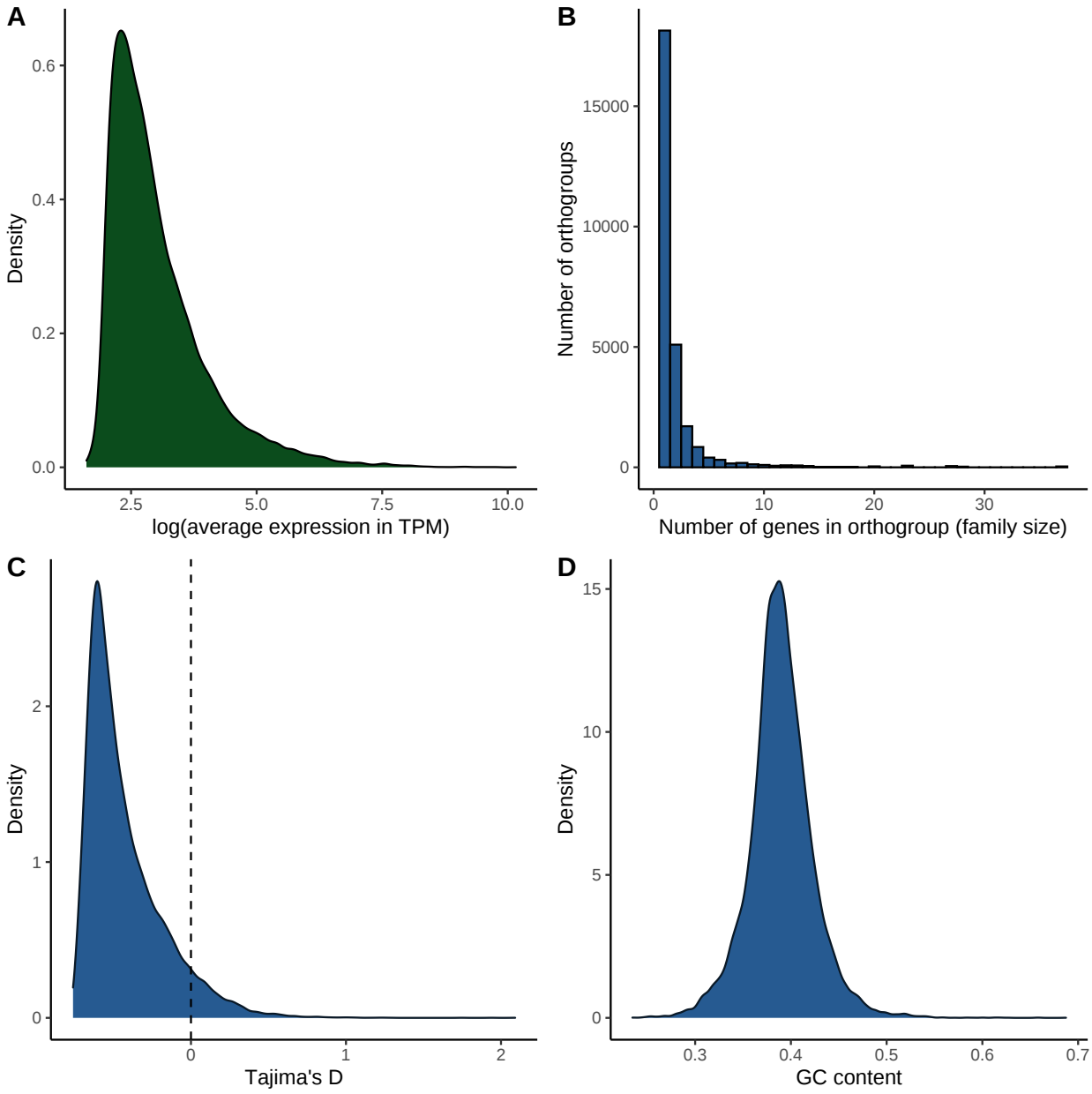

Figure A2: **Histograms of other key variables for the genes included in this study.**

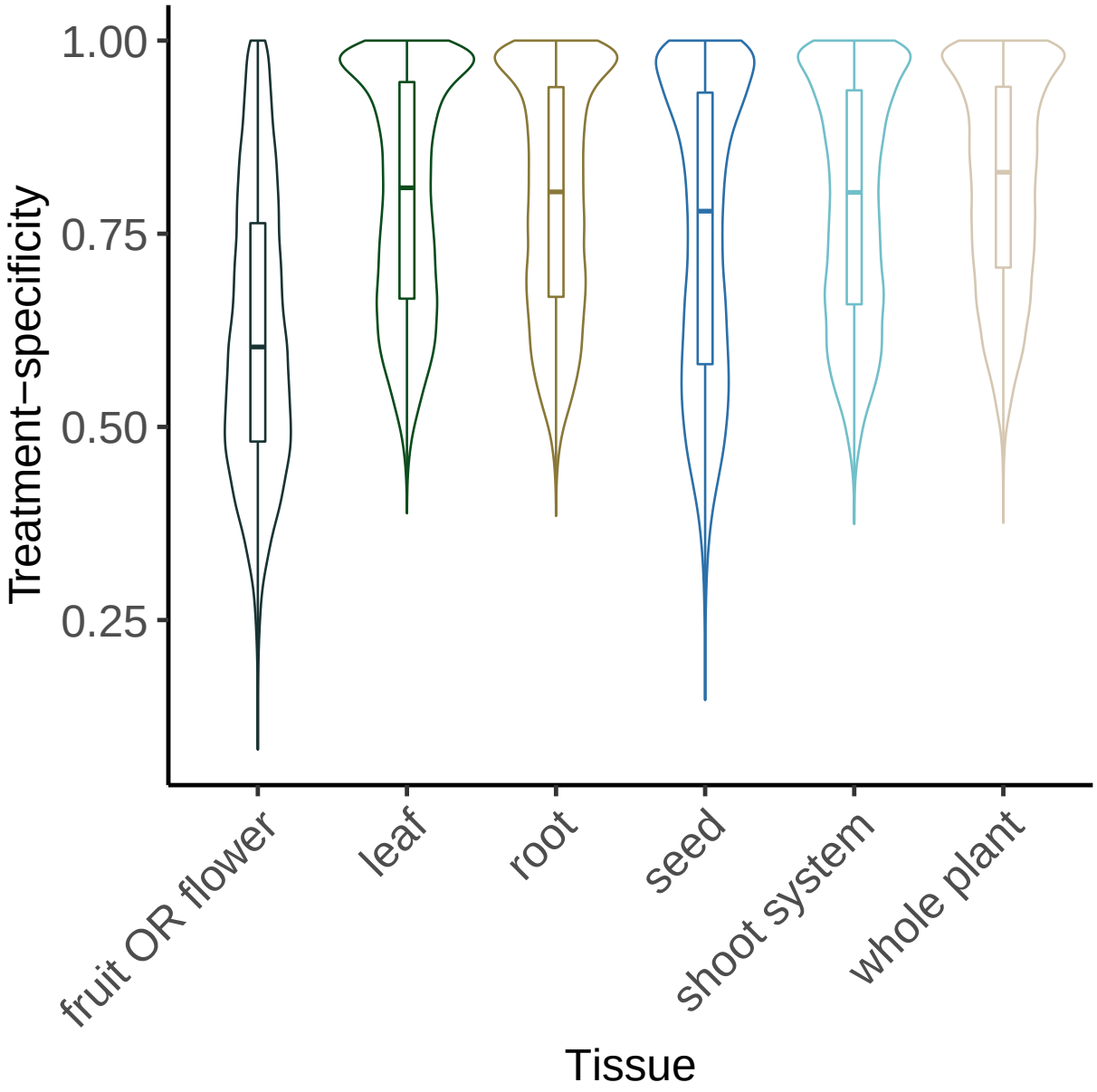

Figure A3: **Violin plot of treatment-specificity in gene expression for six tissue categories.**

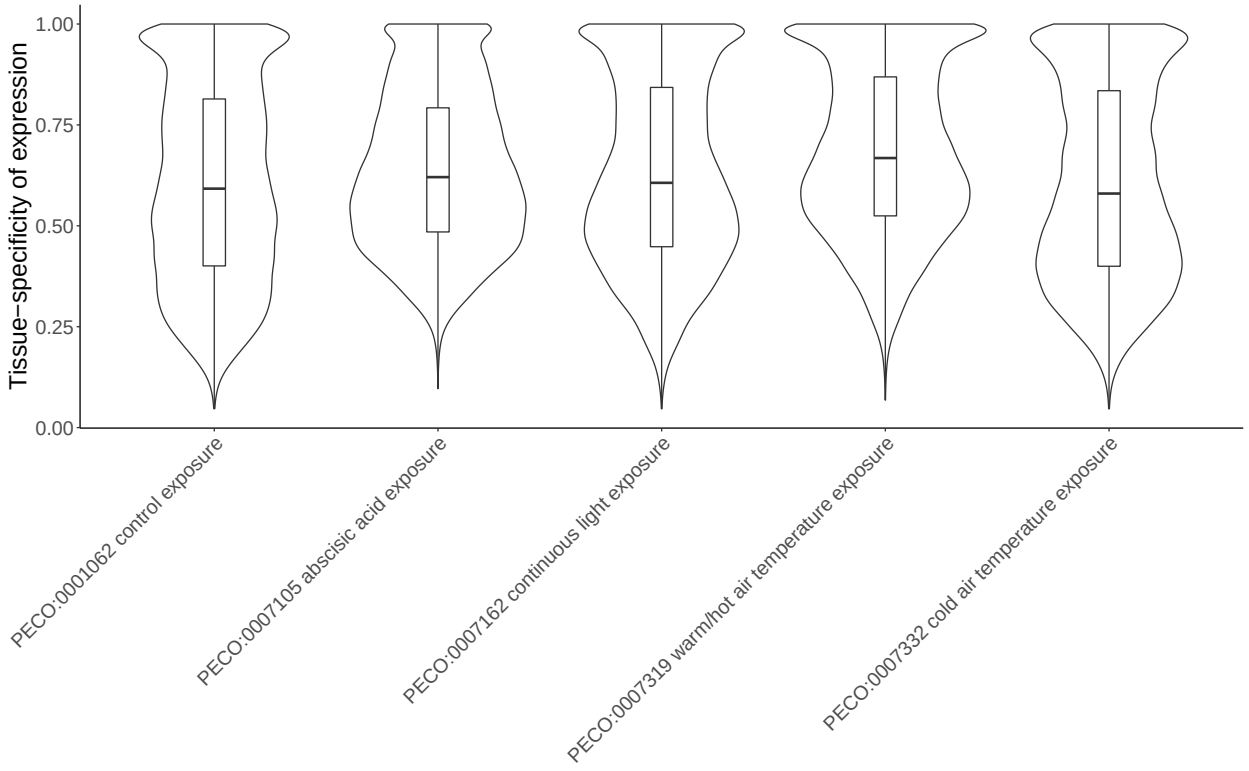

Figure A4: **Violin plot of tissue-specificity in gene expression by treatment type.** Only the five treatments shown had samples from all six tissue types.

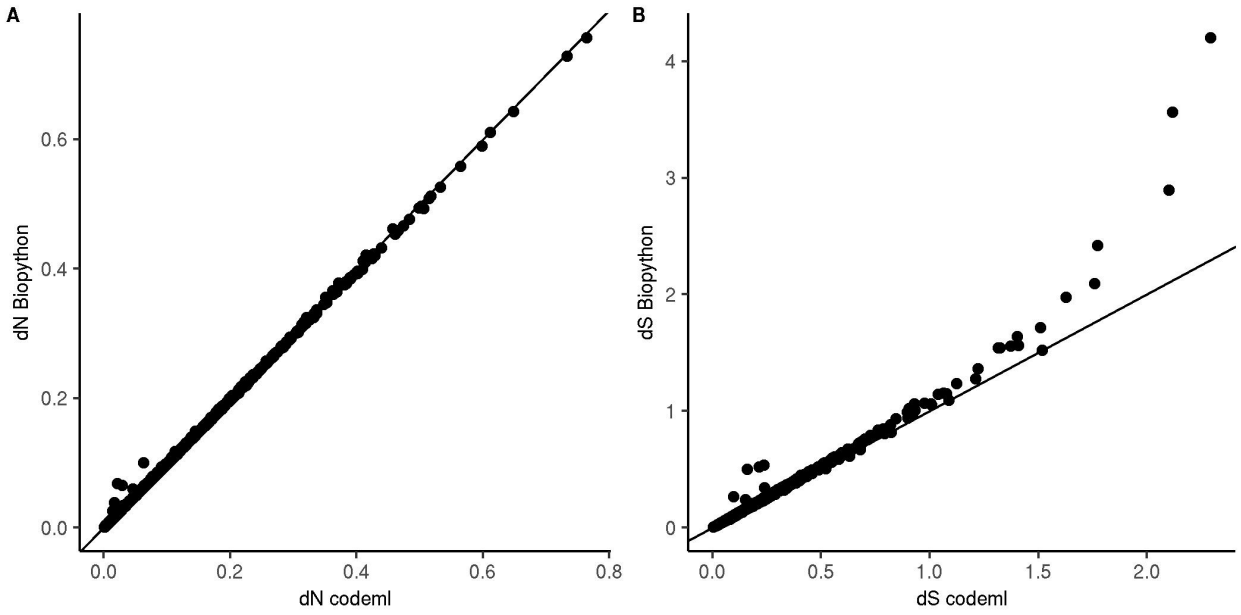

Figure A5: **Comparing dN and dS values returned from codeml and the custom biopython script used in this study.** Line shows exact match between codeml and biopython methods. Genes with saturating divergence ($dS > 1$) were excluded from partial correlation analyses.

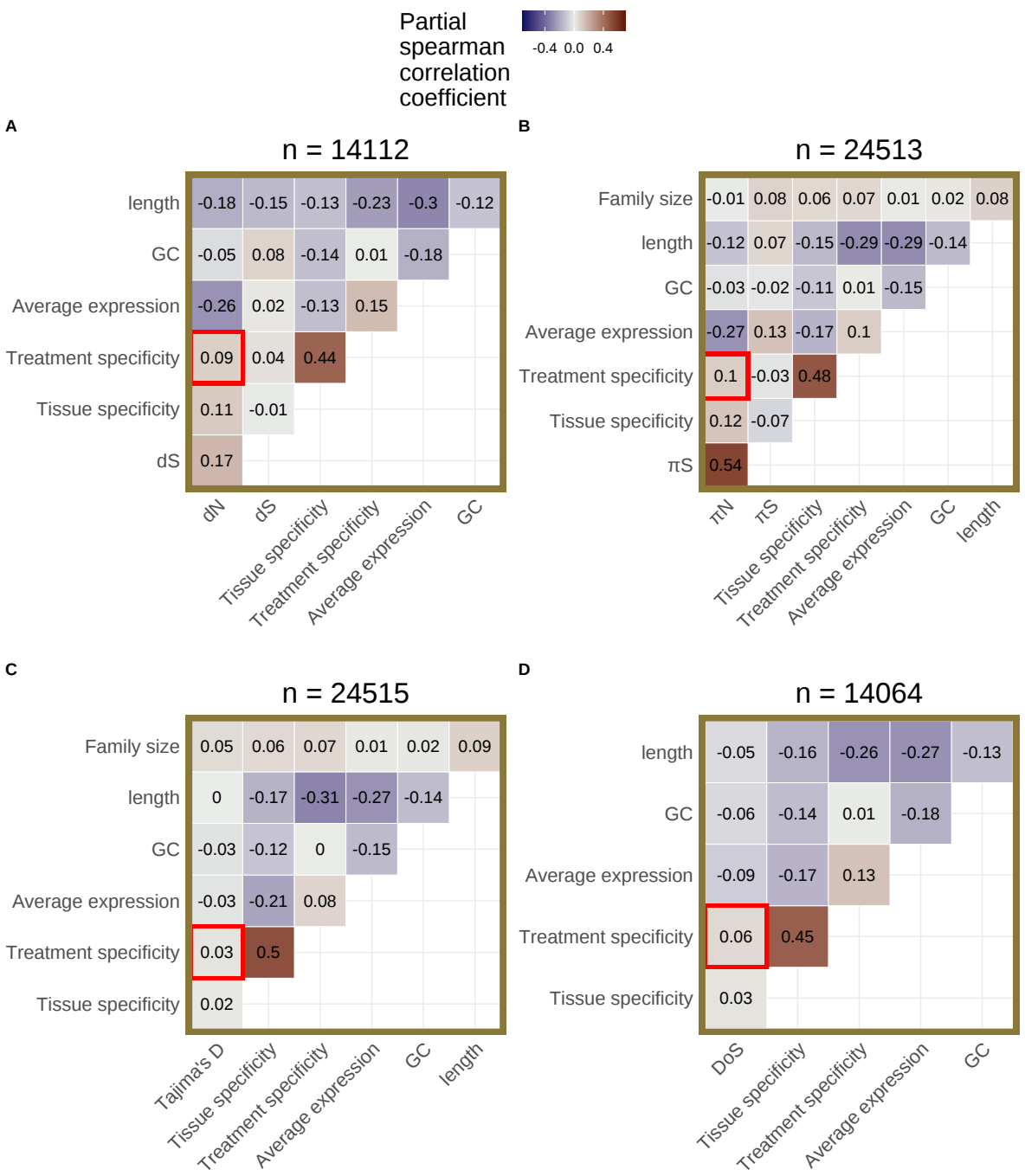

Figure A6: **Partial correlation analysis for root tissue.** Partial correlations for (**A**) dN, (**B**) π_N, (**C**) Tajima's D, and (**D**) direction of selection (DoS) based on root data. Average expression excludes values < 5 TPM and was calculated using only root tissue samples. Tissue-specificity was calculated using only control runs across all tissue categories. Treatment-specificity was calculated using only root tissue runs. The number of genes included in each partial correlation analysis (n) is listed at the top of each heatmap.

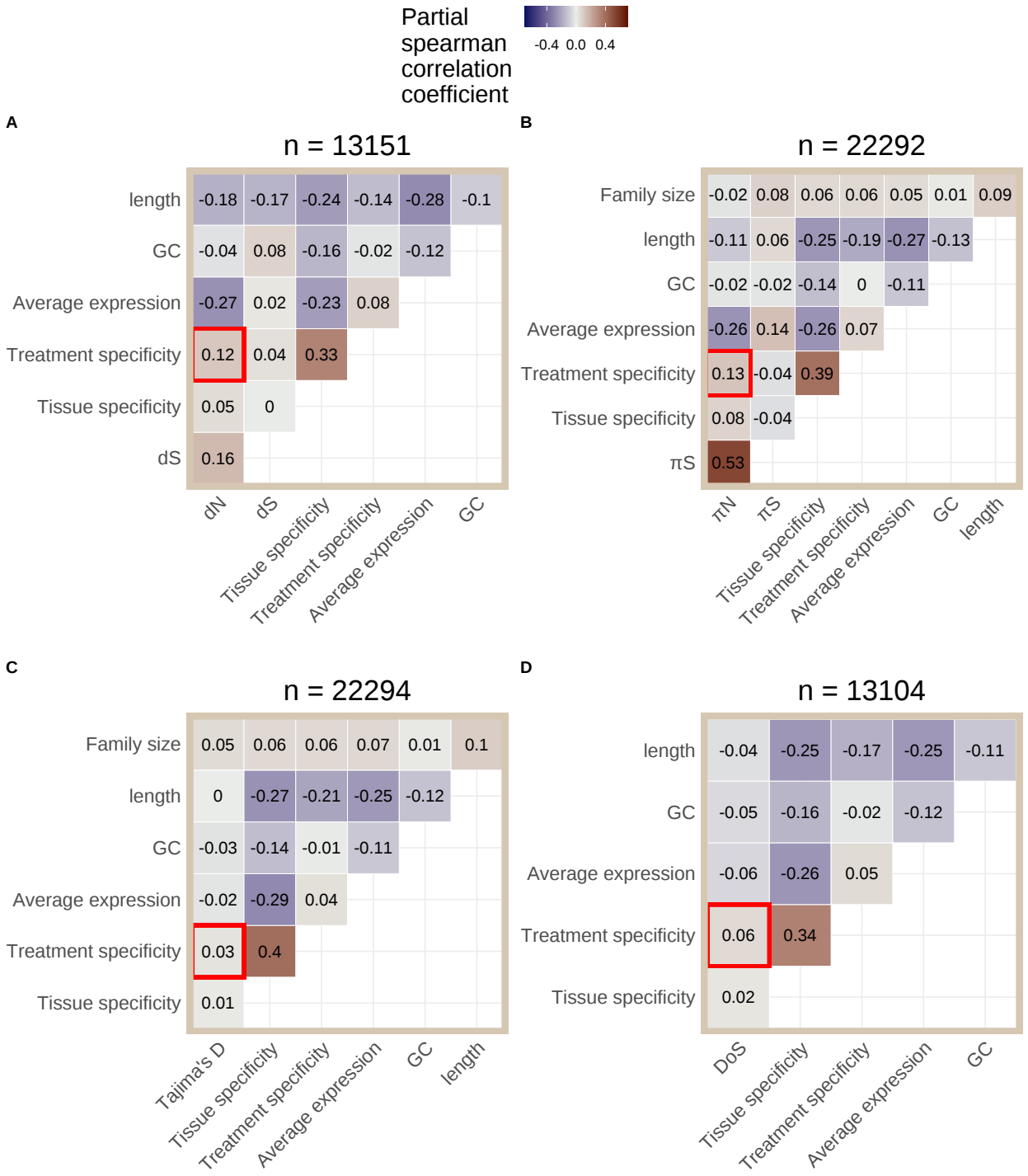

Figure A7: **Partial correlation analysis for whole plant tissue.** Partial correlations for (**A**) dN, (**B**) π_N, (**C**) Tajima's D, and (**D**) direction of selection (DoS) based on whole plant data. Average expression excludes values < 5 TPM and was calculated using only whole plant tissue samples. Tissue-specificity was calculated using only control runs across all tissue categories. Treatment-specificity was calculated using only whole plant tissue runs. The number of genes included in each partial correlation analysis (n) is listed at the top of each heatmap.

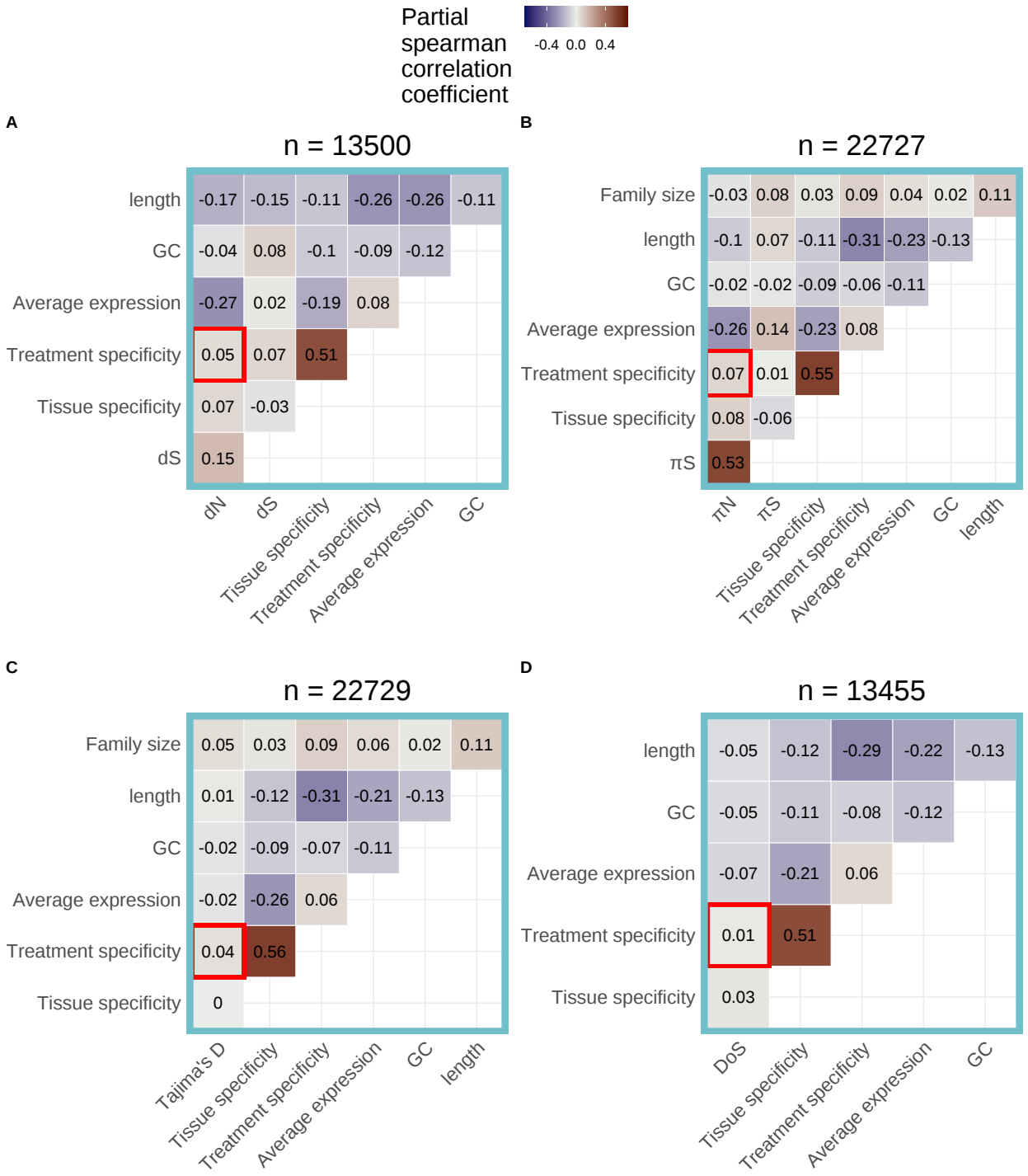

Figure A8: **Partial correlation analysis for shoot tissue.** Partial correlations for (**A**) dN, (**B**) π_N, (**C**) Tajima's D, and (**D**) direction of selection (DoS) based on shoot data. Average expression excludes values < 5 TPM and was calculated using only shoot tissue samples. Tissue-specificity was calculated using only control runs across all tissue categories. Treatment-specificity was calculated using only shoot tissue runs. The number of genes included in each partial correlation analysis (n) is listed at the top of each heatmap.

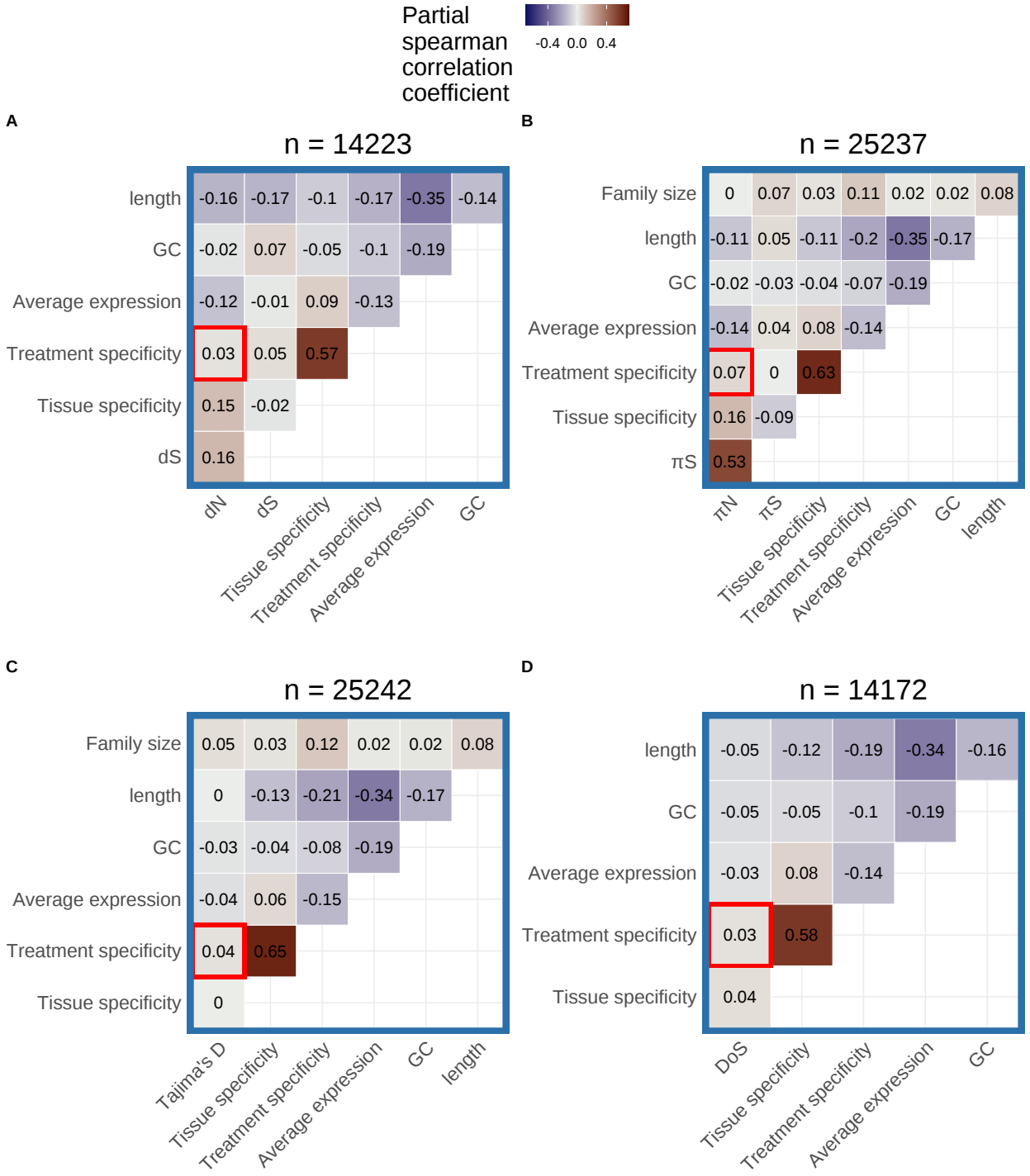

Figure A9: **Partial correlation analysis for seed tissue.** Partial correlations for (**A**) dN, (**B**) π_N, (**C**) Tajima's D, and (**D**) direction of selection (DoS) based on seed data. Average expression excludes values < 5 TPM and was calculated using only seed tissue samples. Tissue-specificity was calculated using only control runs across all tissue categories. Treatment-specificity was calculated using only seed tissue runs. The number of genes included in each partial correlation analysis (n) is listed at the top of each heatmap.

