Polyploidy inference across time scales in the charismatic carnivorous plant genus *Drosera* L. (Droseraceae, Caryophyllales)

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Rebekah Anne Mohn

IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisor: Ya Yang

2023

Acknowledgements

First and foremost, I am deeply grateful Dr. Ya Yang for her outstanding guidance and support as my advisor. Though I am her first graduate student, Ya guided me through graduate school with the wisdom of a seasoned mentor. Ya allowed me to develop my own research project, but I could not have succeeded without her wise guidance. She challenged me to think critically about my research, supported me in my pursuits and collaborations, and advocated for me along the way. Her unfailing encouragement and support through the many challenges of the past six years has been invaluable. I cannot say thank you enough, Ya, for your investment in me!

I am also grateful to my committee members, Dr. Yaniv Brandvain, Dr. Adam Cross, Dr. Clay Carter, Dr. Emma Goldberg, Dr. Dave Moeller, and Dr. George Weiblen for the time they invested in guiding me through the Ph.D. process. They helped me keep in mind both the details and big picture of my research and ensured I got the professional development experience to succeed as a researcher. I especially appreciate the sacrifice of Dr. Adam Cross who attended my committee meeting at odd hours of the night/morning from his location in Australia. I am also grateful to Adam and the Centre for Mine Site Restoration at Curtin University for hosting me during my unfortunately abbreviated Fulbright Futures Fellowship. I want to extend a huge thank you to Dr. Daniel Stanton who stepped in at the last minute to serve on my committee for my defense when Adam was not able to make it. Thank you also to Dr. Peter Tiffin and Dr. Keith Barker for their valuable feedback on my research proposal.

I am deeply grateful to my collaborators for their instruction and contributions to my research. Dr. Rosana Zenil-Ferguson provided essential training and assistance with the BiChrom RevBayes model. She also provided mentoring and support throughout my graduate career. I owe much of my knowledge of *Drosera* to the many hours Thilo Krueger spent reviewing the taxa with me. He also outlined and planned the collection trips for my Fulbright Fellowship in Australia. While those trips were disrupted by COVID-19, he continued to collect plants that will play a critical role when Chapter 3 is published. Dr. Andreas Fleischmann and Dr. Fernando Rivadavia provided valuable assistance with the taxonomy and samples of non-Australian *Drosera* species. While it is

still in progress, Dr. Andreas Houben kindly provided time and feedback on CenH3 staining in *Drosera*. I would not have successfully completed my dissertation without the help of these collaborators, so I would like to sincerely thank each one of them.

I would also like to thank all the Yang Lab members for their camaraderie and support. I am grateful to the lab managers, Yinyin Huang and Zachary Radford, for keeping the lab running smoothly and providing assistance in RNA tissue collection and extraction. Thank you to Dr. Diego Morales-Briones for his assistance setting up the programs I needed and for training me to run the pipeline of analyses used in the lab. Diego also kindly provided feedback as I prepared for the preliminary exams and applied for grants, and he assisted me with collecting cultivated tissue and extracting DNA. Aaron Lee also assisted me with collecting, computational analysis, and feedback on my dissertation chapters. Thank you to Alexandra Crum, Ben Cooper, Dr. Lingyun Chen and Dr. Brett Fredericksen for providing feedback on presentations along the way. Undergraduates Erin Boehme and Nicholas De La Rosa provided assistance in maintaining cultivated plants.

I am deeply grateful to the many people who made my collections possible through helping me locate plants, assisting with permits, and collecting plants with me. For their field assistance, I would like to thank Bruce Mohn, Darren Loomis, Kristen Bednarczyk, Ezra Mohn, Stephen Mohn, Caleb Mohn, and Jason Husveth. The Bell Museum, Welby Smith, Chel Anderson, Ethan Perry, Michael Lee, Nicholas Severson, Shannon Kimball, Andrea Pipp, Barry Rice, John Hayden, Chris Ludwig, John Townsend assisted with locating plants and advice on accessing remote locations. I would also like to thank the land managers at the US Forest Service at the Idaho Panhandle National Forest Bonner's Ferry Ranger District and Priest Lake Ranger District and the Lincoln Ranger District at the Helena National Forest; the Minnesota Department of natural Resources; New Jersey Bureau of Land Management, Division of Fish and Wildlife; and Virginia Department of Natural History, Division of Natural Heritage for maintaining the sites studied.

As cultivation of sundews is challenging, the advice and assistance of Dr. Alex Eilts, the College of Biological Sciences Conservatory, Mark Anderson, Tom Rolf, and Dr. John Brittnacher contributed significantly to my success.

I would also like to thank Dr. Rahul Roy for his mentorship and assistance with RNA extraction from *Drosera.*

A huge thank you to the Plant and Microbial Biology program and Sara Eliason. Sara, your support and friendship relieved a lot of stress along the way. Thank you!

I am grateful the Fulbright Commission, National Science Foundation, Society of Systematic Biology, and Botanical Society of America for enabling this research through their financial support. At the University of Minnesota, I would like to thank the Bell Museum, the Graduate School, the Plant and Microbial Biology program, and Ecology, Evolution, and Behavior for supporting my professional development and research.

I would like to thank my friends and family for their support along the way. The list is too long to name all the friends who helped to make Minnesota my home for the past six years, but I am so thankful for each one of them. I am also deeply grateful to the friends that made my transition to Australia smooth, especially Bryan and Esther Scott and the Tivendale family. They helped immeasurably to give me a home away from home. Finally, I would not be where I am without my family as they have supported and encouraged me along the way. In addition to shaping me into the scientist I am today, my parents have been a sounding board as I navigated graduate school. My family also planned vacations around my plant collection trips so that they could assist me with those collections. Thank you, Dad, Mom, Stephen, Joshua and Heidi, Caleb, Ezra, Jaelynn, and Samuel!

Dedication

To my parents, Drs. Kenneth and Joanna Mohn, for cultivating my curiosity and wonder in plants and the natural world.

Table of Contents

Introduction

Chromosome evolution events, such as duplication, inversion, fusion, and fission are universal across the tree of life (reviewed in Coghlan et al., 2005) and have long been considered a driving force of speciation and lineage diversification (Stebbins, 1971; Grant, 1981; Coyne and Orr, 2004). At the same time, the past few years have seen rapid advances in analytical approaches for detecting whole genome duplication events by synteny, the distribution of synonymous distance among paralogs, elevated rates of gene duplication events from gene trees, and frequently a combination of multiple of these methods (Jiao et al., 2014; McKain et al., 2016; Tiley et al., 2018; Yang et al., 2018; Zwaenepoel and Van de Peer, 2019b; a). Advances in analytical approaches are also seen in the development of macroevolutionary models that explicitly consider single chromosome gain and loss and whole genome duplication events (Mayrose et al., 2010; Zenil-Ferguson et al., 2017). Combined, we now know that in plants the rate of whole genome duplication events is higher in herbaceous compared to woody species (Zenil-Ferguson et al., 2017), are associated with niche evolution and either increased or decreased species diversification rate depending on the dataset analyzed (Mayrose et al., 2011; Vanneste et al., 2014; Tank et al., 2015; Smith et al., 2018; Baniaga et al., 2019). However, there lacks a system with both a well-resolved species-level phylogeny and comprehensive cytological data.

I sought to fill the gap in our understanding of chromosome evolution using the carnivorous plants known as sundews (*Drosera* L., Droseraceae). *Drosera* is particularly interesting in exhibiting both ancient (Walker et al., 2017; Yang et al., 2018) and more recent polyploidy events (Wood, 1955; Rothfels and Heimburger, 1968). It also has high numbers of single chromosome number changes documented (Sheikh et al., 1995; Hoshi and Kondo, 1998; Rivadavia et al., 2003; Shirakawa et al., 2011).

Drosera consists of ~250 species distributed worldwide with over half of the species occurring in Oceania, and Africa and South America representing two additional hotspots of diversity (Rivadavia et al., 2003; Fleischmann et al., 2017). The latest classifications of *Drosera* include four subgenera and 15 sections (Rivadavia et al., 2003; Fleischmann et al., 2017). While an updated phylogeny is in process (Fleischmann *et al.*,

personal communication), the three additional nuclear markers are insufficient to resolve species relationships. A recent study found that 500 loci were necessary to infer the correct simulated phylogeny under high levels of incomplete lineage sorting (Solís-Lemus and Ané, 2016). Similarly, phylogenomic studies with more than 500 loci have been able to detect and critically evaluate the presence of incomplete lineage sorting and hybridization within phylogenies (Yang et al., 2015; Pease et al., 2016; Solís-Lemus and Ané, 2016). In addition to providing enough loci, transcriptomics provide biologically meaningful data that can be mined to understand the evolution of gene families of interest (Pease et al., 2016).

High levels of chromosome number variation in *Drosera* (Rivadavia et al., 2003) provide an opportunity to evaluate various chromosome change and polyploidy inference methods. Chromosome numbers in *Drosera* vary from $2n = 6$ to $2n = 80$ with high levels of chromosome number variation among closely related species (Kondo and Lavarack, 1984; Hoshi and Kondo, 1998). In addition to frequent single chromosome gain and loss events, both recent and ancient polyploidy events have been detected in *Drosera*. Previous phylotranscriptomic studies in Caryophyllales have placed at least one whole genome duplication event early in Droseraceae (Walker et al., 2017). More recent whole genome duplication events are evident from chromosome counts in *Drosera* subclades where chromosome base numbers are relatively stable. For example, of all eight North American species of *Drosera*, seven have the chromosome count of $2n = 20$, with the only exception being *D. anglica* ($2n = 40$), which is likely from a more recent polyploidy event (Wood, 1955). Similarly, the hexaploid *D. tokaiensis* ($2n = 60$) native to Japan has been documented to be from allopolyploidy origin from the two widespread species *D. spatulata* ($2n = 40$) and *D. rotundifolia* (Hoshi et al., 2017). *Drosera*, in this case, provides an opportunity to study both ancient and more recent cases of polyploidy events.

This dissertation seeks to evaluate methods for detecting chromosome number changes of different types and ages, and to develop *Drosera* as a system for chromosome evolution. This is achieved by using chromosome evolution models to test different rates of chromosome evolution between different subgenera in Chapter 1. Chapter 2 infers the parental lineages of *Drosera anglica,* a neopolyploid, and Chapter 3 explores polyploidy and reticulation along the backbone of *Drosera* and infers the backbone relationship in

the genus. The three research topics together address pros and cons of different methods in teasing apart the history of polyploid lineages.

REFERENCES

Baniaga, A. E., H. E. Marx, N. Arrigo, and M. S. Barker. 2019. Polyploid plants have faster rates of multivariate niche differentiation than their diploid relatives. *Ecology Letters* n/a.

Coghlan, A., E. E. Eichler, S. G. Oliver, A. H. Paterson, and L. Stein. 2005. Chromosome evolution in eukaryotes: a multi-kingdom perspective. *Trends in Genetics* 21: 673–682.

Coyne, J. A., and H. A. Orr. 2004. Speciation. Sinauer Associates, Sunderland, MA, USA.

Fleischmann, A., A. T. Cross, R. Gibson, P. M. Gonella, and K. W. Dixon. 2017. Systematics and evolution of Droseraceae. Oxford University Press.

Grant, V. 1981. Plant speciation. 2nd ed. Columbia Univeristy Press, New York City, NY, USA.

Han, T.-S., Q.-J. Zheng, R. E. Onstein, B. M. Rojas-Andrés, F. Hauenschild, A. N. Muellner-Riehl, and Y.-W. Xing. 2019. Polyploidy promotes species diversification of *Allium* through ecological shifts. *New Phytologist*.

Hoshi, Y., M. Azumatani, C. Suyama, and L. Adamec. 2017. Determination of Ploidy Level and Nuclear DNA Content in the Droseraceae by Flow Cytometry. *CYTOLOGIA* 82: 321–327.

Hoshi, Y., and K. Kondo. 1998. Chromosome differentiation in *Drosera*, subgenus *Rorella*, section *Rossolis*.

Jiao, Y. N., J. P. Li, H. B. Tang, and A. H. Paterson. 2014. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *Plant Cell* 26: 2792–2802.

Kondo, K., and P. S. Lavarack. 1984. A cytotaxonomic study of some Australian species of *Drosera* Droseraceae. *Bot. J. Linn. Soc.* 88: 317–334.

Mayrose, I., M. S. Barker, and S. P. Otto. 2010. Probabilistic models of chromosome number evolution and the inference of polyploidy. *Syst Biol* 59: 132–144.

Mayrose, I., S. H. Zhan, C. J. Rothfels, K. Magnuson-Ford, M. S. Barker, L. H. Rieseberg, and S. P. Otto. 2011. Recently formed polyploid plants diversify at lower rates. *Science* 333: 1257.

McKain, M. R., H. Tang, J. R. McNeal, S. Ayyampalayam, J. I. Davis, C. W. dePamphilis, T. J. Givnish, et al. 2016. A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. *Genome Biol Evol* 8: 1150–64.

Pease, J. B., D. C. Haak, M. W. Hahn, and L. C. Moyle. 2016. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. *PLoS biology* 14: e1002379.

Rivadavia, F., K. Kondo, M. Kato, and M. Hasebe. 2003. Phylogeny of the sundews, *Drosera* (Droseraceae), based on chloroplast *rbcL* and nuclear 18S ribosomal DNA Sequences. *American Journal of Botany* 90: 123–130.

Rothfels, K., and M. Heimburger. 1968. Chromosome size and DNA values in sundews (Droseraceae). *Chromosoma* 25: 96–103.

Sheikh, S. A., K. Kondo, and Y. Hoshi. 1995. Study on diffused centromeric nature of *Drosera* chromosomes. *Cytologia* 60: 43–47.

Shirakawa, J., Y. Hoshi, and K. Kondo. 2011. Chromosome differentiation and genome organization in carnivorous plant family Droseraceae. *Chromosome Botany* 6: 111–119.

Smith, S. A., J. W. Brown, Y. Yang, R. Bruenn, C. P. Drummond, S. F. Brockington, J. F. Walker, et al. 2018. Disparity, diversity, and duplications in the Caryophyllales. *New Phytol* 217: 836–854.

Solís-Lemus, C., and C. Ané. 2016. Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting S. Edwards [ed.],. *PLOS Genetics* 12: e1005896.

Stebbins, G. L. 1971. Chromosomal evolution in higher plants. Edward Arnold Ltd., London.

Tank, D. C., J. M. Eastman, M. W. Pennell, P. S. Soltis, D. E. Soltis, C. E. Hinchliff, J. W. Brown, et al. 2015. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytol* 207: 454–467.

Tiley, G. P., M. S. Barker, and J. G. Burleigh. 2018. Assessing the performance of Ks plots for detecting ancient whole genome duplications. *Genome Biol Evol*.

Vanneste, K., G. Baele, S. Maere, and Y. Van de Peer. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome Res* 24: 1334–47.

Walker, J. F., Y. Yang, M. J. Moore, J. Mikenas, A. Timoneda, S. F. Brockington, and S. A. Smith. 2017. Widespread paleopolyploidy, gene tree conflict, and recalcitrant relationships among the carnivorous Caryophyllales. *American Journal of Botany* 104: 858–867.

Wood, C. E. 1955. Evidence for the hybrid origin of *Drosera anglica*. *Rhodora* 57: 105–130.

Yang, Y., M. J. Moore, S. F. Brockington, J. Mikenas, J. Olivieri, J. F. Walker, and S. A. Smith. 2018. Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy events. *New Phytol* 217: 855–870.

Yang, Z., E. K. Wafula, L. A. Honaas, H. Zhang, M. Das, M. Fernandez-Aparicio, K. Huang, et al. 2015. Comparative transcriptome analyses reveal core parasitism genes and suggest gene duplication and repurposing as sources of structural novelty. *Molecular Biology and Evolution* 32: 767–90.

Zenil-Ferguson, R., J. M. Ponciano, and J. G. Burleigh. 2017. Testing the association of phenotypes with polyploidy: An example using herbaceous and woody eudicots. *Evolution*.

Zwaenepoel, A., and Y. Van de Peer. 2019a. Inference of Ancient Whole-Genome Duplications and the Evolution of Gene Duplication and Loss Rates. *Molecular Biology and Evolution* 36: 1384–1404.

Zwaenepoel, A., and Y. Van de Peer. 2019b. wgd—simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* 35: 2153–2155.

Chapter 1: Dramatic difference in rate of chromosome number evolution among sundew (*Drosera* L., Droseraceae) lineages

Running Head: Rate of chromosome evolution in *Drosera*

Rebekah A. Mohn[1*], Rosana Zenil-Ferguson[2], Thilo A. Krueger[3], Andreas S. Fleischmann[4,5], Adam T. Cross[3,6], Ya Yang[1]

Affiliations:
[1]Department of Plant and Microbial Biology, University of Minnesota-Twin Cities, 1445 Gortner Avenue, St. Paul, MN 55108-1095, United States of America
[2]Department of Biology, University of Kentucky, Lexington, Kentucky, United States of America
[3]School of Molecular and Life Sciences, Curtin University, Bentley, Australia
[4]Botanische Staatssammlung München (SNSB-BSM), Munich, Germany
[5]GeoBio-Center LMU, Ludwig-Maximilians-University, Munich, Germany
[3]School of Molecular and Life Sciences, Curtin University, Bentley, Australia
[6]EcoHealth Network, Brookline, Massachusetts, United States of America.

*Corresponding author's email: rebekah.mohn@gmail.com
Author Contribution Statement: RM designed and led the work. TK, AF, and AC contributed to literature search and verified taxonomy. RM and RZF analyzed the data. RM and YY led the writing. All authors provided feedback on the manuscript and approved the final version.

Abstract:

Chromosome number change is a driver of speciation in eukaryotic organisms. Carnivorous sundews, the plant genus *Drosera* L., exhibit single chromosome number variation among and within species, especially in the Australian *Drosera* subg. *Ergaleium* D.C., potentially linked to the presence of holocentromeres. We critically

reviewed literature on chromosome counts in *Drosera*, verified the taxonomy and count quality, and reconstructed chronograms to test alternate models where the chromosome number gain, loss, and doubling rates (+1, −1, ×2) were the same or different between *D.* subg. *Ergaleium* and the other subgenera. The best model for chromosome evolution had equal rates of polyploidy (0.013 per million years; Myr) but higher rates of single chromosome number gain (0.13 and 0.021 per Myr) and loss (0.14 and 0.00040 per Myr) in *D.* subg. *Ergaleium* compared to the other subgenera. We found no evidence for differences in single chromosome evolution to be associated with differences in diploidization after polyploidy, self-compatibility, or to holocentric chromosomes as had been previously proposed. This study highlights the complexity of factors influencing rates of chromosome number evolution.

Keywords: BiChrom model; chromosome number change; diploidization; RevBayes; holocentric chromosomes; carnivorous plants

Data Accessibility: All data and analysis setting files not available in the supplementary material will be archived and made available at DOI:10.5281/zenodo.6081366 upon manuscript acceptance.

INTRODUCTION

Chromosome evolution events, such as duplication, inversion, fusion, and fission, are universal across the eukaryotic tree of life but appear to be more common in some lineages than others (reviewed in Coghlan et al., 2005). These chromosomal changes have long been considered driving forces of speciation and lineage diversification (Stebbins, 1971; Grant, 1981; Coyne and Orr, 2004). Therefore, identifying lineages with unusually high or low rates of chromosome change and the intrinsic and environmental factors influencing these rates is critical to our understanding of evolutionary processes in general.

Recent developments in macroevolutionary modeling approaches have explored the association of chromosome evolution with trait evolution and lineage diversification (Mayrose et al., 2011; Freyman and Höhna, 2018; Baniaga et al., 2019; Zenil-Ferguson et al., 2019; Román-Palacios et al., 2020; Zhan et al., 2021). However, most of this work has focused on the role of chromosome doubling. Putative factors influencing the occurrence of single chromosome change include post-polyploidy dysploidy and rediploidization (Mandáková and Lysak, 2018), as well as centromere type (Luceño and Guerra, 1996; Mayrose and Lysak, 2020; Ruckman et al., 2020). Factors influencing the establishment of a new karyotype have only been explored in relation to polyploidy but likely impact single chromosome evolution as well (Husband et al., 2013; Weiss-Schneeweiss et al., 2013; Van Drunen and Husband, 2019). For example, self-incompatibility is expected to hinder proliferation and fixation of the new karyotype due to the formation of deleterious heterozygote karyotypes (Husband et al., 2013; Van Drunen and Husband, 2019). However, the relative importance of selfing in the establishment of single chromosome changes remains largely unknown.

Despite the importance of chromosome change to understanding evolution, obtaining a dataset of chromosome numbers with a matching phylogenetic tree to model the rates of chromosome change is challenging. A well-resolved phylogeny with a comprehensive species-level sampling is not always available. Further, chromosome counts require fresh root tips or flower buds, and counts are often incomplete for lineages with broad geographic distributions. In addition to incomplete sampling, the quality of chromosome count datasets is eroded by chromosome counting errors (Windham and

Yatskievych, 2003), reporting errors in chromosome number databases (Rivero et al., 2019), and taxonomic issues from misidentifications and taxonomic changes.

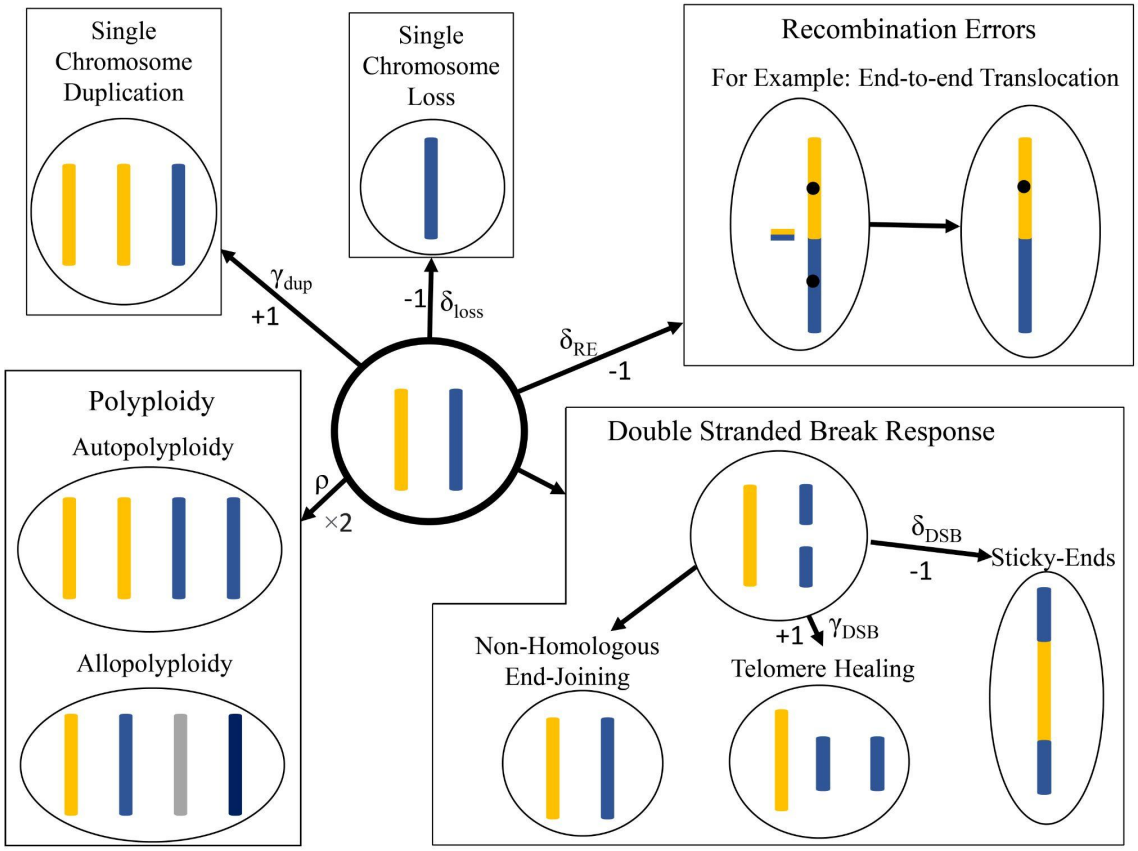


Figure 1: Processes that give rise to changes in chromosome number. Each cell is depicted in haploid form. The original cell (center) starts with two haploid chromosomes. Arrows indicate changes in chromosomes and, where possible, are labeled with the type of change (+1, -1, ×2) and the symbol used in BiChrom models (γ, δ, and ρ respectively; Mayrose & Lysak, 2020). Since +1 and -1 can occur via multiple mechanisms with different impacts on gene copy number, a subscript is used to distinguish the cause of change. Therefore, $\gamma_{DSB} + \gamma_{dup} = \gamma$, and $\delta_{DSB} + \delta_{RE} + \delta_{loss} = \delta$. The centromere is shown as a black spot in the "Recombination Error" box to emphasize the steps required to handle an additional centromere. An increase in one chromosome can be due to telomere healing after a chromosome break or a single chromosome duplication; a single chromosome decrease can be due to a recombination error (Mayrose & Lysak, 2020), two chromosomes fusing after a breakage, or the loss of a single chromosome. Single chromosome loss is unlikely except after polyploidy (Luceno & Guerrra, 1996). A doubling of all chromosomes can be due to an auto- or allo-polyploidy. Holocentromeres are expected to alleviate issues caused by acentric fragments after double stranded breaks and tangling of bicentric chromosomes after fusion (Cuacos et al., 2015).

The carnivorous plants known as sundews (genus *Drosera* L.; family Droseraceae; order Caryophyllales) are exceptionally well-studied cytologically, with

chromosome counts available for about half of its ca. 260 species. *Drosera* species are widely distributed and occur in a wide variety of habitats from boreal peatlands to tropical savannahs and subtropical sandplain heathlands and rock outcrops (Fleischmann et al., 2018). Hotspots of species diversity include Australia (ca. 170 species), Africa (ca. 40 species), and South America (ca. 40 species; Fleischmann et al., 2018). *Drosera* consists of four well-supported subgenera (Fleischmann et al., 2018): two subgenera *D.* subg. *Regiae* Seine & Barthlott and *Arcturia* (Planch.) Schlauer that include only one and two species each, respectively; and two subgenera *D.* subg. *Drosera* L. and *Ergaleium* D.C. comprising ca. 110 and ca. 150 species, respectively. Cytological studies on *Drosera* have been undertaken for over 120 years (Huie, 1897; Rosenberg, 1903), resulting in a rich literature record comprising more than 600 individual chromosome counts for ca. 140 species (e.g., Rothfels and Heimburger, 1968; Kress, 1970; Sheikh and Kondo, 1995; Chen, 1998; Rivadavia, 2005).

Previous cytological studies in *Drosera* have suggested strikingly elevated levels of single chromosome number variation in *D.* subg. *Ergaleium* (almost every haploid number from 1*n* = 3 to 23, with numbers up to 45; tuberous, pygmy, and wooly sundews of Australia; Table S1; Sheikh and Kondo 1995; Hoshi and Kondo, 1998; Rivadavia et al., 2003; Shirakawa, et al., 2011). In contrast, the other three subgenera exhibit primarily polyploid chromosome number series (*n* = 10, 14, 15, 20, 30, 40); Hoshi and Kondo, 1998; Rivadavia et al., 2003). The increased single chromosome number variation has been attributed to the presence of holocentric chromosomes in *Drosera* (Sheikh et al., 1995). Holocentric chromosomes have a single centromere groove (holocentromere) that extends across the majority of the length of the chromosome rather than the localized centromere in the typical monocentric chromosome (Wanner et al., 2015). Holocentric chromosomes have been observed to segregate properly even in individuals heterozygous for a chromosome break (Luceño and Guerra, 1996; Jankowska et al., 2015; Ruckman et al., 2020) and therefore have been associated with increased chromosome fission producing a higher number of smaller chromosomes (Cuacos et al., 2015; Ruckman et al., 2020). In *Drosera*, multiple indirect experimental approaches have been used to investigate the presence of holocentromeres in mitotic tissues. Depending on the experimental approach, all of the species investigated in each study either showed

evidence for having monocentric (three species; Demidov et al., 2014) or holocentric chromosomes (eight species; Sheikh et al., 1995; Furuta and Kondo, 1999; Shirakawa, et al., 2011; Zedek et al., 2016; Kolodin et al., 2018). This suggests that the experimental approaches may be inconclusive, and/or the elevated levels of chromosome number variation in *Drosera* may not correspond to the presence of holocentromeres. Contrasting levels of chromosome number variation could also result from different ages of the lineages, uneven taxon sampling, counting errors, and taxonomic issues (e.g., the misidentification of *D. spatulata* as *D. aliciae* due to morphological similarity; see Kress 1970; of *D. montana* and closely allied taxa due to taxonomic revisions; see Rivadavia, 2005). A critical evaluation of chromosome count data across all records is required to lay the foundations for subsequent analyses. Furthermore, the rate of chromosome number change has yet to be tested using a modeling framework that considers both the phylogenetic history and different modes of chromosome evolution. This phylogenetic modeling framework would also allow the investigation of associations between rates of chromosome number evolution and traits such as centromere type, life history, and mating system.

In this study, we quantified the rate of chromosome doubling and single chromosome gain and loss on dated phylogenies of *Drosera*. We tested whether the rates of chromosome evolution differ significantly between *D.* subg. *Ergaleium* and the other three subgenera, by critically evaluating previously published chromosome counts, verifying voucher specimens to identify possible taxonomic updates or misidentifications, and using the BiChrom (binary state linked to chromosome number change) models (Zenil-Ferguson et al., 2017) and Bayes factors to compare models of subgeneric differences in rates of chromosome evolution in a genus-wide phylogenetic context. An ancestral state reconstruction based on the resulting best-fit model was compared with genome size, self-compatibility, and centromere type to explore potential factors associated with different chromosome evolution rates between *Drosera* subgenera.

METHODS

Literature review and evaluation of chromosome counts

Lists of original references for *Drosera* chromosome counts were obtained from the Chromosome Counts Database (Rice et al., 2015), Index of Plant Chromosome Numbers (Goldblatt and Johnson, 1979), citations referenced by publications on karyotypes in *Drosera* (Kondo, 1969; Dawson, 2000; Rivadavia et al., 2003; Veleba et al., 2017), and searches on Google Scholar and the library databases of the University of Minnesota, Curtin University, and University of Western Australia. Voucher specimen information, chromosome count methodology, and provenance data were recorded for every chromosome count either from the original publication or from subsequent literature in the case of 14 counts (six publications) where the original data could not be obtained.

