Title

Polyploidy in Andropogon gerardi: A Series of Happy Accidents

Author

Phillips, Alyssa

Publication Date

2024

Peer reviewed|Thesis/dissertation

Polyploidy in *Andropogon gerardi*: A Series of Happy Accidents

By

ALYSSA PHILLIPS

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Plant Biology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Jeffrey Ross-Ibarra, Chair

Elisabeth Forrestel

Jennifer Gremer

Committee in Charge

2024

Abstract

This dissertation explores the ecology and evolution of polyploid species, or organisms with multiple sets of chromosomes from whole genome duplication (WGD). Polyploidy is a widespread phenomenon across the tree of life, significantly impacting both animal and plant evolution. Recent advancements in genomics technology has made the study of polyploid species substantially more feasible, but the complexity of polyploid biology is still limiting. As a result, many questions on the evolutionary role of polyploidy remain unanswered. Here, I study the origin and ecological role of polyploidy in *Andropogon gerardi* Vitman, a mixed-ploidy and ecologically dominant prairie grass, using a novel reference genome and whole genome sequencing data. We found mixed-ploidy in *A. gerardi* is a result of recurrent polyploid formation, or happy little accidents. Further, we found WGD in *A. gerardi* confers immediate adaptive phenotypic changes. In the course of this research, we fortuitously assembled a reference genome for *Poa pratensis*, an economically valuable and popular turfgrass, instead of a second *A. gerardi* genome. Finally, I synthesize the challenges of variant calling in polyploids due to extensive genomic diversity and a lack of genomic resources and I propose a variant calling pipeline to addresses key challenges. Overall, this dissertation enhances our understanding of polyploidy's role in plant evolution and environmental adaptation.

Dedication

To my grandma, Dorothy Schwartz, for her unfaltering love and support. Thank you for reminding me the pursuit of your passions is important even in the most challenging times.

Acknowledgements

My dissertation was made possible by numerous people and funding organizations. I would like to express my deepest appreciation to my advisor, Jeffrey Ross-Ibarra, for encouraging me to follow my interests in polyploidy and keeping me on track when things went awry. I could also not have accomplished this work without my dissertation committee, Elisabeth Forrestel and Jennifer Gremer, who shared knowledge and expertise on plant ecophysiology and adaptation that greatly improved my common garden. Additionally, my research was made possible by support from the National Science Foundation, the University of California, Davis, the Botanical Society of America, and the Davis Botanical Society.

This dissertation was data-intensive and would not have been possible without phenomenal collaborators. In particular, Elizabeth A. Kellogg and Taylor AuBuchon-Elder led a massive plant collection effort that provided all of the plant material for my research. Further, Dr. Kellogg was inspired to submit a proposal for a *Andropogon gerardi* genome to the Department of Energy Joint Genome Institute (JGI) in 2017 and generously allowed me to use the genome and the associated data in my research. I am also grateful to M. Cinta Romay, Robert J. Soreng, and Qi Sun for the generation and maintenance of the PanAnd whole genome sequence dataset. Special thanks to my co-first author Arun S. Seetharam for your extensive knowledge of genome assembly and salvaging the *Poa* genome.

I had the pleasure of working with and learning from many wonderful people in the Ross-Ibarra lab. Each person's mentorship and friendship have left a lasting impression. In particular, I would like to thank Elli Cryan for her friendship and for driving a minivan full of plants halfway across the country with me, Sarah Turner-Hissong and Silas Tittes for teaching me bioinformatics, and Catherine Rushworth for her continued mentorship.

I would also like to mention the importance of my friends and family throughout this journey. Thank you for keeping me grounded and providing space for me to decompress away from my research. Lastly, I have endless gratitude for Andrew L. Murray, my number one fan. Thank you, Andrew, for listening to my practice talks, helping me re-pot plants, quizzing me on my qualifying

exam material, going on very slow wildflower hikes, talking through new research ideas, and importantly, reminding me that I am more than my work. This dissertation would not have been possible without you.

Contents

Introduction **1**

Chapter 1: The consequences of polyploidy in adaptation of a dominant prairie grass **3**

1. Abstract 4
2. Introduction 4
3. Results 7
   - 3.1 Assembly of a subgenome- and haplotype- resolved reference assembly . . 7
   - 3.2 Assessment of the impact of habitat fragmentation on population structure and genetic diversity 7
   - 3.3 Origins of mixed-ploidy populations 9
   - 3.4 The effect of polyploidy on growth and reproductive effort 10
   - 3.5 The effect of polyploidy on leaf morphology and economics 13
4. Discussion 15
   - 4.1 Minimal population structure and high genetic diversity in *Andropogon gerardi* despite modern habitat fragmentation 16
   - 4.2 Mixed-ploidy is maintained by recurrent polyploidization 18
   - 4.3 The consequences of neopolyploidy in adaptation 19
5. Conclusion 21
6. Methods 22
   - 6.1 Sample collection 22
   - 6.2 Short-read sequencing of population panel 23
   - 6.3 Genome sequencing for the *A. gerardi* reference genome 25
   - 6.4 Genome size estimation 25
   - 6.5 Genome assembly and construction of pseudomolecule chromosome 26

6.6 Genome annotation . . . 29
6.7 Variant calling and genotyping . . . 30
6.8 Assessment of population structure and diversity . . . 32
6.9 Estimation of hexaploid genetic diversity . . . 33
6.10 Common garden experiment . . . 33
6.11 Phenotyping leaf functional traits . . . 34
6.12 Phenotyping performance traits . . . 35
6.13 Trait data analysis . . . 36
6.14 Data availability . . . 38
7. Acknowledgements . . . 38
8. Supporting Information . . . 40

Chapter 2: A happy accident: A novel turfgrass reference genome 64

1. Abstract . . . 65
2. Background . . . 65
3. Materials & Methods . . . 67
3.1 Sample collection . . . 67
3.2 PacBio sequencing . . . 68
3.3 Bionano optical map generation . . . 69
3.4 Preparation and imaging of metaphase spreads . . . 69
3.5 Illumina sequencing of the *Poa* population panel . . . 70
3.6 Species identification . . . 71
3.7 Genome assembly . . . 73
3.8 Genome annotation . . . 74
3.9 Assessment of the assembly . . . 75
3.10 Population genetics of *Poa* . . . 76
4. Results and Discussion . . . 77
4.1 Species identification and validation . . . 77

4.2 Genome size and ploidy estimation . . . 78
4.3 Genome assembly . . . 78
4.4 Genome annotation . . . 79
4.5 Application of the reference genome . . . 80
4.6 Population genetics of North American Poa . . . 80
5. Conclusion . . . 82
6. Data availability . . . 83
7. Acknowledgments . . . 83
8. Funding . . . 83
9. Supplement . . . 84

Chapter 3: Variant calling in polyploids for population and quantitative genetics **85**

1. Abstract . . . 85
2. Background . . . 85
3. Challenges to variant calling in polyploid systems . . . 88
3.1 Resource requirements scale with genome size . . . 88
3.2 Genome-wide redundancy and elevated polymorphism increase errors in read mapping . . . 91
3.3 Incomplete or misassembled polyploid reference genomes increase genotyping error . . . 92
3.4 Allele dosage cannot be determined if ploidy and inheritance mode are unknown . . . 94
3.5 Existing tools cannot account for further biological complexity . . . 95
4. Proposed solutions to incorporate polyploid complexity in variant calling . . . 96
4.1 Balancing sequencing depth and precision may reduce sequencing costs . . . 96
4.2 Alternative read alignment approaches, genotype callers, and variant filters may reduce errors caused by poor read mapping . . . 97

4.3 Information on ploidy, chromosome inheritance mode, and reference quality can be integrated to determine allele dosage 100
4.4 Current accepted practices for navigating polyploid data with additional biological complexity 101
5. Conclusion 102
6. Acknowledgments 104
7. Appendix 104
7.1 A brief overview of variant calling 104

Conclusion **108**

References **109**

Introduction

My dissertation aims to study the ecology and evolution of polyploid species. Polyploids are organisms with two or more sets of chromosomes from whole genome duplication (WGD). Although we typically discuss genomes as diploid or haploid, polyploids are incredibly common throughout the tree of life. In animals, polyploidy is a well-described feature of cancer cells and common in some lineages of fish, frogs, salamanders and hexapods (Zack *et al.* 2013; Román-Palacios *et al.* 2021; David 2022; Li *et al.* 2018). The role of polyploidy in plants cannot be understated; it is thought to contribute to plant diversification and speciation (One Thousand Plant Transcriptomes Initiative 2019). Over 35% of extant plant species are polyploids, but the entire plant kingdom shares an ancestral WGD event (Wood *et al.* 2009a; One Thousand Plant Transcriptomes Initiative 2019).

Polyploid species are highly diverse in their evolutionary origins and genome structure. Polyploids are broadly described as belonging to two groups defined by the evolutionary origins of the subgenomes: allopolyploids, which form through hybridization of two or more species, and autopolyploids, which are formed through genome doubling within a single species. Allopolyploids are generally defined as having disomic inheritance, where meiosis is similar to diploids and chromosomes preferentially pair with their sister homolog from the ancestral genome. Comparatively, autopolyploids are defined as having non-preferential pairing. In reality, the cytological definitions of these groups are less discrete and fall along a gradient (Stebbins 1950; Mason and Wendel 2020; Meirmans and Van Tienderen 2013). Further, many lineages were formerly polyploids, having undergone genome fractionation and reorganization to return to two sets of chromosome with preferential pairing (i.e. diploidization; Ma and Gustafson 2005). These lineages, known as paleopolyploids, have remnant polyploid features in their genomes, such as duplicate gene copies that contribute to novel gene functions (Ohno 2013). Polyploid species can vary further in traits such aschromosome count (aneuploidy), haploid genome size, severity of initial WGD bottleneck, and age since polyploidization.

Study of diverse polyploid species has been limited due to a lack of genomic resources and methods. As a result, a number of evolutionary questions remain unanswered: How do polyploid species establish? Under what conditions is polyploidy adaptive and maladaptive? What are the demographic consequences of polyploidization? Can polyploid species rapidly respond to climate change? I investigated these questions in *Andropogon gerardi* Vitman, a mixed-ploidy species and the dominant prairie grass in endangered North American tallgrass prairies (Chapter 1). This study utilized a novel *A. gerardi* reference genome, whole genome sequence (WGS) and genome size data for 180 genotypes, and a two-year common garden experiment.

During my study of *A. gerardi*, we initially accidentally sequenced and assembled the genome of a weedy grass that had entangled itself within the pot of the *A. gerardi* reference plant. This weed was Kentucky bluegrass (*Poa pratensis*), one of the most globally popular turfgrass species. We were able to salvage the accidental assembly, as well as some other contaminated WGS data, create the first genomic resources for this polyploid C3 grass (Chapter 2).

After having navigated multiple polyploid WGS datasets, I synthesized the barriers I encountered in variant calling in a review. Variant calling is the first step in any genomics, population genetics, or quantitative genetics study and has unique challenges in polyploids. I proposed a variant calling pipeline that addresses the identified barriers provided a compressive guide for researchers beginning to work in polyploid systems or new to WGS datasets (Chapter 3).

Chapter 1: The consequences of polyploidy in adaptation of a dominant prairie grass

Alyssa R. Phillips[a,b,1], Taylor AuBuchon-Elder[c], Edward S. Buckler[d,e,f], Robert Bukowski[g], Brenda Cameron[a], Elli Cryan[a,b,h], Elisabeth Forrestel[i], Jane Grimwood[j], John T. Lovell[j,k], Patrick Minx[c], Julianna Porter[a], Jeremy Schmutz[j,k], Britney Solomon[a], Qi Sun[l], Sherry Flint-Garcia[m], M. Cinta Romey[d], Elizabeth A. Kellogg[c], and Jeffrey Ross-Ibarra[a,b,n]

[a]Department of Evolution and Ecology, University of California, Davis, Davis, CA 95616

[b]Center for Population Biology, University of California, Davis, Davis, CA 95616

[c]Donald Danforth Plant Science Center, Olivette, MO

[d]School of Integrative Plant Sciences, Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY

[e]Institute for Genomic Diversity, Cornell University, Ithaca, NY

[f]Agricultural Research Service, United States Department of Agriculture, Ithaca, NY

[g]Bioinformatics Facility, Institute of Biotechnology, Cornell University, Ithaca, NY

[h]Department of Plant Sciences, University of California, Davis, Davis, CA 95616

[i]Department of Viticulture and Enology, University of California, Davis, Davis, CA

[j]HudsonAlpha Institute for Biotechnology, Huntsville, AL

[k]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA

[l]Genomic Diversity Facility, Cornell University, Ithaca, New York

[m]USDA-ARS, Plant Genetics Research Unit, Columbia, MO

[n]Genome Center, University of California, Davis, Davis, CA

1. Abstract

Polyploidy is ubiquitous across the tree of life but has played an outsized role in the evolution of plants. It has been well-documented that whole genome duplication (WGD) can alter gene expression, biochemistry, physiology, and morphology. The benefits of WGD are highly dependent on the environment and may explain why polyploids are associated with disturbance, invasiveness, and stress. Progress has been made in understanding the conditions in which ploidy is beneficial primarily in ecological simulations and synthetic polyploids, but studies in natural mixed-ploidy species are lacking. To address this gap, we are studying the role of mixed-ploidy in the local adaptation of *Andropogon gerardi* Vitman, the dominant grass species in endangered North American tallgrass prairies. *A. gerardi* is composed of hexaploids ($2n = 6x$) and enneaploids ($2n = 9x$), which are equally abundant but have distinct ranges; previous research has found the $9x$ cytotype is more common in regions with reduced precipitation and increased variation in temperature range. We have assembled a novel reference genome, whole genome sequenced 25 populations, and measured fitness and morphological variation in a common garden containing 14 populations. We found the $9x$ cytotype is produced through recurrent WGD and existing $9x$ genotypes are likely first-generation polyploids. Additionally, we find polyploidy significantly affects growth rate and stomatal traits, which may make the $9x$ cytotype more competitive in arid climates. Together with previous research documenting the $9x$ have low reproductive viability, our results indicate the $9x$ cytotype is an adaptive 'dead-end' but may locally outcompete the $6x$ cytotype in the short term in some environments. More broadly, our results suggest the benefits of polyploidy depend on the environment and mixed-ploidy is likely an ephemeral state of *A. gerardi* populations.

2. Introduction

Whole genome duplication (WGD) has played an outsized role in the evolution of plants (Wood *et al.* 2009b; One Thousand Plant Transcriptomes Initiative 2019; Li and Barker 2020) and some animal lineages (Román-Palacios *et al.* 2021; Zack *et al.* 2013; Li *et al.* 2018). The prevalence

of polyploidy, the state of having more than two sets of chromosomes, has confounded scientists as new polyploids are not favored to survive. The establishment of new polyploids is challenged by a demographic bottleneck and a frequency-dependent mating disadvantage that arises when the only mates available are the diploid progenitor (minority cytotype exclusion, Levin 1975). Further, new polyploids must overcome challenges associated with higher DNA content, increased resource requirements, meiotic abnormalities, epigenetic instability, and altered gene dosage and allele number (Bird *et al.* 2018; Ramsey and Schemske 2002). If initial barriers are overcome, this genomic novelty may create immediate phenotypic changes that aid establishment and adaptation to novel environments (Clo and Kolář 2021; Porturas *et al.* 2019).

Mixed-ploidy species can be leveraged to answer questions on the ecology of polyploid establishment. Mixed-ploidy, also referred to as intraspecific variation in ploidy, is common within plant species where at least 16% are estimated to contain multiple ploidy levels (Rice *et al.* 2015). The ploidy levels, known as cytotypes, are most commonly parapatric and inhabit separate niches; completely sympatric cytotypes are rare (Kolář *et al.* 2017). It is unclear whether mixed-ploidy species are in a transient state or represent long-term coexistence of multiple cytotypes. Multiple theoretical models suggest coexistence is stabilized by processes that provide reproductive assurance, like selfing and clonality, and reproductive isolation (Gaynor *et al.* 2023; Levin 1975). Perenniality may also aid cytotype coexistence as multiple reproductive cycles increase the chance of producing successful offspring (Van Drunen and Friedman 2022). Without one or more of these mechanisms, coexistence is unfavorable and one cytotype is expected to overcome the other. Our current understanding of the relative extent of these strategies in maintaining natural mixed ploidy populations is limited.

Here, we study mixed-ploidy in *Andropogon gerardi* Vitman (formerly *A. gerardii*), a prairie grass species composed of hexaploids ($2n = 6x = 60$) and enneaploids ($2n = 9x = 90$). The $6x$ cytotype is an allohexaploid with the parental species of the three subgenomes unknown but suspected to belong to the genera *Andropogon* (now *Anatherum*, Vorontsova *et al.* 2023) and *Schizachyrium* (Nagahama and Norrmann 2012; Estep *et al.* 2014). The $9x$ cytotype is an autopolyploid formed

by a reduced ($1n$) and unreduced ($2n$) $6x$ gamete creating a viable embryo (Norrmann *et al.* 1997). As a result, the $9x$ cytotype contains three copies of each chromosome resulting in a high rate of univalents and multivalents in meiosis and subsequently lower seed viability (Norrmann *et al.* 1997; Tompkins *et al.* 2015). Interestingly, the two cytotypes are found sympatrically range-wide (Tompkins *et al.* 2015; Keeler *et al.* 1987; Keeler 2004; Norrmann *et al.* 1997; Keeler 1990), but the $9x$ cytotype is more abundant in drier climates with higher diurnal and annual temperature ranges, such as the southwestern United States (McAllister *et al.* 2015). Other ploidy levels and aneuploids are found less than 5% range-wide (McAllister *et al.* 2015; Keeler 1990). The differing distributions of the two cytotypes suggest that $9x$ establishment or persistence is dependent on the environment.

Composing up to 80% of the biomass in endangered North American tallgrass prairies (Weaver 1968), *A. gerardi* is essential to tallgrass prairie structure and function. As less than 5% of tallgrass prairies remain (Samson and Knopf 1994) and climate change forecasts suggest suitable habitat for *A. gerardi* will be limited within the current range by 2070 (Smith *et al.* 2017), understanding the consequences of polyploidy in climate adaptation of *A. gerardi* populations is essential. Here, we aim to evaluate the effect of polyploidy on the genetic diversity and adaptive potential of *A. gerardi* populations. First, we utilize a range-wide whole genome sequence (WGS) dataset to assess whether genetic diversity of *A. gerardi* populations reflects the contemporary loss and fragmentation of North American prairies. Then, we test three interrelated and non-mutally exclusive hypotheses for how mixed-ploidy persists in *A. gerardi*: (Hypothesis 1) $9x$ individuals are continually input into the population through recurrent WGD events, (Hypothesis 2) $9x$ cytotype is maintained via vegetative growth, enabling it to spread clonally through the prairie, and (Hypothesis 3) the two cytotypes are or have undergone ecological differentiation preventing intraspecific competition. Given the low rates of gene flow expected between cytotypes, we are unable to test the role of interploidy reproduction in maintaining mixed-ploidy. Finally, we investigate the interaction between polyploidy and local adaptation by testing the effect of natural ploidal variation on ecologically relevant traits in a common garden experiment.

3. Results

We collected 180 plants from 25 *A. gerardi* populations, representing the most geographically and environmentally dispersed WGS sample of *A. gerardi* to date (Fig. 2.1A). The 9*x* and 6*x* cytotypes of *A gerardi* are nearly indistinguishable in the field, but our subsequent evaluation of ploidy identified 9*x* individuals in five populations, concentrated in the western half of the species range (S1). Three of the five populations contained multiple 9*x* and 6*x* genotypes (BOU, KON, AUS; Fig. 2.1A).

3.1 Assembly of a subgenome- and haplotype- resolved reference assembly

Previous analyses of *A gerardi* genetics have relied on reduced representation sequencing approaches and commonly ignored ploidy in genotyping. To facilitate our evolutionary analyses of *A. gerardi*, we assembled a new, haplotype-resolved reference genome for a 6*x* genotype from the center of the species range. Using a combination of 92x coverage Illumina short-reads, 54x OmniC reads, and 85x PacBio HiFi long-reads (Tables S5, S4), we assembled 60 chromosome-scale scaffolds, with 30 scaffolds per haplotype (see Methods; Tables S8, S9). The three subgenomes were identified by clustering the chromosome-scale scaffolds into groups of 10 based on shared enriched 12-mer content. Subgenome 'A' was identified as most closely related to *Anatherum virginicum* (formerly *Andropogon virginicus*, Vorontsova *et al.* 2023; Nagahama and Norrmann 2012), a suspected progenitor species (Estep *et al.* 2014), at it had the highest number of shared 18-mers. The reference was annotated with 89,426 gene models in the haploid genome and approximately 70% of the genome is repetitive elements.

3.2 Assessment of the impact of habitat fragmentation on population structure and genetic diversity

Widespread land-use change has resulted in severe fragmentation and loss of North American prairies, where less than 1% of tallgrass prairies are estimated to remain in some regions (Sam-

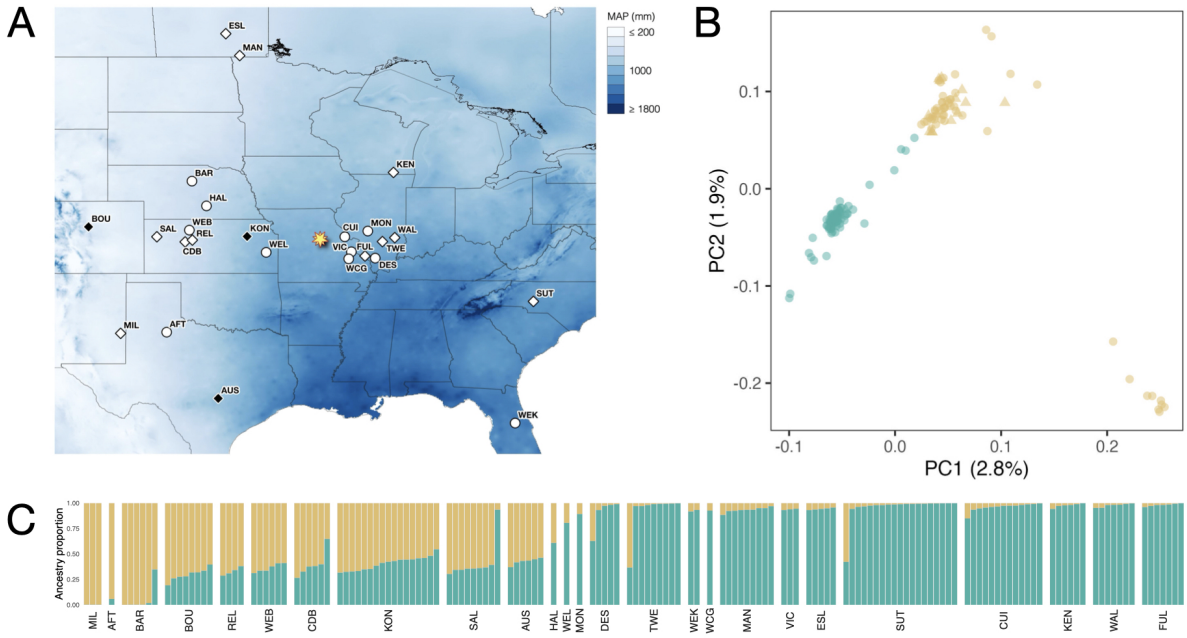

Figure 1.1: **Population structure is limited among sampled *A. gerardi* populations. (A)** Geographic distribution of sampled populations relative to the common garden location (yellow star) across a gradient of mean annual precipitation (MAP). Populations where only 6*x* genotypes were sampled are depicted in white and populations with at least one 9*x* genotype and 6*x* sampled genotypes are shown in black. Diamond-shaped points indicate populations that were included in the common garden. Population codes refer to those in Table S1. **(B)** The first two PCs of a PCA on SNPs for all genotypes. Each point represents a genotype where circles are 6*x* and triangles are 9*x*. The color of each point is the majority ancestry estimated by **(C)** STRUCTURE when K = 2, where tan is the West admixture group and teal is the East.

son and Knopf 1994). To assess the impact of habitat loss on *A. gerardi* population structure and genetic diversity, we sequenced the sampled *A. gerardi* plants to high ($>$ 20x) or low ($<$ 5x) coverage and aligned reads to our reference assembly, resulting in nearly 12 million SNPs after quality control. We used a principal component analysis (PCA) and STRUCTURE, a genetic-clustering algorithm (Pritchard *et al.* 2000). In the principal component analysis (PCA), the first two principal components (PC1 and PC2) separated the sampled genotypes into two broadly discernible geographic groups, referred to here as East and West (Fig. 2.1B). These two clusters were also resolved by STRUCTURE, where K = 2 is the best-supported model (Fig. 2.1C, S3).

Notably, within the West genetic group, a third cluster of populations was distinguished in

the PCA (Fig. 2.1B). The populations in this cluster (MIL, BAR, AFT) had high relatedness and elevated inbreeding. BAR had a significantly higher average inbreeding coefficient than other tested populations (F_{BAR} = 0.28±0.01, p < 0.05; $\bar{F}$ = 0.18, Fig. S6) and lower per-base genetic diversity (Fig. S5). MIL did not have significantly higher inbreeding or lower nucleotide diversity but kinship was elevated (Fig. S4). AFT also shows high within-population kinship but individual inbreeding coefficient could not be estimated for this population as it was only sequenced to low-coverage.

While genetic data allows the identification of these broad clusters, the PCs explain very little of the total genetic variation (2.8% and 1.9%, respectively). Indeed, genetic differentiation between the East and West genetic groups was very low with a F_{ST} of 0.023 (SD = 0.040) indicating that most genetic diversity is shared between genetic groups; metrics specifically designed to partition genetic variation in polyploid taxa give similar results (ρ = 0.043, SD = 0.068). Overall, genetic diversity within the 6*x* cytotype, the base ploidy level, is high ($\bar{\theta}_P$ per population = 0.007, Fig. S5).

3.3 Origins of mixed-ploidy populations

To understand whether recurrent polyploidization or asexual reproduction is contributing to the maintenance of the 9*x* cytotype, we evaluated the relatedness of genotypes and populations within and between cytotypes to determine the number of origins of the 9*x* cytotype. If asexual reproduction is the predominant mechanism, 9*x* genotypes would have the highest relatedness to other 9*x* and high genetic differentiation from genotypes that are 6*x*. Alternatively, if the 9*x* cytotype arose multiple times by recurrent polyploidization, analysis of genetic relatedness would detect genetic groups composed of 9*x* and their 6*x* progenitors. We tested these hypotheses by examining the relatedness of genotypes and populations in the West genetic group as it contained the majority of sampled 9*x* genotypes (Fig. S2).

Grouping genotypes by population and cytotype, we first measured pairwise genetic differentiation between all pairs with ρ, an alternative genetic differentiation statistic comparable to F_{ST} which was developed to overcome bias introduced by ploidy (Ronfort *et al.* 1998; Fig. S7).

Differentiation between and within cytotypes was low overall ($\bar{\rho} = 0.068$, SD = 0.0088). Within-cytotype differentiation was significantly higher for the $9x$ ($\bar{\rho} = 0.096 \pm 0.006$) than the $6x$ cytotype ($\bar{\rho} = 0.060 \pm 0.0017$; $p < 0.0001$), although the difference between mean genetic differentiation is small. Additionally, genetic differentiation between the $6x$ and $9x$ cytotypes ($\bar{\rho} = 0.075 \pm 0.002$) was higher than genetic differentiation among $9x$ ($p < 0.0001$) and lower than divergence among $6x$ ($p < 0.0001$). Differentiation estimated with F_{ST} was consistent with these results (S7). The low overall genetic differentiation, as well as the small but significantly higher differentiation among $9x$, is consistent with the multiple origins hypothesis. Further, the lack of differentiation between $9x$ and $6x$ genetic groups rejects the single-origin hypothesis.

Locally, individuals of the $9x$ cytotype may share a single origin and are maintained by either clonal growth or sexual reproduction, though the latter is expected to be quite rare due to meiotic errors (Norrmann *et al.* 1997). We calculated kinship (F_{ij}) between all $6x$ and $9x$ genotypes in the west genetic group following (VanRaden 2008), which was found to be the best estimator for polyploid populations (Amadeu *et al.* 2020; Bilton *et al.* 2024). We hypothesized that a single local origin should lead to elevated kinship among $9x$ individuals. Consistent with our finding among populations, the majority of sampled genotypes had very low relatedness ($\bar{F}_{ij}$ = 0.0078, SD = 0.021; Fig. 1.2, S4). The average kinship among $9x$ was 0.017 (SD = 0.037) indicating the $9x$ are not clonal stands. The maximum relatedness among $9x$ was 0.18, approximately the expected relatedness of a half-sibling or grandparent-grandchild relationship (F_{ij} = 0.125). Together, these findings suggest the $9x$ cytotype has multiple origins at both the local and regional scale. Mixed-ploidy in *A. gerardi* populations is a product of recurrent polyploidization.

3.4 The effect of polyploidy on growth and reproductive effort

Although recurrent polyploidization may help explain the high overall abundance of $9x$ cytotypes, it does not explain the observed association of the $9x$ cytotype with high temperature variability and low mean annual precipitation (McAllister *et al.* 2015). Instead, we hypothesized that selective differences in survival or growth could explain this pattern. To test this hypothesis, we evaluated

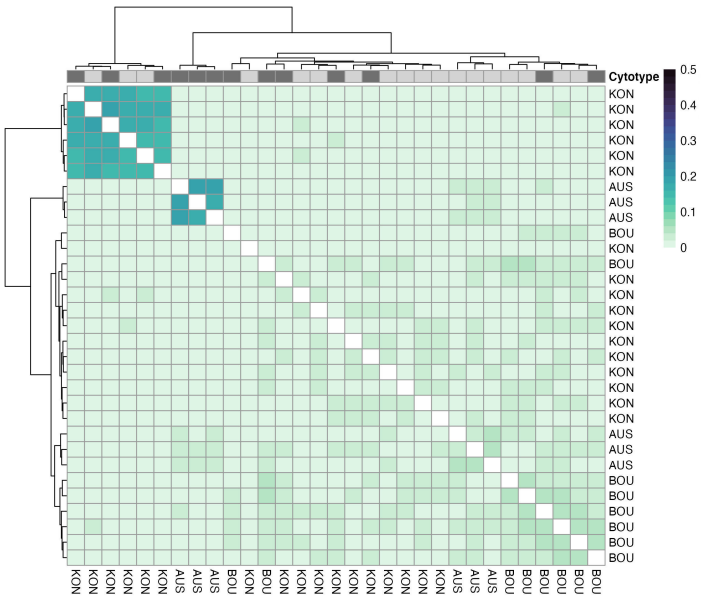

Figure 1.2: **Kinship is low among genotypes in mixed-ploidy populations.** Kinship among 31 genotypes from the mixed-ploidy populations is plotted where genotypes are labeled with the population they were sampled from (AUS, BOU, or KON). Genotypes are hierarchically clustered by Euclidean distance in kinship values and have the same order on both axes. Columns are annotated with the cytotype of each genotype where 9x are dark gray and 6x are light gray. No data is plotted for the diagonal as estimates of self-relatedness are unreliable with low-coverage data. This kinship matrix is a subset of a larger kinship matrix between all genotypes in West genetic group (Fig. S4).

15 ecologically relevant phenotypes in genotypes from 14 populations over two years in a common garden (Fig. 2.1A, S12). Of the 14 populations planted in the common garden, three populations contained multiple 9x and 6x genotypes (BOU, KON, AUS).

We first assessed the effect of ploidy on fitness. We measured differences in reproductive effort (number of tillers and percent of tillers that are flowering) and growth between the two measurement years (change in aboveground biomass, height, basal area, and tillers). Using linear mixed models that control for relatedness among genotypes (Model 1), we found population of origin significantly affects change in aboveground biomass (χ^2 = 12.16, p = 0.00049) but no other measures of reproductive effort or growth (Table S2). Individual plant genotype significantly affected all reproductive effort and growth traits, but not change in plant height and basal area. Notably, we found ploidy significantly affected change in aboveground biomass ($\hat{\delta}$ = 1.02, 95% CI $[0.422, 1.61]$, p = 0.0008). Genotypes that are 9x grew on average 3.25 times (95% CI $[2.58, 3.92]$) larger, which is

approximately 45% more growth than the 6x ($\hat{\mu}$ = 2.23, 95% CI $[1.81, 2.66]$). As all genotypes were sequenced and replicated within the common garden, we could estimate narrow-sense heritability (h^2) and genetic correlations (r_g) among traits to evaluate the contribution of genetic variation to trait variation and phenotypic trade-offs. We found change in aboveground biomass was positively genetically correlated with plant height ($r_{g,Y1}$ = 0.16, $r_{g,Y2}$ = 0.19) and leaf dry matter content (r_g = 0.28, Fig. 1.3B), which is a measure of investment in individual leaves, mechanical resistance, and tissue density (Wright *et al.* 2004), but we did not find a significant effect of ploidy on either trait.

We regressed the population mean change in aboveground biomass against the geographic distance each population was transferred to the common garden (i.e. transfer distance) to assess whether growth rate variation can be attributed to local adaptation. We found change in aboveground biomass steeply declines with increasing transfer distance (r^2 = -0.69, p = 0.0064, Fig. 1.3A). We further broke down transfer distance as the difference in MAP and MAT. We found a significant negative correlation between growth and MAT (r^2 = -0.61, p = 0.020; Fig. S10) and MAP transfer distance (r^2 = -0.82, p = 0.00034). Populations from climates with a lower MAP than the common garden site performed the best while populations with a higher home MAP performed the worst. Additionally, growth declined with an increasing difference in MAT. These results support previous findings of local adaptation in *A. gerardi* (Galliart *et al.* 2019).

Given ploidy can increase growth, it may interact with local adaptation to enable *A. gerardi* populations to withstand greater local environmental variation and stress. To further understand this relationship, we use linear mixed models that include the home climate of each population as a fixed effect (Model 6.13). The home climate was described by PC1 and PC3 from a PCA run on all ClimateNA variables (Wang *et al.* 2016) averaged across 1961 to 1990, which explained 91.7% and 2.7% of variation respectively (see Methods). PC1 was associated with variation in growing degree days where the highest loadings of PC1 were growing degree days below 18°C and above 5 and 10°C with loadings of -0.66, 0.46, and 0.33, respectively (Fig. S9). PC3 was positively associated with mean annual precipitation (MAP) and negatively associated with Hargreaves climatic

moisture deficit (CMD) with loadings of 0.68 and -0.61. In this model, we continued to find a significant effect of ploidy on change in aboveground biomass ($\hat{\delta}$ = 0.79, 95% CI $[0.274, 1.31]$, p = 0.0027) and found a change in aboveground biomass was positively associated with both PC1 ($\hat{\delta}$ = 0.00021, SE = .000052, p = 0.00) and PC3 ($\hat{\delta}$ = 0.0013, SE = .00033, p = 0.00). Overall, these results demonstrate ploidy has a significant effect on fitness and suggest a role for ploidy in local adaptation of *A. gerardi* populations.

3.5 The effect of polyploidy on leaf morphology and economics

At face value, higher fitness of the 9*x* cytotype might suggest the 9*x* cytotype would outcompete the 6*x* cytotype. Rather, the co-dominance of cytotypes might be explained by ploidy-inducing ecological trade-offs. Previous work has identified ecotypic variation in *A. gerardi* leaf anatomy, plant height, and physiology (Olsen *et al.* 2013; Caudle *et al.* 2014; Galliart *et al.* 2020; Bachle and Nippert 2021), but these studies only examined variation across a limited range and precipitation cline or failed to consider ploidal variation. Using linear mixed models, we find the sampled *A. gerardi* populations differ significantly in a number of ecophysiological and morphological traits including in mean stomata length, stomatal density, stomatal pore index (SPI), leaf thickness, leaf length, leaf width, and plant height (Table S2). Nonetheless, trait variation was continuous across populations, and a PCA of best linear unbiased predictions (BLUPs) estimated for each genotype and all phenotypes could not identify clusters that would indicate the presence of ecotypes or morphogroups (Fig. S8). Further, the 90% confidence ellipses for cytotypes or for West and East genetic groups substantially overlap, indicating little differences in overall morphology (Fig. S8). The h^2 we estimated for the measured traits further supports the dominant role of the environment, rather than genetics, in generating ecotypes (Fig. 1.3B, S13).