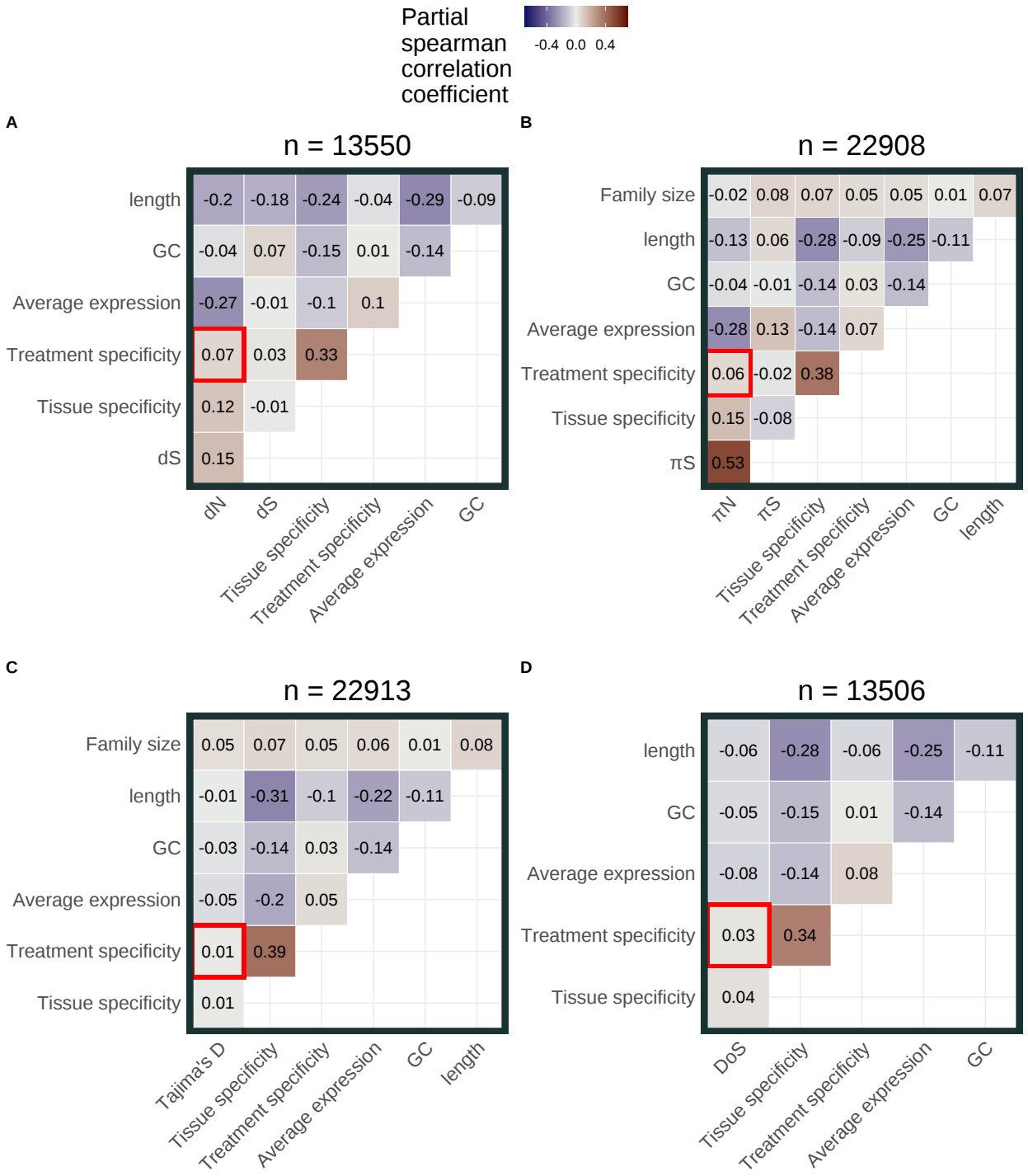

Figure A10: **Partial correlation analysis for fruit or flower tissue.** Partial correlations for (**A**) dN, (**B**) π_N, (**C**) Tajima's D, and (**D**) direction of selection (DoS) based on flower and fruit data. Average expression excludes values < 5 TPM and was calculated using only fruit and flower tissue samples. Tissue-specificity was calculated using only control runs across all tissue categories. Treatment-specificity was calculated using only fruit and flower tissue runs. The number of genes included in each partial correlation analysis (n) is listed at the top of each heatmap.

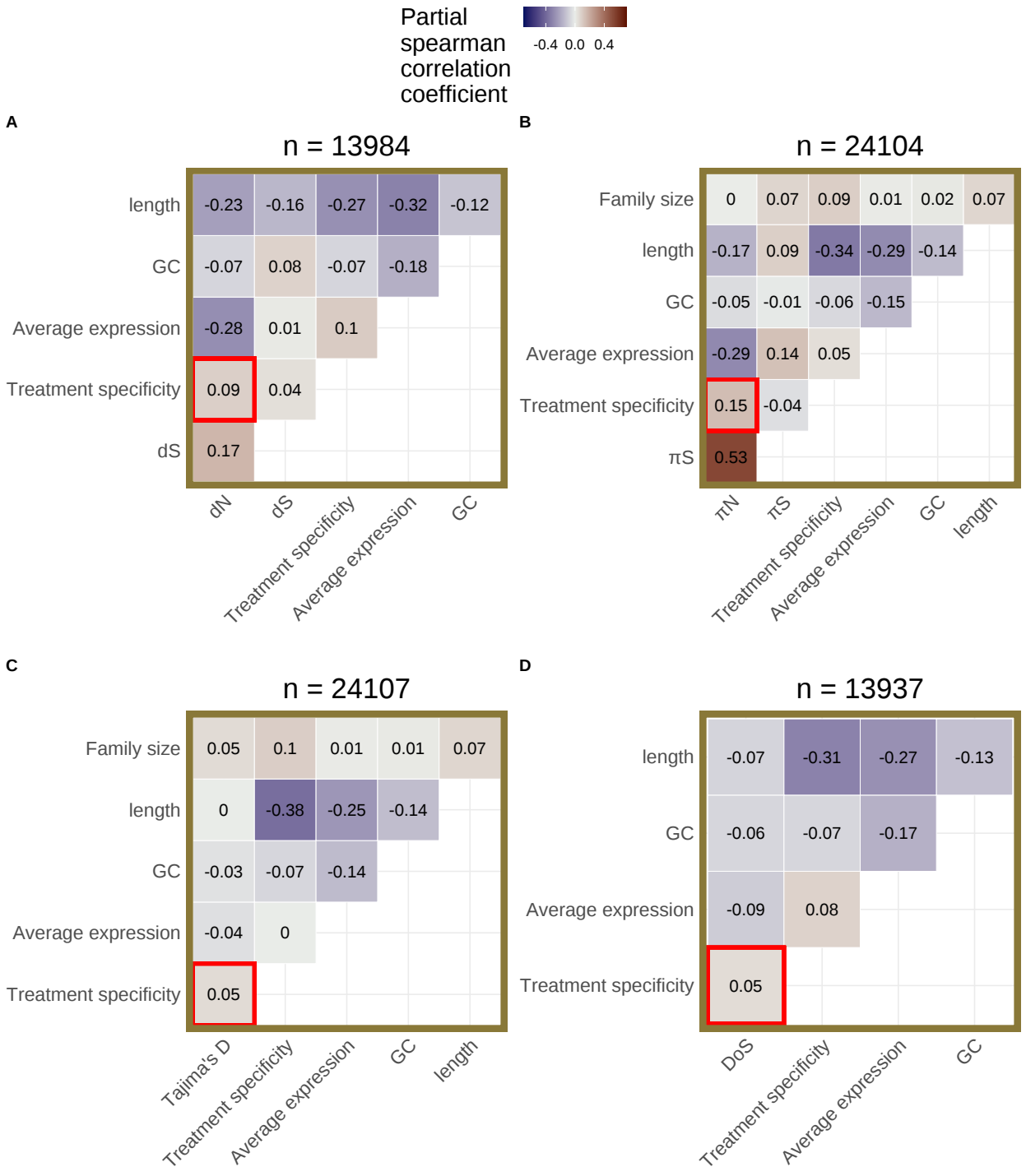

Figure A11: **Partial correlation analysis of root tissue after SVA.** Partial correlations for (**A**) dN, (**B**) π_N, (**C**) Tajima's D, and (**D**) direction of selection (DoS) based on root tissue data after applying SVA. Data was further subset to include only treatment groups with data from more than one study before applying SVA. The number of genes included in each partial correlation analysis (n) is listed at the top of each heatmap.

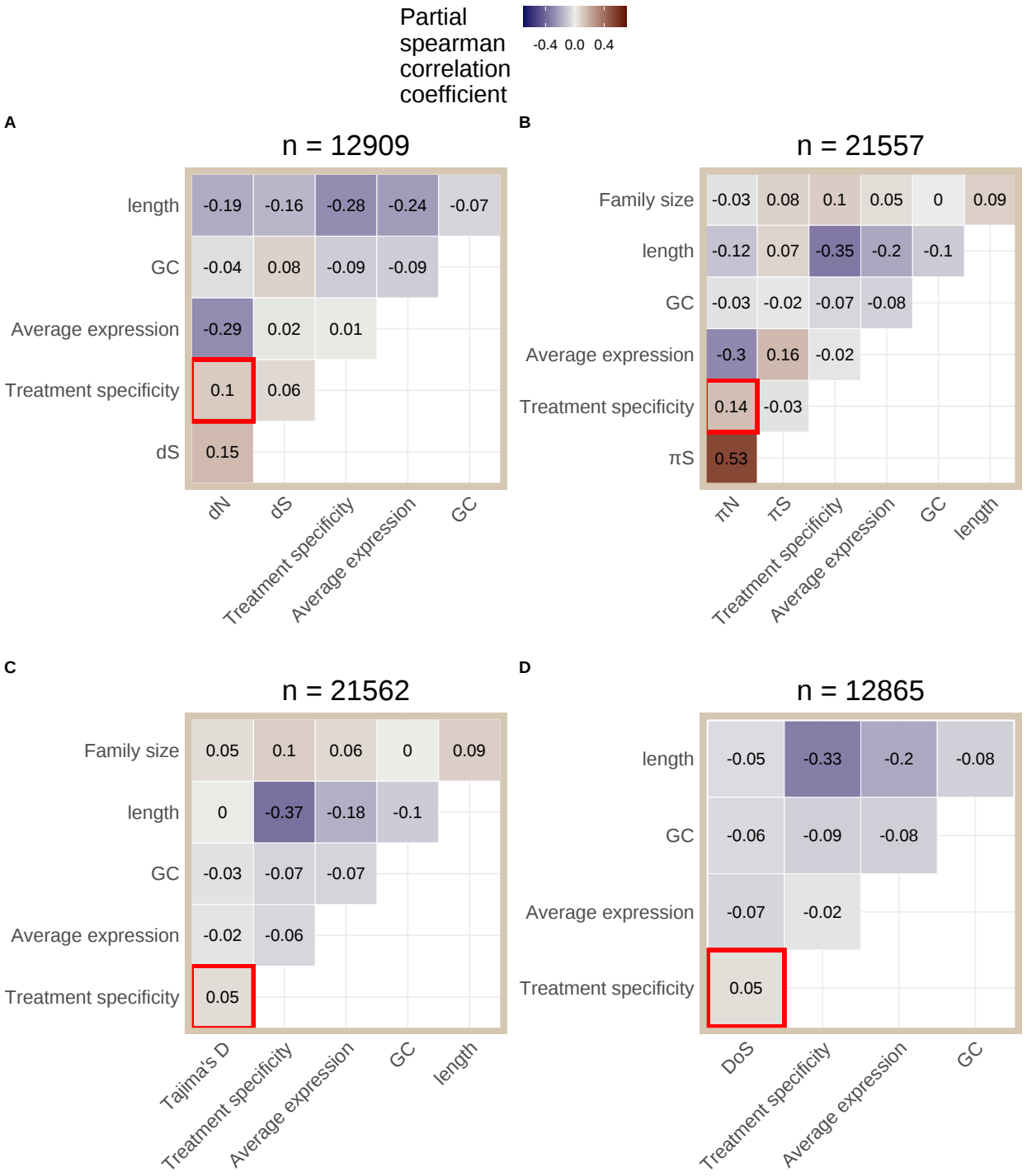

Figure A12: **Partial correlation analysis of whole plant tissue after SVA.** Partial correlations for (**A**) dN, (**B**) π_N, (**C**) Tajima's D, and (**D**) direction of selection (DoS) based on whole plant tissue data after applying SVA. Data was further subset to include only treatment groups with data from more than one study before applying SVA. The number of genes included in each partial correlation analysis (n) is listed at the top of each heatmap.

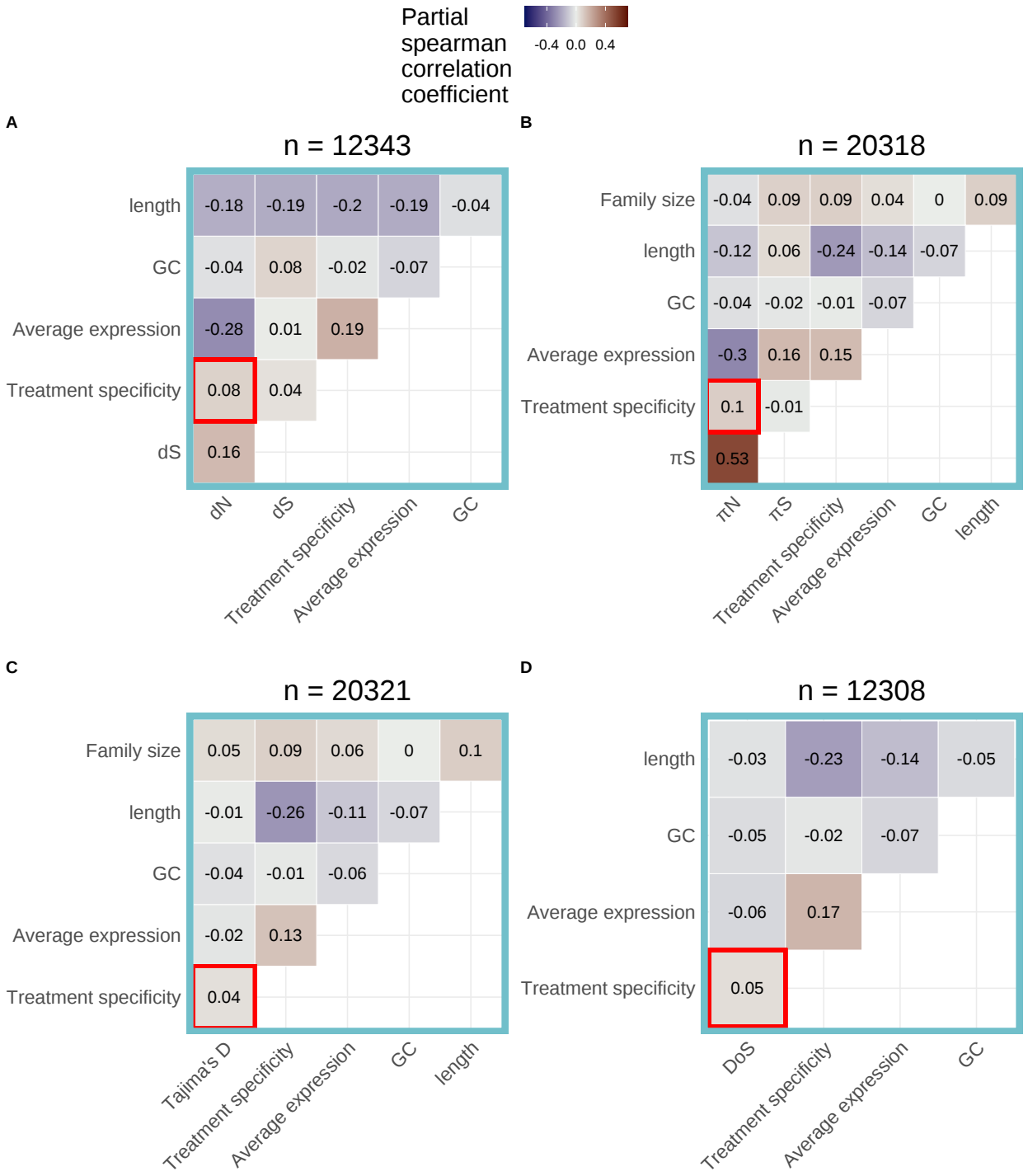

Figure A13: **Partial correlation analysis of shoot tissue after SVA.** Partial correlations for (**A**) dN, (**B**) π_N, (**C**) Tajima's D, and (**D**) direction of selection (DoS) based on shoot tissue data after applying SVA. Data was further subset to include only treatment groups with data from more than one study before applying SVA. The number of genes included in each partial correlation analysis (n) is listed at the top of each heatmap.

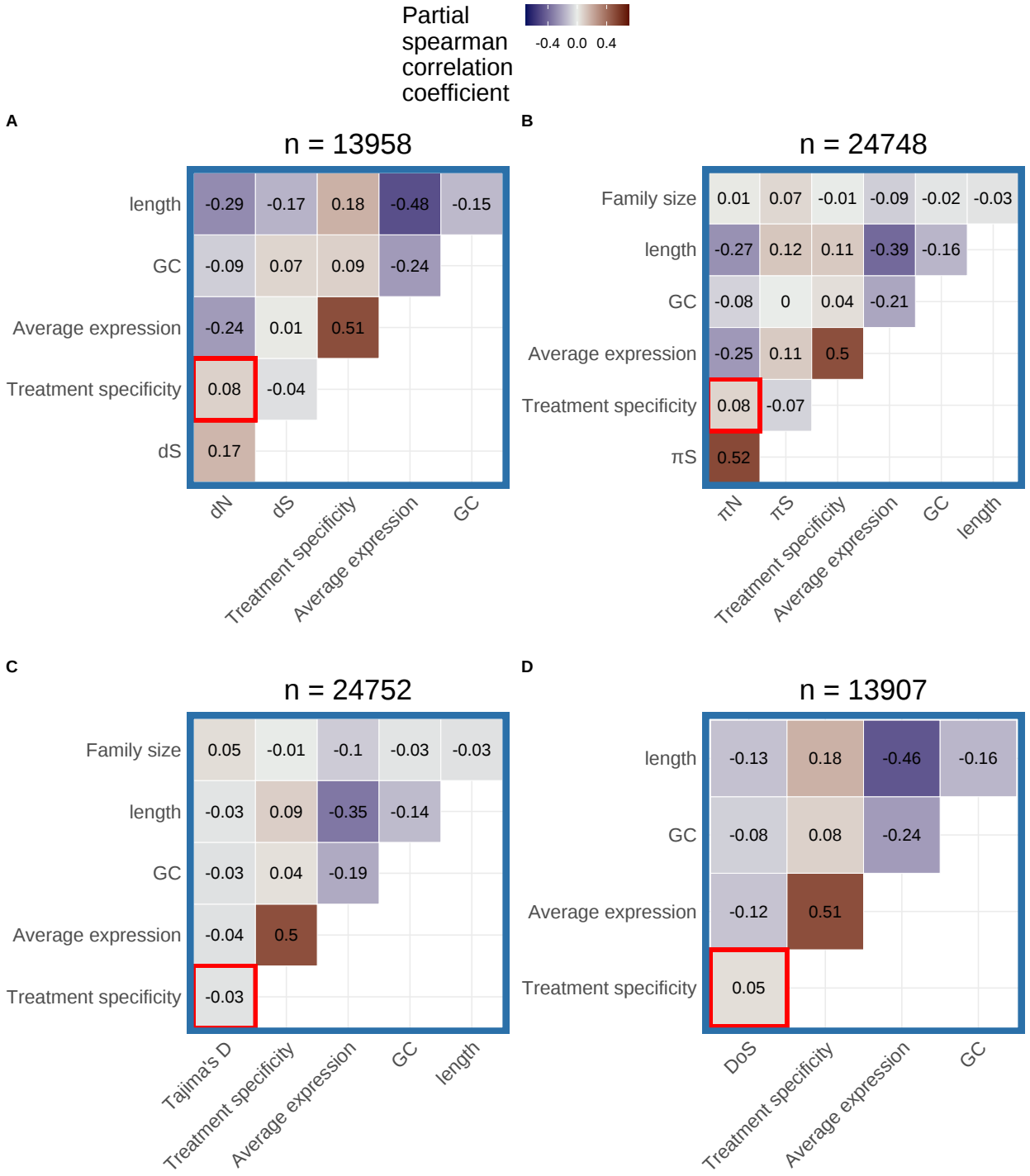

Figure A14: **Partial correlation analysis of seed tissue after SVA.**Partial correlations for (**A**) dN, (**B**) π_N, (**C**) Tajima's D, and (**D**) direction of selection (DoS) based on seed tissue data after applying SVA. Data was further subset to include only treatment groups with data from more than one study before applying SVA. The number of genes included in each partial correlation analysis (n) is listed at the top of each heatmap.

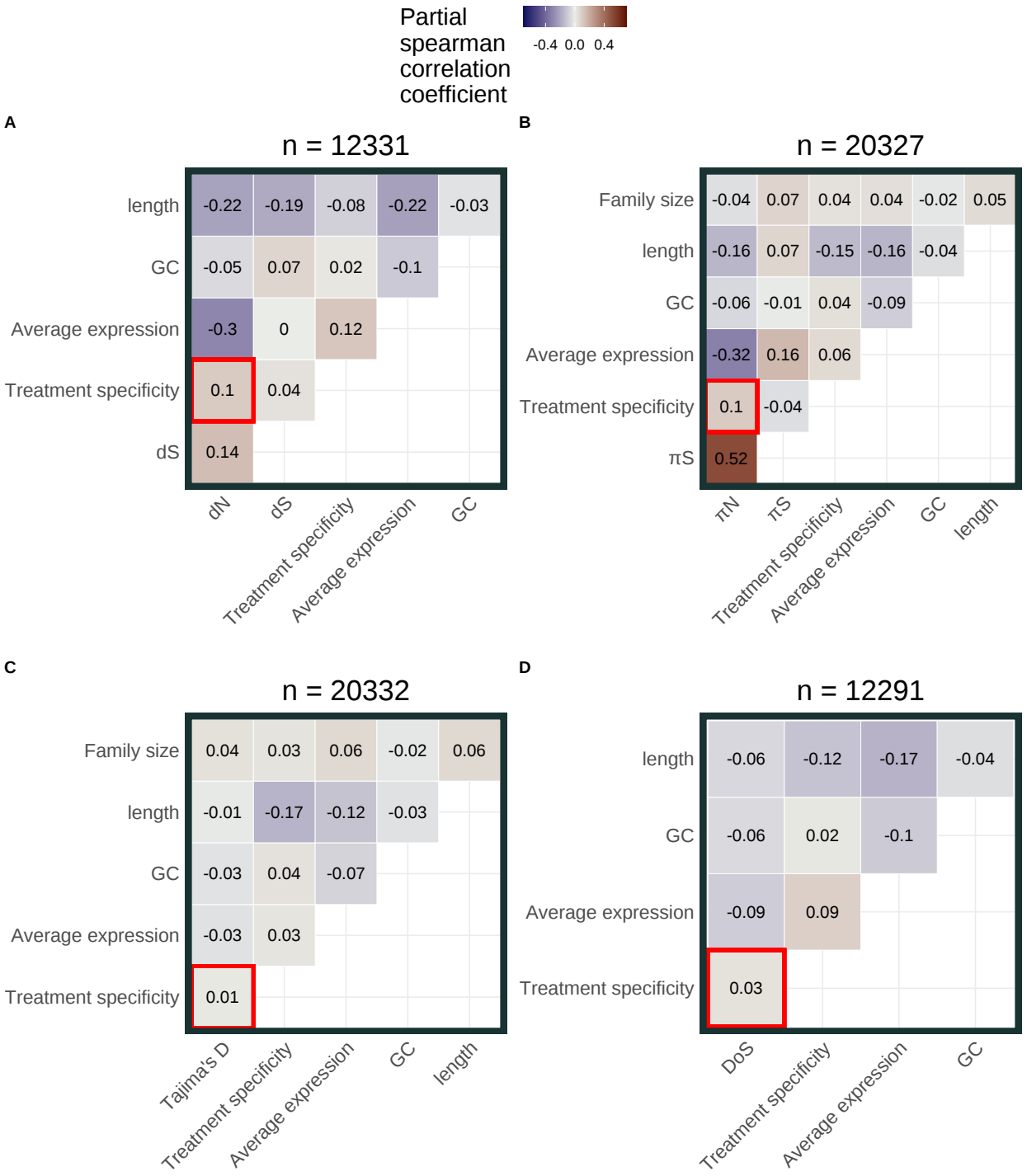

Figure A15: **Partial correlation analysis of fruit or flower tissue after SVA.**Partial correlations for (**A**) dN, (**B**) π_N, (**C**) Tajima's D, and (**D**) direction of selection (DoS) based on flower and fruit tissue data after applying SVA. Data was further subset to include only treatment groups with data from more than one study before applying SVA. The number of genes included in each partial correlation analysis (n) is listed at the top of each heatmap.

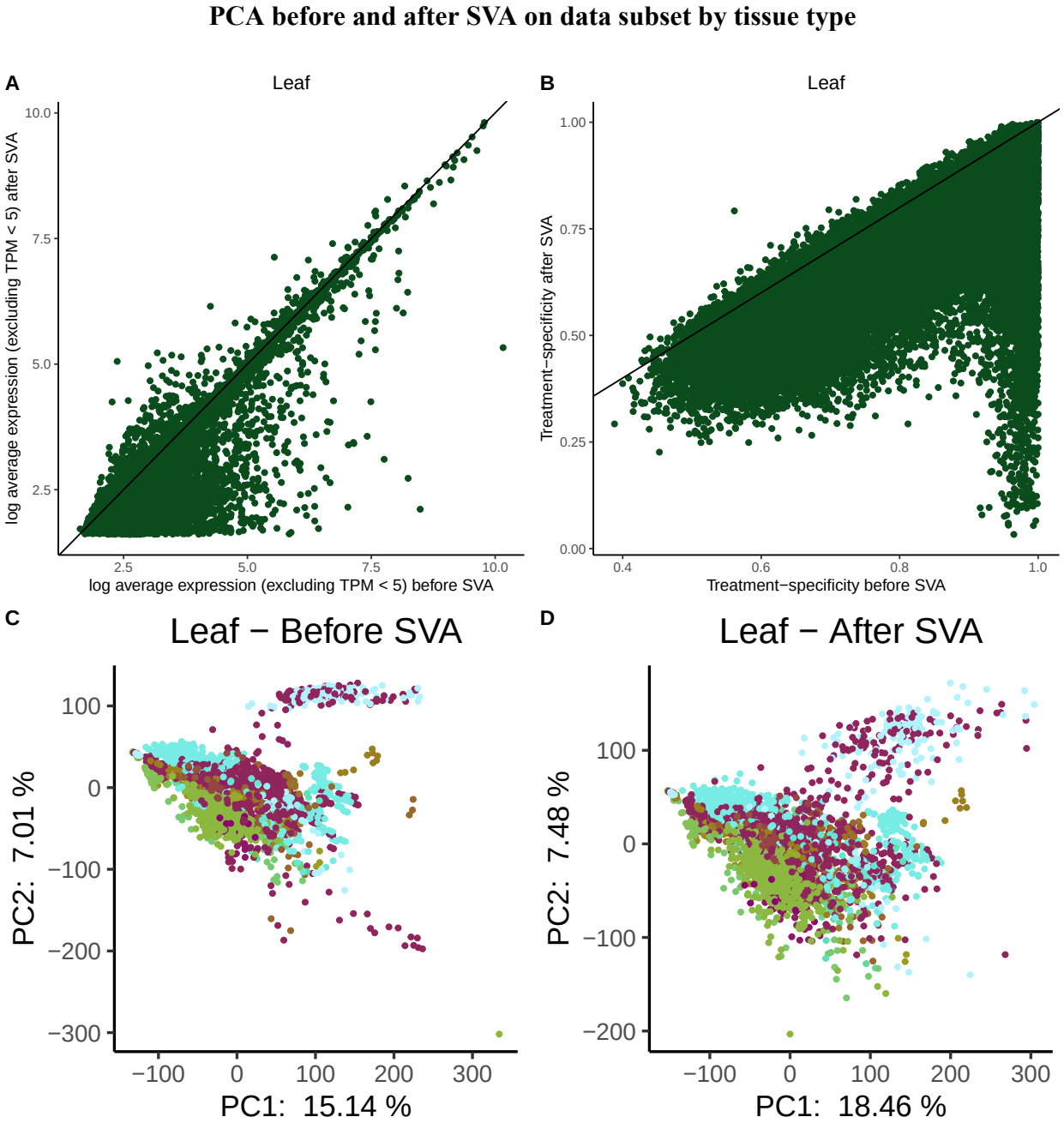

Figure A16: **Effect of surrogate variable analysis on structure of leaf tissue data.** (**A**) Effect of SVA on average expression. (**B**) Effect of SVA on treatment-specificity. (**C**) Principal component analysis of leaf tissue data before SVA. (**D**) Principal component analysis of leaf tissue data after SVA. Color in the PCA plots represent different experimental treatments - legend is Figure A22.

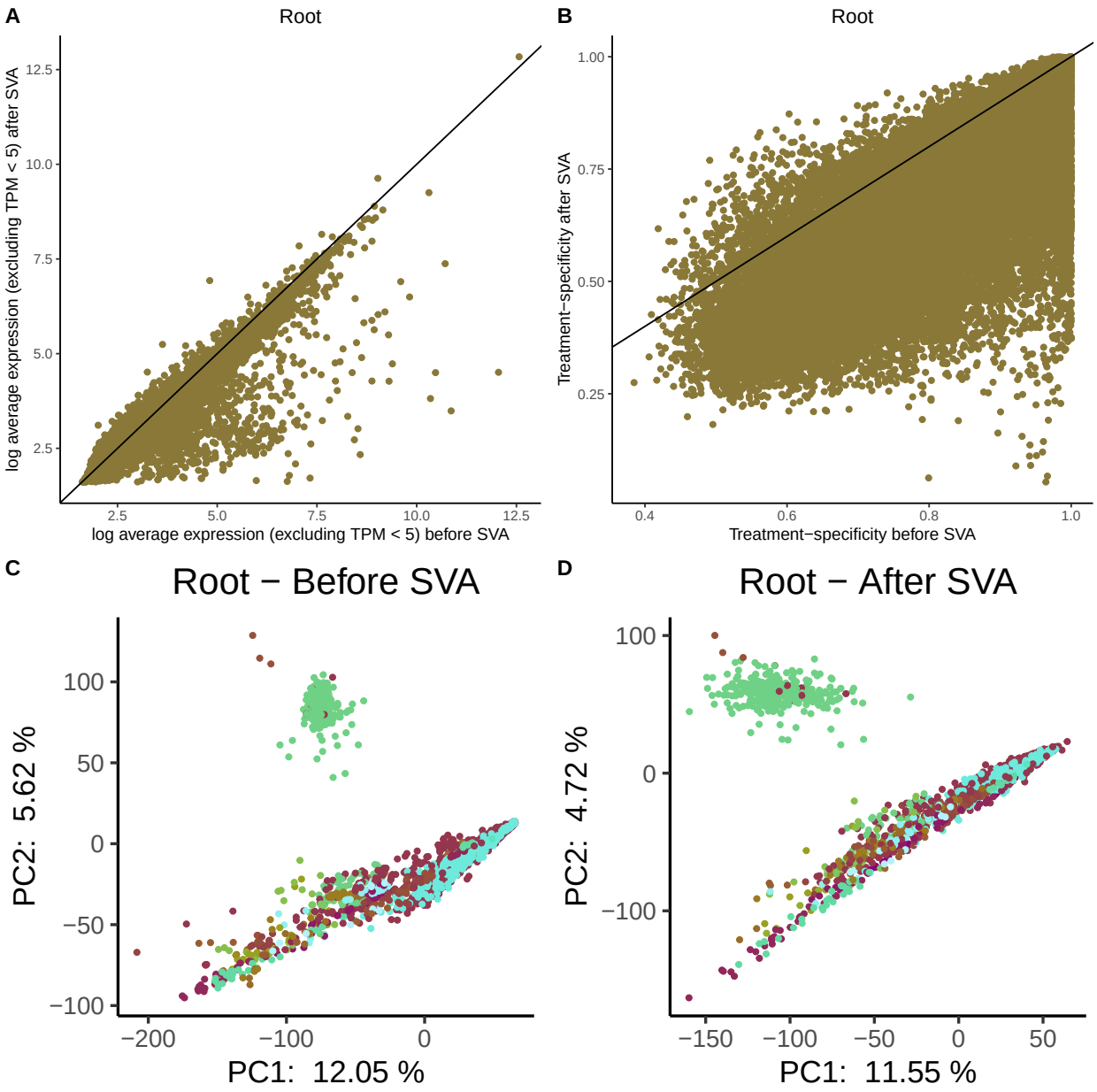

Figure A17: **Effect of surrogate variable analysis on structure of root tissue data.** (**A**) Effect of SVA on average expression. (**B**) Effect of SVA on treatment-specificity. (**C**) Principal component analysis of root tissue data before SVA. (**D**) Principal component analysis of root tissue data after SVA.Color in the PCA plots represent different experimental treatments - legend is Figure A23.