We excluded chromosome counts from subsequent analyses where the count was uncertain (12 counts), where counts were made from first-generation hybrids (31 counts; we kept allopolyploid species), or where taxonomic issues existed (72 counts). Count uncertainty included chromosome number uncertainty expressed by the original publication (8 counts), a count based on a single cell, and a different chromosome number cited by the voucher vs. the corresponding publication (2 counts). Taxonomic issues included 1) counts that lack both species identification and voucher specimen; 2) species with taxonomy updates after the karyotype publication (especially in the case of species complexes) that lack sufficient provenance, character description, or any voucher specimen with which to assign the taxon to the updated species name; 3) counts made from cultivated material of a species often misidentified in cultivation; or 4) a mismatch between the voucher specimen and the name associated with the count. See Supplemental Information S1 for details on evaluating published chromosome count data, and see Table S1 for how extraneous situations were filtered.

After filtering, if multiple chromosome numbers were reported for a species, all chromosome numbers with more than one count were used for subsequent modeling analyses. In cases where all chromosome numbers for a species had only one count, all counts for that species were used.

Phylogenetic reconstruction for comparative analyses

In order to estimate chloroplast and nuclear chronograms for modeling chromosome number evolution, *rbc*L and ITS sequences for *Drosera* species and outgroup taxa from non-core Caryophyllales were retrieved from the GenBank (Table S2).

For *rbc*L, five sequences were removed due to ambiguous nucleotide sites. The taxonomy for sequences with herbarium vouchers at M and SPF (herbarium acronyms following Index Herbariorum) were updated as noted in Table S2. For species with multiple *rbc*L sequences, the longest sequence was kept.

Sequences were aligned with default settings using the MAFFT (Katoh and Standley, 2013) plug-in for Geneious version 11.1.5 (Kearse et al., 2012). The ends of sequences that were only present in two outgroup species were trimmed. Priors for molecular dating in BEAST version 2.6.4 (Bouckaert et al., 2014) followed previous molecular dating analysis across the Caryophyllales (Yao et al., 2019) using a lognormal relaxed molecular clock and the birth-death model of speciation. For each fossil constraint, the prior was set to a lognormal distribution with a mean of 1.0, a standard deviation of 0.5, and an offset based on the age of the fossil. As in Yao et al. (2019), fossil *Aldrovanda intermedia* and *A. ovata* (family Droseraceae) were used to set the prior for the most recent common ancestor (MRCA) of *Dionaea* and *Aldrovanda* with an offset of 41.2 Ma, and *Polygonocarpum johnsonii* was used to constrain the MRCA of the Polygonoideae (family Polygonaceae) included with an offset of 66.0 Ma. The MRCA of non-core Caryophyllales was constrained to 115 Ma with a normal distribution and a standard deviation of 4.0 Ma, representing the 95% confidence interval in the posterior distribution of the dating analysis of Yao et al. (2019). The Markov-Chain Monte Carlo (MCMC) was run for 100,000,000 generations, sampling every 1000 generations. The BEAST input file and data are available at 10.5281/zenodo.6081366. The resulting summary statistics were visualized in Tracer version 1.7.1 (Rambaut et al., 2018).

Similarly to *rbc*L, for species with multiple ITS sequences, the longest sequence was kept. Alignment and BEAST settings followed those above except that the *Polygonocarpum johnsonii* fossil was not used due to different taxon sampling for ITS,

and the root constraint was placed at the divergence of the carnivorous Caryophyllales from other non-core Caryophyllales represented by *Psylliostachys suworowii* (Family Plumbaginaceae).

For both *rbc*L and ITS, the obtained phylogenetic trees were summarized in TreeAnnotator version 2.6.2 (Drummond and Rambaut, 2007) with a 10% burn-in, and the maximum clade credibility tree was visualized in FigTree version 1.4.4 (Rambaut, 2018). The chronograms (ape R package; Paradis and Schliep, 2019) and chromosome count matrices were trimmed to species shared by both the gene and the chromosome datasets for subsequent analyses.

Modeling chromosome number evolution

We used the binary trait linked to chromosome number change model (BiChrom; Zenil-Ferguson et al. 2017) and implemented it in RevBayes software version 1.1.0 (Höhna et al., 2016) to estimate the differences in three rates of chromosome number evolution for each binary state (Fig. 1): γ (a single chromosome gain, by duplication or fission), δ (a single chromosome loss, by rearrangement, fusion, or loss), and ρ (a polyploidy event). The binary state is defined as whether a taxon belongs to *D.* subg. *Ergaleium* (state E) or not, in which case it belongs to *D.* subg. *Drosera*, *Arcturia*, or *Regiae* (state D). By defining our binary state in this fashion, we estimate a transition rate q, which is a nuisance parameter but allows us to correctly compare rates of chromosome change between the two groups using the phylogenetic structure of our estimated trees. Species were assigned as state E or state D *sensu* Fleischmann et al. (2018).

We first defined a Q-matrix describing the dynamics of chromosome number change between two chromosome numbers within a given state (E or D) or a change between the E and D states given a fixed chromosome number (Fig. S1; Mayrose et al., 2010; Zenil-Ferguson et al., 2017). The Q-matrix allows us to define a continuous-time Markov chain for the discrete trait of chromosome number. However, this Q-matrix can be numerically difficult to use because of its large dimensions and many rates being equal to zero (e.g., the instantaneous transition rate between $1n = 10$ to $1n = 17$ is zero since the change is not a doubling, or a single increase or decrease in chromosome number). Therefore, limiting the maximum number of chromosomes, hence smaller matrix

dimensions, is necessary for convergence of estimates (Zenil-Ferguson et al., 2018). Since our dataset had 2*n* chromosome numbers ranging from 8 to 80, we set the haploid (1*n*) chromosome number as states for the Q-matrix to range from 1 to 40 and a 40+ state for taxa with more than 1*n* = 40 to make it computationally feasible (Fig. S1; Zenil-Ferguson et al., 2017, 2018). We removed *Drosera lanata* (2*n* = 19) to avoid non-integer haploid chromosome numbers and records of B-chromosomes, as these small satellite chromosomes do not segregate normally during cell division. The resulting matrix had 82 rows and 82 columns reflecting 1 to 40 and more than 40 chromosome numbers for both states E and D (Fig. S1). Since we expect the chromosome evolution rate in *Drosera* outside of *D.* subg. *Ergaleium* to be more similar to the rate in most angiosperms, we considered state D the ancestral state and E the derived state and only allowed transitions from state D to state E. The probabilities of the root being 1 to more than 40 chromosomes in either state were set equal.

Three nested models were used for testing the difference of chromosome number evolution between *D.* subg *Ergaleium* (state E) and the rest of the genus (state D). The full model (H2) allowed rates (ρ = chromosome doubling, δ = chromosome loss, γ = chromosome gain) to vary independently in states D and E. The fixed-polyploid model (denoted as H1: $\rho_D = \rho_E$) constrained the rate of chromosome doubling to be the same between states D and E. Finally, the null model (H0) constrained all three rates to be equal for states D and E (H0: $\rho_D = \rho_E$, $\gamma_D = \gamma_E$, $\delta_D = \delta_E$). Rate prior distributions for all chromosome transition rates were defined using an exponential distribution with a mean equal to 3 changes per million years (Myr). The prior distribution had a large variance allowing for a wide range of initial potential values for transition rates.

We ran our custom MCMC scripts in RevBayes (Höhna et al., 2016) for 1,000,000 generations. Using Tracer (Rambaut et al., 2018), we ensured convergence had been reached and verified that effective sample sizes for all the parameters were above 200. Concurrently, for the best model, we reconstructed ancestral states using marginal posterior probabilities for each of the internal nodes as part of the inference following Freyman and Höhna (2018) and Zenil-Ferguson et al. (2019). The RevBayes input data and scripts are available from 10.5281/zenodo.6081366.

The three models were compared by estimating their marginal log-likelihoods to calculate the test statistic κ representing the Bayes factors, done in RevBayes as well (Höhna et al., 2016). The marginal likelihood, which is the probability of a model integrated over all the parameter space, allows us to assess model fit in a Bayesian framework similar to the AIC statistic in a likelihood framework (Xie et al., 2011). To compare models, we subtracted the marginal log-likelihood of a given pair of models, which is κ = log marginal likelihood of model 1 – log marginal likelihood of model 2. We consider $\kappa > 6$ as evidence for strong support for model 1, $\kappa > 1$ as moderate support for model 1, a value of κ between -1 and 1 as no evidence in favor of either model, and $\kappa < -1$ as support for model 2 (Kass and Raftery, 1995).

All the MCMC outputs were analyzed using Tracer with the first 10% discarded as burn-in. The ancestral state reconstruction results for the best supported model were visualized with the RevGadgets R package (Fig. 4; Tribble et al., 2021).

Branch length and topology uncertainty

To evaluate the effect of phylogenetic uncertainty on the estimated rates, the best BiChrom model (H1) was fitted to the last ten *rbc*L trees sampled in BEAST and on the ITS chronogram. Before running, *D. indica* and *D. collinsiae* were removed from the ITS phylogeny due to their placement outside the corresponding phylogenetically defined sections (Fleischmann et al., 2018). The MCMC outputs of both analyses were analyzed using Tracer with a burn-in of 10% discarded.

Genome size and mating system

Self-compatibility data for 98 species of *Drosera* were obtained from publications (Table S3). Recent studies (Fleischmann, in press) suggest all *D. auriculata* are self-compatible, contrary to a doubtful previously-published report by Chen et al. (1997). *Drosera* genome sizes were obtained from Veleba et al. (2017) or newly generated in this study for 17 species at the Flow Cytometry Core Lab at the Benaroya Research Institute (Seattle, WA, U.S.A.). For each genome size, four flow cytometry measurements were taken against a known size standard. Source, voucher, and size standards used for

generating new flow cytometry data are listed in Table S3. We used the average genome size for each species for subsequent analyses.

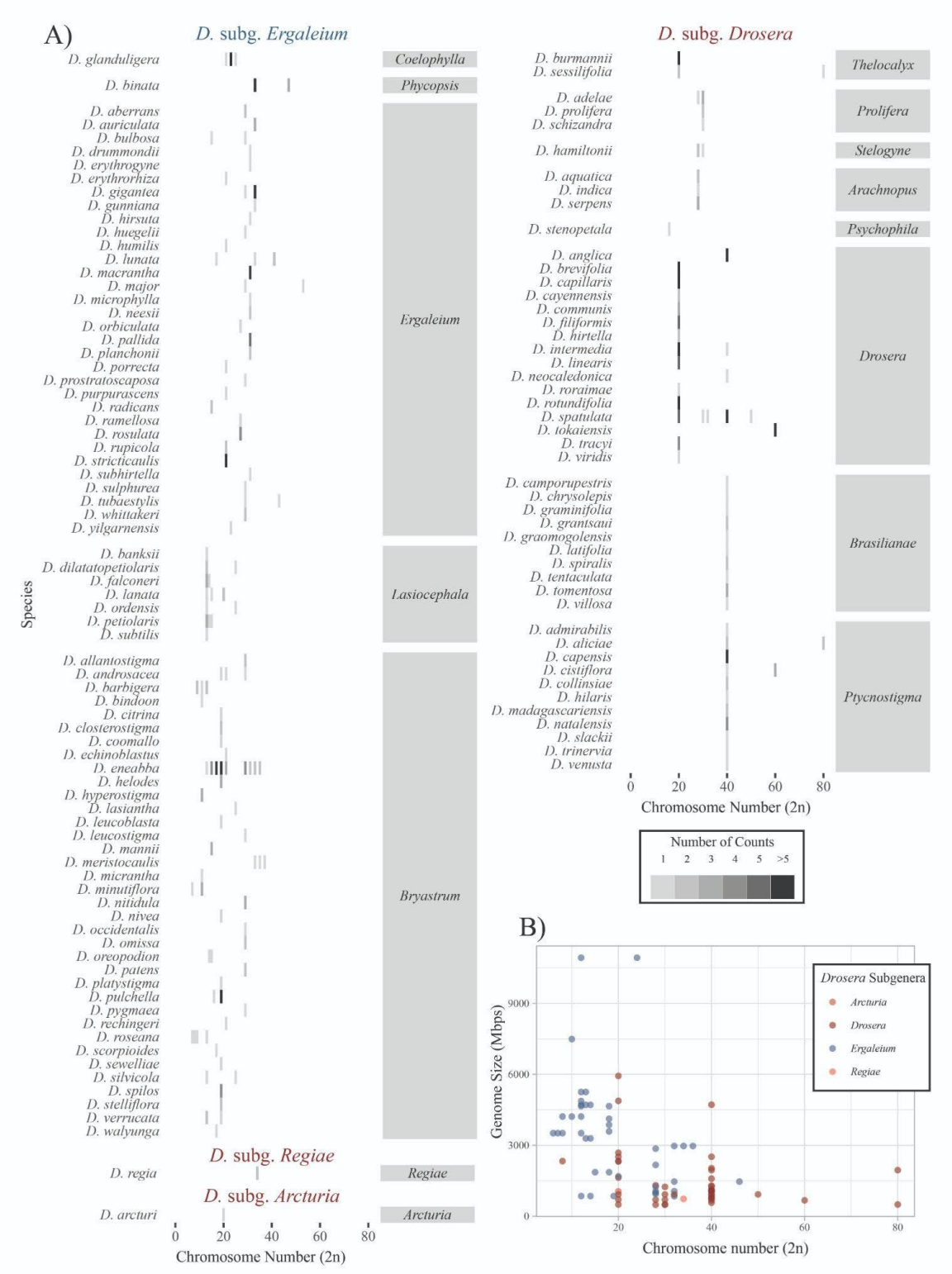


Figure 2: Chromosome and genome size variation in *Drosera.* (A) *Drosera* subg. *Ergaleium* (left) exhibited marked single chromosome number variation both among and within species. In contrast, both among- and within-species chromosome number variation in *D.* subg. *Drosera* (right) fell primarily into polyploidy series. The shade of the bar indicates the number of samples for each species, emphasizing that the lower level of variation in *D.* subg. *Drosera* is not due to a lack of counts. (B) *Drosera* species with larger chromosome numbers tend to have smaller genome sizes.

RESULTS

Chromosome Counts for 127 Drosera *species show distinctive patterns of variation between* D. *subgenus* Ergaleium *and other subgenera*

An initial dataset of 676 chromosome counts in *Drosera* from 150 species or hybrids was compiled (Table S1). After removing hybrids and low-quality counts, 510 counts from 127 species were used for downstream analyses. These counts included 48% of all named species in *Drosera*. Across its geographic distributions, the filtered counts included 32% of named species from Africa, 45% from South America, 51% from Australia, 60% from Asia, and all species from North America and Europe.

Among the four subgenera, *Drosera* subg. *Arcturia*, *D.* subg. *Drosera*, *D.* subg. *Ergaleium*, and *D.* subg. *Regiae* each had 50%, 43%, 51%, and 100% of named species represented. Almost every even chromosome number from $2n = 6$ to 46 was reported from *D.* subg. *Ergaleium*, and similar scattered chromosome number variation was observed within 21 species (Fig. 2A). In contrast, *D.* subg. *Drosera* has chromosome numbers from $2n = 16$ to 80 with variation primarily in polyploid series both within and among species ($2n = 20$, 30, 40, 60, 80; Fig. 2A). Despite more counts have been reported in *D.* subg. *Drosera*, only seven species have within-species chromosome number variation reported. Chromosome number for *D. arcturi* (*D.* subg. *Arcturia*) was $2n = 20$ and for *D. regia* (*D.* subg. *Regiae*) was $2n = 34$.

Chronogram Reconstruction

The trimmed *rbc*L matrix included 1,440 bases with 478 variable sites across the 17 outgroup and 79 ingroup species. The trimmed ITS matrix included 1,133 bases with 783 variable sites across 7 outgroup and 50 ingroup species. After burn-in, the ESS was greater than 200 for all statistics in both ITS and *rbc*L analyses. The *rbc*L tree placed *D.*

regia in a clade with *Aldrovanda* and *Dionaea* with strong to moderate support. The ITS tree placed *D. regia* sister to the rest of *Drosera*, consistent with cladogram from Fleischmann et al. (2018). BEAST analyses estimated the crown age of *Drosera* (including *D. regia*) at around 69.9 Mya based on *rbc*L and 80.1 Mya based on ITS with overlapping confidence intervals.

Drosera *subgenus* Ergaleium *differs from other subgenera in single-chromosome evolution rates*

The chromosome counts and *rbc*L data overlapped for 59 species: 25 from *D.* subg. *Ergaleium*, 32 from *D.* subg. *Drosera,* and one species each from *D.* subg. *Arcturia* and *D.* subg. *Regiae*.

In the full model (H2), the mean posterior rate of gaining ($\gamma_E = 0.16$ per one million year) or losing ($\delta_E = 0.17$) one chromosome in *D.* subg. *Ergaleium* was 7.3-fold and 370-fold higher than other subgenera ($\gamma_D = 0.022$; $\delta_D = 0.00046$; Table S4; Fig. 3). However, the rate of chromosome gain for *D.* subg. *Drosera*, *Arcturia,* and *Regiae* fell within the first quartile of the rate of chromosome gain for *D.* subg. *Ergaleium* and only the 95% credible interval for the rates of single chromosome loss was distinct (95% HPD $\delta_E = 0.036$ to 0.36; 95% HPD $\delta_D = 3.6\times10^{-7}$ to 1.0×10^{-3}; Table S4; Fig. 3). The rates of polyploidy largely overlapped (Fig. 3).

Compared to rates estimated in the full model, the null model (H0) estimated an intermediate rate for losing one chromosome, while the estimated rate of polyploidy doubled and the rate for gaining a chromosome decreased (Fig. 3). Comparing Bayes factors for the full model and null model on the *rbc*L results strongly favored the full model ($\kappa = 15.0$), supporting that chromosome evolution rates were different between *D.* subg. *Ergaleium* and the other subgenera.

Given the largely overlapping polyploidy rates for state D vs. E, we tested an additional model H1, which linked the polyploidy rates for state D and E, but estimated rates for chromosome loss and gain for the two states separately. We found a moderate preference for H1 over the full model (H2; $\kappa = 5.9$; Table S4; Fig. 3).

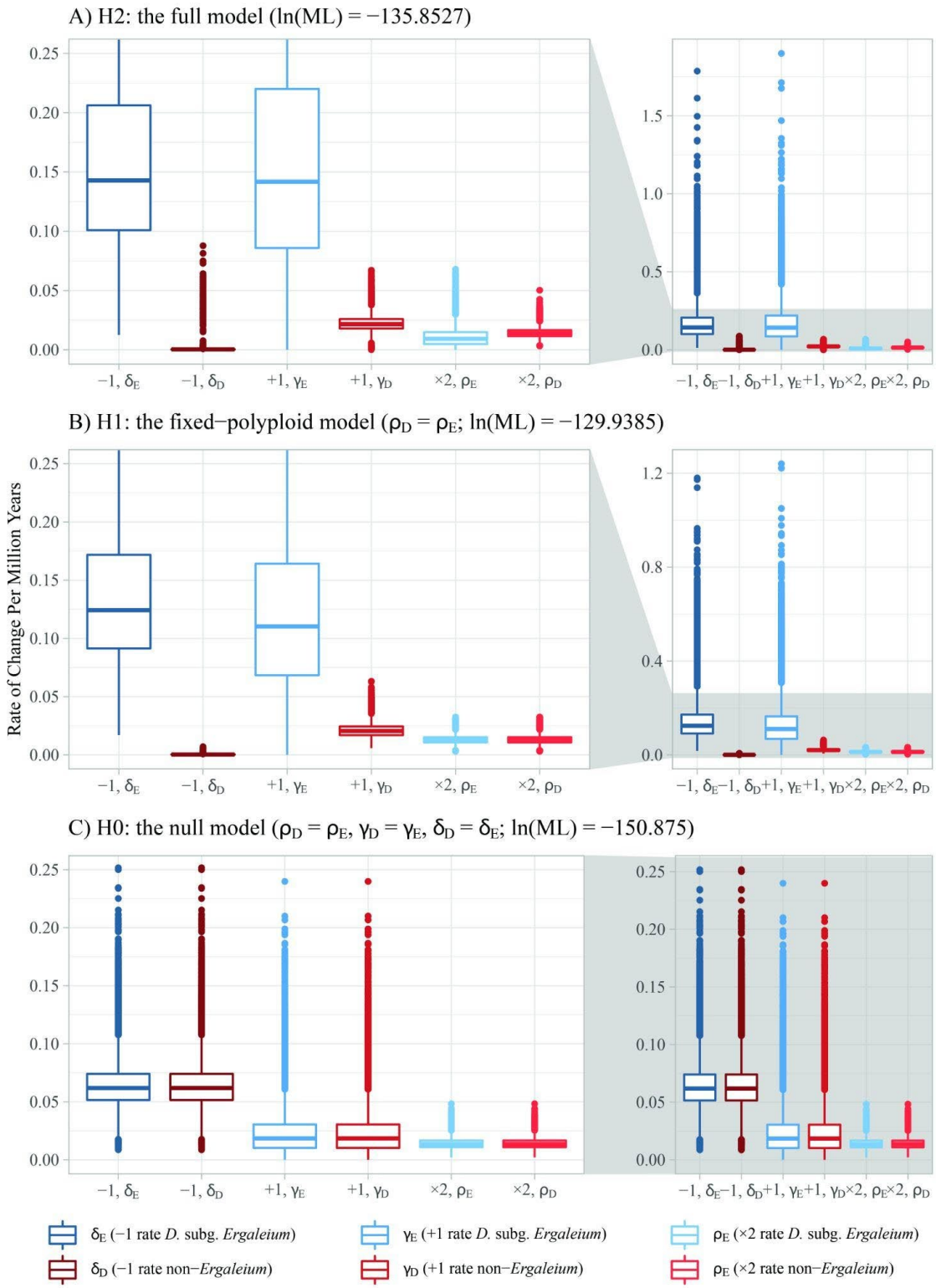


Figure 3: The posterior distribution of chromosome evolution rates for three BiChrom models. These models are (A) H2, where all rates were estimated independently for *Drosera* subg. *Ergaleium* (state E) versus the other three *Drosera* subgenera (state D); (B) H1, where all rates

were independent except ρ (polyploidy) being equal across *Drosera*; and (C) H0, where all rates were equal across *Drosera*. δ_D and δ_E were significantly distinct in H2 and H1. All remaining rates were not significantly different between state E and D.

The best fit model (H1) with separate chromosome loss (δ) and gain (γ) rates for state D vs. E but equal ploidy, showed both higher chromosome loss and chromosome gain rates in *D*. subg. *Ergaleium* and 95% credible intervals similar to the full model (Table S4; Fig. 3). The mean δ_E was 358-fold higher than δ_D, and the 95% HPD did not overlap (Table S4; Fig. 3). With overlapping 95% HPDs, the mean γ_E was over 6.0-fold higher than γ_D (Table S4; Fig. 3).

Under the H1 model, the ancestral state reconstruction estimated the most probable value of the MRCA of *Drosera* to be a haploid chromosome number of eight. The base of *D*. subg. *Ergaleium* also had a most probable haploid chromosome number of eight. The difference in single chromosome change between subgenera is supported across the reconstruction by the stability of chromosome number in *D*. subg. *Drosera* and repeated changes in *D*. subg. *Ergaleium.* Based on the reconstruction, polyploidization events occurred five times in *D*. subg. *Ergaleium*, three times in *D*. subg. *Drosera,* and once in *D*. subg. *Regiae* (Fig. 4).

Results from rbcL were robust when considering phylogenetic uncertainty and when using the ITS dataset

The results of the H1 model (fixed-polyploidy) on the ten *rbc*L trees from the BEAST MCMC sampling all found higher single chromosome gain and loss in *D.* subg. *Ergaleium* than the other subgenera despite differences in branch lengths and topology (Fig. S3). The ITS BiChrom results had higher levels of uncertainty, likely due to only 47 species overlapping between the chromosome count and ITS data after filtering. Nonetheless the ITS dataset once again found higher rates of single chromosome gain and loss in *D.* subg. *Ergaleium* than the other subgenera. The rates of gaining ($\gamma_E = 0.11$) or losing ($\delta_E = 0.11$) one chromosome in *D.* subg. *Ergaleium* were both 8-fold higher than those of the other subgenera ($\gamma_D = 0.014$; $\delta_D = 0.013$; Fig S3).

Self-compatibility differs between **Drosera** *subgenera*

In *D*. subg. *Ergaleium,* 48 of the 60 (80%) species with known mating systems are self-incompatible in at least some populations (Fig. 4; Table S3.2). In contrast, only three distantly related species of the 38 species (8%) in the remaining three subgenera are self-incompatible (Fig. 4; Table S3.2).

Genome size decreases as chromosome number increases across **Drosera**

Our newly generated genome size estimates ranged from 630 to 5249 Mbps (Table S3.1). Many were similar to previous publications, but a few appear to be polyploids such as *D. spatulata.* Across *Drosera,* genome size remained the same or decreased as chromosome number increases (Fig. 2B). By visually comparing the genome size of polyploid taxa and those of the closely related diploid taxa, the polyploid taxa generally have similar or smaller genome sizes except in the more recent polyploid event of *D. anglica* (Fig. 4).

DISCUSSION

Rates of single chromosome number change significantly differ among **Drosera** *subgenera*

In this study, we carefully reviewed primary cytological literature and voucher information to correct for counting and taxonomic issues. We then modeled chromosome evolution taking both time and phylogenetic history into consideration. We found that the rate of polyploidy in *Drosera* (0.014 per Myr) did not significantly differ between subgenera and was very similar to the polyploidy rate previously reported for perennial angiosperms (0.015 per Myr; Van Drunen and Husband, 2019) and across angiosperm families (median 0.025 per Myr; Zhan et al., 2021). The single chromosome gain (0.021) and loss rate (0.00040) for *Drosera* subgenera other than *D.* subg. *Ergaleium* fell higher and lower, respectively, than the average rate (0.0061 and 0.016 respectively) across angiosperm families (Zhan et al., 2021). In contrast, the rate of single chromosome number shifts in *D.* subg. *Ergaleium* was 6-fold (gain) and 350-fold (loss) higher than in the remainder of the genus, and the single chromosome evolution rates in *D.* subg. *Ergaleium* are likely even higher with increased species sampling.

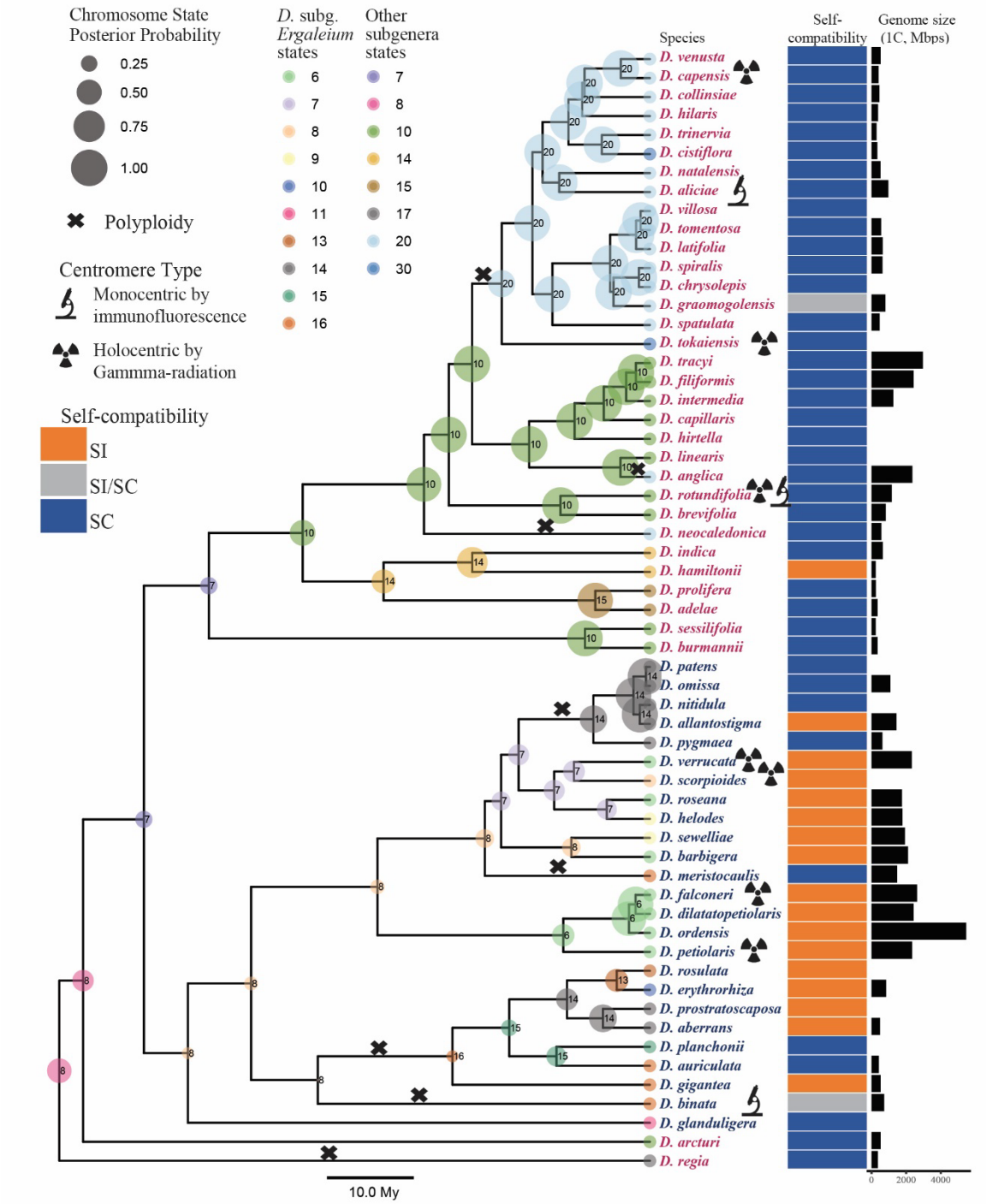


Figure 4: Ancestral state reconstruction of chromosome number evolution using model H1 in RavBayes. In addition to having higher rates of single chromosome change, *Drosera* subg. *Ergaleium* (species name in blue) have more species that are self-incompatible than the other three subgenera (species name in red). Lineages with a polyploidy history (black ×) tend to have smaller genome sizes than their sister lineages in our sampling. Species with experimental evidence for their centromere type are distributed across the genus, but results from gamma-radiation versus immunofluorescence disagree about the type of centromere in *Drosera*, or even within the same species as in the case of *D. rotundifolia.*

The pattern of elevated single chromosome evolution rate, especially single chromosome loss in *D*. subg. *Ergaleium* is robust to phylogenetic uncertainty, taxon sampling, and gene tree discordance. This pattern remains unchanged in rates estimated from the last ten trees of the *rbc*L MCMC sampling. Despite a smaller taxon sampling and difference in tree topology in the nuclear ITS dataset compared to the chloroplast *rbc*L dataset, the rates estimated from the two loci are similar and, more importantly, follow the same trend. Our *rbc*L analysis included 23% of all currently recognized *Drosera* species, and ITS analysis had 18%, resulting in wider credible intervals in estimated rates. If additional loci support the polyphyly of *Drosera* with *D. regia* being more closely related to *Dionaea* and/or *Aldrovanda* than the rest of *Drosera*, the rate estimations and ancestral chromosome state is unlikely to change significantly. This is because the only other species in the family Droseraceae*, Dionaea muscipula* (2*n* = 30 or 32; Rivadavia et al., 2003) and *Aldrovanda vesicula* (2*n* = 38; Rivadavia et al., 2003), have chromosome counts similar to that of *Drosera regia* (2*n* = 34; Table S1.1). Increasing taxon sampling will likely recover additional single chromosome change events and reduce the number of inferred polyploidy events in *D. subg. Ergaleium*. On the other hand, reticulate evolution, especially allopolyploidy events that result in ×1.5 chromosome change (e.g., *D. tokaiensis*, *D.* subg. *Drosera*; Nakamura and Ueda, 1991), may lead to underestimation of polyploidy rate and overestimation of single chromosome change rate. The different modes of chromosome number changes among subgenera of *Drosera* is further supported by the pattern of within-species variation being primarily single chromosome changes in *D*. subg. *Ergaleium* in contrast to being primarily polyploidy series in *D*. subg. *Drosera*, and the lack of within-species variation in the two remaining subgenera. Therefore, considering the caveats of our taxon sampling and modeling approach, increasing the taxon sampling and using additional nuclear genes will likely narrow the credible intervals but unlikely to draw a different conclusion on the drastic difference in single chromosome changes among subgenera of *Drosera*.