Even though we found no overall morphological patterns, leaf-level traits are significantly genetically correlated (Fig. 1.3B). Specific leaf area, leaf thickness, leaf length, and leaf width are significantly genetically correlated in year 1 (Fig. 1.3B). The direction of the relationships is consistent in year 2, although not significantly different from zero. We found environmental PC1

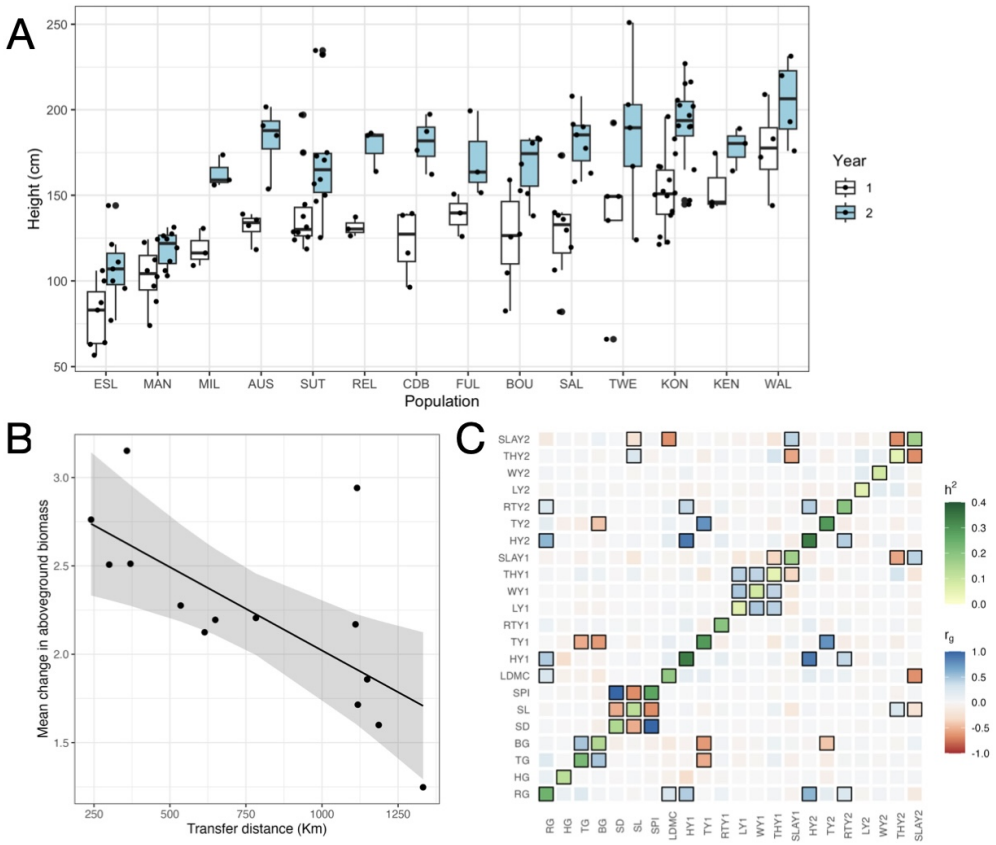

Figure 1.3: **Phenotypic diversity is continuous across *A. gerardi* populations.** (A) Plant height is plotted for each population in each year by mean plant height. The mean plant height for each genotype is overlaid on the boxplots. Population codes refer to those in Figure 2.1A. (B) Population mean change in aboveground biomass, predicted as a BLUP, declines with increasing geographic transfer distance. The gray area indicates the 95% confidence interval of the predicted values. (C) The genetic correlation (r_g) between all measured phenotypes is plotted with narrow-sense heritability (h^2) of each trait on the diagonal. Genetic correlations were estimated independently for traits measured in year 1 (Y1) and year 2 (Y2). Axis labels refer to abbreviations for the measured phenotypes: change in aboveground biomass (RG), growth in height (HG), growth in number of tillers (TG), growth in basal area (BG), stomatal density (SD), stomata length (SL), stomatal pore index (SPI), leaf dry matter content (LDMC), height (H), number of tillers (T), percent of tillers flowering (RT), specific leaf area (SLA), and leaf length (L), width (W) and thickness (TH). Tiles with a black border have a 95% credible interval that does not cross zero.

explains a significant proportion of variation in leaf traits and PC3 significantly explains variation in all leaf traits except leaf thickness (Table S3). Populations at the northern edge of the range (ESL, MAN, Fig. 2.1A), which have the lowest MAT and GDD, have the smallest and thinnest leaves while populations at the southern edge of the range (AUS, MIL) in driest and hottest climates have the thickest and largest leaves. Ploidy only had a significant effect on leaf thickness, but the effect size was within our measurement error (Table S2).

Stomatal trait variation was also significantly associated with climate (Table S3). Stomatal density and stomata length were negatively genetically correlated (r_g = -0.53; Fig. 1.3B), which is the expected relationship (Taylor *et al.* 2012). Environmental PC1 explained a significant proportion of variation of all stomatal traits whereas PC3 significantly explained variation in stomatal density and stomatal pore index (Table S3). Stomatal density increased with PC3 ($\hat{\delta}$ = 0.040, SE = 0.023, p = 0.00; Fig. 1.4A) and decreased with PC1 ($\hat{\delta}$ = 0.0032, SE = 0.0036, p = 0.00). Accordingly, stomata length increased with PC1 ($\hat{\delta}$ = 0.00022, SE = 0.00039, p = 0.00). As a result, populations in regions with lower MAP and higher MAT had the largest and fewest stomata (Fig. 1.4). We found 9x genotypes had significantly larger stomata ($\hat{\delta}$ = -4.3 μm, 95% CI $[-2.38, -4.395]$, p = < 0.0001), lower stomatal density (40 μm^{-2}, 95% CI $[19.4, 60.7]$, p = 0.0001, Fig. 1.4C), and a lower stomatal pore index ($\hat{\delta}$ = 8.63, 95% CI $[4.28, 13]$, p = 0.0001) compared to 6x genotypes (Table S2). This relationship is consistent across the abaxial leaf surface (Fig. S14). When considering home climate in our model, we found the estimate of the effect of ploidy is similar (Table S3).

4. Discussion

Polyploidy is ubiquitous across the green tree of life, yet our understanding of how polyploids establish and adapt to novel environments is limited. Mixed ploidy species provide an opportunity to study the natural conditions in which WGD is beneficial, or maladaptive. We investigated these questions in *A. gerardi*, a mixed-ploidy species and the ecologically dominant bunchgrass in North American prairies.

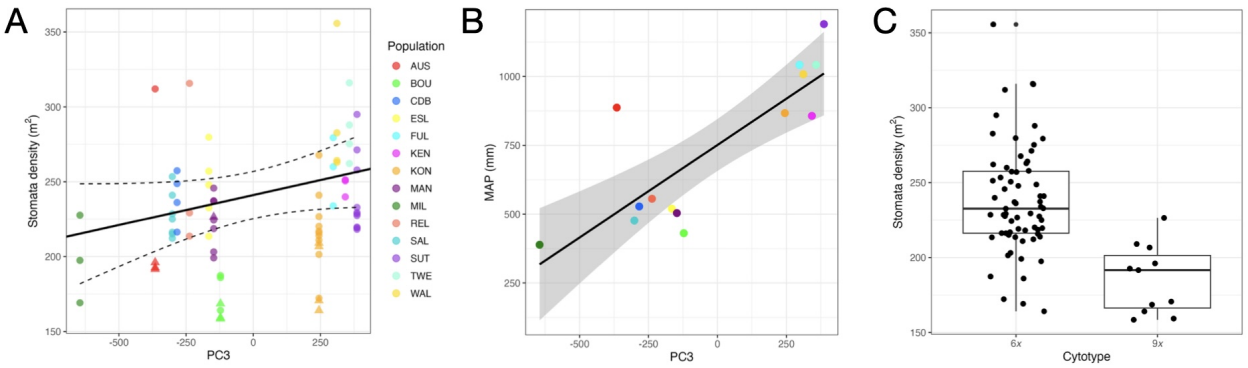

Figure 1.4: **Stomatal density decreases with increasing ploidy and mean annual precipitation.** (A) Stomatal density increases with PC3. The solid line is the effect of PC3 on stomata density estimated with Model 6.13 and the dashed lines are two standard errors from the mean. The average stomatal density is plotted for each genotype where triangles are 9*x* and circles are 6*x* genotypes. (B) Mean annual precipitation (MAP) is positively correlated with PC3. The gray area indicates the 95% confidence interval of the linear regression of MAP and PC3. Values for each population are overlaid. Points in A and B are colored by population where the population codes refer to those in Figure 2.1A. (C) The two cytotypes significantly differ in mean stomata length. The mean stomatal density of each genotype is overlaid on top of the boxplots as black points.

4.1 Minimal population structure and high genetic diversity in *Andropogon gerardi* despite modern habitat fragmentation

Using the first WGS dataset generated for *A. gerardi* populations, we found *A. gerardi* population structure and genetic diversity do not reflect the drastic loss and fragmentation of North American prairies. Broadly, the populations are structured into two genetic groups, with a longitudinal divide into East and West groups. The two groups are highly admixed and the majority of genetic diversity is shared between the groups (Fig. 2.1). These large genetic groups are similar to those described in previous studies using genotype-by-sequencing (GBS, McAllister and Miller 2016; Galliart *et al.* 2020) and amplified fragment length polymorphism (ALFP) markers (Gray *et al.* 2014). Nonetheless, our estimates of genetic differentiation between genetic groups are lower than previous reports, likely due to the larger number of markers and reduced bias of whole genome sequencing.

Genetic diversity in *A. gerardi* populations is high (Fig. S5) and comparable to diverse crops

like maize (Chen *et al.* 2022). We found only three populations had signatures of inbreeding or low N_e; genotypes belonging to AFT, BAR, and MIL have high relatedness across populations (Fig. S4) even though they are geographically disconnected where BAR is separated approximately 1000 km from both AFT and MIL (Fig. 2.1A). Although these sites were selected as they were suspected to be remnant, unseeded prairies, the high relatedness among these populations may suggest the BAR population was seeded or contaminated with foreign seed from the southwestern United States. High relatedness within BAR is consistent with recent inbreeding, as evidenced by significantly higher inbreeding coefficients and lower nucleotide diversity (Fig. S6, S5). High relatedness within MIL is not associated with inbreeding and lower diversity suggesting the population has a low N_e but no recent inbreeding.

A. gerardi populations exhibit local adaptation, with significant variation in growth and morphology associated with climate of origin. For a population to be considered locally adapted, the population must perform best in its home environment and outcompete foreign genotypes at home (Kawecki and Ebert 2004). We found growth declines with geographic transfer distance (Fig. 1.3A), which is consistent with local adaptation. The presence of local adaptation and lack of population structure suggests local selective pressures are strong enough to overcome gene flow or recent shared ancestry. The local selective pressures include MAT and MAP (Fig. S10). The populations from the highest latitude and lowest MAT (Fig. 2.1A; ESL and MAN) had the shortest stature and thin and small leaves. Plant height and leaf thickness increase with both MAT and MAP. These patterns are consistent with previous studies documenting precipitation is a major driver of trait (Olsen *et al.* 2013; Caudle *et al.* 2014; Galliart *et al.* 2019) and genetic variation across *A. gerardi* populations (Avolio *et al.* 2013; Galliart *et al.* 2019). Trait variation along a temperature or latitudinal cline was less explored prior to our study; McMillan (1959) found phenology and plant height was associated with daylength, which is negatively correlated with latitude and positively correlated MAT in North America. Further, Bachle and Nippert (2021) found the best models of leaf anatomical variation included both precipitation and temperature.

Previous researchers have defined *A. gerardi* genetic and phenotypic variation in two or three

ecotypes that follow the longitudinal precipitation cline across the center of the United States (Fig. 2.1A; Galliart *et al.* 2019; Gray *et al.* 2014). Using the definition of ecotype from Lowry (2012), ecotypes are defined as, "groups of populations, which are distinguished by a composite of variation in many traits and allele frequencies across loci over space". We find that *A. gerardi* populations cannot be delineated into ecotypes due to limited genetic differentiation and continuous phenotypic variation across both a latitudinal and longitudinal cline. Rather, the stark trait contrasts previously described may have resulted from sampling bias along a precipitation cline, measurement of traits outside a common garden experiment, and bias in sequencing markers.

4.2 Mixed-ploidy is maintained by recurrent polyploidization

We found the coexistence of multiple cytotypes in *A. gerardi* is primarily supported by recurrent WGD events regenerating the 9*x* cytotype. Our results agree with previous findings by McAllister and Miller (2016), who described a minimum of three origins of the 9*x* cytotype. Recurrent polyploidization may result from continual or environmental-induced bursts of unreduced gamete production in the 6*x* cytotype (Ramsey and Schemske 1998). Unreduced gametes have been estimated to form at a rate of less than 1% in autopolyploid plants (Ramsey and Schemske 1998; Harlan and deWet 1975). This rate may increase under temperature and water stress (Wang *et al.* 2017; Sax 1936; Belling 1925), which may also contribute to 9*x* being more common in regions with higher temperature variability.

Gene flow between cytotypes is likely rare and has a minimal contribution to genetic diversity in mixed-ploidy populations due to multiple reproductive barriers faced by the 9*x* cytotypes. The 9*x* have irregular meiosis which produces unbalanced gametes leading to reduced seed viability compared to the 6*x* (Norrmann *et al.* 1997; Tompkins *et al.* 2015). Although the probability of 9*x* meiosis producing a balanced 1*n* gamete across 30 chromosomes is extremely low, pollen is produced in abundance, individual plants produce as many as 180 flowering tillers in a year, and the generation time is estimated at 50 to 100 years (Keeler 2004). Thus while gene flow among 9*x* individuals or between cytotypes is not impossible, it likely contributes little to observed diversity.

Indeed, the offspring of 9*x* that successfully germinate are typically aneuploids (Norrmann *et al.* 1997) and are found in less than 5% frequency in the field (Keeler 1992; McAllister *et al.* 2015).

Recurrent polyploidization, the limited reproductive success of the 9*x* cytotype, and low occurrence of intermediate and aneuploid cytotypes suggest the existing genotypes that are 9*x* may be first-generation polyploids, referred to as neopolyploids. Individual 9*x* plants persist in *A. gerardi* populations due to vegetative growth through an underground rhizome. Neopolyploidy is further supported by the recent loss and local recolonization of the majority of North American prairies due to the severe 'Dust Bowl' drought lasting from 1932 to 1938 (Schubert *et al.* 2004). Through the drought, *A. gerardi* persisted in substantially smaller stands as the majority of the prairie was converted to C3 grass species (Weaver and Albertson 1943; Knapp *et al.* 2020). In prairies that experienced the greatest decline, there was up to a 60% loss in *A. gerardi* abundance (Weaver and Albertson 1943). The plants that persisted had a shortened stature and root depth greater than 6 ft, in order to reach the lowered water table (Weaver and Albertson 1943). It took as long as 20 years for C4 grasses to return to dominance, suggesting many of the sampled populations have undergone a recent population expansion (Weaver 1968). Further, we see the 9*x* and 6*x* cytotypes are intermixed within populations and do not form large clonal stands (Keeler 1992) supporting recent polyploidization as the origins of mixed-ploidy, rather than recolonization via asexual reproduction.

4.3 The consequences of neopolyploidy in adaptation

In *A. gerardi*, we have shown 9*x* individuals are neopolyploids and polyploidization confers changes in growth and physiology that are likely adaptive in arid conditions. Together, these results suggest the abundance of 9*x* in climates with low MAP and increased temperature variability (McAllister *et al.* 2015) may be largely driven by ecological adaptations due entirely to the impacts of WGD.

We found the 9*x* cytotype has a significantly higher change in aboveground biomass than the 6*x* cytotype. This trait is a measure of both relative growth rate and fitness in bunchgrasses (Keeler and Davis 1999; Aspinwall *et al.* 2013; Lovell *et al.* 2021) as most of the yearly reproduction is

asexual, via vegetative propagation, rather than from seed (Benson and Hartnett 2006). We found the change in aboveground biomass could not be attributed to a difference in lateral growth (change in basal area), tiller density (change in number of tillers), or plant height (Table S2). Rather, change in aboveground biomass was positively genetically correlated with leaf dry matter content and plant height suggesting a greater investment in tissue density and leaf mechnaical resistance (Wright *et al.* 2004, Fig. 1.3B). The change in tissue density may be attributed to altered composition of the cell wall (Corneillie *et al.* 2019) or an increased number of organelles (Fernandes Gyorfy *et al.* 2021), which may affect plant metabolism. Increased tissue density may provide greater structural support enabling changes in plant height. Lastly, the increased growth of 9*x* may be explained by the diversion of resources from seed development to the growth of non-reproductive tissues as the 9*x* have lower seed viability (Tompkins *et al.* 2015). Regardless of the home climate, we found the 9*x* cytotype always has higher growth than the 6*x* cytotype (Table S2, S3).

We also detected a change in stomata size, where the 9*x* have larger but fewer stomata than the 6*x*, suggesting an increase in some cell types, although not all cell types may be affected due to developmental and function constraint (Snodgrass *et al.* 2017). We found the 6*x* have smaller, more dense stomata suggesting they have higher maximum stomatal conductance than the 9*x* cytotype. As the 9*x* cytotype is more common in regions with lower MAP, a lower stomatal conductance would be beneficial to decrease water loss. Further, we see that *A. gerardi* populations from climates with lower MAP, when measured in a shared environment and controlling for population structure, have the largest and lowest density of stomata. Increased stomata size in polyploids compared to their diploid progenitors or lower ploidy levels has been well documented in both synthetic and natural polyploids (Beaulieu *et al.* 2008). Stomata size and stomatal density are indicative of the stomatal pore area on the leaf, which, in combination with the physiological process of stomatal closure, regulate the amount of CO_2 that can enter the leaf for photosynthesis and control water loss via transpiration (i.e. stomatal conductance). In grasses, higher maximum stomatal conductance is generally associated with smaller stomata length, higher stomatal density, and wet habitats (high MAP) (Taylor *et al.* 2012). This relationship is consistent with a previous study in

the closely related sand big bluestem (*A. hallii* Vitman), a grass adapted to sandy soils with low soil moisture (Awada *et al.* 2002).

The effect of ploidy on stomata size, as an approximation for cell size, is consistent with the change in growth we detected. Increasing genome size is expected to so slow the cell cycle (Francis *et al.* 2008) and therefore slow growth rate, but this is can compensated for by making fewer, larger cells (Doyle and Coate 2019). The strategy of making fewer, larger cells was shown to be adaptive for geophytes, plants with underground storage organs, in environments with shorter growing seasons or high seasonality where fast development is needed (Veselỳ *et al.* 2012). The correlation of 9*x* abundance with increased temperature variability (McAllister *et al.* 2015) is consistent with larger cells and a slow cell cycle being an adaptive trait.

Studies evaluating the effect of ploidy on relative growth rate and water use efficiency in multiple environments, as well as research into population variation in unreduced gamete formation, are needed. For example, nitrogen and phosphorus requirements have been shown to increase with ploidy due to increased material costs from synthesizing more DNA and phospholipid bilayer (Walczyk and Hersch-Green 2023; Roddy *et al.* 2020). Additionally, our estimation of stomata traits may be limited by only looking at abaxial stomata as adaxial stomata have been shown to balance the overall maximum stomatal conductance limitations from abaxial stomata (Muir *et al.* 2023). However, the majority of *A. gerardi* are hypostomatous (stomata only present on the abaxial side) and stomata size is similar on both sides of the leaf (Varvel *et al.* 2018; Knapp *et al.* 1994).

5. Conclusion

Polyploidy may provide an adaptive advantage to novel climates. Natural mixed-ploidy species are advantageous systems for studying role of polyploidy in adaptation as they are not confounded by the chemical induction of WGD. However, considering the age of polyploidization in mixed-ploidy species is necessary to separate the effects of WGD from evolution post-polyploidization. Here, we studied the coexistence of 6*x* and 9*x* cytotypes in *A. gerardi*, the dominant species in endangered North American tallgrass prairies. We found the 9*x* cytotype is continuously created by recurrent

polyploidization, where each 9*x* individual is likely a recent, novel WGD event. Asexual reproduction extends the lifespan of 9*x* individuals, creating overlap across WGD events and increasing the abundance of the 9*x* cytotype. Our results support theoretical models that suggest perennially and clonality enables cytotype coexistence (Van Drunen and Friedman 2022) and we are the first to empirically demonstrate the role of recurrent polyploidization in maintaining multiple cytotypes at high frequency.

We also find WGD confers changes to growth and physiology that are likely adaptive to arid climates. These adaptive changes explain the abundance of the 9*x* cytotype in regions with low MAP and high temperature variability (McAllister *et al.* 2015). Further, these adaptations may allow the 9*x* cytotype to locally outcompete 6*x* and lower population adaptive potential due to low rates of seed viability (Norrmann *et al.* 1997; Tompkins *et al.* 2015). We see evidence of 9*x* dominance in some populations being reported as 100% 9*x* (McAllister *et al.* 2015). A 9*x* population could persist with asexual reproduction and very low rates of sexual reproduction, but adaptation will be limited. This is concerning given phenotypic species distribution models predict *A. gerardi* biomass will decrease as much as 60% by 2070 (Smith *et al.* 2017). As a result, ploidy may be a conservation concern in *A. gerardi* populations, as initially proposed by Tompkins *et al.* (2015) in North Carolina and South Carolina populations. Intraspecific variation in ploidy has previously been proposed as a conservation concern and suggested to be considered in seed selection (Kramer *et al.* 2018) and populations prioritization (Wickell *et al.* 2024, but see Almeida and Santos Leal 2024). Our results underscore the need to consider polyploidy in conservation and restoration and emphasize the importance of understanding the role of polyploidy in adaptation.

6. Methods

6.1 Sample collection

A. gerardi plants were either collected as rhizomes or grown from seed (Supplemental Data). Necessary permissions and permits were obtained before collecting. Plants were sampled from 29 sites

in the United States and Canada (Fig. 2.1A) and plants were brought back to the United States from Canada under phytosanitary certificate #3193417. Nearby sites were considered a single site in analyses resulting in a total of 25 sampled populations. Plants from rhizomes were collected following methods described in Phillips *et al.* (2023). Briefly, the plants were dug up with a shovel late in the growing season in 2016 through 2020. Any soil was washed off, the leaves were cut back to about 4 in in height to reduce transpiration, and the rhizomes were wrapped in wet paper towels for transportation to the Donald Danforth Plant Science Center in St. Louis, MO, USA. Plants grown from seed were grown from population bulk seed collected by Loretta Johnson at Kansas State University, KS, USA. The plants were potted in and maintained in greenhouses with average conditions of 8°C day, 22°C night, 50% relative humidity (RH), and a 16 hrs daylength. Once mature, the plants were maintained on outdoor benches year-round. The leaves were cut back and the pots were covered with straw mulch each winter. Voucher specimens were created for each population and have been deposited at the Missouri Botanical Garden (St. Louis, MO, USA.).

6.2 Short-read sequencing of population panel

A set of 148 genotypes were processed for low-coverage sequencing at Cornell University (Supplemental Data). DNA was extracted using approximately 100 mg of lyophilized leaf tissue and a DNeasy Plant Kit (Qiagen Inc., Germantown, MD). High throughput Illumina Nextera libraries were constructed and samples were sequenced with other plant samples in pools of 96 individuals in one lane of an S4 flowcell in an Illumina NovaSeq 6000 System with paired-end 150-bp reads, providing approximately 1.8X coverage for each sample.

Sequencing was re-attempted at the University of California, Davis (UCD) for 15 genotypes, which failed initial sequencing attempts (Supplemental Data). For resequencing, young leaves were re-collected from plants maintained in a greenhouse at UCD, fixed in liquid nitrogen, and stored in a -80°C freezer until use. DNA was again extracted with a DNeasy Plant kit (Qiagen Inc., Germantown, MD) from approximately 10 mm^2 of leaf tissue. Illumina Nextera libraries

were constructed using an epMotion® 5073 (Eppendorf, Hamburg, Germany) following the Nextera Lite protocol (Rowan *et al.* 2019). The samples were pooled samples were run in one lane of a S4 flowcell in an Illumina NovaSeq 6000 System with paired-end 150-bp reads, providing approximately 2.5X coverage for each sample.

An additional subset of 50 genotypes were whole genome sequenced to high coverage by the Department of Energy Joint Genome Institute.

Samples sequenced at Cornell University were part of a larger sequencing across Andropogoneae species. To assess the quality of the short-read sequencing from Cornell and detect possible contamination from other species, reads from each sample were aligned to Sorghum reference genome (NCBI GenBank ID GCF_000003195.3) using `bwa mem` (Li 2013), as the *A. gerardi* reference genome was still in assembly. Alignment statistics were collected including the fraction of mapped reads, duplication rate, the fraction of bases of the whole genome and of the coding sequence portion covered at depths of >0X, >1X, and >5X, and the fraction of reads mapping with various numbers of mismatches (0, 5, 10, 15) were also reported. The Kraken pipeline (Wood and Salzberg 2014) was used to quantify contamination for each sample with sequences originating from bacteria, the human genome, and plants outside of the Poaceae family. To further confirm the taxonomy of the analyzed sequences, a custom database of five plastid genes (*matK*, *ndhF*, *rbcL*, *rpoB* and *rpoC1*) was constructed from 4755 plant plastid genomes downloaded from NCBI RefSeq (Supplemental Data). All these genomes have species-level taxonomy information. Nucleotide sequences of the five genes were extracted from the plastid genomes of each species based on NCBI Refseq gene annotation. For genomes without gene annotations, TBLASTN was used to identify the coordinates of these genes within the respective genomes. For each sample, `bwa aln` (Li and Durbin 2009) was then used to map 1000 randomly selected reads to the plastid genes database. Up to five such genes with the most reported hits were selected, and the Phylum, Class, Order, Family, Subfamily, Tribe, Genus, and Species of these genes were reported. While the categories Phylum through Tribe were typically as expected (Streptophyta, Magnoliopsida, Poales, Poaceae, Panicoideae, Andropogoneae, respectively), Genus was often ambiguous between

Schizachyrium and *Andropogon* due to the species' allopolyploidy. Samples that were identified as anything other than *Schizachyrium* or *Andropogon* were discarded.

6.3 Genome sequencing for the *A. gerardi* reference genome

We sequenced *A. gerardi* (var. Kellogg-1272) using a whole genome shotgun sequencing strategy and standard sequencing protocols. Sequencing reads were collected using Illumina and PacBio platforms. Illumina and PacBio reads were sequenced at the Department of Energy (DOE) Joint Genome Institute (JGI) in Berkeley, California and the HudsonAlpha Institute in Huntsville, Alabama. Illumina reads were sequenced using the Illumina NovoSeq6000 platform, and the PacBio reads were sequenced using the SEQUEL II platform. One 400 bp insert 2x250 Illumina fragment library (92.80x coverage) was sequenced along with one 2x150 OmniC library (54.38x; Table S5). Prior to assembly, Illumina fragment reads were screened for PhiX contamination. Reads composed of >95% simple sequence were removed. Illumina reads <50 bp after trimming for adapter and quality (q<20) were removed. The final read set consists of 1,601,302,817 reads for a total of 92.80x of high-quality Illumina bases. For the PacBio sequencing, the total circular consensus sequencing (CCS) sequence yield consisted of 16,785,606 reads (average size 19,680 bp) that produced 343.73 Gbp (85.93x; Table S4).

6.4 Genome size estimation

We attempted to estimate genome size with flow cytometry for all sampled individuals. Genome size could not be estimated for all individuals as some plants died prior to estimation. Flow cytometry methods were previously described in Phillips *et al.* (2023). Briefly, maize B73 inbred line (5.16 pg/2C) was used as an internal standard. Three replicates were prepared and analyzed separately for each individual. The cell count, coefficient of variation of FL2-A, and mean FL2-A were recorded for the target and reference sample with no gating. The three replicates were averaged to calculate the genome size (Supplemental Data).

Of the individuals for which genome size couldn't be estimated with flow cytometry, sixteen

had sufficient sequencing coverage for ploidy to be determined using nQuire (Weiß *et al.* 2018). Thirty high-coverage genotypes with genome sizes successfully estimated using flow cytometry were included in the nQuire analysis as a control. Of the 30 control genotypes, 29 were 6*x* and one was 9*x*. nQuire utilizes a Gaussian Mixture Model to model the read frequency histogram expected for diploids, triploids, and tetraploids. As all three subgenomes are resolved in the *A. gerardi* genome described below, we expect the 6*x* genotypes to have a diploid distribution and the 9*x* genotypes to have a triploid distribution. Maximized log-likelihoods were estimated for each genotype under the diploid, triploid, and tetraploid models and normalized to the maximized log-likelihood of the data modeled under a free model. The ploidy model with the highest normalized log-likelihood was assigned the ploidy of the genotype. To assess noise and error in this method, the normalized maximized log-likelihoods for each of the three ploidy models were plotted against each other in R (v4.2.2, R Core Team 2017; Figure S1).

Of the remaining individuals for which genome size could not be estimated with flow cytometry or sequenced-based approaches, ploidy could be inferred for one population based on a previous study. The population CUI (Cuivre River State Park, MO, USA) was previously sampled for cytotypic composition by McAllister *et al.* (2015) and found to be 100% 6*x*. Given this study is relatively recent, we assumed all individuals sampled from CUI are 6*x*.

6.5 Genome assembly and construction of pseudomolecule chromosome

A total of 16,785,606 PacBio CCS reads (85.93x) were assembled using HiFiAsm+HIC assembler v15.1, Cheng *et al.* 2021) and subsequently polished using the 1,601,302,817 Illumina fragment 2x250 reads (92.80x) were used to resolve homozygous SNP/indel errors in the consensus with RANCON (v1.4.10; Vaser *et al.* 2017). This produced initial assemblies of both haplotypes. The haplotype 1 (HAP1) assembly consisted of 1,064 scaffolds (1,064 contigs), with a contig N50 of 55.6 Mbp, and a total genome size of 2,780.5 Mbp (Table S6). The haplotype 2 (HAP2) assembly consisted of 731 scaffolds (731 contigs), with a contig N50 of 59.9 Mbp, and a total genome size of 2,701.5 Mbp (Table S7).

Hi-C Illumina reads (54.38x) from *A. gerardi* (var. Kellogg-1272) were separately aligned to the HAP1 and HAP2 contig sets with Juicer (v1.8.9, Durand *et al.* 2016), and chromosome-scale scaffolding was performed with 3D-DNA (v180922, Dudchenko *et al.* 2017). No misjoins were identified in either the HAP1 or HAP2 assemblies. The contigs were then oriented, ordered, and joined together into 30 chromosomes per haplotype using the Hi-C data. A total of 38 joins were applied to the HAP1 assembly, and 38 joins for the HAP2 assembly. Each chromosome join is padded with 10,000 Ns. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the production assembly.

Scaffolds that were not anchored in a chromosome were classified into bins depending on sequence content. Contamination was identified using blastn against the NCBI non-redundant nucleotide collection (NR/NT) and blastx using a set of known microbial proteins. Additional scaffolds were classified in HAP1 as repetitive (>95% masked with 24-mers that occur more than 4 times in the chromosomes; 786 scaffolds, 87.8 Mb), redundant (unanchored scaffolds composed of ≥95% 24-mers >2x in all scaffolds; 2 scaffolds, 31.1 kb), mitochondria (177 scaffolds, 10.5 Mb), and prokaryote (13 scaffolds, 727.0 kb). Scaffolds were also classified as repetitive (>95% masked with 24-mers that occur more than 4 times in the chromosomes; 554 scaffolds, 66.8 Mb), redundant (unanchored scaffolds composed of ≥95% 24-mers >2x in all scaffolds; 5 scaffolds, 115.9 kb), and mitochondria (83 scaffolds, 5.1 Mb).

After forming the chromosomes, it was observed that some small (<20kb) redundant sequences were present on adjacent contig ends within chromosomes. To resolve this issue, adjacent contig ends were aligned to one another using BLAT (v35, Kent 2002), and duplicate sequences were collapsed to close the gap between them. A total of 2 adjacent contig pairs were collapsed in the HAP1 assembly and 6 in the HAP2 assembly. The three subgenomes were then clustered into groups of 10 chromosomes based on shared enriched 12-mer content. For each triplet of chromosomes, all 12-mers were identified from a frequency of 20 – 5000, with a minimum of 50 occurrences being required on one of the triplets. A positively enriched 12-mer had one of

the three with at least 3 times the count of the other two. All occurrences of the enriched 12-mers were counted across the chromosomes and converted to a binary call. Binary k-mer calls were subset to sites observed in $\geq$ 5 chromosomes and used to construct a symmetric binary distance matrix. Clustering was then accomplished on the distance matrix by partitioning around medoids (PAM) with the R package cluster (v2.1.4, Maechler *et al.* 2012). The 3 subgenomes were designated as A, B, and C. The "A" subgenome was identified as being most closely related to *A. virginicus*. This was determined using HipMer (Georganas *et al.* 2015) assembly of *A. virginicus* to mask the 3 subgenomes using 18-mers. The A-genome consistently shared more content with *A. virginicus* than the other two subgenomes. Chromosomes were numbered and oriented within the 3 subgenomes using *Sorghum bicolor*, and the resulting sequence was screened for retained vector and contaminants (Tables S8, S9).

Heterozygous SNP/indel phasing errors were corrected using the 85.93x CCS data. A total of 1,903 heterozygous SNPs/indels were corrected in both haplotypes. Homozygous SNPs and indels were corrected in the tremula and alba releases using ~62X of Illumina reads (2x150, 400 bp insert) by aligning the reads using `bwa mem` (v0.7.17-r1188, Li 2013) and identifying homozygous SNPs and indels with the GATK's UnifiedGenotyper tool (v3.6-0-g89b7209, McKenna *et al.* 2010). A total of 987 homozygous SNPs and 10,691 homozygous indels were corrected in the HAP1 release, while a total of 882 homozygous SNPs and 9,487 homozygous indels were corrected in the HAP2 release. The final version 1.0 HAP1 release contained 2,669.3 Mbp of sequence, consisting of 88 contigs with a contig N50 of 63.1 Mbp and a total of 99.90% of assembled bases in chromosomes. The final version 1.0 HAP2 release contained 2,588.3 Mbp of sequence, consisting of 68 contigs with a contig N50 of 59.2 Mbp and a total of 99.93% of assembled bases in chromosomes.

Completeness of the euchromatic portion of the version 1.0 assemblies was assessed using existing RNASeq reads (library JLJB). The aim of this analysis is to obtain a measure of completeness of the assembly, rather than a comprehensive examination of gene space. The transcripts were aligned to the assembly using `bwa mem` (v0.7.17-r1188, Li 2013) and the screened alignments

indicate that 99.880% of the RNASeq reads aligned to the HAP1 release, and 99.883% aligned to the HAP2 release.

6.6 Genome annotation

Transcript assemblies were made for each haplotype from about 1.2 billion pairs of 2x150 stranded paired-end Illumina RNA-seq reads using PERTRAN, which conducts genome-guided transcriptome short read assembly via GSNAP (Wu and Nacu 2010) and builds splice alignment graphs after alignment validation, realignment, and correction. Approximately 14.8 million PacBio Iso-Seq CCS reads were corrected and collapsed by a genome-guided correction pipeline, which aligns CCS reads to the respective haplotype with GMAP (Wu and Nacu 2010) and corrects introns for small indels in splice junctions when all introns are the same or 95% overlap for single exon, to obtain about 774,000 and 766,000 putative full-length transcripts for HAP1 and HAP2, respectively.

Subsequently, 829,257 (HAP1) and 820,009 (HAP2) transcript assemblies were constructed using PASA (Haas *et al.* 2003) from the RNA-seq transcript assemblies and the respective haplotype. Loci were determined by transcript assembly alignments, EXONERATE alignments, and Swiss-Prot proteomes to repeat-soft-masked *A. gerardi* respective genomes using RepeatMasker (Smit *et al.* 2013–2015) with up to 2,000 bp extension on both ends unless extending into another locus on the same strand. EXONERATE alignments used protein sequences from *Arabidopsis thaliana*, *Glycine max*, *Oryza sativa*, *Sorghum bicolor*, *Brachipodium*, *Aquilegia coerulea*, *Solanum lycopersicum*, *Vitis vinifera*, *Panicum hallii*, *Joinvillea ascendens*, *Acorus americanus*, *Paspalum vaginatum*, *Phoenix dactylifera*, *Musa acuminata*, *Ananas comosus*, *Asparagus officinalis*, *Phalaenopsis equestris*. The repeat library consists of *de novo* repeats by RepeatModeler (Smit *et al.* 2008–2015) on *A. gerardi* HAP1 and repeats in RepBase (Bao *et al.* 2015). Gene models were predicted by homology-based predictors, FGENESH+ (Salamov and Solovyev 2000), FGENESH_EST (similar to FGENESH+, but using EST to compute splice site and intron input instead of protein/translated open reading frame, ORFs), EXONERATE (Slater and Birney 2005), PASA assembly ORFs (in-house homology constrained ORF finder) and AUGUSTUS (Stanke

et al. 2006) trained by the high confidence PASA assembly ORFs and with intron hints from short read alignments. The best-scored predictions for each locus were selected using multiple positive factors, including EST and protein support, and one negative factor of overlap with repeats. PASA improved the selected gene predictions by adding untranslated regions, splicing correction, and alternative transcripts.

PASA-improved gene model proteins were subject to protein homology analysis to the above-mentioned proteomes to obtain the C-score, a protein BLASTP score ratio to the mutual best hit BLASTP score, and protein coverage, the highest percentage of protein aligned to the best of homologs, for each transcript. PASA-improved transcripts were selected if their Cscore was $\geq$ 0.5 and protein coverage $\geq$ 0.5, or it had EST coverage, but its CDS overlapping with repeats is $<$ 20%. Gene models with CDS that overlap repeats more than 20% must have a C-score $\geq$ 0.9 and homology coverage $\geq$ 70% to be selected. Additionally, gene models were subject to Pfam analysis (Mistry *et al.* 2021) and gene models with $>$ 30% TE domains were removed. Gene models that were incomplete, had low homology support, and a short single exon ($<$ 300 bp CDS) without a protein domain nor good expression were manually filtered out. Transposable elements were annotated using the Extensive de novo TE Annotator (EDTA; Ou *et al.* 2019).

6.7 Variant calling and genotyping

The quality of the raw short read sequence data was assessed using FastQC (v0.11.6, `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`). Samples re-sequenced at UC Davis were trimmed using fastp to remove adapters, polyG trails, and the first 9 bp on each read while requiring reads to have a minimum length of 36 bp (`fastp -l 36 -Q --trim_front1 9 --trim_front2 9`; v0.20.1, Chen *et al.* 2018). Samples sequenced by JGI and Cornell did not require trimming to improve alignment quality.

The sequence data were aligned to the *A. gerardi* reference genome with bwa-mem2 (v2.2; Vasimuddin *et al.* 2019). Reads from samples that were sequenced on multiple lanes were combined into a single fasta file prior to alignment. The BAM files were sorted using SAMtools (v1.7;

Danecek *et al.* 2021), read groups were added using Picard AddOrReplaceReadGroups, and duplicates were removed with Picard MarkDuplicates (v2.27, `http://broadinstitute.github.io/picard`) using default settings. Alignment quality was assessed using QualiMap (`qualimap bamqc -nt 1000 -nt 12 -nw 400 --skip-duplicated`; v.2.1.1, Okonechnikov *et al.* 2016). Samples with less than 98% of reads mapping were discarded. BAMs from genotypes that were independently sequenced multiple times were merged into a single BAM (`samtools merge`). High coverage samples sequenced by JGI were subsampled to approximately 2-4X coverage (`samtools view -b -s 0.06`). Genotypes with less than 90% reads mapping or 0.5X coverage and missing genome size data were excluded from downstream analyses.

For analyses that included 6x and 9x genotypes, we identified and filtered variable sites using BCFtools (Li 2011), then called single-read genotypes with ANGSD (Korneliussen *et al.* 2014). Variable sites were called using BCFtools mpileup and BCFtools call (v1.16; Li 2011). After identifying variable sites, sites were filtered to exclude multialleic sites and sites with low mapping quality and sequencing quality with GATK (`gatk VariantFiltration -filter "QUAL` $\leq$ `30" -filter "MQ` $\leq$ `30"` and default `gatk SelectVariants --restrict-alleles-to BIALLELIC`; v4.2; Van der Auwera and O'Connor 2020). SNPs were additionally filtered for less than 20% missing data and a minimum genotype depth of 1 using a custom R script. A maximum depth filter was applied in order to exclude sites where paralogs may be mapping (Phillips 2024). We defined the maximum depth cutoff at each site as the 99^{th} percentile assuming coverage follows a Poisson distribution. Single-read genotypes were estimated for the filtered sites directly from the BAMs using ANGSD (`angsd -doIBS 1 -doMajorMinor 3 -doCounts 1` ; v0.934 ; Korneliussen *et al.* 2014). Single-read genotypes are generated by randomly drawing a read at each site. If the read draw has the reference allele, the genotype is "1" while the alternate allele is "0".