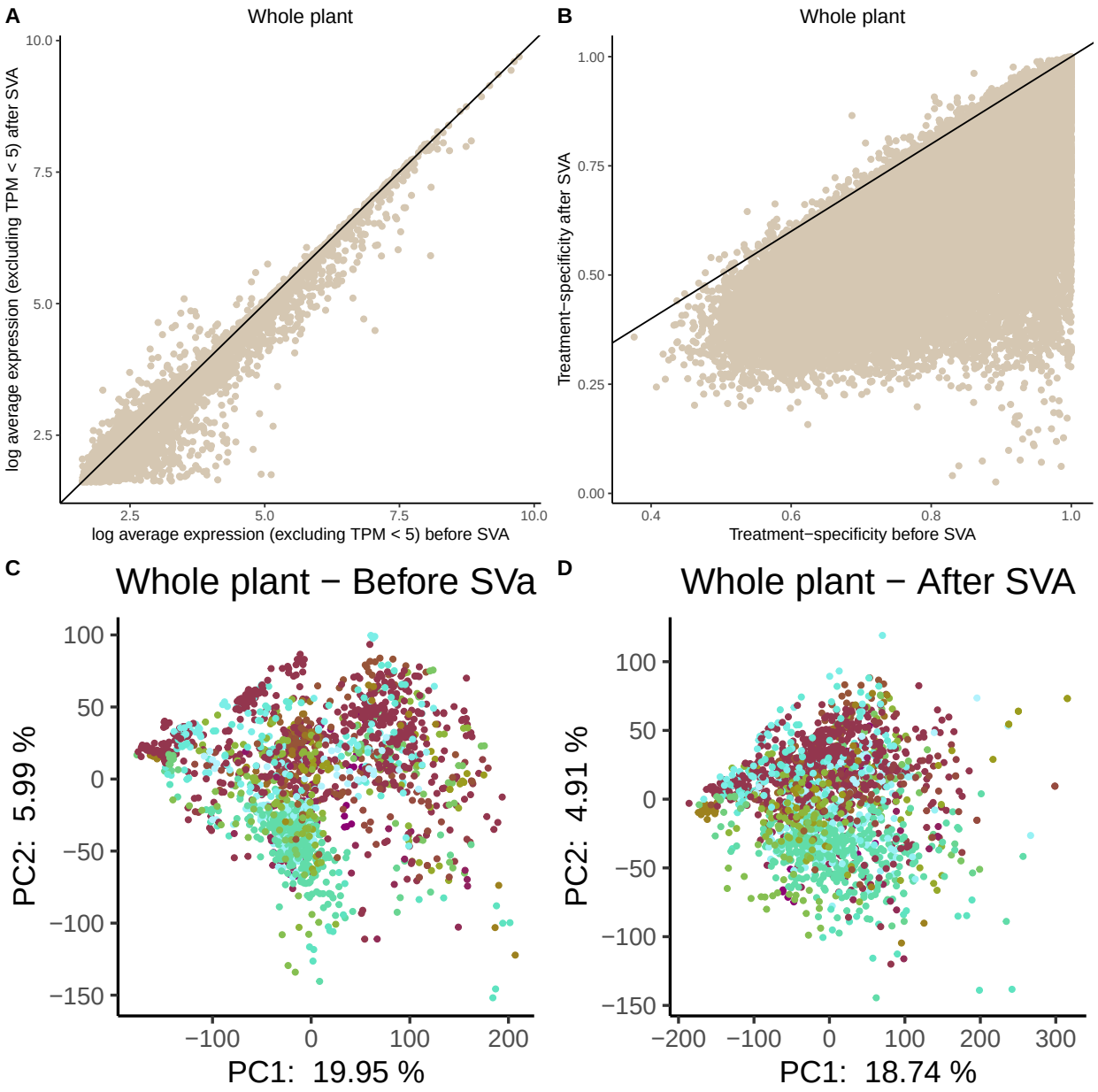

Figure A18: **Effect of surrogate variable analysis on structure of whole plant tissue data.** (**A**) Effect of SVA on average expression. (**B**) Effect of SVA on treatment-specificity. (**C**) Principal component analysis of whole plant tissue data before SVA. (**D**) Principal component analysis of whole plant tissue data after SVA.Color in the PCA plots represent different experimental treatments - legend is Figure A24.

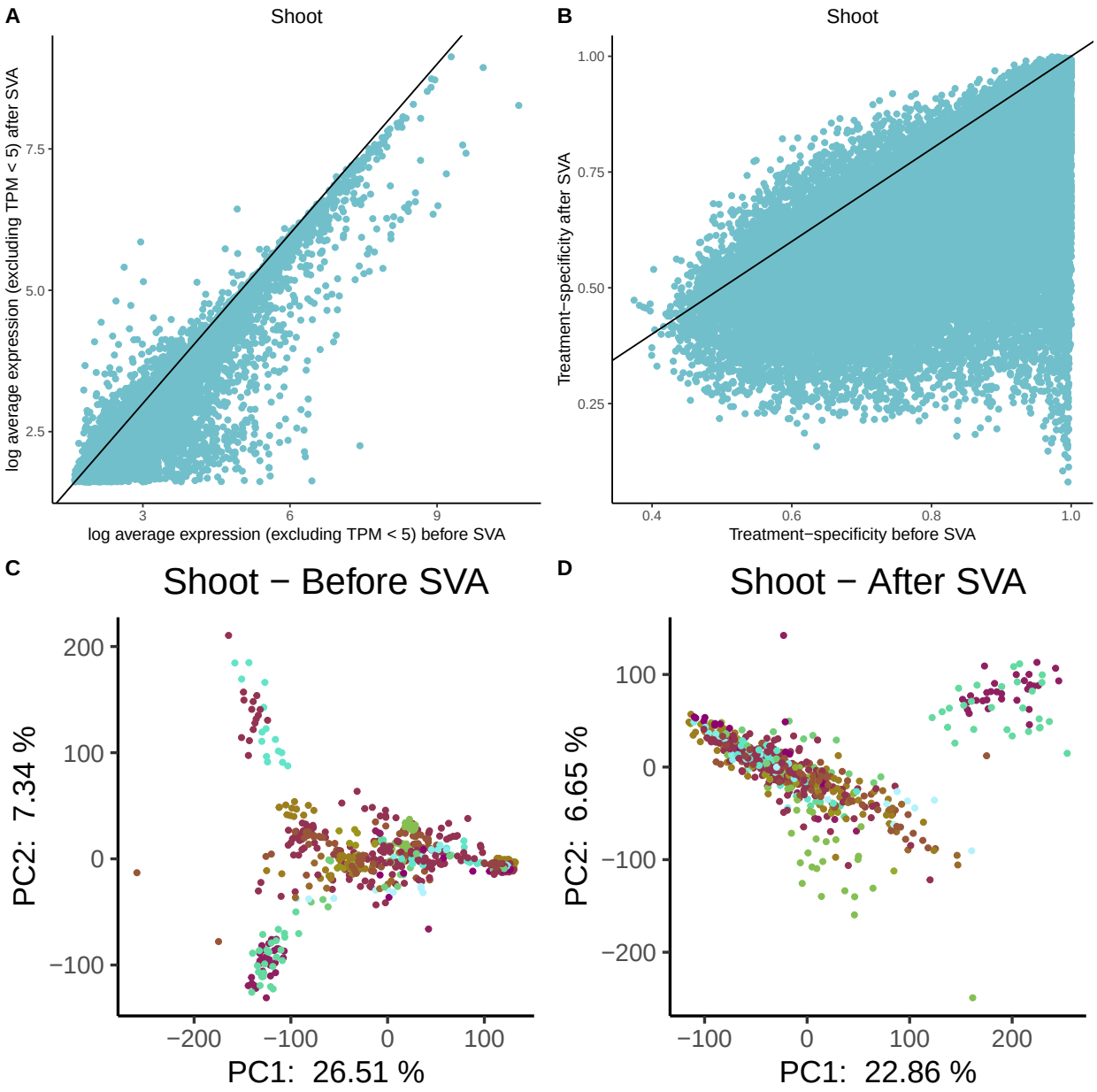

Figure A19: **Effect of surrogate variable analysis on structure of shoot tissue data.** (**A**) Effect of SVA on average expression. (**B**) Effect of SVA on treatment-specificity. (**C**) Principal component analysis of shoot tissue data before SVA. (**D**) Principal component analysis of shoot tissue data after SVA. Color in the PCA plots represent different experimental treatments - legend is Figure A25.

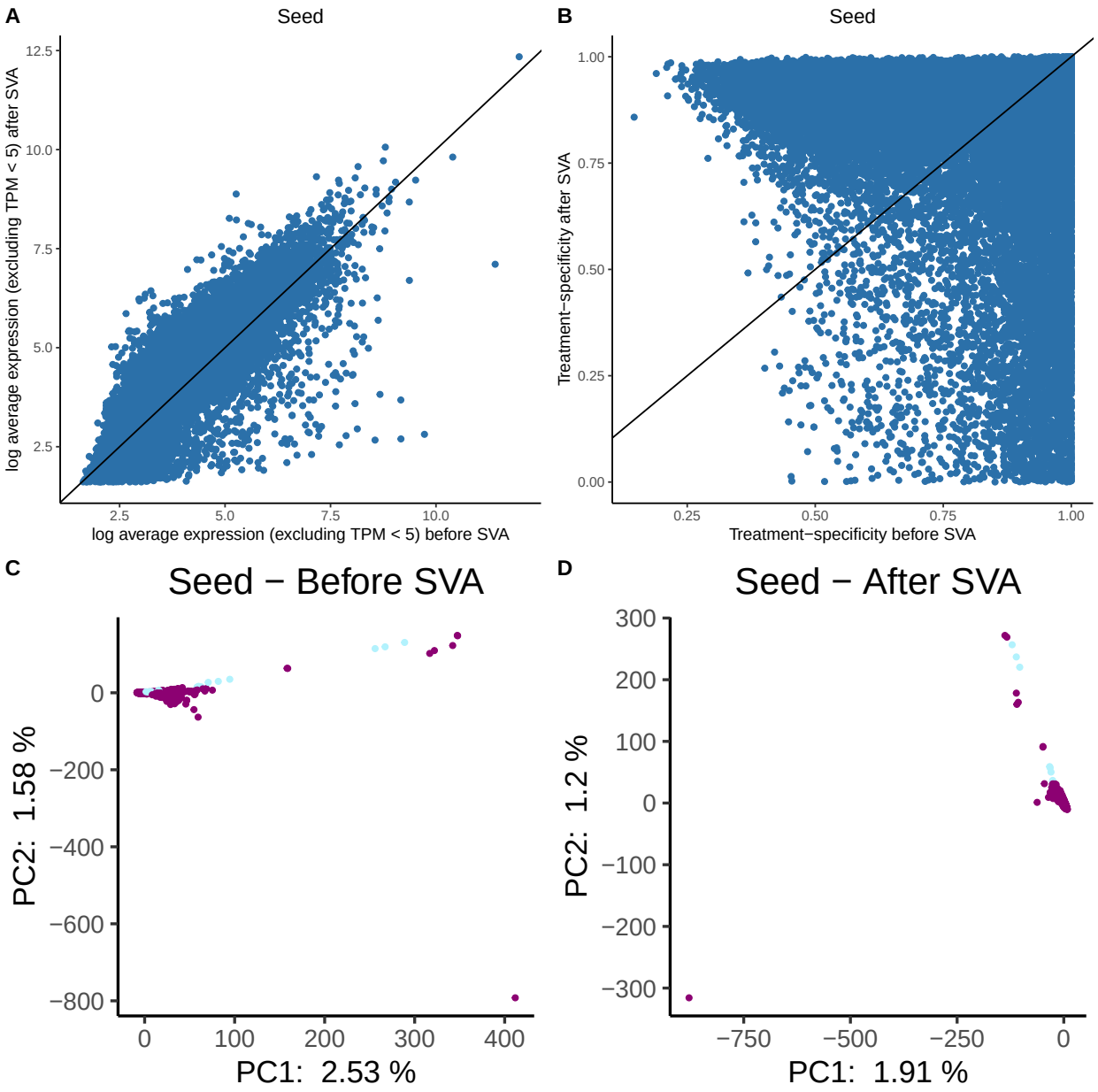

Figure A20: **Effect of surrogate variable analysis on structure of seed tissue data.** (**A**) Effect of SVA on average expression. (**B**) Effect of SVA on treatment-specificity. (**C**) Principal component analysis of seed tissue data before SVA. (**D**) Principal component analysis of seed tissue data after SVA. Color in the PCA plots represent different experimental treatments - legend is Figure A26.

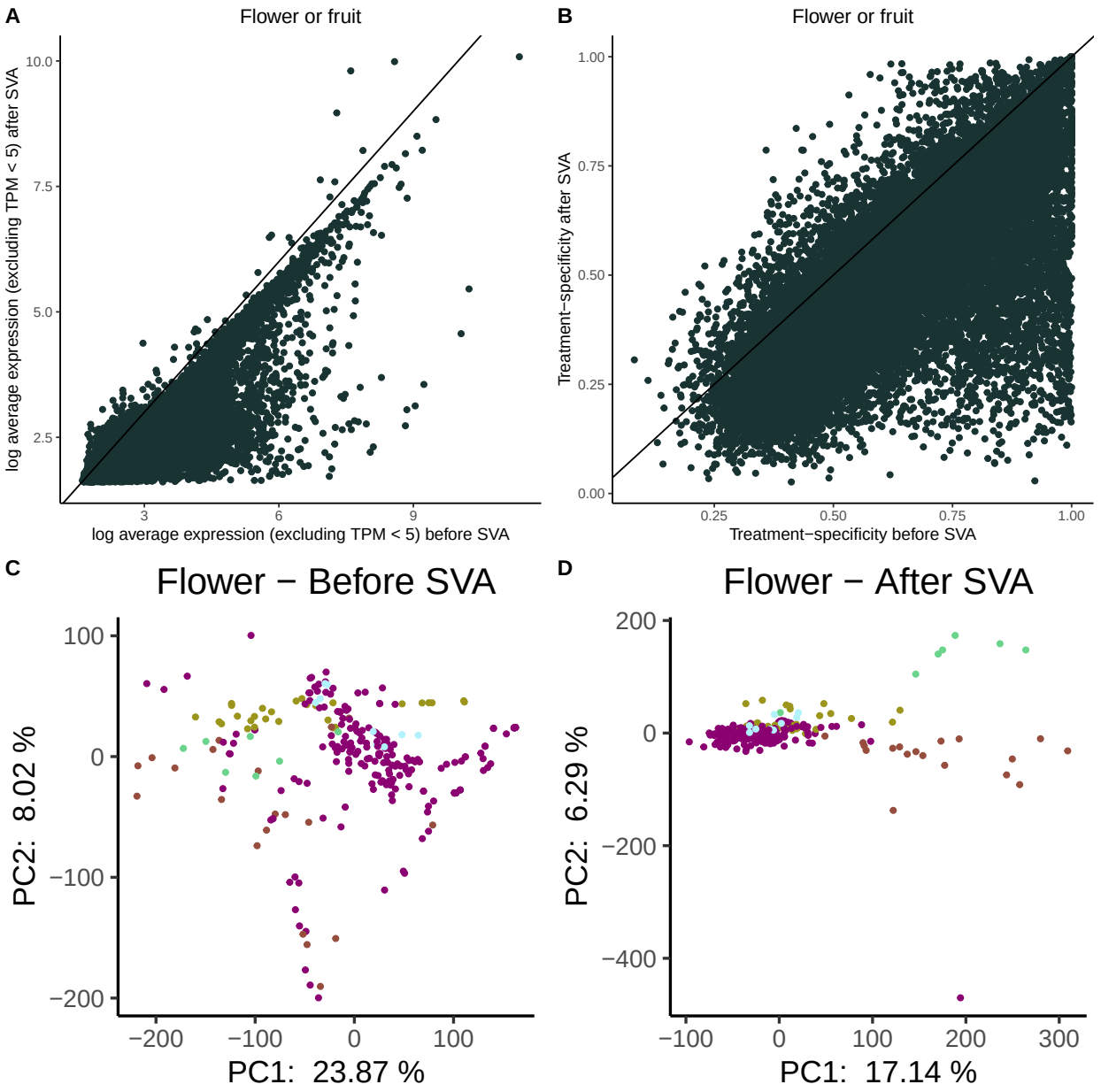

Figure A21: **Effect of surrogate variable analysis on structure of fruit and flower tissue data.** (**A**) Effect of SVA on average expression. (**B**) Effect of SVA on treatment-specificity. (**C**) Principal component analysis of fruit and flower tissue data before SVA. (**D**) Principal component analysis of fruit and flower tissue data after SVA. Color in the PCA plots represent different experimental treatments - legend is Figure A27.

Legends - PCA before and after SVA on data subset by tissue type

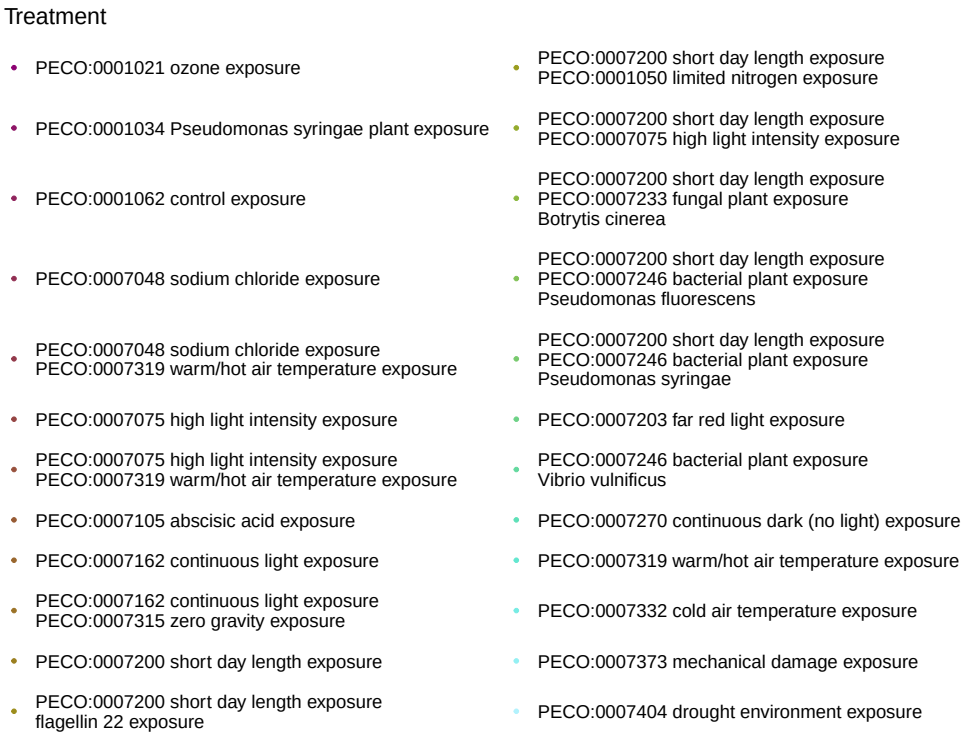

Figure A22: **Legend for PCA before and after SVA, leaf data.**

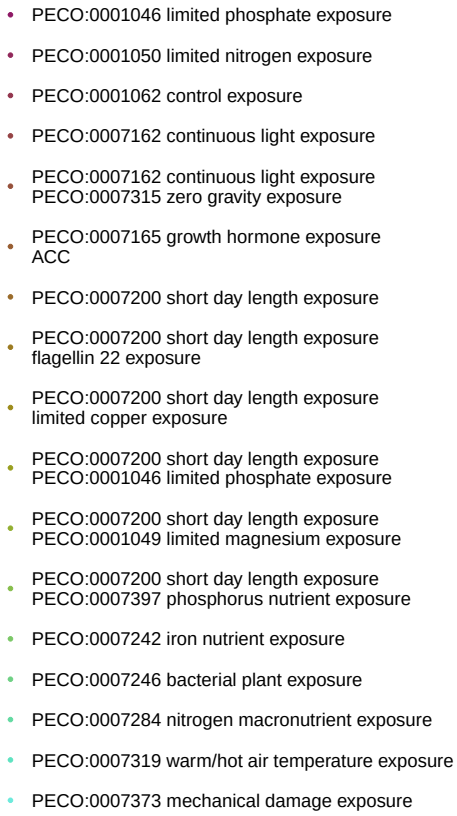

Figure A23: **Legend for PCA before and after SVA, root data.**

Figure A24: **Legend for PCA before and after SVA, whole plant data.**

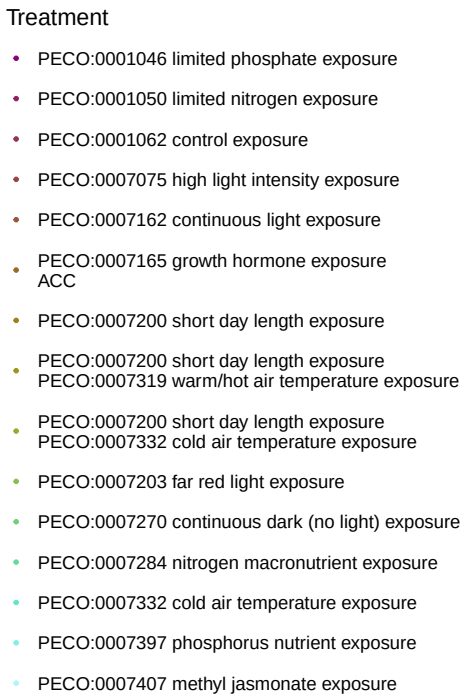

Figure A25: **Legend for PCA before and after SVA, shoot data.**

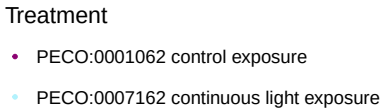

Figure A26: **Legend for PCA before and after SVA, seed data.**

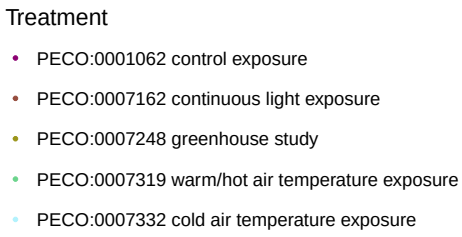

Figure A27: **Legend for PCA before and after SVA, flower data.**

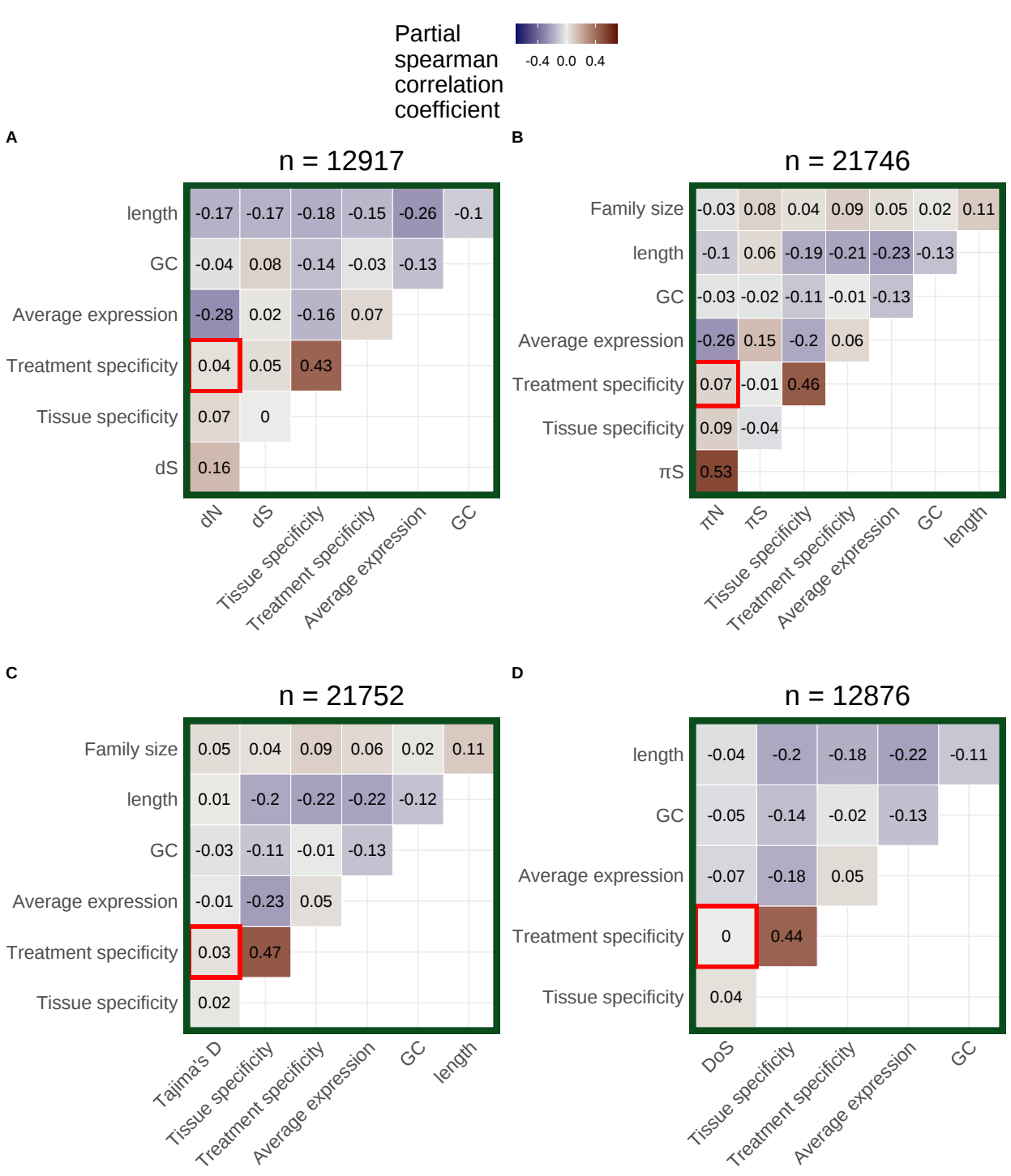

Figure A28: **Partial correlation analysis of leaf tissue on balanced data subset after SVA.** Partial correlations for (**A**) dN, (**B**) π_N, (**C**) Tajima's D, and (**D**) direction of selection (DoS) after applying SVA. Average expression and tissue specificity are calculated using only leaf tissue data. The number of genes included in each partial correlation analysis (n) is listed at the top of each heatmap.

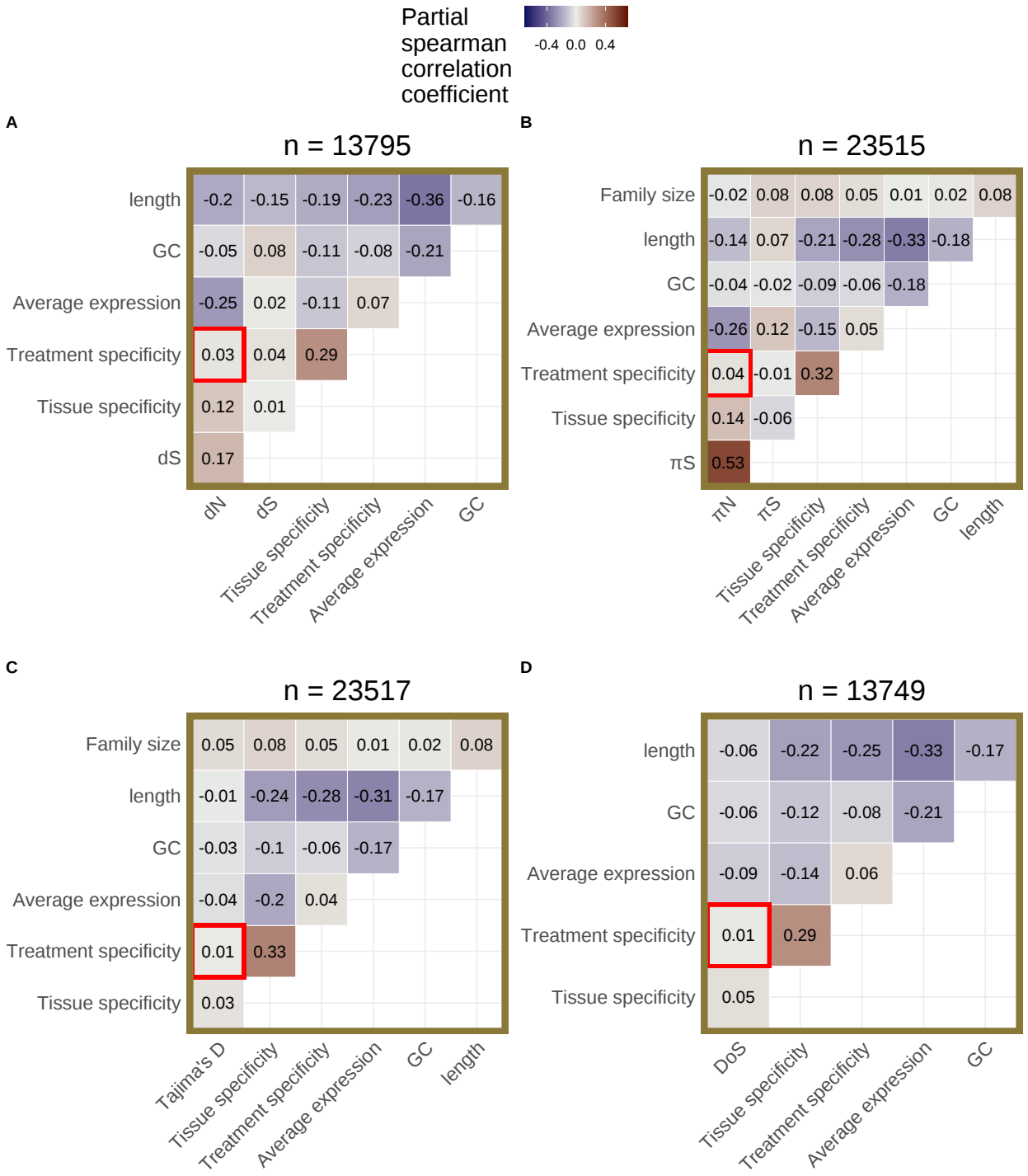

Figure A29: **Partial correlation analysis of root tissue on balanced data subset after SVA.** Partial correlations for (**A**) dN, (**B**) π_N, (**C**) Tajima's D, and (**D**) direction of selection (DoS) after applying SVA. Average expression and tissue specificity are calculated using only root tissue data. The number of genes included in each partial correlation analysis (n) is listed at the top of each heatmap.

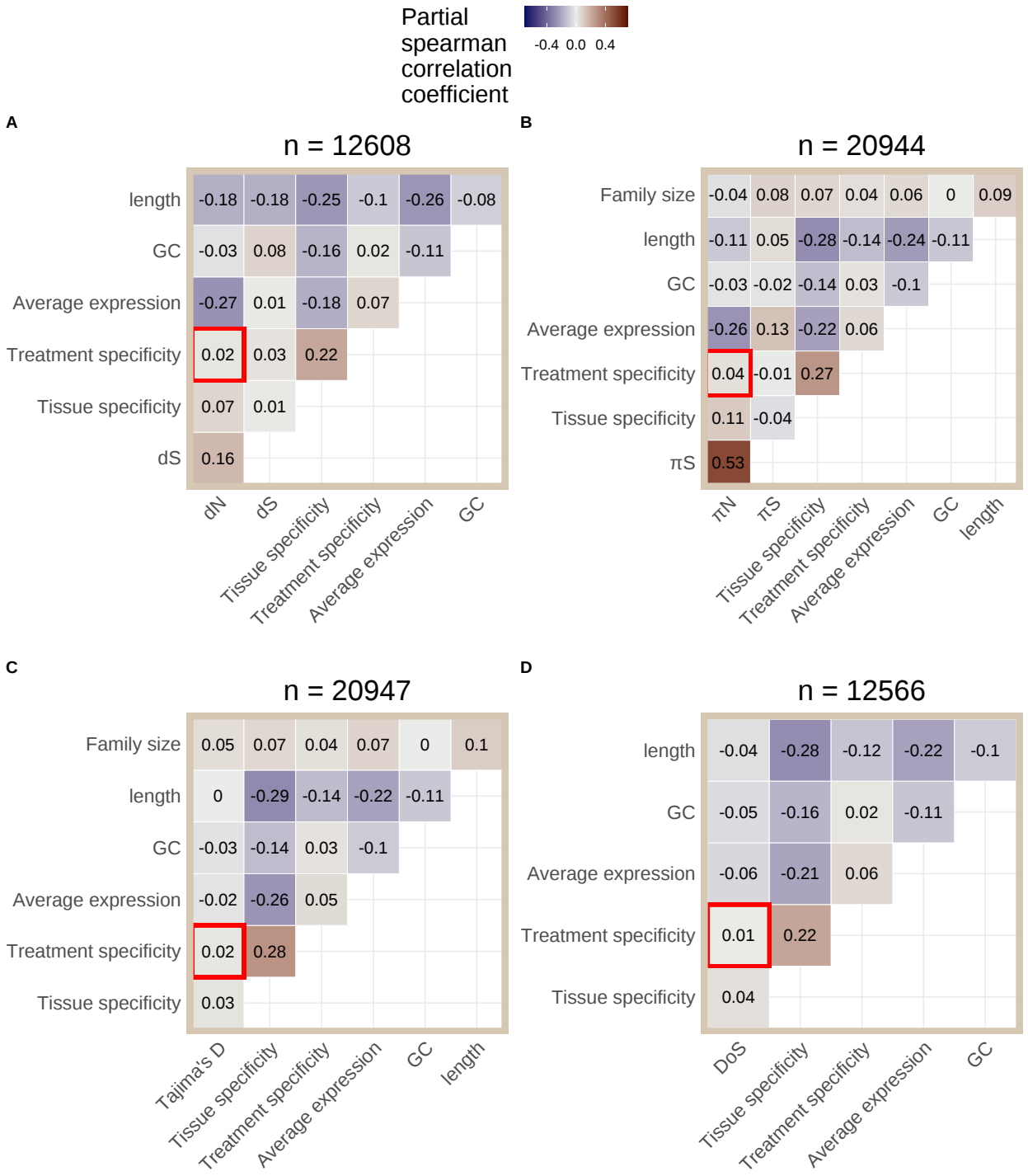

Figure A30: **Partial correlation analysis of whole plant tissue on balanced data subset after SVA.** Partial correlations for (**A**) dN, (**B**) π_N, (**C**) Tajima's D, and (**D**) direction of selection (DoS) after applying SVA. Average expression and tissue specificity are calculated using only whole plant tissue data. The number of genes included in each partial correlation analysis (n) is listed at the top of each heatmap.