Elevated rates of single chromosome evolution can be due to increased rates of polyploidy and subsequent rediploidization (Mandáková and Lysak, 2018). However, we did not find evidence for difference in rates of polyploidy among subgenera in *Drosera*. Although polyploid species in *D.* subg. *Drosera* were considered stable polyploids as

their chromosome numbers follow polyploid series (Hoshi and Kondo, 1998; Shirakawa, et al., 2011), we found evidence for genome downsizing after polyploidy across *Drosera.* Of the nine polyploidy events inferred, the most recent has a genome size close to double that of the sister lineage, while the remaining eight more ancient polyploid lineages have similar or, in seven cases, smaller genome sizes than their closely related diploid lineages (Fig. 4; Table S3.1; Veleba et al., 2017). Therefore, our analysis did not recover any difference in rates of polyploidy or post-polyploidy diploidization patterns among *Drosera* subgenera, and there is no evidence to support either being the major cause of single chromosome number shifts in the genus.

Potential drivers of chromosome evolution rate shift

Similar orders of magnitude differences in chromosome loss and gain rates have also been documented between herbaceous versus woody plants, and also among some *Carex* lineages and among some insect lineages that have holocentric chromosomes (Escudero et al., 2014; Zenil-Ferguson et al., 2017; Ruckman et al., 2020; Sylvester et al., 2020). Holocentromeres have been associated with increased tolerance of chromosome fission (Cuacos et al., 2015; Ruckman et al., 2020) as the resulting chromosome fragments with centromeres can pair and segregate properly even in heterozygous individuals (Luceño and Guerra, 1996; Jankowska et al., 2015; Ruckman et al., 2020). Experimental investigation of centromere type in *Drosera* has been limited to a small number of species (Fig. 4) using indirect methods. Two indirect methods, response to gamma-radiation-induced breakages and the distribution of a histone commonly associated with the centromeric or pericentric region, both supported all tested species having the same centromere type. Therefore, so far no evidence supports the presence of holocentromeres as the cause of the heterogeneity in chromosome evolution rate in *Drosera*, and more direct experimental investigations are needed. A similar lack of association between holocentric chromosomes and significant differences in chromosome evolution rates has also been documented in a study across 22 orders of insects (Ruckman et al., 2020).

A newly formed karyotype may be eliminated due to drift or selection against the deleterious nature of heterozygous individuals, especially in monocentric plants

(Husband et al., 2013). Species that are self-compatible (or have other reproductive assurances such as clonal propagation) may avoid these issues as the proportion of individuals in the population with the new chromosome number can increase without producing heterozygous individuals (Husband et al., 2013; Van Drunen and Husband, 2019; Spoelhof, Keeffe, et al., 2020). While a perennial life history and clonal propagation are common across *Drosera* (Fleischmann et al., 2018), contrary to expectation, a higher percentage of species studied in *D.* subg. *Ergaleium* are self-incompatible compared to the other subgenera (Fig 4; Table S3). Interestingly, this supports the hypothesis of Spoelhof, Keeffe, et al. (2020) that sexual reproduction, especially outcrossing, is important for the long-term maintenance of genetic diversity after the bottleneck when a new karyotype forms. These two seemingly contradictory arguments related to mechanisms underlying new karyotype establishment await future studies at the intraspecific level investigating factors including population size, spatial distribution, and meiotic drive (Reed et al., 2013; Bureš and Zedek, 2014; Blackmon et al., 2019; Ruckman et al., 2020; Spoelhof, Soltis, et al., 2020; Griswold, 2021).

Conclusion

In this study we found highly elevated rates in single chromosome evolution but not polyploidy in *Drosera* subg. *Ergaleium* compared to the rest of the genus. This pattern is robust to taxon sampling and the phylogeny used, and is not an artifact of errors or clade age. In addition to the 6-fold and 358-fold higher rates of gain and loss compared with other subgenera, respectively, *Drosera* subg. *Ergaleium* harbors a higher percentage of self-incompatible species (80% compared with 8% for species from other subgenera). More broadly, our findings suggest that factors other than holocentromeres and genome downsizing after polyploidy impact the rate of single chromosome number evolution. Because chromosome number change is a key driver of speciation, future work to tease apart the natural history and molecular mechanisms underlying lineages with highly elevated rates of chromosome number change would further our understanding of evolution at both the macro- and microevolutionary scales.

SUPPLEMENTAL MATERIALS:

Figure S1: The transition matrix in the BiChrom model. See Fig. 1 for definition of chromosome transition parameters. The parameter q_{DE} represents the instantaneous transition rate from state D to state E.

Figure S2: Maximum clade credibility tree from BEAST analyses of the *rbc*L (A) and the ITS (B) dataset. Bars on nodes represent the 95% HPD intervals for the age of the node.

Figure S3: Posterior probability distribution of chromosome evolution rates estimated from the 10 *rbc*L trees and the consensus ITS chronogram.

Table S1: The full chromosome count data matrix with notes. Table S1.1 is the matrix itself, Table S1.2 contains the header key and additional information, and Table S1.3 contains the references.

Table S2: Source for sequence data, the species name used, the GenBank ID, the originally reported species name, and notes for taxonomic updates. Table S2.1: *rbc*L; Table S2.2: ITS; Table S2.3: naming authority for each *Drosera* species.

Table S3: The genome size (Table S3.1), self-compatibility (Table S3.2), experimental evidence for centromere type (Table S3.3) and reference (Table S3.4). The genome size matrix included species names, locality, and herbarium voucher information (visit 10.5281/zenodo.6081366 for photos), size standard, and reference. The self-compatibility data included species, reference, and notes on changes in taxonomy.

Table S4: Summary statistics of the three BiChrom models estimated from the rbcL chronogram. All rates are reported in the probability of change per Myr and followed by their 95% HPD distributions in parentheses.

Supplemental Information S1: Methods for the chromosome count scoring and filtering.

REFERENCES:

Baniaga, A. E., H. E. Marx, N. Arrigo, and M. S. Barker. 2019. Polyploid plants have faster rates of multivariate niche differentiation than their diploid relatives. *Ecology Letters* 23: 68–78.

Blackmon, H., J. Justison, I. Mayrose, and E. E. Goldberg. 2019. Meiotic drive shapes rates of karyotype evolution in mammals. *Evolution* 73: 511–523.

Bouckaert, R., J. Heled, D. Kühnert, T. Vaughan, and C.-H. Wu. 2014. BEAST 2: A

software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* 10: 1003537.
Bureš, P., and F. Zedek. 2014. Holokinetic drive: centromere drive in chromosomes without centromeres. *Evolution* 68: 2412–2420.
Chen, L. 1998. Chromosome numbers, breeding systems and genetic diversity in several species of *Drosera* (Droseraceae) from the south-west of Western Australia. unpublished PhD Thesis, The University of Western Australia.
Chen, L., S. H. James, and H. M. Stace. 1997. Self-incompatibility, seed abortion and clonality in the breeding systems of several Western Australian *Drosera* species (Droseraceae). *Australian Journal of Botany* 45: 191.
Coghlan, A., E. Eichler, S. Oliver, A. Paterson, and L. Stein. 2005. Chromosome evolution in eukaryotes: a multi-kingdom perspective. *Trends in Genetics* 21: 673–682.
Coyne, J. A., and H. A. Orr. 2004. *Speciation*. Sinauer Associates, Inc., Sunderland.
Cuacos, M., H. Franklin, F. Chris, and S. Heckmann. 2015. Atypical centromeres in plants—what they can tell us. *Frontiers in Plant Science* 6: 913.
Dawson, M. I. 2000. Index of chromosome numbers of indigenous New Zealand spermatophytes. *New Zealand Journal of Botany* 38: 47–150.
Demidov, D., V. Schubert, K. Kumke, O. Weiss, R. Karimi-Ashtiyani, J. Buttlar, S. Heckmann, et al. 2014. Anti-phosphorylated histone H2AThr120: a universal microscopic marker for centromeric chromatin of mono- and holocentric plant species. *Cytogenetic and genome research* 143: 150–156.
Drummond, A. J., and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7: 1–8.
Van Drunen, W. E., and B. C. Husband. 2019. Evolutionary associations between polyploidy, clonal reproduction, and perenniality in the angiosperms. *New Phytologist* 224: 1266–1277.
Escudero, M., S. Martín-Bravo, I. Mayrose, M. Fernández-Mazuecos, O. Fiz-Palacios, A. L. Hipp, M. Pimentel, et al. 2014. Karyotypic changes through dysploidy persist longer over evolutionary time than polyploid changes. *PLoS ONE* 9: e85266.
Fleischmann, A. (in press). Flower biology of the carnivorous sundews (*Drosera*, Droseraceae). Frontiers in Plant Science (in press).
Fleischmann, A., A. T. Cross, R. Gibson, P. M. Gonella, and K. W. Dixon. 2018. Systematics and evolution of Droseraceae. *In* A. Ellison, and L. Adamec [eds.], *Carnivorous Plants: Physiology, Ecology, and Evolution*, 45–57. Oxford University Press, Oxford.
Freyman, W. A., and S. Höhna. 2018. Cladogenetic and anagenetic models of chromosome number evolution: A Bayesian model averaging approach. *Systematic Biology* 67: 195–215.
Furuta, T., and K. Kondo. 1999. Effects of gamma-rays on diffused-centromeric chromosomes of *Drosera falconerii* [*sic*] in vitro. *Chromosome Science* 3: 93–100.
Goldblatt, P., and D. E. Johnson [eds.]. 1979–. *Index to plant chromosome numbers*. Missouri Botanical Garden, St. Louis.
Grant, V. 1981. *Plant Speciation.* Columbia University Press.
Griswold, C. K. 2021. The effects of migration load, selfing, inbreeding depression, and the genetics of adaptation on autotetraploid versus diploid establishment in

peripheral habitats. *Evolution* 75: 39–55.
Höhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology* 65: 726–736.
Hoshi, Y., and K. Kondo. 1998. Chromosome differentiation in *Drosera*, subgenus *Rorella*, section *Rossolis*. *Cytologia* 63:199-211
Huie, L. 1897. Changes in the cell-organs of *Drosera rotundifolia* produced by feeding with egg-albumen. *Quarterly Journal of Microscopical Science* 39: 387–425.
Husband, B. C., S. J. Baldwin, and J. Suda. 2013. The incidence of polyploidy in natural plant populations: major patterns and evolutionary processes. *In* I. J. Leitch, J. Greilhuber, J. Dolezel, and J. F. Wendel [eds.], *Plant Genome Diversity*, 255–276. Springer.
Jankowska, M., J. Fuchs, E. Klocke, M. Fojtová, P. Polanská, J. Fajkus, V. Schubert, and A. Houben. 2015. Holokinetic centromeres and efficient telomere healing enable rapid karyotype evolution. *Chromosoma* 124: 519–528.
Kass, R. E., and A. E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90: 773–795.
Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* 30: 772–80.
Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, et al. 2012. Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28: 1647–1649.
Kolodin, P., H. Cempírková, P. Bureš, L. Horová, A. Veleba, J. Francová, L. Adamec, and F. Zedek. 2018. Holocentric chromosomes may be an apomorphy of Droseraceae. *Plant Systematics and Evolution* 304: 1289–1296.
Kondo, K. 1969. Chromosome numbers of carnivorous plants. *Bulletin of the Torrey Botanical Club* 96: 322–328.
Kress, A. 1970. Zytotaxonomische Untersuchungen an einigen Insektenfängern (Droseraceae, Byblidaceae, Cephalotaceae, Roridulaceae, Sarraceniaceae). *Berichte der Deutschen Botanischen Gesellschaft* 83: 55–62.
Luceño, M., and M. Guerra. 1996. Numerical variations in species exhibiting holocentric chromosomes: a nomenclatural proposal. *Caryologia* 49: 301–309.
Mandáková, T., and M. A. Lysak. 2018. Post-polyploid diploidization and diversification through dysploid changes. *Current Opinion in Plant Biology* 42: 55–65.
Mayrose, I., M. S. Barker, and S. P. Otto. 2010. Probabilistic models of chromosome number evolution and the inference of polyploidy. *Syst. Biol* 59: 132–144.
Mayrose, I., S. Zhan, C. Rothfels, K. Magnuson-Ford, M. Barker, L. Rieseberg, and S. Otto. 2011. Recently formed polyploid plants diversify at lower rates. *Science* 333: 1257.
Mayrose, I., and M. A. Lysak. 2020. The evolution of chromosome numbers: mechanistic models and experimental approaches. *Genome Biology and Evolution* 13: evaa220
Nakamura, T., and K. Ueda. 1991. Phytogeography of Tokai Hilly Land Element II. Taxonomic study of *Drosera tokaiensis* (Komiya & C. Shibata) T. Nakamura & Ueda (Droseraceae). *Acta phytotaxonomica et geobotanica* 42: 125–137.

Paradis E. & Schliep K. 2019. ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35: 526–528.
Rambaut, A. 2018. FigTree v. 1.4.4. Website http://tree.bio.ed.ac.uk/software/figtree/ [accessed 18 June 2021].
Rambaut, A., A. J. Drummond, D. Xie, G. Baele, and M. A. Suchard. 2018. Posterior Summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology* 67: 901–904.
Reed, F. A., A. Traulsen, and P. M. Altrock. 2013. Underdominance. *Brenner's Encyclopedia of Genetics: Second Edition*: 247–249.
Rice, A., L. Glick, S. Abadi, M. Einhorn, N. M. Kopelman, A. Salman-Minkov, J. Mayzel, et al. 2015. The Chromosome Counts Database (CCDB) – a community resource of plant chromosome numbers. *New Phytologist* 206: 19–26.
Rivadavia, F. 2005. New chromosome numbers for *Drosea* L. (Droseraceae). *Carnivorous Plant Newsletter* 34: 85–91.
Rivadavia, F., K. Kondo, M. Kato, and M. Hasebe. 2003. Phylogeny of the sundews, *Drosera* (Droseraceae), based on chloroplast *rbc*L and nuclear 18S ribosomal DNA sequences. *American Journal of Botany* 90: 123–130.
Rivero, R., E. B. Sessa, and R. Zenil-Ferguson. 2019. EyeChrom and CCDBcurator: Visualizing chromosome count data from plants. *Applications in Plant Sciences* 7(1): e1207.
Román-Palacios, C., Y. F. Molina-Henao, and M. S. Barker. 2020. Polyploids increase overall diversity despite higher turnover than diploids in the Brassicaceae. *Proceedings of the Royal Society B: Biological Sciences* 287.
Rosenberg, O. 1903. Das Verhalten der Chromosomen in einer hybriden Pflanze. *Berichte der Deutschen Botanischen Gesellschaft* 21: 110–119.
Rothfels, K., and M. Heimburger. 1968. Chromosome size and DNA values in sundews (Droseraceae). *Chromosoma* 25: 96–103.
Ruckman, S. N., M. M. Jonika, C. Casola, and H. Blackmon. 2020. Chromosome number evolves at equal rates in holocentric and monocentric clades. *PLOS Genetics* 16: e1009076.
Sheikh, S. A., and K. Kondo. 1995. Differential staining with orcein, giemsa, CMA, and DAPI for comparative chromosome study of 12 species of Australian *Drosera* (Droseraceae). *American Journal of Botany* 82: 1278–1286.
Sheikh, S. A., K. Kondo, and Y. Hoshi. 1995. Study on diffused centromeric nature of *Drosera* chromosomes. *Cytologia* 60: 43–47.
Shirakawa, J., Y. Hoshi, and K. Kondo. 2011. Chromosome differentiation and genome organization in carnivorous plant family Droseraceae. *Chromosome Botany* 6: 111–119.
Spoelhof, J. P., R. Keeffe, and S. F. McDaniel. 2020. Does reproductive assurance explain the incidence of polyploidy in plants and animals? *New Phytologist* 227: 14–21.
Spoelhof, J. P., D. E. Soltis, and P. S. Soltis. 2020. Habitat Shape Affects Polyploid Establishment in a Spatial, Stochastic Model. *Frontiers in Plant Science* 0: 1770.
Stebbins, G. L. 1971. *Chromosome Evolution in Higher Plants*. Edward Arnold Ltd, London.
Sylvester, T., C. E. Hjelmen, S. J. Hanrahan, P. A. Lenhart, J. S. Johnston, and H.

Blackmon. 2020. Lineage-specific patterns of chromosome evolution are the rule not the exception in Polyneoptera insects. *Proceedings of the Royal Society B* 287: 20201388

Tribble, C. M., W. A. Freyman, M. J. Landis, J. Y. Lim, J. Barido-Sottani, B. T. Kopperud, S. Höhna, and M. R. May. 2021. RevGadgets: an R Package for visualizing Bayesian phylogenetic analyses from RevBayes. *bioRxiv*: 2021.05.10.443470.

Veleba, A., P. Šmarda, F. Zedek, L. Horová, J. Šmerda, and P. Bureš. 2017. Evolution of genome size and genomic GC content in carnivorous holokinetics (Droseraceae). *Annals of Botany* 119: 409–416.

Wanner, G., E. Schroeder-Reiter, W. Ma, A. Houben, and V. Schubert. 2015. The ultrastructure of mono- and holocentric plant centromeres: an immunological investigation by structured illumination microscopy and scanning electron microscopy. *Chromosoma* 124: 503–517.

Weiss-Schneeweiss, H., K. Emadzade, T.-S. Jang, and G. M. Schneeweiss. 2013. Evolutionary consequences, constraints and potential of polyploidy in plants. *Cytogenetic and Genome Research* 140: 137–150.

Windham, M. D., and G. Yatskievych. 2003. Chromosome studies of cheilanthoid ferns (Pteridaceae: Cheilanthoideae) from the western United States and Mexico. *American Journal of Botany* 90: 1788–1800.

Xie, W., P. O. Lewis, Y. Fan, L. Kuo, and M. H. Chen. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology* 60: 150–160.

Yao, G., J.-J. J. Jin, H.-T. T. Li, J.-B. B. Yang, V. S. Mandala, M. Croley, R. Mostow, et al. 2019. Plastid phylogenomic insights into the evolution of Caryophyllales. *Molecular Phylogenetics and Evolution* 134: 74–86.

Zedek, F., P. Veselý, L. Horová, and P. Bureš. 2016. Flow cytometry may allow microscope-independent detection of holocentric chromosomes in plants. *Scientific Reports* 6: 27161.

Zenil-Ferguson, R., J. M. Ponciano, and J. G. Burleigh. 2017. Testing the association of phenotypes with polyploidy: An example using herbaceous and woody eudicots. *Evolution* 71: 1138–1148.

Zenil-Ferguson, R., J.G. Burleigh, and J.M. Ponciano, 2018. chromploid: An R package for chromosome number evolution across the plant tree of life. *Applications in Plant Sciences*, 6: p.e1037.

Zenil-Ferguson, R., J. G. Burleigh, W. A. Freyman, B. Igić, I. Mayrose, and E. E. Goldberg. 2019. Interaction among ploidy, breeding system and lineage diversification. *New Phytologist* 224: 1252–1265.

Zhan, S. H., S. P. Otto, and M. S. Barker. 2021. Broad variation in rates of polyploidy and dysploidy across flowering plants is correlated with lineage diversification. *bioRxiv*: 2021.03.30.436382.

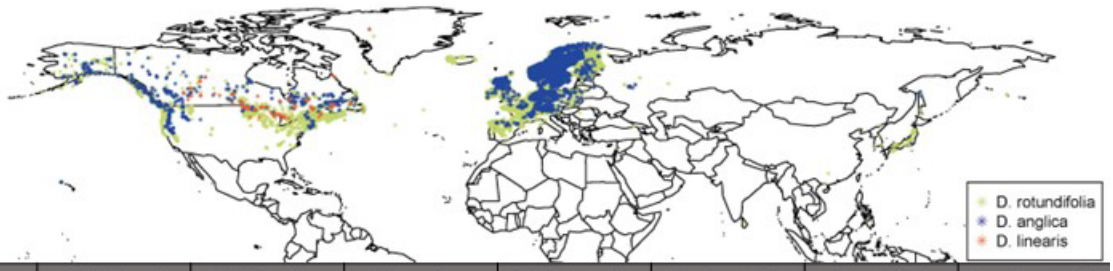
Chapter 2: Phylogenomic analyses of North American species of *Drosera* L. (Droseraceae) with a special emphasis on the origin of the allopolyploid *Drosera anglica*

INTRODUCTION

Drosera L. (Droseraceae, Caryophyllales) is a carnivorous plant genus of ~250 species found around the world. Twenty-eight of those species are classified in *D.* sect. *Drosera* based on molecular, cytological, and morphological evidence (Fleischmann et al., 2018). This section is the most geographically diverse of the genus, being found in Europe, Asia into Oceania, North America, and South America, but its highest species diversity is in North and South America (Fleischmann et al., 2018; Lowrie et al., 2017). There are eight species of *Drosera* that are native to North America, all of them belong to *D.* sect. *Drosera* (Fig. 1). These eight species have two general geographic distributions. Five species have a north-south distribution along the East Coast of North America with three of these extending into South America. Among them, *Drosera intermedia* is found in eastern North America, South America, and Europe. The remaining three species have east-west distributions in the boreal zone: two of which are circumboreal with disjunct populations in tropical mountains, and the third species, *Drosera linearis*, is restricted to boreal North America (Fig. 1).

All North American *Drosera* species have a diploid chromosome number of $2n$ = 20 except *D. anglica*, which has a tetraploid chromosome count of $2n$ = 40. In 1903 with the burgeoning study of chromosomes in hybrids, Rosenberg noticed that during meiosis in hybrids between *D. rotundifolia* ($2n$ = 20) and *D. anglica* ($2n$ = 40), there were 10 bivalent (II) and 10 univalent (I) chromosome pairs (from now on 10II, 10I; Rosenberg, 1903, 1904, 1909). The 10 bivalent chromosomes are likely properly paired homologous chromosomes, while the 10 univalent chromosomes are unpaired. This observation led to the hypothesis that *D. anglica* was an allopolyploid between *D. rotundifolia* and another species (Winge, 1917). This raised the question of the identity of the other parental species and led to subsequent cytological work on hybrids between *D. anglica* and most

of the other North American *Drosera* species (Fig. 1; Gervais & Gauthier, 1999; Kondo & Segawa, 1988).



	D. filiformis	*D. linearis*	*D. anglica*	*D. intermedia*	*D. brevifolia*	*D. capillaris*	*D. rotundifolia*
Leaf Shape							
Distribution	East Coast North America	Boreal North America	Circumboreal, Hawaii	Eastern North America, Europe, South America	Eastern North America, South America	Eastern North America, South America	Circumboreal, Southeast Asia
Chromosome count (Meiotic pairing when hybridizing with *D. anglica*)	2n = 20 (12-13,17 I, 5-7 II, 1-2 III)	2n = 20 (9-13 I, 7-10 II, 1 III)	2n = 40 (20 II)	2n = 20 (10 I, 10 II)	2n = 20	2n = 20 (13-14 I, 7-8 II, 1 III)	2n = 20 (10 I, 10 II)
Reference	Kondo & Segawa 1988	Gervais & Gauthiere 1999	Kondo & Segawa 1988	Kondo & Segawa 1988	Wood, Jr., 1955	Kondo & Segawa 1988	Kondo & Segawa 1988

Figure 1: Distribution and cytological data of seven out of the eight North American *Drosera* species. The eighth species, *D. tracyi*, is morphologically similar to *D. filiformis* with a more restricted distribution than *D. filiformis*. The chromosome pairing of meiotic counts in hybrids between *D. anglica* and five diploid species are listed in number of univalent (I), bivalent (II), and trivalent (III) chromosome pairs.

Cytological studies looking for the other parental species of *D. anglica* especially focused on *D. intermedia* and *D. linearis* because of their similarity in distribution, habitat, and morphology compared to *D. anglica* (Fig. 1). *Drosera anglica* and *D. intermedia* are easily confused, especially in herbarium specimens, as their leaf shapes are very similar, and the plants, especially early in the vegetative stage, have few distinguishing features. *Drosera intermedia* occurs in the Eastern United States, Europe, and South America, with a disjunct population in northern Idaho and a second disjunct population in southern Idaho more than 1000 km from the nearest populations. The Idaho

populations have been protected by the state as a rare plant with potentially unique genetics, but others suspect that they may be misidentified *D. anglica* populations (Lowrie et al., 2017). While *D. rotundifolia* and *D. intermedia* rarely form hybrids with each other in nature (Grima, 2020), *D. anglica* and *D. intermedia* form 10 II and 10 I pairs when hybridized artificially (Kondo & Segawa, 1988). Work comparing isozymes of European *Drosera* found *D. intermedia* and *D. anglica* did not share allozymes, unlike *D. anglica* and *D. rotundifolia* (Seeholzer, 1993).

Alternatively, *D. linearis* has been hypothesized as the other parent of *D. anglica* due to *D. anglica* occurring in fen lines (ridges in patterned fens), fen edges, and wetter regions of bogs that are intermediate between the often calcium-rich fen flark (a depression in a patterned fen) microhabitat of *D. linearis* and the drier, more acidic, sphagnum hummock microhabitat of *D. rotundifolia. Drosera anglica* is also intermediate in leaf morphology *D. linearis* and *D. rotundifolia* (Fig. 1; Wood, Jr., 1955). However, *D. linearis* has a more restricted distribution in boreal North America and its chromosomes do not pair properly with those of *D. anglica* (Fig. 1). When hybridized in the wild, *D. anglica* × *D. linearis* forms 9-13 I, 7-10 II, 1 III pairs during meiosis (Gervais & Gauthier, 1999). Chromosomes in hybrids between *D. rotundifolia* and *D. linearis* do not pair properly either (2-10 I + 3-7 II + 1-3 III), and although fertile neo-allopolyploid hybrids have been found in nature, they are slightly different in appearance compared to *D. anglica* (Wood, Jr., 1955). Additional work on the flower structure of *D. anglica* found that it was not intermediate between *D. rotundifolia* and *D. linearis* (Gervais & Gauthier, 1999).

Despite a long history and large body of cytological work, previous molecular phylogenetic work on *Drosera* has been inadequate to disentangle *D. anglica*'s parentage. Rivadavia et al. found that the *rbc*L sequence of *D. anglica* and *D. rotundifolia* only differed in three base pairs (2003), supporting that *D. rotundifolia* was the maternal parent. However, this and other recent phylogenetic studies of *Drosera* have all relied on two to three loci and lacked phylogenetic signal for resolving the relationships within the recently diversified *D.* sect. *Drosera* (Rivadavia et al., 2003; Veleba et al., 2017). The absence of *D. linearis* in previous phylogenetic or allozyme

studies due to its restricted and remote habitat also left the parentage of *D. anglica* unresolved.

In order to detect both maternal and paternal parents of *D. anglica* in the context of the recently diverged *D.* sect. *Drosera,* increased sampling of nuclear loci is needed. Transcriptomes provide thousands of loci each with relatively long sequences informative for evaluating discordance among taxa with short branch lengths (Yang et al., 2015). It is also an effective genome subsampling approach to obtain adequate sequencing depth to tease apart subgenomes. Additionally, transcriptome datasets can be re-analyzed and combined with genome resequencing and target enrichment datasets or used later to address molecular evolution questions.

With this information in mind, we sequenced transcriptomes from one to four populations in 12 species from across *D. sect. Drosera*. We reconstructed gene trees and species trees, phased subgenomes, and analyzed genetic distance and allele diversity to answer the following questions:

1. What are the maternal and paternal species of *Drosera anglica,* and how does that compare to previous cytological findings?
2. Is the northern Idaho population of '*D. intermedia*', *D. intermedia* or *D. anglica*?
3. Since *Drosera anglica* is circumboreal, did *Drosera anglica* originate on one continent and spread subsequently or is there any evidence for multiple origins of *Drosera anglica* in North America and Eurasia?