6.8 Assessment of population structure and diversity

One-hundred thousand sites were randomly sampled from the single-read genotype matrix for assessment of population structure. The principal component analysis (PCA) was run using ANGSD (`angsd -doCov 1`). The kinship matrix was estimated following (VanRaden 2008) using a custom script in R. The diagonal elements were set to 1 for use in downstream analyses, as an accurate estimate of kinship within an individual cannot be made using a draw of a single read. Downstream analyses produced the same results given a diagonal of 1, 0, or random values. Clones were identified as having a kinship coefficient greater than 0.4; only two hexaploid genotypes were identified as clones. Population structure analyses were run with and without clones and were similar.

Population admixture was assessed by estimating the individual ancestry coefficients and number of genetic clusters (K) using the STRUCTURE admixture model (v2.3.4, Pritchard *et al.* 2000). STRUCTURE was run for a K of 2 through 24 for 3 replicates of 85,000 iterations per model (including a 10,000 burn-in). We specified `PLOIDY` as 1 because the single-read genotypes only sample one haplotype. Convergence was confirmed by consistent results between replicates (Fig. S3).

Population differentiation was estimated with F_{ST} and ρ (Ronfort *et al.* 1998; Meirmans *et al.* 2018). Values were estimated between the East and West genetic groups identified in previous analyses and pairwise between a cytotypes within each population. For the pairwise population analysis, three genotypes were randomly selected from each cytotype-population groups containing at least three genotypes. Then, population allele frequencies were calculated from the single-read genotypes. Subsequently, pairwise F_{ST} was calculated as $F_{ST} = \frac{H_T - H_S}{H_T}$, where H_T is the expected heterozygosity and H_S is the observed heterozygosity within the two populations. Pairwise ρ was calculated as $\rho = \frac{H_T - H_S}{H_T - H_{SP}}$, where H_{SP} is the ploidy-corrected H_S, following Meirmans *et al.* (2018). F_{ST} and ρ were estimated per-site for each pairwise comparison and then averaged. To determine if within 9*x*, within 6*x*, and between cytotype genetic differentiation was statistically different, we estimated the mean F_{ST} and ρ for the three comparison types (6*x*-6*x*, 9*x*-9*x*, 6*x*-9*x*) using the R function `emmeans` and tested whether the means differed using a Tukey pairwise comparison

implemented in `contrast` (v1.1, Searle *et al.* 1980) and a 95% confidence level.

6.9 Estimation of hexaploid genetic diversity

Nucleotide diversity (θ_P), Watterson's theta (θ_W), and Tajima's D were estimated for all 6*x* genotypes in 10,000 and 50,000 bp windows using ANGSD (v0.935, Korneliussen *et al.* 2013). The folded site frequency spectrum (SFS), thetas, and Tajima's D were estimated independently for each population. The two identified clones were excluded from estimation of hexaploid genetic diversity. Individual inbreeding coefficients were estimated only in 6*x* genotypes with high coverage WGS data, as low coverage data can result in significant bias (Bilton *et al.* 2024). The inbreeding coefficients were estimated in parallel for each chromosome using ngsF (v1.2.0, Vieira *et al.* 2013). Results for all analyses were plotted in R using ggplot2 (v3.4; Wickham 2016).

6.10 Common garden experiment

A subset of 85 genotypes from 14 populations were planted in a common garden in Columbia, Missouri at the University of Missouri Genetics Farm in May 2021. The populations and genotypes were selected to maximize diversity in home environment and representation of the 9*x* cytotype. Additionally, the selected genotypes were required to have WGS and flow cytometry data. The selected genotypes were vegetatively propagated by splitting the rhizomes to produce three clonal replicates. Clonal replicates were bulked at the Donald Danforth Plant Science Center and the University of California, Davis greenhouse facilities. The genotypes were planted in a randomized block design with three blocks, placing one random clonal replicate per genotype in each block (Fig. S12). Random positions within the block were modified only when two genotypes from the same population were neighbors.

Prior to planting, the field was covered with landscaping cloth (DeWitt Sunbelt Woven Ground Cover) held down with anchor pins in order to prevent competition from weedy annual grasses and broadleaf weeds. Plants were spaced 4.5 ft (1.37 m) apart in a grid pattern and the landscape cloth extended for at least 4.5 ft (1.37 m) around the edge of the blocks. The three blocks were separated

by 9 ft (2.74 m). The holes cut in the landscaping cloth were cut in an X-shape at least 3 ft wide in each direction to ensure lateral growth was not constrained. After planting, the field was irrigated once to promote establishment and was not irrigated for the rest of the experiment, relying only on rainfall. The field was hand-weeded and broadleaf weeds were sprayed with 2,4-D herbicide early in the spring when the *A. gerardi* plants were small and likely to be outcompeted by weeds.

6.11 Phenotyping leaf functional traits

The field was phenotyped in early September of 2021 and 2022. Six leaves were collected from each plant each year, selecting the youngest fully expanded leaves on 6 different tillers (Perez-Harguindeguy *et al.* 2016; Garnier *et al.* 2001). Leaves were cut at the ligule with scissors, placed in a plastic bag with a damp paper towel, then stored overnight in a 4°C fridge before phenotyping. Within 72 hours of collection, fresh weight (g) and leaf lamina thickness (mm) was measured. Fresh weight was only measured in 2022. Leaf lamina thickness was measured at the base of the leaf using digital calipers, taking care to avoid the midrib. The leaves were scanned for measurement of width, length, and one-sided area using ImageJ (v1.53-1.54; Schneider *et al.* 2012). Leaf length was measured as the length of the leaf sheath from the ligule to the leaf tip and leaf width was measured at the widest part of the leaf. Specific leaf area (SLA), leaf dry matter content (LDMC; 2022 only), and leaf density (ρ_F) were calculated for each leaf (Vile *et al.* 2005; Garnier *et al.* 2001).

After scanning, abaxial leaf impressions were taken on the fresh leaves using dental putty (Zhermack elite HT+ light body fast set) for analysis of stomatal traits. Three impressions were taken per leaf (bottom, middle, top) to capture developmental variation. Results were similar across leaf impressions (Fig. S14). After the dental putty impressions were taken, the fresh leaves were placed in a 7 in manilla envelope and dried at 70°C for at least 72 hours. After the leaves were dried, dry mass was measured (mg).

To measure stomatal traits, a negative of the dental putty impression was taken using clear nail polish and then imaged using a Leica DM1000 microscope and Leica MC170 HD digital camera

at 10X magnification. Using ImageJ, guard cell length was measured along the longest portion of the guard cell for five stomata per impression. Additionally, stomatal density was measured for each impression. As stomatal density is very high in *A. gerardi*, the image was cropped to a smaller area containing at least 10 stomata. In the cropped image, the area and number of stomata were measured (Supplemental Data). Stomatal pore index (SPI) was measured as stomatal density divided by the square root of mean guard cell length. Stomatal traits were measured in both years but were measured incompletely in 2021. As a result, we analyze and present only the 2022 data. Variation between impressions was assessed using a linear model regressing stomatal traits against impression. We tested for differences in mean stomatal traits between each group of impressions using a Tukey test with a 95% confidence interval implemented in `emmeans` and `contrast` (v1.1, Searle *et al.* 1980).

6.12 Phenotyping performance traits

Survival (dead or alive) was recorded each year, although overall mortality was low (11%). Additionally, we measured the number of tillers, percent flowering tillers, plant height, basal area, and above-ground biomass. Plant height was measured as the length of the longest tiller from the ground to the tip of the inflorescence. To estimate basal area, we measured the diameter of the base in two perpendicular directions and then calculated the area as an ellipse (cm^2, Aspinwall *et al.* 2013). After all phenotypes were collected for the year, all above-ground biomass was cut off the plant approximately 4 in above the crown using grass shears. The biomass was placed in a brown paper bag and dried at 37°C until the dry weight stabilized (3 to 5 days). Once dry, the total dry above-ground biomass was measured (g). Root and below-ground traits were not measured as they require destructive sampling. Basal growth, tiller growth, and relative growth were calculated as the difference between year 2 and year 1 measurements to account for the varying sizes of the plants when transplanting.

6.13 Trait data analysis

The effect of ploidy and the environment on trait variation was specified using two independent linear mixed models for each trait using the R package Sommer (v4.3.3, Giovanny 2016). First, the effect of ploidy on a given phenotype (Y) was tested with the following model, where ploidy was specified as a fixed effect and population (z) and genotype (g) were random effects. Covariance among genotypes was specified with a kinship matrix ($\mathbf{K}$). Year (t) was included as a random effect if the phenotype was measured in multiple years.

$$Y = \mu + \beta_P P + g + t + z + \varepsilon \tag{1}$$

$$g \sim MVN(0, V_A \mathbf{K})$$

$$t, z, \varepsilon \sim N(0, \sigma^2)$$

To test the effect of a population's home environment on trait variation, yearly climate data for each population and the common garden site was extracted from ClimateNA using ClimateNAr (v1.1, Wang *et al.* 2016). Climate variables were averaged from 1961 to 1990. Average yearly growing degree days at 10°C (GDD) were separately estimated using daymetr (v1.7, Hufkens *et al.* 2018) for the period of 1980 to 2000. A PCA was run on the climate data to describe the environmental distance between populations. After examining the correlation between the principal components (PCs) and the raw climate data, PC1 and PC3 were selected to represent the home environment of each model. PC1 (E_1) and PC3 (E_3) were specified as fixed effects and added to the previously described model:

$$Y = \mu + \beta_P P + \beta_{E_1} E_1 + \beta_{E_3} E_3 + g + t + z + \varepsilon \tag{2}$$

The residuals of each model were qualitatively assessed for normality, homogeneity of variances, and independence. Transformations were applied where needed; a log, exponential, and

inverse normal transformation were applied to basal growth, percent of flowering tillers, and SLA, respectively. The significance of fixed effects was tested with an ANOVA. If ploidy was significant, we used `emmeans` and `contrast` (v1.1, Searle *et al.* 1980) to estimate cytotype means and test if the the cytotype means were significantly different. Log likelihood ratio tests were used to test the significance of random effects in the model using a confidence level of 95%. Significance was evaluated for all tests using a 95% adjusted for multiple tests by Bonferroni corrections where the corrected alpha is 0.0033.

The genetic correlation and heritability of the phenotypes were estimated with MegaLMM (v0.1.0 ,Runcie *et al.* 2021) using Model 1 without year in the model. Rather, we treated traits measured in multiple years as separate traits. We specified 20 latent factors and used the default priors. We extracted the posterior means for lambda, genetic variance, and genetic covariances from a single model run for 1000 iterations after a burn-in of 500 iterations. We also estimated the 95% credible interval of the posterior distributions for the genetic covariances and heritabilities.

Change in aboveground biomass was regressed against geographic and climate transfer distance to assess location adaptation. Using Sommer, best linear unbiased predictors (BLUPs) were first estimated for each population using Model 1. The grand mean was added to the estimated BLUPs to improve interpretation. Geographic transfer distance was estimated as the Haversine distance between the population and common garden using `geodist` (v0.0.8, Padgham 2021). Climate distance was measured by the difference in average mean annual precipitation (MAP) and mean annual temperature (MAT) between the common garden environment and home environment. Average MAP and MAT was estimated for each population by averaging the ClimateNA data described above. Daily precipitation and temperature data for the common garden in 2021 and 2022 was downloaded for the Columbia-Jefferson Farm and Gardens (Boone County, MO, USA) weather station from the Missouri Historical Agricultural Weather Database. The average daily temperature was averaged across years to estimate the common garden MAT. Total daily precipitation was summed for each year then averaged to estimate the common garden MAP. Finally, the three measure of transfer distance (geographic, MAP, and MAT) were regressed against population

BLUPS with `lm` in R.

6.14 Data availability

The short-read WGS data is available on NCBI Sequence Read Archive (SRA) under BioProject PRJNA1109389. Supplementary data including genotype metadata, raw phenotype data, and flow cytometry data is available on Dryad at `https://doi.org/10.5061/dryad.gxd2547v1`. The *A. gerardi* reference genome is available on Phytozome under genome IDs 784 and 783. All scripts for genotype calling, population genetic analyses, and trait data analysis can be found at `https://github.com/phillipsar2/andro_snakemake`.

7. Acknowledgements

This project was funded by the National Science Foundation (NSF) grant numbers 1822330 and 1934384, the Davis Botanical Society, and the Botanical Society of America. JRI also acknowledges funding from USDA Hatch project CA-D-PLS-2066-H 548. We thank Christine McAllister, Michael McKain, and Loretta Johnson for providing germplasm for whole genome sequencing and the common garden experiment. We would also like to thank the agencies and people that provided permits and aided collections: Matt McCaw and the City of Austin Water Quality Protection Lands (WQPL), North Carolina Department of Agriculture and Consumer Services, NC Plant Conservation Program, the Suther family and Rev. Dennis Testerman, Chris Matson and Florida Park Service and Department of Environmental Protection, Kyle Dillard and the Milnesand Prairie Preserve (Creamer Ranch), City of Boulder Open Space and Mountain Parks, Lynn Riedel, Brian Anacker, Bess Bookout, Tim Teetaert and the Manitoba Tallgrass Prairie Preserve, Robert D. Bradley, Ken McCarty and Cuiver River State Park, Malissa Briggler and Victoria Glades Conservation Area. We thank the Genomics Facility (RRID: SCR_021727) of the Biotechnology Resource Center of Cornell Institute of Biotechnology and the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786–01, for their help with sequencing experiments. Short-read sequencing was par-

tially carried out at the The Cornell Institute of Biotechnology completed the DNA extraction and library preparation for the population panel samples. We would like to thank Christopher Browne for assistance in planting and maintaining the common garden.

8. Supporting Information

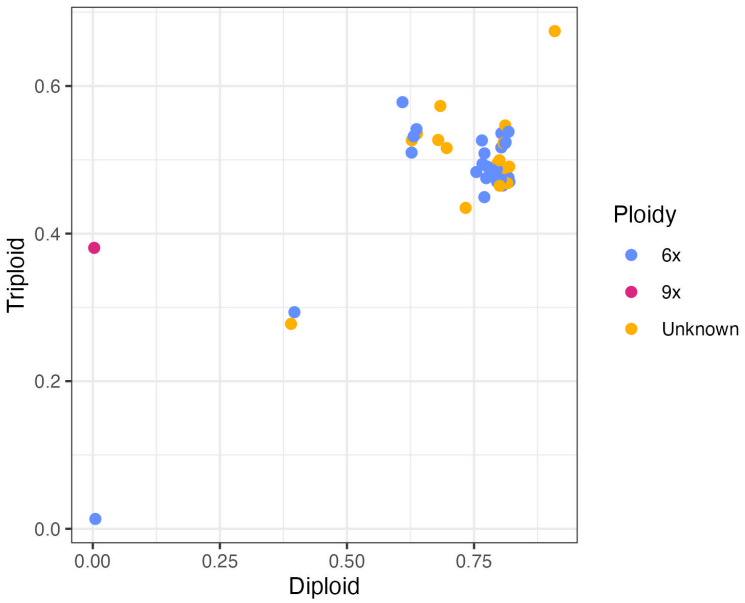

Figure S1: **The normalized maximized log-likelihood of the diploid and triploid nQuire models for genotypes sequenced with high coverage.** The color of the points represents whether the ploidy is unknown or known via flow cytometry. Of the genotypes for which ploidy was known, all were 6*x* except for one genotype. Tested unknown genotypes had similar values to known 6*x* genotypes except for two outliers. In the upper left and lower right corners are genotypes with unusually high (45X) and low (16X) coverage, respectively.

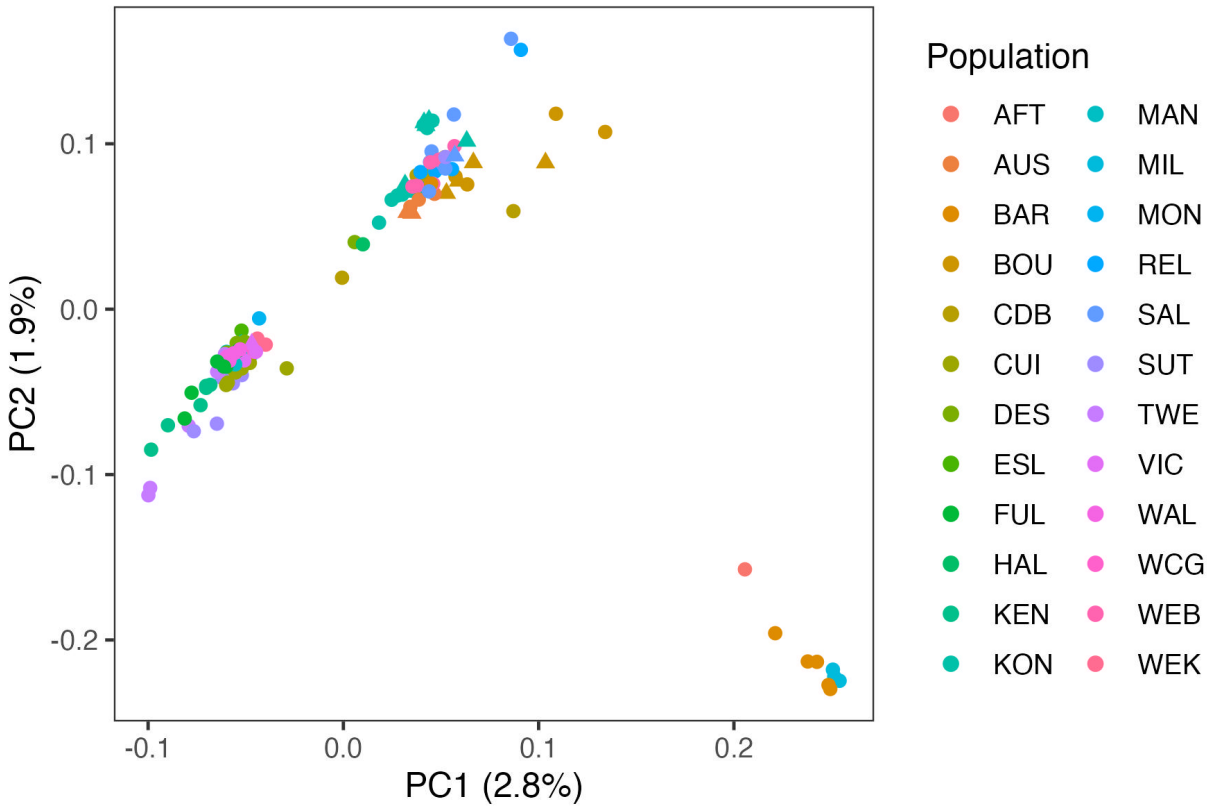

Figure S2: **Principal component analysis of single read genotypes for all sequenced genotypes.** The first two principal components are plotted for each genotype with the color of each point indicating the population of origin. The shape of each point indicates the ploidy of the sample where 6*x* are circles and 9*x* are triangles.

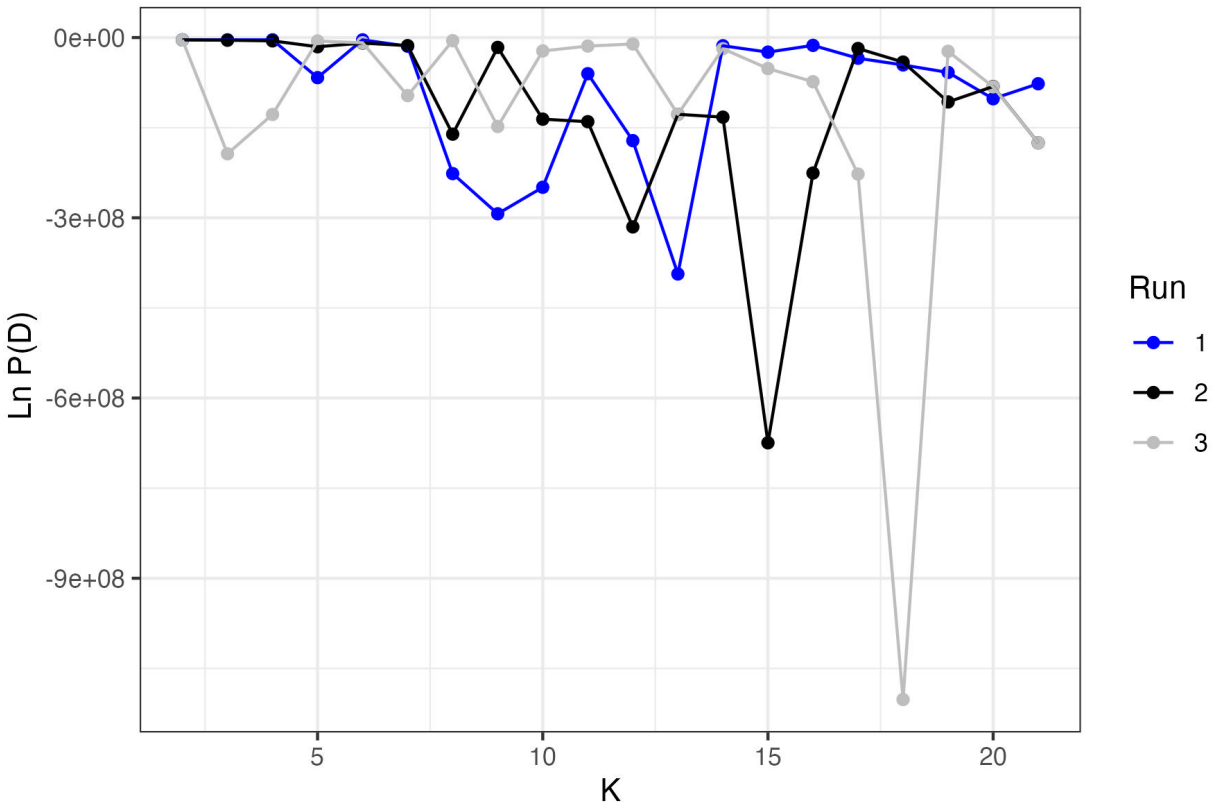

Figure S3: **The estimated log probability of the sequence data given K admixture groups as estimated by STRUCTURE.** Three runs were estimated for each K. The estimated probability and consistency across runs declines with an increasing value of K.

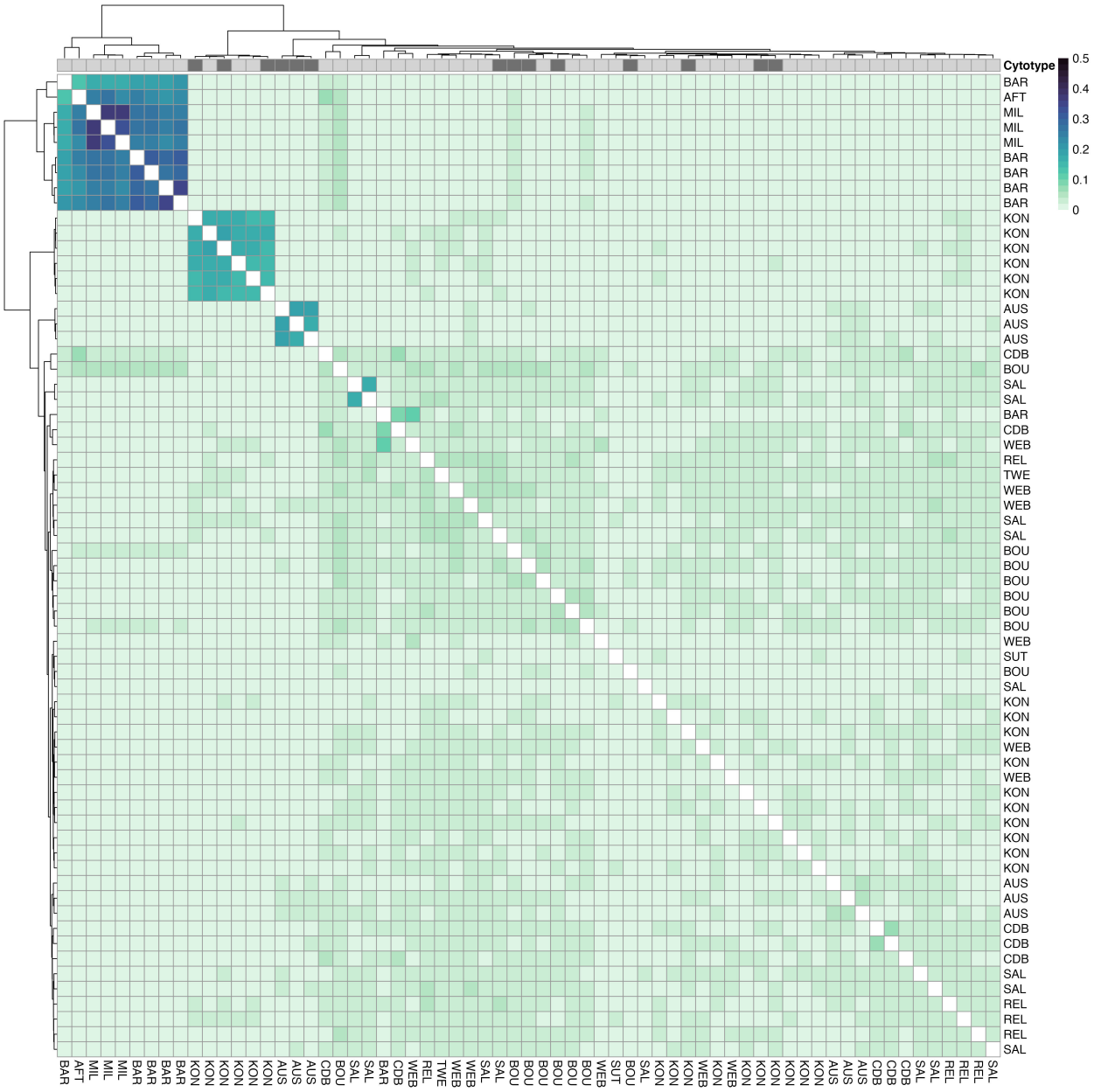

Figure S4: **Kinship of all sequenced genotypes in the West genetic group.** Columns are annotated with the cytotype of each genotype where 9x are dark gray and 6x are light gray. Genotypes are hierarchically clustered by Euclidean distance in kinship values and have the same order on both axis. Genotypes are labeled with their population code which follows Figure 2.1A. No data is plotted for the diagonal as estimates of self-relatedness are unreliable with low-coverage data.

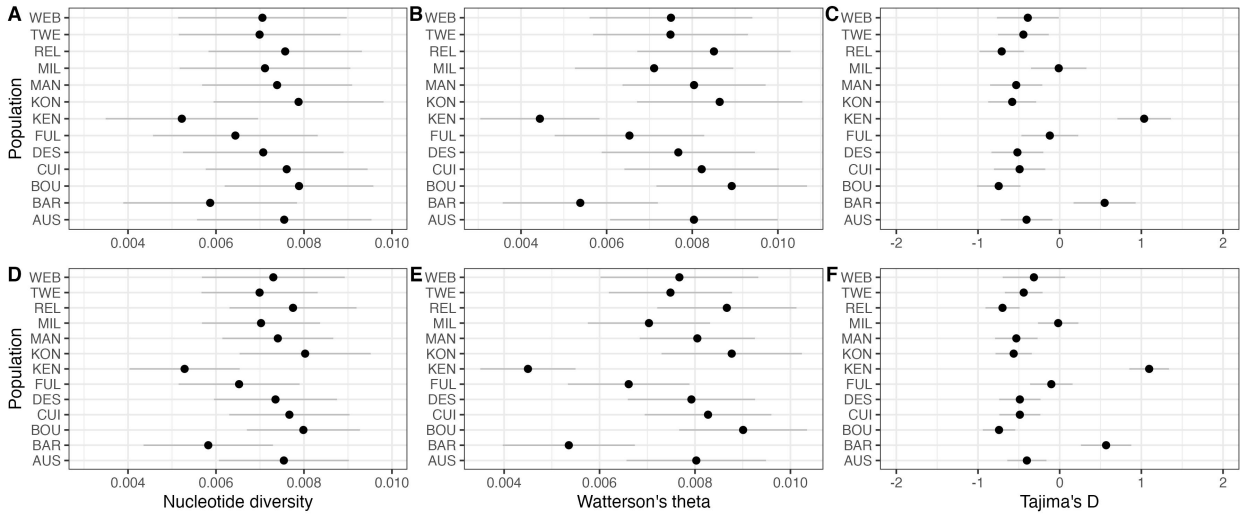

Figure S5: **Population estimates of hexaploid genetic diversity.** Nucleotide diversity (A,D), Watterson's theta (B,E), and Tajima's D (C, F) were estimated in 10 kbp (A, B, C) and 50 kbp (D, E, F) windows. The mean values for each population are shown as a black dot with an error bar indicating one standard deviation from the mean.

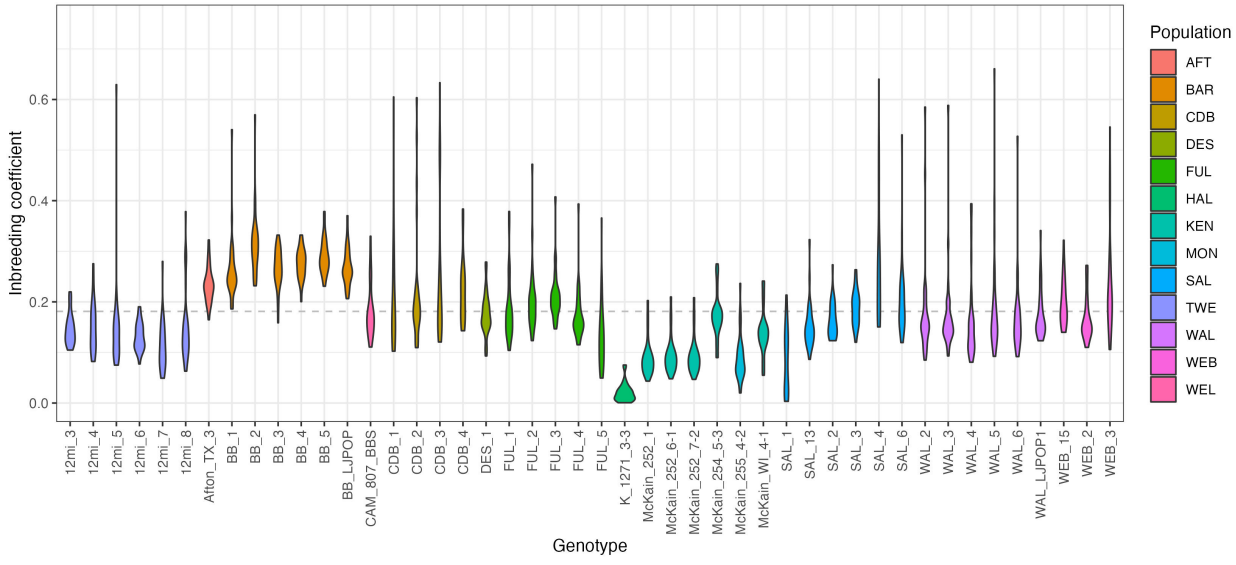

Figure S6: **Individual inbreeding coefficient estimated for hexaploid genotypes with high coverage WGS data.** The violin plots show the distribution of the inbreeding coefficients estimated for each chromosome (n = 30) per genotype and are colored by population. The gray dashed line is the average inbreeding coefficient across these genotypes.

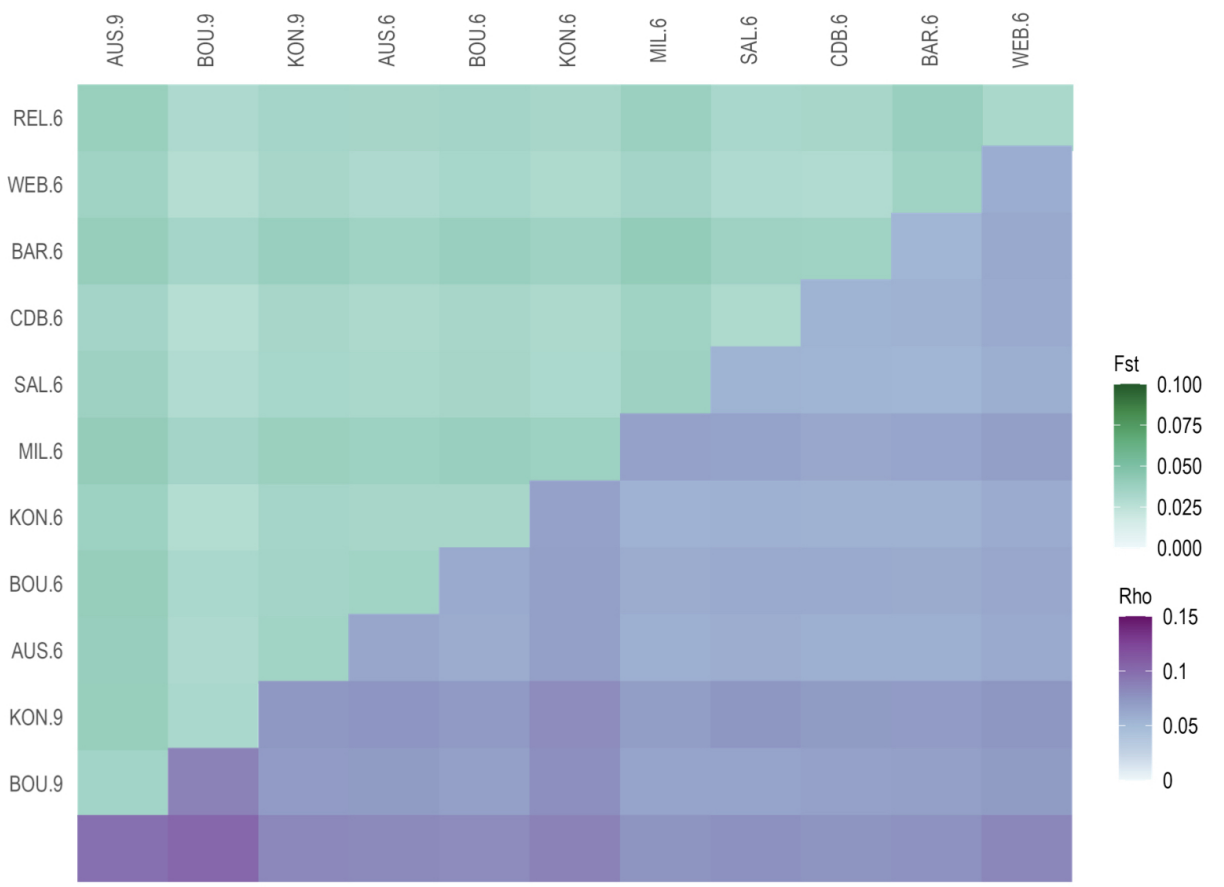

Figure S7: **Pairwise F_{ST} and ρ estimated between 6*x* and 9*x* population pairs in the West genetic group.** The upper diagonal is pairwise F_{ST} (shades of green) and the lower panel is pairwise ρ (shades of purple). F_{ST} is dependent on ploidy level and is expected to be elevated in comparisons among 9*x* compared to comparisons among 6*x* (Ronfort *et al.* 1998).

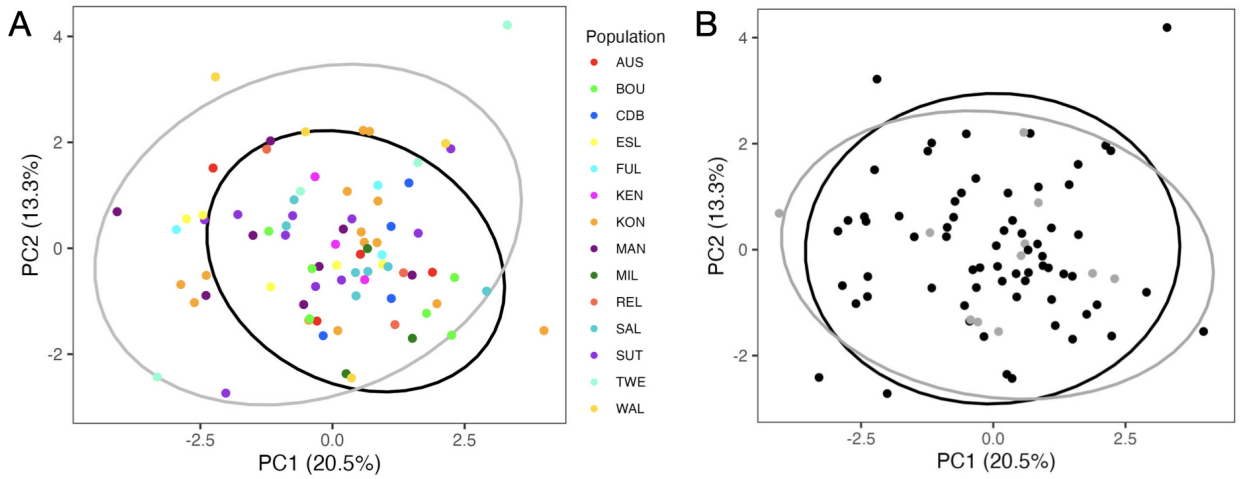

Figure S8: **PCA on best linear unbiased predictors of common garden phenotypes.** The first two principal components are plotted for each genotype overlain with the vectors for each standardized phenotype and 90% confidence ellipses. **(A)** Points are colored by population and the confidence ellipses enclose the East (black) and West (gray) genetic groups. **(B)** Points and ellipses are colored by ploidy where the 6*x* cytotype is black and 9*x* cytotype is gray.