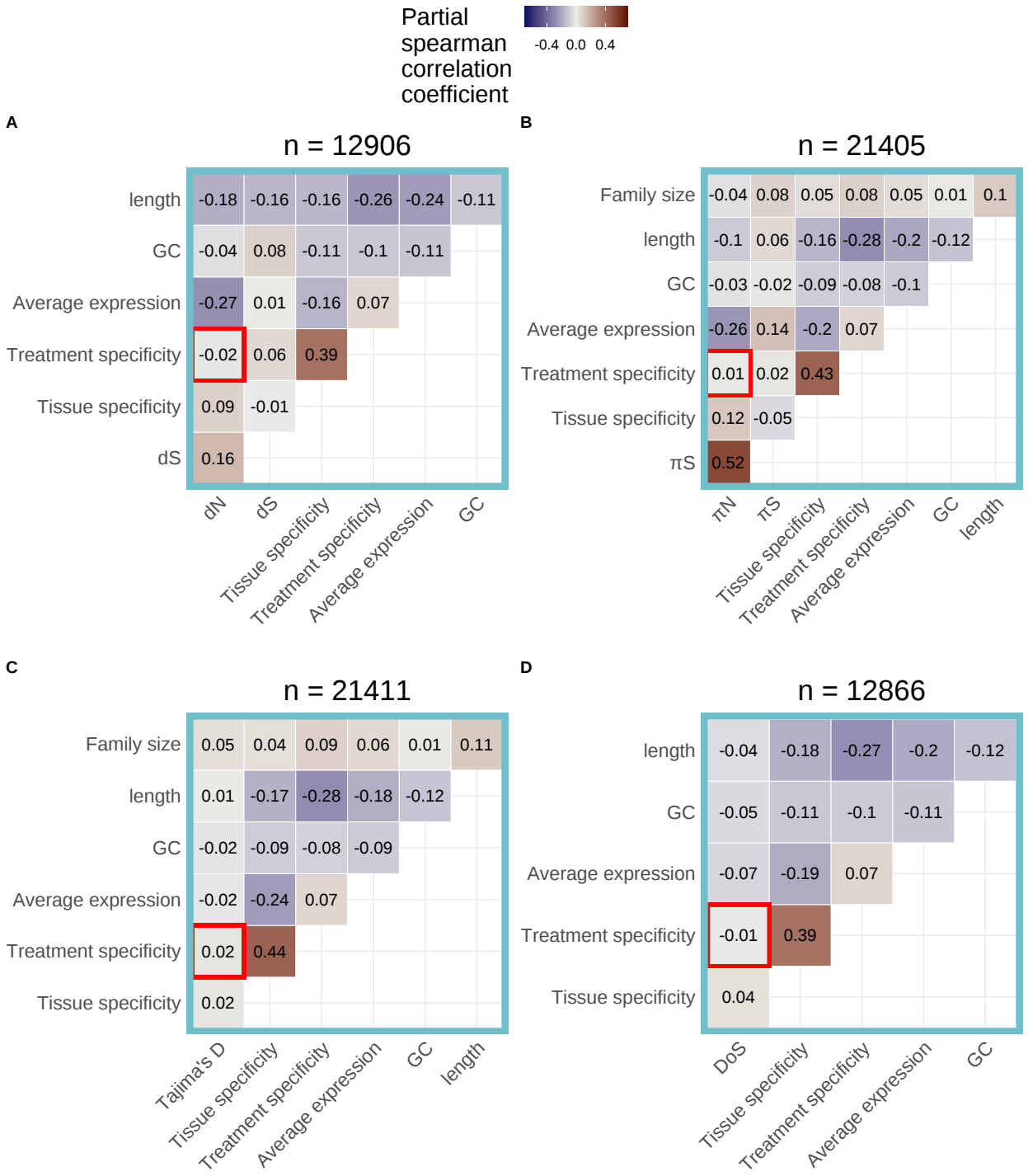

Figure A31: **Partial correlation analysis of shoot tissue on balanced data subset after SVA.** Partial correlations for (**A**) dN, (**B**) π_N, (**C**) Tajima's D, and (**D**) direction of selection (DoS) after applying SVA. Average expression and tissue specificity are calculated using only shoot tissue data. The number of genes included in each partial correlation analysis (n) is listed at the top of each heatmap.

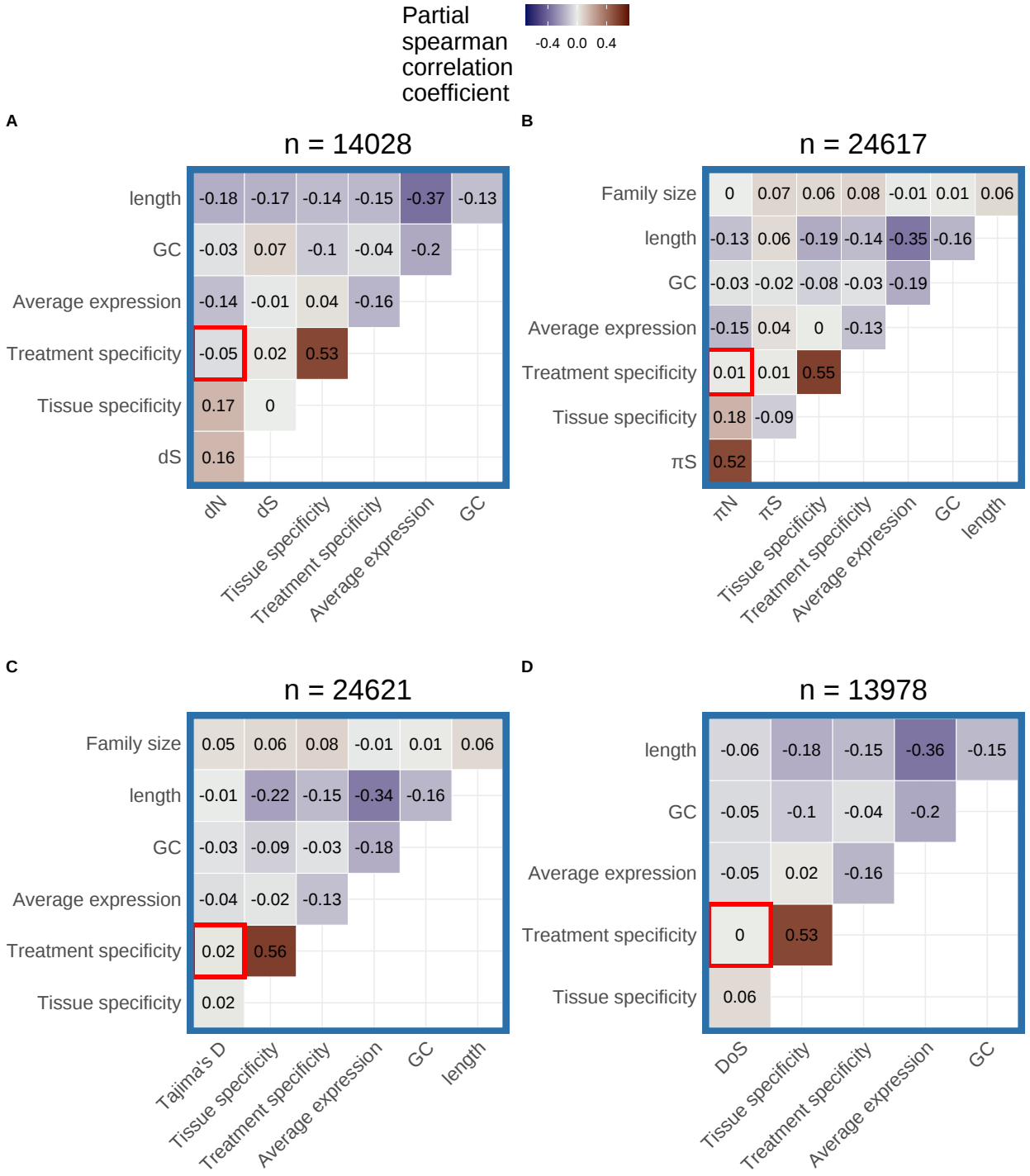

Figure A32: **Partial correlation analysis of seed tissue on balanced data subset after SVA.** Partial correlations for (**A**) dN, (**B**) π_N, (**C**) Tajima's D, and (**D**) direction of selection (DoS) after applying SVA. Average expression and tissue specificity are calculated using only seed tissue data. The number of genes included in each partial correlation analysis (n) is listed at the top of each heatmap.

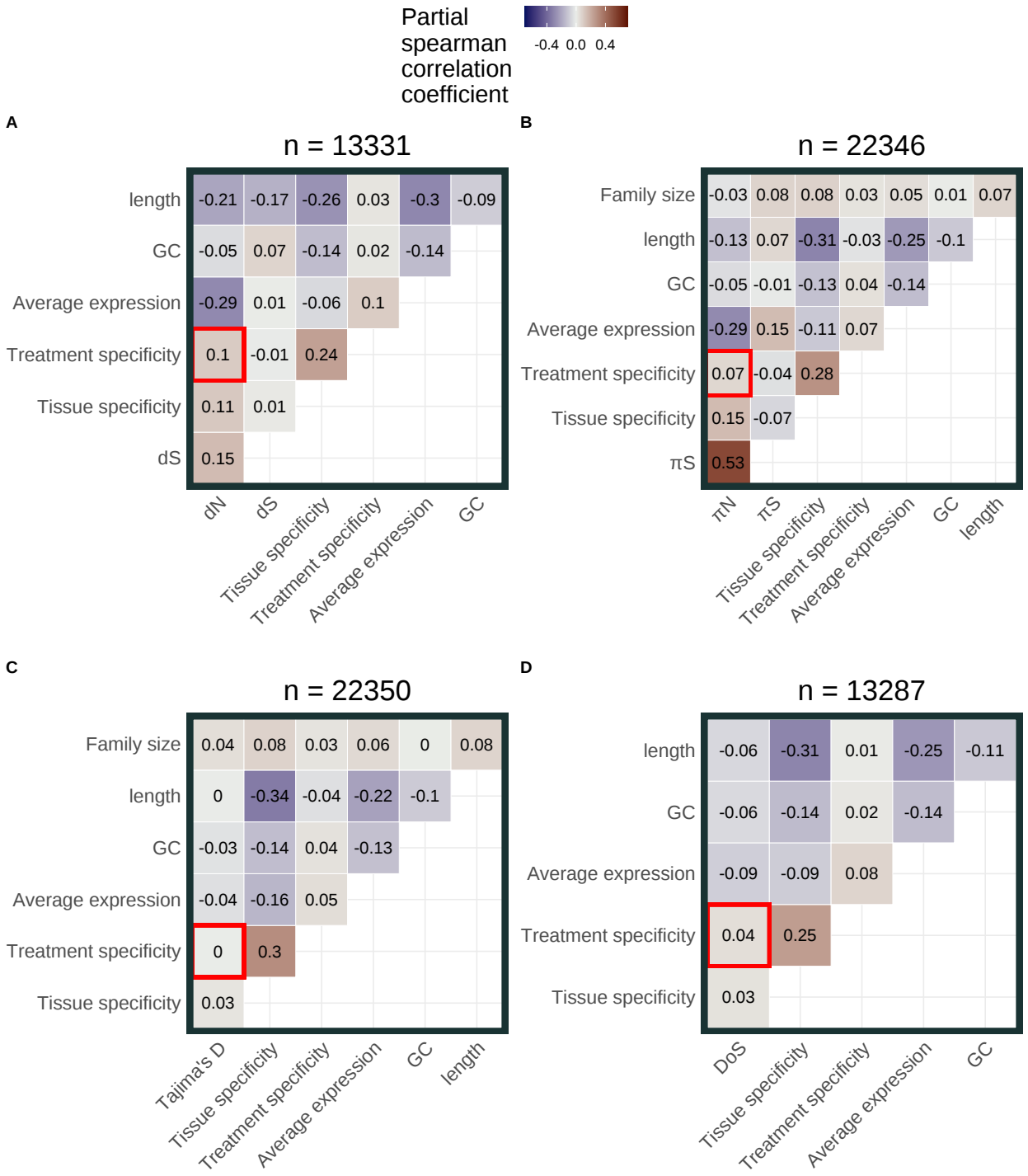

Figure A33: **Partial correlation analysis of fruit or flower tissue on balanced data subset after SVA.** Partial correlations for **(A)** dN, **(B)** π_N, **(C)** Tajima's D, and **(D)** direction of selection (DoS) after applying SVA. Average expression and tissue specificity are calculated using only fruit and flower tissue data. The number of genes included in each partial correlation analysis (n) is listed at the top of each heatmap.

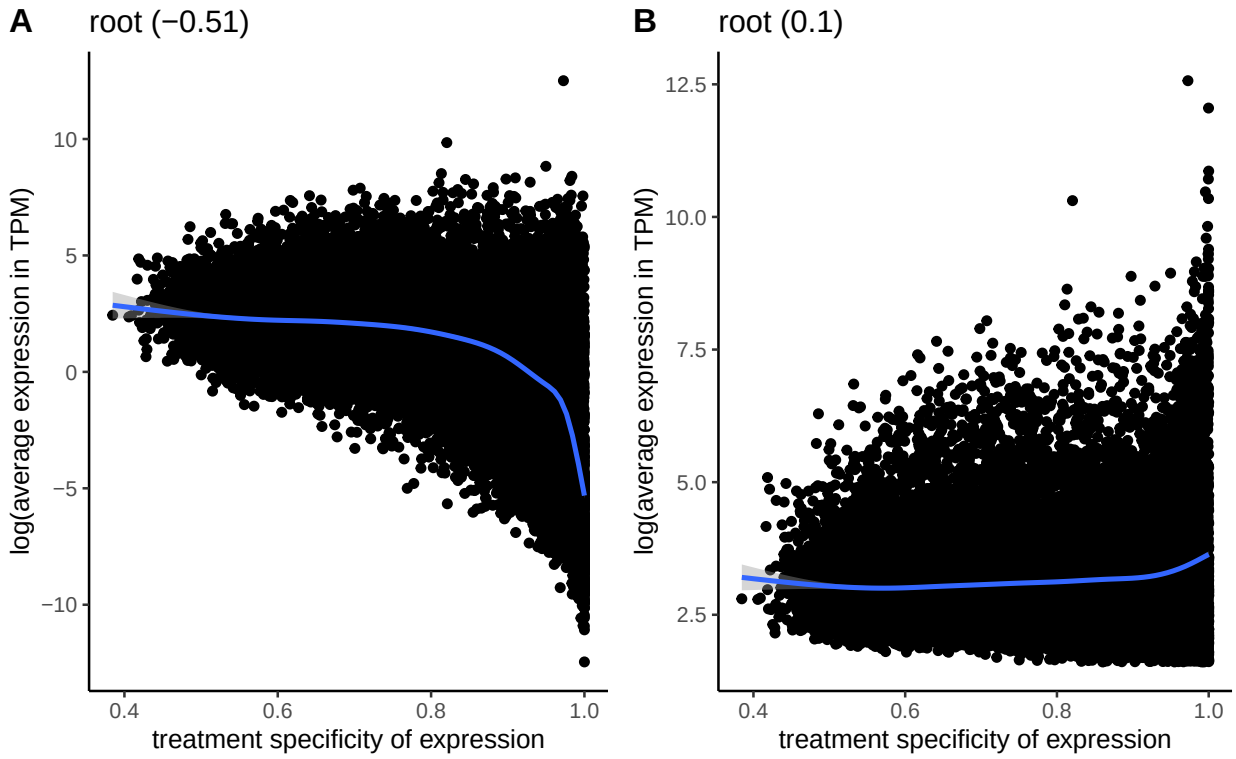

Figure A34: **Treatment specificity vs average expression correlation in root tissue.** Correlation between average expression in transcripts per million (TPM) and treatment specificity of genes when low expression values (< 5 TPM) are included (**A**) vs excluded (**B**). Expression level and treatment specificity were calculated using only data from root tissue samples. Line is a smoothing line with 95 % confidence intervals and values in parentheses give spearman correlation.

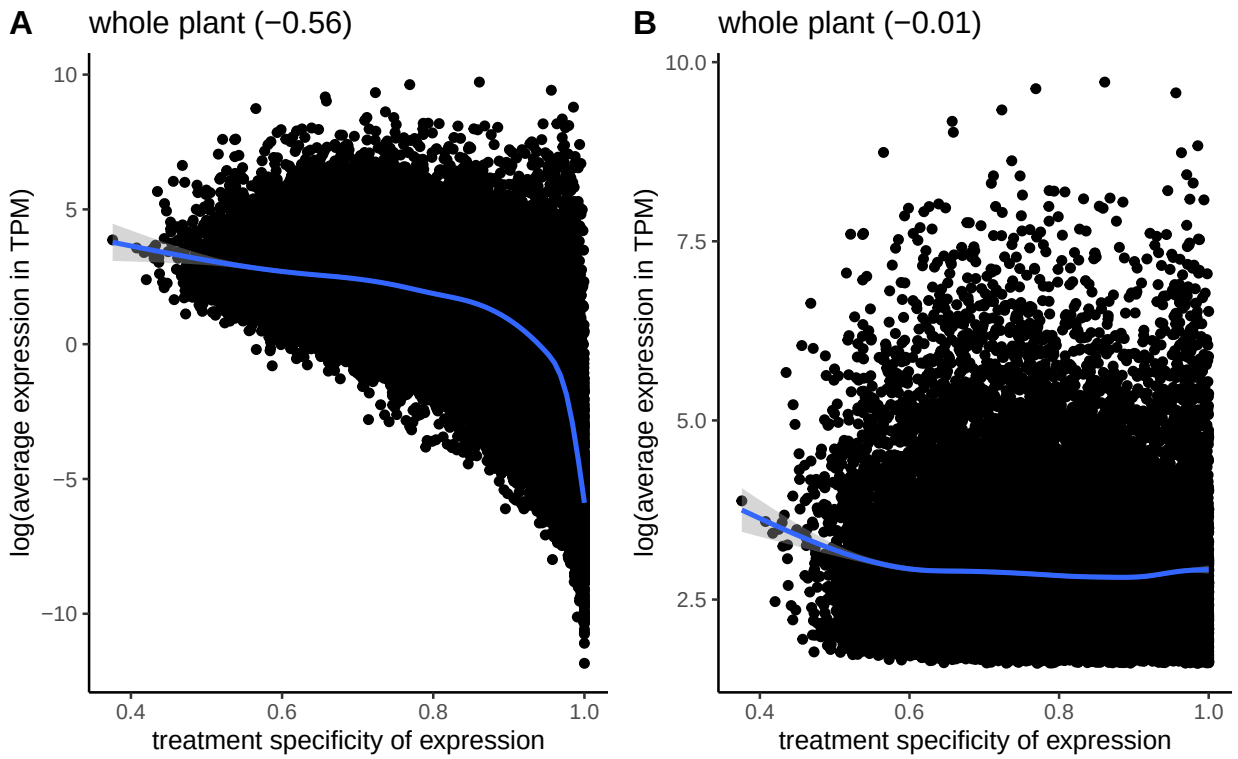

Figure A35: **Treatment specificity vs average expression correlation in whole plant tissue.** Correlation between average expression in transcripts per million (TPM) and treatment specificity of genes when low expression values (< 5 TPM) are included (**A**) vs excluded (**B**). Expression level and treatment specificity were calculated using only data from whole plant tissue samples. Line is a smoothing line with 95 % confidence intervals and values in parentheses give spearman correlation.

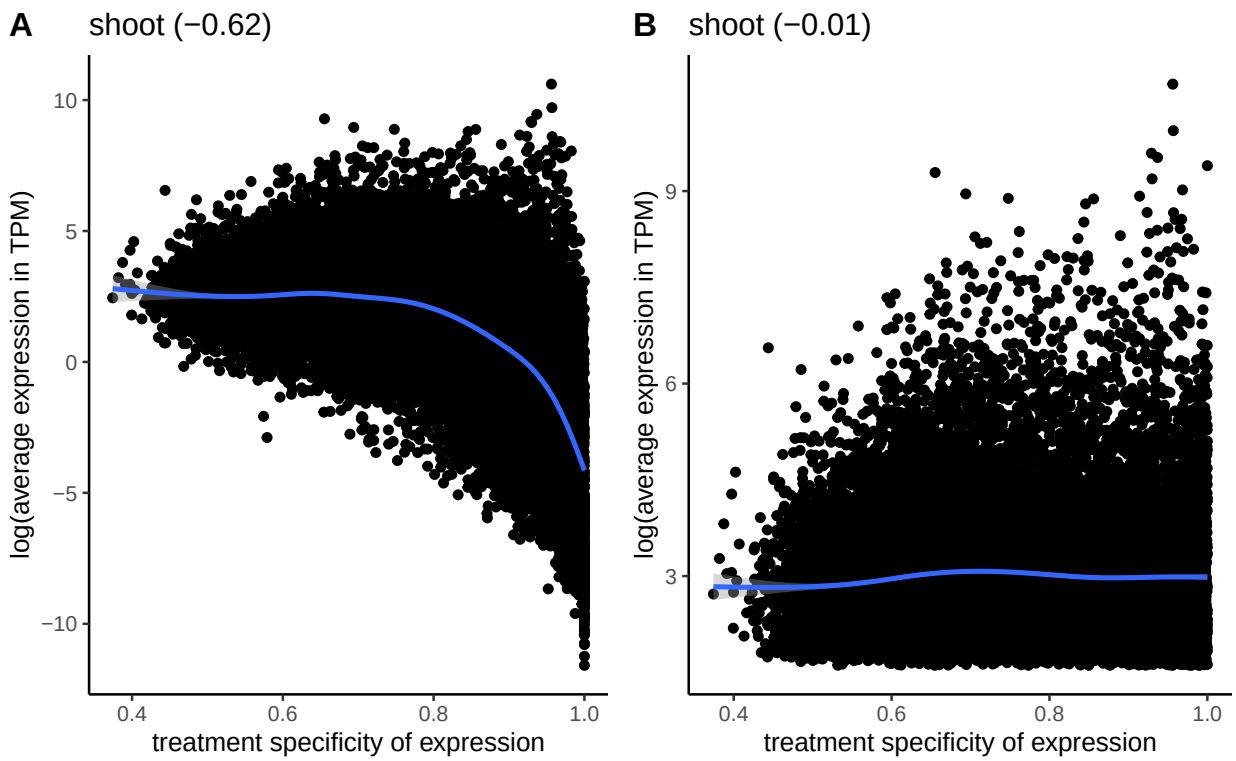

Figure A36: **Treatment specificity vs average expression correlation in shoot tissue.** Correlation between average expression in transcripts per million (TPM) and treatment specificity of genes when low expression values (< 5 TPM) are included (**A**) vs excluded (**B**). Expression level and treatment specificity were calculated using only data from shoot tissue samples. Line is a smoothing line with 95 % confidence intervals and values in parentheses give spearman correlation.

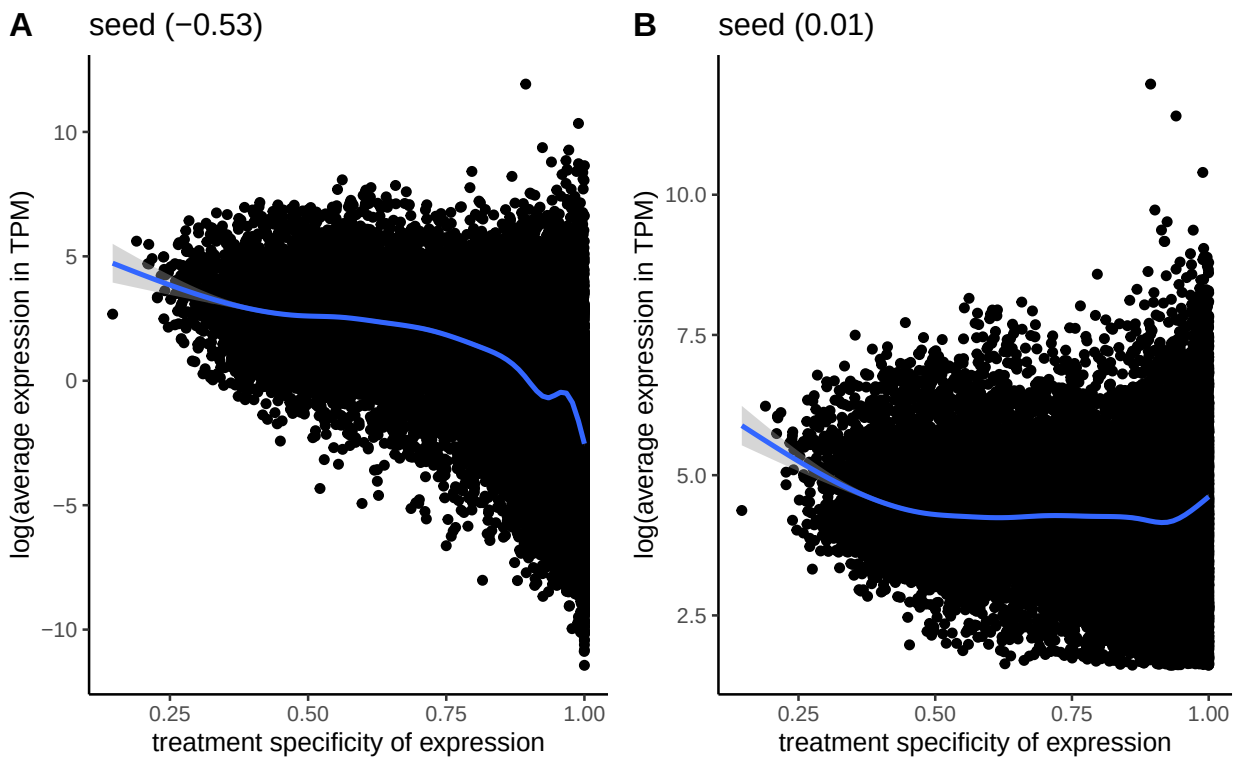

Figure A37: **Treatment specificity vs average expression correlation in seed tissue.** Correlation between average expression in transcripts per million (TPM) and treatment specificity of genes when low expression values (< 5 TPM) are included (**A**) vs excluded (**B**). Expression level and treatment specificity were calculated using only data from seed tissue samples. Line is a smoothing line with 95 % confidence intervals and values in parentheses give spearman correlation.

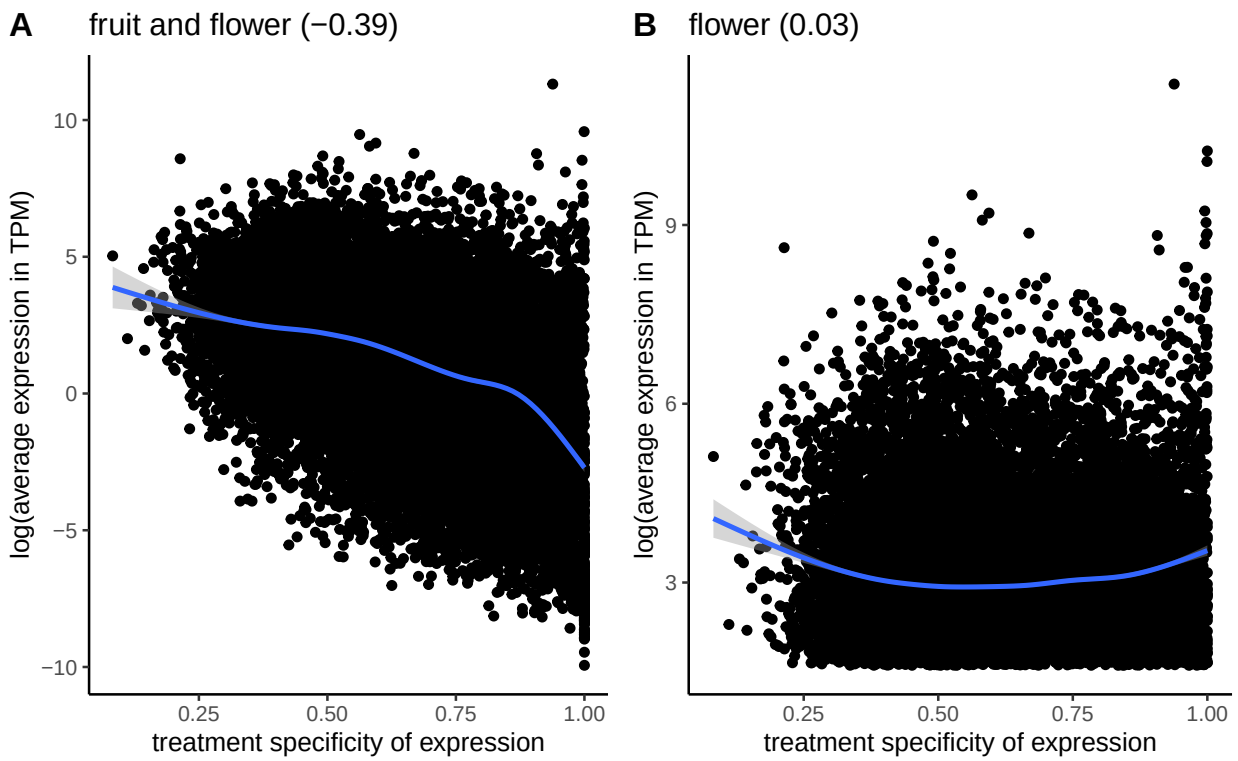

Figure A38: **Treatment specificity vs average expression correlation in fruit or flower tissue.** Correlation between average expression in transcripts per million (TPM) and treatment specificity of genes when low expression values (< 5 TPM) are included (**A**) vs excluded (**B**). Expression level and treatment specificity were calculated using only data from fruit and flower tissue samples. Line is a smoothing line with 95 % confidence intervals and values in parentheses give spearman correlation.