METHODS

Sampling and Collection

To maximize the chance that our sampling included the parents of *D. anglica*, we included species from across *D.* sect. *Drosera*, with an emphasis on potential parents of *D. anglica.* From across *D. sect. Drosera*, we selected a diversity of species based on the previously published *rbc*L phylogeny (Rivadavia et al., 2003), the geographic distribution, and the morphology. For *D. anglica,* we sampled across its range of distribution and included multiple populations from its hypothesized parents.

We sampled one European and two North American populations of *D. anglica*, two populations of *D. linearis* and *D. rotundifolia*, and four of the five other North American species from five populations, '*D. intermedia*' from Idaho, and four South American species. In addition to our newly generated transcriptome datasets, we downloaded a publicly available transcriptome from a Russian population of *D. rotundifolia* (NCBI SRA: SRR8948654; Gruzdev et al., 2019). For clarity, we will refer to each sample by the species name, and for species with multiple collections, the location abbreviation in parentheses will follow the species name (Table 1). As an outgroup we used the published *D. spatulata* genome assembly, coding sequences (CDS) from genome annotation, and raw reads from its transcriptome (Palfalvi et al., 2020). We chose *D. spatulata* as the outgroup as it is diploid, is sister to the rest of *D.* sect. *Drosera* (see Chapter 3; Veleba et al., 2017), and has a long-read based genome assembly.

Newly generated transcriptomes consisted of field-collected and cultivated samples. For most North American populations, we collected the whole plants in the field, transferred the plants into 8 mL Nalgene bottles and flash froze them in liquid nitrogen. For South American and European populations, tissue from cultivated plants was collected, immediately placed in a 2 mL lysing tube with Lysing Matrix A (MP Biomedicals), and flash frozen in liquid nitrogen. To avoid contamination while collecting the samples, we wore gloves, and between species cleaned tweezers with Kimwipes and ethanol followed by RNase Zap. If our gloves came in contact with the plant, we also changed our gloves. The samples were ground using the FastPrep-24™ 5G bead beating grinder and lysis system (MP Biomedicals) in dry ice with the CoolPrep™ adapter. RNA was extracted following a modified PureLink protocol (see supplemental methods for lab methods; Yang et al., 2017). Library preparation and sequencing was done by either the University of Minnesota Genomics Center or Novogene Corporation, Inc. (Table S1). At University of Minnesota Genomics Center, the libraries were prepared using Illumina Ribo-Zero Plus rRNA Depletion Kit and were sequenced either with 125 paired-end reads on the Illumina HiSeq 2500 or with 150 single-end reads on the NextSeq 550. We requested paired-end reads but the sequencing facility accidentally did single-end reads, and the remaining samples were discarded. Novogene used the New

England Biological NEBNext Ultra II Directional RNA Library Prep kit and sequenced 150 paired-end reads on the Illumina NovaSeq 6000 platform.

Table 1: Sample information and genome sizes.

Sample name	Source (sample location abbreviation)	Collection number (voucher)	Diploid genome size (Mb)
D. anglica (CZ)	(cult.) Best Carnivorous Plants. Locality: Sumava Mts, Southern Bohemia, Czech Republic	RM298A (Photo)	4715[b]
D. anglica (MN)	Lost River Peatlands SNA, Minnesota (MN)	RM230 (MIN)	4715[b]
D. anglica (WA)	Petit Lake, Priest Lake District, Washington (WA)	RM217 (MIN)	4640
D. brevifolia	Cherry Orchard Natural Area Preserve, Virginia	RM211 (MIN)	1636[a]
D. capillaris (FL)	(cult.) Best Carnivorous Plants. Locality: Florida (FL) Panhandle, USA	RM240 (Photo)	-
D. capillaris (VA)	Cherry Orchard Natural Area Preserve, Virginia (VA)	RM210 (MIN)	-
D. esmeraldae	(cult.) Best Carnivorous Plants. Locality: Cerro Duida,Venezuela	RM241 (Photo)	-
D. felix	(cult.) Best Carnivorous Plants. Locality: Tuku Muruku, Gran Sabana	RM245 (Photo)	-
D. filiformis	Webb's Mill Bog, New Jersey	RM208 (MIN)	4877[b]
'*D. intermedia*' (ID)	Grass Creek, Bonner's Ferry, Idaho (ID)	RM218 (MIN)	5303
D. intermedia (NJ)	Webb's Mill Bog, New Jersey (NJ)	RM207 (MIN)	2516[b]
D. linearis (MN)	Lost River Peatlands SNA, Minnesota (MN)	RM228 (MIN)	-
D. linearis (MT)	Indian Meadows RNA, Montana (MT)	RM219 (MIN)	-
D. roraimae	(cult.) Best Carnivorous Plants. Locality: Summit of Mt. Roraima	RM242 (Photo)	2683[b]
D. rotundifolia (ID)	Hager Lake, Priest Lake District, Idaho (ID)	RM214 (MIN)	1933
D. rotundifolia (NJ)	Webb's Mill Bog, New Jersey (NJ)	RM209 (MIN)	2331[b]
D. rotundifolia (RUS)	Russia: Moscow region, wetland (RUS) SRR8948654, Gruzdev et al., 2019	-	2331[b]
D. solaris	(cult.) Best Carnivorous Plants. Locality: Mt Yakontipu, Pakaraima Mountains, Guyana	RM237 (Photo)	2429[a]
D. spatulata	Palfalvi et al 2022	-	646[c]

[a]genome sizes from Mohn et al., 2022
[b]genome sizes from Veleba et al., 2017
[c]genome sizes from Palfalvi et al., 2020

Genome size estimation

Fresh samples were collected and mailed to the Flow Cytometry Core Lab at the Benaroya Research Institute (Seattle, WA, U.S.A.) for genome size estimation. For each genome size, four flow cytometry measurements were taken against a known size standard. We used the average genome size for each species for subsequent analyses. Source, voucher, and size standards used for generating new flow cytometry data are listed in Table S1.

Read cleaning and trimming

We roughly followed the previously published pipeline https://bitbucket.org/yanglab/phylogenomic_dataset_construction/ (Morales-Briones et al., 2021; Yang & Smith, 2014) to clean reads, assemble transcripts, and carry out phylogenomic analysis. Programs Rcorrector (Song & Florea, 2015), Trimmomatic (Bolger et al., 2014), Bowtie2 (Langmead & Salzberg, 2012), and FastQC (Andrew, 2010) were used to clean, trim, map and filter out organellar reads, and detect and filter over-represented reads.

Distribution of synonymous distance (Ks) estimated using de novo assembled transcriptomes

The cleaned transcriptome reads were *de novo* assembled with Trinity version 2.5.1 (Haas et al., 2013). To test for cross-contamination, the cleaned reads and the Trinity assembly were fed into CroCo version 1.1 (Simion et al., 2018) in two groups: one from paired-end and the other from single-end datasets as CroCo can only take one type of read configuration in each run.

To estimate the timing of polyploidy events with Ks plots, assembled transcripts from Trinity were translated with TransDecoder (https://github.com/TransDecoder/TransDecoder; Haas, BJ, n.d.) without any filtering. Within-species Ks plots were calculated following (Yang et al., 2015, 2018). We removed Ks values < 0.01 before visualizing Ks distributions as heterozygosity and a large number of isoforms often contribute to Ks values less than 0.01 and make visualizing the signal from paralogs difficult.

Visual inspection of reads mapped to initial assemblies indicated that *D. rotundifolia* (ID), *D. anglica* (WA), and '*D. intermedia*' (ID) reads, which were from the same sequencing batch, often had a 'T' on the 3' end, likely as part of the adapter sequence. Therefore, one base pair was removed from the 3' end of these three samples for subsequent analyses unless otherwise stated.

Synthetic in-silico hybrid

To evaluate our ability to tease apart subgenomes in allopolyploids, after cleaning the raw reads, we combined 6,666,667 paired-end reads from *D. linearis* (MT) and 8,000,000 from *D. rotundifolia* (NJ) to make a synthetic *in-silico* hybrid that roughly resembled *D. anglica* in read coverage. This difference in the number of reads made up for the 125 PE reads of *D. rotundifolia* (NJ) and 150 PE reads of *D. linearis* (MT). Since single-end reads may suffer from additional challenges in phasing, we made a second synthetic hybrid with the same reads but without indication of pairing. These hybrids served as positive controls for phasing subgenomes and will be referred to as synthetic hybrids from now on.

Selecting targets for HybPiper

To choose target genes that are single-copy and well supported by transcriptome data, and to reduce computational time, an initial round of phylogenomic analysis was performed with transcriptome assemblies from two *D. linearis*, two *D. rotundifolia* (before the 3' end of *D. rotundifolia* (ID) was trimmed), and *D. intermedia* (NJ) and the CDS file from the *D. spatulata* genome annotation (Palfalvi et al., 2020). This was done following the Yang & Smith pipeline (Morales-Briones et al., 2021; Yang & Smith, 2014).

We used TransRate version 1.0.3 (Smith-Unna et al., 2016) to quantify the quality of the Trinity assemblies and removed transcripts with nucleotides of mapped reads poorly matching the assembled transcript ($s(C_{nuc}) \leq 0.25$), low read coverage ($s(C_{cov}) \leq 0.25$); and paired-end reads misaligned ($s(C_{ord}) \leq 0.5$). Additionally, chimeric transcripts with multiple open reading frames stitched together in opposite directions, each with at least 30% similarity in at least 100 bp compared to *Beta vulgaris* were removed (Yang &

Smith, 2013). The resulting transcripts were translated with TransDecoder with *Arabidopsis thaliana* and *Beta vulgaris* reference proteomes. Finally, the CDS were further reduced with CD-HIT (W. Li & Godzik, 2006) to remove sequences with > 99% similarity using a 10 base pair word length.

To cluster the resulting CDS, hits from an all-by-all BLASTn (Altschul et al., 1990; Camacho et al., 2009) search with a hit fraction cut-off of 0.3 were input into mcl (Van Dongen, 2008) with an inflation value of 1.4. The resulting clusters were each aligned with MAFFT version 7.475 (Katoh & Standley, 2013) using the generalized affine gap cost for pairwise alignments with 1000 iterations, alignments trimmed with Phyx (Brown et al., 2017) removing columns with >90% missing data, and gene trees estimated using RAxML (version 8.2.11). Using TreeShrink (Mai & Mirarab, 2018), we trimmed spurious tips that were in the 0.4 quantile, then we removed monophyletic tips of the same taxa, leaving only one with the highest number of aligned characters in the trimmed alignment. We then visually inspected the resulting gene trees and cut long internal branches that were more than 0.1 substitutions per site, as internal branches among our sampled species were mostly < 0.06 in length and branches longer than 0.1 were due to either deeper paralogs or spurious sequences. We retained trees with all six taxa, re-aligned the sequences with MAFFT. Terminal branches 10× longer than their sister clade or more than 1.0 substitutions per site were trimmed. We masked monophyletic and paraphyletic tips from the same sample. Finally, we selected one-to-one orthologs present in all six samples and used the *D. spatulata* coding sequences for these genes in downstream analyses.

Targeted assembly with HybPiper

We chose targeted assembly for phylogenetic analyses given the tools available for phasing subgenomes. As part of the HybPiper 2.0 (Johnson et al., 2016) pipeline with default settings, we used BWA (H. Li, 2013) to map reads to the 6443 *D. spatulata* target genes selected as above, and SPAdes (Bankevich et al., 2012) to assemble the mapped reads into transcripts where read depth ≥ 8. Assembled transcripts for each gene were compared to each other and the reference using Exonerate version 2.2.0 (Slater et al. (http://www.ebi.ac.uk/~guy/exonerate/; Slater & Birney, 2005). If multiple transcripts

were assembled by SPAdes that each covered >75% of the length of the target, HybPiper throws a long paralog warning. For genes without an assembled ("long") paralog but with a second contig that covered <75% of the target length and with a read depth of at least 1 (1−7 where unassembled) for 75% of the gene, HybPiper throws a depth paralog warning. Paralogs that are too similar may not be detected and may be assembled into chimeras.

Reference-based phasing with HybPhaser

We used the HybPhaser pipeline to identify polyploids and phase the subgenomes using references from putative parents (Nauheimer et al., 2021). This consists of four steps: visualizing the distribution of single nucleotide polymorphisms in each sample to detect potential hybrids and polyploids, determining whether samples belong to multiple clades by mapping to clade references, phasing reads to the appropriate clade references, and then re-assembling the phased reads in HybPiper. First, loci with less than 20% of samples or <10% of the target sequence length covered were removed. Using only the reads mapped to the target by BWA in HybPiper, HybPhaser version 2.1 (Nauheimer et al., 2021) generated a consensus sequence for each gene and each species. For an ambiguity to be called at a site, there must be a read depth of at least 10 at the site and the allele must be supported by at least 4 reads and 15% of the reads. Allele divergence, the percentage of SNPs per gene length, and loci heterozygosity, the percentage of loci with SNPs, were calculated by HybPhaser. In single-copy genes, allele divergence is equal to the nucleotide diversity per site (π). In genes with multiple copies, especially in polyploid species, paralogs also contribute to allele divergence. We identified samples with high heterozygosity and sequence divergence as evidence of polyploidy and/or hybrids with potential need for subgenome phasing. To select which transcriptome dataset to use as clade references, we gathered the genes from the diploid individuals of the initial HybPiper run, aligned with MAFFT v7.475 (Katoh & Standley, 2013), trimmed alignment with Phyx (Brown et al., 2017) removing columns with >90% missing data, estimated gene trees in RAxML version 8.2.11 (Stamatakis, 2014), and estimated the species tree in ASTRAL version 5.7.8 (Zhang et al., 2017). We selected references that had low heterozygosity, low sequence divergence, and represent different clades within

D. sect. *Drosera*. Every sample was then mapped to the clade references. Samples that had both high heterozygosity and sequence divergence and mapped to multiple clade references at approximately equal rates were then phased. To phase subgenomes, HybPhaser mapped reads from a sample to both references using BBMap version 38.96 (Bushnell, 2014). If the reads mapped unambiguously to one of the references, the reads are sorted to that reference. If they map to both equally, they go to both. If they did not map to either, they were removed. This produced two files with phased reads that were assembled to the original 6443 *D. spatulata* genes using HybPiper.

The resulting phased homeologs from HybPhaser and unphased genes from the remaining samples were again aligned using PRANK v.170427 (Löytynoja, 2014), and alignments trimmed using Phyx (Brown et al., 2017) requiring a minimum of 5% column occupancy. A subset of resulting alignments were visually inspected to ensure proper assembly and phasing. RAxML was used to estimate gene trees with 100 bootstrap replicates. ASTRAL was then used to estimate the species tree from the gene trees. Gene tree discordance was then calculated by PhyParts (Smith et al., 2015) requiring a minimum local bootstrap of 50. Genes with at least 50 base pairs were concatenated for phylogenetic reconstruction using RAxML.

Genetic Distance

In addition to tree-based methods, we also calculated the pairwise genetic distance between samples. A distance-based method is informative especially when relationships among samples are not strictly tree-like and when samples are very closely related. We filtered the PRANK alignments by 98.0% or greater pairwise sequence identity, which removed alignments with large segments of ambiguous characters or little overlap between alignments. After visually inspecting the remaining alignments, we removed aligned columns with any gap or ambiguous characters and kept only alignments longer than 1000 bps to ensure enough signal. One additional gene was removed because most of the variation was due to one sample suggesting that it may be the result of a chimeric assembly. We then used bio3d version 2.4-4 (Grant et al., 2006) in R version 4.2.3 to calculate the pairwise genetic distance of each sample or subgenome per alignment. We

then calculated the mean and median genetic distance between each pair of samples or subgenomes.

Haplotype-based phasing and SNAPP coalescent estimation of population divergence

As the reference-based phasing by HybPhaser may bias the assembly toward the reference genotypes, we also used a haplotype-based phasing approach to evaluate divergence among subgenomes and parents. The *de novo* assembled *Drosera rotundifolia* (NJ) transcripts filtered by Transrate and with chimeras removed were de-duplicated with Corset version 1.07 (Davidson & Oshlack, 2014). This transcriptome was selected as the reference to call SNPs because the RNA had the highest RIN, the transcriptome had the lowest redundancy (number of transcripts per gene) indicating a contiguous assembly, and the peptides matched to the highest number of sequences in the *Beta vulgaris* reference proteome. Therefore, the transcriptome of *D. rotundifolia* (NJ) was indexed with Bowtie2.

Cleaned reads from all the samples went through a second round of more stringent trimming by Trimmomatic to ensure that error would not bias the SNP calling. We followed the parameters of (Conover & Wendel, 2022): "LEADING:28 TRAILING:28 SLIDINGWINDOWS:8:28 SLIDINGWINDOW:1:10 MINLEN:65". Subsequently, reads from all samples were aligned to *D. rotundifolia* (NJ) using Bowtie 2 version 2.3.5.1 (Langmead & Salzberg, 2012) with --end-to-end settings. The pair-end samples used the --no-mixed setting. The SAM files were then sorted and indexed with Samtools version 1.15 (H. Li et al., 2009).

We took four steps to obtain phased haplotypes: 1) HAPLOSWEEP called phased haplotypes in *D. anglica,* '*D. intermedia*', and the synthetic hybrid samples, 2) The haplotypes of *D. rotundifolia* (NJ) and *D. linearis* (MT) were called at the same locations, 3) The phased haplotypes were sorted into the subgenome by whether the synthetic hybrid matched *D. rotundifolia* (NJ) or *D. linearis* (MT), and 4) The haplotype of *D. rotundifolia* (ID), *D. linearis* (MN), *D. brevifolia,* and *D. capillaris* populations were called at the same location.

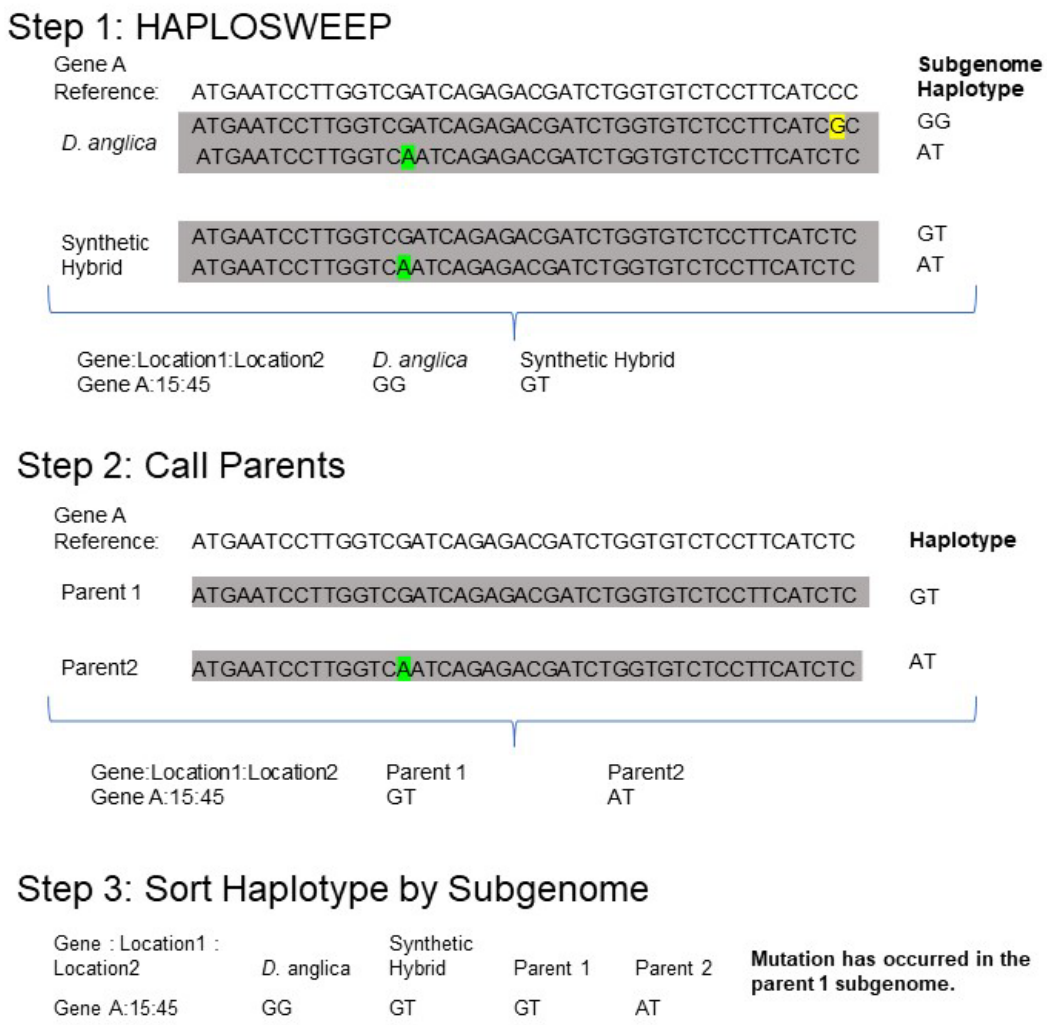


Figure 2: Haplotype calling, phasing, and sorting. Step 1: Single nucleotide polymorphisms (SNPs) were called with BCFtools and phased with HAPLOSWEEP in the *D. anglica,* synthetic hybrid*,* and '*D. intermedia*' (ID; not shown here) samples. Step 2: We then called the matching haplotypes in the parents and Step 3: sorted the haplotype based which the parent matched the synthetic hybrid. The two SNPs, one distinguishing the subgenome (green) and a second SNP distinguishing among *D. anglica* samples, must be within 125 basepairs so that the second SNP can be phased to the correct subgenome.

For step 1, we used BCFtools version 1.16 mpileup to call variants for the *D. anglica*, '*D. intermedia*' (ID), and synthetic hybrid samples with up to 8000 reads and bcftools to convert the BCF file to VCF file format. The VCF file of SNPs and indexed BAM files were fed into HAPLOSWEEP (Clevenger et al., 2018) to call haplotypes within the length of a read. HAPLOSWEEP is a pipeline specifically designed to phase subgenomes in polyploid transcriptomes and detect SNPs in each subgenome that

distinguish between samples. Each haplotype is made up of one SNP that distinguishes the two subgenomes in each sample, and one SNP that varies within the subgenome across samples. In HAPLOSWEEP, we used the settings --genotype_parents --polyploid with parameters for single-end read length set to 125 base pairs (Fig. 2). Since HAPLOSWEEP was designed for identifying SNPs between parent genotypes of polyploid species (--genotype_parents) and then genotyping hybrid populations by calling haplotypes at those sites (--call_population), the --genotype_parents setting allowed us to identify SNPs differentiating between polyploid individuals instead of subgenomes. This resulted in phased haplotypes without information on which subgenome they belonged to. To sort the haplotypes to subgenomes in step 2, we wrote a script to call the haplotypes of *D. rotundifolia* (NJ) and *D. linearis* (MT) at the same locations using the bam files for these samples (Fig. 2). Step 3: Where the haplotype of the synthetic hybrid (PE) matched either *D. rotundifolia* (NJ) or *D. linearis* (MT), the subgenome was sorted to the respective parent (Fig. 2). Haplotypes that could not be matched to a parent because of missing data for at least one parent or the synthetic hybrid or because of no variation between the two parents were ignored in subsequent analyses. Step 4: We called haplotypes at these locations for *D. rotundifolia* (ID) and *D. linearis* (MN), as well as *D. brevifolia* and *D. capillaris* populations as outgroups.

Phased and sorted haplotypes were used to infer the population history among subgenomes and putative parents in RAxML and SNAPP. Haplotypes were concatenated and the maximum likelihood phylogeny was estimated using RAxML with 100 bootstrap replicates and *D. brevifolia* as outgroup. Only one haplotype per locus was called and so they were treated as haploid alleles and converted to 0 for the primary allele and 2 for the alternate allele using PGDSpider 2.1.1.5 (Lischer & Excoffier, 2012). We used the default settings of SNAPP version 1.6.1 (Bryant et al., 2012) implemented in BEAST version 2.7.3 (Bouckaert et al., 2014) with 5,000,000 MCMC steps. To include within group variance for the coalescence estimation, within species, populations were grouped for *D. linearis, D. rotundifolia,* and outgroup *D. capillaris.* Because of the previous tree topologies of *D. anglica* and to include heterozygosity within samples for coalescence estimations, *D. anglica* (MN) and *D. anglica* (CZ) were treated as one population and *D. anglica* (WA) and '*D. intermedia'* (ID) as a separate population. We visualized the

sampling in Tracer 1.7.1 (Rambaut et al., 2018) and produced a summary tree with TreeAnnotator (Drummond & Rambaut, 2007). Because the synthetic hybrids were identical to one of the parental samples and would violate assumptions in SNAPP, they were not included in topology inference.

rbcL and ITS

To compare our results with previously published sequences, we extracted the chloroplast *rbc*L and nuclear ITS sequences from our *D. linearis*, *D. rotundifolia*, *D. anglica,* '*D. intermedia*' transcriptomes. For ITS, we used Bowtie2 to map reads with the second round of Trimmomatic trimming to a *D. rotundifolia* ITS sequence (MT784099.1 from NCBI GenBank). For *rbc*L, we used the extracted organellar reads and Bowtie2 to map the reads to a reference *rbc*L sequence from *D. rotundifolia* (AB29809.1 from NCBI GenBank). In both cases we used the Bowtie2 --end-to-end setting for all samples and --no-mixed setting for paired-end samples. We called variants using BCFtools mpileup with default settings except that the max depth was increased to 30,000 and visualized the mapping and variants in the Integrative Genomics Viewer version 2.12.3 (IGV).

RESULTS

Drosera anglica *and* D. filiformis *have doubled in genome size compared to other diploid North American species*

Diploid genome sizes ranged from 0.6 Gb in *D. spatulata* to 5.9 Gb in *D. tracyi.* North and South American *Drosera* with a chromosome count of 2*n* = 20 had diploid genome sizes mostly between 1.6 Gb and 2.7 Gb with the exception of *D. filiformis* and *D. tracyi*, which were 4.9 and 5.9 Gb, respectively. Genome size of *D. anglica* ranged from 4.6 to 4.7 Gb, over twice that of *D. rotundifolia*. Genome size did vary between studies with our newly generated *D. anglica* 'WA' being 75 Mb less than previous estimations of *D. anglica*, and *D. rotundifolia* 'ID' was 400 Mb less than previous studies. The *'D. intermedia'* (ID) population had a genome size more than twice that of *D. intermedia* and similar to, though approximately 660 Mb higher than, the *D. anglica* 'WA' population. There are no genome size estimates for *D. linearis.*

Sampling and sequencing

Of the 28 *D.* sect. *Drosera* species (Fleischmann et al., 2018), our dataset included transcriptomes from 12 species and 19 individuals representing different populations. Of these, 17 were newly generated transcriptome datasets. The RIN numbers ranged from 2.0 to 7.3. Sequencing ranged from 22 million to 26 million single-end reads and 20 to 44 million paired-end reads (Table S1). Initial, unphased targeted assembly with HybPiper recovered 4610–4880 genes from samples with single-end reads, compared to >6099 genes from samples with paired-end reads. After reference-based phasing with HybPhaser, the second round of HybPiper assembly recovered >5950 genes for subgenomes from both single-end and paired-end samples, suggesting that the initial lower numbers of genes recovered from single-end datasets were not due to read coverage. *De novo* assembly with Trinity, on the other hand, produced proteome set with the highest redundancy (8.2–11.9) in *Drosera roraimae* and *D. esmeraldae*, followed by *D. anglica* (CZ; 7), with remaining samples (except *D. anglica* (MN) as it was not calculated) < 5. This suggests that *D. roraimae* and *D. esmeraldae* both had fragmented assemblies from Trinity, whereas their HybPiper assemblies did not have such issues.

CroCo found no evidence of cross-contamination among paired-end samples. For the single-end samples, CroCo flagged 2-13% of assembled transcripts from three samples with potential cross contamination. These three samples, *D. rotundifolia* (ID), *D. anglica* (WA), and *'D. intermedia'* (ID), are too genetically similar to distinguish cross-contamination from genetic similarity. Therefore, we proceeded in the analysis with all the samples.

No Ks peak between 0 to 0.5 was observed in any samples, except that of *D. anglica* which had a slightly broader distribution close to zero (Fig. S1).

Few **Drosera anglica** *paralogs detected by HybPiper despite high nucleotide diversity statistics*

Gene clustering and tree-based ortholog inference using *D. spatulata* CDS from genome annotation and *de novo* assembled transcripts from *D. intermedia, D. rotundifolia,* and *D. linearis* resulted in 6443 one-to-one orthologs to be used as targets for HybPiper assembly. Of the 6443 targets, 4610 to 6425 genes were assembled in each

sample. Of these, 37–476 genes had paralog warnings for multiple long assembled transcripts. In addition, 76–1239 genes had paralog warnings for paralogs with too little depth to assemble (unassembled from here). Paired-end *D. anglica* samples had the highest number of assembled (456–476 versus < 303) and unassembled (1081–1239 versus < 550) paralogs. However, in the *D. anglica* sample with single-end reads only 86 assembled and 171 unassembled paralogs were detected. Both synthetic hybrids had similar numbers of paralogs as diploid species (291–297 assembled; 546–582 unassembled), indicating that homeologs were not properly separated in allopolyploid species.