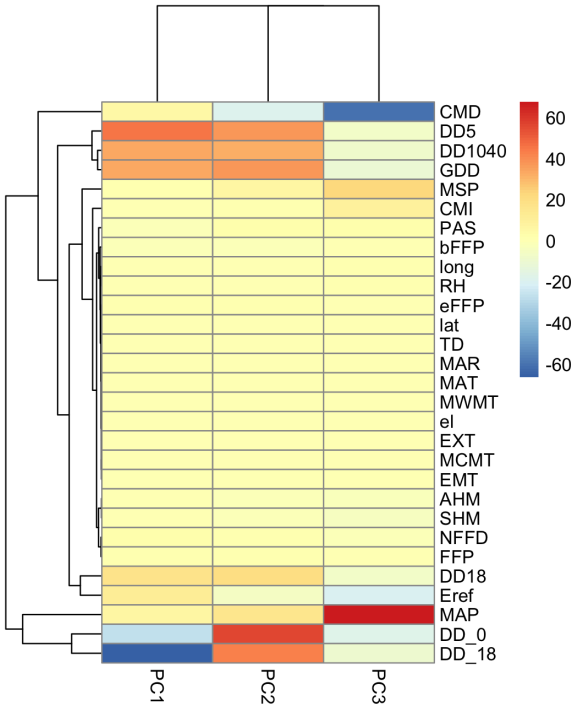

Figure S9: **Principal component loadings of the PCA on climate variables.** The principal component loading is plotted for each comparison between principal component (PCA) and climate averages. All variables are from ClimateNA except GDD, which is the growing degree days at 10°C estimated with daymetr.

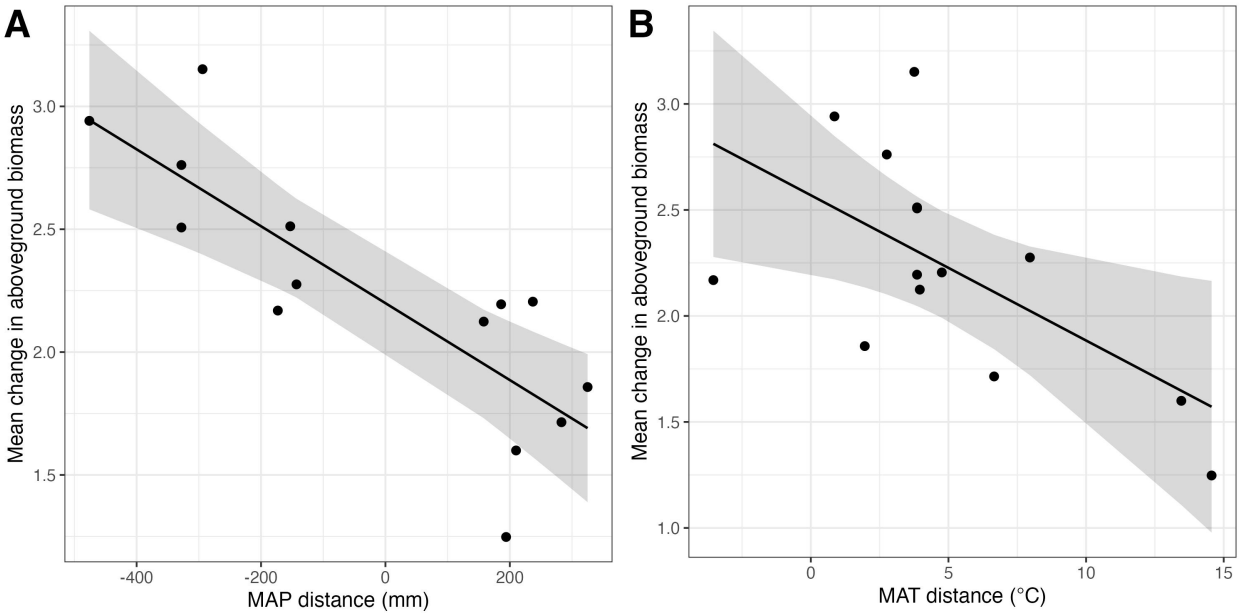

Figure S10: **Population mean change in aboveground biomass regressed against climate transfer distance.** Population means are best linear unbiased predictors (BLUPs) estimated using linear mixed models (see Methods). Predicted values (black line) are plotted with 95% confidence intervals in gray and population BLUPs overlaid as black dots. Climate transfer distance was estimate as **(A)** The difference in mean annual precipitation (MAP) and **(B)** mean annual temperature (MAT).

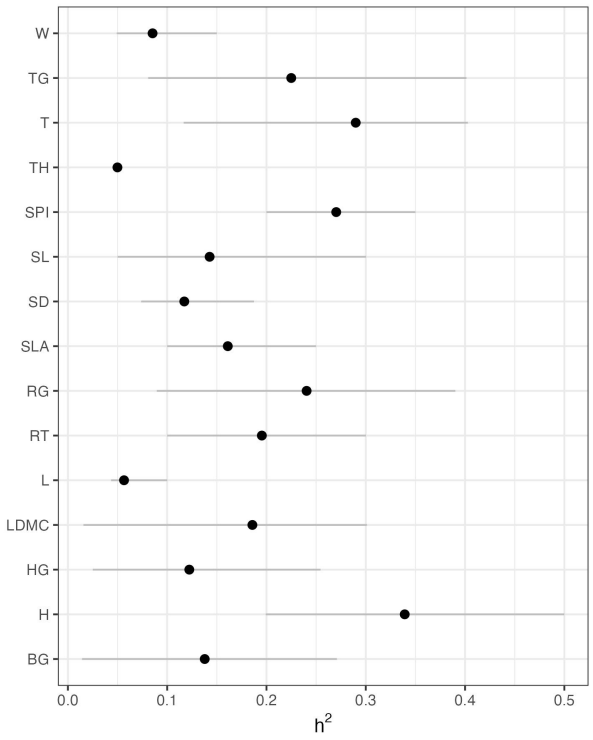

Figure S11: **Posterior mean narrow sense heritability for all measured phenotypes.** The black dot indicates the posterior mean and the gray bar 95% credible interval. See Figure 1.3B for trait codes.

Figure S12: **The common garden prior to phenotyping in 2022.** The field runs north to south and the image shows the field looking north. The common garden was located in Columbia, Missouri.

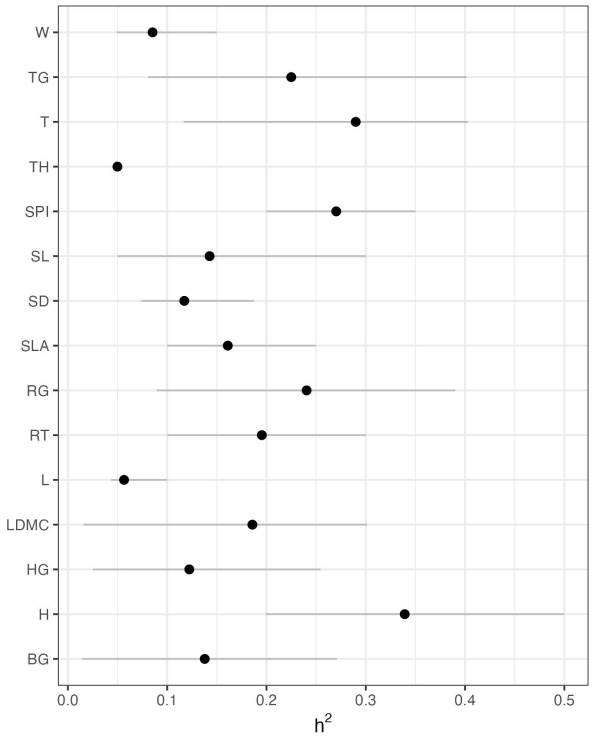

Figure S13: **Posterior mean narrow sense heritability for all measured phenotypes.** The black dot indicates the posterior mean and the gray bar 95% credible interval. See Figure 1.3B for trait codes.

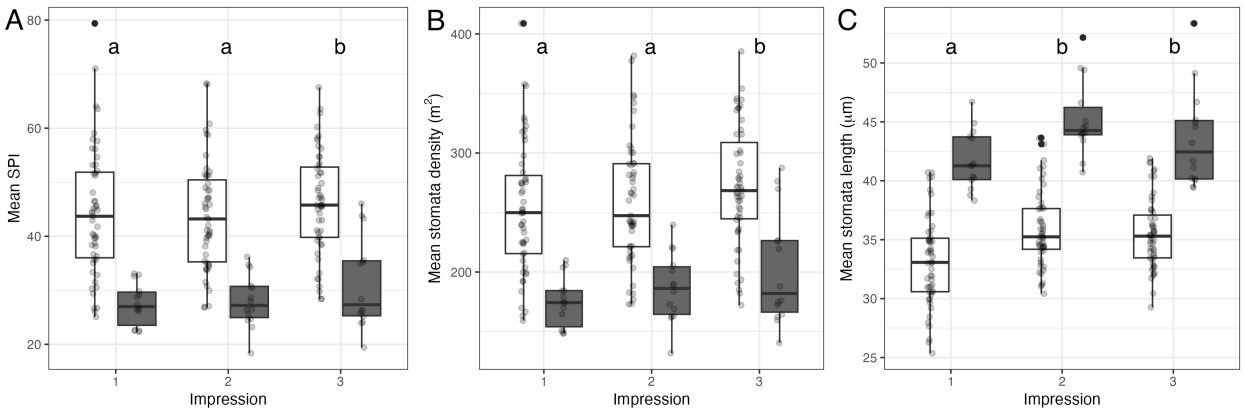

Figure S14: **The effect of cytotype on stomatal traits is consistent across the leaf.** We measured stomatal traits at three locations on the abaxial side of each leaf to assess developmental variation. Impression 1 is an impression taken within the bottom 1 inch of the leaf, impression 2 is the middle of the leaf, and impression 3 is within the top inch of the leaf. We assessed developmental variation in (A) stomatal pore index (SPI), (B) stomatal density, and (C) stomata length on a subset of genotypes (n = 23). The boxplot depict the distribution of genotype means for each ploidy where white is 6*x* and gray is 9*x* genotypes. The mean values for each genotype are overlaid as points. Lowercase letters above the boxplots indicate which impressions means are significantly different using a Tukey test and 95% confidence level.

Table S1: Population metadata. Population ID refers to sites in Figure 2.1A. CG indicates whether or not a population was included in the common garden experiment.

Population ID	Subpopulation ID	Location	Latitude	Longitude	CG
AFT	Big Bluestem Texas	Afton, Dickens County	33.748371	-100.7992	N
AUS	AUB 79	Austin, TX	30.0763595	-97.9433	Y
AUS	AUB 4	Austin, TX	30.0798056	-97.9344	Y
BAR	BB	Barta Brothers' Ranch, NE	42.2333	-99.6500	N
BOU	Kellogg 1284	Boulder, CO	39.8673889	-105.2427	Y
BOU	Kellogg 1283	Boulder, CO	39.9339	-105.2075	Y
CDB	CDB	Cedar Bluffs Reserve, KS	38.75	-99.7666	Y
CUI	Cuivre River SP	Cuivre River SP, MO	39.030639	-90.9619	N
CUI	Kellogg 1298		39.03333	-90.9174	N
DES	DES	Desoto Railroad Prairie, IL	37.85	-89.2333	N
ESL	ESL	East Shoal Lake, Manitoba	50.28583	-97.5144	Y
FUL	FUL	Fults Hill Prairie, IL	37.9666	-89.8000	Y
HAL	Kellogg 1271	Hall, NE	40.7436	-98.5851	N
KEN	Mckain WI Div	Kenosha, WI	42.5999	-88.2230	Y
KON	Konza	Konza Research Station	39.0886	-96.5533	Y
MAN	MBTGPP	Manitoba Tallgrass Prairie Preserve, Manitoba	49.07426	-96.7431	Y
MIL	Kellogg 1277	Milnesand Prairie Reserve, NM	33.6833	-103.3405	Y
MON	Mckain ILL Div	Montgomery, IL	39.350439	-89.6435	N
REL	REL	Relict Prairie, KS	38.85	-99.3666	Y
SAL	SAL	Saline Expt. Range, KS	39.0333	-101.3333	Y
SUT	Suther	Suther Prairie, NC	35.451238	-80.4673	Y
SUT	AUB 94	Suther Prairie, NC	35.4486526	-80.4672	N
TWE	12mi	Twelve Mile Railroad Prairie, IL	38.7666	-88.8333	Y
VIC	Victoria Glades glade 1	Victoria Glades glade 1	38.2021462	-90.5549	N
WAL	WAL	Walters Prairie, IL	38.9833	-88.1500	Y
WCG	Wallen creek glades	Wallen creek glades, MO	37.812797	-90.7006	N
WEB	WEB	Webster Reserve, KS	39.4	-99.5333	N
WEK	AUB 121	Wekiya State Park, FL	28.7124841	-81.4796	N
WEL	807 BBS	Welda Prairies, Anderson county, KS	38.16897	-95.2748	N

Table S2: Model summaries for Model 1. Significance of fixed effects was evaluated with an ANOVA and signficance of random effects was evaluated with a likelihood ratio test. Significance was determined by a threshold of p = 0.0033. Significant p-values are bolded. Estimates for SLA, basal growth, and percent flowering tillers are transformed as described in 6..

Trait	**n**	**Fixed effects**	**Estimate**	**SE**	**p**	**Random effects**	**Variance**	**SE**	χ^2	**p**
Leaf width	2712	Intercept	0.649	0.093		Population	0.009	0.004	21.3	**3.85E-06**
		Ploidy	-0.031	0.032	0.325	Block	0.001	0.001	57.5	**3.40E-14**
						Year	0.015	0.021	591.0	**1.55E-130**
						Genotype	0.006	0.001	268.6	**2.31E-60**
						Residual	0.028	8.05E-04		
Leaf length	2494	Intercept	18.9	4.5		Population	6.94	3.30	22.5	**2.12E-06**
		Ploidy	0.00074	0.906	0.999	Block	1.45	1.51	48.8	**2.80E-12**
						Year	38.26	54.15	927.6	**9.96E-204**
						Genotype	4.01	0.96	88.8	**4.28E-21**
						Residual	41.58	1.20		
Leaf thickness	2552	Intercept	0.150	0.0074		Population	4.28E-04	1.88E-04	30.0	**4.34E-08**
		Ploidy	-0.016	0.0053	0.0019	Block	1.43E-05	1.51E-05	18.3	**2.00E-05**
						Year	2.95E-05	4.25E-05	322.5	**4.16E-72**
						Genotype	1.56E-04	3.15E-05	322.5	**4.16E-72**
						Residual	6.27E-04	1.78E-05		

Table S2 Continued from previous page.

Trait	n	Fixed effects	Estimate	SE	p	Random effects	Variance	SE	χ^2	p
SLA	2530	Intercept	0.0561	0.114		Population	0.073	0.048	6.3	0.0121
		Ploidy	0.264	0.163	0.105	Block	0.008	0.009	14.1	**1.80E-04**
						Year	0.000	7.96E-04	0.00	1.000
						Genotype	0.165	0.033	292.6	**1.35E-65**
						Residual	0.707	0.020		
LDMC	1260	Intercept	379.0	7.81		Population	97.0	157.6	0.25	0.616
		Ploidy	-2.56	12.80	0.841	Block	63.6	70.4	13.73	**2.10E-04**
						Genotype	1132.2	230.0	221.65	**3.95E-50**
						Residual	2721.6	112.0		
Change in	199	Intercept	2.23	0.216		Population	0.387	0.213	12.168	**4.90E-04**
aboveground		Ploidy	-1.02	0.303	**0.0008**	Block	0.013	0.022	1.131	0.288
biomass						Genotype	0.362	0.110	20.405	**6.27E-06**
						Residual	0.561	0.072		
Height growth	210	Intercept	1.28	0.04		Population	0.008	0.005	4.881	0.027
		Ploidy	-0.12	0.06	0.0441	Block	0.001	0.002	1.822	0.177
						Genotype	0.007	0.005	2.956	0.086
						Residual	0.045	0.005		

Table S2 Continued from previous page.

Trait	n	Fixed effects	Estimate	SE	p	Random effects	Variance	SE	χ^2	p
Height	438	Intercept	150.5	20.0		Population	488.60	243.66	22.71	**1.88E-06**
		Ploidy	-0.956	8.52	0.911	Block	0.58	3.75	0.03	0.864
						Year	700.34	993.35	232.82	**1.45E-52**
						Genotype	396.05	85.65	130.72	**2.86E-30**
						Residual	452.51	33.87		
Number of	421	Intercept	61.1	23.2		Population	0.00	59.81	1.00E-05	0.998
tillers		Ploidy	9.55	9.17	0.298	Block	4.38	9.59	0.46	0.5
						Year	1043.02	1479.89	214.47	**1.45E-48**
						Genotype	593.41	127.23	92.41	**7.03E-22**
						Residual	705.93	54.12		
Tiller growth	201	Intercept	2.51	0.16		Population	0.142	0.118	2.406	0.121
		Ploidy	-0.21	0.30	0.493	Block	0.000	0.013	0.000	0.993
						Genotype	0.319	0.122	9.692	**0.002**
						Residual	0.827	0.106		
Basal growth	204	Intercept	1.44	0.133		Population	0.029	0.031	1.428	0.232
		Ploidy	-0.183	0.168	0.276	Block	0.033	0.039	7.963	0.005
						Genotype	0.079	0.041	4.877	0.027

Table S2 Continued from previous page.

Trait	n	Fixed effects	Estimate	SE	p	Random effects	Variance	SE	χ^2	p
						Residual	0.350	0.044		
Flowering tillers (%)	420	Intercept	0.592	0.093		Population	0.0026	0.0031	0.3098	0.5778
		Ploidy	0.005	0.054	0.931	Block	0	2.57E-04	1.00E-05	0.997
						Year	0.016	0.023	79.350	**5.20E-19**
						Genotype	0.016	0.004	48.765	**2.89E-12**
						Residual	0.035	0.003		
Stomatal density	1246	Intercept	240.5	7.89		Population	492.2	273.0	8.7	**3.21E-03**
		Ploidy	40	10.6	**0.0001**	Block	20.3	25.8	4.2	0.0407
						Genotype	612.3	134.5	152.6	**4.80E-35**
						Residual	2164.4	89.6		
Stomata length	1247	Intercept	33.9	0.826		Population	6.40	3.21	15.13	**1.00E-04**
		Ploidy	-4.3	0.979	<**0.0001**	Block	0.14	0.16	8.46	0.00364
						Genotype	5.80	1.13	360.27	**2.46E-80**
						Residual	8.47	0.35		
SPI	1245	Intercept	41.9	1.75		Population	26.83	14.09	11.66	**6.40E-04**
		Ploidy	8.63	2.22	**0.0001**	Block	0.68	0.88	3.99	0.0458
						Genotype	28.12	5.90	210.30	**1.18E-47**

Table S2 Continued from previous page.

Trait	n	Fixed effects	Estimate	SE	p	Random effects	Variance	SE	χ^2	p
						Residual	75.25	3.12		

Table S3: Model summaries for Model 6.13. Significance of fixed effects was evaluated with an ANOVA and signficance of random effects was evaluated with a likelihood ratio test. Significance was determined by a threshold of p = 0.0033. Significant p-values are bolded. Estimates for SLA, basal growth, and percent flowering tillers are transformed as described in 6..

Trait	**n**	**Fixed effects**	**Estimate**	**SE**	**p**	**Random effects**	**Variance**	**SE**	χ^2	**p**
Leaf width	2712	Intercept	0.653	0.091		Population	0.0040	0.0025	7.9	**0.0049**
		Ploidy	0.028	0.032	**0**	Block	0.0010	0.0011	51.4	**7.40E-13**
		PC1	2.56E-05	1.04E-05	**0**	Year	0.0150	0.0212	580.7	**2.61E-128**
		PC3	1.37E-04	6.67E-05	**0**	Genotype	0.0059	0.0012	252.1	**9.06E-57**
						Residual	0.0286	0.0008		
Leaf length	2494	Intercept	18.94	4.47		Population	3.28	2.00	7.35	**0.00669**
		Ploidy	-0.15	0.89	1	Block	1.46	1.51	49.05	**2.49E-12**
		PC1	9.98E-04	2.95E-04	**0**	Year	38.26	54.14	927.59	**9.84E-204**
		PC3	1.46E-04	1.88E-03	**0.0012**	Genotype	4.05	0.96	88.51	**5.07E-21**
						Residual	41.58	1.20		
Leaf thickness	2552	Intercept	0.150	0.006		Population	2.05E-04	1.08E-04	14.8	**1.20E-04**
		Ploidy	0.014	0.005	**0**	Block	8.98E-06	9.76E-06	18.2	**2.00E-05**
		PC1	5.48E-06	2.15E-06	**0**	Year	2.86E-05	4.11E-05	52.9	**3.43E-13**
		PC3	-3.45E-05	1.36E-05	1	Genotype	1.59E-04	3.20E-05	315.6	**1.30E-70**
						Residual	6.28E-04	1.79E-05		

Table S3 Continued from previous page.

Trait	n	Fixed effects	Estimate	SE	p	Random effects	Variance	SE	χ^2	p
SLA	2530	Intercept	0.045	0.105		Population	0.048	0.040	4.568	0.033
		Ploidy	-0.212	0.160	**0**	Block	0.008	0.009	14.163	**1.70E-04**
		PC1	-9.28E-05	4.28E-05	**0**	Year	0	7.96E-04	0	0.9999
		PC3	4.19E-04	2.79E-04	**0**	Genotype	0.163	0.033	292.31	**1.56E-65**
						Residual	0.707	0.020		
LDMC	1260	Intercept	375.92	7.15		Population	0	108.6	0	0.9995
		Ploidy	9.33	11.89	0.0341	Block	64.7	71.6	14.0	**1.80E-04**
		PC1	-0.008	0.002	1	Genotype	1031.6	211.4	222.1	**3.15E-50**
		PC3	0.024	0.016	1	Residual	2722.1	112.1		
Change in	199	Intercept	2.28	0.13		Population	0	0.051	0	**0.9995**
aboveground		Ploidy	0.79	0.26	**1.00E-04**	Block	0.012	0.021	1.04	0.307
biomass		PC1	2.14E-04	5.25E-05	**0**	Genotype	0.342	0.104	19.93	**8.03E-06**
	PC3	0.0013	3.30E-04	**0**	Residual	0.562	0.072			
Height growth	210	Intercept	1.280	0.035		Population	0.0034	0.0038	0.771	0.380
		Ploidy	0.080	0.058	**3.70E-03**	Block	0.0014	0.0020	1.809	0.179
		PC1	2.35E-05	1.31E-05	**1.00E-04**	Genotype	0.0074	0.0047	2.970	0.0848
		PC3	-1.91E-04	8.32E-05	1	Residual	0.0448	0.0055		

Table S3 Continued from previous page.

Trait	n	Fixed effects	Estimate	SE	p	Random effects	Variance	SE	χ^2	p
Height	438	Intercept	151.48	19.31		Population	142.4	112.8	4.22	**0.040**
		Ploidy	-0.205	8.203	1	Block	0.47	3.66	0.02	0.888
		PC1	0.008	0.002	**0**	Year	699.9	992.5	232.5	**1.70E-52**
		PC3	0.033	0.015	**0**	Genotype	395.9	85.4	130.5	**3.24E-30**
						Residual	452.7	33.9		
Number of	421	Intercept	60.80	23.31		Population	28.9	83.2	0.08	0.772
tillers		Ploidy	-8.92	9.50	0.063	Block	4.52	9.74	0.48	0.488
		PC1	-9.72E-04	2.04E-03	1	Year	1043.2	1480.1	214.7	**1.33E-48**
		PC3	-9.41E-04	1.33E-02	1	Genotype	593.2	128.9	92.7	**6.20E-22**
						Residual	705.6	54.1		
Tiller growth	201	Intercept	2.50	0.16		Population	0.156	0.134	2.48	0.116
		Ploidy	0.25	0.31	0.109	Block	0	0.013	5.00E-05	0.994
		PC1	-2.45E-05	7.72E-05	1	Genotype	0.320	0.123	9.71	**0.002**
		PC3	5.37E-04	4.94E-04	**0**	Residual	0.828	0.106		
Basal growth	204	Intercept	1.436	0.126		Population	0.0096	0.0237	0.188	0.664
		Ploidy	0.201	0.162	0.066	Block	0.0329	0.0383	7.73	0.0054
		PC1	2.69E-05	3.36E-05	0.050	Genotype	0.0737	0.0405	4.24	0.0394

Table S3 Continued from previous page.

Trait	n	Fixed effects	Estimate	SE	p	Random effects	Variance	SE	χ^2	p
		PC3	4.86E-04	2.12E-04	**0**	Residual	0.3536	0.0444		
Flowering tillers (%)	420	Intercept	0.591	0.094		Population	0.00524	0.00456	1.09	0.296
		Ploidy	-0.002	0.055	1	Block	0	2.57E-04	1.00E-05	0.997
		PC1	2.33E-06	1.44E-05	0.311	Year	0.0158	0.0226	79.2	**5.71E-19**
		PC3	3.75E-05	9.24E-05	0.0046	Genotype	0.0151	0.00391	48.1	**4.12E-12**
						Residual	0.0349	0.00268		
Stomatal density	1246	Intercept	241.10	7.88		Population	499.1	295.2	10.8	**0.00103**
		Ploidy	-39.24	10.45	**1**	Block	20.2	25.7	4.16	0.0415
		PC1	0.0032	0.0036	**0**	Genotype	586.4	130.1	145.5	**0**
		PC3	0.0400	0.0228	**0**	Residual	2164.7	89.6		
Stomata length	1247	Intercept	33.89	0.83		Population	6.56	3.53	19.41	**1.00E-05**
		Ploidy	4.15	0.98	**0**	Block	0.135	0.157	8.46	0.00364
		PC1	2.21E-04	3.89E-04	**0**	Genotype	5.68	1.11	357.76	**8.67E-80**
		PC3	-0.0040	0.0025	1	Residual	8.47	0.351		
SPI	1245	Intercept	42.03	1.76		Population	27.9	15.6	14.8	**1.20E-04**
		Ploidy	-8.40	2.21	**0**	Block	0.681	0.873	3.97	0.0464
		PC1	4.60E-04	8.19E-04	**0**	Genotype	27.2	5.74	203.3	**4.02E-46**

Table S3 Continued from previous page.

Trait	n	Fixed effects	Estimate	SE	p	Random effects	Variance	SE	χ^2	p
		PC3	0.0091	0.0052	**0**	Residual	75.3	3.12		

Table S4: PacBio library statistics for the libraries included in the *Andropogon gerardi* (var. Kellogg-1272) genome assembly and their respective assembled sequence coverage levels.

Cutoff	Number of Reads	Basepairs	Average Read Length	Coverage
0	16,785,606	343,726,563,435	19,680	85.93x
1,000	16,784,941	343,726,421,728	19,680	85.93x
2,000	16,782,985	343,723,282,795	19,681	85.93x
3,000	16,781,398	343,719,397,953	19,681	85.93x
4,000	16,780,091	343,714,820,784	19,682	85.93x
5,000	16,777,997	343,705,307,473	19,682	85.93x
6,000	16,774,286	343,684,717,768	19,683	85.92x
7,000	16,768,962	343,650,000,372	19,684	85.91x
8,000	16,763,136	343,606,292,932	19,686	85.90x
9,000	16,757,799	343,560,982,110	19,687	85.89x
10,000	16,753,026	343,515,663,833	19,688	85.88x
11,000	16,748,251	343,465,530,496	19,689	85.87x
12,000	16,743,169	343,406,976,519	19,690	85.85x
13,000	16,734,980	343,303,965,973	19,692	85.83x
14,000	16,704,343	342,886,573,490	19,699	85.72x
15,000	16,566,666	340,874,186,260	19,732	85.22x
16,000	16,066,795	333,079,495,791	19,853	83.27x
17,000	14,753,317	311,321,403,055	20,189	77.83x
18,000	12,463,595	271,193,752,445	20,850	67.80x
19,000	9,938,051	224,485,620,806	21,750	56.12x

Table S5: Genomic libraries included in the *Andropogon gerardi* (var. Kellogg-1272) genome assembly and their respective assembled sequence coverage levels in the final release. *Average read length of PacBio reads.

Library	Sequencing Platform	Average Read/Insert Size	Read Number	Assembled Sequence Coverage
JEXY	Illumina	400	1,601,302,817	92.80x
GOXCG	Illumina-HiC	–	938,370,083	54.38x
	PacBio	19,680*	16,785,606	85.93x
Total		–	2,556,458,506	233.11x

Table S6: Summary statistics of the initial output of the HAP1 RACON polished HiFiAsm+HIC assembly. The table shows total contigs and total assembled basepairs for each set of scaffolds greater than the size listed in the left hand column.

Minimum Scaffold Length	**Number of Scaffolds**	**Number of Contigs**	**Scaffold Size**	**Basepairs**	**Non-gap Basepairs**
5 Mb	61	61	2,664,415,677	2,664,415,677	100.00%
2.5 Mb	65	65	2,677,111,928	2,677,111,928	100.00%
1 Mb	68	68	2,681,267,982	2,681,267,982	100.00%
500 kb	77	77	2,687,222,053	2,687,222,053	100.00%
250 kb	117	117	2,700,371,179	2,700,371,179	100.00%
100 kb	344	344	2,732,853,625	2,732,853,625	100.00%
50 kb	1,064	1,064	2,780,484,710	2,780,484,710	100.00%
25 kb	1,064	1,064	2,780,484,710	2,780,484,710	100.00%
10 kb	1,064	1,064	2,780,484,710	2,780,484,710	100.00%
5 kb	1,064	1,064	2,780,484,710	2,780,484,710	100.00%
2.5 kb	1,064	1,064	2,780,484,710	2,780,484,710	100.00%
1 kb	1,064	1,064	2,780,484,710	2,780,484,710	100.00%
0 bp	1,064	1,064	2,780,484,710	2,780,484,710	100.00%

Table S7: Summary statistics of the initial output of the HAP2 RACON polished HiFiAsm+HIC assembly. The table shows the total contigs and total assembled basepairs for each set of scaffolds greater than the size listed in the left-hand column.

Minimum Scaffold Length	**Number of Scaffolds**	**Number of Contigs**	**Scaffold Size**	**Basepairs**	**Non-gap Basepairs**
5 Mb	61	61	2,608,260,179	2,608,260,179	100.00%
2.5 Mb	65	65	2,622,080,605	2,622,080,605	100.00%
1 Mb	67	67	2,625,544,079	2,625,544,079	100.00%
500 kb	78	78	2,633,098,987	2,633,098,987	100.00%
250 kb	115	115	2,645,787,790	2,645,787,790	100.00%
100 kb	290	290	2,671,355,876	2,671,355,876	100.00%
50 kb	731	731	2,701,527,046	2,701,527,046	100.00%
25 kb	731	731	2,701,527,046	2,701,527,046	100.00%
10 kb	731	731	2,701,527,046	2,701,527,046	100.00%
5 kb	731	731	2,701,527,046	2,701,527,046	100.00%
2.5 kb	731	731	2,701,527,046	2,701,527,046	100.00%
1 kb	731	731	2,701,527,046	2,701,527,046	100.00%
0 bp	731	731	2,701,527,046	2,701,527,046	100.00%

Table S8: Final summary assembly statistics for the HAP1 chromosome scale assembly.

Scaffold Total	48
Contig Total	88
Scaffold Sequence Total	2,669.7 Mb
Chromosome Sequence Total	2,666.8 Mb
Contig Sequence Total	2,669.3 Mb (0.001% gap)
Scaffold N/L50	13 / 86.8 Mb
Contig N/L50	16 / 63.1 Mb

Table S9: Final summary assembly statistics for the HAP2 chromosome scale assembly.

Scaffold Total	39
Contig Total	68
Scaffold Sequence Total	2,588.6 Mb
Chromosome Sequence Total	2,586.5 Mb
Contig Sequence Total	2,588.3 Mb (0.05% gap)
Scaffold N/L50	13 / 86.4 Mb
Contig N/L50	17 / 59.2 Mb

Chapter 2: A happy accident: A novel turfgrass reference genome

Published in G3 June 2023

Alyssa R. Phillips[*,†,1], Arun S. Seetharam[*,2], Patrice S. Albert[3], Taylor AuBuchon-Elder[4], James A. Birchler[3], Edward S. Buckler[5,6,7], Lynn J. Gillespie[8], Matthew B. Hufford[3], Victor Llaca[9], M. Cinta Romay[6], Robert J. Soreng[10], Elizabeth A. Kellogg[4], and Jeffrey Ross-Ibarra[†,1,11]

[*]Co-first authors

[†]Corresponding authors

[1]Department of Evolution and Ecology and Center for Population Biology, University of California, Davis, Davis, CA 95616, USA

[2]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011, USA

[3]Division of Biological Sciences, University of Missouri, Columbia, MO 65201, USA

[4]Donald Danforth Plant Science Center, Olivette, MO 63132, USA

[5]School of Integrative Plant Sciences, Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY 14850, USA

[6]Institute for Genomic Diversity, Cornell University, Ithaca, NY 14850, USA

[7]Agricultural Research Service, United States Department of Agriculture, Ithaca, NY 14850, USA

[8]Botany Section, Research and Collections, Canadian Museum of Nature, Ottawa, ON K2P 2R1, Canada

[9]Corteva Agriscience, Johnston, IA 50131, USA

[10]Deptartment of Botany, Smithsonian Institution, Washington, DC 20560, USA

[11]Genome Center, University of California, Davis, Davis, CA 95616, USA

1. Abstract

Poa pratensis, commonly known as Kentucky bluegrass, is a popular cool-season grass species used as turf in lawns and recreation areas globally. Despite its substantial economic value, a reference genome had not previously been assembled due to the genome's relatively large size and biological complexity that includes apomixis, polyploidy, and interspecific hybridization. We report here a fortuitous *de novo* assembly and annotation of a *P. pratensis* genome. Instead of sequencing the genome of a C4 grass, we accidentally sampled and sequenced tissue from a weedy *P. pratensis* whose stolon was intertwined with that of the C4 grass. The draft assembly consists of 6.09 Gbp with an N50 scaffold length of 65.1 Mbp, and a total of 118 scaffolds, generated using PacBio long reads and Bionano optical map technology. We annotated 256K gene models and found 58% of the genome to be composed of transposable elements. To demonstrate the applicability of the reference genome, we evaluated population structure and estimated genetic diversity in *P. pratensis* collected from three North American prairies, two in Manitoba, Canada and one in Colorado, USA. Our results support previous studies that found high genetic diversity and population structure within the species. The reference genome and annotation will be an important resource for turfgrass breeding and study of bluegrasses.

2. Background

Poa pratensis L., commonly known as Kentucky bluegrass, is an economically valuable horticultural crop grown globally on lawns and recreational areas as turf (Haydu *et al.* 2006). Native to Europe and Asia, it was introduced to North America in the seventeenth century by European colonizers as a forage crop (Carrier and Bort 1916; Raggi *et al.* 2015). Today, Kentucky bluegrass is the most popular cool-season grass used for turf due to it's vigorous growth and quick establishment that creates a dense, strong sod with a long lifespan (Casler and Duncan 2003).

Today, there are 40 million acres of managed turf in the United States (U.S.), an area approximately the size of the state of Florida (Milesi *et al.* 2005). While this massive area has the

potential to serve as an important carbon sink, the large water and fertilization resources required currently outweigh the benefits (Milesi *et al.* 2005). Breeding efforts are underway to improve environmental-stress tolerances, disease and insect resistance, seed quality and yield, as well as uniformity and stability of traits (reviewed in Bonos and Huff 2013). While the economic value of *P. pratensis* is high, it is highly invasive, and in the last 30 years has aggressively invaded the North American Northern Great Plains, altering ecosystem function by reducing pollinator and plant diversity and altering nutrient dynamics (Kral-O'Brien *et al.* 2019; DeKeyser *et al.* 2015; Hendrickson *et al.* 2021). Continued research into the genetic diversity of wild *P. pratensis* is needed to understand how invasive populations are rapidly adapting, and the study of wild populations may enable identification of disease or environmentally tolerant ecotypes for use in turfgrass breeding.

Previous studies using RAPD, ISSR, and SRR markers demonstrated high genetic diversity in both developed cultivars and wild populations but limited population structure between these groups (Bushman *et al.* 2013; Raggi *et al.* 2015; Honig *et al.* 2012, 2018, but see Dennhardt *et al.* 2016). Population divergence has been detected amongst some wild populations (Dennhardt *et al.* 2016) but the extent of population structure is unclear. There are a number of potential reasons for finding a lack of population structure, including gene flow, the independent development of cultivated lines from locally adapted ecotypes (Raggi *et al.* 2015; Bonos and Huff 2013), and geographic heterogeneity in patterns of genetic diversity. Repeated reversion of cultivars to wild forms has also been suggested, but is unlikely (Dennhardt *et al.* 2016). Alternatively, previous studies may simply not have had sufficient marker resolution to detect population structure in a highly heterozygous polyploid species like *P. pratensis*.

Genetic analysis and improvement of turfgrass are challenging because of apomixis and polyploidy (Bushman and Warnke 2013). *Poa pratensis* is a facultative apomict, meaning it can reproduce sexually or asexually by aposporous apomixis, and it is a polyploid with frequent aneuploidy (Brown 1939). Although apomixis is a highly valued trait for seed production, high rates of apomixis stymie the recombination needed to genetically analyze traits or recombine beneficial

traits into one cultivar (Bonos and Huff 2013). Polyploidy and aneuploidy further these difficulties due to copy number variation of regions of interest and non-Mendelian inheritance resulting from double reduction. While some progress has been made in managing apomixis (Funk *et al.* 1967; Pepin and Funk 1971; Matzk 1991), including the discovery of its genetic basis (Albertini *et al.* 2004; Marconi *et al.* 2020), the development of additional molecular and genomic tools in *P. pratensis* are needed to move genetic analysis and breeding efforts forward in the face of its complex biology.

Here, we report the first *P. pratensis* genome. While attempting to assemble the genome for a C4 prairie grass, *Andropogon gerardi*, we unknowingly sequenced and assembled a wild *Poa* growing in the same pot. Fortunately, this resulted in a highly contiguous, near complete genome assembly. We utilized the reference genome and wild *Poa* from three prairies to investigate the genetic diversity and population structure of North American *Poa*. The reference genome and annotation presented here are an important advancement for Kentucky bluegrass breeding. Additionally, this reference genome provides an important resource for the study of closely related bluegrasses including *P. trivialis* L., *P. annua* L., and *P. arachnifera* Torr.

3. Materials & Methods

3.1 Sample collection

Rhizomes of *Poa* species were collected fortuitously as part of a different project aimed at collecting major C4 prairie grasses (*Andropogon gerardi* Vitman, *Sorghastrum nutans* (L.) Nash, and *Schizachyrium scoparium* (Michx.) Nash) in moist prairies in Colorado, USA and two prairies in Manitoba, Canada and (Table S1). Necessary permissions and permits were obtained prior to collecting. Plants were brought back to the United States from Canada under phytosanitary certificate 3193417.