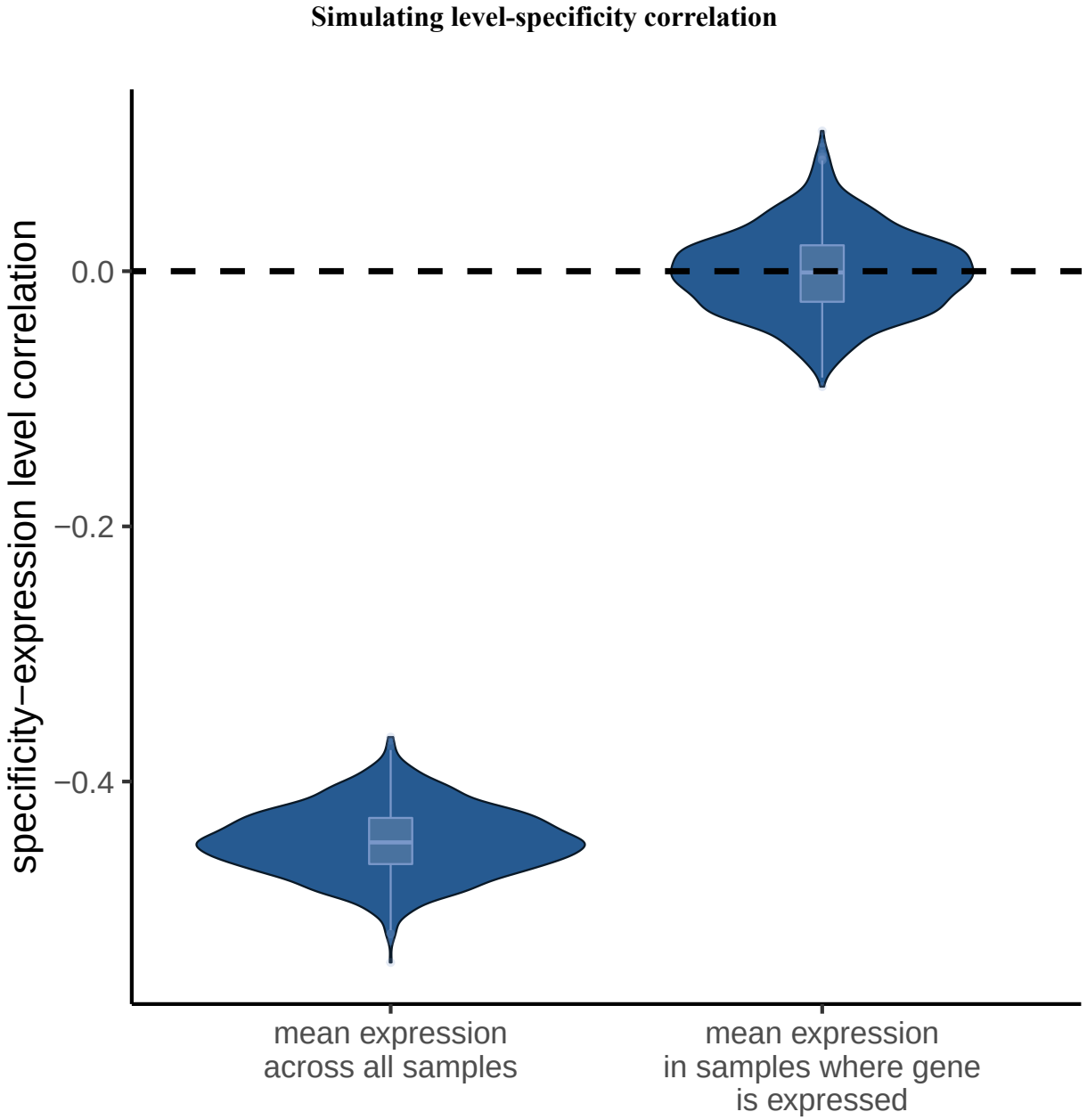

Figure A39: **Violin plot of correlations between expression specificity and expression level across 1000 simulated expression matrices.** In the two different simulations shown here, expression level was quantified either as average expression across all samples or average expression across samples with expression values > 0. Each expression matrix was simulated by sampling from a zero-inflated negative binomial distribution.

APPENDIX B: SUPPLEMENTAL METHODS FOR CHAPTER 2

The records that we used to estimate range size from GBIF occurrence data could not have any issue codes except for the following:

- “AMBIGUOUS_COLLECTION”
- “AMBIGUOUS_INSTITUTION”
- “COLLECTION_MATCH_FUZZY”
- “COLLECTION_MATCH_NONE”
- “CONTINENT_DERIVED_FROM_COORDINATES”
- “COORDINATE_ROUNDED”
- “COUNTRY_DERIVED_FROM_COORDINATES”
- “COUNTRY_MISMATCH”
- “DEPTH_MIN_MAX_SWAPPED”
- “DEPTH_NON_NUMERIC”
- “DEPTH_NOT_METRIC”
- “DEPTH_UNLIKELY”
- “DIFFERENT_OWNER_INSTITUTION”
- “ELEVATION_MIN_MAX_SWAPPED”
- “ELEVATION_NON_NUMERIC”
- “ELEVATION_NOT_METRIC”
- “ELEVATION_UNLIKELY”
- “GEODETIC_DATUM_ASSUMED_WGS84”
- “INSTITUTION_COLLECTION_MISMATCH”
- “INSTITUTION_MATCH_FUZZY”
- “INSTITUTION_MATCH_NONE”
- “OCCURRENCE_STATUS_INFERRED_FROM_BASIS_OF_RECORD”
- “OCCURRENCE_STATUS_INFERRED_FROM_INDIVIDUAL_COUNT”

APPENDIX C: SUPPLEMENTAL FIGURES FOR CHAPTER 2

Exploring relationships in data before outlier removal

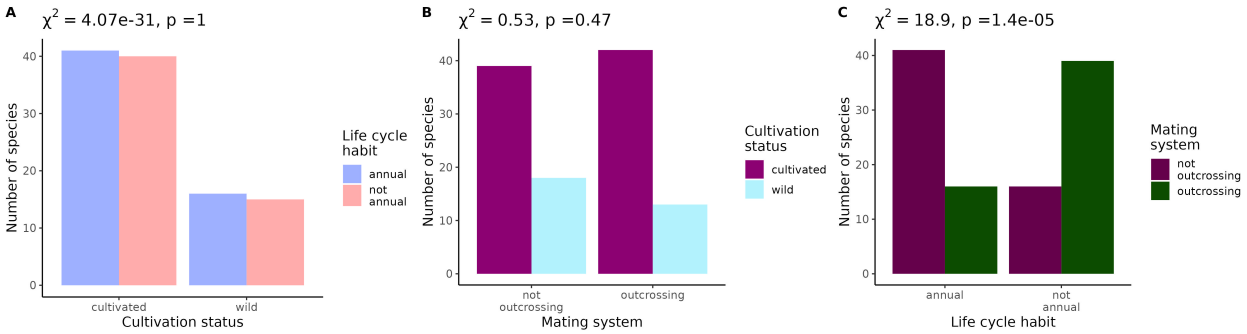

Figure C1: **Tests of independence between life-history traits included in this study.** Values across the top of each plot give the results of a χ^2 test of independence between each pair of life-history traits.

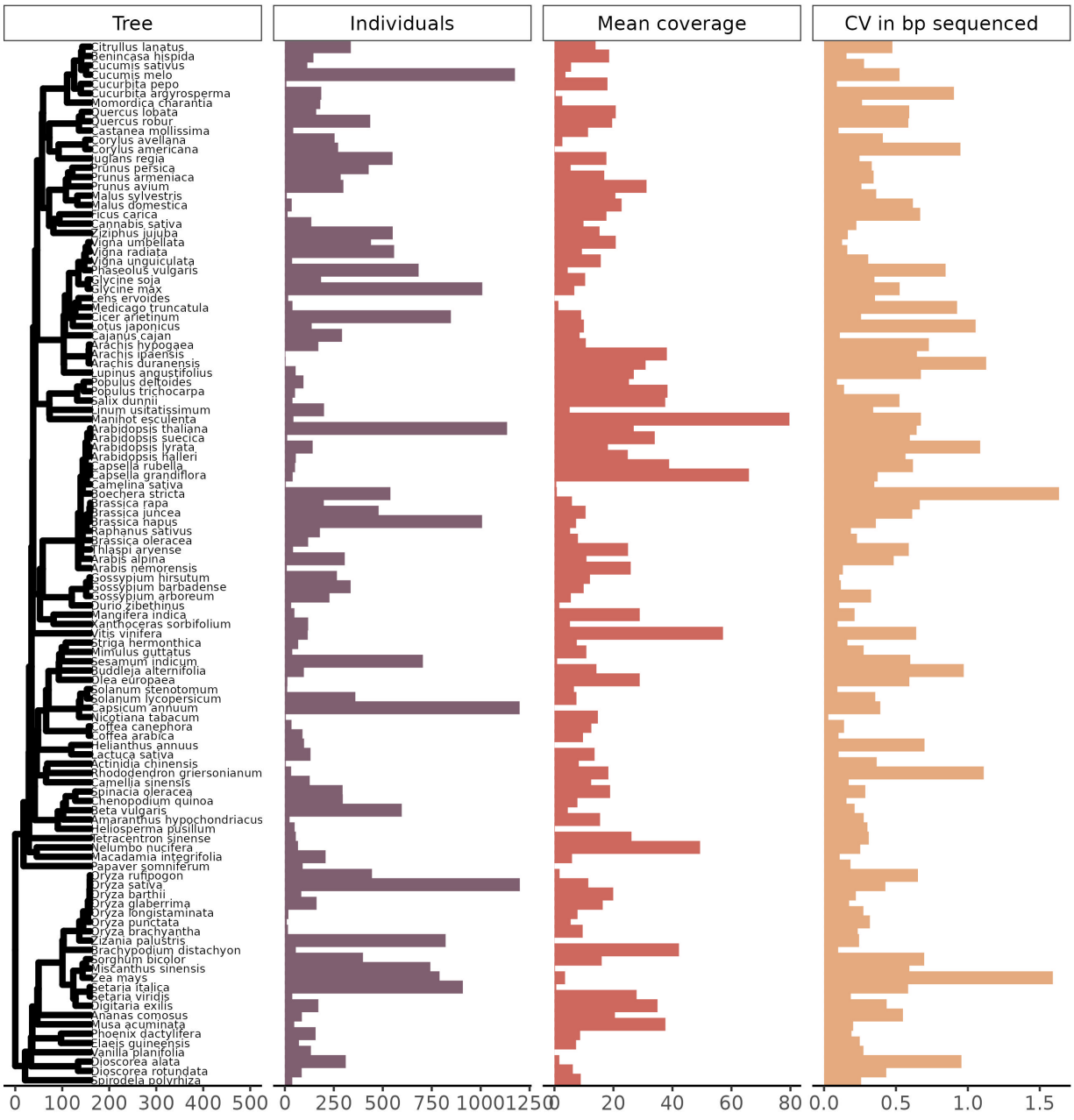

Figure C2: **Phylogenetic tree of species included in this study plotted alongside technical sequencing variables.** The phylogenetic tree scale is in millions of years. We limited the number of individuals sequenced per species to no more than 1200. CV stands for coefficient of variation and bp for base pairs.

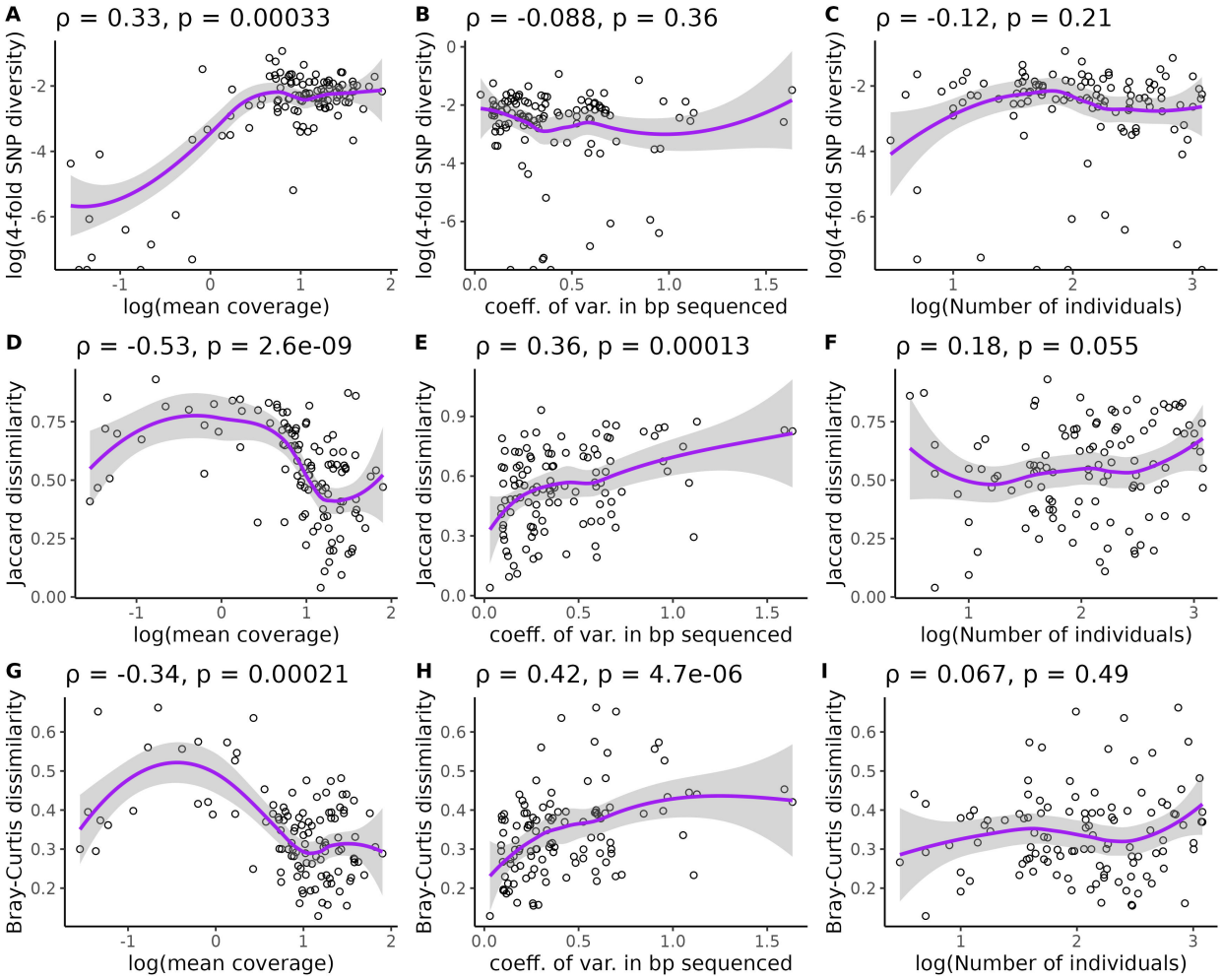

Figure C3: **Correlations between technical sequencing variables and diversity.** Each point is a species. Values across the top of each plot give the Spearman's correlation coefficient and p-value (testing whether the correlation differs from zero) for each pairwise relationship. Each pairwise plot includes one of three measures of diversity (Nucleotide diversity: A-C, Jaccard Dissimilarity: D-F, Bray-Curtis dissimilarity: G-I) and one of three technical sequencing variables (mean coverage: A, D, G; Coefficient of variation in bp sequenced: B, E, H; Number of individuals sequenced: C, F, I). Purple line is a loess smoothing line with 95% confidence intervals shaded in gray. All logarithms are base 10. Three species with nucleotide diversity values of 0 are omitted from plots involving nucleotide diversity.

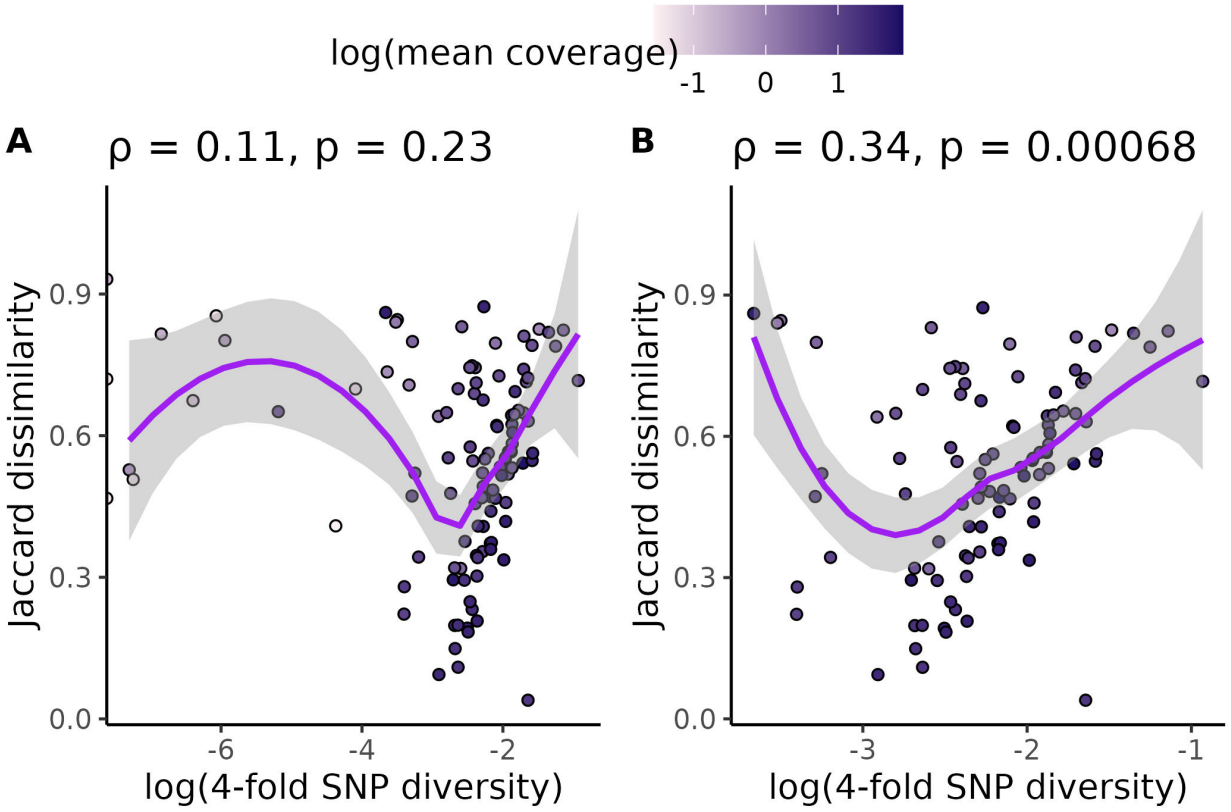

Figure C4: **Effect of coverage on relationship between π and Jaccard dissimilarity.** (A) shows the relationship between k-mer diversity (Jaccard dissimilarity) and nucleotide diversity without omitting species with $\leq$ 0.5x coverage or $\leq$ 1000 SNP calls. (B) shows the same relationship, except these species with low coverage or low numbers of SNP calls are omitted. Each data point is a species. All species' points are colored by the log of average genome-wide coverage per individual (base 10) for that species. Purple lines are loess smoothing curves with 95% confidence intervals shaded in gray. Values across the top of each plot are Spearman correlation coefficients (ρ) and p-values that test whether each correlation coefficient differs from zero. Three species with nucleotide diversity values of 0 are omitted from these plots.

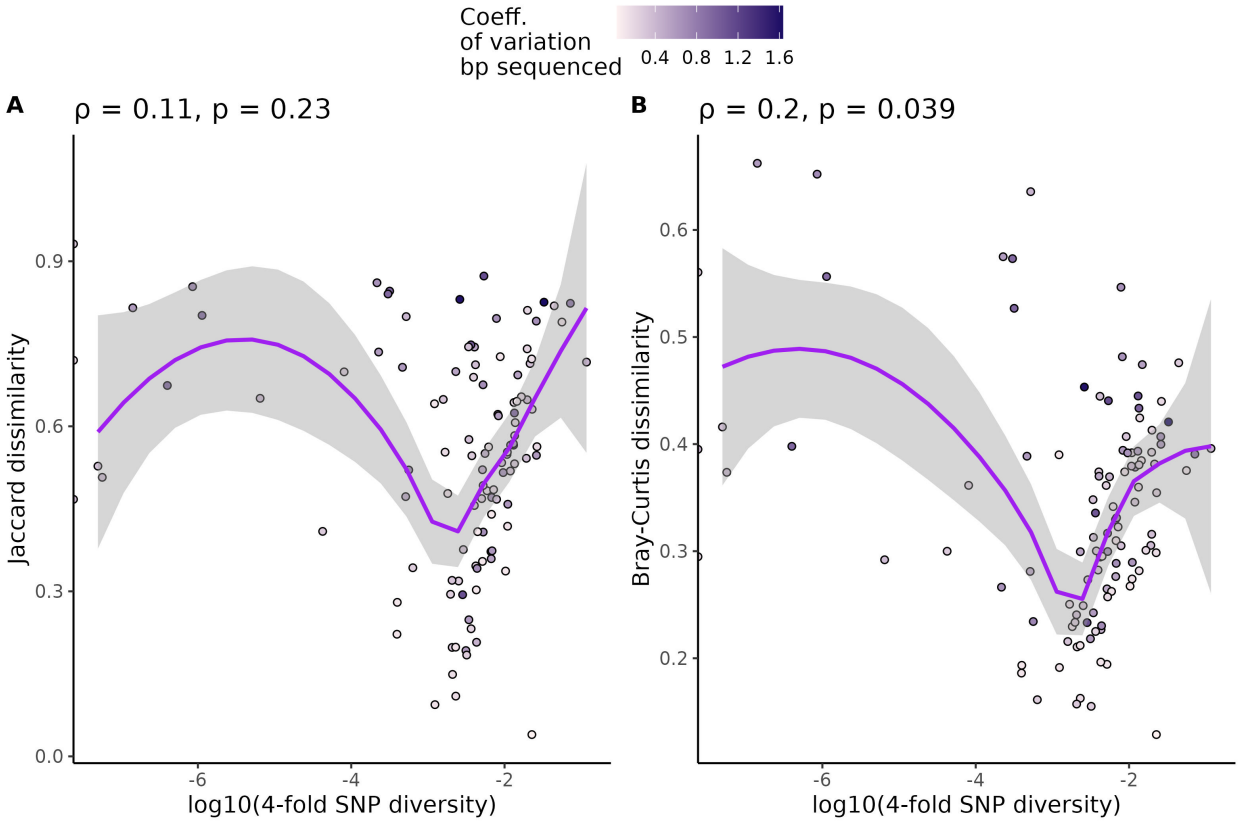

Figure C5: **Relationship between k-mer diversity, nucleotide diversity and variation in sequencing coverage.** Each point is a species. Values across the top of each plot give the Spearman's correlation coefficient and p-value (testing whether the correlation differs from zero) for each pairwise relationship. Purple line is a loess smoothing line with 95% confidence intervals shaded in gray. Points are colored by the coefficient of variation (mean/standard deviation) in bp sequenced for each species. Three species with nucleotide diversity values of 0 are omitted from the plots.

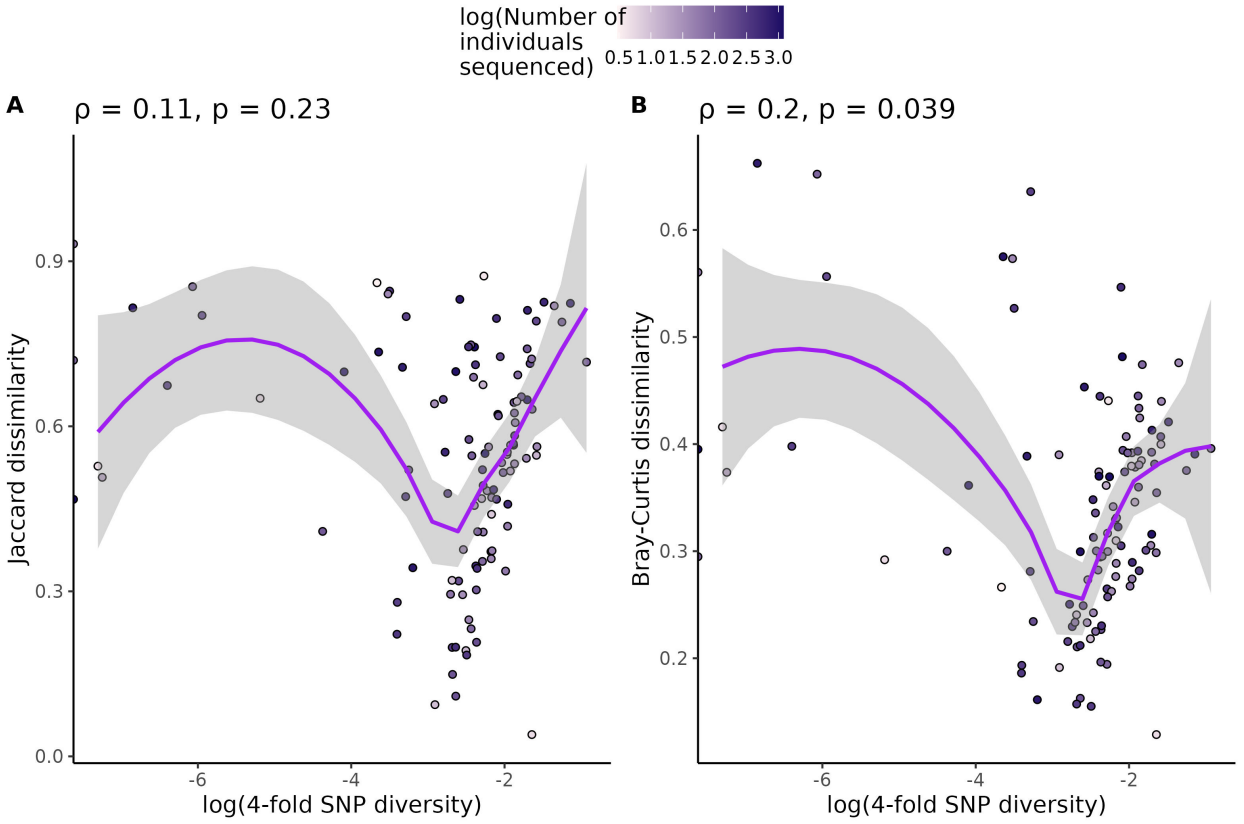

Figure C6: **Relationship between k-mer diversity, nucleotide diversity, and number of individuals sequenced.** Each point is a species. Values across the top of each plot give the Spearman's correlation coefficient and p-value (testing whether correlation differs from zero) for each pairwise relationship. Purple line is a loess smoothing line with 95% confidence intervals shaded in gray. Points are colored by the logarithm of the number of individuals sequenced within each species (base 10). Three species with nucleotide diversity values of 0 are omitted from the plots.

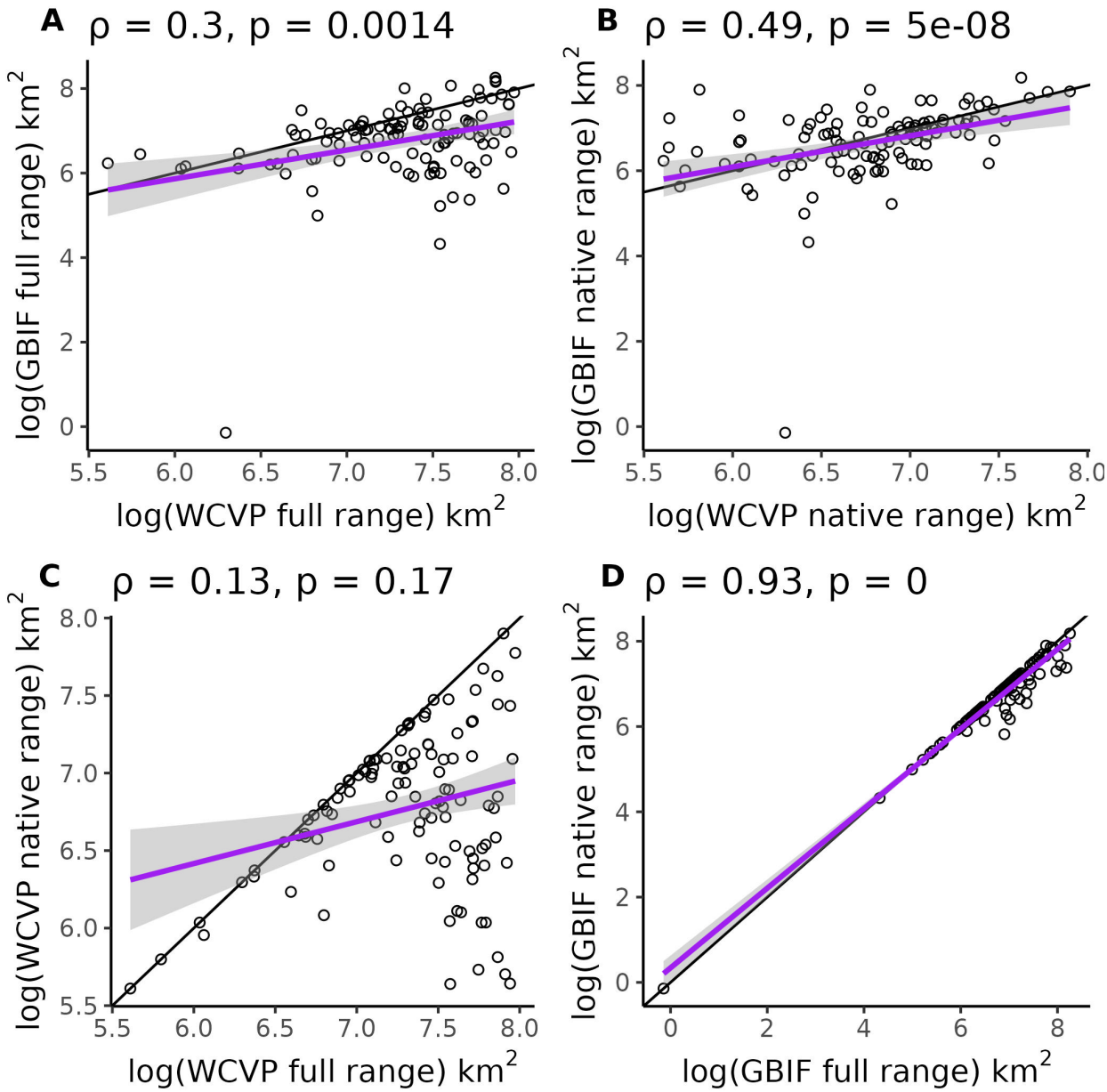

Figure C7: **Independent methods of range size estimation correlate with each other.** (A) Compares the size of total size of native and invaded ranges from GBIF occurrence data and WCVP range maps. (B) Compares native range estimates only (invaded ranges excluded) from GBIF occurrence data and WCVP range maps. (C) and (D) compare the full and native range estimates from WCVP range maps and GBIF occurrence data, respectively.Every data point is a species. The solid black lines are 1:1 reference lines where the different measures of range size are equal. The purple lines are linear regression lines with 95 % confidence intervals in grey shading. Values across the top of each plot give the Spearman's correlation coefficient (ρ) and p-value testing whether correlation differs from zero. One p-value is reported as zero because it was $< 2.2 \times 10^{-16}$, which is the limit of precision for doubles in R. All logarithms are base 10.