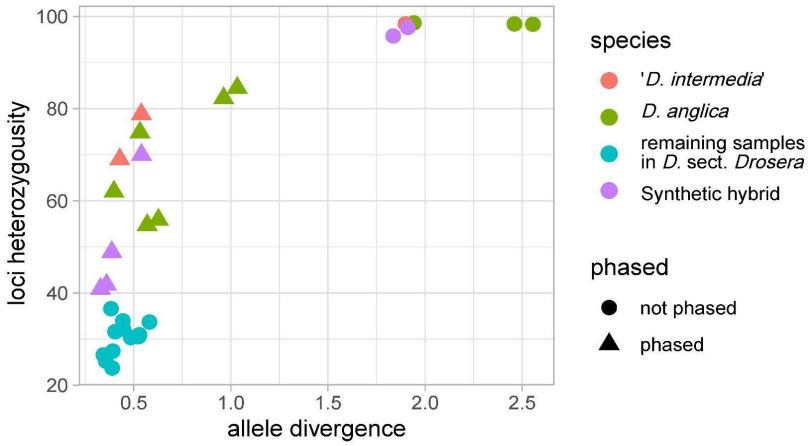


Figure 2: *Drosera anglica* and the *'D. intermedia'* (ID) samples had increased allele divergence and loci heterozygosity. After phasing, the allele divergence decreased to similar to diploid species, but the loci heterozygosity remained higher than the diploid species.

All *Drosera anglica* samples, '*D. intermedia*' (ID), and the synthetic hybrids had similar high loci heterozygosity and allele divergence compared to diploid samples of *D.* sect. *Drosera*, further suggesting that homeologs were not assembled correctly. Loci heterozygosity, the percentage of genes with SNPs, ranged from 98% to 99% in *Drosera anglica* and '*D. intermedia*' (ID) while in all other samples ranged from 24% to 48% (Fig. 3). Similarly, allele divergence, the percentage of SNPs per gene length, was 1.9–2.6 for *D. anglica* and '*D. intermedia*' (ID) and < 0.6 for all other samples (Fig. 3). The synthetic hybrids had both loci heterozygosity and allele divergence very similar to, albeit slightly lower than *D. anglica*.

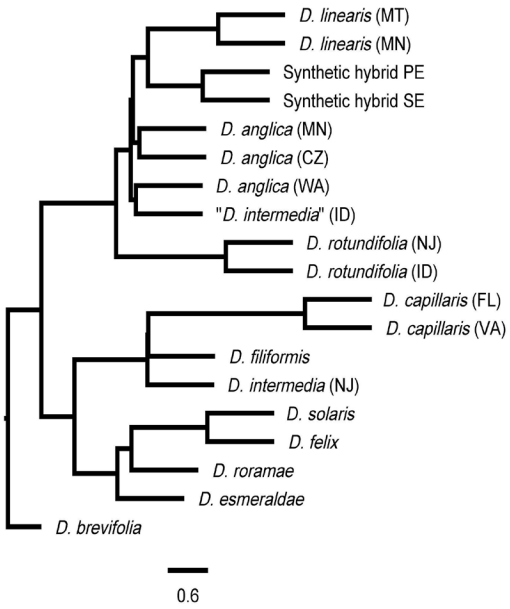


Figure 3: ASTRAL tree estimated from HybPiper assembly. *Drosera brevifolia* was used to root the tree as *D. spatulata* was not included and because of its placement in previous *Drosera* phylogenies. *Drosera brevifolia, D. linearis, D. rotundifolia, D. capillaris,* and *D. esmeraldae* were chosen as clade references. The scale bar indicates internal branch lengths in coalescent unit. Terminal branch lengths were artificially chosen.

Given the high loci heterozygosity and allele divergence in *D. anglica* and *'D. intermedia'* (ID) samples, the relatively few paralogs detected, especially in the synthetic hybrids and single-end read *D. anglica* and *'D. intermedia'* (ID) samples, suggested that paralogs were too similar to distinguish *de novo* and may be forming chimeras. Subsequently, we used a reference-based phasing approach in HybPhaser and a haplotype phasing approach in HAPLOSWEEP.

Reference-based phasing showed Drosera anglica *and* 'D. intermedia' *(ID) subgenomes being most similar to* D. rotundifolia *and* D. linearis

D. rotundifolia (NJ)*, D. linearis* (MN)*, D. esmeraldae, D. brevifolia,* and *D. capillaris* (FL) were chosen as clade references in HybPhaser based on initial species tree inference (Fig. 4). Most samples mapped strongly to one reference or weakly to multiple references except *D. anglica* and '*D. intermedia*' (ID) samples, which mapped strongly to both *D. rotundifolia* (NJ) and *D. linearis* (MN). After phasing of *D. anglica* and '*D. intermedia*' (ID) reads against *D. rotundifolia* (NJ) and *D. linearis* (MN), a total of 3569 genes were assembled in all subgenomes and diploid samples with HybPiper. Allele

divergence in the phased *D. anglica* and '*D. intermedia*' (ID) samples decreased to ranges similar to diploid species except for *D. anglica* (MN) mapped to *D. linearis* (MN) and *D. anglica* (CZ) mapped to *D. linearis* (MN). Loci heterozygosity remained high for most *D. anglica* and '*D. intermedia'* (ID) samples after phasing. This is likely due to mutations that occurred since the divergence of *D. linearis* and *D. rotundifolia* from *D. anglica* and "*D. intermedia"* (ID) samples.

Overall, phylogenomic analysis with subgenomes and diploid species showed discordance

From the HybPiper-HybPhaser pipeline, 3569 genes were retrieved that had been assembled in every sample or phased sample. Phylogenetic analyses using ASTRAL and RAxML recovered very short internal branch lengths (< 0.0005) and discordance between the RAxML and ASTRAL trees among *D. brevifolia, D. linearis, D. rotundifolia,* and the longitudinally spread North American (except *D. brevifolia)* and South American species (Fig. 5). All *Drosera* sect. *Drosera* species that occurred exclusively in South America (*D. felix, D. solaris, D. esmeraldae,* and *D. roraimae*) were monophyletic with strong support (supported by 1470/2409 genes each with >50 bootstrap; informative from now on; Fig. 5). Sister to this clade was a clade of eastern North American species, many reaching South America (*D. filiformis, D. intermedia,* and *D. capillaris;* 2054/2780 informative genes; Fig. 5). While some closely related species show similar distributions, neither the boreal nor longitudinally distributed taxa were monophyletic.

Branch lengths in* Drosera anglica *subgenomes were short

In both the RAxML and ASTRAL results, the *rotundifolia* subgenomes of *D. anglica* were monophyletic (1069/1780 informative trees) and sister to *D. rotundifolia* + phased subgenomes of the synthetic hybrids. On the other hand, the *D. linearis* subgenomes of *D. anglica* were paraphyletic (ASTRAL) or monophyletic with *D. linearis* being paraphyletic (RAxML). Between the two subgenomes, the subgenome with affinity to *D. linearis* had shorter branch lengths, less informative genes, and had more gene tree discordance than the subgenome with affinity to *D. rotundifolia* (Fig. 5). In fact,

the first and second topologies for each node in the subgenome with affinity to *D. linearis* had a similar number of genes supporting each (Fig. 5B).

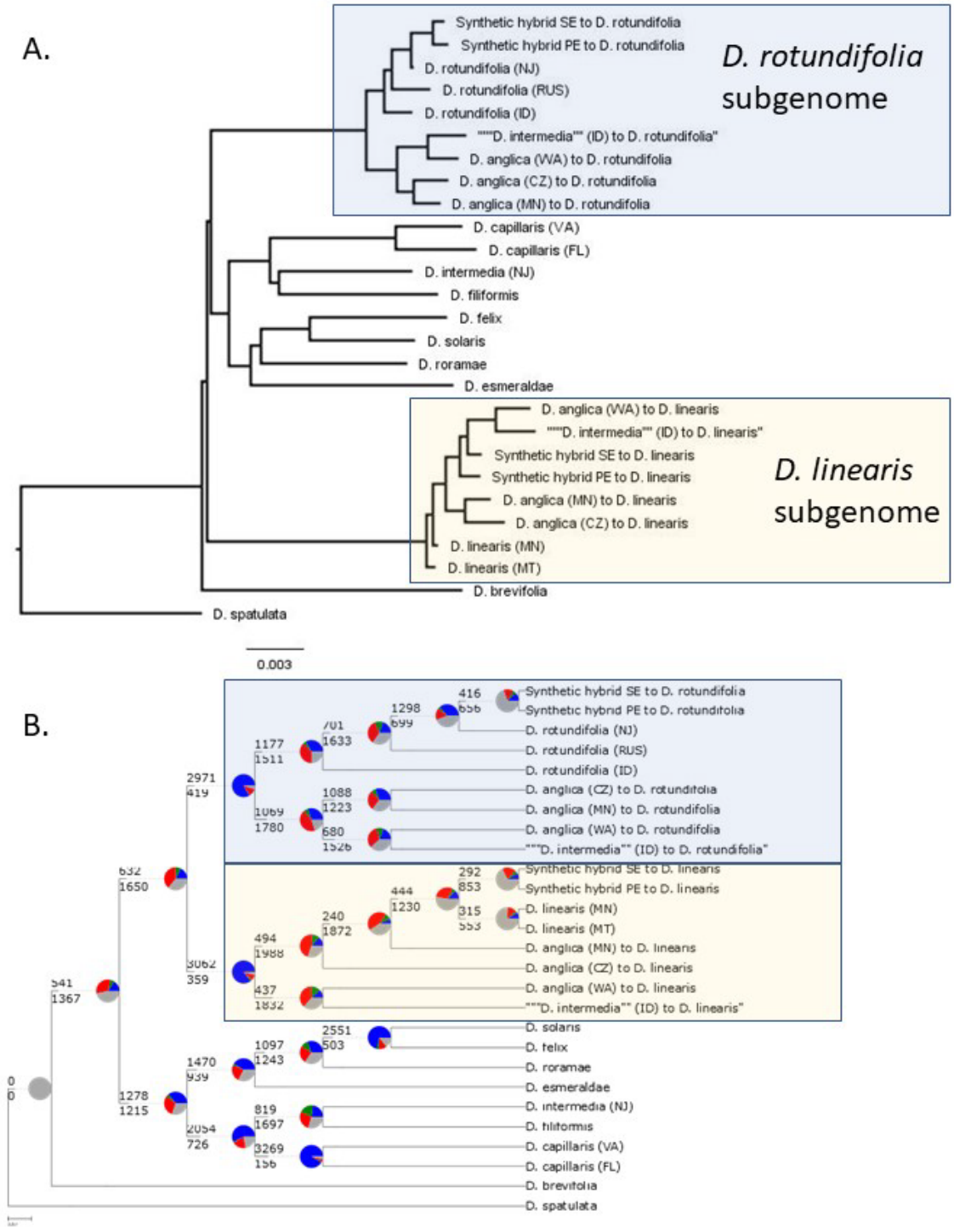


Figure 5. *Drosera* sect. *Drosera* tree topology estimated from 3569 genes assembled in HybPiper with the subgenomes of *D. anglica* phased in HybPhaser. A. The RAxML tree. The

scale bar indicates branch length. All nodes had a 100 bootstrap. B. the ASTRAL tree with gene tree support from PhyParts mapped on the nodes.

Visual inspection of the alignments after phasing found that a pairwise genetic distance of 98.0% seemed to eliminate issues with chimeric sequences. After removing gaps in the alignment and alignments with <1000 bps, 334 genes remained. Overall, the median distance between samples ranged from 0 to 0.022 with the highest distance being between *D. spatulata* and *D. brevifolia* (Table S2). The median distance between *D. rotundifolia* and *D. linearis* samples was 0.014–0.015 (Table S2). The median genetic distance between samples of *D. anglica* subgenome *linearis* or between *D. anglica* subgenome *linearis* and *D. linearis* was 0.000 with the exception of *D. anglica* (WA) to *D. anglica* (MN), which was 0.001(Table S2).

On the other hand, there was more variation in the *D. anglica* subgenome *rotundifolia*. The median genetic distance between *D. anglica* subgenome *rotundifolia* and *D. rotundifolia* was 0.002 (Table 2). The genetic distance between the *rotundifolia* subgenomes of *D. anglica* samples ranged from 0.000 to 0.001 with *D. anglica* (MN) and *D. anglica* (CZ) being genetically similar and *'D. intermedia'* (ID) and *D. anglica* (WA) genetically similar (Table 2).

Table 2: The median pairwise genetic distance of the *D. rotundifolia* and *D. anglica* subgenome *rotundifolia* samples from 334 phased genes assembled in HybPiper.

	D. rotundifolia (NJ)	Synthetic hybrid PE	*D. rotundifolia* (ID)	*D. rotundifolia* (RUS)	*D. anglica* (WA)	*'D. intermedia'* (ID)	*D. anglica* (MN)	*D. anglica* (CZ)
D. rotundifolia (NJ)	0	0	0.001	0	0.002	0.002	0.002	0.002
Synthetic hybrid PE	0	0	0.001	0	0.002	0.002	0.002	0.002
D. rotundifolia (ID)	0.001	0.001	0	0.001	0.002	0.002	0.002	0.002
D. rotundifolia (RUS)	0	0	0.001	0	0.002	0.002	0.002	0.002
D. anglica (WA)	0.002	0.002	0.002	0.002	0	0	0.001	0.001
'D. intermedia' (ID)	0.002	0.002	0.002	0.002	0	0	0.001	0.001
D. anglica (MN)	0.002	0.002	0.002	0.002	0.001	0.001	0	0
D. anglica (CZ)	0.002	0.002	0.002	0.002	0.001	0.001	0	0

Haplotype phasing found little divergence among Drosera anglica *populations*

HAPLOSWEEP returned 1058 phased haplotypes from the *D. anglica* populations, '*D. intermedia*' (ID) and the synthetic hybrids. Of these, 466 haplotypes phased to the *D. rotundifolia* subgenome, and 156 haplotypes phased to the *D. linearis* subgenome.

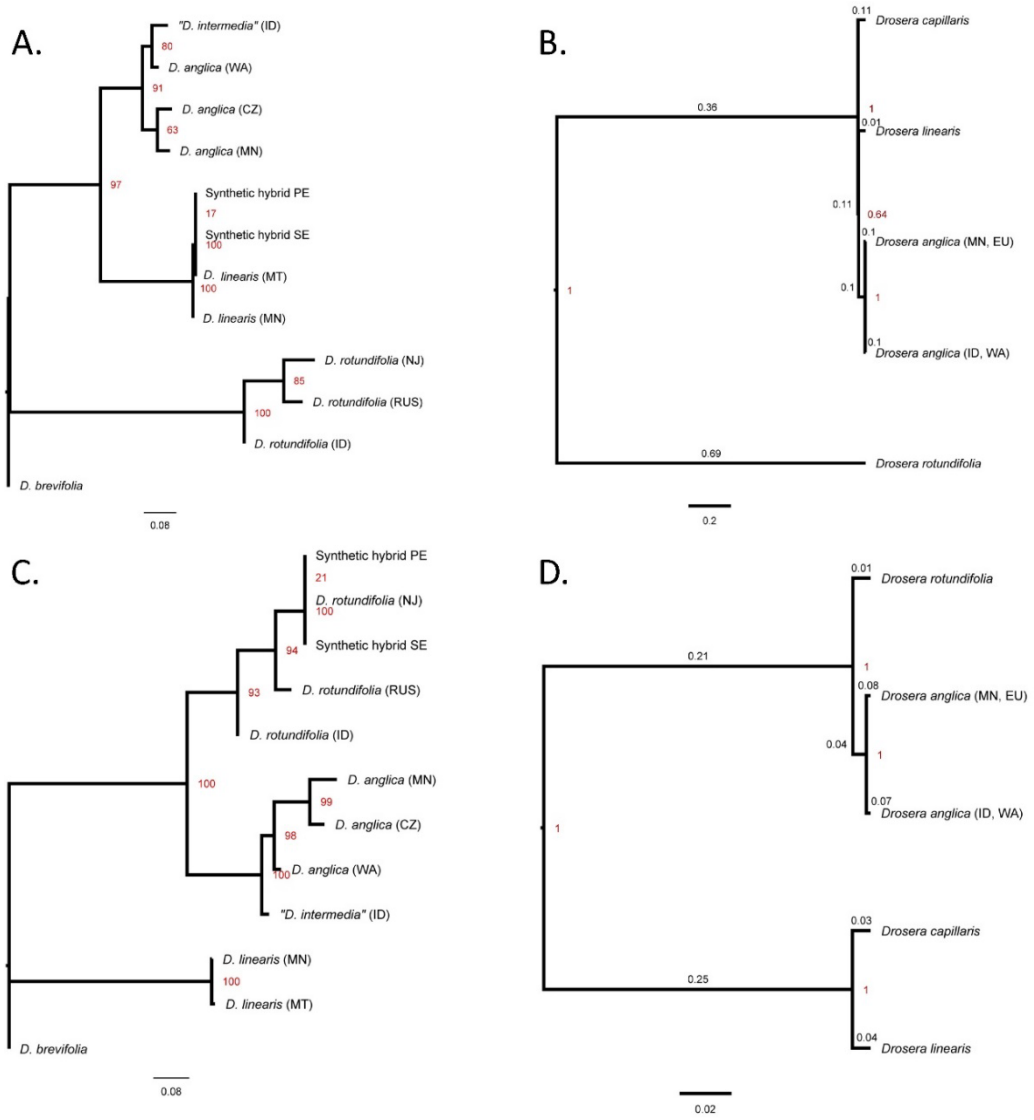


Figure 6: Relationship between *D. anglica* and '*D. intermedia*' (ID) populations and parental species using haplotypes phased by HAPLOSWEEP, including only SNPs that are variable between subgenomes or between the populations of *D. anglica,* '*D. intermedia*' (ID), and the synthetic hybrids. A. & C. RAxML tree for the *linearis* subgenome (A) and the *rotundifolia* subgenome (C). Numbers at nodes are bootstrap percentages. The scale bars indicate branch length. B & D. SNAPP results for the *linearis* subgenome (B) and the *rotundifolia* subgenome (D) with the number above the branch indicating estimated population mutation parameter (θ; amount of variation at loci). Next to each node is the posterior probability for the tree topology. Both

subgenomes in both analyses support the monophyly of *D. anglica*. In the RAxML tree, the Minnesota population was more closely related to the Czech Republic population than to the remaining North American populations.

The RAxML analysis using concatenated haplotypes recovered *D. anglica* and '*D. intermedia*' (ID) as monophyletic but different topologies between the two subgenomes. However, both subgenomes supported *D. anglica* from Minnesota as more closely related to the Czech Republic population (bootstrap support 63 and 99, respectively; Fig. 6) than to other North American populations. In both subgenome trees, *D. brevifolia* had very short branch lengths. As the SNPs in the haplotypes specifically distinguished either *D. anglica* samples, '*D. intermedia*' (ID), and the synthetic hybrid or the parental subgenomes, few SNPs included in the analyses had information on the divergence of *D. brevifolia*.

SNAPP recovered the two populations of *D. anglica* as sister in both the *D. linearis* and *D. rotundifolia* subgenomes (Fig. 6). The population mutation parameter (θ; amount of variation at loci) leading to *D. linearis* or *D. rotundifolia* was small on the corresponding subgenome tree compared to the remainder of the tree. This is likely due to biases introduced in calling haplotypes. *Drosera capillaris*, which served as an outgroup in the SNAPP analysis, was sister to *D. linearis* or *D. linearis* and the *linearis* subgenome of *D. anglica*.

rRNA and rbcL sequences in D. anglica *were nearly identical to* Drosera linearis

The *rbc*L of *Drosera anglica* matched that of *D. linearis.* The maximum read depth for *rbc*L ranged from 32 to 606,594, likely due to different library preparation approaches. Four SNPs distinguished *D. linearis* from *D. rotundifolia*, and all samples of *D. anglica* matched *D. linearis* at all four SNP sites. Our results suggested that the *D. linearis* lineage represented the maternal parent of *D. anglica*.

Despite adequate read depth, the ribosomal RNA and ITS locus was homozygous for SNPs matching *D. linearis*. Read depths reached over 19,000 for all *D. anglica* samples at regions across the ribosomal RNA locus. There were 34 SNPs locations, primarily in the ITS region, where *D. linearis* and *D. anglica* were all present and homozygous for one variant while *D. rotundifolia* was homozygous for the other variant.

Five of these SNPs had read depths greater than 2000 for all *D. anglica* samples. At only one SNP *D. rotundifolia* and *D. anglica* were homozygous for the same allele while *D. linearis* was homozygous for a different allele. Two SNPs that exclusively were found in *D. anglica* (MN) and (CZ) populations, were homozygous in *D. anglica* (MN) and homozygous or heterozygous in *D. anglica* (CZ).

DISCUSSION

By sampling transcriptomes in multiple populations across its range of distribution, we found strong evidence for the origin of *Drosera anglica* from *D. linearis,* representing the maternal lineage, and *D. rotundifolia,* representing the paternal lineage. Additionally, we confirmed that the disjunct '*D. intermedia'* population in Northern Idaho is genetically *D. anglica.* Comparing *D. anglica* populations with parental lineages, we found no evidence for a different origin of the European and the North American populations. Visualization of our assemblies and alignments played an important role in identifying noise and chimeric assemblies and interpreting the data.

Drosera rotundifolia ***and*** **D. linearis** ***are the paternal and maternal parents respectively of*** **D. anglica**

Using transcriptomic data and phasing homeologs by mapping to references in HybPhaser or phasing SNPs in HAPLOSWEEP, we found a high similarity of *D. anglica* to *D. rotundifolia* and *D. linearis.* Additionally, by calling SNPs on *rbc*L we determined that chloroplast of *D. anglica* originated from *D. linearis*, which likely represents the maternal lineage. A previous phylogenetic study proposed *D. rotundifolia* being the maternal parent of *D. anglica* as their *rbc*L sequence only differed by three base pairs (Rivadavia et al., 2003). However, previous taxon sampling did not include *D. linearis* and this similarity is reflective of the few SNPs between the *rbc*L sequences of *D. rotundifolia* and *D. linearis.*

Previous work in *Drosera* has primarily focused on commonly used loci like *rbc*L and ITS. While *rbc*L is expected to be uniparentally inherited, as a nuclear marker ITS is, at least initially, expected to represent both subgenomes in polyploids. Interestingly, only the *D. linearis* copy of the ribosomal subunits and intergenic spacers was expressed in all

our *D. anglica* transcriptomes. A similar pattern of gene conversion to a single ribosomal copy or expression of a single subgenome has been observed in both *Gossypium* (Cronn et al., 1999) and *Brassica napus* (Adams et al., 2003). Analyses of additional genes in the transcriptome data are needed to determine whether the maternal subgenome is dominant transcriptome-wide. Additionally, genomic data is needed to infer whether the ribosomal/ITS copies are only differentially expressed or whether gene conversion has occurred. The inability to detect both parental lineages in rRNA emphasizes the value of sampling a large number of nuclear genes in teasing apart subgenomes.

Despite the genetic similarity of *D. linearis* and *D. anglica,* their homologous chromosomes do not pair properly in hybrids, unlike the homologous chromosomes of *D. rotundifolia* and *D. anglica* (Gervais & Gauthier, 1999; Kondo & Segawa, 1988). The improper pairing of chromosomes suggests chromosome rearrangement events in *D. linearis*, but synteny analysis is needed to ascertain whether this change is epigenetic or genetic. Gervais and Gauthier expressed concern that hybridization might dilute and ultimately replace *D. linearis* (1999). The linearis-specific chromosome rearrangement suggests a potential mechanism that maintains species boundaries between *D. linearis* and the multiple *Drosera* species with which it co-occurs.

The northern Idaho population of Drosera intermedia *is* D. anglica

All evidence, including genome size, loci heterozygosity, allele divergence, and transcriptomic analysis based on both reference-guided and *de novo* assemblies, supports that the Idaho population of '*D. intermedia*' is indeed *D. anglica.* The diploid genome size of '*D. intermedia*' (ID) was about twice that of *D. intermedia* and *D. rotundifolia.* The '*D. intermedia*' (ID) sample had loci heterozygosity and allele divergence similar to the known *D. anglica* populations and much higher than the diploid *D. intermedia*. When mapped to clade representatives, like *D. anglica* and unlike *D. intermedia*, it had a strong affinity to both *D. linearis* and *D. rotundifolia*. In phylogenetic analyses with phased SNPs and haplotypes, '*D. intermedia*' (ID) was nested among the *D. anglica* samples, resulting in the conclusion that it is a misidentified population of *D. anglica.* While the leaf shape of this population is similar to *D. intermedia*, the flowering stalks rise vertically from the rosette and its leaves are mostly rising instead of spreading. This

population of '*D. intermedia'* in northern Idaho and another one in south central Idaho are over 1000 km west of the nearest *D. intermedia* populations, which has made them a conservation priority. With the identification of the northern Idaho population as *D. anglica*, we expect the southern Idaho population has also been misidentified, and at least the resources to protect this northern Idaho population can be reallocated elsewhere.

We observed some genome size variation within *D. anglica.* The genome size of the Idaho population of *D. anglica* (*D. intermedia* (ID) previously) was 660 Mb larger than *D. anglica* (WA). Some variation in genome size may be due to water loss due to varying time between collection and genome size estimation, but the estimated genome size for the *D. anglica* (ID) sample was more than the 10% larger than *D. anglica* (WA), greater than the variation observed between fresh and silica dried samples (Bainard et al., 2011; Wang & Yang, 2016).

Interestingly, the *D. anglica* (ID) population occurs in a different habitat than other nearby *D. anglica* populations. While *D. anglica* populations in the region occur on the lake side edges of floating bogs among *Sphagnum* and Buck bean (*Menyanthes trifoliata*) and on floating logs, this population occurs on a sloped fen. This raises the question whether the leaf shape difference between the Idaho and neighboring populations of *D. anglica* may be due to selective forces of the habitat.

The leaf shape, angle of the petiole, and shape of the peduncle can assist with identification among *D. anglica* and related or often confused species (Fig. 7). *Drosera rotundifolia*'s leaf blade is wider than long, the leaves are generally flat against the ground or slightly raised, and the peduncle rises directly from the middle of the plant. *Drosera anglica* and *D. linearis* leaves are generally raised, although older leaves may spread some. The peduncle originates vertically from the basal rosette in both species. Leaf blades of *Drosera linearis* are linear, as its name suggests, with the two sides of the leaf being parallel and the ends stopping abruptly instead of tapering. *Drosera anglica*'s leaf shape is quite variable ranging from oblong to linear-spatulate (Lowrie et al., 2017; Mellichamp, 2016). In mature *D. intermedia* plants, the leaves are spread out evenly and may be reflexed when the plant has a stem. Like the shorter of *D. anglica*'s leaves, they tend to be spatulate (Lowrie et al., 2017).

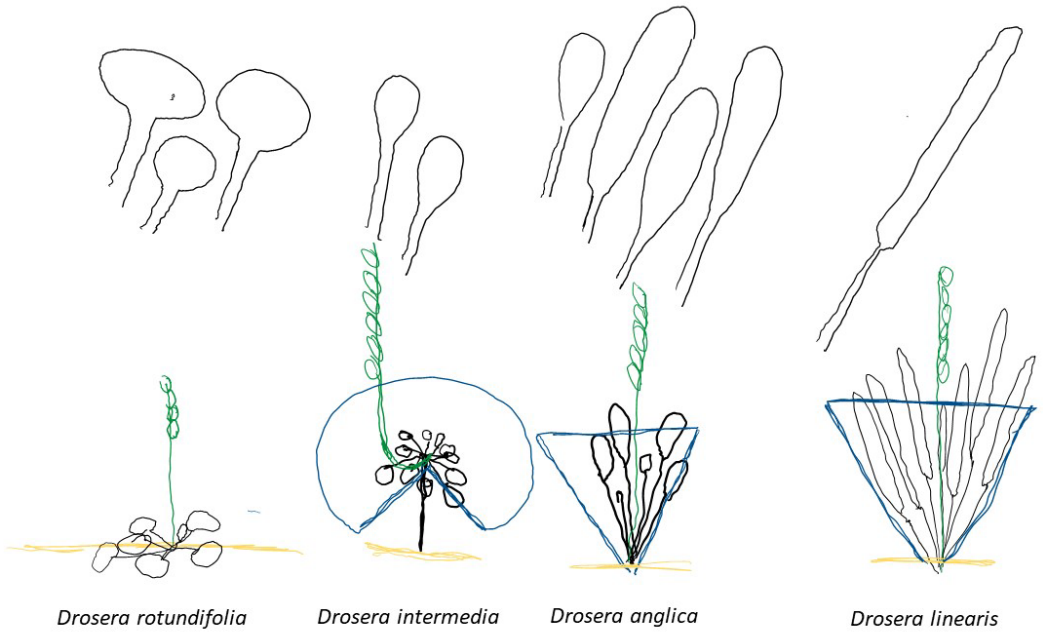


Figure 7: Comparison of leaf shape, angle of the petiole, and shape of the peduncle among *D. rotundifolia*, *D. intermedia, D. anglica,* and *D. linearis.*

No evidence supports multiple origins of Drosera anglica

Given the circumboreal distribution of both *D. anglica* and its paternal parent *D. rotundifolia*, it is possible that *D. anglica* originated multiple times across its range of distribution. However, despite sampling populations of both species from North America and Europe, we did not find evidence supporting a distinct origin of the European *D. anglica*. While *D. anglica* subgenome *linearis* was polyphyletic in the RAxML tree estimated from the phased genes assembled in HybPiper, this is likely due to the low levels of divergence between the subgenomes. The genetic distance between each subgenome and its parental lineage as between 0.001 and 0.002, averaging to only a few mutations at most per gene and therefore a lack of phylogenetic information in gene trees. Regions with higher evolutionary rates may be of assistance for further testing multiple origins. The contrasting genetic diversity between the two subgenomes among *D. anglica* populations suggests multiple origins, different diversity between the parental genomes, or that the *D. rotundifolia* subgenome has experienced relaxed selection.

The widespread distribution of *Drosera rotundifolia* in the Pleistocene has been supported by multiple lines of evidence. Pleistocene fossils of *D. rotundifolia* have been

found in Canada (Penhallow, 1890, 1896). In addition, population genetics of Korean populations of *D. rotundifolia* recovered low within population diversity but high between population divergence suggestive of micro-refugia during the last glacial maximum (Chung et al., 2013). On the other hand, *D. linearis* is more restricted in its geography and habitat, and the flarks where it occurs expand and contract more rapidly based on climate (Kolari et al., 2022). This may explain the lower genetic diversity in *D. linearis* than in *D. rotundifolia. Drosera anglica*, the allopolyploid hybrid, occurs in an intermediate habitat between bogs and fen flarks that is more abundant than the flarks of *D. linearis.*

Visualizing raw data is important in analyzing large datasets

With the small genetic distance among samples, visualizing the assemblies and alignments was necessary to catch unexpected issues. In this manner, we identified sequence processing errors, violated assembly assumptions, and issues in the quality of reference genome annotations.