The C4 focal plants were dug up with a shovel late in the growing season in 2018 (when the *Poa* was dormant and thus invisible), soil was washed off, rhizomes were wrapped in wet paper

towels, and leaves were cut back to about 4 inches height to reduce transpiration. The focal C4 plant was placed in a 1-gallon Ziploc bag and returned to the plant growth facility at the Donald Danforth Plant Science Center in St. Louis, MO, USA. Plants were potted in 2:1 BRK20 promix soil to turface. The previously dormant *Poa* plants produced fresh green leaves in this setting and grew faster than the C4 plant with which it was entwined. Once it was discovered that *Poa* had interpolated itself into the rhizome and root area of the C4 plants, the *Poa* plants were extricated and placed in separate pots.

One *Poa* was found inside the pot for an *Andropogon gerardi* genotype which was used to attempt assembly of a reference genome. Instead of collecting tissue from the *A. gerardi* plant, tissue was accidentally sampled from the *Poa* plant. This *Poa* individual is referred to as the *Poa* reference individual (Table S1). Eight additional *Poa*, referred to here as the *Poa* population panel, were discovered in various pots for C4 grasses whose genomes we attempted to sequence.

As *Poa* species generally require vernalization to flower, several plants were over-wintered outside under mulch and flowered in spring 2020 and/or 2021; voucher specimens were taken from these plants to verify species identity and have been deposited at the Smithsonian Institution (Washington, District of Columbia, U.S.A) and the Missouri Botanical Garden (St. Louis, MO, U.S.A.) (Heide 1994). Not all *Poa* individuals survived, so some specimens lack vouchers. Additionally, not all surviving *Poa* flowered so vegetative vouchers were submitted (Table S1).

3.2 PacBio sequencing

Approximately 4.1 g fresh tissue from the reference individual was extracted for PacBio sequencing using a High Molecular Weight (HMW) DNA approach based on the Circulomics Big DNA Kit (Circulomics, USA). This method yields DNA with a center of mass at 200 Kb, which is sufficient to construct PacBio CLR 20 Kb+ libraries. Sequencing was completed on the Sequel II across four SMRTCells. DNA extraction and sequencing was completed by Corteva Agriscience^{TM}.

3.3 Bionano optical map generation

DNA was extracted from 0.7 g of fresh leaf tissue from the reference individual using agarose embedded nuclei and the Bionano Prep^{TM} Plant Tissue DNA Isolation kit. DNA extraction, labeling, imaging, and optical map assembly followed the methods previously described in Hufford *et al.* (2021) and was completed by Corteva Agriscience^{TM}.

3.4 Preparation and imaging of metaphase spreads

Metaphase spreads were utilized to estimate chromosome count and ploidy of the reference individual. Root tips were harvested from a recent off-shoot of the reference individual, treated with nitrous oxide (3 hr at 160 psi) to stop mitosis in metaphase (Kato 1999), then processed as previously described in Kato *et al.* (2004) and Kato *et al.* (2011) with minor modification. Specifically, the root tips were fixed in 90% acetic acid for 15 min, then rinsed with and stored in 70% ethanol at -20°C. Ethanol was removed from the root tips prior to enzymatic digestion by soaking in water for 10 min. About 1 mm of the tip (meristem and root cap) was excised and transferred to a tube containing 20 µL of 3% cellulase R-10 (Desert Biologicals, Phoenix, AZ) and 1.25% pectolyase Y-23 (Desert Biologicals) in citrate buffer (10 mM sodium citrate, 10 mM EDTA, adjusted to pH 5.5 with citric acid) on ice. The tissue was digested for approximately 1 hr at 37°C. Seventy percent ethanol was used to inactivate the enzymes and rinse the samples. The ethanol was replaced with approximately 7 µL of a solution of 90% acetic acid and 10% methanol. The tissue was broken and cells dispersed using a blunted dissecting probe. The entire volume was dropped from a height less than 1 cm onto a microscope slide in a container lined with wet paper towels and allowed to dry.

Preparations were counterstained with a 1/20 dilution of Vectashield with DAPI (Vector Laboratories, Burlingame, CA). Images were captured using Applied Spectral Imaging software (Carlsbad, CA) on an Olympus BX 61 fluorescence microscope. Photoshop Brightness/Contrast and Curves functions were used to decrease background noise and better define the chromosomal arms.

Genome size estimation Genome size was estimated for the *Poa* reference individual and 4 of

the population panel individuals (Table S5). Not all population panel individuals were sampled as some plants died prior to estimation. Genome size estimation methods using an internal standard are modified from Doležel *et al.* (2007). Two internal standards were used for the reference: maize B73 inbred line (5.16 pg/2C) and *Andropogon gerardi* accession CAM 1351 (6.13 pg/2C). Only the maize B73 internal standard was used for the population panel. Approximately 10x1 cm of fresh leaf tissue for the target and sample standard were placed in a plastic square petri dish. A chopping solution composed of 1 mL LB01 buffer solution, 250µL PI stock (2 mg/mL), and 25 µL RNase (1 mg/mL) added to the dish (1.25 mL; Doležel *et al.* 2007). The tissue was then chopped into 2-4 mm lengths and the chopping solution was mixed through the leaves by pipetting. The solution was then pipetted through a 30µm sterile single-pack CellTrics® filter into a 2 mL Rohren tube on ice. Three replicates were chopped separately and analyzed for each *Poa* population panel genotype and 9 replicates were analyzed for the reference. The samples were left to chill for 20 min before analysis with a BD Accuri^{TM} C6 flow cytometer. Samples were run in Auto Collect mode with a 5-min run limit, slow fluidics option, a FSC-H threshold with less than 200,000 events, and a 1-cycle wash. The cell count, coefficient of variation of FL2-A, and mean FL2-A were recorded for the target and reference sample with no gating. Results were analyzed separately for each replicate and manually annotated to designate the set of events. The replicates for each *Poa* genotype were averaged (Table S6).

3.5 Illumina sequencing of the *Poa* population panel

DNA was extracted from the *Poa* population panel using approximately 100 mg of lyophilized leaf tissue and a DNeasy® Plant Kit (Qiagen Inc., Germantown, MD). High throughput Illumina Nextera ® libraries were constructed and samples were sequenced with other plant samples in pools of 96 individuals in one lane of an S4 flowcell in an Illumina Novaseq 6000 System with paired-end 150-bp reads, providing approximately 0.80X coverage for each sample.

3.6 Species identification

Species identification was completed using both morphological and DNA sequence data. Morphological assessment was completed for the *Poa* reference genome and three of the population panel samples using flowering and vegetative vouchers. Phylogenetic inference was completed for species identification of all samples using one plastid and two nuclear ribosomal DNA loci: *trn*T-*trn*L-*trn*F (TLF), external transcribed spacer (ETS), and internal transcribed spacer (ITS), respectively. Trees for *mat*K and *rpo*B-*trn*C were also evaluated but the sequences showed little variation across sampled species.

Sequences for these loci were extracted from the *Poa* population panel whole genome sequence data by aligning reads to a *P. pratensis* sequence for each locus downloaded from Genbank (Table S2) using the default options of bwa mem (v0.7.17; Li 2013). The alignment files were sorted using SAMtools (v1.7; Danecek *et al.* 2021), read groups were added using Picard AddOrReplaceReadGroups, and duplicates removed with Picard MarkDuplicates using default settings (`http://broadinstitute.github.io/picard`). We identified variable sites for each sample separately using GATK (v4.1) HaplotypeCaller with default options (Van der Auwera and O'Connor 2020). SNPs were filtered to remove sites with low mapping quality and low sequencing quality (`gatk VariantFiltration -filter "QUAL < 40.0" -filter "MQ < 40.0"` and default `gatk SelectVariants`). A consensus sequence for each locus and sample was generated using GATK FastaAlternateReferenceMaker, which replaces the gene reference bases at variable sites with the alternate allele.

Sequences were extracted from the reference genome by aligning the *P. pratensis* reference sequences downloaded from Genbank to the reference genome with bwa mem using default options (v0.7.17; Li 2013). This allowed us to identify the position of each locus in the reference. Each locus only mapped to a single region in the reference genome, which was extracted using bioawk (`https://github.com/lh3/bioawk`).

Sequences from the reference genome and the population panel were included in a dataset with 119 *Poa* samples from previous work (Table S3; Cabi *et al.* 2016, 2017; Gillespie *et al.* 2007,

2008, 2009, 2018; Giussani *et al.* 2016; Refulio-Rodriguez *et al.* 2012; Soreng and Gillespie 2018; Soreng *et al.* 2015, 2017, 2020; Sylvester *et al.* 2021). These samples were chosen to represent the phylogenetic diversity of the genus *Poa*, and include all seven currently recognized subgenera as well as 29 of 38 sections and several unclassified species groups (classification according to Gillespie *et al.* (2007), with updates by Cabi *et al.* (2017); Gillespie *et al.* (2008, 2018); Soreng and Gillespie (2018); Soreng *et al.* (2020)). Since formal infrageneric taxonomic delimitations are often imperfect, and the genus *Poa* is large and highly complex, genotype codes are used in Table S3 as shorthand for the plastid and nrDNA clades found in a sample or species (see Soreng *et al.* (2020) for the most recent iterations).

Sequences were aligned using the auto-select algorithm and default parameters in the MAFFT plugin (v7.017; Katoh and Standley 2013) in Geneious (v8.1.9; `http://www.geneious.com`) followed by manual adjustment. *Poa* sect. *Sylvestres* was used as the outgroup to root trees based on its strongly supported position as sister to all other *Poa* species in previous plastid analyses (Gillespie *et al.* 2007, 2009, 2018). Bayesian Markov chain Monte Carlo analyses were conducted in MrBayes (v3.2.6; Ronquist *et al.* 2012). Optimal models of molecular evolution were determined using the Akaike Information Criterion (AIC; Akaike 1974) conducted through likelihood searches in jModeltest (Darriba *et al.* 2012) with default settings. Models were set at `GTR + Γ` for ETS and `GTR + I + Γ` for ITS and TLF based on the AIC scores and the models allowed in MrBayes. Two independent runs of four chained searches were performed for three or four million generations, sampling every 500 generations, with default parameters. Analyses were stopped when an average standard deviation of split frequencies of 0.007001, 0.006350, and 0.006490 was reached for ITS, ETS, and TLF, respectively. A 25% burn-in was implemented prior to summarizing a 50% majority rule consensus tree and calculating Bayesian posterior probabilities. Trees were visualized and annotated in R using ggtree (v2.0.4) with ape (v5.4) and treeio (v1.10) (Yu 2020; R Core Team 2017; Wang *et al.* 2020; Paradis and Schliep 2019).

3.7 Genome assembly

PacBio subreads obtained as BAM files were converted to FASTA format using SAMtools (v1.10; Danecek *et al.* 2021) and error-correction was performed using overlap detection and error correction module (first stage) of Falcon (v1.8.0; Chin *et al.* 2016). For running Falcon, the following options were used: the expected genome size was set to 6.4 Gbp (`-genome_size = 6400000000`), a minimum of two reads, maximum of 200 reads, and minimum identity of 70% for error corrections (`--min_cov 2 --max_n_read 200, --min_idt 0.70`), using the 40x seed coverage for auto-calculated cutoff. The average read correction rate was set to 75% (`-e 0.75`) with local alignments at a minimum of 3000 bp (`-l 3000`) as suggested by the Falcon manual. For the DAligner step, the exact matching length of k-mers between two reads was set to 18 bp (`-k 18`) with a read correction rate of 80% (`-e 0.80`) and local alignments of at least 1000 bp (`-l 1000`). Genome assembly was performed with Canu (v1.9: Koren *et al.* 2017) using the error-corrected reads from Falcon. For sequence assembly, the corrected reads had over 70x coverage for the expected genome size of *Poa* and were characterized by N50 of 25.6 Kbp and average length of 16.3 Kbp. These reads were trimmed and assembled with Canu using the default options except for `ovlMerThreshold=500`.

The Canu generated contig assembly was further scaffolded utilizing the Bionano optical map with Bionano Solve (v3.4) and Bionano Access (v1.3.0), as described previously by Hufford *et al.* 2021. The default config file (hybridScaffold_DLE1_config.xml) and the default parameters file (optArguments_nonhaplotype_noES_noCut_DLE1_saphyr.xml) were used for the hybrid assembly. The scaffolding step of Bionano Solve incorporates three types of gaps: 1) gaps of estimated size (varying N-size, but not 100bp or 13bp), using calibrated distance conversion of optical map to basepair (cases when contiguous optical map connects two contigs); 2) gaps of unknown sizes (100-N gaps), when distance could not be estimated (cases when large repeat regions like rDNA or centromeres interrupt the optical map but evidence to connect the map is present); and 3) 13-N gaps, in regions where two or more independently assembled contigs align to the same optical map, overlapping at the ends. The 13-N gaps are usually caused by sequence similarity sufficient

for aligning to the optical map, but less than required to merge contigs. This could be caused by either high heterozygosity in that region, highly repetitive sequence, paralogous regions of the sub-genomes, or assembly errors. The contig overlaps, regardless of the size, are connected end-to-end by adding 13-N gaps when processed using Bionano Solve. Due to the polyploid nature of Poa as well as its high heterozygosity, these 13-N gaps had to be manually curated. We inspected the contig alignments to the optical map using using Bionano Access (v1.3.0), either to trim the overlapping sequence or to remove exact duplicates to generate error-free assembly.

3.8 Genome annotation

Gene prediction was carried out using a comprehensive method combining *ab initio* predictions (from BRAKER v2.1.6; Brůna *et al.* 2021) with direct evidence (inferred from transcript assemblies) using the BIND strategy (Li *et al.* 2021). Briefly, 58 RNA-seq libraries were downloaded from NCBI (Table S4) and mapped to the genome using a STAR (v2.5.3a; Dobin *et al.* 2013)-indexed genome and an iterative two-pass approach under default options to generate mapped BAM files. BAM files were used as input for multiple transcript assembly programs to assemble transcripts: Class2 (v2.1.7; Song *et al.* 2016), Cufflinks (v2.2.1; Trapnell *et al.* 2012), Stringtie (v2.1.4; Pertea *et al.* 2015) and Strawberry (v1.1.2; Liu and Dickerson 2017). Redundant assemblies were collapsed and the best transcript for each locus was picked using Mikado (v2.3.3; Venturini *et al.* 2018) by filling in the missing portions of the ORF using TransDecoder (v5.5.0; Haas *et al.* 2013) and homology as informed by the NCBI BLASTX (v2.10.1+; Altschul *et al.* 1990) results to the SwissProtDB (Duvaud *et al.* 2021). Splice junctions were also refined using Portcullis (v1.2.1; Mapleson *et al.* 2018) to identify isoforms and to correct misassembled transcripts. Both *ab initio* and direct evidence predictions were analyzed with TESorter (v1.3.0; Zhang *et al.* 2019a) to identify and remove any TE-containing genes before merging them. Merging was done using the GeMoMa (v1.8) Annotation Filter tool, to combine and filter gene predictions from BRAKER, Mikado and additional homology-based gene predictions generated by the GeMoMa pipeline using *Hordeum vulgare* annotations (Mascher *et al.* 2021; Keilwagen *et al.* 2016, 2018). The predictions

were prioritized using weights, with highest for homology (1.0), followed by direct evidence (0.9) and lowest for gene predictions from *ab initio* methods (0.1). Homology is defined by GeMoMa as protein sequence similarity and and intron position conservation relative to *Hordeum vulgare*. The Annotation Filter tool was run with settings to enforce the completeness of the prediction (`start=='M' stop=='*'`), external evidence support (`score/aa>=0.75`), and RNAseq support (`evidence>1` or `tpc==1.0`). The final predictions were subjected to phylostratiography analyses using phylostratr (v0.20; Arendsee *et al.* 2019). The focal species were set as '4545' for *Poa pratensis*, and default options were used. The program creates a clade tree of species based on the current NCBI tree of life, trims the tree to maximize evolutionary diversity, retrieves the species proteome from Uniprot, and compares the proteins of the focal species to those of other species in the tree using pairwise BLASTs (Diamond search). Each gene is then assigned to the deepest clade in which it has an inferred homolog. Genes found only in the focal species are considered orphan genes and assigned to the phylostratum '*Poa pratensis*.' Final gene-level annotations were saved in GFF3 format and the predicted peptides/CDS sequences were extracted using `gffread` of the Cufflinks package (v2.2.1; Trapnell *et al.* 2012).

3.9 Assessment of the assembly

Genome contiguity statistics were computed using the Assemblathon script (Bradnam *et al.* 2013). Gene space completeness was measured using BUSCO (v4.0; Manni *et al.* 2021) using the liliopsida_odb10 profile (n = 3278) and poales_odb10 profile (n = 4896) with default options. The contiguity of TE assembly was then assessed using the LTR Assembly Index (LAI; Ou *et al.* 2018). To compute LAI, we first annotated repeats using the Extensive *de-novo* TE Annotator (EDTA; v1.9.6; Ou *et al.* 2019), and intact LTR retrotransposons (LTR-RT) were identified using LTRharvest (v1.6.1; Manchanda *et al.* 2020), and LTR_FINDER_parallel (v1.1; Ellinghaus *et al.* 2008). LTR_retriever (v2.9.0; Ou *et al.* 2018) was then used to filter the intact LTRs and computed the LAI score for the genome.

3.10 Population genetics of *Poa*

The population panel was mapped to the scaffold assembly, excluding the alternate scaffolds, using bwa mem (v0.7.17; Li 2013). Reads were sorted using SAMtools (v1.7; Danecek *et al.* 2021), read groups were added using Picard AddOrReplaceReadGroups, and duplicates removed with Picard MarkDuplicates (`http://broadinstitute.github.io/picard`) using default settings.

Site filtering and genotyping was completed with ANGSD (v0.934; Korneliussen *et al.* 2014). Reads were filtered, retaining unique reads, reads with a flag below 255, and proper pairs (`angsd -uniqueOnly 1 -remove_bads 1 -only_proper_pairs 1 -trim 0`), as well as a minimum mapping and base quality of 30 (`angsd -minMapQ 30 -minQ 30`). Sites were filtered with a strict maximum depth cutoff in order to exclude sites where paralogs may be mapping. Assuming read depth follows a Poisson distribution with a mean of 0.8, we expect 99% of reads to have a depth of 4 or less. We included sites with a minimum depth of 1 and a maximum depth of 4 and required all genotypes to have data at a site (`angsd -doCounts 1 -setMinDepthInd 1 -setMaxDepthInd 4 -minInd 8`). Sites were also filtered for a minor allele frequency greater than 5% in the principal component analysis (PCA; `angsd -doMajorMinor 4 -doCounts 1 -doMaf 1 -minMaf 0.05`).

After filtering, a single-read was randomly sampled at each base to serve as the genotype (`angsd -doIBS 1`). This genotyping approach is discussed in Results and Discussion. A genotype matrix was sampled three independent times for each of the following analyses in order to assess sampling error.

Population structure and nucleotide diversity were evaluated to demonstrate the utility of the *P. pratensis* reference genome. Population structure was assessed using a principal component analysis (PCA) implemented in ANGSD (`angsd -doCov`). A PCA was run with all *Poa* and only *P. pratensis*. The covariance matrices were plotted with ggplot2 (v3.4) in R (R Core Team 2017; Wickham 2016).

Nucleotide diversity was estimated for each *P. pratensis* genotype in the *Poa* population panel as nucleotide diversity per genome using a custom R script. We are defining nucleotide diversity

per genome as the number of sites with the reference allele divided by the total number of sites. Only sites that met our filtering criteria and contained no missing data across *P. pratensis* genotypes were included. Results were plotted with ggplot2 in R.

4. Results and Discussion

4.1 Species identification and validation

Herbarium vouchers for the *Poa* reference genome and two of the population panel genotypes were identified as *P. pratensis* by their morphology (Table S1). The *Poa* reference genotype can be further classified as subspecies *angustifolia*, characterized by narrower and involute leaf blades, usually with strigose hairs on the adaxial surface of blades. The blades of *P. pratensis* subspecies *angustifolia* are firmer and tend to be more consistently glaucous. The intravaginal shoots are often disposed in fascicles of more than one shoot, the inflorescences are generally narrower, and the spikelets are smaller than other *P. pratensis* subspecies (Soreng and Barrie 1999; Soreng 2007; Cope and Gray 2009). *P. pratensis* subspecies *angustifolia* is the most likely classification for the reference genotype, although the infraspecies structure is complex and the subspecies genetically and morphologically grade into one another (Soreng and Barrie 1999; Soreng 2007; Cope and Gray 2009).

The remaining *Poa* population genotypes did not survive long enough for detailed morphological identification. We identified the remaining genotypes, and confirmed the morphological IDs, using phylogenetic inference with three commonly used loci (ETS, ITS, TLF). The reference genome was identified as *P. pratensis* by all three loci (Figures S1-3). Seven of the 8 genotypes in the *Poa* population panel were identified as *P. pratensis* by two of the three loci (ITS and ETS; Figures S1-2; Table S1) and held an unresolved position within the subgenus *Poa* in the third tree (TLF; Figure S3). The eighth population panel genotype was identified as *P. compressa* L. by all three loci. Phylogenetic identification thus supports our morphological identification of the reference genome as *P. pratensis*.

4.2 Genome size and ploidy estimation

The reference individual was estimated to be octoploid given a genome size estimate of 3,525 Mbp and chromosome count of 54, assuming a basic chromosome number of $x = 7$ and a loss of two chromosomes (Figure S4; Table S5; Avdulov 1931; Phylogenty Working Group 2001). Further cytological studies are required to understand whether the chromosome loss is due to deletion or rearrangement. Our genome size estimate falls within the large range of genome sizes reported for *P. pratensis*, 2 to 9 pg/1C (Eaton *et al.* 2004; Huff and Bara 1993; Barcaccia *et al.* 1997; Raggi *et al.* 2015).

We also estimated the genome size of four of the eight population panel individuals. Genome size ranged from 3,248 to 4,856 Mbp with genotypes from the same population having similar genome sizes (Table S5). The substantial range in genome size variation in the population panel is not unexpected as *P. pratensis* is a polyploid series with common aneuploidy (Huff 2010). Given the range in the population panel, it is likely the genotypes have different chromosome counts and ploidy.

4.3 Genome assembly

Error-corrected PacBio reads (100 Gb; 70X coverage) were assembled into 27,953 contigs. The contig assembly was oriented and further scaffolded using a Bionano optical map resulting in 118 primary scaffolds and 10 alternate scaffolds (Table 2.1).

The assembly is approximately 173% of the genome size (Table 2.1). Completeness of the assembly was assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO) and the LTR Assembly Index (LAI). The assembly contains 99% of the expected conserved genes (BUSCOs), 98% of which were duplicated, and a LAI value of 25.8 indicates the transposable element assembly is also complete (Ou *et al.* 2018). Given the assembled genome size is approximately two-times the size of the estimated genome size and nearly all detected BUSCOs are duplicated, two unphased haplotypes are likely present in the assembly. Additionally, the high rate of duplicated BUSCOs may also be due to similarity among *Poa* subgenomes.

Table 2.1: **Assembly statistics.**

Variable	Description
Scaffolds	118
Contigs	8,391
Estimated genome size	3.521 Gbp
Assembled genome size	6.09 Gbp
Scaffold N50	65,127,037 bp
Scaffold L50	31
Contig N50	1,095,498 bp
Contig L50	1548
Longest scaffold	177,118,352 bp
Scaffolds $>$ 1 Mb	110
Scaffolds $>$ 10 Mb	98
Average scaffold length	51,622,171 bp
Average length of gaps	44,233 bp
Complete BUSCOs	99.2%
LAI	25.8

4.4 Genome annotation

We identified 256,281 gene models, approximately 32K per subgenome assuming octoploidy, using a hybrid gene prediction pipeline that combined *ab initio* gene models with direct evidence annotations. Phylostrata demonstrated approximately 13% of the gene models are species-specific, which is higher than would be expected from orphan genes alone (Arendsee *et al.* 2014). Since the phylostratr program uses full proteomes from Uniprot to classify genes to their phylostrata, and there is lack of high-quality representative genomes for this clade, we observed an excess of species-specific genes. This demonstrates the important gap a *P. pratensis* reference genome fills in the green tree of life.

Transposable elements were comprehensively annotated using EDTA (Ou *et al.* 2019) and found to compose 58% of the genome. More specifically, Class I LTR retrotransposons and Class II DNA transposons comprise 36% and 15% of the genome, respectively. At the level of superfamily, the RLG (*Ty3*) LTR retrotransposon superfamily was the most common at 18% of the genome.

4.5 Application of the reference genome

The reference genome contains multiple unphased haplotypes, and care should be taken in analyses that require genotypes or allele frequencies. Briefly, we discuss an alternative framework for estimating allele frequencies and potential pitfalls. Diploid genotypes (AA, Aa, aa) should not be called, as at least two haplotypes are assembled for many reference positions. Instead, we utilized an approach in which we randomly sampled a read from each position (Green *et al.* 2010). The randomly sampled read can then be used to calculate population allele frequencies and pairwise genetic distance matrices that are unbiased to sequencing depth or ploidy (Green *et al.* 2010; van der Valk *et al.* 2021; Pečnerová *et al.* 2021). Although we don't detect a bias due to ploidy or chromosome count in our analyses (see below), these factors should always be considered in interpretation of results.

4.6 Population genetics of North American Poa

Here, we demonstrate the effectiveness of the reference genome and a single-read genotyping approach in the estimation of population structure, using PCA and nucleotide diversity.

A PCA was run separately for all *Poa* genotypes, using 74,876 sites, and only *P. pratensis* genotypes, using 140,458 sites. The single-read genotypes were generated three times for the same set of sites and demonstrated similar results. We present the results for one run here. In the PCA with all *Poa* samples, most genetic variation was explained by species (27.9%) followed by population (16.2%; Figure 2.1A). *P. compressa* is distantly related to *P. pratensis* (Figure S1-S3) therefore we would expect the first principal component (PC) to separate by species. The second PC separates the *P. pratensis* genotypes in the Colorado population from two Manitoba *P. pratensis* genotypes (Figure 2.1A), while genotypes from the Colorado population remain clustered. The third principal component further separates the three *P. pratensis* populations.

The *P. pratensis*-only PCA demonstrates similar results with the first PC (24.6%) separating the Colorado genotypes from the two genotypes from Manitoba (Figure S5). The second PC (15.8%) separates the two genotypes from Manitoba and separates one Colorado genotype from the cluster.

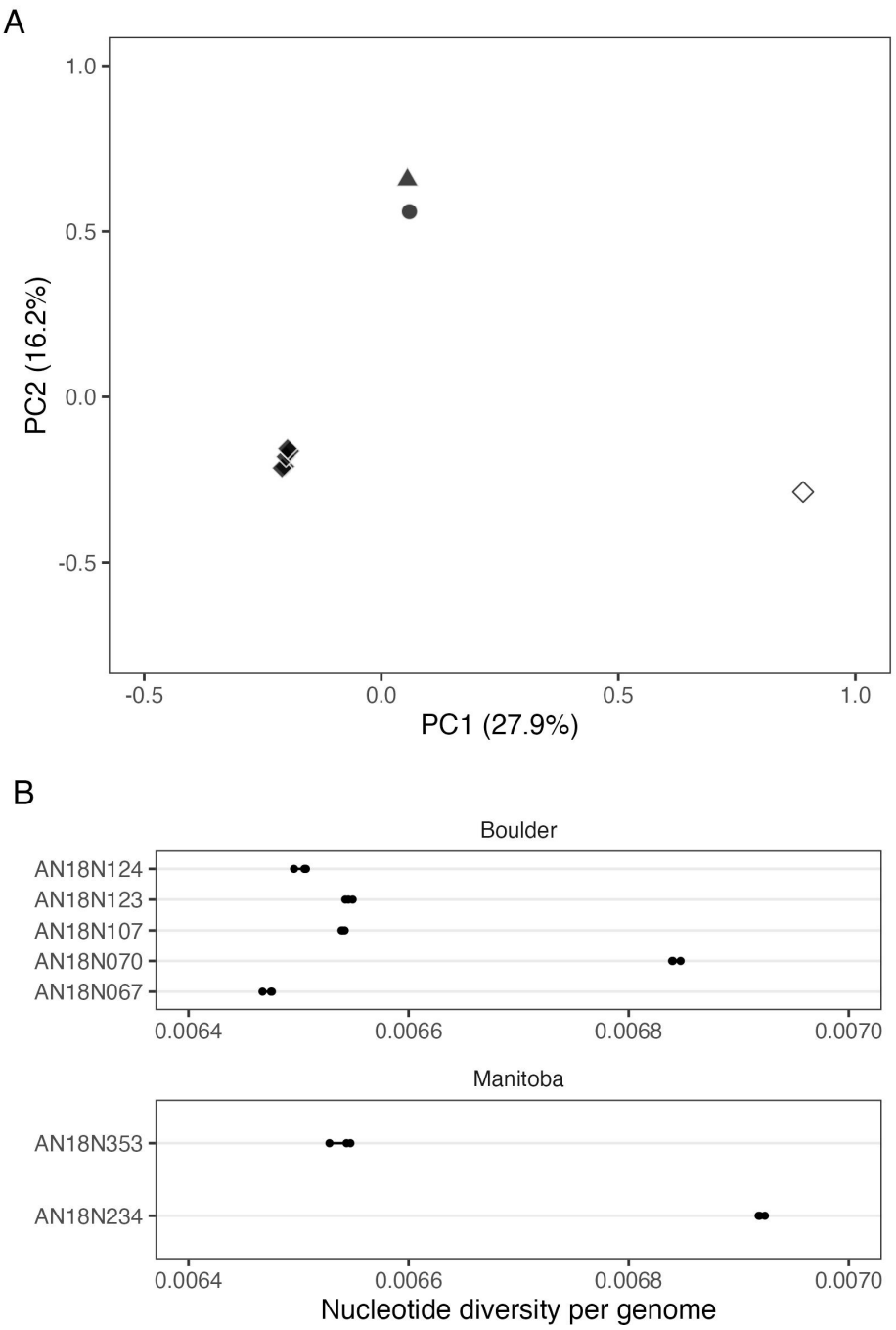

Figure 2.1: **Population structure of *Poa* and nucleotide diversity in *P. pratensis*.** (A) The first two PCs of a PCA of all sequenced *Poa* genotypes. The percent of genetic variation explained by each PC is reported in parenthesis on each axis. Sample locations are indicated by shape (circle = Argyle, Manitoba, triangle = Tolstoi, Manitoba, diamond = Boulder, Colorado) and species are colors (white = *P. compressa*, black = *P. pratensis*). (B) Mean nucleotide diversity per genome for only *P. pratensis* genotypes. Mean diversity of each run is plotted as a black circle for all genotypes.

These results suggest North American *P. pratensis* populations are genetically differentiated and exhibit population structure, rather than being highly homogeneous or clonal. Our results support previous findings of population divergence in Northern Great Plains populations (Dennhardt *et al.* 2016).

To further understand the structure of genetic diversity across *P. pratensis* populations and the clustering within the Colorado population, we estimated nucleotide diversity per genome using 20,149,358 sites. Single-read genotypes were randomly drawn and nucleotide diversity was calculated three times with little variation between runs (Figure 2.1B; average variation between runs = $2.85x10^{-11}$). Mean diversity across *P. pratensis* genotypes is high (π = 0.0066, SD = 0.00017), which is consistent with previous studies of *P. pratensis* (Bonos and Huff 2013; Raggi *et al.* 2015; Bushman *et al.* 2013; Honig *et al.* 2018, 2012). The range of mean nucleotide diversity per genome within the Colorado population (0.0065 - 0.0068) and between the Manitoba genotypes (0.065 - 0.0069) is large, suggesting high within-population diversity.

5. Conclusion

Poa pratensis is a globally popular turfgrass species used in lawns and recreation areas. Despite its economic value, progression of molecular tools to aid breeding has been slow compared to other turfgrasses as a result of polyploidy and apomixis (Bushman and Warnke 2013). Utilizing long read technology and a Bionano optical map, we have assembled and annotated the first high quality *P. pratensis* reference genome. We demonstrated the utility and application of the reference genome by evaluating the genetic diversity and population structure of wild North American *Poa*. As a result, we provided the first estimate of nucleotide diversity in *P. pratensis*.

Since our initial manuscript submission and preprint, Robbins *et al.* (2023) have published the genome of *P. annua*, a distantly related *Poa* species known as a weed and turfgrass worldwide. Future analyses, beyond the scope of this paper, comparing the two genomes will likely be fruitful for understanding the global success of *P. pratensis* and *P. annua*. As such, the *P. pratensis* reference genome and annotation will serve as an important resource in the study of bluegrasses.

6. Data availability

The genome assembly and annotation are available from the European Nucleotide Archive (ENA) under BioProject PRJEB51672. The raw Illumina sequence data for the *Poa* population panel is available from NCBI Sequence Read Archive (SRA) under BioProject ID PRJNA730042. The code for the entirety of assembly, annotation, and population genetic analyses is documented at `https://github.com/phillipsar2/poa_genome`.

7. Acknowledgments

Thank you to Dr. Chrissy McAllister and Bess Bookout for sharing samples collected with permission from Nature Conservancy Canada properties and Lynn Riedel for collection of samples with permission from the City of Boulder Open Space and Mountain Parks. The Texas Advanced Computing Center (TACC) at The University of Texas at Austin, HPC@ISU equipment at Iowa State University (partially funded by NSF under MRI grant number 1726447) provided HPC resources that have contributed to the research results reported within this paper. We thank Dr. Kevin Fengler (for providing assembly instructions) and Dr. Gina Zastrow-Hayes (for establishing sequencing contracts), of Corteva Agriscience for their help in this project. A.R. Phillips would like to thank Andrew L. Murray for his support throughout the duration of this project. Additionally, the authors would like to thank our *Andropogon gerardi* reference plant for being contaminated with *Poa* and Felix Andrews for his alleged role in the happy accident that led to this work. Finally, thank you to Bob Ross for inspiring a generation of scientists to persevere.

8. Funding

This project was funded by the National Science Foundation (NSF) grant number 1822330. HPC resources at TACC were partially funded by NSF under MRI grant number 1726447.

9. Supplement

Supplementary Information is available with the published manuscript at https://doi.org/10.1093/g3journal/jkad073.

Chapter 3: Variant calling in polyploids for population and quantitative genetics

Accepted to Applications in Plant Sciences April 2024

Alyssa Phillips[1,2]

[1]Department of Evolution and Ecology, University of California Davis

[2]Center for Population Biology, University of California Davis

1. Abstract

Advancements in genome assembly and sequencing technology have made whole genome sequence (WGS) data and reference genomes accessible to study polyploid species. The genome-wide coverage and greater marker density provided by WGS data, compared to popular reduced-representation sequencing approaches, can greatly improve our understanding of polyploid species and polyploid biology. However, biological features that make polyploid species interesting also pose challenges in read mapping, variant identification, and genotype estimation. Accounting for characteristics, like allelic dosage uncertainty, homology between subgenomes, and variance in chromosome inheritance mode, in variant calling can reduce errors. Here, I discuss the challenges of variant calling in polyploid WGS data and discuss where potential solutions can be integrated into a standard variant calling pipeline.

2. Background

Recent progress in genome assembly and sequencing technology has increased accessibility to study the genomics of polyploids, or organisms that have experienced whole genome duplication and have more than two sets of chromosomes (Formenti *et al.* 2022; Gladman *et al.* 2023). Notably,

improvements in long-read sequencing and the accuracy of scaffolding technology have enabled the assembly of highly heterozygous and polyploid reference genomes at a chromosome-scale (Kyriakidou *et al.* 2018; Hotaling *et al.* 2023). In parallel, the cost of short-read sequencing has continued to decline causing whole genome resequencing of polyploid populations to become increasingly feasible (Fuentes-Pardo and Ruzzante 2017). As polyploidy is a critical character of cancer cells, common in fish, amphibians, and insects, and ubiquitous in the plant kingdom, including many economically important crops, the extension of modern genomics technologies to polyploid systems is important for our broader understanding of medicine, and biodiversity, agriculture (Udall and Wendel 2006; Wood *et al.* 2009a; Zack *et al.* 2013; One Thousand Plant Transcriptomes Initiative 2019; Román-Palacios *et al.* 2021; David 2022). These advances have already begun to improve our understanding of the origins of polyploid species (Bertioli *et al.* 2019; Edger *et al.* 2019; Goeckeritz *et al.* 2023), genome reorganization and stabilization after polyploidization (Chen *et al.* 2020; Bohutínská *et al.* 2021; Wang *et al.* 2022; Session and Rokhsar 2023), and the role of polyploidy in adaptation of wild and domesticated species(Hollister *et al.* 2012; Chen *et al.* 2021; Lovell *et al.* 2021; Ebadi *et al.* 2023; Hämälä *et al.* 2023). Nevertheless, these studies have only scratched the surface of polyploid biology.

Population and quantitative genetics particularly benefit from the availability of reference genomes and whole genome sequence (WGS) data. These fields use variable loci, loci with two or more alleles segregating in a population, to study the genetic composition of populations and complex traits over space and time in response to selection, genetic drift, mutation, and migration. WGS data in combination with a reference genome offers genome-wide coverage and the ability to identify variable loci, also referred to as variants, at a higher density than reduced representation sequencing (RRS) approaches. RRS approaches, such as genotype-by-sequencing (GBS) and restriction site-associated DNA sequencing (RADseq), are currently used in the majority of polyploid population and quantitative genetics studies due to their comparatively low cost and the growing number of user-friendly software packages for analysis (Poland and Rife 2012). RRS approaches are useful for sampling a portion of the genome to, for example, characterize population structure or com-

plete quantitative trait locus (QTL) analysis. However, RRS does not have high enough marker density for genome-wide analyses central to studying patterns of selection, identifying the genetic basis of adaptive traits, and genomic prediction (Tiffin and Ross-Ibarra 2014; Lowry *et al.* 2017, but see de Bem Oliveira *et al.* 2020). Additionally, WGS data improves the detection of structural variants (SVs) and transposable elements (TEs), although both are still challenging even in diploid systems (Ewing 2015; Baduel *et al.* 2019; Mahmoud *et al.* 2019; Cooke *et al.* 2022; Ramakrishnan *et al.* 2022). Detection and inclusion of SVs and TEs are important because they affect gene expression and function and are signatures of the stabilization and reorganization of the genome post-polyploidization (Lisch 2013; Kosugi *et al.* 2019).