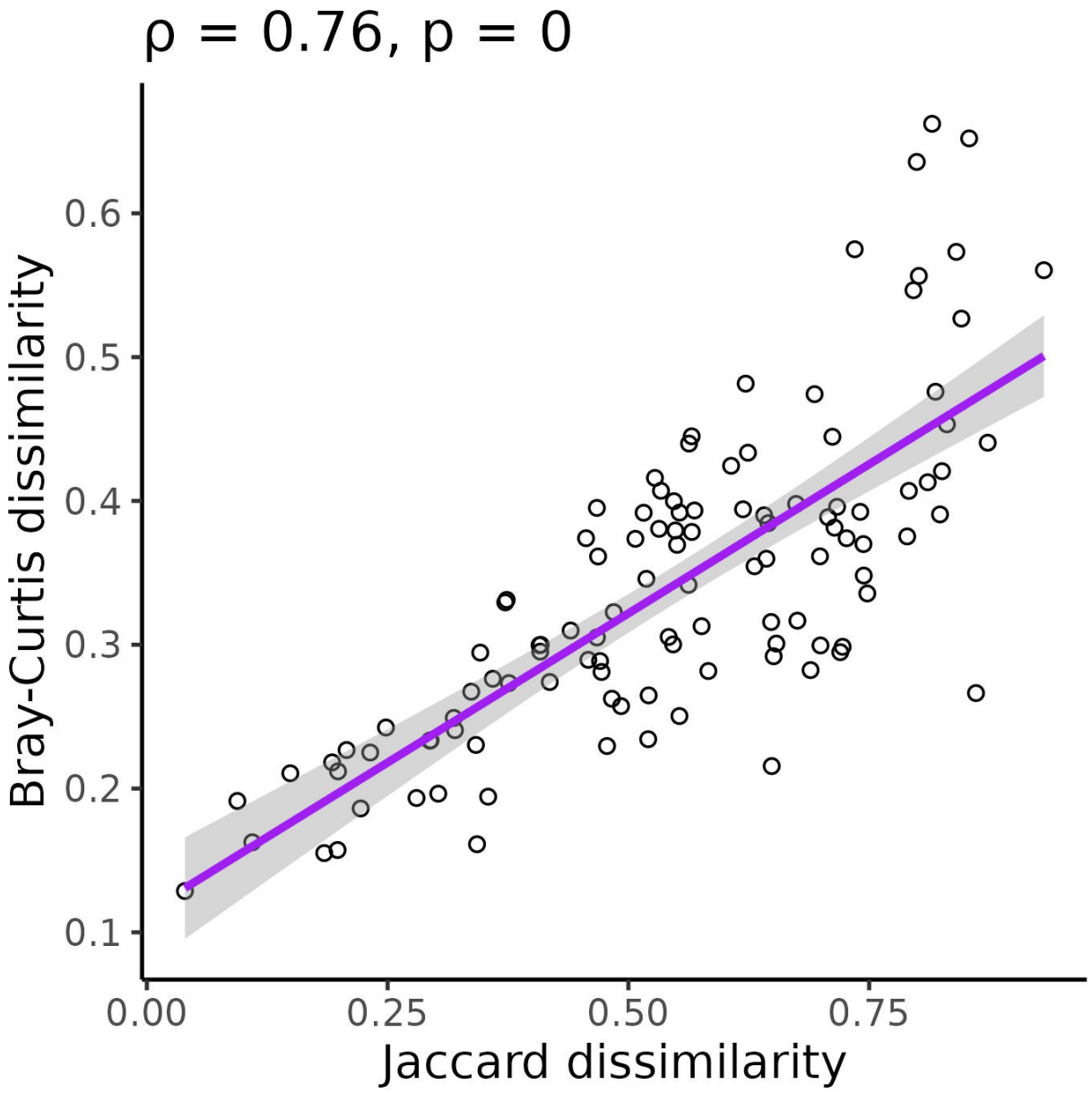

Figure C8: **Relationship between Bray-Curtis and Jaccard dissimilarity.** Each point is a species. Values across the top of each plot give the Spearman's correlation coefficient and p-value (testing whether correlation differs from zero). The p-value is reported as zero because it was $< 2.2 \times 10^{-16}$, which is the limit of precision for doubles in R. Line is a linear regression with 95% confidence intervals shaded in gray.

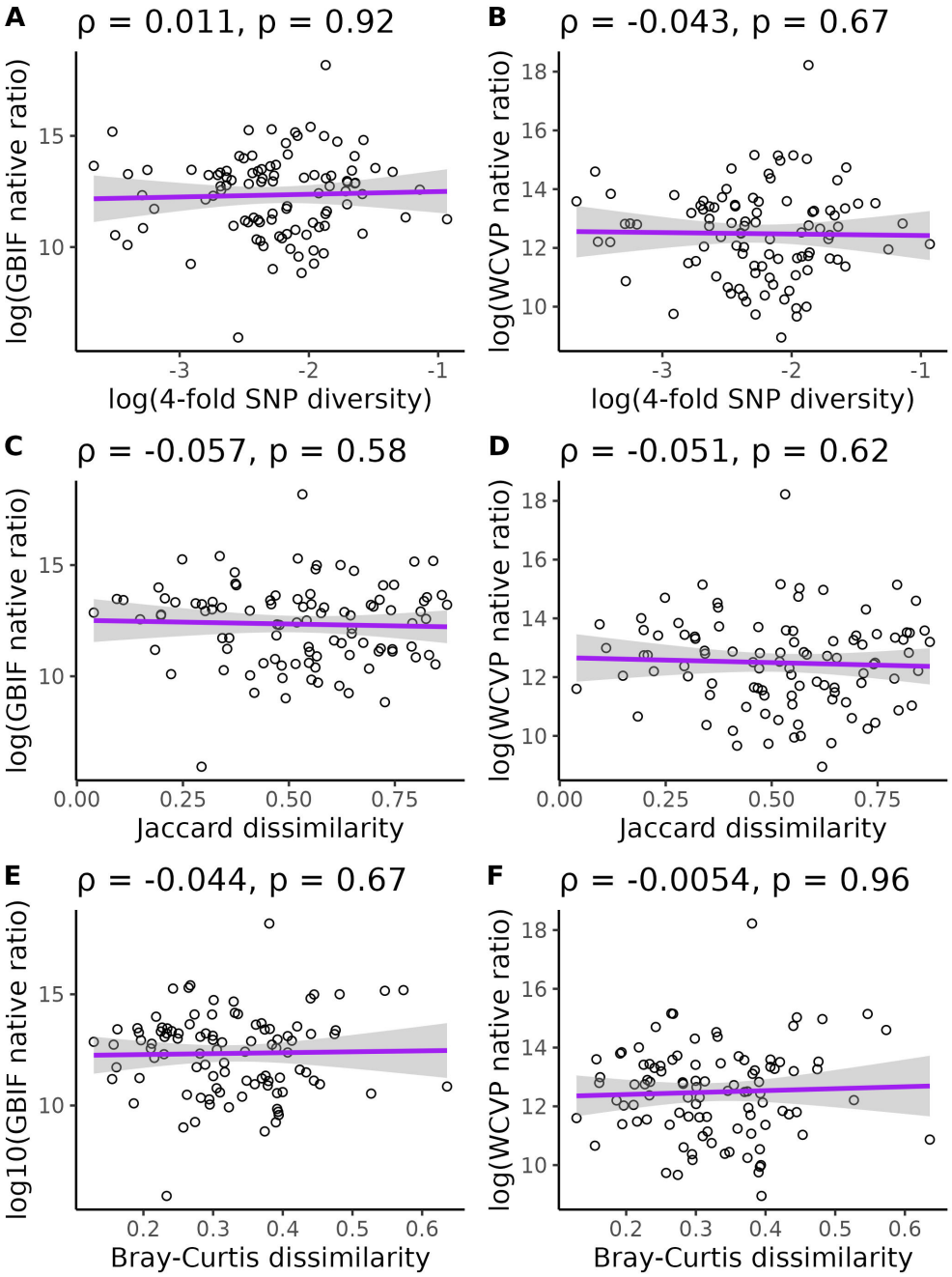

Figure C9: **Relationship between the range size-squared height ratio and diversity without correcting for evolutionary history, genome size, life cycle habit, mating system, or cultivation status.** Each point is a species. Values across the top of each plot give the Spearman's correlation coefficient and p-value (testing whether correlation differs from zero) for each pairwise relationship. Each pairwise relationship involves the native range size-squared height ratio, where native range size was estimated from GBIF occurrences (A, C, E), or WCVP range maps (B, D, F), and one of three diversity measures (Nucleotide diversity: A-B, Jaccard Dissimiarity: C-D, Bray-Curtis dissimilarity: E-F). Purple lines are loess smoothing lines with 95% confidence intervals shaded in gray. Three species with nucleotide diversity values of 0 are omitted from plots involving nucleotide diversity. All logarithms are base 10.

Population size proxy vs diversity relationships after controlling for phylogeny and life-history variables, but not controlling for genome size

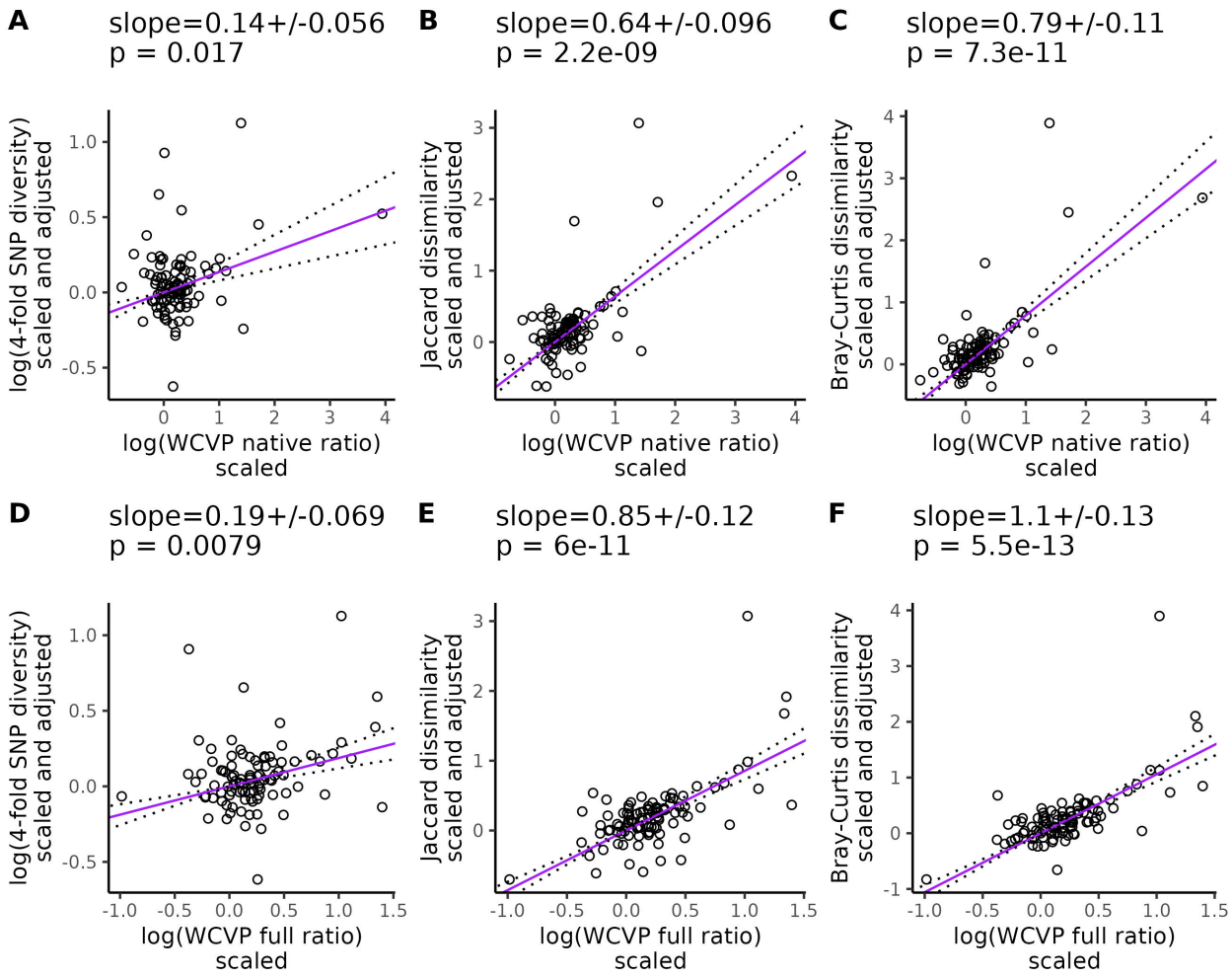

Figure C10: **Partial phylogenetic regression between WCVP population size proxy and diversity.** Each point is a species and only species with $>$ 0.5x mean coverage and $>$ 1000 variant sites were included in the regression. The regressions are organized according to whether invaded ranges were excluded (A-C) or included (D-F) in the range size-squared height ratio. Range size was estimated from WCVP range maps. Lines give the relationship between the pairs of plotted variables after controlling for mating system, life cycle habit, cultivation status, and evolutionary history. Before fitting the line, each response variable was scaled to a standard normal distribution (mean = 0, variance = 1), then multiplied by the inverse of the Cholesky decomposition of the phylogenetic variance-covariance matrix to correct for phylogenetic relationships. The values at the top of each plot give the slope of the partial regression $\pm$ one standard error and p-values testing whether the slopes differ from zero. Dotted lines show the partial regression slope $\pm$ one standard error. All logarithms are base 10.

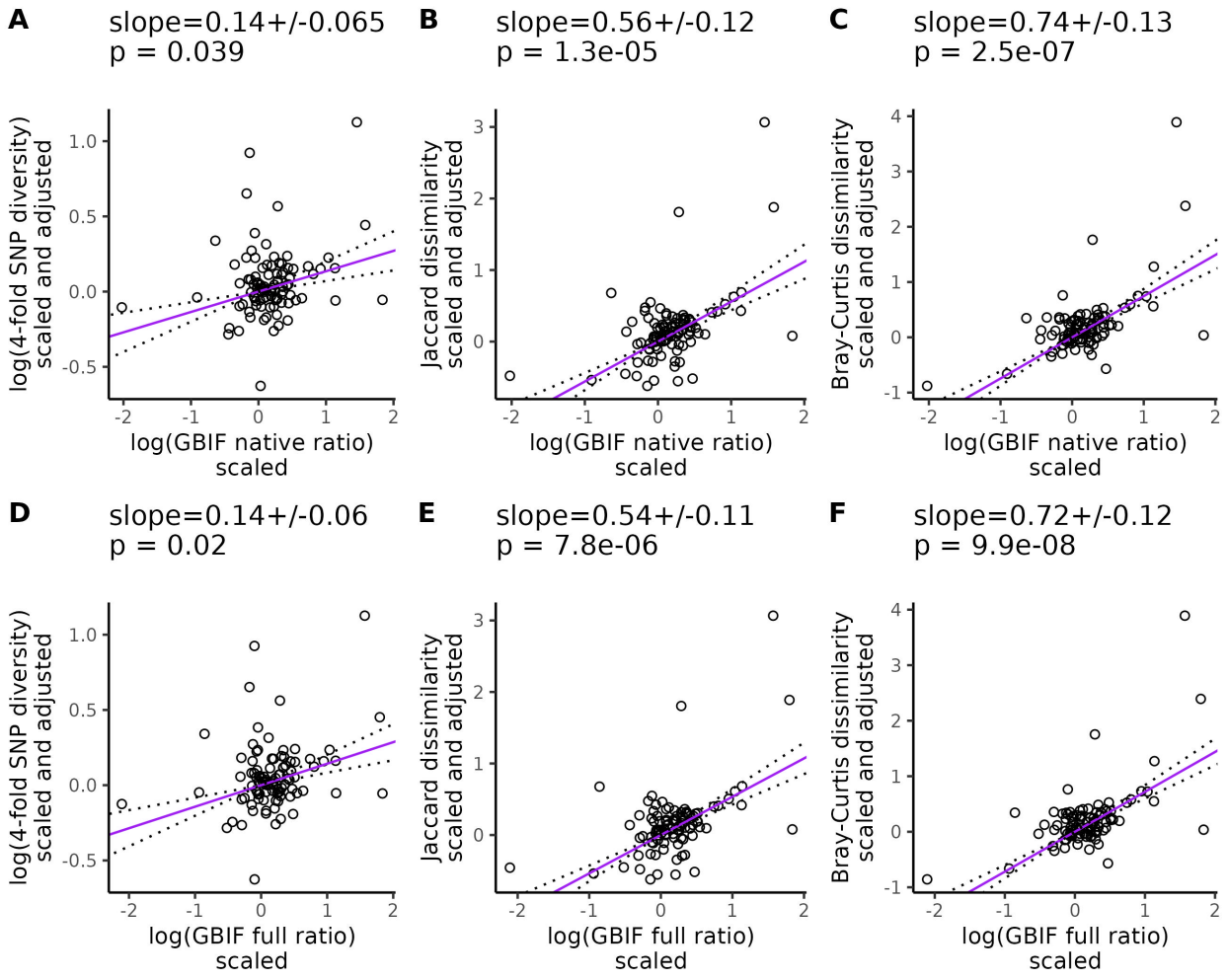

Figure C11: **Partial phylogenetic regression between GBIF population size proxy and diversity.** Each point is a species and only species with $>$ 0.5x mean coverage and $>$ 1000 variant sites were included in the regression. The regressions are organized according to whether invaded ranges were excluded (A-C) or included (D-F) in the range size-squared height ratio. Range size was estimated from GBIF occurrence data. Lines give the relationship between the pairs of plotted variables after controlling for mating system, life cycle habit, cultivation status, and evolutionary history. Before fitting the line, each response variable was scaled to a standard normal distribution (mean = 0, variance = 1), then multiplied by the inverse of the Cholesky decomposition of the phylogenetic variance-covariance matrix to correct for phylogenetic relationships. The values at the top of each plot give the slope of the partial regression $\pm$ one standard error and p-values testing whether the slopes differ from zero. Dotted lines show the partial regression slope $\pm$ one standard error. All logarithms are base 10.

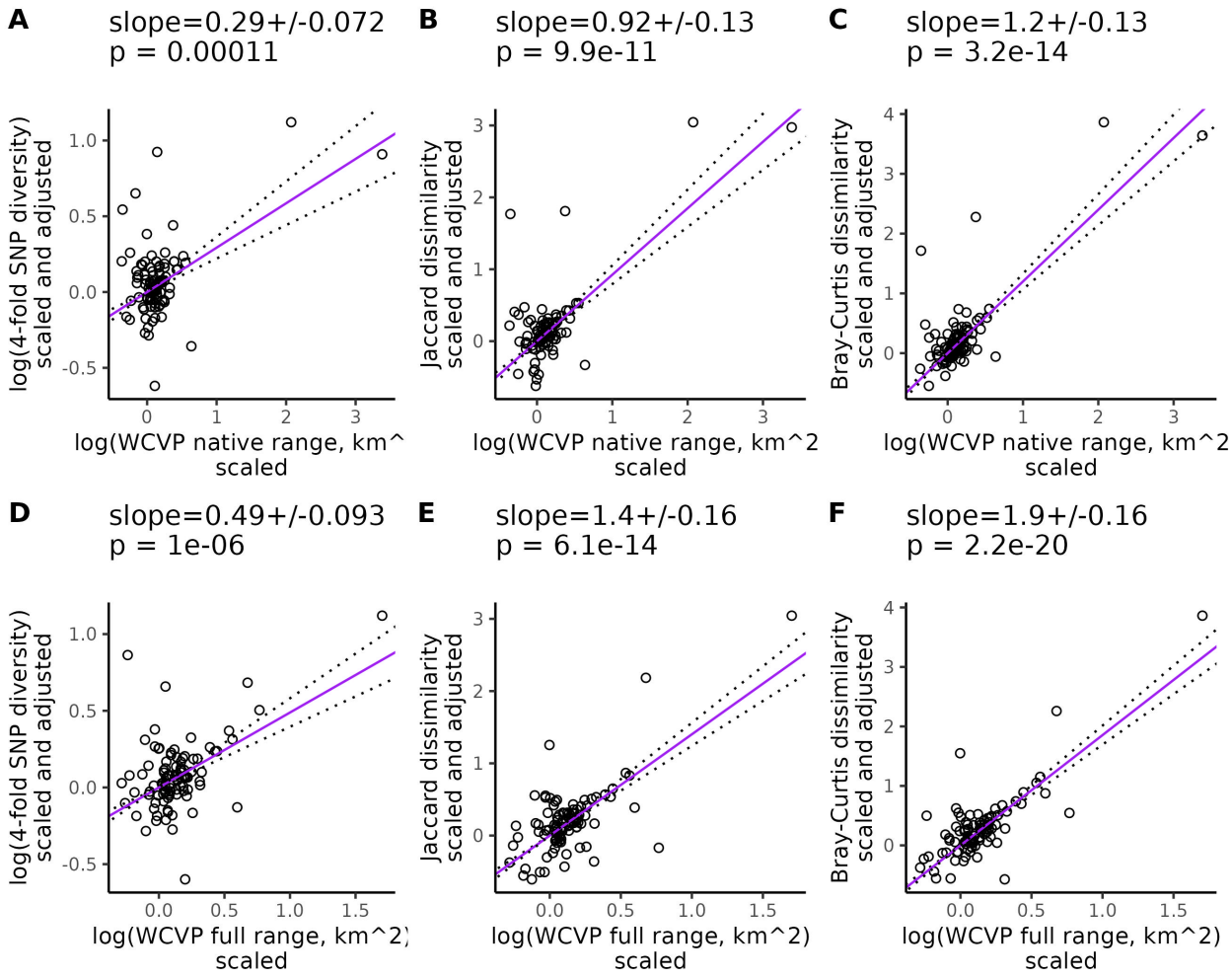

Figure C12: **Partial phylogenetic regression between WCVP range size and diversity.** Each point is a species and only species with $>$ 0.5x mean coverage and $>$ 1000 variant sites were included in the regression. The regressions are organized according to whether invaded ranges were excluded (A-C) or included (D-F) in the range size estimates. Range size was estimated from WCVP range maps. Lines give the relationship between the pairs of plotted variables after controlling for mating system, life cycle habit, cultivation status, and evolutionary history. Before fitting the line, each response variable was scaled to a standard normal distribution (mean = 0, variance = 1), then multiplied by the inverse of the Cholesky decomposition of the phylogenetic variance-covariance matrix to correct for phylogenetic relationships. The values at the top of each plot give the slope of the partial regression $\pm$ one standard error and p-values testing whether the slopes differ from zero. Dotted lines show the partial regression slope $\pm$ one standard error. All logarithms are base 10.

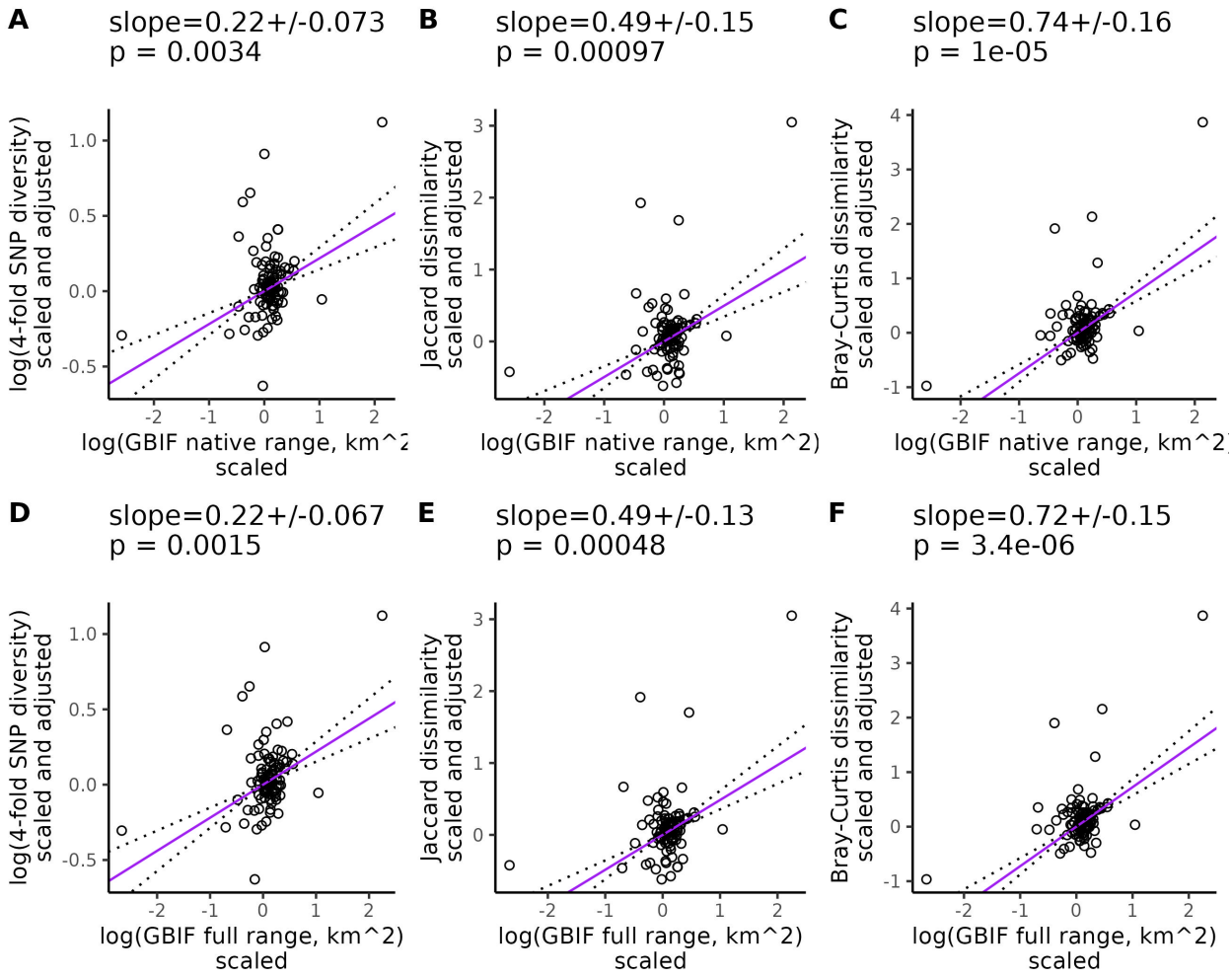

Figure C13: **Partial phylogenetic regression between GBIF range size and diversity.** Each point is a species and only species with $>$ 0.5x mean coverage and $>$ 1000 variant sites were included in the regression. The regressions are organized according to whether invaded ranges were excluded (A-C) or included (D-F) in the range size estimates. Range size was estimated from GBIF occurrence data. Lines give the relationship between the pairs of plotted variables after controlling for mating system, life cycle habit, cultivation status, and evolutionary history. Before fitting the line, each response variable was scaled to a standard normal distribution (mean = 0, variance = 1), then multiplied by the inverse of the Cholesky decomposition of the phylogenetic variance-covariance matrix to correct for phylogenetic relationships. The values at the top of each plot give the slope of the partial regression $\pm$ one standard error and p-values testing whether the slopes differ from zero. Dotted lines show the partial regression slope $\pm$ one standard error. All logarithms are base 10.

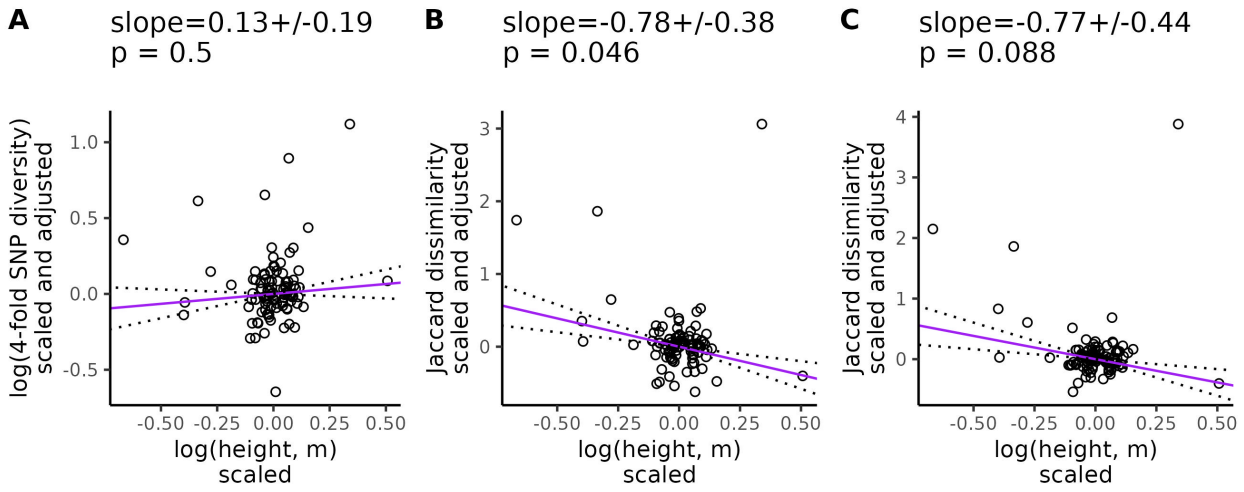

Figure C14: **Partial phylogenetic regression between height and diversity.** Each point is a species and only species with > 0.5x mean coverage and > 1000 variant sites were included in the regression. Lines give the relationship between the pairs of plotted variables after controlling for mating system, life cycle habit, cultivation status, and evolutionary history. The statistics across the top of each plot give the value of the slope of the lines ($\pm$ the standard error), the p-value testing whether the slope differs from zero. Before fitting the line, each response variable was scaled to a standard normal distribution (mean = 0, variance = 1), then multiplied by the inverse of the Cholesky decomposition of the phylogenetic variance-covariance matrix to correct for phylogenetic relationships. All logarithms are base 10.

Population size proxy vs diversity relationships after controlling for phylogeny, life-history variables, and genome size

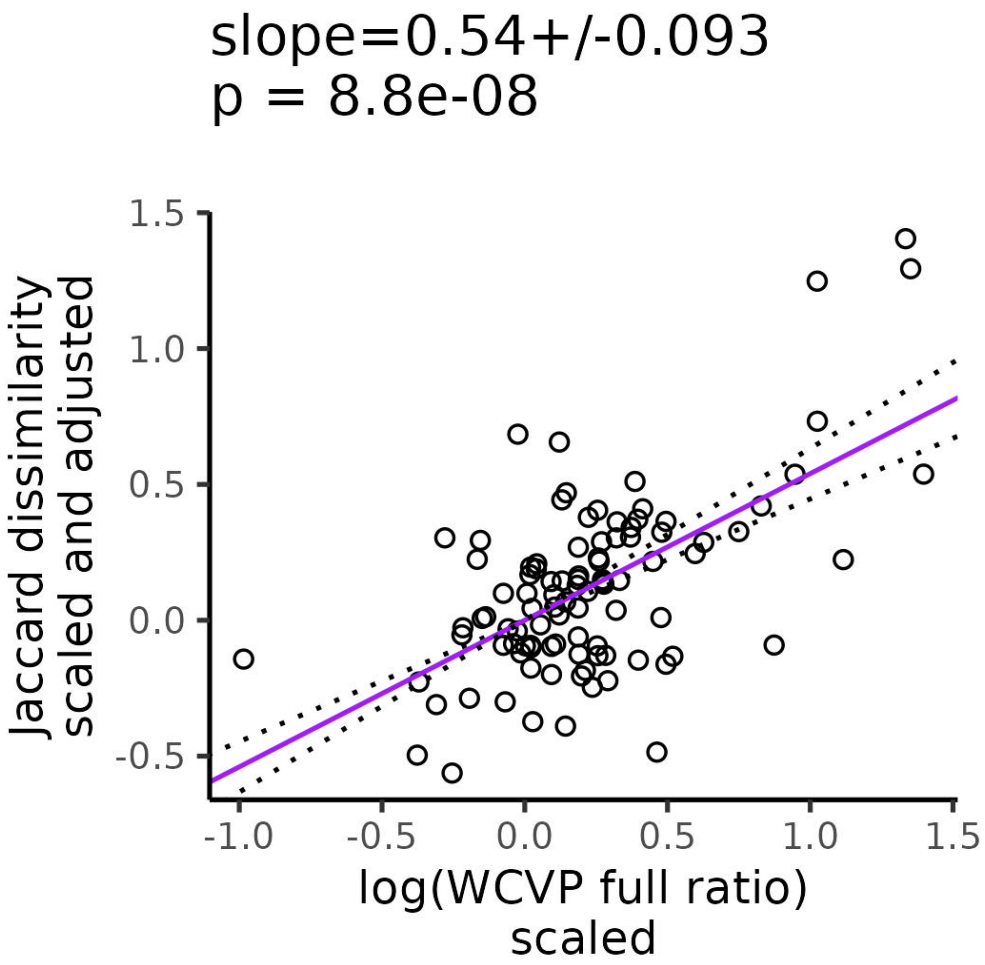

Figure C15: **Partial phylogenetic regression between WCVP full range size-squared height ratio and Jaccard dissimilarity, controlling for genome size.** Each point is a species and only species with $>$ 0.5x mean coverage and $>$ 1000 variant sites were included in the regression. WCVP full ratio gives the ratio of range size to squared plant height, where range size includes invaded ranges and is estimated from WCVP range maps. The partial regression controls for genome size, mating system, life cycle habit, cultivation status, and evolutionary history. Before fitting the line, each response variable was scaled to a standard normal distribution (mean = 0, variance = 1), then multiplied by the inverse of the Cholesky decomposition of the phylogenetic variance-covariance matrix to correct for phylogenetic relationships. The values at the top of the plot give the slope of the partial regression $\pm$ one standard error and p-values testing whether the slope differ from zero. Dotted lines show the partial regression slope $\pm$ one standard error. All logarithms are base 10.