Visualizing assemblies resulted in the detection of a sequence processing error. After cleaning and assembling genes, we observed that all single-end read samples from the same sequencing batch had an increased number of SNPs on the 3' end. In addition to residual primers, we observed a single 'T' nucleotide on the 3' end of many of the reads. When enough reads with a terminal 'T' ended at the same place, this resulted in a SNP being called erroneously.

HybPiper assembly was designed to assemble genomic reads to coding sequences, but we detected issues when we assembled transcriptomic reads to coding sequences with HybPiper. Because genomic reads will include parts of introns that are missing from coding sequences, when ends of reads do not map to the target, they are trimmed. While splice variants may result in some read ends being trimmed, we observed assemblies that appeared to have even coverage of the gene, but the reads were trimmed, and there were no reads bridging two adjacent regions. For target enrichment, where genomic reads are mapped to the coding sequence of a gene, no reads may bridge two adjacent exons of a gene, but this is problematic when mapping transcriptomic reads to a coding sequence.

Relatedly, when we visualized the reads mapped to the *D. spatulata* reference, we observed that some genes had one or two regions with read depths in the thousands while other regions had much lower read depths. Often there were no reads spanning these two regions suggesting the presence of chimeric genes in the *D. spatulata* genome annotation.

By visualizing our assemblies, we identified issues with errors in three samples, issues with chimeras in the reference transcripts from a genome assembly, and issues with applying a target enrichment pipeline to transcriptomic data. These issues could be easily overlooked without the visualization of assemblies and alignments and could have propagated error in our results by overestimating the divergence of sequences.

Conclusion

Both reference-guided and *de novo* based methods supported *D. rotundifolia* and *D. linearis* as the paternal and maternal lineages of *D. anglica.* We also found *D. anglica* from Minnesota, United States of America and the Czech Republic to be more similar to each other than to other North American populations of *D. anglica*. Future work should further explore subgenomic dominance in *D. anglica,* include *D. anglica* from Hawaii to determine its origin, and evaluate the presence of a chromosomal rearrangement in *D. linearis*.

SUPPLEMENTAL MATERIALS

Figure S1: Ks plots with Ks values 0 to 0.5.

Table S1: Genome sizes estimation and newly sequenced sample information.

Table S2: Pairwise genetic distance between samples.

Supplementary Methods: Modified PureLink RNA extraction protocol.

Supplementary Materials: Photo vouchers.

REFERENCES

Adams, K. L., Cronn, R., Percifield, R., & Wendel, J. F. (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences*, *100*(8), 4649–4654. https://doi.org/10.1073/pnas.0630618100

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Andrew, S. (2010). *FastQC: A Quality Control tool for High Throughput Sequence Data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bainard, J. D., Husband, B. C., Baldwin, S. J., Fazekas, A. J., Gregory, T. R., Newmaster, S. G., & Kron, P. (2011). The effects of rapid desiccation on estimates of plant genome size. *Chromosome Research*, *19*(6), 825–842. https://doi.org/10.1007/s10577-011-9232-5

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., & Wu, C.-H. (2014). BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Comput Biol*, *10*(4), 1003537. https://doi.org/10.1371/journal.pcbi.1003537

Brown, J. W., Walker, J. F., & Smith, S. A. (2017). Phyx: Phylogenetic tools for Unix. *Bioinformatics*, *33*(12), 1886–1888. https://doi.org/10.1093/bioinformatics/btx063

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., & RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, *29*(8), 1917–1932. https://doi.org/10.1093/molbev/mss086

Bushnell, B. (2014). *BBMap: A Fast, Accurate, Splice-Aware Aligner* (LBNL-7065E). Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States). https://www.osti.gov/biblio/1241166

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Chung, M. Y., López-Pujol, J., & Chung, M. G. (2013). Population history of the two carnivorous plants *Drosera peltata* var. *Nipponica* and *Drosera rotundifolia* (Droseraceae) in Korea. *American Journal of Botany*, *100*(11), 2231–2239. https://doi.org/10.3732/ajb.1200486

Clevenger, J. P., Korani, W., Ozias-Akins, P., & Jackson, S. (2018). Haplotype-Based Genotyping in Polyploids. *Frontiers in Plant Science*, *9*. https://www.frontiersin.org/articles/10.3389/fpls.2018.00564

Conover, J. L., & Wendel, J. F. (2022). Deleterious Mutations Accumulate Faster in Allopolyploid Than Diploid Cotton (*Gossypium*) and Unequally between Subgenomes. *Molecular Biology and Evolution*, *39*(2), msac024. https://doi.org/10.1093/molbev/msac024

Cronn, R. C., Small, R. L., & Wendel, J. F. (1999). Duplicated genes evolve independently after polyploid formation in cotton. *Proceedings of the National Academy of Sciences*, *96*(25), 14406–14411. https://doi.org/10.1073/pnas.96.25.14406

Davidson, N. M., & Oshlack, A. (2014). Corset: Enabling differential gene expression analysis for *de novo* assembled transcriptomes. *Genome Biology*, *15*(7), 410. https://doi.org/10.1186/s13059-014-0410-6

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*(1), 1–8. https://doi.org/10.1186/1471-2148-7-214

Fleischmann, A., Cross, A. T., Gibson, R., Gonella, P. M., & Dixon, K. W. (2018). Systematics and evolution of Droseraceae. In A. Ellison & L. Adamec (Eds.), *Carnivorous Plants: Physiology, Ecology, and Evolution* (pp. 45–57). Oxford University Press.

Gervais, C., & Gauthier, R. (1999). Etude cytotaxonomique des espèces et des hybrides naturels du genre *Drosera* (Droseraceae) au Québec. *Acta Botanica Gallica*, *146*(4), 387–401. https://doi.org/10.1080/12538078.1999.10515825

Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A., & Caves, L. S. D. (2006). Bio3d: An R package for the comparative analysis of protein structures. *Bioinformatics*, *22*(21), 2695–2696. https://doi.org/10.1093/bioinformatics/btl461

Grima, P. (2020). The Natural Hybrid between *Drosera intermedia* and *Drosera rotundifolia* in Massachusetts. *Rhodora*, *122*(989), 23. https://doi.org/10.3119/20-08

Gruzdev, E. V., Kadnikov, V. V., Beletsky, A. V., Kochieva, E. Z., Mardanov, A. V., Skryabin, K. G., & Ravin, N. V. (2019). Plastid Genomes of Carnivorous Plants *Drosera rotundifolia* and *Nepenthes ×ventrata* Reveal Evolutionary Patterns Resembling Those Observed in Parasitic Plants. *International Journal of Molecular Sciences*, *20*(17), Article 17. https://doi.org/10.3390/ijms20174107

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Haas, BJ. (n.d.). *TransDecoder* [Perl]. TransDecoder. Retrieved April 24, 2023, from https://github.com/TransDecoder/TransDecoder (Original work published 2015)

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., Zerega, N. J. C., & Wickett, N. J. (2016). HybPiper: Extracting Coding Sequence and Introns for Phylogenetics from High-Throughput Sequencing Reads Using Target Enrichment. *Applications in Plant Sciences*, *4*(7). https://doi.org/10.3732/apps.1600016

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kolari, T. H. M., Sallinen, A., Wolff, F., Kumpula, T., Tolonen, K., & Tahvanainen, T. (2022). Ongoing Fen–Bog Transition in a Boreal Aapa Mire Inferred from Repeated Field Sampling, Aerial Images, and Landsat Data. *Ecosystems*, *25*(5), 1166–1188. https://doi.org/10.1007/s10021-021-00708-7

Kondo, K., & Segawa, M. (1988). A cytotaxonomic study in artificial hybrids between *Drosera anglica* Huds. and its certain closely related species in series *Drosera*, section *Drosera*, subgenus *Drosera*, *Drosera*. *La Kromosomo II*, 1697–1709.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM* (arXiv:1303.3997). arXiv. http://arxiv.org/abs/1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Lischer, H. E. L., & Excoffier, L. (2012). PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, *28*(2), 298–299. https://doi.org/10.1093/bioinformatics/btr642

Lowrie, A., Robinson, A., Nunn, R., Rice, B., Bourke, G., Gibson, R., McPherson, S., Fleischmann, A., & Gonella, P. (2017). *Drosera of the World*.

Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. In D. J. Russell (Ed.), *Multiple Sequence Alignment Methods* (Vol. 1079, pp. 155–170). Humana Press. https://doi.org/10.1007/978-1-62703-646-7_10

Mai, U., & Mirarab, S. (2018). TreeShrink: Fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics*, *19*. https://doi.org/10.1186/s12864-018-4620-2

Mellichamp, T. L. (2016). Drosera. In F. of N. A. E. Committee (Ed.), *Flora of North America North of Mexico* (p. 545).

Mohn, R. A., Zenil-Ferguson, R., Krueger, T. A., Fleischmann, A. S., Cross, A. T., & Yang, Y. (2022). Over two orders of magnitude difference in rate of single chromosome loss among sundew (*Drosera* L., Droseraceae) lineages (p. 2022.10.24.513289). *bioRxiv*. https://doi.org/10.1101/2022.10.24.513289

Morales-Briones, D. F., Kadereit, G., Tefarikis, D. T., Moore, M. J., Smith, S. A., Brockington, S. F., Timoneda, A., Yim, W. C., Cushman, J. C., & Yang, Y. (2021). Disentangling Sources of Gene Tree Discordance in Phylogenomic Data Sets: Testing Ancient Hybridizations in Amaranthaceae s.l. *Systematic Biology*, *70*(2), 219–235. https://doi.org/10.1093/sysbio/syaa066

Nauheimer, L., Weigner, N., Joyce, E., Crayn, D., Clarke, C., & Nargar, K. (2021). HybPhaser: A workflow for the detection and phasing of hybrids in target capture data sets. *Applications in Plant Sciences*, *9*(7). https://doi.org/10.1002/APS3.11441

Palfalvi, G., Hackl, T., Terhoeven, N., Shibata, T. F., Nishiyama, T., Ankenbrand, M., Becker, D., Förster, F., Freund, M., Iosip, A., Kreuzer, I., Saul, F., Kamida, C., Fukushima, K., Shigenobu, S., Tamada, Y., Adamec, L., Hoshi, Y., Ueda, K., … Hedrich, R. (2020). Genomes of the Venus Flytrap and Close Relatives Unveil the Roots of Plant Carnivory. *Current Biology*. https://doi.org/10.1016/j.cub.2020.04.051

Penhallow, D. P. (1890). Notes on the Pleistocene Plants. *Bulletin of the Geological Society*, *I*, 321–334.
Penhallow, D. P. (1896). *Contributions to the Pleistocene Flora of Canada*. 29.
Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology*, *67*(5), 901–904. https://doi.org/10.1093/sysbio/syy032
Rivadavia, F., Kondo, K., Kato, M., & Hasebe, M. (2003). Phylogeny of the sundews, *Drosera* (Droseraceae), based on chloroplast *rbc*L and nuclear 18S ribosomal DNA sequences. *American Journal of Botany*, *90*(1), 123–130.
Rosenberg, O. (1903). Das Verhalten der Chromosomen in einer hybriden Pflanze. *Berichte Der Deutschen Botanischen Gesellschaft*, *21*, 110–119.
Rosenberg, O. (1904). Uber die Tetradenteilung eines *Drosera*-Bastardes. *Berichte Der Deutschen Botanischen Gesellschaft*, *22*, 300–300. https://doi.org/10.1007/bf01547069
Rosenberg, O. (1909). Cytologiche und morphologische Studien an *Drosera longifolia* x *rotundifolia*. *Zeitschrift Für Induktive Abstammungs- Und Vererbungslehre*, *3*(1), 217–219. https://doi.org/10.1007/bf02047738
Seeholzer, C. (1993). Biosystematische Untersuchungen an schweizerischen Drosera-Arten. *Botanica Helvetica*, *103*(1), 39–53.
Simion, P., Belkhir, K., François, C., Veyssier, J., Rink, J. C., Manuel, M., Philippe, H., & Telford, M. J. (2018). A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data. *BMC Biology*, *16*(1), 28. https://doi.org/10.1186/s12915-018-0486-7
Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-31
Smith, S. A., Moore, M. J., Brown, J. W., & Yang, Y. (2015). Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology*, *15*(1), 150. https://doi.org/10.1186/s12862-015-0423-0
Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2016). TransRate: Reference-free quality assessment of *de novo* transcriptome assemblies. *Genome Research*, *26*(8), 1134–1144. https://doi.org/10.1101/gr.196469.115
Song, L., & Florea, L. (2015). Rcorrector: Efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience*, *4*(1), 48. https://doi.org/10.1186/s13742-015-0089-y
Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033
Van Dongen, S. (2008). Graph Clustering Via a Discrete Uncoupling Process. *SIAM Journal on Matrix Analysis and Applications*, *30*(1), 121–141. https://doi.org/10.1137/040608635
Veleba, A., Šmarda, P., Zedek, F., Horová, L., Šmerda, J., & Bureš, P. (2017). Evolution of genome size and genomic GC content in carnivorous holokinetics (Droseraceae). *Annals of Botany*, *119*(3), 409–416. https://doi.org/10.1093/aob/mcw229

Wang, G., & Yang, Y. (2016). The effects of fresh and rapid desiccated tissue on estimates of Ophiopogoneae genome size. *Plant Diversity*, *38*(4), 190–193. https://doi.org/10.1016/j.pld.2016.08.001

Winge, O. (1917). The chromosomes: Their numbers and general importance. In *Comptes Rendus des Travaux du Laboratoire Carlesberg* (Vol. 13).

Wood, Jr., C. E. (1955). Evidence for the Hybrid Origin of *Drosera anglica*. *Rhodora*, *57*, 105–130. https://doi.org/10.2307/23305068

Yang, Y., Moore, M. J., Brockington, S. F., Mikenas, J., Olivieri, J., Walker, J. F., & Smith, S. A. (2018). Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy events. *New Phytologist*, *217*(2), 855–870. https://doi.org/10.1111/nph.14812

Yang, Y., Moore, M. J., Brockington, S. F., Soltis, D. E., Ka-Shu Wong, G., Carpenter, E. J., Zhang, Y., Chen, L., Yan, Z., Xie, Y., Sage, R. F., Covshoff, S., Hibberd, J. M., Nelson, M. N., Smith, S. A., Wong, G. K.-S., Carpenter, E. J., Zhang, Y., Chen, L., … Smith, S. A. (2015). Dissecting Molecular Evolution in the Highly Diverse Plant Clade Caryophyllales Using Transcriptome Sequencing. *Molecular Biology and Evolution*, *32*(8), 2001–2014. https://doi.org/10.1093/molbev/msv081

Yang, Y., Moore, M. J., Brockington, S. F., Timoneda, A., Feng, T., Marx, H. E., Walker, J. F., & Smith, S. A. (2017). An Efficient Field and Laboratory Workflow for Plant Phylotranscriptomic Projects. *Applications in Plant Sciences*, *5*(53). https://doi.org/10.3732/apps.1600128

Yang, Y., & Smith, S. A. (2013). Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics*, *14*(1), 328. https://doi.org/10.1186/1471-2164-14-328

Yang, Y., & Smith, S. A. (2014). Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics. *Molecular Biology and Evolution*, *31*(11), 3081–3092. https://doi.org/10.1093/molbev/msu245

Zhang, C., Sayyari, E., & Mirarab, S. (2017). *ASTRAL-III: Increased Scalability and Impacts of Contracting Low Support Branches* (pp. 53–75). Springer, Cham. https://doi.org/10.1007/978-3-319-67979-2_4

Chapter 3: Polyploidy and discordance in the backbone of Droseraceae (Caryophyllales)

INTRODUCTION

Droseraceae (Caryophyllales) is a globally distributed family of carnivorous plants that was famously studied by Charles Darwin (Darwin, 1875, 1985). The family consists of three genera that are each morphologically well defined: the monotypic terrestrial snap-trap genus, *Dionaea* J. Ellis, the monotypic aquatic snap-trap genus, *Aldrovanda* L., and the species-rich, fly-paper trap genus, *Drosera* L. Droseraceae is not only known for its ability to catch and digest insect preys, but also has a long history of cytological studies since the beginning of the 1900's (Rosenberg, 1903). Previous cytological studies recovered many polyploidy events and single chromosome number changes across the family. However, to date, molecular phylogenetic investigations have been limited to between two to five loci (Rivadavia et al., 2003; Veleba et al., 2017; Fleischmann et al., 2018). These loci showed discordance with some suggesting the polyphyly of *Drosera,* but they did not have enough signal to determine the causes of discordance (Rivadavia et al., 2003).

A previous literature review recovered 510 chromosome counts from 127 out of ~250 species of *Drosera* (Chapter 1, Mohn et al., 2022). By modeling chromosome evolution in a phylogenetic framework, eight whole genome duplication events were inferred in *Drosera.* However, the modeling approach was unable to differentiate between chromosome breakage and chromosomal duplication events. When the chromosome number doubles, we assume the occurrence of a whole genome duplication, but chromosome number doubling due to multiple chromosome fission events has been detected in monkey flowers*,* for example (Fishman et al., 2014). Additionally, chromosome evolution models are also unable to infer whether a polyploidy event occurred with the hybridization of two species (allopolyploid) or within a species (autopolyploid). A phylogenomic approach using a large number of nuclear loci is necessary to both test the phylogenetic location and infer the nature of chromosome doubling events previously identified.

In Droseraceae, the monotypic *Dionaea* occurs only in North and South Carolina while the globally distributed *Aldrovanda* and *Drosera* have one and 250 species respectively. Within *Drosera*, there are four subgenera. *Drosera* subg. *Regiae* and *D.* subg. *Arcturia* consist of one and two species respectively. The two species-rich subgenera in *Drosera*, *D.* subg. *Drosera* and *D.* subg. *Ergaleium*, having been consistently strongly supported as each being monophyletic and sister to each other, together have been referred to as the core *Drosera* (Rivadavia et al., 2003; Fleischmann et al., 2018). The further classification of *D.* subg. *Drosera* and *D.* subg. *Ergaleium* into 7 and 5 sections respectively was based on phylogenetic analysis of 2-5 loci and supported by cytology where branch support was weak (Rivadavia et al., 2003; Fleischmann et al., 2018).

Sequencing transcriptomes provides thousands of genes and sufficient phylogenetic signal to detect discordance and test for reticulation. In addition, transcriptome data can be analyzed to address questions regarding gene family evolution and nucleotide diversity. For example, in Chapter 1, the difference in single chromosome evolution and self-compatibility in *D.* subg. *Ergaleium* versus the other subgenera, raised the question of whether we would see other differences in molecular evolution across the genome.

In Chapter 1 (Mohn et al., 2022), we inferred eight polyploid events across *Drosera*, and in Chapter 2, we focused on *D.* sect. *Drosera* and pinpointed the parental lineages of the allopolyploid circumboreal species *D. anglica*. Of the remaining seven polyploidy events, four were inferred to be at the MRCA of one or more sections, and a fifth polyploidy event occurred in *D.* sect. *Bryastrum* and subsequently diversified to approximately seven species. Given the discordance and polyploidy previously inferred along the backbone of Droseraceae, we sought to resolve the backbone relationship of Droseraceae and test whether transcriptomic data supported the polyploid events inferred by chromosome number reconstructions.

METHODS

Taxon sampling and sample processing

We selected species from across *Droseraceae* to represent each genus, subgenus, and section, and when possible, at least two species representing the diversity of species-rich sections. Given our work in *D.* sect. *Drosera* (see Chapter 2), in this chapter we only included four species from across that section. To ensure reusability of data, we utilized cultivated material from known locations and made herbarium vouchers when possible. For samples without herbarium vouchers, a photo voucher has been included in the supplemental materials. Tissue was collected from cultivated plants and immediately placed in 2 mL lysing tubes with Lysing Matrix A (MP Biomedicals) and flash frozen in liquid nitrogen. To avoid cross contamination while collecting the samples, we wore gloves and between species cleaned tweezers with Kimwipes, ethanol, and RNase Zap. We also changed our gloves if they came in contact with the plant. The samples were ground using the FastPrep-24™ 5G bead beating grinder and lysis system (MP Biomedicals) with lysing tubes in the CoolPrep™ adapter with dry ice. RNA extraction followed a modified PureLink protocol (see supplemental methods for further detail; Yang et al., 2017). As indicated in the Table S1, either Illumina Ribo-Zero Plus rRNA Depletion Kit or New England Biological NEBNext Ultra II Directional RNA Library Prep kit was used for library preparation. Either 125 base pair, paired-end reads were sequenced on the Illumina HiSeq 2500 platform at the University of Minnesota Genomics Center, or 150 base pair, paired-end reads were sequenced on the NovaSeq 6000 at Novogene Corporation, Inc.

In addition to our newly sequenced datasets, we included four transcriptomes from Chapter 2 for *D.* sect. *Drosera* and the published genome assemblies and annotations from *D. spatulata, Dionaea muscipula,* and *Aldrovanda vesiculosa* (Palfalvi et al., 2020). We used previously published CDS files for *Drosera binata* and *Nepenthes alata* (Yang et al., 2018) as well as the additional publicly available transcriptome reads for four Droseraceae species from NCBI SRA (Table 1).

Table 1: Samples used from other works.

Species	Section and subgenus	Reference	Notes
D. filiformis	*Drosera Drosera*	Mohn et al., Chapter 2	-
D. rotundifolia	*Drosera Drosera*	Mohn et al., Chapter 2	-
D. brevifolia	*Drosera Drosera*	Mohn et al., Chapter 2	-
D. linearis	*Drosera Drosera*	Mohn et al., Chapter 2	-
D. felix	*Drosera Drosera*	Mohn et al., Chapter 2	-
D. spatulata	*Drosera Drosera*	Palfalvi et al 2022	CDS from genome and annotations. Raw reads from DRR220142 for HybPiper and HybPhaser and DRR220131 for Ks Plot.
D. binata	*Phycopsis Ergaleium*	(Walker et al., 2017)	CDS from Walker et al., 2017 (combined reads from 3 developmental stages: SRR4450408, SRR4450409, SRR4450411); Raw reads from SRR4450409 for Ks Plot.
Aldrovanda vesiculosa		Palfalvi et al 2020; Walker et al., 2017	CDS from Palfalvi et al. genome and annotations. Raw reads from SRR1979677 for Ks Plot and HybPhaser from Walker et al., 2017.
Dionaea muscipula		Palfalvi et al 2021	Palfalvi et al. genome and annotation for assembled transcripts and SRR20631684 for Ks Plot and HybPhaser.
Nepenthes alata		(Walker et al., 2017)	CDS from Walker et al., 2017 (combined reads from 3 developmental stages: SRR4450413, SRR4450412, SRR4450410)

Phylogenomic analyses

We roughly followed the previously established analysis pipeline https://bitbucket.org/yanglab/phylogenomic_dataset_construction/ (Yang and Smith, 2014; Morales-Briones et al., 2021) for read processing, *de novo* assembly, and phylogenomic analysis. Briefly, Programs Rcorrector version 1.02 or 1.04 (Song and Florea, 2015), Trimmomatic version 0.36 (Bolger et al., 2014), Bowtie2 version 2.3.4.1 or 2.3.5.1 (Langmead and Salzberg, 2012), and FastQC version 0.11.7 (Andrew, 2010) were used, respectively, to clean, trim, map and filter out organellar reads, and detect and filter over-represented reads. The cleaned reads were then *de novo* assembled with Trinity version 2.5.1 (Haas et al., 2013). To test for cross-contamination, the cleaned reads and Trinity assembly were all fed into CroCo version 1.1 (Simion et al., 2018).

Assembled transcripts were cleaned and processed. We used TransRate version 1.0.3 (Smith-Unna et al., 2016) to quantify the quality of the Trinity assemblies. Transcripts with reads poorly matching the assembled transcript (s(Cnuc) ≤ 0.25), contigs with low read coverage (s(Ccov) ≤ 0.25), and contigs with paired-reads misaligned (s(Cord) ≤ 0.5) were removed. Additionally, chimeric transcripts with multiple open reading frames stitched together in opposite directions, each with at least 30% similarity in at least 100 base pairs compared to *Beta vulgaris* were removed (Yang and Smith, 2013). Isoforms were de-duplicated with Corset version 1.07 (Davidson and Oshlack, 2014), and the longest isoform was retained. While Corset resulted in a decrease in the number of *Beta vulgaris* genes detected in each dataset by ~10%, the de-duplication removed redundant isoforms that would otherwise have slow phylogenomic analysis. Coding sequences (CDS) were identified and translated by TransDecoder (Haas, BJ, n.d.; https://github.com/TransDecoder/TransDecoder) with *Arabidopsis thaliana* and *Beta vulgaris* reference proteomes. To evaluate the quality and completeness of each translated sequence dataset, each peptide was queried against *Beta vulgaris* with BLASTp (Altschul et al., 1990; Camacho et al., 2009) returning only one top hit per peptide with an *E* value cutoff set to 10. The hits were filtered by a minimum of 60% identity. We then calculated the total number of *Beta vulgaris* genes and amino acids recovered and the average number of transcripts per *Beta vulgaris* gene. Finally, the CDS were further reduced with CD-HIT (Li and Godzik, 2006) to remove sequences with > 99% similarity using a 10 base pair word length.

To sort CDS into homologous gene clusters, we performed an all-by-all BLASTn (Altschul et al., 1990; Camacho et al., 2009) search and filtered results with a hit fraction cut-off of 0.3. The resulting hits were clustered using MCL (Van Dongen, 2008) with an inflation value of 1.4. For clusters with at least 25 species, FASTA files were written. These were aligned with MAFFT (version 7.475), alignments trimmed with Phyx (Brown et al., 2017) removing columns with >90% missing data, and trees estimated with RAxML (version 8.2.11). Terminal branches more than 10× longer than their sister clade or more than 1.0 substitutions per site were trimmed. Then monophyletic and paraphyletic tips that belonged to the same sample were reduced, retaining the tip with the highest number of characters in the trimmed alignment.

Sequences from the resulting gene trees were realigned with MAFFT, alignment trimmed with Phyx (Brown et al., 2017) removing columns with >90% missing data, and the trees were re-estimated in RAxML with 100 bootstrap replicates. Terminal branches more than 10 times longer than their sister or > 1.0 substitutions per site were trimmed. Then monophyletic and paraphyletic tips from the same sample were reduced with the tip with the most aligned characters retained. These gene trees were further filtered to retain those with only one gene copy per species (one-to-one orthologs) and at least 25 species. A second ortholog dataset was constructed following the "monophyletic outgroup" algorithm (MO; Yang and Smith, 2014). Briefly, unrooted homologous gene trees with all outgroup species being represented by a single-copy of the gene and being monophyletic were rooted by the outgroup. Then the rooted homolog trees were searched from root to tip. When gene duplication was detected, the duplicated copy with a smaller number of taxa was removed. The resulting MO orthologs were filtered to retain those with at least 30 species. The sequences were realigned with PRANK (Löytynoja, 2014), columns with >70% missing data were removed, and RAxML trees with bootstrap values were estimated.

A species tree was estimated from the MO ortholog gene trees in ASTRAL (Zhang et al., 2017). Trimmed alignments of the MO orthologs with more than 50 base pairs were concatenated and a species tree was estimated with RAxML.

Evaluating gene tree discordance

The 1900 MO ortholog gene trees were rooted with *Nepenthes alata* using Phyx (Brown et al., 2017). Gene tree discordance as compared to the ASTRAL MO species tree was evaluated with PhyParts (Smith et al., 2015) with a local bootstrap filter of 80%.

To visualize gene tree discordance, we used the MO ortholog trees with all taxa (412 trees) to estimate dated gene trees using treePL (Smith and O'Meara, 2012). We constrained the MRCA of *Nepenthes alata* and *D. rotundifolia* to 98 to 102 Mya based on approximate divergence in (Yao et al., 2019). We used Densitree (Bouckaert, 2010) to analyze gene tree discordance and to construct the cloudogram.

Detecting genome duplication events

We used two approaches to detect large-scale gene duplication events that are putative whole genome duplication events. To calculate Ks plots, assembled transcripts from Trinity were translated with Transdecoder (Haas, BJ, n.d.; https://github.com/TransDecoder/TransDecoder) without any filtering. Within-species Ks plots and, for taxa pairs of interest, between species Ks plots were calculated following (Yang et al., 2015, 2018).

Rooted clades were extracted from the trees estimated from the original gene clusters. On each rooted clade, when two or more taxa overlapped between two daughter clades, we mapped a gene duplication event to the most recent common ancestor (MRCA) on the RAxML MO species tree (Yang and Smith, 2014). In addition, we also carried out a similar analysis requiring the gene tree and species tree to have a concordant topology at the node. Because the concordant and bootstrap filtering scripts returned similar numbers of gene duplications per node, we report only the results for gene trees with an average bootstrap value of 80.

Subsampling taxa for reticulation analyses

Given the strong support for the monophyly of Droseraceae and *D.* subg. *Drosera* (Yang et al., 2018; Palfalvi et al., 2020) and the polyphyly we observed within these groups, we carried out targeted analyses with a reduced taxon sampling for each to investigate reticulation along the backbone of *Drosera*. To represent each subgenus (for the reduced Droseraceae dataset) or section (for the reduced *D.* subg. *Drosera* dataset), we chose the sample with the lowest redundancy (the average number of translated sequences with top BLASTp match per *Beta vulgaris* gene) and high reference gene coverage compared to *Beta vulgaris* after Corset. This minimized selecting samples that were polyploid or had fragmented or incomplete assemblies. For sections with evidence of non-monophyly, we chose multiple samples to represent each lineage. We also chose a high-quality genome dataset over a transcriptome dataset when possible. The two datasets were as follows:

The Droseraceae subsampling dataset (all genera and subgenera in Droseraceae represented): *D.* subg. *Drosera* (represented by *D. hamiltonii*), *D.* subg. *Ergaleium* (*D.*

porrecta), *D.* subg. *Arcturia* (*D. murfetii*), *D.* subg. *Regiae* (*D. regia*), *Dionaea muscipula* (genome), *Aldrovanda vesiculosa* (genome). In addition, we included the outgroup *Nepenthes alata*.