The improvement in variant detection offered by WGS data is useful only when variants can be confidently called and genotypes accurately estimated. Typical sources of error in diploid variant calling include sequencing errors, misalignment of reads to the reference genome, misassembly of the reference genome, and natural structural variation (Li 2014; Mahmoud *et al.* 2019; Lou and Therkildsen 2022). Polyploidy exacerbates these sources of error and introduces additional challenges due to the associated characteristics like large haploid genome sizes, homology between subgenomes, genome fractionation, and elevated polymorphism (Bennett and Leitch 2011; Page and Udall 2015; Blischak *et al.* 2018). As a result, there may be higher variant calling errors in polyploids. Errors in the variant calling pipeline will subsequently be carried into all downstream analyses leading to misestimation of metrics like allele frequencies, heterozygosity, and linkage.

Universal solutions to reduce errors in variant calling are challenging to identify as polyploids are not a uniform group. Polyploids are generally categorized as allopolyploids, which form through hybridization of two or more species, or autopolyploids, which derive from genome doubling of a single species. Further, they can be described by their chromosome inheritance patterns. Allopolyploids have disomic inheritance, like diploids where chiasma for between only homologous chromosomes, and autopolyploids have polysomic chromosome inheritance, where there is no preferential pairing among chromosomes and chiasmata may form between more than two homologous chromosomes (Stift *et al.* 2008). However, the rate of preferential pairing and

chromosome inheritance mode may vary across the genome in allo- and autopolyploids depending on the relatedness amongst subgenomes and the time since polyploidization (Stebbins 1947; Mason and Wendel 2020). This distinction between inheritance modes is important because even low rates of recombination between subgenomes can bias allele frequencies to be more homozygous than expected (Meirmans and Van Tienderen 2013). Polyploids may additionally vary in haploid genome size, mating system, repeat content, and degree of diploidization, all of which may impact variant calling and genotype estimation.

In this review, I identify significant challenges of variant calling in polyploid WGS data and, where available, propose potential solutions that can be integrated into standard variant calling pipelines (Fig. 2.1, Appendix 7.1, reviewed in Van der Auwera *et al.* 2013; De Summa *et al.* 2017; Fuentes-Pardo and Ruzzante 2017; Therkildsen and Palumbi 2017; O'Leary *et al.* 2018; Lou *et al.* 2021). The scope of this discussion is limited to WGS data aligned to the study species' reference genome, although aspects of this discussion may apply to RRS and reference-free approaches. Additionally, I focus on the identification of single nucleotide variants (SNVs) as well as small SVs ($<$ 50 bp) that can be identified by some polyploid variant calling software (Cooke *et al.* 2022). As the genomics of polyploids is a rapidly growing area of research, established best practices are limited. By highlighting barriers in variant calling, I aim to raise readers' awareness of potential sources of error and motivate the innovation of new and effective solutions.

3. Challenges to variant calling in polyploid systems

3.1 Resource requirements scale with genome size

The foremost barrier to polyploid genomics remains the cost of sequencing and high-performance computing (HPC) resources for analysis. Sequencing cost increases with both haploid genome size and ploidy level while computational costs primarily scale with haploid genome size. Sequencing large genomes is expensive as more sequencing runs are required to reach a target coverage, or the genome-wide average number of reads sequenced for a given site. For example, Chen *et al.*

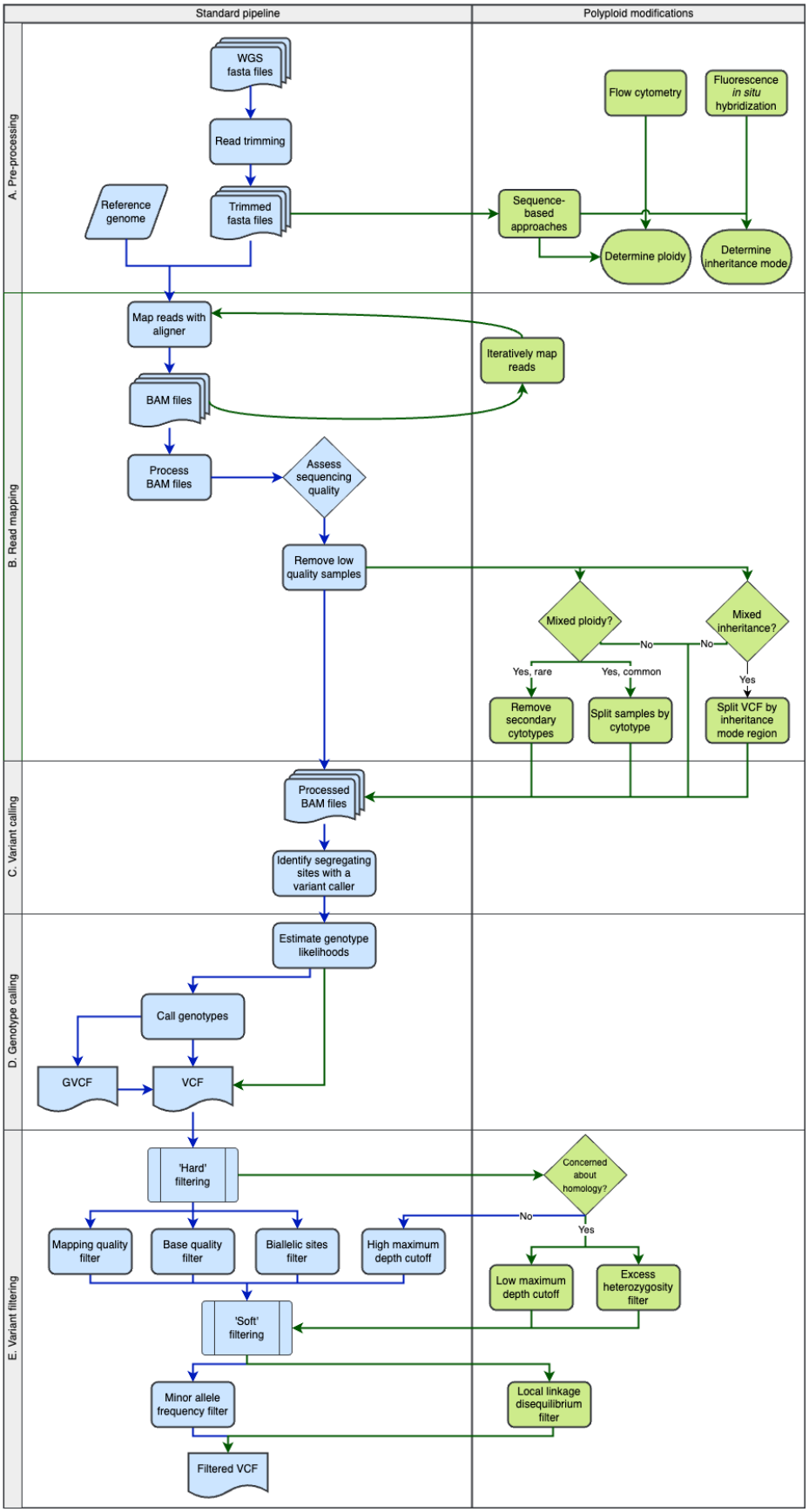
Standard pipeline
Polyploid modifications
A. Pre-processing
B. Read mapping
C. Variant calling
D. Genotype calling
E. Variant filtering
WGS fasta files
Read trimming
Reference genome
Trimmed fasta files
Flow cytometry
Fluorescence in situ hybridization
Sequence-based approaches
Determine ploidy
Determine inheritance mode
Map reads with aligner
BAM files
Iteratively map reads
Process BAM files
Assess sequencing quality
Remove low quality samples
Mixed ploidy?
Mixed inheritance?
No
No
Yes, rare
Yes, common
Yes
Remove secondary cytotypes
Split samples by cytotype
Split VCF by inheritance mode region
Processed BAM files
Identify segregating sites with a variant caller
Estimate genotype likelihoods
Call genotypes
GVCF
VCF
'Hard' filtering
Concerned about homology?
No
Yes
Mapping quality filter
Base quality filter
Biallelic sites filter
High maximum depth cutoff
Low maximum depth cutoff
Excess heterozygosity filter
'Soft' filtering
Minor allele frequency filter
Local linkage disequilibrium filter
Filtered VCF

Figure 2.1: A standard variant calling pipeline (blue) can be adapted for polyploid systems (modifications in green). (A) Before beginning variant calling, raw sequence data may need trimming to remove adapters and low-quality bases. An effort should be made to determine the ploidy and chromosome inheritance mode of the sequenced genotypes, as this information will be incorporated later in the pipeline. Multiple approaches can be used to determine ploidy and inheritance mode depending on the researcher's skillset. (B) Reads are mapped to the reference genome using an aligner. Binary alignment maps (BAMs) are output from the aligners and processed by adding read groups, removing duplicate reads, and then sorting. Sequencing and alignment quality are assessed so low-quality samples may be identified and removed before variant calling. Samples should be split by ploidy and regions by inheritance mode, if necessary, at this stage. (C) Variants are called (D) and then genotype likelihoods and genotypes are estimated. Variant calling and genotyping are often completed using the same software but can be run separately. Genotype calling can be skipped if genotype likelihoods will be used downstream. A variant call file (VCF) is output if invariant sites are discarded, otherwise the output is a genomic variant call file (GVCF). (E) Variants are filtered first by removing low-quality sites (i.e. hard filtering). Then, variants are filtered to prioritize variants specific to downstream analyses (i.e. soft filtering). A more detailed description of the standard pipeline, including useful polyploid aligners and genotype calling software, is provided in Appendix 7.1.

(2024) have found sequencing the allohexaploid bread wheat genome to 5X coverage currently costs 473 times that of diploid rice and 21 times that of maize, a diploidized paleotetraploid (Gaut and Doebley 1997). This disparity in sequencing cost at low coverage is increased by many existing polyploid genotyping algorithms requiring high coverage to overcome allelic dosage uncertainty, which is the ambiguity in the number of alternate allele copies in polyploid genotypes (Gerard *et al.* 2018; Clark *et al.* 2019; Cooke *et al.* 2022). The minimum coverage requirement to obtain high-confidence genotypes may range from 10 to over 50X depending on the ploidy level and genotyping software, whereas diploids need only 8X coverage (Cooke *et al.* 2022; Jighly 2022). After sequencing has been accomplished, access to HPC is needed for data storage and analysis because the size of sequence alignment files (BAMs) and variant call files (VCFs) produced in the variant calling pipeline scale with genome size and sample size (Muir *et al.* 2016; Weiß *et al.* 2018). Failing to sequence to sufficient coverage or limiting sample size to meet budget constraints may result in insufficient sampling of alleles and rare variants, the misestimation of allele frequencies, and low power in analyses like admixture analysis and genome wide association (Jighly 2022).

3.2 Genome-wide redundancy and elevated polymorphism increase errors in read mapping

Aligning reads to polyploid genomes is challenging because polyploids have an elevated level of polymorphism and multiple occurrences of related sequences (Otto and Whitton 2000; Page and Udall 2015). Both of these biological features violate assumptions of read mapping algorithms that assume divergence among loci is larger than divergence among alleles at a single locus (Musich *et al.* 2021); polymorphism creates an excess of divergence while repeated sequences are too similar. Violation of this assumption results in the incorrect and failed mapping of reads. I will briefly describe how these two biological features may create genotyping errors.

As the density of SNVs and SVs in a locus increases, sequence similarity among alleles declines and reads containing alternate alleles are less likely to align (Nielsen *et al.* 2011; Brandt *et al.* 2015). This is an issue in polyploids as they are expected to have higher diversity than their diploid progenitors due to functional redundancy between subgenomes enabling the accumulation of mutations. Additionally, the post-polyploidization process of fractionation, which is gene loss leading to stabilization of the polyploid genome or diploidization, increases structural variation (Haldane 1933; Otto and Whitton 2000; Ma and Gustafson 2005; Emery *et al.* 2018; Beric *et al.* 2021). As an example in the 1000 Genomes Project (H*omo sapiens*), 18.6% of SNV calls in highly polymorphic *HLA* genes were incorrect due to failed mapping of the alternate allele creating bias towards the reference allele, known as allele bias (Brandt *et al.* 2015). Alternate reads may also fail to align to inversions due to disagreement at the inversion boundaries, and reads mapping to presence-absence variants (PAVs) will fail to align if the reference contains the 'absence' variant (Sun *et al.* 2018; Gui *et al.* 2022). As a result, the reference genotype selected for read mapping and time since whole genome duplication will determine the extent of allele bias and the variants detected. Allele bias will be highest in autopolyploids, where reads are aligned to only one copy of the duplicated genome (see Section 3.4). Allele bias is likely an issue genome-wide, although the effect of increased polymorphism on read mapping has yet to be quantified in a polyploid system.

Analogously, genomic features like loci of common ancestry, repetitive elements, and copy number variants (CNVs) promote mismapping because there are multiple occurrences of similar

sequences across the genome. In autopolyploids, whole genome duplication produces duplicate loci between subgenomes that are indistinguishable immediately after duplication. Whereas in allopolyploids, loci of common ancestry are brought back together by hybridization. Both diploids and polyploids contain repeat dense regions and CNVs caused by small-scale duplications and retrotransposons (Brandt *et al.* 2015). As a result, reads may have equal similarities to multiple positions in the reference genome causing reads to equally map to multiple loci (i.e. multiply mapping reads) or improperly align to a closely related locus (Li *et al.* 2008). The extent of error in read mapping due to these redundant genomic features is dependent on the divergence among the loci of common ancestry, known as homologous loci, the age of the polyploidization event, the divergence between parental genomes, mutation rate, and strength of selection on a given locus. Given these factors, read mapping will be most challenging where loci of common ancestry have not accumulated mutations, such as immediately after whole genome duplication or in genes under purifying selection. Additionally, read mapping may be challenging in recently formed polyploids if purifying selection is relaxed genome-wide post-polyploidization allowing rapid TE expansion (McClintock 1984).

If the errors in read mapping discussed here are not resolved, failed alignment of reads may lead to the undercalling of variants, overestimation of homozygosity, and underestimation of population alternative allele frequencies. The mismapping of reads further exacerbates these issues in addition to creating false variants which could create false signals of allele sharing and alter patterns of genome-wide heterozygosity. This can significantly increase downstream errors in the estimation of population divergence, gene flow, genome-wide diversity, and identification of causal variants in GWAS and selection scans.

3.3 Incomplete or misassembled polyploid reference genomes increase genotyping error

Undetected errors in the assembly of polyploid genomes create genotyping errors similar to homologous loci and SVs. For instance, chimeric subgenome assemblies, where scaffolds from one subgenome are misassembled into another subgenome, cause reads to fail to map at misassem-

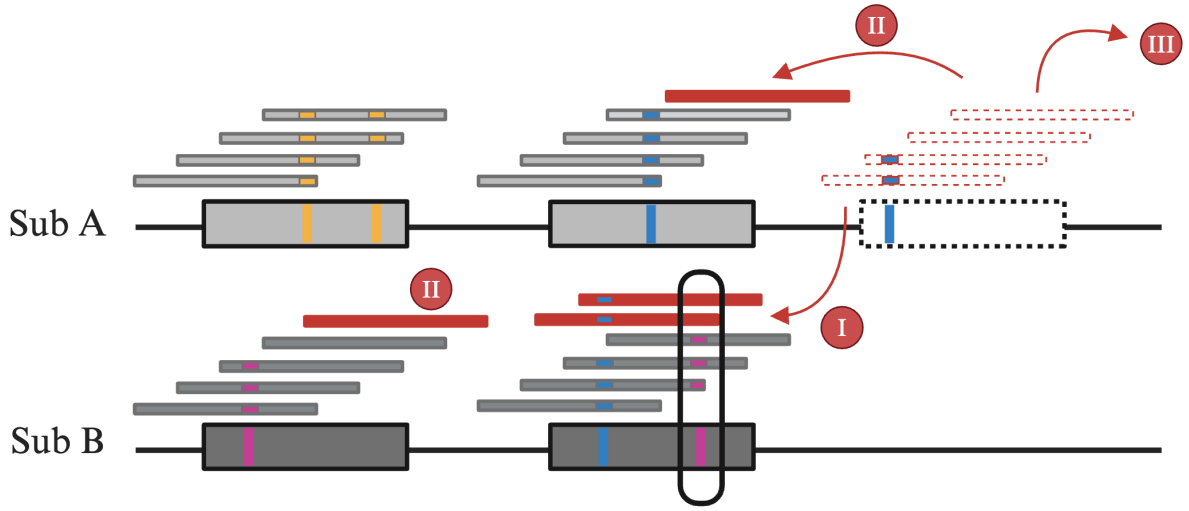

Figure 2.2: A syntenic block between subgenome A and subgenome B in an allotetraploid is depicted. This region in subgenome A contains three genes (light gray) while subgenome B (dark gray) contains two. The genes contain one or two segregating sites, with alleles depicted as yellow, pink, and blue. The assembly of subgenome A is incomplete, missing the farthest right gene (dashed line). Reads that should have aligned to the missing gene (red reads) instead may (I) align to a homolog in subgenome B resulting in a false heterozygote call, (II) map equally to other homologs within or across subgenomes, or (III) fail to align. This figure was created with BioRender.com.

bled scaffold junctions. This leads to genotyping errors at scaffold junctions and incorrect variant positions that impact analyses using linkage information, such as genome scan approaches and estimating runs of homozygosity. In an incomplete reference genome, reads belonging to missing regions will either not align or map to homologous loci (Fig. 2.2). Reads that successfully map to a homolog are likely to be biased toward the reference allele. However, if reads with the alternative allele do align to a homolog, false heterozygotes may be called (Fig. 2.2A). Comprehensively addressing the challenge of poor read mapping caused by low reference genome quality will require continued improvement of the reference genome. As comprehensive reviews on genome assembly are available elsewhere (Zhang *et al.* 2019b; Zhou *et al.* 2022; Gladman *et al.* 2023), I later discuss practical solutions to mitigate these issues and enhance the accuracy of genotyping when using existing genome assemblies.

3.4 Allele dosage cannot be determined if ploidy and inheritance mode are unknown

Determining the allele dosage, the number of reference and alternate alleles, present at each sequenced site for a given individual is imperative for accurate genotyping. In diploids, the reference genome is ideally phased, meaning the maternal and paternal copy of each chromosome is assembled so each chromosome in the assembly has two 'haplotypes' (Gladman *et al.* 2023). All reads are aligned to only one of the two haplotypes and, as a result, the possible genotype values at a site are 0, 1, and 2 corresponding to the number of alternate alleles. The range of potential genotypes for a polyploid is less clear as there are multiple factors to consider: ploidy level, chromosome inheritance mode, and the reference genome quality. This is because autopolyploids and allopolyploids have distinct reference genome structures (Kihara and Ono 1926; Kyriakidou *et al.* 2018; Zhang *et al.* 2019b). Ideally, autopolyploid assemblies are phased so all copies (i.e. haplotypes) of the genome are assembled. Assuming the autopolyploid has no preferential pairing amongst chromosomes (i.e. complete polysomic inheritance), all reads should be aligned to only one haplotype, similar to diploids, and the maximum allele dosage would be equal to the ploidy (Fig. 2.3B). In allopolyploids, the paternal and maternal haplotypes of each ancestral subgenome are assembled and reads are aligned to one haplotype of each subgenome simultaneously (Fig. 2.3A). Here, the maximum allele dosage would be the ploidy divided by the number of subgenomes. As an example, consider the allotetraploid switchgrass (*Panicum virgatum*) reference genome, which contains two phased subgenomes (Napier *et al.* 2022). Switchgrass is a mixed-ploidy species composed of tetraploids ($2n = 4x$) and octoploids ($2n = 8x$). As both subgenomes were successfully assembled, Napier *et al.* (2022) concurrently aligned reads to one haplotype of each subgenome and called genotypes for the tetraploid and octoploid samples as diploid (0, 1, 2) and tetraploid genotype values (0, 1, 2, 3, 4), respectively. If the switchgrass reference genome was not phased, the ploidy of each sample was unknown, or if it was unclear whether the species is allo- or autopolyploid, the correct allele dosage could not be determined. Unknown or incorrect allele dosage can result in the misestimation of allele frequencies and heterozygosity, similar to co-dominant markers like AFLPs (Dufresne *et al.* 2014).

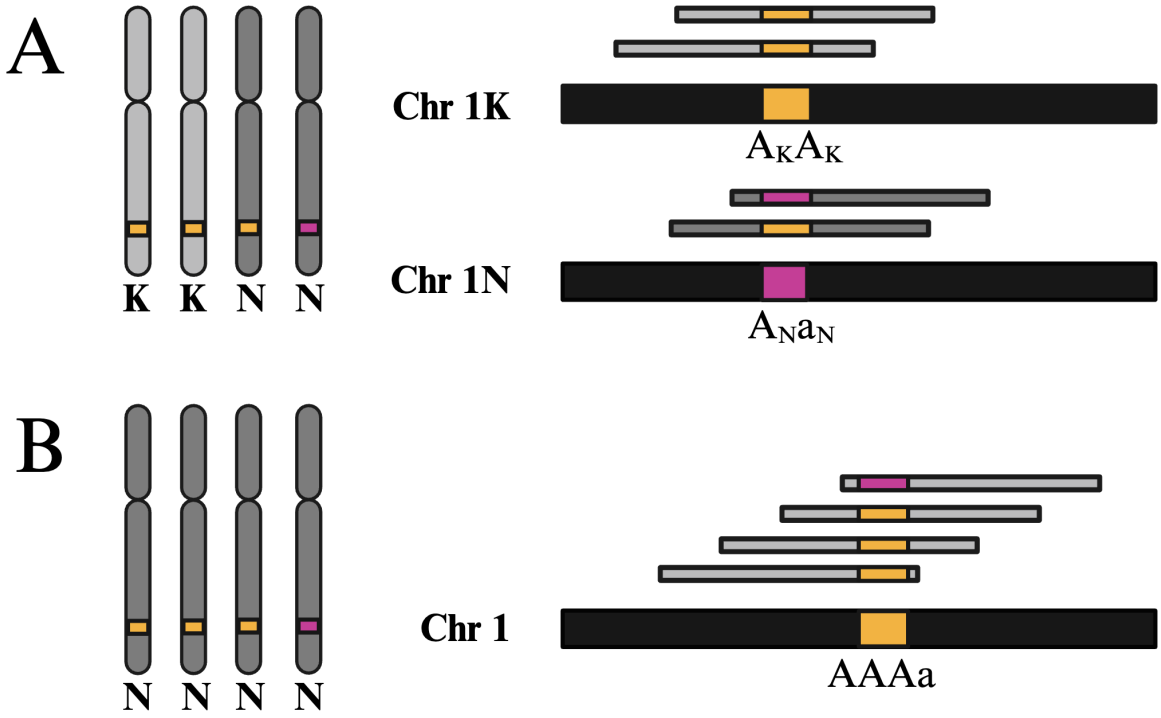

Figure 2.3: Read mapping and the called allele dosage in allo- and autopolyploids differs due to the structure of the reference genome. Reads (gray) are shown aligning the reference genome (black) with alleles for the focal variant in pink or yellow. (A) In an allotetraploid with two subgenomes (subgenome K in light gray and subgenome N in dark gray), reads are mapped to one haplotype of each parental subgenome, and diploid genotypes are called. (B) In an autotetraploid with no preferential pairing, all reads are mapped to a single haplotype. Here, reads are aligned to a haplotype carrying the yellow A allele at the focal variant.

3.5 Existing tools cannot account for further biological complexity

The reach of polyploid population and quantitative genetics is limited by further biological complexities. Commonly, populations may be mixed-ploidy, meaning they contain genotypes of varying ploidy levels (Kolář *et al.* 2017). Additionally, inheritance mode may vary along the genome (Allendorf *et al.* 2015). Variance in inheritance mode occurs because, following whole genome duplication, it is likely that all homologs pair together, and thus experience polysomic inheritance. However, over time, sequence divergence among homologous chromosomes may lead to preferential pairing and allow the return of disomic inheritance in some regions of the genome

(Allendorf *et al.* 2015). In addition to mixed ploidy and inheritance mode, polyploid species may have multiple origins (Holloway *et al.* 2006; Soltis *et al.* 2009) and often hybridize (Alix *et al.* 2017), which makes population and quantitative genetics challenging. It is difficult to develop a variant calling pipeline that considers this complexity in a meaningful way while also producing genotypes that can be used in existing downstream tools. For example, existing software packages that estimate genotypes for mixed-ploidy populations require separate estimations for each ploidy (Blischak *et al.* 2018; Gerard *et al.* 2018; Clark *et al.* 2019; Van der Auwera and O'Connor 2020; Cooke *et al.* 2021). In multi-sample variant calling, which incorporates information from multiple samples to improve genotype estimates, the separation of samples by ploidy reduces the utility and power of this approach (Liu *et al.* 2013). The mismapping of reads further exacerbates these issues in addition to creating false variants which could create false signals of allele sharing and alter patterns of genome-wide heterozygosity. Alternative approaches such as estimating genotypes at the same allele dosage for all cytotypes will result in underestimating heterozygous genotypes for higher ploidy levels and inaccurate allele frequency estimations.

4. Proposed solutions to incorporate polyploid complexity in variant calling

4.1 Balancing sequencing depth and precision may reduce sequencing costs

Careful experimental design, consideration of downstream analysis, and alternative genotyping approaches can be leveraged to reduce the cost of working with polyploid WGS data. Although a certain level of sequencing coverage is required to overcome allelic dosage uncertainty, high sequencing depth is not required for all analyses. Jighly (2022) argues that sequencing depth should be selected depending on the research question and analysis plan, in conjunction with the ploidy level, as sequencing depth has diminishing returns. Analyses that require the detection of low-frequency and rare variants, such as inferring novel alleles, will require a higher depth. In contrast, studies examining population structure and differentiation, which rely on common alleles to differentiate groups, may accommodate a lower sequencing depth. Therefore, considering the research

question and analysis plan when determining the target coverage will prevent over-sequencing and extend a budget.

The increased allele dosage uncertainty that comes from low sequencing depth (<10X) can be partially mitigated by the use of genotype likelihoods (GLs) or continuous genotypes in place of categorical genotypes. A GL is the probability of the sequencing data given the possible genotypes. GLs can be directly used in some software or they can be used to infer genotypes. Polyploid-capable software such as GATK, EBG, Updog, and polyRAD (Blischak *et al.* 2018; Gerard *et al.* 2018; Clark *et al.* 2019; Van der Auwera and O'Connor 2020), infer categorical genotypes from GLs. Updog and polyRAD can also estimate continuous genotypes, which are continuous values of the likely allele count (Gerard *et al.* 2018; Clark *et al.* 2019; Njuguna *et al.* 2023). The combination of low-coverage data and GLs or continuous genotypes is becoming increasingly popular in large-scale studies due to its affordability (Korneliussen *et al.* 2014; Grandke *et al.* 2016; Batista *et al.* 2022). Further, GLs and continuous genotypes reduce allelic dosage uncertainty by incorporating genotyping certainty and may be beneficial in moderate or high-coverage sequence data. These alternative genotypes have been shown to provide more accurate estimates than categorical genotypes in numerous population and quantitative genetics analyses (Korneliussen *et al.* 2014; Grandke *et al.* 2016; Gerard 2021b; Shastry *et al.* 2021; Batista *et al.* 2022; Rasmussen *et al.* 2024). Continuous genotypes can be easily integrated into existing software, however, software for downstream population and quantitative genetic analysis with polyploid GLs is still limited.

4.2 Alternative read alignment approaches, genotype callers, and variant filters may reduce errors caused by poor read mapping

Several strategies can be applied to reduce read mapping errors caused by homology, high polymorphism, or low reference genome quality throughout the variant calling pipeline. First, alternative alignment approaches could be applied to improve read mapping and assignment to subgenomes. For example, iterative read mapping is a promising strategy. Here, all reads are mapped to the reference genome but only reads that map to exactly one place in the genome (i.e. uniquely mapped

reads) are retained. Then, a pseudo-reference genome is generated by replacing variable sites with the alternate alleles from the uniquely mapping reads, reads are re-mapped to the pseudo-reference, and, again, only uniquely mapped reads are retained (Rozowsky *et al.* 2011; Xu *et al.* 2020). When applied to maize whole-genome bisulfite sequencing data to reduce mapping bias, this approach was found to increase the detection of methylated cytosines by 5% (Xu *et al.* 2020). Alternatively, the software WASP alters the mapped reads, instead of the reference genome, to have the opposite allele. The altered reads are remapped and only kept if they map in the same location (van de Geijn *et al.* 2015). Both iterative read mapping approaches are particularly useful for reducing the number of multiply mapping reads and reducing false heterozygotes. Other alternative read mapping solutions have been developed specifically to identify subgenome differences in allopolyploids by either comparing polymorphisms to modern diploid progenitors (Mithani *et al.* 2013; Page *et al.* 2013; Peralta *et al.* 2013; Khan *et al.* 2016) or competitively mapping reads between subgenomes (Page and Udall 2015). The former approach requires knowledge of the diploid progenitors and the ladder approach has limited benefits if both subgenomes of the allopolyploid are assembled. As a result, iterative read mapping is currently the most promising solution for improving read mapping.

Second, a genotype caller that considers allele bias and read-mapping errors could be used in addition to iterative read mapping to reduce the extent of false heterozygous or homozygous calls. The popular polyploid genotype caller Updog estimates the degree of allele bias simultaneously with genotype estimation (Gerard *et al.* 2018). No other polyploid genotype callers, to my knowledge, account for allele bias. Emerging solutions to reducing genotyping error from poor read mapping include the modification of variant calling algorithms developed for CNVs (Layer *et al.* 2014; Prodanov and Bansal 2022) or ancient DNA (Günther and Nettelblad 2019). For example, the software ancient DNA software, snpAD (Prüfer 2018), iteratively estimates genotype probabilities and r, the frequency at which the sequences are sampled from the reference allele at heterozygous sites, to account for reference bias. Although snpAD is not currently able to estimate polyploid GLs, algorithms such as this have the potential to improve uncertainty in polyploid

genotyping caused by poor read mapping.

Third, variant filters may be applied to exclude any remaining false-positive variants and genotyping errors caused by mismapped reads. Filters that have been used for this purpose discriminate variants by mapping quality, maximum coverage, and local linkage disequilibrium (Fig. 2.1E). I will briefly review these filters. To begin, mapping quality is a commonly applied 'hard' filter (Appendix 7.1) and is estimated as the phred-scaled probability a read is aligned to the wrong position. It is determined by the number of mismatches in the alignment while considering the quality of all other possible alignments (Li *et al.* 2008). Reads that map equally to multiple homologs (i.e. multiply mapping reads; Fig. 2.2C) will have a mapping quality of zero and be removed in standard variant filtering pipelines. Typically, a mapping quality is applied to remove reads below a quality of 10 to 40 (Van der Auwera *et al.* 2013; Korneliussen *et al.* 2014; Puritz *et al.* 2014), which is equivalent to removing sites with greater than 0.01-10% probability of alignment error.

Exclusion of mismapped reads could also be accomplished using a maximum coverage filter. If reads improperly map to a given site, the site would have higher coverage than expected given the average genome-wide coverage (Fig. 2.2A). Applying this logic, maximum depth filters are commonly used to exclude false heterozygotes in repetitive regions of the genome (Li 2014), but these are generally set too high to exclude reads mismapping in non-repetitive regions. In polyploid systems, this approach has been adopted to set a low per-site maximum depth threshold using models of expected read depth (Bohutínská *et al.* 2021; Korani *et al.* 2021; Phillips *et al.* 2023; Yu *et al.* 2023), although the efficacy of this filter and the best read depth model has not been determined.

A promising novel approach to exclude false-positive variants is to leverage the expectation that two true neighboring variants may have correlated allele frequencies within a population, known as local linkage disequilibrium (LD; Bukowski *et al.* 2018). Variants in low LD with nearby variants would be excluded. This approach may also be useful in resolving the alignment of multiply-mapping reads by measuring local LD at each site the read is aligned to determine the most likely position, although this is likely computationally time-consuming and is yet to be tested

in diploids or polyploids. LD estimates are biased by genotype uncertainty, which is exaggerated in polyploid genotypes, but this can be remedied with the recently developed R package ldsep that provides computationally efficient methods to estimate LD from diploid and polyploid GLs (Gerard 2021a,b).

Other variant filters, such as the removal of loci with excess heterozygosity or departure from Hardy-Weinberg equilibrium (HWE), have also been explored for removing false-positive variants. If the mismapped reads carry the alternate allele, these filters may be able to remove false heterozygous sites (Keller *et al.* 2013; McKinney *et al.* 2017; Ahrens *et al.* 2020; Clark *et al.* 2022; Bohutínská *et al.* 2023). Researchers should exercise caution in applying filters that assume populations are at HWE because many biological factors, such as a non-panmictic population structure, small population sizes, and genetic drift, cause deviations from HWE (Pearman *et al.* 2022). Polyploidy itself deviates from diploid HWE therefore methods developed in Gerard (2022b) and Gerard (2023) should be used to properly account for unknown rates of double reduction (Gerard 2022a).

4.3 Information on ploidy, chromosome inheritance mode, and reference quality can be integrated to determine allele dosage

Investment in the determination of ploidy level and inheritance mode of the reference genotype and sequenced genotypes towards the beginning of an experiment, although potentially time-intensive, is strongly recommended to identify the correct allele dosage. Traditionally, ploidy and inheritance mode have been determined using chromosome squashes (Goldblatt and Lowry 2011), flow cytometry (Bennett and Leitch 2011; Pellicer and Leitch 2020) and fluorescence *in situ* hybridization (FISH), where fluorescent probes are used to label specific DNA sequences to identify and track chromosome pairings (Szadkowski *et al.* 2010; Chester *et al.* 2013; Parra-Nunez *et al.* 2020). Unfortunately, these approaches are time-intensive, require specialized equipment, and are an uncommon skill set. With the advent of next-generation sequencing, there has been a large research effort to determine ploidy from allele frequency distributions (Margarido and Heckerman 2015;

Augusto Corrêa Dos Santos *et al.* 2017; Weiß *et al.* 2018; Ranallo-Benavidez *et al.* 2020; Soraggi *et al.* 2022; Sun *et al.* 2023; Viruel *et al.* 2023; Gaynor *et al.* 2024). Sequence-based approaches have also begun to be explored for determining inheritance mode. One approach proposed by Scott *et al.* (2023) compares estimated allelic depth distributions to those expected under disomic and tetrasomic inheritance, although this approach is sensitive to demography. Other approaches include leveraging divergence among genes duplicated during whole genome duplication to detect windows of disomic or tetrasomic inheritance along the genome (Campbell *et al.* 2019; Scott *et al.* 2023) and the joint inference of inheritance mode and demography (Blischak *et al.* 2023; Roux *et al.* 2023) or genotypes (discussed in Section 4.4; Gerard *et al.* 2018; Clark *et al.* 2019). Sequence-based approaches are exceptionally promising for determining ploidy and inheritance mode in systems where flow cytometry and FISH are especially difficult or impossible, such as succulents and herbarium samples.

In cases where allele dosage cannot be determined because the ploidy and inheritance mode of the reference genotype is unknown, the reference scaffolds could be filtered to only one copy of syntenic scaffolds for read mapping. If the scaffolds can be assigned into subgenomes, such as in an allopolyploid, scaffolds would be filtered within each subgenome. This is a strategy applied in many systems with contig assemblies (Hellsten *et al.* 2013; Neale *et al.* 2022; Phillips *et al.* 2023). The risk of aligning to only a subset of scaffolds is that a large proportion of reads may not align and variants could be underdetected.

4.4 Current accepted practices for navigating polyploid data with additional biological complexity

Existing tools are limited in their ability to incorporate complexity such as mixed ploidy and inheritance mode, but variant calling pipelines have the potential to accommodate this additional axis of diversity in several ways. For datasets with mixed ploidy, the current best practice is to call genotypes separately for each cytotype, if using a joint genotyping approach (Napier *et al.* 2022; Bohutínská *et al.* 2023; De Luca *et al.* 2023). In cases where the secondary cytotype is rare or un-

dersampled, it is advisable to exclude the minority cytotypes from the study because variability in downstream analyses attributable to cytotype differences may not be detectable with small sample sizes. If multiple cytotypes are included in the study, it should be noted that polyploid genotypes have inherently different expected variations in allele frequencies which can significantly impact downstream analyses (Faske 2023). Similarly to mixed-ploidy analyses, allele dosage should be specified per-site in species with mixed inheritance modes. If the regions of the genome with polysomic inheritance are known, the per-site specification can be accomplished with any polyploid genotype caller, although this has rarely been applied outside of the Salmonids (Campbell *et al.* 2019). Alternatively, if polysomic regions are known, sites could be filtered to include only disomic or polysomic regions (Bourret *et al.* 2013). In the majority of cases, the rate of preferential pairing or the regions undergoing polysomic inheritance will be unknown. Here, the genotype calling software Updog (Gerard *et al.* 2018) and polyRAD (Clark *et al.* 2019) may be useful as their approaches determine inheritance mode during genotype estimation. Updog accomplishes this by simultaneously estimating genotypes and the rate of preferential pairing in a population, assuming bivalent pairing only. Comparatively, polyRAD determines inheritance mode by estimating genotypes for all possible user-specified genotypes and then uses a χ^2 statistic to determine the best genotype at each site. The polyRAD approach is particularly useful as it allows both ploidy and inheritance mode to vary among genotypes. There is no current best practice for mixed inheritance mode among these approaches, but they should be considered as even low rates of polysomic inheritance can affect allele frequencies across subgenomes (Meirmans and Van Tienderen 2013). Consequently, careful consideration is required when analyzing populations with biological complexity beyond polyploidy.

5. Conclusion

Complex polyploid biology may produce errors in read mapping, variant calling, and genotyping. The extent of error often depends on the quality of the reference genome and biological reasons like the age of the polyploidization event, extent of fractionation, divergence between parental

genomes, and strength of selection at a given locus. As such, bioinformatic solutions can be selectively applied to resolve sources of error prevalent in a given polyploid system. In Figure 2.1, I summarize where existing solutions can be integrated into a standard variant calling pipeline. The study of polyploid genomes is a growing field and, as such, there may be additional solutions in active development.