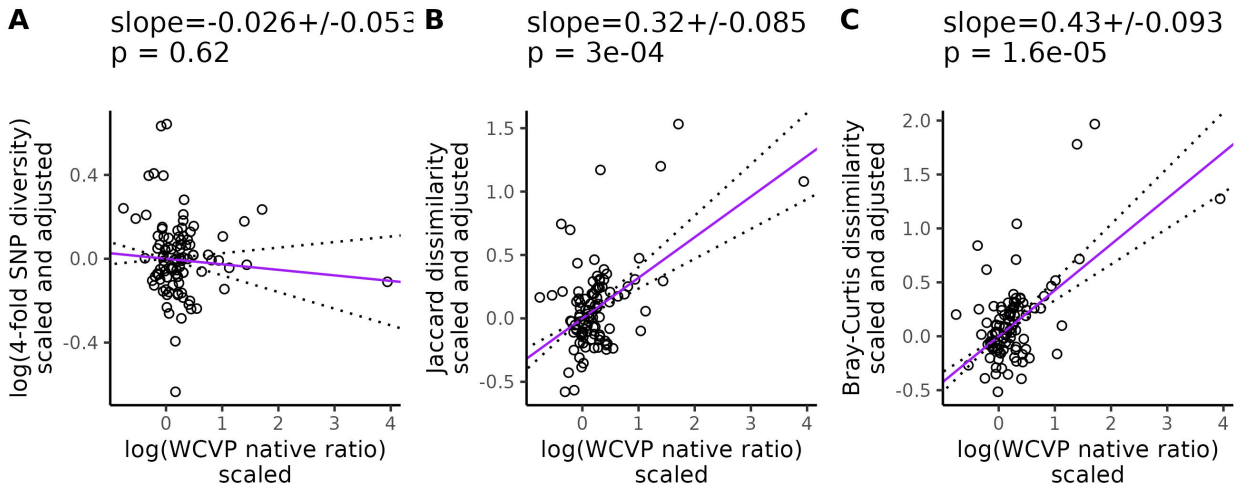

Figure C16: **Partial phylogenetic regression between WCVP native range size-squared height ratio and diversity, controlling for genome size.** Each point is a species and only species with $>$ 0.5x mean coverage and $>$ 1000 variant sites were included in the regression. WCVP native ratio gives the ratio of range size to squared plant height, where range size excludes invaded ranges and is estimated from WCVP range maps. The partial regression controls for genome size, mating system, life cycle habit, cultivation status, and evolutionary history. Before fitting the line, each response variable was scaled to a standard normal distribution (mean = 0, variance = 1), then multiplied by the inverse of the Cholesky decomposition of the phylogenetic variance-covariance matrix to correct for phylogenetic relationships. The values at the top of the plot give the slope of the partial regression $\pm$ one standard error and p-values testing whether the slope differ from zero. Dotted lines show the partial regression slope $\pm$ one standard error. All logarithms are base 10.

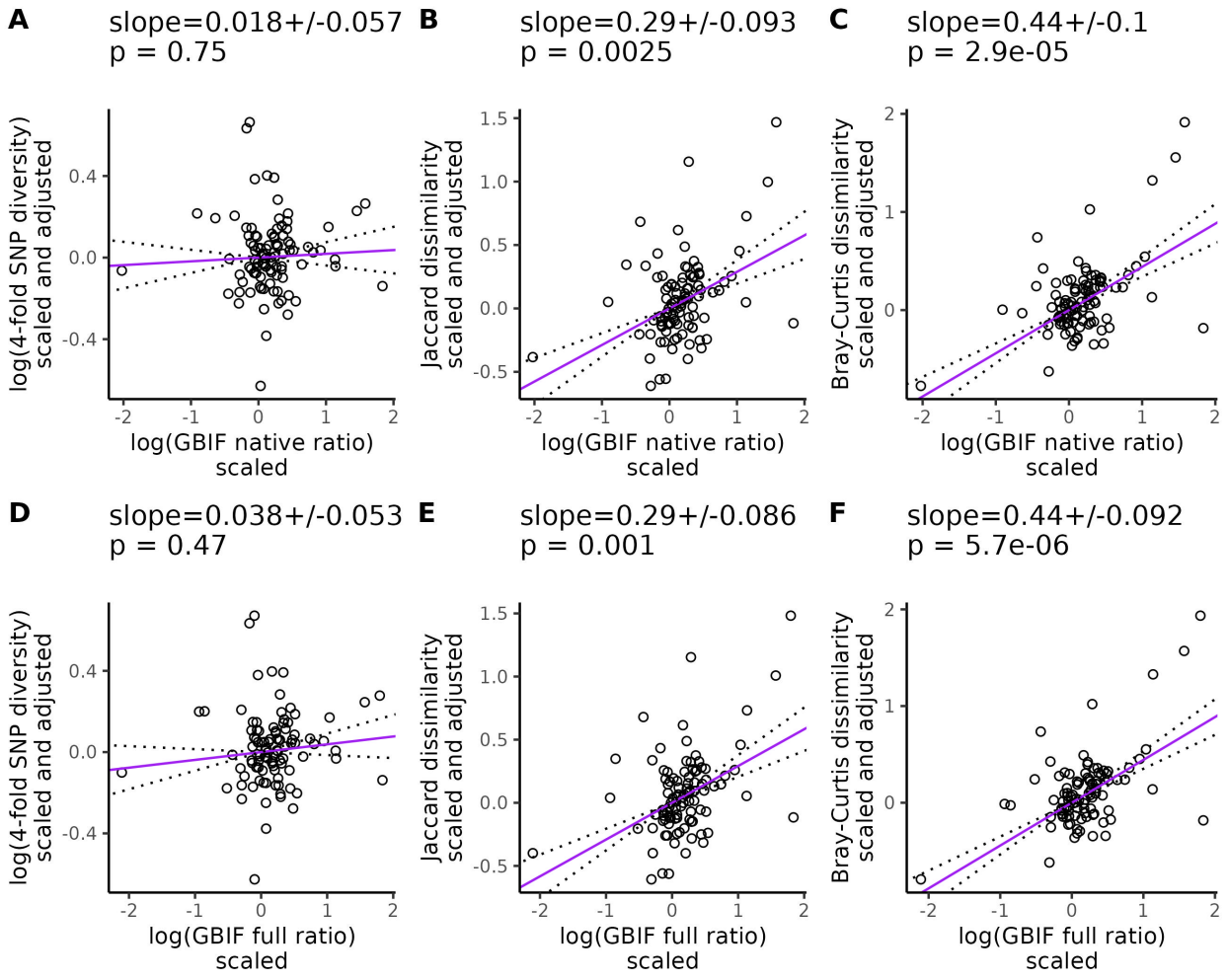

Figure C17: **Partial phylogenetic regression between GBIF range size-squared height ratio and diversity, controlling for genome size.** Each point is a species and only species with > 0.5x mean coverage and > 1000 variant sites were included in the regression. The regressions are organized according to whether invaded ranges were excluded (A-C) or included (D-F) in the range size-squared height ratio. Lines give the relationship between the pairs of plotted variables after controlling for mating system, life cycle habit, cultivation status, genome size, and evolutionary history. The statistics across the top of each plot give the value of the slope of the lines ($\pm$ the standard error) and the p-value testing whether the slope differs from zero. Before fitting the line, each response variable was scaled to a standard normal distribution (mean = 0, variance = 1), then multiplied by the inverse of the Cholesky decomposition of the phylogenetic variance-covariance matrix to correct for phylogenetic relationships. Dotted lines show the partial regression slope $\pm$ one standard error. All logarithms are base 10.

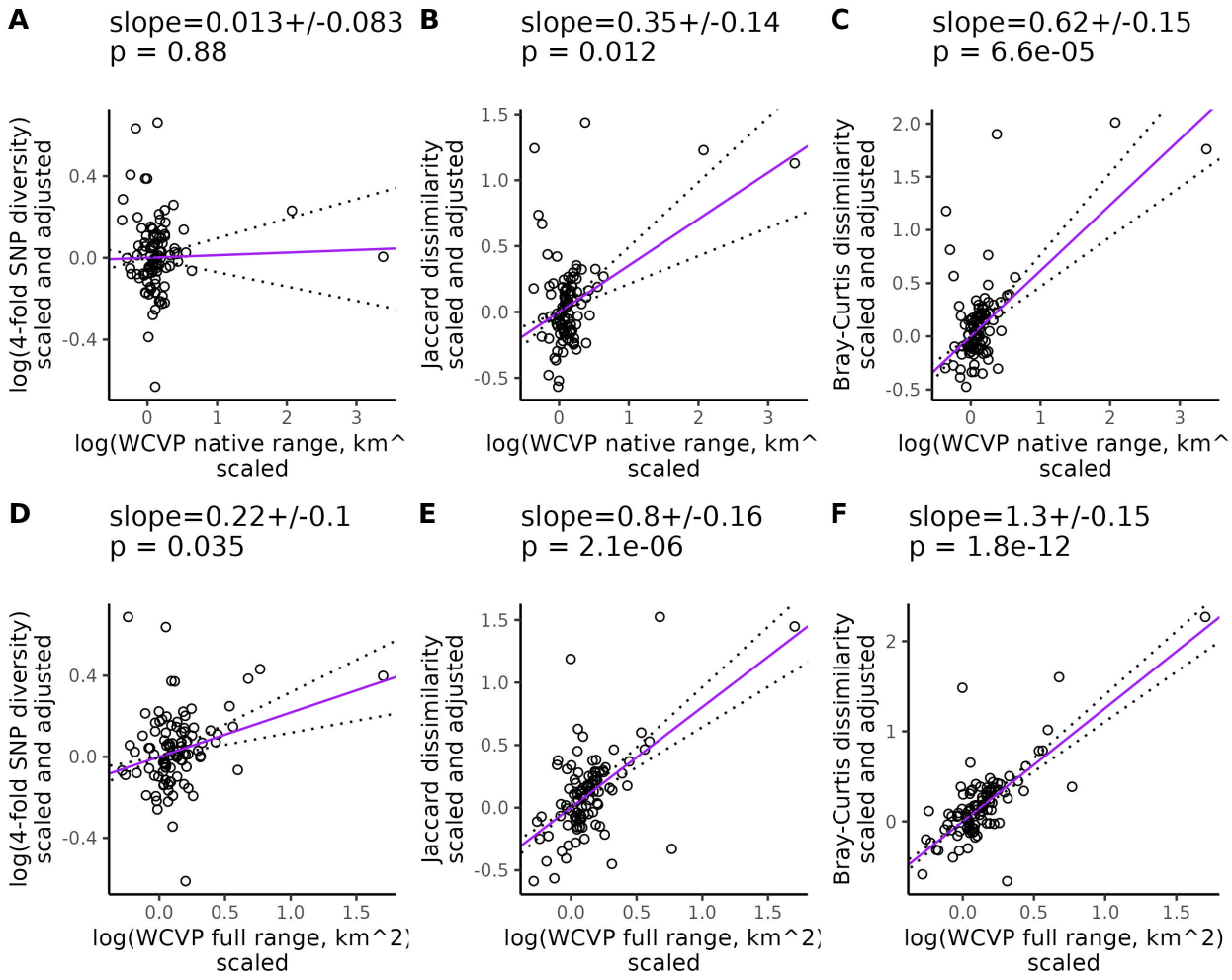

Figure C18: **Partial phylogenetic regression between WCVP range size and diversity, controlling for genome size.** Each point is a species and only species with > 0.5x mean coverage and > 1000 variant sites were included in the regression. The regressions are organized according to whether invaded ranges were excluded (A-C) or included (D-F) in the range size estimates. Lines give the relationship between the pairs of plotted variables after controlling for mating system, life cycle habit, cultivation status, genome size, and evolutionary history. The statistics across the top of each plot give the value of the slope of the lines ($\pm$ the standard error) and the p-value testing whether the slope differs from zero. Before fitting the line, each response variable was scaled to a standard normal distribution (mean = 0, variance = 1), then multiplied by the inverse of the Cholesky decomposition of the phylogenetic variance-covariance matrix to correct for phylogenetic relationships. Dotted lines show the partial regression slope $\pm$ one standard error. All logarithms are base 10.

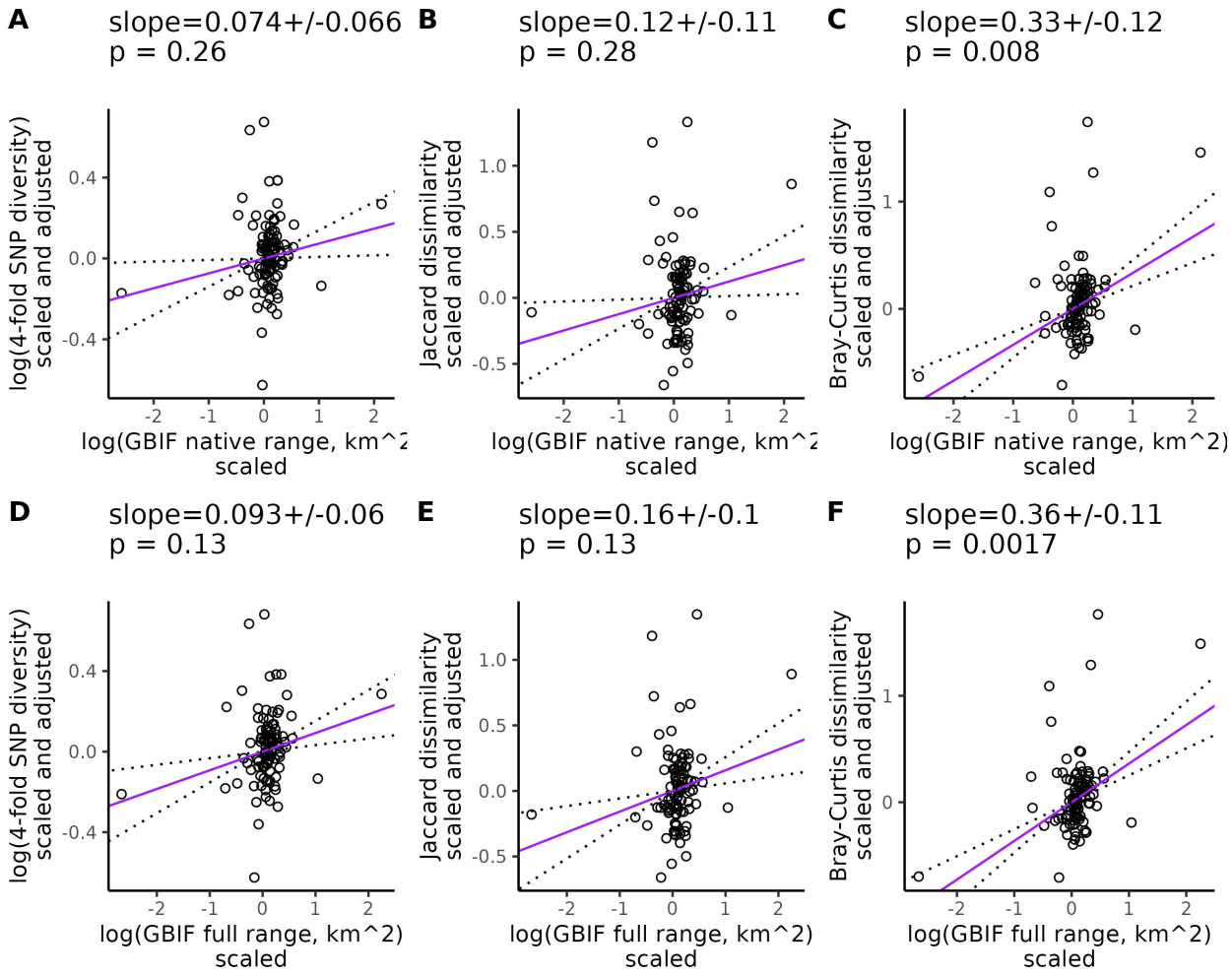

Figure C19: **Partial phylogenetic regression between GBIF range size and diversity, controlling for genome size.** Each point is a species and only species with $>$ 0.5x mean coverage and $>$ 1000 variant sites were included in the regression. The regressions are organized according to whether invaded ranges were excluded (A-C) or included (D-F) in the range size estimates. Lines give the relationship between the pairs of plotted variables after controlling for mating system, life cycle habit, cultivation status, genome size, and evolutionary history. The statistics across the top of each plot give the value of the slope of the lines ($\pm$ the standard error) and the p-value testing whether the slope differs from zero. Before fitting the line, each response variable was scaled to a standard normal distribution (mean = 0, variance = 1), then multiplied by the inverse of the Cholesky decomposition of the phylogenetic variance-covariance matrix to correct for phylogenetic relationships. Dotted lines show the partial regression slope $\pm$ one standard error. All logarithms are base 10.

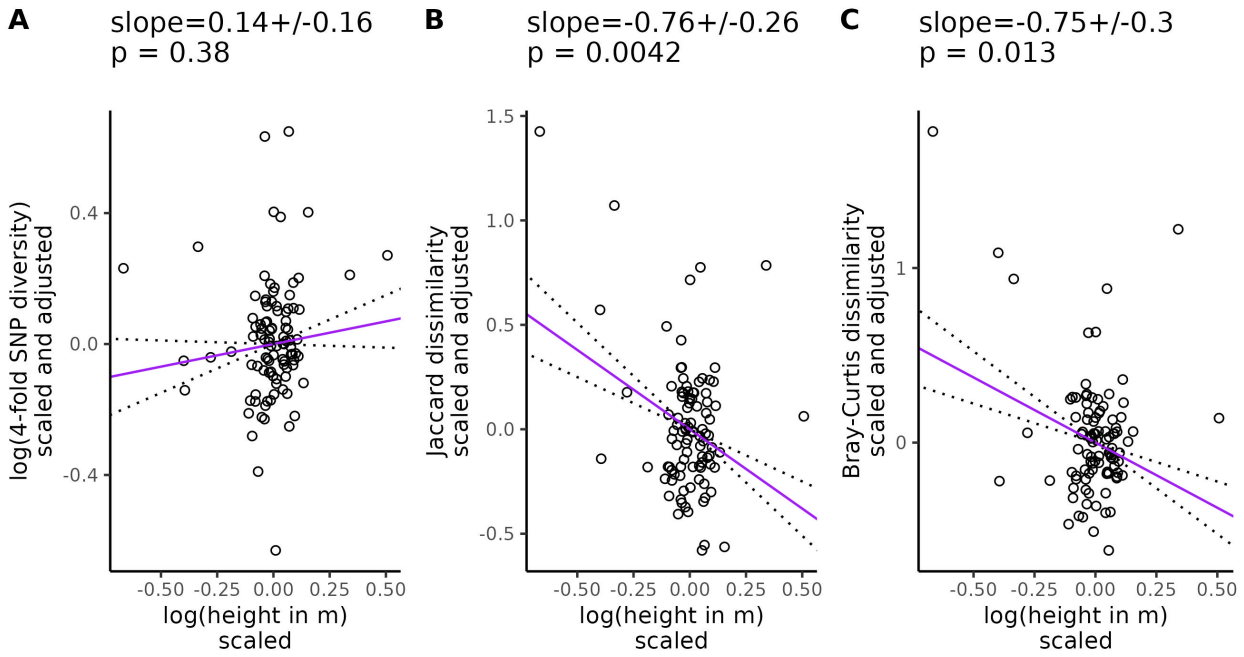

Figure C20: **Partial phylogenetic regression between height and diversity, controlling for genome size.** Each point is a species and only species with $>$ 0.5x mean coverage and $>$ 1000 variant sites were included in the regression. Lines give the relationship between the pairs of plotted variables after controlling for mating system, life cycle habit, cultivation status, genome size, and evolutionary history. The statistics across the top of each plot give the value of the slope of the lines ($\pm$ the standard error) and the p-value testing whether the slope differs from zero. Before fitting the line, each response variable was scaled to a standard normal distribution (mean = 0, variance = 1), then multiplied by the inverse of the Cholesky decomposition of the phylogenetic variance-covariance matrix to correct for phylogenetic relationships. Dotted lines show the partial regression slope $\pm$ one standard error. All logarithms are base 10.

Genome size vs diversity relationships after controlling for population size proxies, phylogeny, and life-history variables

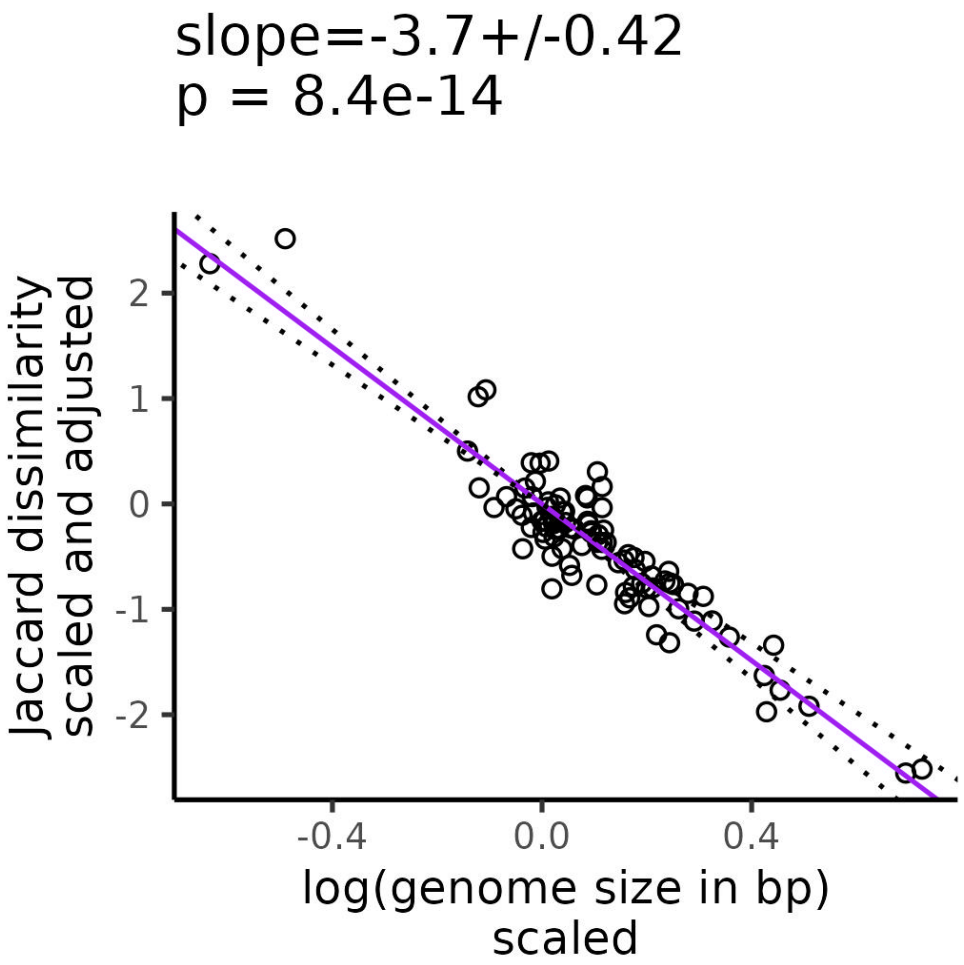

Figure C21: **Partial phylogenetic regression between Jaccard dissimilarity and genome size, controlling for WCVP full range size-height ratio.** Each point is a species and only species with > 0.5x mean coverage and > 1000 variant sites were included in the regression. WCVP full ratio gives the ratio of range size to squared plant height, where range size includes invaded ranges and is estimated from WCVP range maps. The partial regression controls for range size-squared height ratio, mating system, life cycle habit, cultivation status, and evolutionary history. Before fitting the line, each response variable was scaled to a standard normal distribution (mean = 0, variance = 1), then multiplied by the inverse of the Cholesky decomposition of the phylogenetic variance-covariance matrix to correct for phylogenetic relationships. The values at the top of the plot give the slope of the partial regression $\pm$ one standard error and p-values testing whether the slope differ from zero. Dotted lines show the partial regression slope $\pm$ one standard error. All logarithms are base 10.

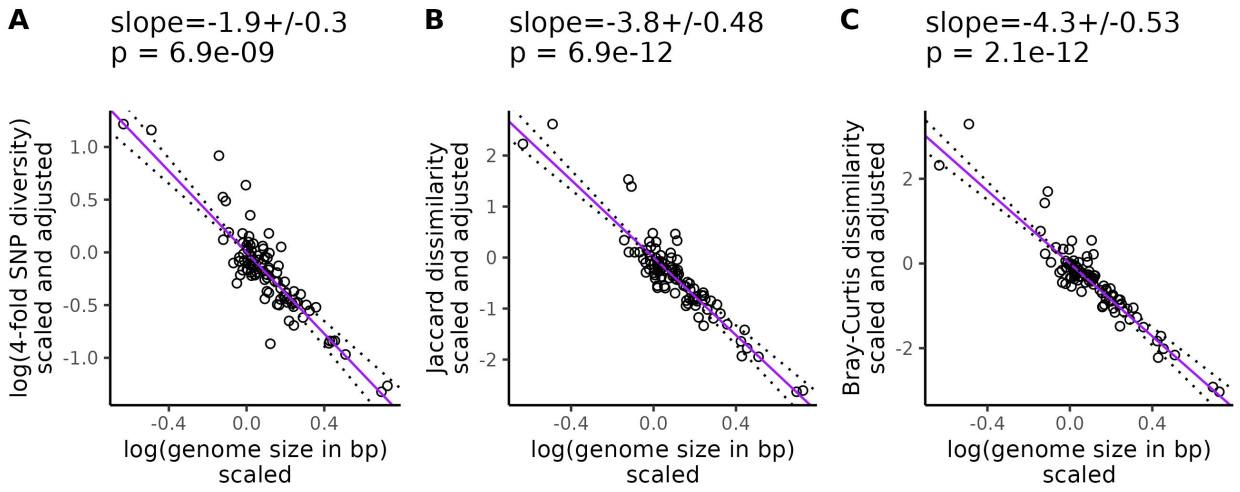

Figure C22: **Partial phylogenetic regression between diversity and genome size, controlling for WCVP native range size-height ratio.** Each point is a species and only species with > 0.5x mean coverage and > 1000 variant sites were included in the regression. WCVP native ratio gives the ratio of range size to squared plant height, where range size excludes invaded ranges and is estimated from WCVP range maps. The partial regression controls for genome size, mating system, life cycle habit, cultivation status, and evolutionary history. Before fitting the line, each response variable was scaled to a standard normal distribution (mean = 0, variance = 1), then multiplied by the inverse of the Cholesky decomposition of the phylogenetic variance-covariance matrix to correct for phylogenetic relationships. The values at the top of the plot give the slope of the partial regression $\pm$ one standard error and p-values testing whether the slope differ from zero. Dotted lines show the partial regression slope $\pm$ one standard error. All logarithms are base 10.

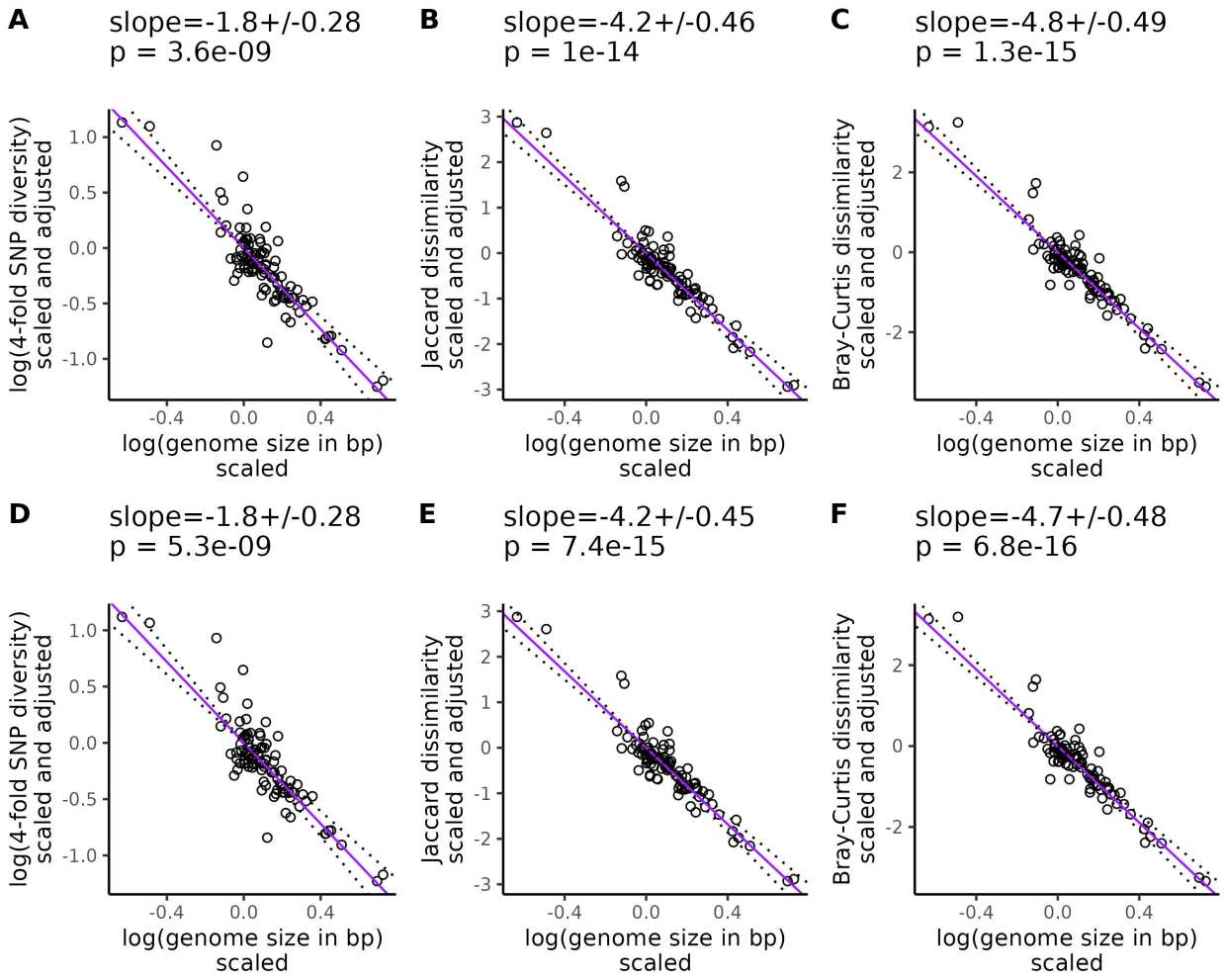

Figure C23: **Partial phylogenetic regression between diversity and genome size, controlling for GBIF range size-height ratio.** Each point is a species and only species with > 0.5x mean coverage and > 1000 variant sites were included in the regression. The regressions are organized according to whether invaded ranges were excluded (A-C) or included (D-F) in the range size-squared height ratio. Lines give the relationship between the pairs of plotted variables after controlling for mating system, life cycle habit, cultivation status, range size-squared height ratio, and evolutionary history. The statistics across the top of each plot give the value of the slope of the lines ($\pm$ the standard error) and the p-value testing whether the slope differs from zero. Before fitting the line, each response variable was scaled to a standard normal distribution (mean = 0, variance = 1), then multiplied by the inverse of the Cholesky decomposition of the phylogenetic variance-covariance matrix to correct for phylogenetic relationships. Dotted lines show the partial regression slope $\pm$ one standard error. All logarithms are base 10.