The *Drosera* subg. *Drosera* subsampling dataset (6 of 8 sections represented): *D.* sect. *Drosera* (*D. filiformis*), *D.* sect. *Brasilianae* (*D. tomentosa*), *D.* sect. *Ptycnostigma* (*D. admirabilis*), *D.* sect. *Psychophila* (*D. stenopetala*), *D.* sect. *Thelocalyx* (*D. hamiltonii*), *D.* sect. *Prolifera* (*D. prolifera*), outgroup (*D. porrecta*). We also included *D. spatulata* (genome) from *D.* sect. *Drosera* due to the non-monophyly of *D.* sect. *Drosera* in the full phylogeny.

For each subsampled dataset, the gene clustering, alignment, tree estimation, and tree filtering steps were the same as for the full data set with the following exceptions: only monophyletic (instead of both mono- and paraphyletic) tips were trimmed and only one round of tree branch and tip trimming was done. Paraphyletic tips were not trimmed as they may be due to polyploidy and not isoforms. Since more than 1000 one-to-one ortholog gene clusters containing all taxa were recovered in both subsampled datasets and one-to-one orthologs are not biased by trimming like MO orthologs, one-to-one orthologs were used for subsequent ASTRAL species tree estimation, concatenated for RAxML tree estimation, used in gene tree discordance calculations with PhyParts, and used for constructing Densitree cloudograms. For the Droseraceae subsampled dataset, the MRCA of *Nepenthes alata* and *D. porrecta* was constrained to 98 to 102 Mya based on approximate divergence in (Yao et al., 2019). Likewise, for the *D.* subg. *Drosera* the MRCA of *D. porrecta* and *D. filiformis* was constrained to 50-55 Mya.

Targeted assembly and heterozygosity using HybPiper

We carried out targeted assembly and analysis of SNPs to pursue additional evidence for ploidy levels and reticulate history in *Drosera*. Currently, the only publicly available reference genome in *Drosera* using long-reads is *D. spatulata* (Palfalvi et al., 2020). To maximize gene and specifically single-copy gene recovery and minimize unnecessary computational time, we selected a subset of 6443 genes from the *D. spatulata* annotation that are one-to-one ortholog gene clusters and recovered in all representative members of *D.* subg. *Drosera* (*D. spatulata, D. rotundifolia, D. linearis,*

and *D. intermedia*; see Chapter 2). This subset of 6443 *D. spatulata* genes that were single copy and well-supported by a cross-species transcriptome data were used as targets for HybPiper2. By using targeted assembly of single-copy genes and a SNP-based approach, we minimized issues due to isoforms and differences in the number of genes resolved. For each species of Droseraceae, HybPiper version 2.0 (Johnson et al., 2016) mapped reads to CDS of the 6443 genes of *D. sptatulata* using BWA mem (Li, 2013) with default settings and then used SPAdes (Bankevich et al., 2012) with a minimum coverage of 8 to assemble the genes. Using the reads mapped to the target by BWA mem, HybPhaser version 2.1 (Nauheimer et al., 2021) generated a consensus sequence for each gene and each species. For an ambiguity to be called at a site in the consensus sequence, we required a depth of at least 10 reads at the site and each allele supported by at least 4 reads and 15% of the reads. Based on the number of ambiguous characters in the consensus sequence of each species, HybPhaser calculated the percent of loci that were heterozygous (loci heterozygosity) and percent average number of SNPs per site (allele divergence). In genes without multiple copies, allele divergence is equal to π per site which is a measure of nucleotide diversity. In polyploid species, homeologs also contribute to the calculation of allele divergence.

RESULTS AND DISCUSSION

Sampling and initial quality control of sequencing data

To resolve the backbone topology in *Drosera,* we included one outgroup species and 31 ingroup Droseraceae species, of which 23 were from newly generated transcriptomes (Tables 1–2). These represented all three genera in the family, and all four subgenera and 12 of the 15 sections of *Drosera*. For every section with more than two species, with the exception of *D.* sect. *Arachnopus*, we sampled at least two species. The extracted RNA had an RNA integrity number (RIN) ranging from 2.0 to 8.3 and resulted in transcriptomes with 20 to 45 million paired-end reads (see Table S1). Neither the RIN nor the total number of reads appear to correlate with the number of reference genes recovered in the assembled and filtered transcriptome of each sample (Fig. S1).

Table 2: Newly sequenced samples in this project

Species	Section Subgenus	Collection number (voucher)	Locality/Source (“cult.” for samples from cultivation)
D. menziesii	*Ergaleium Ergaleium*	RM200 / RM204 (MIN)	(cult.) From Alex Eilts
D. capensis	*Ptycnostigma Drosera*	RM221 (Photo)	(cult.) Koue bokkeveld, South Africa
D. nidiformis	*Ptycnostigma Drosera*	RM223 (Photo)	(cult.) Pietermaritzburg, South Africa
D. regia	*Regiae Regiae*	RM236 (Photo)	(cult.) Best Carnivorous Plants. Upper waterfall site, higher altitude form, Bains Kloof, South Africa
D. latifolia	*Brasilianae Drosera*	RM238 (Photo)	(cult.) Best Carnivorous Plants. Giant, Serra do Cabral, Brazil
D. graomogolensis	*Brasilianae Drosera*	RM239 (MIN)	(cult.) Best Carnivorous Plants. Itacambira, Minas Gerais, Brazil
D. adelae	*Prolifera Drosera*	RM244 (MIN)	(cult.) Best Carnivorous Plants. Bishop Peak, Queensland, Australia
D. caduca	*Lasiocephala Ergaleium*	RM248 (Photo)	(cult.) Best Carnivorous Plants. Wide leaf, white flower; Bachsten Creek, Kimberley, Australia
D. madagascariensis	*Ptycnostigma Drosera*	RM247 (Photo)	(cult.) Best Carnivorous Plants. Botswana
D. tomentosa	*Brasilianae Drosera*	RM252 (Photo)	(cult.) Best Carnivorous Plants. Morro do Jambeiro, Grao Mogol, Minas Gerais, Brazil
D. roseana	*Bryastrum Ergaleium*	RM261 (MIN)	(cult.) From Mark Anderson
D. nitidula	*Bryastrum Ergaleium*	RM262 (MIN)	(cult.) From Mark Anderson
D. hamiltonii	*Stelogyne Drosera*	RM263 (MIN)	(cult.) From Mark Anderson
D. porrecta	*Ergaleium Ergaleium*	RM267 (Photo)	(cult.) From Alex Eilts; “southern form”
D. magnifica	*Brasilianae Drosera*	RM290 (Photo)	(cult.) Best Carnivorous Plants. Pico Padre Angelo, eastern Minas Gerais, Brazil 1500m (seed grown individuals)
D. stenopetala	*Psychophila Drosera*	RM291 (Photo)	(cult.) Best Carnivorous Plants. Tararua Ranges, New Zealand
D. uniflora	*Psychophila Drosera*	RM293 (Photo)	(cult.) Best Carnivorous Plants. Costero, Chile 300-900m
D. prolifera	*Prolifera Drosera*	RM294 (Photo)	(cult.) Best Carnivorous Plants.
D. paradoxa	*Lasiocephala Ergaleium*	RM296 (MIN)	(cult.) Best Carnivorous Plants. White to pink flowers, Mount Bomford, Kimberley, Western Australia, Australia
D. admirabilis	*Ptycnostigma Drosera*	RM299 (Photo)	(cult.) Best Carnivorous Plants. Ceres, South Africa
D. murfetti	*Arcturia Arcturia*	RM300 (Photo)	(cult.) Best Carnivorous Plants. Giant form, the Druids, Tasmania
D. meristocaulis	*Bryastrum Ergaleium*	RM301 (Photo)	(cult.) Best Carnivorous Plants. Northwest plateaus of Cerro Neblina, border Brazil-Venezuela

The newly sequenced samples showed no evidence of cross-contamination. In our cloudogram (Fig. 1), no samples showed short branch lengths suggestive of contamination, and CroCo found no evidence of cross-contamination among samples. Therefore, we proceeded in the analysis with all the samples.

Phylogenomic analysis with the full taxon sampling supported the monophyly of each* Drosera *subgenus and recovered extensive gene tree discordance in two areas

Our phylogenomic analyses with all samples recovered low support for the monophyly of *Drosera*. ASTRAL recovered *Drosera* as monophyletic, while RAxML recovered *Drosera* polyphyletic with *D. regia* (*D.* subg. *Regiae*) + *D. murfetii* (*D.* subg. *Arcturia*) being sister to *Dionaea* + *Aldrovanda* (Fig. 1). Only 296 of the 853 MO ortholog gene trees with a bootstrap >80% ("informative genes" hereafter) at this node supported the monophyly of *Drosera* (Fig. 1). This corresponded to the cloud of different gene tree topologies at the base of Droseraceae (Fig. 1C).

Regarding core *Drosera*, gene trees strongly supported the monophyly of *D.* subg. *Drosera* (1859/1885 informative gene trees) and *D.* subg. *Ergaleium* (1761/1801) and their sister relationship (1827/1865).

Except for *D.* sect. *Drosera*, all *Drosera* sections represented by multiple species in our sampling were monophyletic in both the RAxML and ASTRAL phylogenies. Concerning *D.* sect. *Drosera, D. spatulata* was recovered by both RAxML and ASTRAL as sister to *D.* sect. *Brasilianae* + *D.* sect. *Ptycnostigma* instead of forming a clade with the other four species of *D.* sect. *Drosera* (Fig. 1). The paraphyly of *D.* sect. *Drosera* coincided with elevated gene tree discordance and uninformative gene trees both within and among *D.* sect. *Drosera*, *D.* sect. *Ptycnostigma*, and *D.* sect. *Brasilianae* (Fig. 1).

Within-species Ks plots in 16 species across Droseraceae showed evidence of recent polyploidy events (Ks < 0.5; Figs. S2–S3). *Aldrovanda vesiculosa, D. regia,* and *D. murfetii* each had a Ks peak around 0.17–0.23. In *D.* subg. *Drosera,* all species from *D.* sect. *Brasilianae, D.* sect. *Ptycnostigma,* and *D.* sect. *Psychophila* had a slightly broader Ks distribution from 0 to 0.05 than other species in the subgenus. A similarly

slightly broadened Ks distribution is also observed in *D. meristocaulis, D. roseana, D. paradoxa,* and *D. caduca* of *D.* subg. *Ergaleium.*

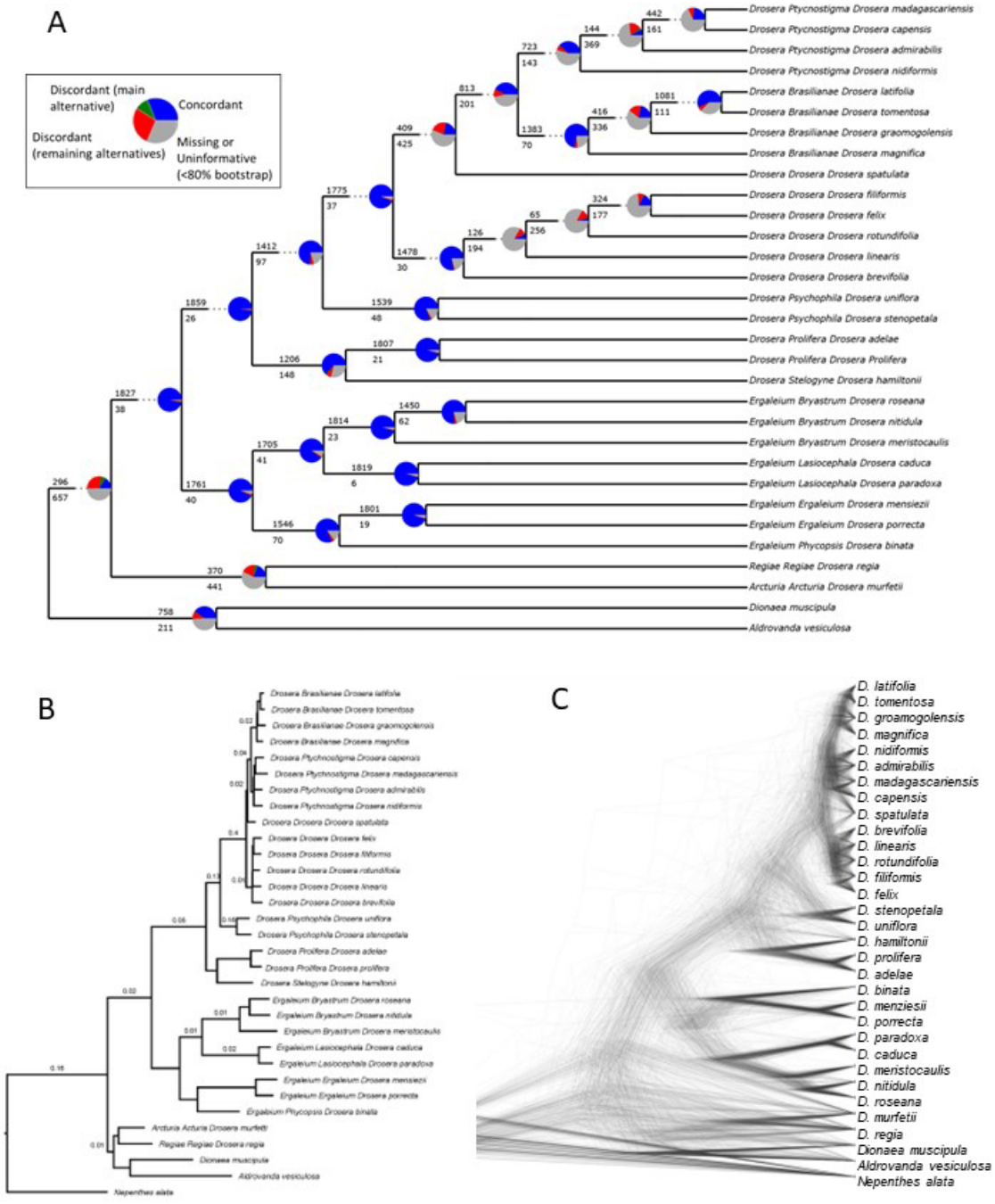


Figure 1: Droseraceae species tree with gene duplications and discordance. A. ASTRAL species tree estimated from 1900 MO ortholog trees with the discordance calculated by PhyParts mapped in pie charts on each node. Above the branch is the number of informative MO orthologs supporting the species tree topology. Below the branch are the number of informative MO orthologs supporting alternate topologies. B. RAxML tree from concatenating 1900 MO orthologs. Above each branch is the percentage of homologs that have a gene duplication event with MRCA

mapped to that node (value missing when zero). C. Cloudogram of 411 MO orthologs with no missing taxa.

Of the clades with evidence for recent polyploidy, only *D.* sect. *Psychophila* had an increased number of genes with duplications mapped to its MRCA (Fig. 1). 16% of informative genes had a gene duplication event mapping to the MRCA of *D.* sect. *Psychophila* (Fig. 1). While no polyploid event was inferred in the chromosome number ancestral state reconstruction (Chapter 1), within and between species Ks plots, and mapping gene duplication events both support a shared polyploidy event at the MRCA of the species we sampled in *D.* sect. *Psychophila*. The MRCA of Droseraceae and the MRCA of *D.* sect. *Brasilianae* + *D.* sect. *Ptycnostigma* + *D.* sect. *Drosera* each had 16% and 40% of homologs with a gene duplication event mapped to that node, respectively. However, not all species included in either clade shared a Ks peak, suggesting the presence of allopolyploid events.

The findings of discordance at the base of Droseraceae, in the placement of *D. spatulata,* and the relationship among *D.* sect. *Drosera, D.* sect. *Ptycnostigma,* and *D.* sect. *Brasilianae* led to subsampling and analysis to test for reticulation among Droseraceae and within *D.* subg. *Drosera.*

Allopolyploid origin of Drosera *subg.* Regiae + D. *subg.* Arcturia

Given the elevated level of gene tree discordance at the base of Droseraceae, we constructed the Droseraceae subsampling dataset that included all three Droseraceae genera and four *Drosera* subgenera with *Nepenthes alata* as the outgroup. We identified 1487 one-to-one orthologs with no missing samples. In this dataset, both ASTRAL and RAxML analyses supported *Drosera* being monophyletic (Fig. 2). However, similar to the all-taxa analysis, the monophyly of *Drosera* was supported by only 262 out of 857 informative genes, and 129 and 116 genes supported *D. regia* or *D. regia* + *D. murfetii*, respectively, being sister to *Aldrovanda vesiculosa* + *Dionaea muscipula* (Fig. 2). Therefore, the high levels of discordance among *D. murfetii*, *D. regia*, and core *Drosera* suggested either incomplete lineage sorting or reticulation (Fig. 2). In contrast, gene trees strongly supported the monophyly of core *Drosera* (*D. hamiltonii* + *D. porrecta*: 1305/1334 informative genes) and *Dionaea* + *Aldrovanda* (777/967).

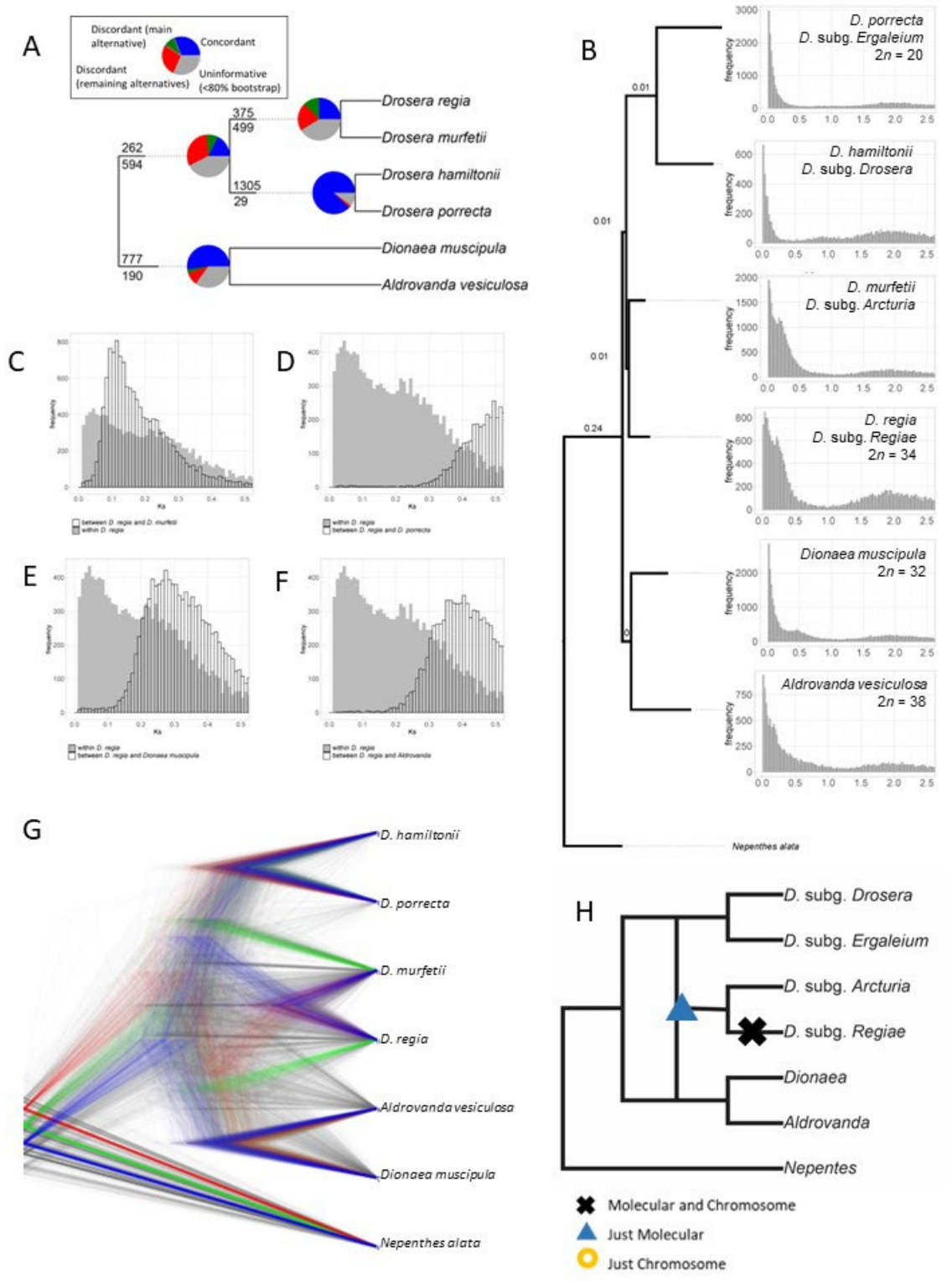


Figure 2. *Drosera regia* and *D. murfetii* share an allopolyploidy event between the stem lineage of core *Drosera* and the stem lineage leading to *Dionaea muscipula* + *Aldrovanda vesiculosa.* A. ASTRAL species tree from 1487 one-to-one ortholog trees with the gene tree discordance calculated by PhyParts mapped in pie charts on each node. B. RAxML tree from concatenating

1487 one-to-one orthologs. Values above branches were the proportion of genes with the MRCA of gene duplications mapped from 6916 homologous gene trees with 80% average bootstrap support. Within-species 0 to 2.5 Ks distribution between are plotted at the tip of each in-group species. C–F. Comparison of within- vs. between-species Ks distribution suggested that only *Drosera murfettii* and *D. regia* shared a polyploid event before their speciation. G. Cloudogram: blue indicates the most common topology, red indicates the second most common topology, and green indicates the third most common topology. Gray represents the remaining topologies. Colors used in cloudogram indicate alternative overall topologies, which are different from the coloring of local topologies in A. H. The reticulations and duplications inferred from the discordance, Ks plots, mapping gene duplications, and cloudogram as compared to those inferred from the chromosome ancestral state reconstruction (Chapter 1, Mohn et al., 2022) for the backbone of Droseraceae.

The Ks peaks and mapped gene duplications suggested that the gene tree discordance was caused by an allopolyploidy event. Both *D. regia* of *D.* subg. *Regiae* and *D. murfetii* of *D.* subg. *Arcturia* had a Ks peak at ~0.23, with a more recent between-species Ks peak at ~ 0.11, supporting a shared polyploid event at a Ks of ~0.23. The Ks peaks between *D. regia* and all other species in our Droseraceae subsampling dataset were older than 0.23 indicating that *D. murfetii* + *D. regia* did not share the polyploid event with any other lineage (Fig. 2). However, branch lengths in *D. regia* and *D. murfetii* are much shorter than core *Drosera* and somewhat shorter than *Dionaea* and *Aldrovanda* making comparison of Ks plots and interpretations of between-species Ks values challenging (Fig. 1-2). Nonetheless, only 1% of genes had gene duplication events mapping to the MRCA of *D. regia* + *D. murfetii,* and another 1% mapped to the MRCA of *Drosera.* Notably, 24% of genes have a gene duplication event mapped to the MRCA of Droseraceae (Fig. 2). As observed here, a high percentage of duplicated genes at the base of a clade in which not all taxa shared a Ks peak is indicative of an allopolyploid event. Specifically, a large percentage of genes had a gene duplication mapped to the MRCA of Droseraceae, but only *D. regia* and *D. murfetii* share a Ks peak (Fig. 2) supporting an allopolyploid event between *Dionaea* + *Aldrovanda* and core *Drosera* giving rise to *D. regia* and *D. murfetii*.

An allopolyploid origin of *D. regia* + *D. murfetii* would also explain the differences in their placement between major gene tree topologies. Depending on which paralog was present, three major topologies were found in one-to-one gene trees (Fig. 2G). The most common tree topology (blue) placed *D. regia* + *D. murfetii* as sister to the core *Drosera* (12% gene trees), and the second placed *D. regia* + *D. murfetii* sister to

Dionaea + *Aldrovanda* (red; 9.4%). The remaining topologies (green and gray) varied in their placement of *D. regia* and *D. murfetii* at the base of Droseraceae. This is not surprising given the short internal branch lengths and that uninformative genes made up 58-65% of genes at all nodes except for the MRCA of the core *Drosera* (Fig. 2B). Patterns of discordance at the MRCA of *Drosera* and these major topologies recovered are consistent with an allopolyploid origin of paralogs in informative gene trees. Therefore, Ks plots, gene tree discordance, mapping gene duplications, and the cloudogram all support an allopolyploidy origin of *D.* subg. *Regiae* + *D.* subg. *Arcturia* from the stem lineage leading to core *Drosera* and the stem lineage leading to *Dionaea* + *Aldrovanda* (Fig. 2).

This polyploidy event aligns with previous findings. While previous ancestral state chromosome number reconstruction did not find evidence to support a shared polyploidy event in *D.* subg. *Regiae* and *D.* subg. *Arcturia* (Chapter 1), it did find evidence for a polyploidy event in *D. regia* alone. This was also supported by a second Ks peak at ~0.05 in this species. Our findings of two polyploidy events in *D. regia,* one shared with *D. murfetii* and a duplication in its own lineages are supported by the results of a *D. regia* genome assembly (Renner et al., 2019). Renner et al. found synteny evidence for two duplications in *D. regia* neither of which were shared by *D. capensis* of *D.* subg. *Drosera* (2019). An allopolyploid event also explains the discordance between ITS and *rbc*L phylogenies described by (Rivadavia et al., 2003). Additionally, previous study of Droseraceae pollen morphology identified similarities between *Dionaea muscipula* and *Drosera regia* (Takahashi and Sohma, 1982). *Drosera* subg. *Arcturia* was not included in these analyses, but its pollen has been reported to be operculate like *Dionaea* and *Drosera regia* (Fleischmann et al., 2018).

Based on Ks plots, previous publications have proposed a polyploidy event shared by all of Droseraceae (Yang et al., 2018; Palfalvi et al., 2020) but genome assemblies of *D. regia* and *D. capensis* did not find evidence for a shared polyploidy (Renner et al., 2019). We also observed a slight Ks uptick in *D. spatulata, D. binata, Aldrovanda vesiculosa, Dionaea muscipula,* and *Drosera hamiltonii* at ~0.5–0.8. However, this uptick was much shorter than the shared triplication of Eudicots and was indistinguishable in the

remaining *Drosera* samples (Fig S3). Further analysis with additional outgroups may be helpful in ascertaining the cause of this Ks slight uptick.

Polyploid origin of* Drosera *sect.* Brasilianae + D. *sect.* Ptycnostigma *from the lineage leading to D. spatulata and an unsampled/extinct lineage

Analyses of both the full sampling dataset and the *D.* subg. *Drosera* subsampling dataset strongly supported the topology among *D.* subg. *Drosera* sections found in previous publications for all sections except *D.* sect. *Brasilianae*, *D.* sect. *Ptycnostigma*, and *D.* sect. *Drosera* (Rivadavia et al., 2003; Fleischmann et al., 2018). While 1373 / 1393 informative gene trees supported the monophyly of *D.* sect. *Brasilianae* + *D.* sect. *Ptycnostigma* + *D.* sect. *Drosera* (represented by *D. spatulata* and *D. filiformis* in the *D.* subg. *Drosera* dataset), PhyParts found a high level of discordance within this clade. Of the 882 informative orthologs at the MRCA of *D. spatulata* + *D. filiformis*, 388 supported *D.* sect. *Drosera* being monophyletic with *D. spatulata* and *D. filiformis* sister to each other (Fig. 3). Another 233 orthologs supported the alternative topology with *D. filiformis* sister to *D. spatulata* + *D*. sect. *Brasilianae* + *D*. sect. *Ptycnostigma* (Fig. 3).

This discordance corresponds to increased gene duplications and evidence of an allopolyploid event. We found 17% of genes had a duplication that mapped to the MRCA of *D.* sect. *Brasilianae* + *D.* sect. *Ptycnostigma* + *D.* sect. *Drosera* (Fig. 3). This also fits with our observation that, of the 7071 homolog gene clusters that after trimming had all 8 species, only 1479 had a single copy per species (Fig. 3). However, within the clade *Brasilianae* + *D.* sect. *Ptycnostigma* + *D.* sect. *Drosera*, only species from *D.* sect *Ptycnostigma* and *D.* sect *Brasilianae* showed a broader Ks distribution at Ks = 0 to 0.05 (Fig. 3), suggesting that a shared allopolyploidy event by these two sections gave rise to the MRCA of the gene duplications mapping to a deeper node.

Given the high level of discordance among *D.* sect. *Drosera, D.* sect. *Ptycnostigma,* and *D.* sect. *Brasilianae,* we visualized both within and between-species Ks distribution across these sections to try to pinpoint parental lineages. Between-species Ks plots showed an almost simultaneous divergence among *D.* sect. *Ptycnostigma, D.* sect. *Brasilianae,* and *D. spatulata* that overlapped with the within-species Ks peak in *D.* sect. *Ptycnostigma* and *D.* sect. *Brasilianae* (Fig. 3, S4). On the other hand, *D. filiformis*

diverged from all these taxa before the polyploidy event (Fig. S4). Thus, *D. spatulata* represents the only potential parental lineage in our sampling. However, Ks plots did not provide strong support for or against a reticulate origin of the polyploidy clade *D.* sect. *Brasilianae* + *D.* sect. *Ptycnostigma*.

An allopolyploid event also explains the two most common tree topologies in the cloudogram. Densitree found that the most frequent tree topology agreed with the topology of the RAxML and ASTRAL trees supporting a monophyletic *D*. sect. *Drosera* (16% of trees; blue), the second most frequent topology had the same overall topology except *D. spatulata* being sister to *D*. sect. *Braslianae* + *D*. sect. *Ptycnostigma* (14% of trees; red), and the third most frequent topology (6%) had *D. spatulata* sister to *D. filiformis* + *D.* sect. *Braslianae* + *D.* sect. *Ptycnostigma* (green; Fig. 3). If this gene tree discordance was due to incomplete lineage sorting alone, the two alternative topologies would have similar frequencies. However, in an allopolyploidy scenario, we would expect the two most frequent topologies having approximately equal frequencies, which is consistent with what we are seeing here. The remaining topologies vary in their placement of these four taxa or their placement of *D.* sect. *Stelogyne*, *D.* sect. *Prolifera,* and *D.* sect. *Psychophila* (Fig. 3).