Further improvements to variant calling in polyploids will require focused research in three primary areas: evaluation of variant filters, development of downstream software that incorporates genotype uncertainty, and high-throughput estimation of ploidy and inheritance mode. First, empirical studies evaluating the efficacy of variant filters are needed to understand when their application is appropriate and which thresholds are effective. It is equally as important to set a threshold that excludes low-quality variants while also not over-filtering the data, as variant classes important in downstream analyses may be unintentionally excluded (Linck and Battey 2019; Pearman *et al.* 2022). Second, continued development of population and quantitative genetics software that utilize GLs is needed (Korneliussen *et al.* 2014; Grandke *et al.* 2016; Gerard 2021b; Shastry *et al.* 2021; Batista *et al.* 2022; Rasmussen *et al.* 2024). The adoption of GLs to reduce sequencing costs is likely to be limited until more user-friendly software becomes available. Theory and tools are also lacking for the analysis of mixed-ploidy and mixed-inheritance mode datasets. Third, continued development of methods for high throughput estimation of ploidy and inheritance mode is greatly needed. While there has been substantial development in this area (see Section 4.3), the majority of approaches still necessitate ample ground truthing (Gaynor *et al.* 2024).

Emerging technologies may have the potential to improve variant detection. Long-read sequencing data overcomes many read mapping challenges as the extended read length increases the information available to determine the best alignment (Chen *et al.* 2024). Similar to short-read sequencing, long-read sequencing is increasingly cost-effective and accurate (De Coster *et al.* 2021; Kim *et al.* 2024). Additionally, pan-genomic approaches, such as haplotype graphs and sequence variation groups, have recently been applied in polyploid systems to detect a diversity of SVs as well as multiallelic sites (Gordon *et al.* 2020; Bayer *et al.* 2021; Della Coletta *et al.*

2021; Lovell *et al.* 2021; Wang *et al.* 2022). The adoption of the variant calling practices reviewed here, continued investment in the assembly of polyploid reference genomes, and early adoption of novel genomic tools will enhance contemporary population and quantitative genetics studies in polyploids.

6. Acknowledgments

The author would like to thank Jeffrey Ross-Ibarra, Elli Cryan, Natasha Dhamrait, Regina Fairbanks, Samantha Snodgrass, Tyler Kent, Elisabeth Forrestel, Jennifer Gremer, Michelle Gaynor, and Trevor Faske for their feedback on this manuscript. The ForBio "Population genetics of polyploids, from theory to practice" workshop provided a useful space to discuss these ideas. This project was funded by the National Science Foundation (NSF) grant numbers 1822330 and 1934384.

7. Appendix

7.1 A brief overview of variant calling

In diploid and polyploid systems, variant calling involves a series of qualitative decisions that depend on the biology of the study system and data quality. A variant calling pipeline, as described here, includes the alignment of reads to the reference genome, variant calling, genotype estimation, and variant filtering. Consideration of ploidy in downstream analyses has been well-reviewed elsewhere (Dufresne *et al.* 2014; Meirmans *et al.* 2018; Ackiss and Balao 2020; Bohutínská *et al.* 2023). Here, I aim to provide an overview of a general variant calling pipeline to support discussions of where this pipeline may be improved for polyploid systems. I provide citations for commonly used software where relevant.

To begin, reads are mapped to a reference genome using a short-read aligner to generate the sequence alignment maps (SAMs) or binary alignment maps (BAMs). The aligner is selected depending on the read length, sequencing method, and divergence of the sequenced sample from

the reference genome (Altmann *et al.* 2012; Bak *et al.* 2021; Musich *et al.* 2021). The Burrow-Wheeler aligner (BWA-MEM and BWA-MEM2) is a highly popular short-read aligner (Liu *et al.* 2013; Md *et al.* 2019). Additionally, the best practice is to use a reference genome closely related to your samples of interest, but how closely related your reference genome needs to be to your samples will depend on the divergence between species and amongst populations (Günther and Nettelblad 2019). For example, in a *Zea mays* RNA-seq study, as much as one-half of alleles with increased gene expression were not detected when reads from the inbred line, B73, were mapped to the reference of a second inbred line, Mo17, because *Z. mays* has high nucleotide diversity and structural variation (Zhan *et al.* 2021).

The SAMs or BAMs are processed to remove duplicate reads and add read groups, which provide an improved evaluation of sequencing and alignment quality but have limited effect on variant detection (Ebbert *et al.* 2016). SAMtools (Danecek *et al.* 2021) and GATK (De Summa *et al.* 2017; Van der Auwera and O'Connor 2020) provide useful guidelines and pipelines for effectively processing the alignment files. The sequencing and alignment quality should be evaluated for attributes such as mapping quality, the percent of reads mapping, and coverage before variant calling (Nielsen *et al.* 2011). Although this can be accomplished with custom scripts, software like Qualimap provides a user-friendly evaluation of sequence quality (García-Alcalde *et al.* 2012; Okonechnikov *et al.* 2016). If the quality is poor, reads may need to be trimmed to remove adapters or low-quality bases and re-mapped (Sewe *et al.* 2022). Trimmomatic (Sewe *et al.* 2022) and fastp (Chen *et al.* 2018; Chen 2023) efficiently detect and trim a wide variety of adaptor sequences.

Variants are then identified using a variant caller, which determines whether a particular site in a sequenced sample is different from the reference genome. Many variant callers, such as GATK (Van der Auwera and O'Connor 2020), were developed for human genomes and have been adopted for use with highly repetitive plant genomes. Before genotype calling, sites that are fixed across sequenced samples, known as invariant sites, are often excluded to improve computational efficiency. It should be noted that the inclusion of invariant sites is important for many population and quantitative genetics analyses, such as the estimation of nucleotide diversity and demographic

history, and they can be added back into the pipeline after variant calling. Genotypes are subsequently called where the most likely genotype is estimated based on the number of references and alternate reads that are mapped to a given site (Nielsen *et al.* 2011).

The same software is often used for both variant calling and genotyping. Importantly, the genotype caller selected should be able to estimate polyploid genotypes. Polyploid genotype callers have been sufficiently compared and reviewed elsewhere (Grandke *et al.* 2016; Blischak *et al.* 2018; Clark *et al.* 2019; Cooke *et al.* 2022). Briefly, polyploid variant and genotype callers that can be applied to whole genome sequence data include GATK, freebayes (Garrison and Marth 2012), EBG (Blischak *et al.* 2018), Updog (Gerard *et al.* 2018), polyRAD (Clark *et al.* 2019), and Octopus (Cooke *et al.* 2021). Additionally, GATK, freebayes, and Octopus can identify small structural variants under 50 bp (Cooke *et al.* 2022). Each polyploid genotype considers different aspects of polyploid biology in their estimation, and as such, researchers should select the caller that fits the biology of their study system the best. For example, Updog considers allele bias (see Section 4.2) and preferential pairing in genotype estimation, while polyRAD considers per-site variance in inheritance mode (see Section 4.4, Gerard *et al.* 2018; Clark *et al.* 2019). Notably, Updog, polyRAD, and Octopus support binomial priors, which are considered 'informative' priors because they assume genotypes follow HWE, unlike GATK which uses uniform priors that assume genotypes have equal probabilities (McKenna *et al.* 2010; Gerard *et al.* 2018; Clark *et al.* 2019; Cooke *et al.* 2021). Additionally, polyRAD offers additional informative priors that consider population structure and mapping populations (Clark *et al.* 2019). Genotype callers and priors should be carefully selected as genotypes will be heavily influenced by the priors at low sequencing coverage (Clark *et al.* 2019).

Finally, variants are filtered to remove sites with false-positive variants and low-confidence genotypes. This is often accomplished using custom scripts, GATK, VCFtools (Danecek *et al.* 2021), or several other packages. Variant filtering is often grouped into two parts: 'hard' and 'soft' filtering (De Summa *et al.* 2017). In hard filtering, sites that fail to pass a set of quality controls are removed to reduce the likelihood of falsely identifying them as polymorphic. The quality

controls may include mapping quality, base quality, depth, and strand bias (defined in Van der Auwera and O'Connor (2020)). Biallelic sites are typically selected when hard filtering, regardless of ploidy, as most empirical and theoretical population and quantitative genetics assume only two alleles (but see Karlin (1990); Balding and Nichols (1995); Ferretti *et al.* (2018); Broman *et al.* (2019) for examples of multi-allelic approaches). After hard filtering, soft filters are applied to prioritize variants specific to downstream analyses, often ad-hoc. For example, a minor allele frequency filter is a soft filter often applied to exclude sites with rare variants. Thresholds for hard and soft filtering are user-defined and formal testing of the significance of a given threshold is uncommon. Researchers often derive thresholds from those previously applied within their study system, review articles (Van der Auwera *et al.* 2013; Clevenger *et al.* 2015), or, less commonly, those tested in an empirical study (Linck and Battey 2019; Pearman *et al.* 2022). Importantly, researchers should take care not to over-filter their datasets as many population and quantitative genetics analyses can be biased by datasets where particular variant classes were excluded (Linck and Battey 2019; Pearman *et al.* 2022).

Conclusion

The collective findings from my dissertation underscore the dynamic nature of genomic research in polyploid plant species. In Chapter 1, I investigated the role of mixed-ploidy in the adaptation of *Andropogon gerardi* Vitman, revealing the influence of polyploidy on growth and physiology. We find mixed-ploidy is a product of recurrent polyploidization where each individual of the higher ploidy level is a new polyploid, or neopolyploid. The study highlights the environment-dependent effect of polyploidy and the need to consider polyploidy in conservation and restoration. Chapter 2 presents a serendipitous assembly of a reference genome for *Poa pratensis*. The accidental assembly provides a valuable resource for turfgrass breeding and showcases the potential for unexpected breakthroughs in genomics. Finally, in Chapter 3 I review the challenges and potential solutions in variant calling for polyploid species, emphasizing the importance of addressing polyploid biology rather than ignoring it. I highlight the ongoing efforts to improve genomic methods for polyploids and propose a literature-informed variant calling pipeline. Overall, my dissertation contributes to a deeper understanding of the interplay between polyploidy, environmental adaptation, and plant evolution.

References

Ackiss, A. S. and F. Balao, 2020 Diving in uncharted waters: An updated genetics toolkit highlights the challenges of polyploidy in landscape genomics analyses. Mol. Ecol. Resour. **20**: 841–843.

Ahrens, C. W., E. A. James, A. D. Miller, F. Scott, N. C. Aitken, *et al.*, 2020 Spatial, climate and ploidy factors drive genomic diversity and resilience in the widespread grass *Themeda triandra*. Mol. Ecol. **29**: 3872–3888.

Akaike, H., 1974 A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**: 716–723.

Albertini, E., G. Marconi, G. Barcaccia, L. Raggi, and M. Falcinelli, 2004 Isolation of candidate genes for apomixis in *Poa pratensis* L. Plant Mol. Biol. **56**: 879–894.

Alix, K., P. R. Gérard, T. Schwarzacher, and J. S. P. Heslop-Harrison, 2017 Polyploidy and interspecific hybridization: Partners for adaptation, speciation and evolution in plants. Ann. Bot. **120**: 183–194.

Allendorf, F. W., S. Bassham, W. A. Cresko, M. T. Limborg, L. W. Seeb, *et al.*, 2015 Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. J. Hered. **106**: 217–227.

Almeida, T. E. and B. S. Santos Leal, 2024 Recurrent allopolyploidy and its implications for conservation in vascular plants: a commentary on 'Population genomics of the *Isoetes appalachiana* (Isoetaceae) complex supports a "diploids-first" approach to conservation'. Annals of Botany **133**: i–ii.

Altmann, A., P. Weber, D. Bader, M. Preuss, E. B. Binder, *et al.*, 2012 A beginners guide to SNP calling from high-throughput DNA-sequencing data. Hum. Genet. **131**: 1541–1554.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. **215**: 403–410.

Amadeu, R. R., L. A. Lara, P. Munoz, and A. A. Garcia, 2020 Estimation of molecular pairwise relatedness in autopolyploid crops. G3: Genes, Genomes, Genetics **10**: 4579–4589.

Arendsee, Z., J. Li, U. Singh, A. Seetharam, K. Dorman, *et al.*, 2019 phylostratr: A framework for phylostratigraphy. Bioinformatics **35**: 3617–3627.

Arendsee, Z. W., L. Li, and E. S. Wurtele, 2014 Coming of age: Orphan genes in plants. Trends in Plant Sci. **19**: 698–708.

Aspinwall, M. J., D. B. Lowry, S. H. Taylor, T. E. Juenger, C. V. Hawkes, *et al.*, 2013 Genotypic variation in traits linked to climate and aboveground productivity in a widespread C4 grass: Evidence for a functional trait syndrome. New Phytologist **199**: 966–980.

Augusto Corrêa Dos Santos, R., G. H. Goldman, and D. M. Riaño-Pachón, 2017 ploidyNGS: Visually exploring ploidy with next generation sequencing data. Bioinformatics **33**: 2575–2576.

Avdulov, N., 1931 Kario-sistematicheskoye issledovaniye semeystva zlakov (Karyosystematic studies in the grass family). Bull. Appl. Bot. Gen. Pl. Breed., Leningrad **44**: 1–428.

Avolio, M. L., J. M. Beaulieu, and M. D. Smith, 2013 Genetic diversity of a dominant C4 grass is altered with increased precipitation variability. Oecologia **171**: 571–581.

Awada, T., L. E. Moser, W. H. Schacht, and P. E. Reece, 2002 Stomatal variability of native warm-season grasses from the nebraska sandhills. Canadian Journal of Plant Science **82**: 349–355.

Bachle, S. and J. B. Nippert, 2021 Microanatomical traits track climate gradients for a dominant C4 grass species across the Great Plains, USA. Annals of Botany **127**: 451–459.

Baduel, P., L. Quadrana, B. Hunter, K. Bomblies, and V. Colot, 2019 Relaxed purifying selection in autopolyploids drives transposable element over-accumulation which provides variants for local adaptation. Nat. Commun. **10**: 5818.

Bak, A., D. Bodziony, G. Migdałek, C. S. Pareek, and K. Żukowski, 2021 Evaluation of analytical protocols of alignment mapping tools using high throughput next-generation genome sequencing data. Translational Research in Veterinary Science **3**: 61.

Balding, D. J. and R. A. Nichols, 1995 A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica **96**: 3–12.

Bao, W., K. K. Kojima, and O. Kohany, 2015 Repbase update, a database of repetitive elements in

eukaryotic genomes. Mobile DNA **6**: 1–6.

Barcaccia, G., A. Mazzucato, A. Belardinelli, M. Pezzotti, S. Lucretti, *et al.*, 1997 Inheritance of parental genomes in progenies of *Poa pratensis* L. from sexual and apomictic genotypes as assessed by RAPD markers and flow cytometry. Theor. and Appl. Genet. **95**: 516–524.

Batista, L. G., V. H. Mello, A. P. Souza, and G. R. A. Margarido, 2022 Genomic prediction with allele dosage information in highly polyploid species. Theor. Appl. Genet. **135**: 723–739.

Bayer, P. E., A. Scheben, A. A. Golicz, Y. Yuan, S. Faure, *et al.*, 2021 Modelling of gene loss propensity in the pangenomes of three *Brassica* species suggests different mechanisms between polyploids and diploids. Plant Biotechnol. J. **19**: 2488–2500.

Beaulieu, J. M., I. J. Leitch, S. Patel, A. Pendharkar, and C. A. Knight, 2008 Genome size is a strong predictor of cell size and stomatal density in angiosperms. New Phytol. **179**: 975–986.

Belling, J., 1925 The origin of chromosomal mutations in *Uvularia*. Journal of Genetics **15**: 245–266.

Bennett, M. D. and I. J. Leitch, 2011 Nuclear DNA amounts in angiosperms: Targets, trends and tomorrow. Ann. Bot. **107**: 467–590.

Benson, E. J. and D. C. Hartnett, 2006 The role of seed and vegetative reproduction in plant recruitment and demography in tallgrass prairie. Plant Ecology **187**: 163–178.

Beric, A., M. E. Mabry, A. E. Harkess, J. Brose, M. E. Schranz, *et al.*, 2021 Comparative phylogenetics of repetitive elements in a diverse order of flowering plants (Brassicales). G3 **11**.

Bertioli, D. J., J. Jenkins, J. Clevenger, O. Dudchenko, D. Gao, *et al.*, 2019 The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. Nat. Genet. **51**: 877–884.

Bilton, T. P., S. K. Sharma, M. R. Schofield, M. A. Black, J. M. Jacobs, *et al.*, 2024 Construction of relatedness matrices in autopolyploid populations using low-depth high-throughput sequencing data. Theoretical and Applied Genetics **137**: 1–18.

Bird, K. A., R. VanBuren, J. R. Puzey, and P. P. Edger, 2018 The causes and consequences of subgenome dominance in hybrids and recent polyploids. New Phytologist **220**: 87–93.

Blischak, P. D., L. S. Kubatko, and A. D. Wolfe, 2018 SNP genotyping and parameter estimation

in polyploids using low-coverage sequencing data. Bioinformatics **34**: 407–415.

Blischak, P. D., M. Sajan, M. S. Barker, and R. N. Gutenkunst, 2023 Demographic history inference and the polyploid continuum. Genetics **224**.

Bohutínská, M., M. Alston, P. Monnahan, T. Mandáková, S. Bray, *et al.*, 2021 Novelty and convergence in adaptation to whole genome duplication. Mol. Biol. Evol. **38**: 3910–3924.

Bohutínská, M., J. Vlček, P. Monnahan, and F. Kolář, 2023 Population genomic analysis of diploid-autopolyploid species. Methods Mol. Biol. **2545**: 297–324.

Bonos, S. A. and D. R. Huff, 2013 Cool-season grasses: Biology and breeding. Turfgrass: Biology, use, and management **56**: 591–660.

Bourret, V., M. P. Kent, C. R. Primmer, A. Vasemägi, S. Karlsson, *et al.*, 2013 SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of atlantic salmon (*Salmo salar*). Mol. Ecol. **22**: 532–551.

Bradnam, K. R., J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, *et al.*, 2013 Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species. GigaScience **2**: 2047–217X.

Brandt, D. Y. C., V. R. C. Aguiar, B. D. Bitarello, K. Nunes, J. Goudet, *et al.*, 2015 Mapping bias overestimates reference allele frequencies at the *HLA* genes in the 1000 Genomes Project Phase I data. G3 **5**: 931–941.

Broman, K. W., D. M. Gatti, P. Simecek, N. A. Furlotte, P. Prins, *et al.*, 2019 R/qtl2: Software for mapping quantitative trait loci with high-dimensional data and multiparent populations. Genetics **211**: 495–502.

Brown, W. L., 1939 Chromosome complements of five species of *Poa* with an analysis of variation in *Poa pratensis*. Am. J. Bot. **26**: 717–723.

Brůna, T., K. J. Hoff, A. Lomsadze, M. Stanke, and M. Borodovsky, 2021 BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics and Bioinf. **3**.

Bukowski, R., X. Guo, Y. Lu, C. Zou, B. He, *et al.*, 2018 Construction of the third-generation *Zea*

mays haplotype map. Gigascience **7**: 1–12.

Bushman, B. S. and S. E. Warnke, 2013 Genetic and genomic approaches for improving turfgrass. Turfgrass: Biology, Use, and Management **56**: 683–711.

Bushman, B. S., S. E. Warnke, K. L. Amundsen, K. M. Combs, and P. G. Johnson, 2013 Molecular markers highlight variation within and among Kentucky bluegrass varieties and accessions. Crop Sci. **53**: 2245–2254.

Cabi, E., R. J. Soreng, and L. Gillespie, 2017 Taxonomy of *Poa jubata* and a new section of the genus (Poaceae). Turk. J. Bot. **41**: 404–415.

Cabi, E., R. J. Soreng, L. Gillespie, and N. Amiri, 2016 *Poa densa* (Poaceae), an overlooked Turkish steppe grass, and the evolution of bulbs in *Poa*. Willdenowia **46**: 201 – 211.

Campbell, M. A., M. C. Hale, G. J. McKinney, K. M. Nichols, and D. E. Pearse, 2019 Long-term conservation of ohnologs through partial tetrasomy following whole-genome duplication in Salmonidae. G3 **9**: 2017–2028.

Carrier, L. and K. S. Bort, 1916 The history of Kentucky bluegrass and white clover in the United States. Agron. J. **8**: 256–267.

Casler, M. D. and R. R. Duncan, 2003 Turfgrass biology, genetics, and breeding .

Caudle, K., L. Johnson, S. Baer, and B. Maricle, 2014 A comparison of seasonal foliar chlorophyll change among ecotypes and cultivars of *Andropogon gerardii* (Poaceae) by using nondestructive and destructive methods. Photosynthetica **52**: 511–518.

Chen, L., J. Luo, M. Jin, N. Yang, X. Liu, *et al.*, 2022 Genome sequencing reveals evidence of adaptive variation in the genus zea. Nature Genetics **54**: 1736–1745.

Chen, S., 2023 Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. Imeta **2**.

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

Chen, X., C. Tong, X. Zhang, A. Song, M. Hu, *et al.*, 2021 A high-quality *Brassica napus* genome reveals expansion of transposable elements, subgenome evolution and disease resistance. Plant

Biotechnol. J. **19**: 615–630.

Chen, Y., W. Wang, Z. Yang, H. Peng, Z. Ni, *et al.*, 2024 Innovative computational tools provide new insights into the polyploid wheat genome. aBIOTECH .

Chen, Z. J., A. Sreedasyam, A. Ando, Q. Song, L. M. De Santiago, *et al.*, 2020 Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. Nat. Genet. **52**: 525–533.

Cheng, H., G. T. Concepcion, X. Feng, H. Zhang, and H. Li, 2021 Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods **18**: 170–175.

Chester, M., M. J. Lipman, J. P. Gallagher, P. S. Soltis, and D. E. Soltis, 2013 An assessment of karyotype restructuring in the neoallotetraploid *Tragopogon miscellus* (Asteraceae). Chromosome Res. **21**: 75–85.

Chin, C.-S., P. Peluso, F. J. Sedlazeck, M. Nattestad, G. T. Concepcion, *et al.*, 2016 Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods **13**: 1050–1054.

Clark, L. V., A. E. Lipka, and E. J. Sacks, 2019 polyRAD: Genotype calling with uncertainty from sequencing data in polyploids and diploids. G3 **9**: 663–673.

Clark, L. V., W. Mays, A. E. Lipka, and E. J. Sacks, 2022 A population-level statistic for assessing mendelian behavior of genotyping-by-sequencing data from highly duplicated genomes. BMC Bioinformatics **23**: 101.

Clevenger, J., C. Chavarro, S. A. Pearl, P. Ozias-Akins, and S. A. Jackson, 2015 Single nucleotide polymorphism identification in polyploids: A review, example, and recommendations. Mol. Plant **8**: 831–846.

Clo, J. and F. Kolář, 2021 Short-and long-term consequences of genome doubling: A meta-analysis. American Journal of Botany **108**: 2315–2322.

Cooke, D. P., D. C. Wedge, and G. Lunter, 2021 A unified haplotype-based method for accurate and comprehensive variant calling. Nat. Biotechnol. **39**: 885–892.

Cooke, D. P., D. C. Wedge, and G. Lunter, 2022 Benchmarking small-variant genotyping in polyploids. Genome Res. **32**: 403–408.

Cope, T. A. and A. J. Gray, 2009 Grasses of the British Isles. Botanical Society of the British Isles.

Corneillie, S., N. De Storme, R. Van Acker, J. U. Fangel, M. De Bruyne, *et al.*, 2019 Polyploidy affects plant growth and alters cell wall composition. Plant Physiology **179**: 74–87.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, *et al.*, 2021 Twelve years of SAMtools and BCFtools. Gigascience **10**.

Darriba, D., G. L. Taboada, R. Doallo, and D. Posada, 2012 jModelTest 2: More models, new heuristics and parallel computing. Nat. Methods **9**: 772–772.

David, K. T., 2022 Global gradients in the distribution of animal polyploids. Proc. Natl. Acad. Sci. U. S. A. **119**: e2214070119.

de Bem Oliveira, I., R. R. Amadeu, L. F. V. Ferrão, and P. R. Muñoz, 2020 Optimizing whole-genomic prediction for autotetraploid blueberry breeding. Heredity **125**: 437–448.

De Coster, W., M. H. Weissensteiner, and F. J. Sedlazeck, 2021 Towards population-scale long-read sequencing. Nat. Rev. Genet. **22**: 572–587.

De Luca, D., E. Del Guacchio, P. Cennamo, L. Paino, and P. Caputo, 2023 Genotyping-by-sequencing provides new genetic and taxonomic insights in the critical group of *Centaurea tenorei*. Front. Plant Sci. **14**: 1130889.

De Summa, S., G. Malerba, R. Pinto, A. Mori, V. Mijatovic, *et al.*, 2017 GATK hard filtering: Tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. BMC Bioinformatics **18**: 119.

DeKeyser, E. S., L. A. Dennhardt, and J. Hendrickson, 2015 Kentucky bluegrass (*Poa pratensis*) invasion in the Northern Great Plains: A story of rapid dominance in an endangered ecosystem. Invasive Plant Sci. Manage. **8**: 255–261.

Della Coletta, R., Y. Qiu, S. Ou, M. B. Hufford, and C. N. Hirsch, 2021 How the pan-genome is changing crop genomics and improvement. Genome Biol. **22**: 3.

Dennhardt, L. A., E. S. DeKeyser, S. A. Tennefos, and S. E. Travers, 2016 There is no evidence of geographical patterning among invasive Kentucky bluegrass (*Poa pratensis*) populations in the Northern Great Plains. Weed Sci. **64**: 409–420.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, *et al.*, 2013 Star: Ultrafast universal RNA-seq aligner. Bioinformatics **29**: 15–21.

Doležel, J., J. Greilhuber, and J. Suda, 2007 Estimation of nuclear DNA content in plants using flow cytometry. Nat. Protocols **2**: 2233–2244.

Doyle, J. J. and J. E. Coate, 2019 Polyploidy, the nucleotype, and novelty: The impact of genome doubling on the biology of the cell. International Journal of Plant Sciences **180**: 1–52.

Dudchenko, O., S. S. Batra, A. D. Omer, S. K. Nyquist, M. Hoeger, *et al.*, 2017 De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science **356**: 92–95.

Dufresne, F., M. Stift, R. Vergilino, and B. K. Mable, 2014 Recent progress and challenges in population genetics of polyploid organisms: An overview of current state-of-the-art molecular and statistical tools. Mol. Ecol. **23**: 40–69.

Durand, N. C., M. S. Shamim, I. Machol, S. S. Rao, M. H. Huntley, *et al.*, 2016 Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Systems **3**: 95–98.

Duvaud, S., C. Gabella, F. Lisacek, H. Stockinger, V. Ioannidis, *et al.*, 2021 Expasy, the swiss bioinformatics resource portal, as designed by its users. Nucleic Acids Res. **49**: W216–W227.

Eaton, T., J. Curley, R. Williamson, and G. Jung, 2004 Determination of the level of variation in polyploidy among Kentucky bluegrass cultivars by means of flow cytometry. Crop Sci. **44**: 2168–2174.

Ebadi, M., Q. Bafort, E. Mizrachi, P. Audenaert, P. Simoens, *et al.*, 2023 The duplication of genomes and genetic networks and its potential for evolutionary adaptation and survival during environmental turmoil. Proc. Natl. Acad. Sci. U. S. A. **120**: e2307289120.

Ebbert, M. T. W., M. E. Wadsworth, L. A. Staley, K. L. Hoyt, B. Pickett, *et al.*, 2016 Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC Bioinformatics **17 Suppl 7**: 239.

Edger, P. P., T. J. Poorten, R. VanBuren, M. A. Hardigan, M. Colle, *et al.*, 2019 Origin and evolution of the octoploid strawberry genome. Nat. Genet. **51**: 541–547.

Ellinghaus, D., S. Kurtz, and U. Willhoeft, 2008 LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinf. **9**: 1–14.

Emery, M., M. M. S. Willis, Y. Hao, K. Barry, K. Oakgrove, *et al.*, 2018 Preferential retention of genes from one parental genome after polyploidy illustrates the nature and scope of the genomic conflicts induced by hybridization. PLoS Genet. **14**: e1007267.

Estep, M. C., M. R. McKain, D. Vela Diaz, J. Zhong, J. G. Hodge, *et al.*, 2014 Allopolyploidy, diversification, and the Miocene grassland expansion. Proceedings of the National Academy of Sciences **111**: 15149–15154.

Ewing, A. D., 2015 Transposable element detection from whole genome sequence data. Mob. DNA **6**: 24.

Faske, T., 2023 2 does not equal 4: Variance dissimilarities in mixed-ploidy genomic data cause irregular patterns in PCA and other clustering analyses. Botany.

Fernandes Gyorfy, M., E. R. Miller, J. L. Conover, C. E. Grover, J. F. Wendel, *et al.*, 2021 Nuclear–cytoplasmic balance: Whole genome duplications induce elevated organellar genome copy number. The Plant Journal **108**: 219–230.

Ferretti, L., A. Klassmann, E. Raineri, S. E. Ramos-Onsins, T. Wiehe, *et al.*, 2018 The neutral frequency spectrum of linked sites. Theor. Popul. Biol. **123**: 70–79.

Formenti, G., K. Theissinger, C. Fernandes, I. Bista, A. Bombarely, *et al.*, 2022 The era of reference genomes in conservation genomics. Trends Ecol. Evol. **37**: 197–202.

Francis, D., M. S. Davies, and P. W. Barlow, 2008 A strong nucleotypic effect on the cell cycle regardless of ploidy level. Annals of Botany **101**: 747–757.

Fuentes-Pardo, A. P. and D. E. Ruzzante, 2017 Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. Mol. Ecol. **26**: 5369–5406.

Funk, C. R., J. H. Sang, *et al.*, 1967 Recurrent intraspecific hybridization – A proposed method of breeding Kentucky bluegrass (*Poa pratensis* L.). New Jersey Agricultural Experiment Station Bulletin .

Galliart, M., N. Bello, M. Knapp, J. Poland, P. St Amand, *et al.*, 2019 Local adaptation, genetic divergence, and experimental selection in a foundation grass across the US Great Plains' climate gradient. Global Change Biology **25**: 850–868.

Galliart, M., S. Sabates, H. Tetreault, A. DeLaCruz, J. Bryant, *et al.*, 2020 Adaptive genetic potential and plasticity of trait variation in the foundation prairie grass *Andropogon gerardii* across the US Great Plains' climate gradient: Implications for climate change and restoration. Evolutionary Applications **13**: 2333–2356.

García-Alcalde, F., K. Okonechnikov, J. Carbonell, L. M. Cruz, S. Götz, *et al.*, 2012 Qualimap: Evaluating next-generation sequencing alignment data. Bioinformatics **28**: 2678–2679.

Garnier, E., B. Shipley, C. Roumet, and G. Laurent, 2001 A standardized protocol for the determination of specific leaf area and leaf dry matter content. Functional Ecology pp. 688–695.

Garrison, E. and G. Marth, 2012 Haplotype-based variant detection from short-read sequencing. arXiv .

Gaut, B. S. and J. F. Doebley, 1997 DNA sequence evidence for the segmental allotetraploid origin of maize. Proc. Natl. Acad. Sci. U. S. A. **94**: 6809–6814.

Gaynor, M. L., N. Kortessis, D. E. Soltis, P. S. Soltis, and J. M. Ponciano, 2023 Dynamics of mixed-ploidy populations under demographic and environmental stochasticities.

Gaynor, M. L., J. B. Landis, T. K. O'Connor, R. G. Laport, J. J. Doyle, *et al.*, 2024 nQuack: An R package for predicting ploidal level from sequence data using site-based heterozygosity.

Georganas, E., A. Buluç, J. Chapman, S. Hofmeyr, C. Aluru, *et al.*, 2015 HipMer: An extreme-scale de novo genome assembler. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–11.

Gerard, D., 2021a Pairwise linkage disequilibrium estimation for polyploids. Mol. Ecol. Resour. **21**: 1230–1242.

Gerard, D., 2021b Scalable bias-corrected linkage disequilibrium estimation under genotype uncertainty. Heredity **127**: 357–362.

Gerard, D., 2022a Comment on three papers about Hardy-Weinberg equilibrium tests in autopoly-

ploids. Front. Genet. **13**: 1027209.

Gerard, D., 2022b Double reduction estimation and equilibrium tests in natural autopolyploid populations. Biometrics .

Gerard, D., 2023 Bayesian tests for random mating in polyploids. Mol. Ecol. Resour. **23**: 1812–1822.

Gerard, D., L. F. V. Ferrão, A. A. F. Garcia, and M. Stephens, 2018 Genotyping polyploids from messy sequencing data. Genetics **210**: 789–807.

Gillespie, L. J., A. Archambault, and R. J. Soreng, 2007 Phylogeny of *Poa* (Poaceae) based on trnT–trnF sequence data: major clades and basal relationships. Aliso **23**: 420–434.

Gillespie, L. J., R. J. Soreng, R. D. Bull, S. W. Jacobs, and N. F. Refulio-Rodriguez, 2008 Phylogenetic relationships in subtribe Poinae (Poaceae, Poeae) based on nuclear ITS and plastid trnT-trnL-trnF sequences. Botany (Ottawa) **86**: 938–967.

Gillespie, L. J., R. J. Soreng, E. Cabi, and N. Amiri, 2018 Phylogeny and taxonomic synopsis of *Poa* subgenus *Pseudopoa* (including *Eremopoa* and *Lindbergella*) (Poaceae, Poeae, Poinae). PhytoKeys **111**: 69–101.

Gillespie, L. J., R. J. Soreng, and S. W. Jacobs, 2009 Phylogenetic relationships of Australian *Poa* (Poaceae: Poinae), including molecular evidence for two new genera, *Saxipoa* and *Sylvipoa*. Aust. Syst. Bot. **22**: 413–436.

Giovanny, C.-P., 2016 Genome assisted prediction of quantitative traits using the R package sommer. PLoS ONE **11**: 1–15.

Giussani, L. M., L. J. Gillespie, M. A. Scataglini, M. A. Negritto, A. M. Anton, *et al.*, 2016 Breeding system diversification and evolution in American *Poa* supersect. *Homalopoa* (Poaceae: Poeae: Poinae). Ann. Bot. **118**: 281–303.

Gladman, N., S. Goodwin, K. Chougule, W. Richard McCombie, and D. Ware, 2023 Era of gapless plant genomes: Innovations in sequencing and mapping technologies revolutionize genomics and breeding. Curr. Opin. Biotechnol. **79**: 102886.

Goeckeritz, C. Z., K. E. Rhoades, K. L. Childs, A. F. Iezzoni, R. VanBuren, *et al.*, 2023 Genome

of tetraploid sour cherry (*Prunus cerasus* l.) 'montmorency' identifies three distinct ancestral *Prunus* genomes. Hortic. Res. **10**: uhad097.

Goldblatt, P. and P. P. Lowry, 2011 The Index to Plant Chromosome Numbers (IPCN): Three decades of publication by the Missouri Botanical Garden come to an end. mobt **98**: 226–227.

Gordon, S. P., B. Contreras-Moreira, J. J. Levy, A. Djamei, A. Czedik-Eysenberg, *et al.*, 2020 Gradual polyploid genome evolution revealed by pan-genomic analysis of *Brachypodium hybridum* and its diploid progenitors. Nat. Commun. **11**: 3670.

Grandke, F., P. Singh, H. C. M. Heuven, J. R. de Haan, and D. Metzler, 2016 Advantages of continuous genotype values over genotype classes for GWAS in higher polyploids: A comparative study in hexaploid *Chrysanthemum*. BMC Genomics **17**: 672.

Gray, M. M., P. S. Amand, N. M. Bello, M. B. Galliart, M. Knapp, *et al.*, 2014 Ecotypes of an ecologically dominant prairie grass (*Andropogon gerardii*) exhibit genetic divergence across the US Midwest grasslands' environmental gradient. Molecular Ecology **23**: 6011–6028.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, *et al.*, 2010 A draft sequence of the Neandertal genome. Science **328**: 710–722.

Gui, S., W. Wei, C. Jiang, J. Luo, L. Chen, *et al.*, 2022 A pan-*Zea* genome map for enhancing maize improvement. Genome Biol. **23**: 178.

Günther, T. and C. Nettelblad, 2019 The presence and impact of reference bias on population genomic studies of prehistoric human populations. PLoS Genet. **15**: e1008302.

Haas, B. J., A. L. Delcher, S. M. Mount, J. R. Wortman, R. K. Smith Jr, *et al.*, 2003 Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research **31**: 5654–5666.

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, *et al.*, 2013 De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. Nat. Protocols **8**: 1494–1512.

Haldane, J. B. S., 1933 The part played by recurrent mutation in evolution. Am. Nat. **67**: 5–19.

Hämälä, T., C. Moore, L. Cowan, M. Carlile, D. Gopaulchan, *et al.*, 2023 Impact of whole-genome

duplications on structural variant evolution in the plant genus *Cochlearia*.

Harlan, J. R. and J. M. deWet, 1975 On Ö. Winge and a prayer: The origins of polyploidy. The Botanical Review **41**: 361–390.

Haydu, J. J., A. W. Hodges, and C. R. Hall, 2006 Economic impacts of the turfgrass and lawncare industry in the United States. EDIS **2006**.

Heide, O., 1994 Control of flowering and reproduction in temperate grasses. New Phytol. **128**: 347–362.

Hellsten, U., K. M. Wright, J. Jenkins, S. Shu, Y. Yuan, *et al.*, 2013 Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing. Proc. Natl. Acad. Sci. U. S. A. **110**: 19478–19482.

Hendrickson, J., M. Liebig, J. Printz, D. Toledo, J. Halvorson, *et al.*, 2021 Kentucky bluegrass impacts diversity and carbon and nitrogen dynamics in a Northern Great Plains rangeland. Rangeland Ecol. Manage. **79**: 36–42.

Hollister, J. D., B. J. Arnold, E. Svedin, K. S. Xue, B. P. Dilkes, *et al.*, 2012 Genetic adaptation associated with genome-doubling in autotetraploid *Arabidopsis arenosa*. PLoS Genet. **8**: e1003093.

Holloway, A. K., D. C. Cannatella, H. C. Gerhardt, and D. M. Hillis, 2006 Polyploids with different origins and ancestors form a single sexual polyploid species. Am. Nat. **167**: E88–101.

Honig, J. A., V. Averello, S. A. Bonos, and W. A. Meyer, 2012 Classification of Kentucky bluegrass (*Poa pratensis* L.) cultivars and accessions based on microsatellite (simple sequence repeat) markers. HortScience **47**: 1356–1366.