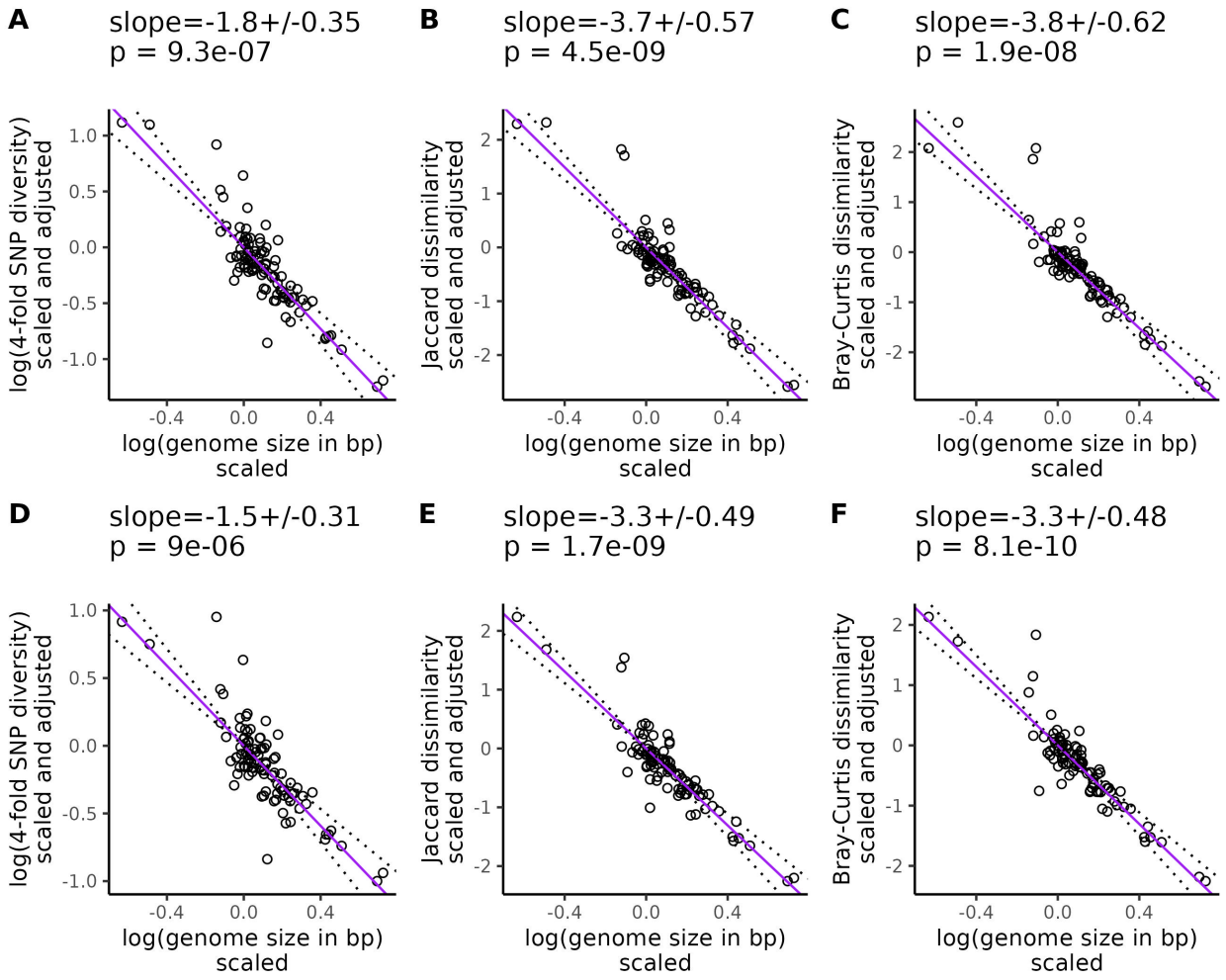

Figure C24: **Partial phylogenetic regression between diversity and genome size, controlling for WCVP range size.** Each point is a species and only species with > 0.5x mean coverage and > 1000 variant sites were included in the regression. The regressions are organized according to whether invaded ranges were excluded (A-C) or included (D-F) in the range size estimates. Lines give the relationship between the pairs of plotted variables after controlling for mating system, life cycle habit, cultivation status, range size, and evolutionary history. The statistics across the top of each plot give the value of the slope of the lines ($\pm$ the standard error) and the p-value testing whether the slope differs from zero. Before fitting the line, each response variable was scaled to a standard normal distribution (mean = 0, variance = 1), then multiplied by the inverse of the Cholesky decomposition of the phylogenetic variance-covariance matrix to correct for phylogenetic relationships. Dotted lines show the partial regression slope $\pm$ one standard error. All logarithms are base 10.

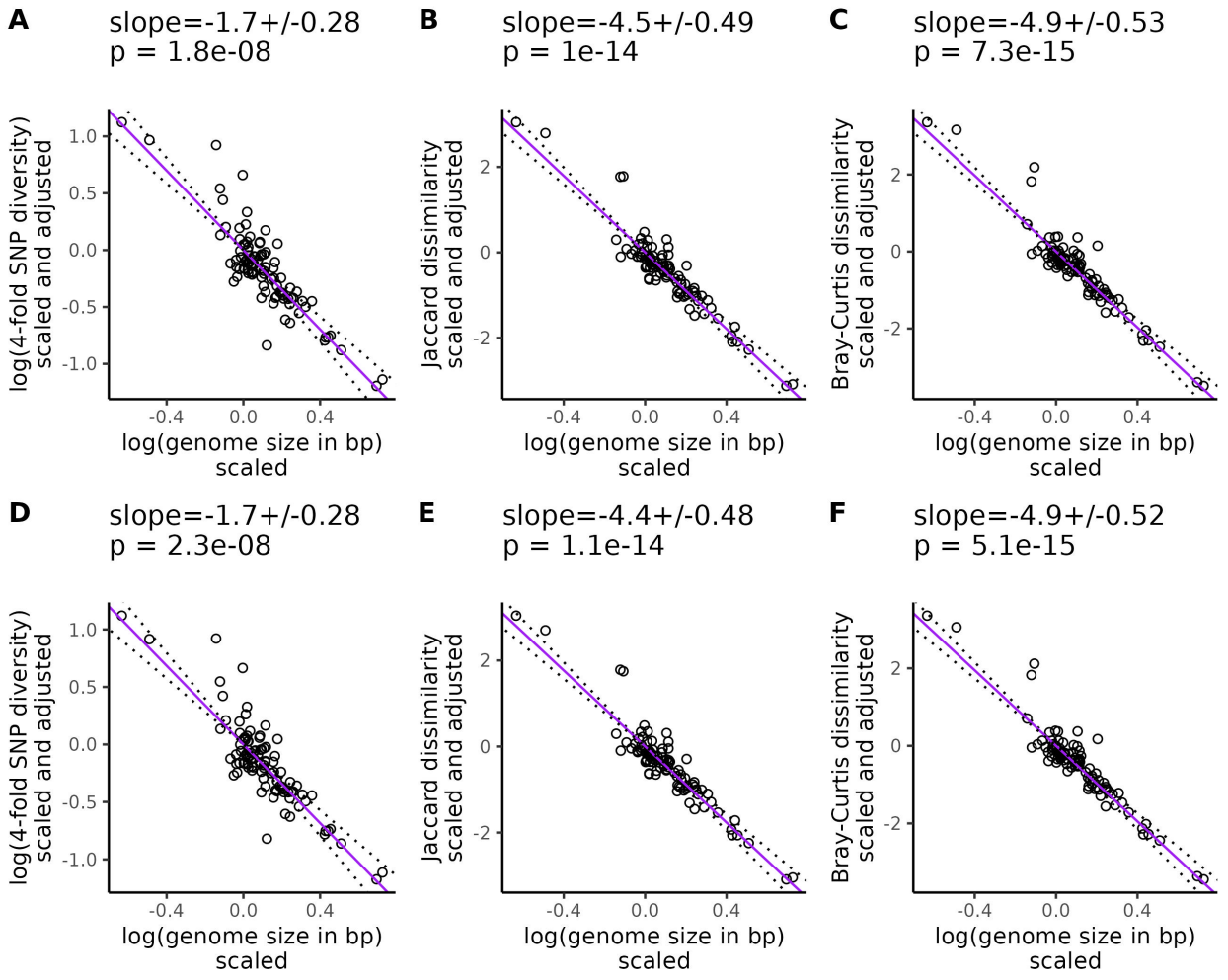

Figure C25: **Partial phylogenetic regression between diversity and genome size, controlling for GBIF range size.** Each point is a species and only species with > 0.5x mean coverage and > 1000 variant sites were included in the regression. The regressions are organized according to whether invaded ranges were excluded (A-C) or included (D-F) in the range size estimates. Lines give the relationship between the pairs of plotted variables after controlling for mating system, life cycle habit, cultivation status, range size, and evolutionary history. The statistics across the top of each plot give the value of the slope of the lines ($\pm$ the standard error) and the p-value testing whether the slope differs from zero. Before fitting the line, each response variable was scaled to a standard normal distribution (mean = 0, variance = 1), then multiplied by the inverse of the Cholesky decomposition of the phylogenetic variance-covariance matrix to correct for phylogenetic relationships. Dotted lines show the partial regression slope $\pm$ one standard error. All logarithms are base 10.

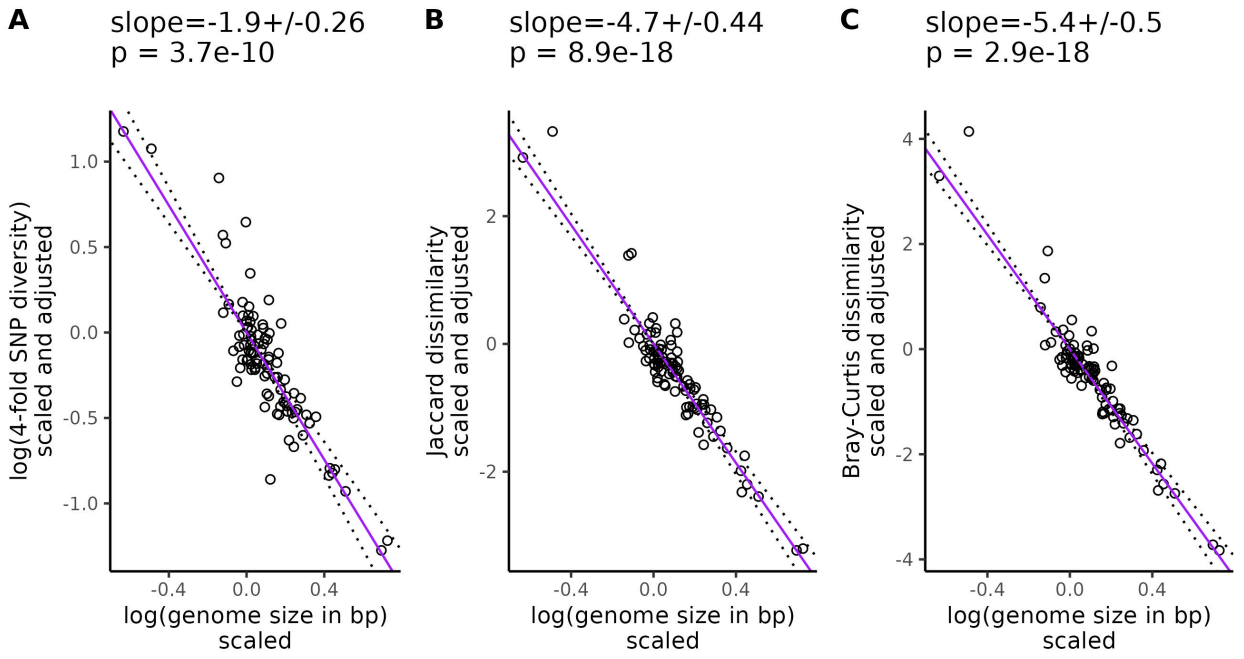

Figure C26: **Partial phylogenetic regression between diversity and genome size, controlling for height.** Each point is a species and only species with $>$ 0.5x mean coverage and $>$ 1000 variant sites were included in the regression. The regressions are organized according to whether invaded ranges were excluded (A-C) or included (D-F) in the range size-squared height ratio. Lines give the relationship between the pairs of plotted variables after controlling for mating system, life cycle habit, cultivation status, height, and evolutionary history. The statistics across the top of each plot give the value of the slope of the lines ($\pm$ the standard error) and the p-value testing whether the slope differs from zero. Before fitting the line, each response variable was scaled to a standard normal distribution (mean = 0, variance = 1), then multiplied by the inverse of the Cholesky decomposition of the phylogenetic variance-covariance matrix to correct for phylogenetic relationships. Dotted lines show the partial regression slope $\pm$ one standard error. All logarithms are base 10.

APPENDIX D: SUPPLEMENTAL FIGURES FOR CHAPTER 3

The effect of coverage on k-mer measures

We observe in Figure D1 a complex interaction between coverage and the resulting Bray-Curtis score depending on the length of k. With 30-mers, there is very little difference in the Bray-Curtis score. However, 10-mers are more affected by changes in coverage. The overall trend and ranks of the scores remain, but the Bray-Curtis score is generally higher with 10x coverage.

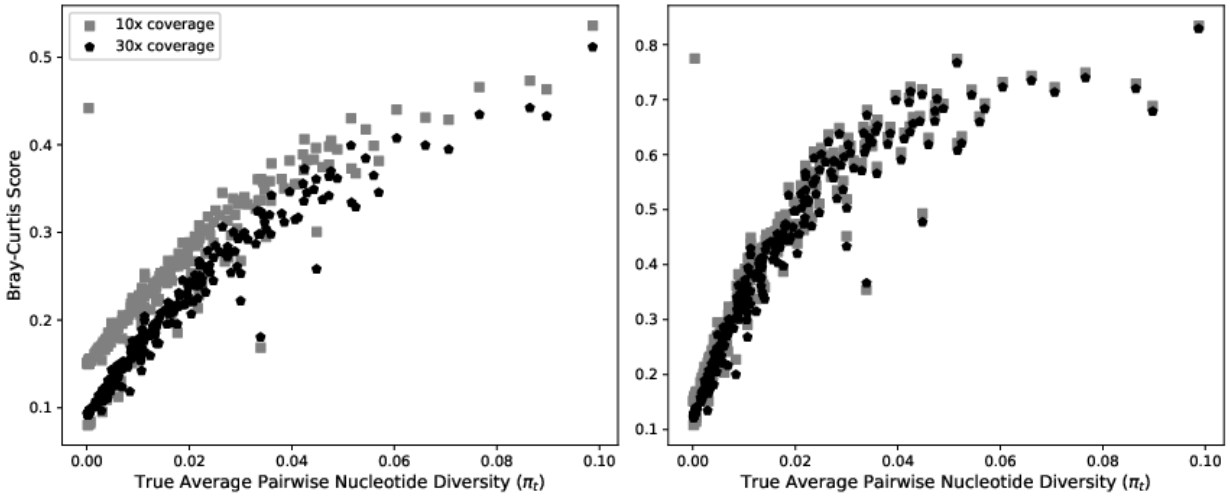

Figure D1: **Effect of k and coverage on Bray-Curtis dissimilarity.** Bray-Curtis scores calculated with 10-mers (left) and 30-mers (right). For each sample, reads with 10x coverage (gray/square) and 30x coverage (black/pentagon) were simulated. After simulating reads, the k-mers were counted and then the Bray-Curtis score is calculated. Each point represents one sample with a specific k and coverage.

Adjusting array size of Counting Bloom Filter

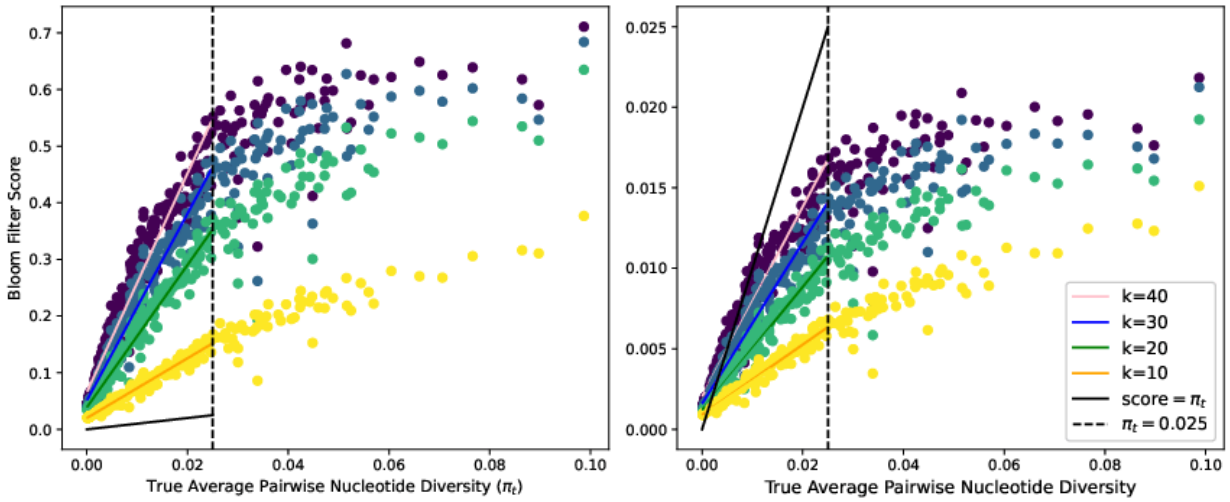

Figure D2: **Effect of bloom filter size on cosine dissimilarity.** A comparison of cosine similarity scores of CBFs with an array size of 20 million (left) and 10,000 (right). The black line shows the 1-to-1 mapping of π_t to the dissimilarity score. Note the difference in the scales of the scores on the y-axes, and that the right panel shows points that are closer to the 1-to-1 line.

APPENDIX E: SUPPLEMENTAL FIGURES FOR CHAPTER 4

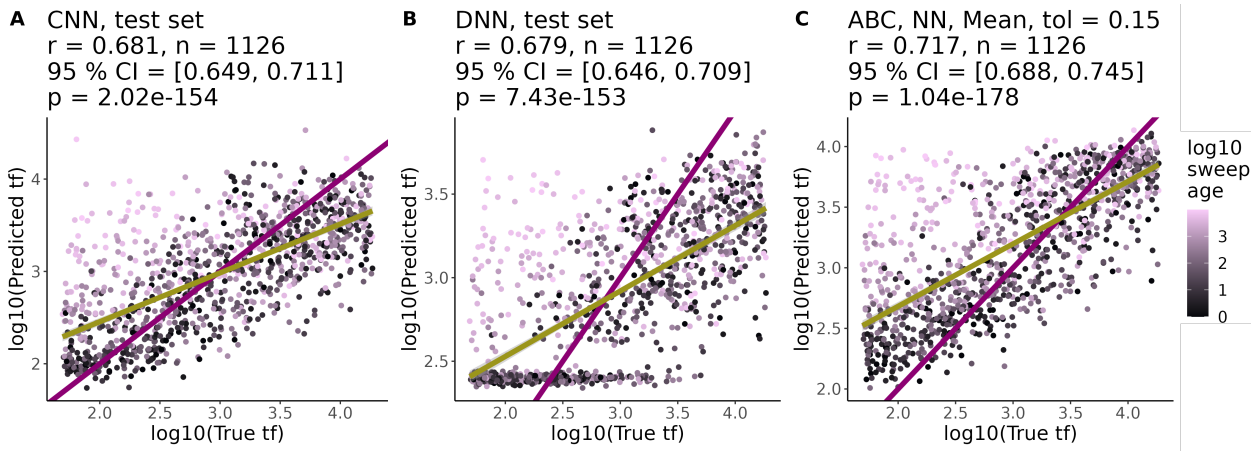

Figure E1: **Comparison of best (A) CNN, (B), DNN, and (C) ABC models at predicting times to fixation for the decay demography scenario.**

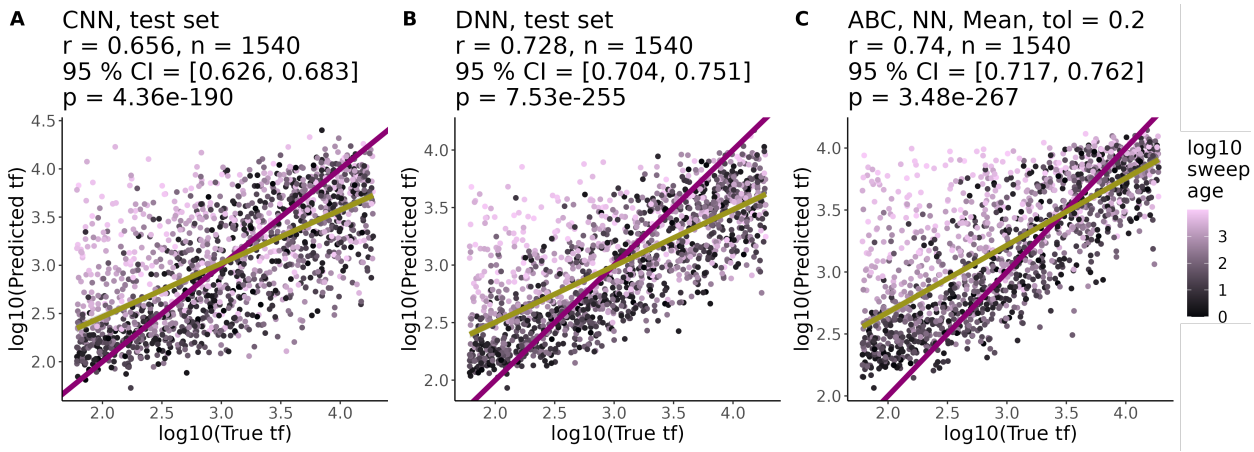

Figure E2: **Comparison of best (A) CNN, (B), DNN, and (C) ABC models at predicting times to fixation for the cycle demography scenario.**

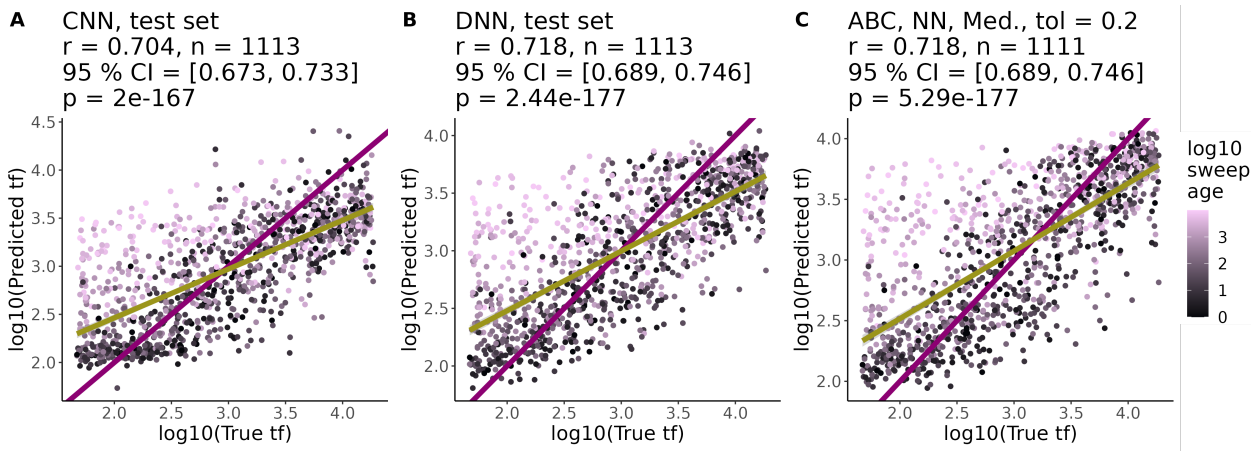

Figure E3: **Comparison of best (A) CNN, (B), DNN, and (C) ABC models at predicting times to fixation for the chaos demography scenario.**

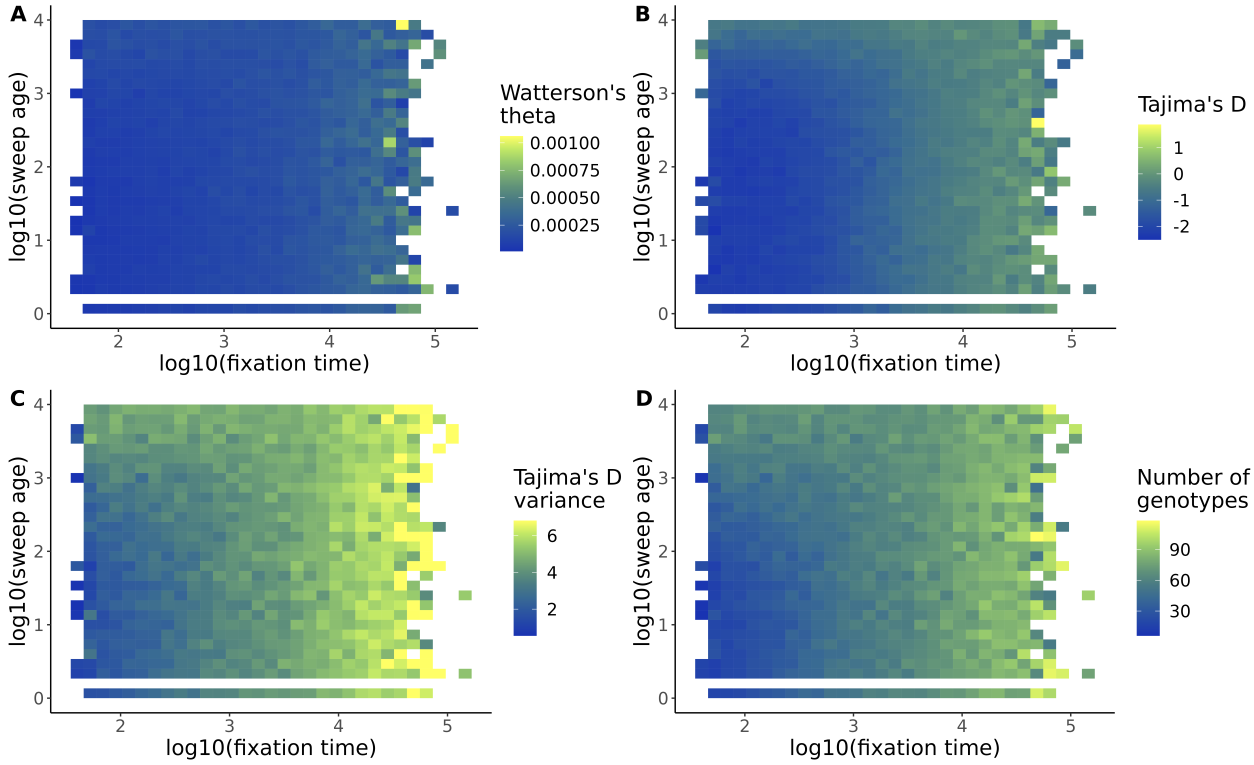

Figure E4: t_f **vs** t_a **for statistics based on the site frequency spectrum for sweeps in constant-sized populations.**

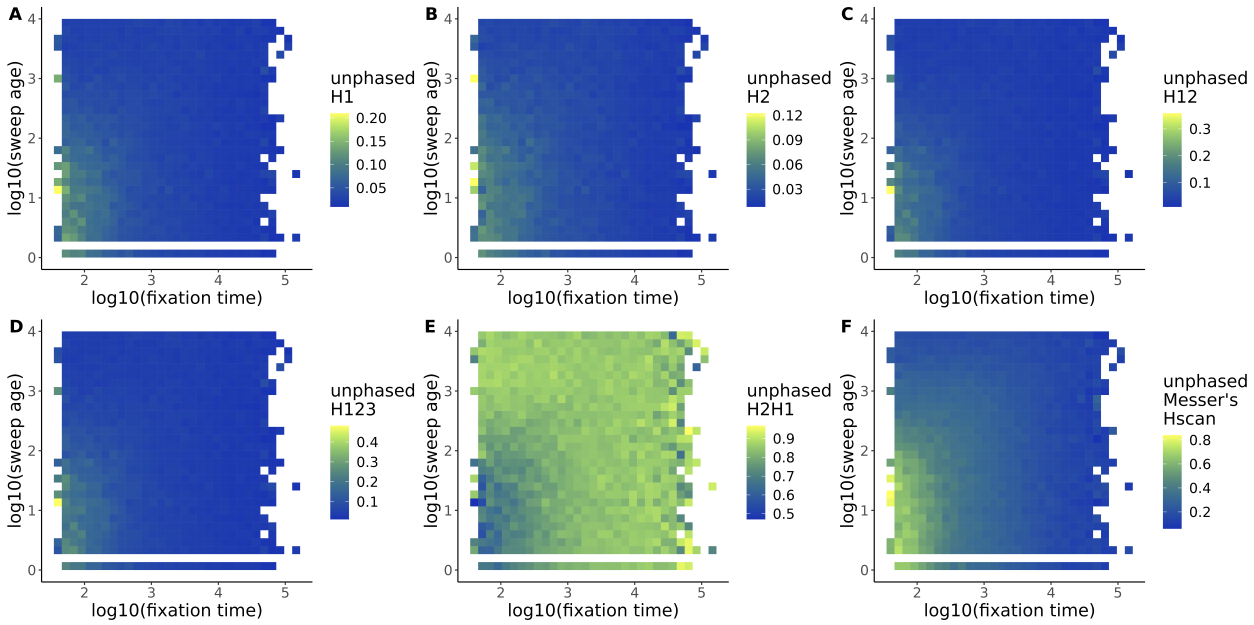

Figure E5: t_f **vs** t_a **for unphased H statistics and unphased Messer's Hscan in constant-sized populations.**

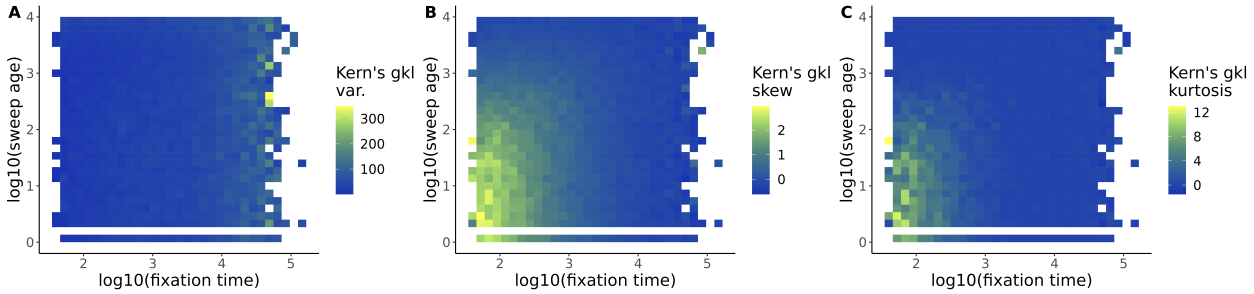

Figure E6: t_f **vs** t_a **for Kern's** g_{kl} **statistics in constant-sized populations.**

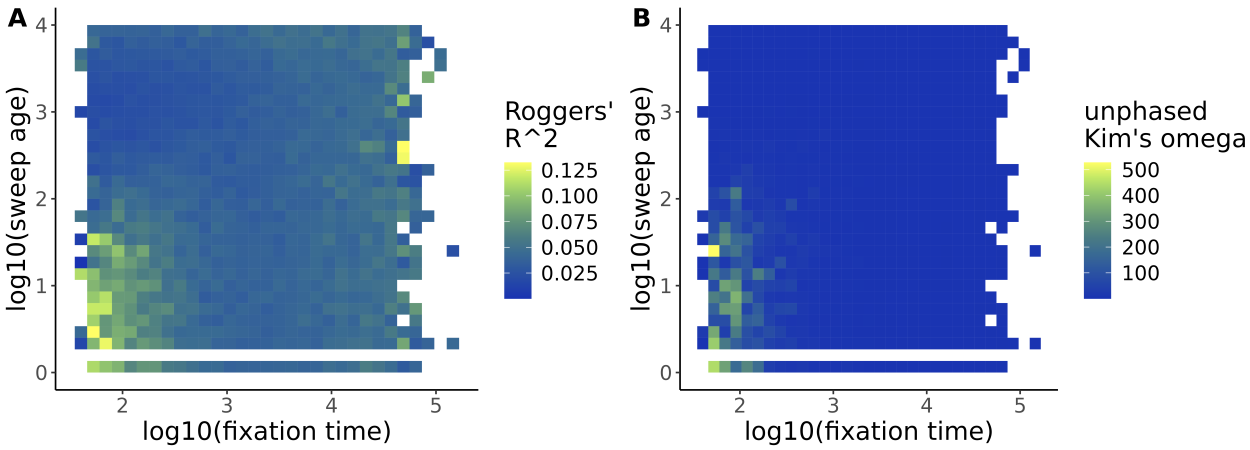

Figure E7: t_f **vs** t_a **for Rogers'** R^2 **and Kim's** ω**.**