Honig, J. A., V. Averello, C. Kubik, J. Vaiciunas, B. S. Bushman, *et al.*, 2018 An update on the classification of Kentucky bluegrass cultivars and accessions based on microsatellite (SSR) markers. Crop Sci. **58**: 1776–1787.

Hotaling, S., E. R. Wilcox, J. Heckenhauer, R. J. Stewart, and P. B. Frandsen, 2023 Highly accurate long reads are crucial for realizing the potential of biodiversity genomics. BMC Genomics **24**: 117.

Huff, D. R., 2010 Bluegrasses. In *Fodder crops and amenity grasses*, pp. 345–379, Springer.

Huff, D. R. and J. M. Bara, 1993 Determining genetic origins of aberrant progeny from facultative apomictic Kentucky bluegrass using a combination of flow cytometry and silver-stained RAPD markers. Theor. and Appl. Genet. **87**: 201–208.

Hufford, M. B., A. S. Seetharam, M. R. Woodhouse, K. M. Chougule, S. Ou, *et al.*, 2021 De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science **373**: 655–662.

Hufkens, K., D. Basler, T. Milliman, E. K. Melaas, and A. D. Richardson, 2018 An integrated phenology modelling framework in R: Modelling vegetation phenology with phenor. Methods in Ecology & Evolution **9**: 1–10.

Jighly, A., 2022 When do autopolyploids need poly-sequencing data? Mol. Ecol. **31**: 1021–1027.

Karlin, S., 1990 Levels of multiallelic overdominance fitness, heterozygote excess and heterozygote deficiency. Theor. Popul. Biol. **37**: 129–149.

Kato, A., 1999 Air drying method using nitrous oxide for chromosome counting in maize. Biotech. Histochem. **74**: 160–166.

Kato, A., J. C. Lamb, P. S. Albert, T. Danilova, F. Han, *et al.*, 2011 Chromosome painting for plant biotechnology. In *Plant chromosome engineering*, pp. 67–96, Springer.

Kato, A., J. C. Lamb, and J. A. Birchler, 2004 Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. PNAS **101**: 13554–13559.

Katoh, K. and D. M. Standley, 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. **30**: 772–780.

Kawecki, T. J. and D. Ebert, 2004 Conceptual issues in local adaptation. Ecology letters **7**: 1225–1241.

Keeler, K., B. Kwankin, P. Barnes, and D. Galbraith, 1987 Polyploid polymorphism in *Andropogon gerardii*. Genome **29**: 374–379.

Keeler, K. H., 1990 Distribution of polyploid variation in big bluestem (*Andropogon gerardii*, Poaceae) across the tallgrass prairie region. Genome **33**: 95–100.

Keeler, K. H., 1992 Local polyploid variation in the native prairie grass *Andropogon gerardii*. Am. J. Bot. **79**: 1229–1232.

Keeler, K. H., 2004 Impact of intraspecific polyploidy in *Andropogon gerardii*(Poaceae) populations. Am. Midl. Nat. **152**: 63–74.

Keeler, K. H. and G. A. Davis, 1999 Comparison of common cytotypes of *Andropogon gerardii* (Andropogoneae, Poaceae). Am. J. Bot. **86**: 974–979.

Keilwagen, J., F. Hartung, M. Paulini, S. O. Twardziok, and J. Grau, 2018 Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC Bioinf. **19**: 1–12.

Keilwagen, J., M. Wenk, J. L. Erickson, M. H. Schattat, J. Grau, *et al.*, 2016 Using intron position conservation for homology-based gene prediction. Nucleic Acids Res. **44**: e89–e89.

Keller, I., C. E. Wagner, L. Greuter, S. Mwaiko, O. M. Selz, *et al.*, 2013 Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of lake victoria cichlid fishes. Mol. Ecol. **22**: 2848–2863.

Kent, W. J., 2002 BLAT—the BLAST-like alignment tool. Genome Research **12**: 656–664.

Khan, A., E. J. Belfield, N. P. Harberd, and A. Mithani, 2016 HANDS2: Accurate assignment of homoeallelic base-identity in allopolyploids despite missing data. Sci. Rep. **6**: 29234.

Kihara, H. and T. Ono, 1926 Chromosomenzahlen und systematische gruppierung der Rumex-Arten. Zeitschrift für Zellforschung und Mikroskopische Anatomie **4**: 475–481.

Kim, C., M. Pongpanich, and T. Porntaveetus, 2024 Unraveling metagenomics through long-read sequencing: A comprehensive review. J. Transl. Med. **22**: 111.

Knapp, A. K., A. Chen, R. J. Griffin-Nolan, L. E. Baur, C. J. Carroll, *et al.*, 2020 Resolving the dust bowl paradox of grassland responses to extreme drought. Proceedings of the National Academy of Sciences **117**: 22249–22255.

Knapp, A. K., M. Cocke, E. P. Hamerlynck, and C. E. Owensby, 1994 Effect of elevated CO2 on stomatal density and distribution in a C4 grass and a C3 forb under field conditions. Annals of Botany **74**: 595–599.

Kolář, F., M. Čertner, J. Suda, P. Schönswetter, and B. C. Husband, 2017 Mixed-ploidy species:

Progress and opportunities in polyploid research. Trends Plant Sci. **22**: 1041–1055.

Korani, W., D. O'Connor, Y. Chu, C. Chavarro, C. Ballen, *et al.*, 2021 De novo QTL-seq identifies loci linked to blanchability in peanut (*Arachis hypogaea*) and refines previously identified QTL with low coverage sequence. Agronomy **11**: 2201.

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, *et al.*, 2017 Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research **27**: 722–736.

Korneliussen, T., I. Moltke, A. Albrechtsen, and R. Nielsen, 2013 Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. BMC Bioinformatics **14**: 289.

Korneliussen, T. S., A. Albrechtsen, and R. Nielsen, 2014 ANGSD: Analysis of next generation sequencing data. BMC Bioinformatics **15**: 356.

Kosugi, S., Y. Momozawa, X. Liu, C. Terao, M. Kubo, *et al.*, 2019 Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. **20**: 117.

Kral-O'Brien, K. C., R. F. Limb, T. J. Hovick, and J. P. Harmon, 2019 Compositional shifts in forb and butterfly communities associated with Kentucky bluegrass invasions. Rangeland Ecol. Manage. **72**: 301–309.

Kramer, A. T., T. E. Wood, S. Frischie, and K. Havens, 2018 Considering ploidy when producing and using mixed-source native plant materials for restoration. Restoration Ecology **26**: 13–19.

Kyriakidou, M., H. H. Tai, N. L. Anglin, D. Ellis, and M. V. Strömvik, 2018 Current strategies of polyploid plant genome sequence assembly. Front. Plant Sci. **9**: 1660.

Layer, R. M., C. Chiang, A. R. Quinlan, and I. M. Hall, 2014 LUMPY: A probabilistic framework for structural variant discovery. Genome Biol. **15**: R84.

Levin, D. A., 1975 Minority cytotype exclusion in local plant populations. Taxon **24**: 35–43.

Li, H., 2011 A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics **27**: 2987–2993.

Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 .

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

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

Li, H., J. Ruan, and R. Durbin, 2008 Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. **18**: 1851–1858.

Li, J., U. Singh, P. Bhandary, J. Campbell, Z. Arendsee, *et al.*, 2021 Foster thy young: Enhanced prediction of orphan genes in assembled genomes. Nucleic Acids Res. .

Li, Z. and M. S. Barker, 2020 Inferring putative ancient whole-genome duplications in the 1000 plants (1KP) initiative: Access to gene family phylogenies and age distributions. Gigascience **9**.

Li, Z., G. P. Tiley, S. R. Galuska, C. R. Reardon, T. I. Kidder, *et al.*, 2018 Multiple large-scale gene and genome duplications during the evolution of hexapods. Proceedings of the National Academy of Sciences **115**: 4713–4718.

Linck, E. and C. J. Battey, 2019 Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. Mol. Ecol. Resour. **19**: 639–647.

Lisch, D., 2013 How important are transposons for plant evolution? Nat. Rev. Genet. **14**: 49–61.

Liu, R. and J. Dickerson, 2017 Strawberry: Fast and accurate genome-guided transcript reconstruction and quantification from RNA-seq. PLoS Computat. Biol. **13**: e1005851.

Liu, X., S. Han, Z. Wang, J. Gelernter, and B.-Z. Yang, 2013 Variant callers for next-generation sequencing data: A comparison study. PLoS One **8**: e75619.

Lou, R. N., A. Jacobs, A. P. Wilder, and N. O. Therkildsen, 2021 A beginner's guide to low-coverage whole genome sequencing for population genomics. Mol. Ecol. **30**: 5966–5993.

Lou, R. N. and N. O. Therkildsen, 2022 Batch effects in population genomic studies with low-coverage whole genome sequencing data: Causes, detection and mitigation. Mol. Ecol. Resour. **22**: 1678–1692.

Lovell, J. T., A. H. MacQueen, S. Mamidi, J. Bonnette, J. Jenkins, *et al.*, 2021 Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. Nature **590**: 438–444.

Lowry, D. B., 2012 Ecotypes and the controversy over stages in the formation of new species. Biological Journal of the Linnean Society **106**: 241–257.

Lowry, D. B., S. Hoban, J. L. Kelley, K. E. Lotterhos, L. K. Reed, *et al.*, 2017 Breaking RAD: An evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. Mol. Ecol. Resour. **17**: 142–152.

Ma, X.-F. and J. P. Gustafson, 2005 Genome evolution of allopolyploids: A process of cytological and genetic diploidization. Cytogenet. Genome Res. **109**: 236–249.

Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik, 2012 Cluster: Cluster analysis basics and extensions. R package version 2.1.4 .

Mahmoud, M., N. Gobet, D. I. Cruz-Dávalos, N. Mounier, C. Dessimoz, *et al.*, 2019 Structural variant calling: The long and the short of it. Genome Biol. **20**: 246.

Manchanda, N., J. L. Portwood, M. R. Woodhouse, A. S. Seetharam, C. J. Lawrence-Dill, *et al.*, 2020 GenomeQC: A quality assessment tool for genome assemblies and gene structure annotations. BMC Genomics **21**: 1–9.

Manni, M., M. R. Berkeley, M. Seppey, F. A. Simao, and E. M. Zdobnov, 2021 BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. arXiv:2106.11799 .

Mapleson, D., L. Venturini, G. Kaithakottil, and D. Swarbreck, 2018 Efficient and accurate detection of splice junctions from RNA-seq with portcullis. GigaScience **7**: giy131.

Marconi, G., D. Aiello, B. Kindiger, L. Storchi, A. Marrone, *et al.*, 2020 The role of apostart in switching between sexuality and apomixis in *Poa pratensis*. Genes **11**: 941.

Margarido, G. R. A. and D. Heckerman, 2015 ConPADE: Genome assembly ploidy estimation from next-generation sequencing data. PLoS Comput. Biol. **11**: e1004229.

Mascher, M., T. Wicker, J. Jenkins, C. Plott, T. Lux, *et al.*, 2021 Long-read sequence assembly: A technical evaluation in barley. The Plant Cell **33**: 1888—1906.

Mason, A. S. and J. F. Wendel, 2020 Homoeologous exchanges, segmental allopolyploidy, and polyploid genome evolution. Front. Genet. **11**: 1014.

Matzk, F., 1991 New efforts to overcome apomixis in *Poa pratensis* L. Euphytica **55**: 65–72.

McAllister, C., R. Blaine, P. Kron, B. Bennett, H. Garrett, *et al.*, 2015 Environmental correlates of cytotype distribution in *Andropogon gerardii* (Poaceae). Am. J. Bot. **102**: 92–102.

McAllister, C. A. and A. J. Miller, 2016 Single nucleotide polymorphism discovery via genotyping by sequencing to assess population genetic structure and recurrent polyploidization in *Andropogon gerardii*. Am. J. Bot. **103**: 1314–1325.

McClintock, B., 1984 The significance of responses of the genome to challenge. Science **226**: 792–801.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research **20**: 1297–1303.

McKinney, G. J., R. K. Waples, L. W. Seeb, and J. E. Seeb, 2017 Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. Mol. Ecol. Resour. **17**: 656–669.

McMillan, C., 1959 The role of ecotypic variation in the distribution of the central grassland of North America. Ecological Monographs **29**: 286–308.

Md, V., S. Misra, H. Li, and S. Aluru, 2019 Efficient architecture-aware acceleration of BWA-MEM for multicore systems. arXiv p. arXiv:1907.12931.

Meirmans, P. G., S. Liu, and P. H. van Tienderen, 2018 The analysis of polyploid genetic data. Journal of Heredity **109**: 283–296.

Meirmans, P. G. and P. H. Van Tienderen, 2013 The effects of inheritance in tetraploids on genetic diversity and population divergence. Heredity **110**: 131–137.

Milesi, C., S. W. Running, C. D. Elvidge, J. B. Dietz, B. T. Tuttle, *et al.*, 2005 Mapping and modeling the biogeochemical cycling of turf grasses in the United States. Environ. Manage. **36**: 426–438.

Mistry, J., S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, *et al.*, 2021 Pfam: The protein families database in 2021. Nucleic acids research **49**: D412–D419.

Mithani, A., E. J. Belfield, C. Brown, C. Jiang, L. J. Leach, *et al.*, 2013 HANDS: A tool for genome-wide discovery of subgenome-specific base-identity in polyploids. BMC Genomics **14**: 653.

Muir, C. D., M. À. Conesa, J. Galmés, V. S. Pathare, P. Rivera, *et al.*, 2023 How important are functional and developmental constraints on phenotypic evolution? An empirical test with the stomatal anatomy of flowering plants. The American Naturalist **201**: 794–812.

Muir, P., S. Li, S. Lou, D. Wang, D. J. Spakowicz, *et al.*, 2016 The real cost of sequencing: Scaling computation to keep pace with data generation. Genome Biol. **17**: 53.

Musich, R., L. Cadle-Davidson, and M. V. Osier, 2021 Comparison of short-read sequence aligners indicates strengths and weaknesses for biologists to consider. Front. Plant Sci. **12**: 657240.

Nagahama, N. and G. A. Norrmann, 2012 Review of the genus *Andropogon* (Poaceae: Andropogoneae) in America based on cytogenetic studies. Journal of Botany **2012**.

Napier, J. D., P. P. Grabowski, J. T. Lovell, J. Bonnette, S. Mamidi, *et al.*, 2022 A generalist-specialist trade-off between switchgrass cytotypes impacts climate adaptation and geographic range. Proc. Natl. Acad. Sci. U. S. A. **119**: e2118879119.

Neale, D. B., A. V. Zimin, S. Zaman, A. D. Scott, B. Shrestha, *et al.*, 2022 Assembled and annotated 26.5 Gbp coast redwood genome: A resource for estimating evolutionary adaptive potential and investigating hexaploid origin. G3 **12**.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song, 2011 Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. **12**: 443–451.

Njuguna, J. N., L. V. Clark, A. E. Lipka, K. G. Anzoua, L. Bagmet, *et al.*, 2023 Impact of genotype-calling methodologies on genome-wide association and genomic prediction in polyploids. Plant Genome **16**: e20401.

Norrmann, G., C. Quarin, and K. Keeler, 1997 Evolutionary implications of meiotic chromosome behavior, reproductive biology, and hybridization in 6x and 9x cytotypes of *Andropogon gerardii*

(Poaceae). Am. J. Bot. **84**: 201.

Ohno, S., 2013 *Evolution by gene duplication*. Springer Science & Business Media.

Okonechnikov, K., A. Conesa, and F. García-Alcalde, 2016 Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics **32**: 292–294.

O'Leary, S. J., J. B. Puritz, S. C. Willis, C. M. Hollenbeck, and D. S. Portnoy, 2018 These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. Mol. Ecol. .

Olsen, J. T., K. L. Caudle, L. C. Johnson, S. G. Baer, and B. R. Maricle, 2013 Environmental and genetic variation in leaf anatomy among populations of *Andropogon gerardii* (Poaceae) along a precipitation gradient. American Journal of Botany **100**: 1957–1968.

One Thousand Plant Transcriptomes Initiative, 2019 One thousand plant transcriptomes and the phylogenomics of green plants. Nature **574**: 679–685.

Otto, S. P. and J. Whitton, 2000 Polyploid incidence and evolution. Annu. Rev. Genet. **34**: 401–437.

Ou, S., J. Chen, and N. Jiang, 2018 Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res. **46**: e126–e126.

Ou, S., W. Su, Y. Liao, K. Chougule, J. R. Agda, *et al.*, 2019 Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biology **20**: 1–18.

Padgham, M., 2021 *geodist: Fast, Dependency-Free Geodesic Distance Calculations*. R package version 0.0.7.

Page, J. T., A. R. Gingle, and J. A. Udall, 2013 PolyCat: A resource for genome categorization of sequencing reads from allopolyploid organisms. G3 **3**: 517–525.

Page, J. T. and J. A. Udall, 2015 Methods for mapping and categorization of DNA sequence reads from allopolyploid organisms. BMC Genet. **16 Suppl 2**: S4.

Paradis, E. and K. Schliep, 2019 ape. Bioinformatics **35**: 526–528.

Parra-Nunez, P., M. Pradillo, and J. L. Santos, 2020 How to perform an accurate analysis of

metaphase I chromosome configurations in autopolyploids of *Arabidopsis thaliana*. In *Plant Meiosis: Methods and Protocols*, edited by M. Pradillo and S. Heckmann, pp. 25–36, Springer New York, New York, NY.

Pearman, W. S., L. Urban, and A. Alexander, 2022 Commonly used Hardy-Weinberg equilibrium filtering schemes impact population structure inferences using RADseq data. Mol. Ecol. Resour. **22**: 2599–2613.

Pečnerová, P., G. Garcia-Erill, X. Liu, C. Nursyifa, R. K. Waples, *et al.*, 2021 High genetic diversity and low differentiation reflect the ecological versatility of the African leopard. Curr. Biol. **31**: 1862–1871.

Pellicer, J. and I. J. Leitch, 2020 The plant DNA c-values database (release 7.1): An updated online repository of plant genome size data for comparative studies. New Phytol. **226**: 301–305.

Pepin, G. W. and C. R. Funk, 1971 Intraspecific hybridization as a method of breeding Kentucky bluegrass (*Poa pratenis* L.) for turf. Crop Sci. **11**: 445–448.

Peralta, M., M.-C. Combes, A. Cenci, P. Lashermes, and A. Dereeper, 2013 SNiPloid: A utility to exploit High-Throughput SNP data derived from RNA-Seq in allopolyploid species. Int. J. Plant Genomics **2013**: 890123.

Perez-Harguindeguy, N., S. Diaz, E. Garnier, S. Lavorel, H. Poorter, *et al.*, 2016 Corrigendum to: New handbook for standardised measurement of plant functional traits worldwide. Australian Journal of Botany **64**: 715–716.

Pertea, M., G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, *et al.*, 2015 StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnol. **33**: 290–295.

Phillips, A., 2024 Variant calling in polyploids for population and quantitative genetics. EcoEvoRxiv .

Phillips, A. R., A. S. Seetharam, P. S. Albert, T. AuBuchon-Elder, J. A. Birchler, *et al.*, 2023 A happy accident: A novel turfgrass reference genome. G3: Genes, Genomes, Genetics **13**: jkad073.

Phylogenty Working Group, G., 2001 Phylogeny and subfamilial classification of the grasses (Poaceae). Ann. Mo. Bot. Gard. pp. 373–457.

Poland, J. A. and T. W. Rife, 2012 Genotyping-by-sequencing for plant breeding and genetics. Plant Genome **5**: 92–102.

Porturas, L. D., T. J. Anneberg, A. E. Curé, S. Wang, D. M. Althoff, *et al.*, 2019 A meta-analysis of whole genome duplication and the effects on flowering traits in plants. American Journal of Botany **106**: 469–476.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics **155**: 945–959.

Prodanov, T. and V. Bansal, 2022 Robust and accurate estimation of paralog-specific copy number for duplicated genes using whole-genome sequencing. Nat. Commun. **13**: 3221.

Prüfer, K., 2018 snpAD: An ancient DNA genotype caller. Bioinformatics **34**: 4165–4171.

Puritz, J. B., C. M. Hollenbeck, and J. R. Gold, 2014 ddocent: A RADseq, variant-calling pipeline designed for population genomics of non-model organisms. PeerJ **2**: e431.

R Core Team, 2017 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raggi, L., E. Bitocchi, L. Russi, G. Marconi, T. F. Sharbel, *et al.*, 2015 Understanding genetic diversity and population structure of a *Poa pratensis* worldwide collection through morphological, nuclear and chloroplast diversity analysis. PLoS One **10**: e0124709.

Ramakrishnan, M., L. Satish, A. Sharma, K. Kurungara Vinod, A. Emamverdian, *et al.*, 2022 Transposable elements in plants: Recent advancements, tools and prospects. Plant Mol. Biol. Rep. **40**: 628–645.

Ramsey, J. and D. W. Schemske, 1998 Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annual Review of Ecology and Systematics **29**: 467–501.

Ramsey, J. and D. W. Schemske, 2002 Neopolyploidy in flowering plants. Annual Review of Ecology and Systematics **33**: 589–639.

Ranallo-Benavidez, T. R., K. S. Jaron, and M. C. Schatz, 2020 GenomeScope 2.0 and smudgeplot

for reference-free profiling of polyploid genomes. Nat. Commun. **11**: 1432.

Rasmussen, M. S., C. Wiuf, and A. Albrechtsen, 2024 Inferring drift, genetic differentiation, and admixture graphs from low-depth sequencing data.

Refulio-Rodriguez, N. F., J. T. Columbus, L. J. Gillespie, P. M. Peterson, and R. J. Soreng, 2012 Molecular phylogeny of *Dissanthelium* (Poaceae: Pooideae) and its taxonomic implications. Syst. Bot. **37**: 122–133.

Rice, A., L. Glick, S. Abadi, M. Einhorn, N. M. Kopelman, *et al.*, 2015 The chromosome counts database (CCDB) - a community resource of plant chromosome numbers. New Phytol. **206**: 19–26.

Robbins, M. D., B. S. Bushman, D. R. Huff, C. W. Benson, S. E. Warnke, *et al.*, 2023 Chromosome-scale genome assembly and annotation of allotetraploid annual bluegrass (*Poa annua* L.). Genome Biol. Evol. **15**: evac180.

Roddy, A. B., G. Théroux-Rancourt, T. Abbo, J. W. Benedetti, C. R. Brodersen, *et al.*, 2020 The scaling of genome size and cell size limits maximum rates of photosynthesis with implications for ecological strategies. International Journal of Plant Sciences **181**: 75–87.

Román-Palacios, C., C. A. Medina, S. H. Zhan, and M. S. Barker, 2021 Animal chromosome counts reveal a similar range of chromosome numbers but with less polyploidy in animals compared to flowering plants. J. Evol. Biol. **34**: 1333–1339.

Ronfort, J., E. Jenczewski, T. Bataillon, and F. Rousset, 1998 Analysis of population structure in autotetraploid species. Genetics **150**: 921–930.

Ronquist, F., M. Teslenko, P. Van Der Mark, D. L. Ayres, A. Darling, *et al.*, 2012 MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. **61**: 539–542.

Roux, C., X. Vekemans, and J. Pannell, 2023 Inferring the demographic history and inheritance mode of tetraploid species using ABC. In *Polyploidy: Methods and Protocols*, edited by Y. Van de Peer, pp. 325–348, Springer US, New York, NY.

Rowan, B. A., D. Heavens, T. R. Feuerborn, A. J. Tock, I. R. Henderson, *et al.*, 2019 An ultra high-

density *Arabidopsis thaliana* crossover map that refines the influences of structural variation and epigenetic features. Genetics **213**: 771–787.

Rozowsky, J., A. Abyzov, J. Wang, P. Alves, D. Raha, *et al.*, 2011 AlleleSeq: Analysis of allele-specific expression and binding in a network framework. Mol. Syst. Biol. **7**: 522.

Runcie, D. E., J. Qu, H. Cheng, and L. Crawford, 2021 MegaLMM: Mega-scale linear mixed models for genomic predictions with thousands of traits. Genome Biology **22**: 1–25.

Salamov, A. A. and V. V. Solovyev, 2000 Ab initio gene finding in *Drosophila* genomic DNA. Genome Research **10**: 516–522.

Samson, F. and F. Knopf, 1994 Prairie conservation in North America. BioScience **44**: 418–421.

Sax, K., 1936 The experimental production of polyploidy. Journal of the Arnold Arboretum **17**: 153–159.

Schneider, C. A., W. S. Rasband, and K. W. Eliceiri, 2012 NIH Image to ImageJ: 25 years of image analysis. Nature Methods **9**: 671–675.

Schubert, S. D., M. J. Suarez, P. J. Pegion, R. D. Koster, and J. T. Bacmeister, 2004 On the cause of the 1930s Dust Bowl. Science **303**: 1855–1859.

Scott, A. D., J. D. Van de Velde, and P. Y. Novikova, 2023 Inference of polyploid origin and inheritance mode from population genomic data. In *Polyploidy: Methods and Protocols*, edited by Y. Van de Peer, pp. 279–295, Springer US, New York, NY.

Searle, S. R., F. M. Speed, and G. A. Milliken, 1980 Population marginal means in the linear model: An alternative to least squares means. The American Statistician **34**: 216–221.

Session, A. M. and D. S. Rokhsar, 2023 Transposon signatures of allopolyploid genome evolution. Nat. Commun. **14**: 3180.

Sewe, S. O., G. Silva, P. Sicat, S. E. Seal, and P. Visendi, 2022 Trimming and validation of Illumina short reads using trimmomatic, trinity assembly, and assessment of RNA-Seq data. In *Plant Bioinformatics: Methods and Protocols*, edited by D. Edwards, pp. 211–232, Springer US, New York, NY.

Shastry, V., P. E. Adams, D. Lindtke, E. G. Mandeville, T. L. Parchman, *et al.*, 2021 Model-based

genotype and ancestry estimation for potential hybrids with mixed-ploidy. Mol. Ecol. Resour. **21**: 1434–1451.

Slater, G. S. C. and E. Birney, 2005 Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics **6**: 1–11.

Smit, A., R. Hubley, and P. Green, 2008–2015 RepeatModeler Open-1.0. http://www.repeatmasker.org .

Smit, A., R. Hubley, and P. Green, 2013–2015 RepeatMasker Open-4.0.

Smith, A. B., J. Alsdurf, M. Knapp, S. G. Baer, and L. C. Johnson, 2017 Phenotypic distribution models corroborate species distribution models: A shift in the role and prevalence of a dominant prairie grass in response to climate change. Global Change Biology **23**: 4365–4375.

Snodgrass, S., J. Jareczek, and J. Wendel, 2017 An examination of nucleotypic effects in diploid and polyploid cotton. AoB Plants **9**: plw082.

Soltis, D. E., R. J. A. Buggs, W. B. Barbazuk, P. S. Schnable, and P. S. Soltis, 2009 On the origins of species: Does evolution repeat itself in polyploid populations of independent origin? Cold Spring Harb. Symp. Quant. Biol. **74**: 215–223.

Song, L., S. Sabunciyan, and L. Florea, 2016 CLASS2: Accurate and efficient splice variant annotation from RNA-seq reads. Nucleic Acids Res. **44**: e98–e98.

Soraggi, S., J. Rhodes, I. Altinkaya, O. Tarrant, F. Balloux, *et al.*, 2022 HMMploidy: Inference of ploidy levels from short-read sequencing data. Peer Community J. **2**.

Soreng, R., L. Gillespie, and L. Consaul, 2017 Taxonomy of the *Poa laxa* group, including two new taxa from Arctic Canada and Greenland, and Oregon, and a re-examination of *P.* sect. *Oreinos* (Poaceae). Nord. J. Bot. **35**: 513–538.

Soreng, R. J., 2007 *Poa* L. Flora of North America, Poaceae, part 1, vol. 24. p. 486–601.

Soreng, R. J. and F. R. Barrie, 1999 (1391) Proposal to conserve the name *Poa pratensis* (Gramineae) with a conserved type. Taxon **48**: 157–159.

Soreng, R. J. and L. J. Gillespie, 2018 *Poa secunda* J. Presl (Poaceae): a modern summary of infraspecific taxonomy, chromosome numbers, related species and infrageneric placement based

on DNA. PhytoKeys p. 101.

Soreng, R. J., L. J. Gillespie, H. Koba, E. Boudko, and R. D. Bull, 2015 Molecular and morphological evidence for a new grass genus, *Dupontiopsis* (Poaceae tribe Poeae subtribe Poinae s.l.), endemic to alpine Japan, and implications for the reticulate origin of *Dupontia* and *Arctophila* within Poinae s.l. J. of Syst. Evol, **53**: 138–162.

Soreng, R. J., M. V. Olonova, N. S. Probatova, and L. J. Gillespie, 2020 Breeding systems and phylogeny in *Poa*, with special attention to Northeast Asia: The problem of *Poa shumushuensis* and sect. *Nivicolae* (Poaceae). J. Syst. Evol. **58**: 1031–1058.

Stanke, M., O. Schöffmann, B. Morgenstern, and S. Waack, 2006 Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics **7**: 1–11.

Stebbins, G. L., 1950 *Variation and evolution in plants*. Columbia University Press.

Stebbins, G. L., Jr, 1947 Types of polyploids; their classification and significance. Adv. Genet. **1**: 403–429.

Stift, M., C. Berenos, P. Kuperus, and P. H. van Tienderen, 2008 Segregation models for disomic, tetrasomic and intermediate inheritance in tetraploids: A general procedure applied to *Rorippa* (yellow cress) microsatellite data. Genetics **179**: 2113–2123.

Sun, M., E. Pang, W.-N. Bai, D.-Y. Zhang, and K. Lin, 2023 ploidyfrost: Reference-free estimation of ploidy level from whole genome sequencing data based on de Bruijn graphs. Mol. Ecol. Resour. **23**: 499–510.

Sun, S., Y. Zhou, J. Chen, J. Shi, H. Zhao, *et al.*, 2018 Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat. Genet. **50**: 1289–1295.

Sylvester, S. P., R. J. Soreng, and L. J. Gilliespie, 2021 Resolving páramo *Poa* (Poaceae): Morphometric and phylogenetic analysis of the 'Cucullata complex'of north-west South America. Bot. J. Linn. Soc. **197**: 104–146.

Szadkowski, E., F. Eber, V. Huteau, M. Lodé, C. Huneau, *et al.*, 2010 The first meiosis of resynthesized *Brassica napus*, a genome blender. New Phytol. **186**: 102–112.

Taylor, S., P. Franks, S. Hulme, E. Spriggs, P. Christin, *et al.*, 2012 Photosynthetic pathway and ecological adaptation explain stomatal trait diversity amongst grasses. New Phytologist **193**: 387–396.

Therkildsen, N. O. and S. R. Palumbi, 2017 Practical low-coverage genome wide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in non-model species. Mol. Ecol. Resour. **17**: 194–208.

Tiffin, P. and J. Ross-Ibarra, 2014 Advances and limits of using population genetics to understand local adaptation. Trends Ecol. Evol. **29**: 673–680.

Tompkins, R. D., C. A. McAllister, and S. Bloom, 2015 Ploidy levels for some remnant eastern big bluestem (*Andropogon gerardii*) populations: Implications for their conservation and restoration. Ecol. Restor. **33**: 289–296.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, *et al.*, 2012 Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protocols **7**: 562–578.

Udall, J. A. and J. F. Wendel, 2006 Polyploidy and crop improvement. Crop Sci. **46**: S–3–S–14.

van de Geijn, B., G. McVicker, Y. Gilad, and J. K. Pritchard, 2015 WASP: Allele-specific software for robust molecular quantitative trait locus discovery. Nat. Methods **12**: 1061–1063.

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, *et al.*, 2013 From FastQ data to high confidence variant calls: The genome analysis toolkit best practices pipeline. Curr. Protoc. Bioinformatics **43**: 11.10.1–11.10.33.

Van der Auwera, G. A. and B. D. O'Connor, 2020 *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media, Incorporated.

van der Valk, T., P. Pečnerová, D. Díez-del Molino, A. Bergström, J. Oppenheimer, *et al.*, 2021 Million-year-old DNA sheds light on the genomic history of mammoths. Nature **591**: 265–269.

Van Drunen, W. E. and J. Friedman, 2022 Autopolyploid establishment depends on life-history strategy and the mating outcomes of clonal architecture. Evolution **76**: 1953–1970.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. Journal of Dairy Sci-

ence **91**: 4414–4423.

Varvel, N. A., C. J. Hilt, L. C. Johnson, M. Galliart, S. G. Baer, *et al.*, 2018 Genetic and environmental influences on stomates of big bluestem (*Andropogon gerardii*). Environmental and Experimental Botany **155**: 477–487.

Vaser, R., I. Sović, N. Nagarajan, and M. Šikić, 2017 Fast and accurate de novo genome assembly from long uncorrected reads. Genome Research **27**: 737–746.

Vasimuddin, M., S. Misra, H. Li, and S. Aluru, 2019 Efficient architecture-aware acceleration of bwa-mem for multicore systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 314–324, IEEE.

Venturini, L., S. Caim, G. G. Kaithakottil, D. L. Mapleson, and D. Swarbreck, 2018 Leveraging multiple transcriptome assembly methods for improved gene structure annotation. GigaScience **7**: giy093.

Veselỳ, P., P. Bureš, P. Šmarda, and T. Pavlíček, 2012 Genome size and DNA base composition of geophytes: The mirror of phenology and ecology? Annals of Botany **109**: 65–75.

Vieira, F. G., M. Fumagalli, A. Albrechtsen, and R. Nielsen, 2013 Estimating inbreeding coefficients from NGS data: Impact on genotype calling and allele frequency estimation. Genome Research **23**: 1852–1861.

Vile, D., E. Garnier, B. Shipley, G. Laurent, M.-L. Navas, *et al.*, 2005 Specific leaf area and dry matter content estimate thickness in laminar leaves. Annals of Botany **96**: 1129–1136.

Viruel, J., O. Hidalgo, L. Pokorny, F. Forest, B. Gravendeel, *et al.*, 2023 A bioinformatic pipeline to estimate ploidy level from target capture sequence data obtained from herbarium specimens. Methods Mol. Biol. **2672**: 115–126.

Vorontsova, M. S., K. B. Petersen, P. Minx, T. M. Aubuchon-Elder, M. C. Romay, *et al.*, 2023 Reinstatement and expansion of the genus *Anatherum* (Andropogoneae, Panicoideae, Poaceae). Systematics and Biodiversity **21**: 2274386.

Walczyk, A. M. and E. I. Hersch-Green, 2023 Genome-material costs and functional trade-offs in the autopolyploid *Solidago gigantea* (giant goldenrod) series. American Journal of Botany **110**:

e16218.

Wang, J., D. Li, F. Shang, and X. Kang, 2017 High temperature-induced production of unreduced pollen and its cytological effects in *Populus*. Scientific Reports **7**: 5281.

Wang, L.-G., T. T.-Y. Lam, S. Xu, Z. Dai, L. Zhou, *et al.*, 2020 Treeio: An R package for phylogenetic tree input and output with richly annotated and associated data. Molecular Biol. Evol. **37**: 599–603.

Wang, M., J. Li, Z. Qi, Y. Long, L. Pei, *et al.*, 2022 Genomic innovation and regulatory rewiring during evolution of the cotton genus *Gossypium*. Nat. Genet. **54**: 1959–1971.

Wang, T., A. Hamann, D. Spittlehouse, and C. Carroll, 2016 Locally downscaled and spatially customizable climate data for historical and future periods for North America. PloS ONE **11**: e0156720.

Weaver, J. E., 1968 Prairie plants and their environment. A fifty-year study in the midwest. .

Weaver, J. E. and F. W. Albertson, 1943 Resurvey of grasses, forbs, and underground plant parts at the end of the great drought. Ecological Monographs **13**: 63–117.

Weiß, C. L., M. Pais, L. M. Cano, S. Kamoun, and H. A. Burbano, 2018 nQuire: A statistical framework for ploidy estimation using next generation sequencing. BMC Bioinformatics **19**: 122.

Wickell, D., J. Landis, E. Zimmer, and F.-W. Li, 2024 Population genomics of the *Isoetes appalachiana* (Isoetaceae) complex supports a 'diploids-first'approach to conservation. Annals of Botany **133**: 261–272.

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

Wood, D. E. and S. L. Salzberg, 2014 Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biology **15**: 1–12.

Wood, T. E., N. Takebayashi, M. S. Barker, I. Mayrose, P. B. Greenspoon, *et al.*, 2009a The frequency of polyploid speciation in vascular plants. Proc. Natl. Acad. Sci. U. S. A. **106**: 13875–13879.

Wood, T. E., N. Takebayashi, M. S. Barker, I. Mayrose, P. B. Greenspoon, *et al.*, 2009b The

frequency of polyploid speciation in vascular plants. PNAS **106**: 13875–13879.

Wright, I. J., P. B. Reich, M. Westoby, D. D. Ackerly, Z. Baruch, *et al.*, 2004 The worldwide leaf economics spectrum. Nature **428**: 821–827.

Wu, T. D. and S. Nacu, 2010 Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics **26**: 873–881.

Xu, G., J. Lyu, Q. Li, H. Liu, D. Wang, *et al.*, 2020 Evolutionary and functional genomics of DNA methylation in maize domestication and improvement. Nat. Commun. **11**: 5539.

Yu, G., 2020 Using ggtree to visualize data on tree-like structures. Curr. Protoc. in Bioinformatics **69**: e96.

Yu, R.-M., N. Zhang, B.-W. Zhang, Y. Liang, X.-X. Pang, *et al.*, 2023 Genomic insights into biased allele loss and increased gene numbers after genome duplication in autotetraploid *Cyclocarya paliurus*. BMC Biol. **21**: 168.

Zack, T. I., S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, *et al.*, 2013 Pan-cancer patterns of somatic copy number alteration. Nat. Genet. **45**: 1134–1140.

Zhan, S., C. Griswold, and L. Lukens, 2021 *Zea mays* RNA-seq estimated transcript abundances are strongly affected by read mapping bias. BMC Genomics **22**: 285.

Zhang, R.-G., Z.-X. Wang, S. Ou, and G.-Y. Li, 2019a TEsorter: Lineage-level classification of transposable elements using conserved protein domains. bioRxiv .

Zhang, X., S. Zhang, Q. Zhao, R. Ming, and H. Tang, 2019b Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat Plants **5**: 833–845.

Zhou, Y., J. Zhang, X. Xiong, Z.-M. Cheng, and F. Chen, 2022 *De novo* assembly of plant complete genomes. Tropical Plants **1**: 1–8.