In summary, while elevated frequencies of gene duplications mapped to the MRCA of *D.* sect. *Drosera* + *D.* sect. *Ptycnostigma* + *D.* sect. *Brasilianae,* only *D.* sect. *Ptycnostigma* and *D.* sect. *Brasilianae* shared a Ks peak and very low frequencies of gene duplications mapped to the MRCA of *D.* sect. *Ptycnostigma* + *D.* sect. *Brasilianae*. The two most common gene tree topologies of either monophyletic *D.* sect. *Drosera* or *D. spatulata* sister to *D.* sect. *Ptycnostigma* + *D.* sect. *Brasilianae* further support that incomplete lineage sorting alone cannot explain the distribution of gene tree topologies. Instead, *D.* sect. *Ptycnostigma* + *D.* sect. *Brasilianae* were likely of allopolyploidy origin with one parent closely related to *D. spatulata*. The other parent is either unsampled or extinct, as all the other samples in our analysis diverged from *D.* sect. *Ptycnostigma* + *D.* sect. *Brasilianae* before the polyploidy event according to Ks peaks.

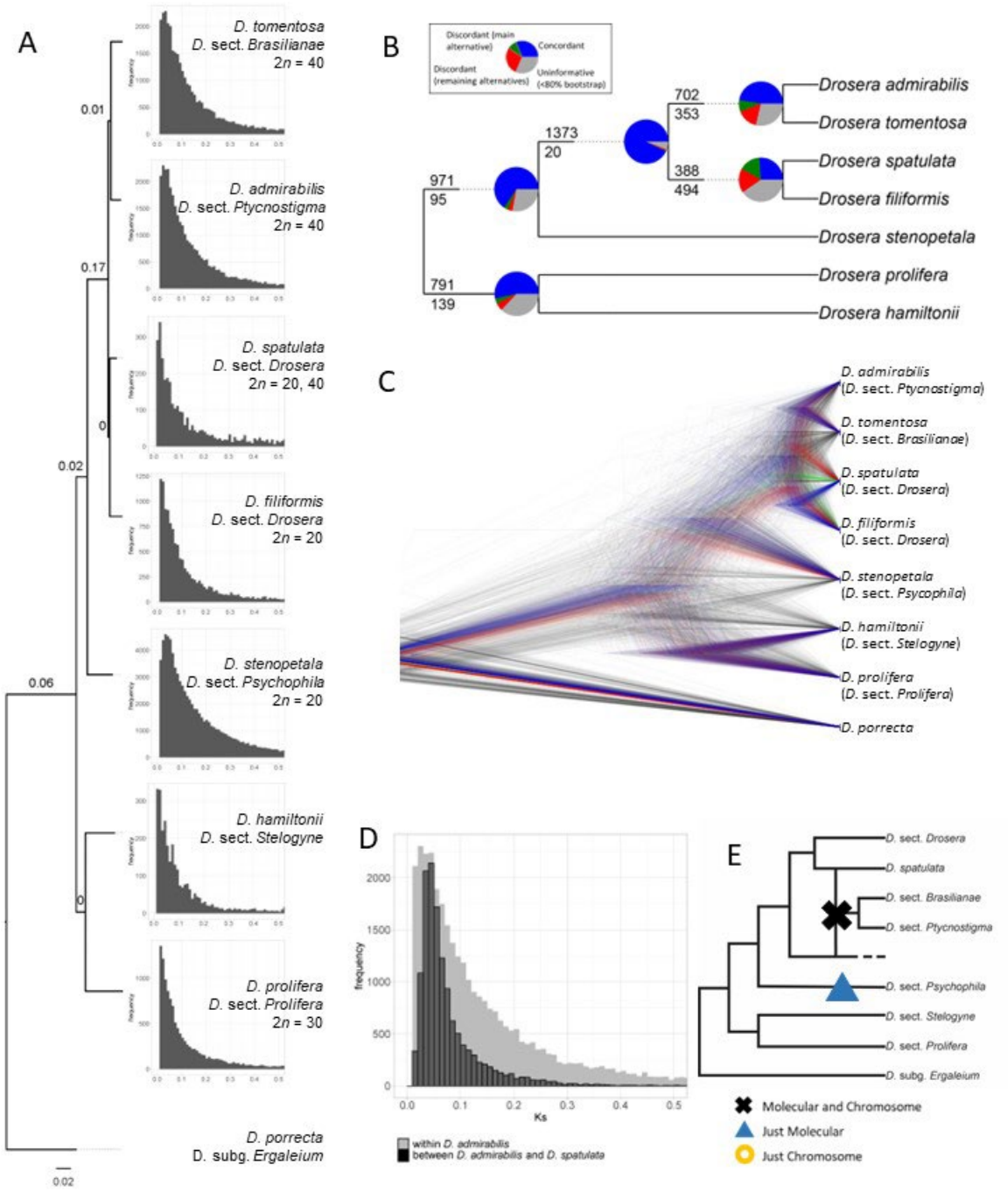


Figure 3: Ks Plots, gene tree discordance, gene duplications, and the gene tree cloudogram all support an allopolyploid event giving rise to *D.* sect. *Brasilianae* and *D.* sect. *Ptycnostigma*. A. RAxML concatenated tree from 1479 one-to-one orthologs with percent of duplications mapped from 7313 gene trees with an 80% bootstrap filter above the branches. Within-species Ks plots 0 to 0.5 are plotted beneath the label of each in-group taxa. B. ASTRAL tree with 1479 ortholog trees with the gene tree discordance calculated by PhyParts mapped in pie charts on each node. C. Cloudogram: blue most common topology, red second most common topology, and green third most common topology. D. Ks plot of *D. spatulata*, *D. admirabilis*, and the pairwise comparison. E. The reticulations and duplications inferred from the discordance, Ks plots, duplications most

recent common ancestor, and cloudogram as compared to those inferred from the chromosome ancestral state reconstruction (Chapter 1, Mohn et al., 2022) for all of *D.* subg. *Drosera*.

The polyploidy event at the MRCA of *D.* sect. *Brasilianae* + *D.* sect. *Ptycnostigma* was further supported by our previous chromosome ancestral state reconstruction (Chapter 1, Mohn et al., 2022). Most species in *D.* sect. *Brasilianae* and *D.* sect. *Ptycnostigma* have chromosome counts of $2n = 40$ while most species in *D.* sect. *Drosera* has a chromosome count of $2n = 20$ (Mohn et al., 2022). The ancestral state reconstruction of chromosome numbers inferred a polyploidy event at the node giving rise to *D.* sect. *Brasilianae,* and *D.* sect. *Ptycnostigma* (Mohn et al., 2022). Interestingly, *D. spatulata* has small chromosomes more similar to *D.* sect. *Brasilianae* and *D.* sect. *Ptycnostigma* than to other *D.* sect. *Drosera* (Nakamura and Ueda, 1991; Hoshi et al., 2017). Though a well-supported polyploidy event, the allopolyploidy origin of *D.* sect. *Brasilianae* + *D.* sect. *Ptycnostigma* is moderately supported from gene tree topology alone, with synonymous distribution being equivocal in allo- vs. autopolyploidy. Further hypothesis testing would be helpful to distinguish the two scenarios.

The well-supported clade consisting of *D.* sect. *Drosera, D.* sect. *Ptycnostigma,* and *D.* sect. *Brasilianae* is the most geographically widespread in the genus. *Drosera* sect. *Brasilianae* is restricted to South America, *D.* sect. *Ptycnostigma* is restricted to Africa, *D. spatulata* is widespread throughout Southeast Asia and Oceania, and *D.* sect. *Drosera* is widespread across Oceania, Asia, Europe, North America, and South America. Further research could explore the role of habitat specialization, biogeography, etc. in the context of a rapid radiation in this group.

Phylogenomic analyses did not support three previously inferred polyploidy events in Drosera *subg.* Ergaleium

In the full phylogenomic analysis (Fig. 1), *D.* subg. *Ergaleium* had little gene tree discordance, and all nodes had less than 2% of genes with duplications. The lack of deep polyploidy events in the subgenus was also supported by the lack of distinct Ks peaks in species of *D.* subg. *Ergaleium* (Fig. 4). This suggests that the chromosome number doubling event leading to *D.* sect. *Ergaleium* previously inferred by modeling

chromosome evolution is likely due to chromosome fission instead of polyploidy events (Mohn et al., 2022).

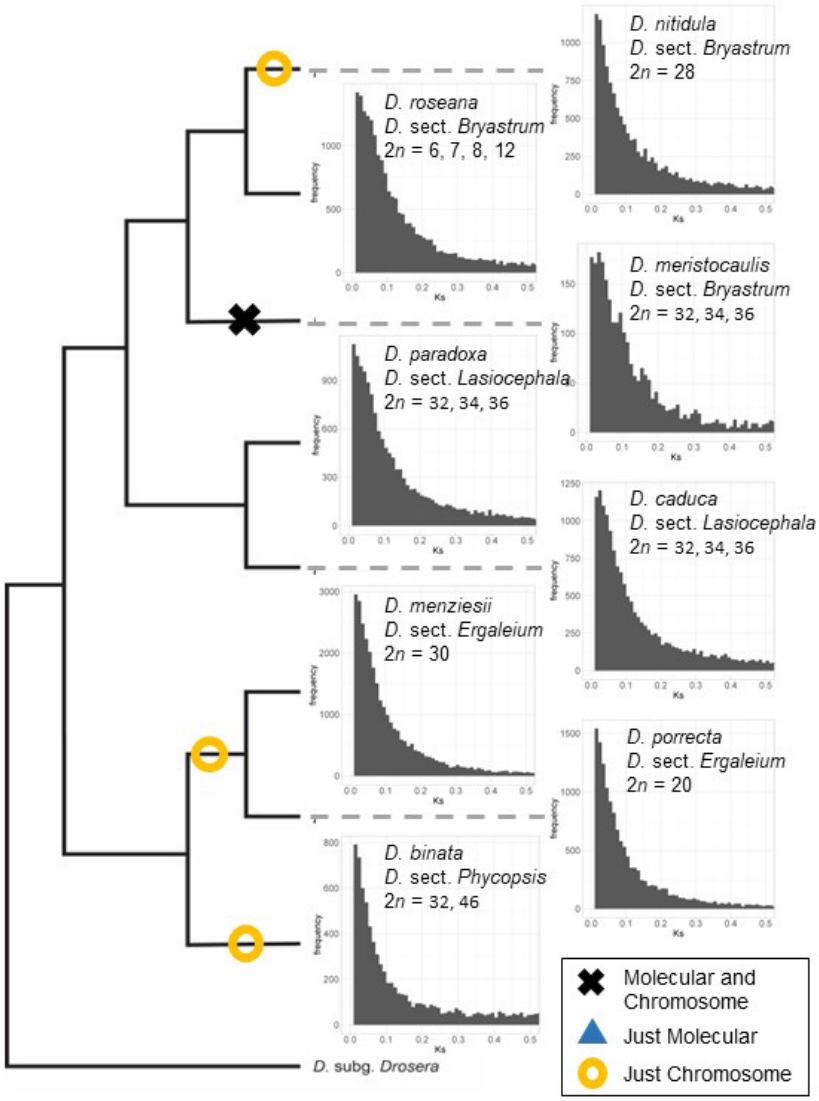


Figure 4: *Drosera* subg. *Ergaleium* species tree topology with within-species Ks plots. *Drosera meristocaulis*, *D. roseana*, *D. caduca,* and *D. paradoxa* show some broadening Ks distribution at ~0.4. In *D.* subg. *Ergaleium,* three of the whole genome duplication events inferred from chromosome counts were not supported by our molecular data, suggesting that chromosome fission may play an important role in this subgenus than the other groups.

Of the other three polyploidy events inferred from chromosome number evolution (Fig. 4), only one was supported by molecular evidence. Previous modeling analyses using chromosome numbers reconstructed a polyploidy event leading to *D. binata* and another event within *D.* sect. *Bryastrum* that is represented by *D. nitidula* in our taxon

sampling (Mohn et al., 2022). However, neither the Ks plot of *D. binata* nor of *D. nitidula* had a peak corresponding to the putative polyploidy event. The discrepancy between modeling chromosome number vs. phylogenomics may be due to either very recent polyploidy events that are indistinguishable with Ks plots, or alternatively, the extraordinarily high chromosome fission rate in *D.* subg. *Ergaleium* (Mohn et al., 2022). Importantly, *D. nitidula* has a low allele divergence and loci heterozygosity further supporting its diploidy and a chromosome fission event (Fig. 5). *Drosera meristocaulis* on the other hand, has a slightly wider Ks distribution than closely related species at ~0.05 (Fig. 4). With its high allele divergence and loci heterozygosity (Fig. 5), our molecular data supports its polyploid ancestral state reconstruction in Chapter 1 (Mohn et al., 2022).

The slight broadening of Ks distribution near zero in *D. roseana, D. caduca, and D. paradox* may be due to increased heterozygosity associated with self-incompatibility (see discussion below), aneupolyploidy, or very recent autopolyploid events. For example, *D. roseana* has been reported to have a chromosome number of $2n = 6, 7, 8,$ and 12 (Sheikh and Kondo, 1995; Chen et al., 1997), and we do not have chromosome count data on the sample sequenced.

Loci heterozygosity and allele divergence correlates with both ploidy levels and mating systems

Loci heterozygosity and allele divergence both provide additional evidence to detect more recent polyploidy events and insight into other factors impacting the Ks distributions in *D.* subg. *Ergaleium.* Read mapping to single copy genes in *D. spatulata* recovered seven species with low loci heterozygosity (<50%) and low allele divergence (<0.75%), seven species with high loci heterozygosity (>50%) and medium allele divergence (0.75-2%), and 14 species with high loci heterozygosity (>50%) and high allele divergence (>2.5%; Fig. 5). Targeted assembly of *D. binata* using HybPiper failed and so was not included in these results.

All 12 species with both Ks plots and gene duplication mapping supporting a polyploid history had high allele divergence (>2.5%) and high loci heterozygosity (>50%; Fig. 5). This is expected as π (nucleotide diversity) and divergence between both

paralogs/homeologs contributed to allele divergence, so species with an increased number of paralogs or homeologs should have an increased allele divergence. Two additional species, *Aldrovanda vesiculosa* and *D. meristocaulis*, both of which had recent Ks peaks, also had high allele divergence and high loci heterozygosity supporting their polyploid history.

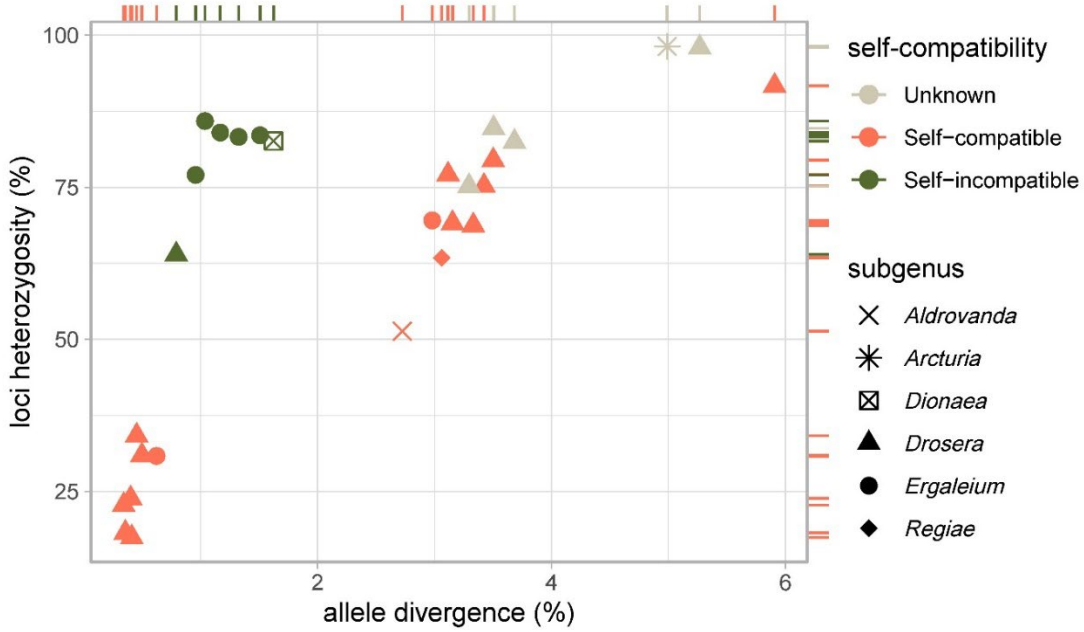


Figure 5: Allele divergence (paralog divergence and π per site) and loci heterozygosity split Droseraceae species into three groups: the self-compatible diploids, all polyploids, and the self-incompatible diploids. Mating system (data from Chapter 1) was noted as unknown in species without data and in species where populations differ in self-compatibility so is not known for our sample.

The much higher allele divergence in *D. murfetii* than *D. regia,* suggests that *D. murfetii* may have undergone its own polyploid recent polyploid event not detected on the Ks plot. So far, no chromosome count has been reported for *D. murfetii*. Chromosome counts of $2n = 58$ (Shirakawa et al., 2011) and $2n=20$ (Kondo and Whitehead, 1971) have both been reported for its close relative *D. arcturi*. However, because of the change in taxonomy and the lack of specificity in location information specifically for the $2n = 58$ count, it remains unclear whether this chromosome count belongs to *D. arcturi* or *D. murfetii*.

Of the remaining species, the seven species with high loci heterozygosity and medium allele divergence are all either self-incompatible or monoecious, and the seven species with both low allele divergence and loci heterozygosity are self-compatible (Fig. 5; Chapter 1; Mohn et al., 2022). If these species are diploids, the allele divergence would be contributed to less by paralog divergence and is approximately equal to the nucleotide diversity statistic, π. Breeding system impacts π with outcrossing species having less than twice the level of π as selfing species (Glémin et al., 2006). This is approximately the difference observed in allele divergence between the medium allele divergence and the low allele divergence groups supporting that these species, notably most *D.* subg. *Ergaleium,* are all diploids and not recent polyploids.

The increased single chromosome evolution of *D.* subg. *Ergaleium* found in Chapter 1 was further supported by our phylogenomic analyses. Of the six polyploidy events from Chapter 1 that overlapped with the taxon sampling in this chapter, three were further supported by phylogenomic analyses evidence. The remaining three with no evidence from phylogenomic data were all in *D.* subg. *Ergaleium* and were likely due to chromosome fission instead of whole genome duplication. This further supports the high chromosome fission rate we observed in *D.* subg. *Ergaleium* (Chapter 1, Mohn et al., 2022).

While *D.* sect. *Bryastrum* species (like the rest of *D.* subg. Ergaleium) are mostly self-incompatible (Chapter 1; Chen et al., 1997; Mohn et al., 2022) and have highly elevated rates of single chromosome number changes and within-species chromosome number variations, a clade within *D.* sect. *Bryastrum* that was represented by *D. nitidula* in our sampling (the *D. nitidula* clade from now on) is a notable exception. Species in the *D. nitidula clade* are mostly self-compatible and do not appear to have as much single chromosome number variation (Chen et al., 1997; Mohn et al., 2022). Modeling chromosome number evolution (Chapter 1, Mohn et al., 2022) and the genomic work presented here supported a chromosome number doubling event likely due to fission early in the section, with no subsequent chromosome number changes or any within-species chromosome number variation reported (Mohn et al., 2022). This raises the question for future research of whether self-incompatibility and chromosome number are linked in a mechanism such as meiotic drive.

CONCLUSION

Our phylogenomic analyses using transcriptomes and genomes across *Droseraceae* found strong evidence for two reticulation events along the backbone of *Drosera:* one leading to *D.* subg. *Regiae* and *D.* subg. *Arcturia* and the other leading to *D.* sect. *Brasilianae* and *D.* sect. *Ptycnostigma.* These reticulation events yield both *Drosera* and *D.* sect. *Drosera* polyphyletic and potentially require updates to the taxonomy.

A total of eight polyploid events were investigated in both this work and Chapter 1 (Mohn et al., 2022) given the overlapping taxon sampling. Only three were supported by both molecular evidence and chromosome ancestral state reconstructions (Fig. 2, 3, 4). Two were supported by molecular data alone, and three were only supported by chromosome count reconstructions (Fig. 2, 3, 4), likely due to highly elevated chromosome fission instead of polyploidy (Mohn et al., 2022).

Moving forward we will increase the taxon sampling to include samples from the remaining three of the 15 sections in *Drosera* and carry out phylogenetic network analysis and hypothesis testing to test for reticulation events. Future work should further explore the genomic rearrangements in *Droseraceae*, and the role of meiotic drive and centromere evolution in diversification of this charismatic group of plants.

SUPPLEMENTAL MATERIALS

Figure S1: Neither the number of the base pairs nor the RNA integrity number appear to impact the number of reference sequence identities found after filtering.

Figure S2: Ks plots with Ks values 0 to 0.5.

Figure S3: Ks plots with Ks values 0 to 2.5.

Figure S4: Between species Ks plots for *D. admirabilis, D. tomentosa, D. filiformis,* and *D. spatulata*.

Table S1: Newly sequenced samples for this publication.

Supplementary Methods: Modified PureLink RNA extraction protocol.

Supplementary Materials: Photo vouchers.

REFERENCES

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

Andrew, S. 2010. FastQC: A Quality Control tool for High Throughput Sequence Data.

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, et al. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* 19: 455–477.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* 30: 2114–20.

Bouckaert, R. R. 2010. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* 26: 1372–1373.

Brown, J. W., J. F. Walker, and S. A. Smith. 2017. Phyx: Phylogenetic tools for Unix J. Kelso [ed.],. *Bioinformatics* 33: 1886–1888.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST+: architecture and applications. *BMC bioinformatics* 10: 421.

Chen, L., S. H. James, and H. M. Stace. 1997. Self-incompatibility, seed abortion and clonality in the breeding systems of several Western Australian *Drosera* species (Droseraceae). *Australian Journal of Botany* 45: 191.

Darwin, C. 1875. Insectivorous Plants. 1st ed. John Murray, London.

Darwin, C. 1985. Letter no. 2996. *In* F. Burkhardt and et al [eds.], The Correspondence of Charles Darwin Vol. 8, Cambridge University Press.

Davidson, N. M., and A. Oshlack. 2014. Corset: Enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biology* 15: 410.

Fishman, L., J. H. Willis, C. A. Wu, and Y.-W. Lee. 2014. Comparative linkage maps suggest that fission, not polyploidy, underlies near-doubling of chromosome number within monkeyflowers (Mimulus; Phrymaceae). *Heredity* 112: 562–568.

Fleischmann, A., A. T. Cross, R. Gibson, P. M. Gonella, and K. W. Dixon. 2018. Systematics and evolution of Droseraceae. *In* A. Ellison, and L. Adamec [eds.], Carnivorous Plants: Physiology, Ecology, and Evolution, 45–57. Oxford University Press, Oxford.

Glémin, S., E. Bazin, and D. Charlesworth. 2006. Impact of mating systems on patterns of sequence polymorphism in flowering plants. *Proceedings of the Royal Society B: Biological Sciences* 273: 3011–3019.

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* 8: 1494–512.

Haas, BJ. TransDecoder.

Hoshi, Y., M. Azumatani, C. Suyama, and L. Adamec. 2017. Determination of Ploidy Level and Nuclear DNA Content in the Droseraceae by Flow Cytometry.

Johnson, M. G., E. M. Gardner, Y. Liu, R. Medina, B. Goffinet, A. J. Shaw, N. J. C. Zerega, and N. J. Wickett. 2016. HybPiper: Extracting Coding Sequence and

Introns for Phylogenetics from High-Throughput Sequencing Reads Using Target Enrichment. *Applications in Plant Sciences* 4.
Kondo, K., and B. Whitehead. 1971. Chromosome number of *Drosera arcturi* Hook. *Journal of Japanese botany* 46: 344.
Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* 9: 357–359.
Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
Li, W., and A. Godzik. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658–1659.
Löytynoja, A. 2014. Phylogeny-aware alignment with PRANK. *In* D. J. Russell [ed.], Multiple Sequence Alignment Methods, Methods in Molecular Biology, 155–170. Humana Press, Totowa, NJ.
Mohn, R. A., R. Zenil-Ferguson, T. A. Krueger, A. S. Fleischmann, A. T. Cross, and Y. Yang. 2022. Over two orders of magnitude difference in rate of single chromosome loss among sundew *(Drosera* L., Droseraceae) lineages. 2022.10.24.513289.
Morales-Briones, D. F., G. Kadereit, D. T. Tefarikis, M. J. Moore, S. A. Smith, S. F. Brockington, A. Timoneda, et al. 2021. Disentangling Sources of Gene Tree Discordance in Phylogenomic Data Sets: Testing Ancient Hybridizations in Amaranthaceae s.l. *Systematic Biology* 70: 219–235.
Nakamura, T., and K. Ueda. 1991. Phytogeography of Tokai Hilly Land Element : II. Taxonomic study of *Drosera tokaiensis* (Komiya & C. Shibata) T. Nakamura & Ueda (Droseraceae). *Acta phytotaxonomica et geobotanica* 42: 125–137.
Nauheimer, L., N. Weigner, E. Joyce, D. Crayn, C. Clarke, and K. Nargar. 2021. HybPhaser: A workflow for the detection and phasing of hybrids in target capture data sets. *Applications in Plant Sciences* 9.
Palfalvi, G., T. Hackl, N. Terhoeven, T. F. Shibata, T. Nishiyama, M. Ankenbrand, D. Becker, et al. 2020. Genomes of the Venus Flytrap and Close Relatives Unveil the Roots of Plant Carnivory. *Current Biology*.
Renner, T., V. A. Albert, D. Sankoff, C. Zheng, Q. W. Tan, K. Fukushima, C. P. Drummond, et al. 2019. Deep-time morphological stasis in the carnivorous plant genus *Drosera* despite different trajectories of genomic upheaval. VIB Conference Series, Ghent, BE.
Rivadavia, F., K. Kondo, M. Kato, and M. Hasebe. 2003. Phylogeny of the sundews, *Drosera* (Droseraceae), based on chloroplast rbcL and nuclear 18S ribosomal DNA sequences. *American Journal of Botany* 90: 123–130.
Rosenberg, O. 1903. Das Verhalten der Chromosomen in einer hybriden Pflanze. *Berichte der Deutschen Botanischen Gesellschaft* 21: 110–119.
Sheikh, S. A., and K. Kondo. 1995. Differential staining with orcein, giemsa, CMA, and DAPI for comparative chromosome study of 12 species of Australian Drosera (Droseraceae). *American Journal of Botany* 82: 1278–1286.
Shirakawa, J., K. Nagano, and Y. Hoshi. 2011. A chromosome study of two centromere differentiating *Drosera* species, *D. regia* and *D. arcturi*. *Caryologia* 64: 453–463.

Simion, P., K. Belkhir, C. François, J. Veyssier, J. C. Rink, M. Manuel, H. Philippe, and M. J. Telford. 2018. A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data. *BMC Biology* 16: 28.

Smith, S. A., M. J. Moore, J. W. Brown, and Y. Yang. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* 15: 150.

Smith, S. A., and B. C. O'Meara. 2012. TreePL: Divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* 28: 2689–2690.

Smith-Unna, R., C. Boursnell, R. Patro, J. M. Hibberd, and S. Kelly. 2016. TransRate: Reference-free quality assessment of de novo transcriptome assemblies. *Genome Research* 26: 1134–1144.

Song, L., and L. Florea. 2015. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience* 4: 48.

Takahashi, H., and K. Sohma. 1982. Pollen Morphology of the Droseraceae and Its Related Taxa. Tohoku University.

Van Dongen, S. 2008. Graph Clustering Via a Discrete Uncoupling Process. *SIAM Journal on Matrix Analysis and Applications* 30: 121–141.

Veleba, A., P. Šmarda, F. Zedek, L. Horová, J. Šmerda, and P. Bureš. 2017. Evolution of genome size and genomic GC content in carnivorous holokinetics (Droseraceae). *Annals of Botany* 119: 409–416.

Walker, J. F., Y. Yang, M. J. Moore, J. Mikenas, A. Timoneda, S. F. Brockington, and S. A. Smith. 2017. Widespread paleopolyploidy, gene tree conflict, and recalcitrant relationships among the carnivorous Caryophyllales. *American Journal of Botany* 104: 858–867.

Yang, Y., M. J. Moore, S. F. Brockington, J. Mikenas, J. Olivieri, J. F. Walker, and S. A. Smith. 2018. Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy events. *New Phytologist* 217: 855–870.

Yang, Y., M. J. Moore, S. F. Brockington, D. E. Soltis, G. Ka-Shu Wong, E. J. Carpenter, Y. Zhang, et al. 2015. Dissecting Molecular Evolution in the Highly Diverse Plant Clade Caryophyllales Using Transcriptome Sequencing. *Molecular Biology and Evolution* 32: 2001–2014.

Yang, Y., M. J. Moore, S. F. Brockington, A. Timoneda, T. Feng, H. E. Marx, J. F. Walker, and S. A. Smith. 2017. An Efficient Field and Laboratory Workflow for Plant Phylotranscriptomic Projects. *Applications in Plant Sciences* 5.

Yang, Y., and S. A. Smith. 2013. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics* 14: 328.

Yang, Y., and S. A. Smith. 2014. Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics. *Molecular Biology and Evolution* 31: 3081–3092.

Yao, G., J.-J. J. Jin, H.-T. T. Li, J.-B. B. Yang, V. S. Mandala, M. Croley, R. Mostow, et al. 2019. Plastid phylogenomic insights into the evolution of Caryophyllales. *Molecular Phylogenetics and Evolution* 134: 74–86.

Zhang, C., E. Sayyari, and S. Mirarab. 2017. ASTRAL-III: Increased Scalability and Impacts of Contracting Low Support Branches. 53–75. Springer, Cham.

Conclusion

The extraordinary chromosome variation of *Drosera* has long intrigued scientists (Rosenberg, 1903; Kress, 1970; Kondo and Lavarack, 1984). In this dissertation I have examined whether *D.* subg. *Ergaleium* has a different rate of chromosome evolution than the rest of *Drosera*, phylogenetically inferred the parental species of the allopolyploid *D. anglica*, and evaluated polyploidy and reticulation on the backbone of *Drosera* while inferring the sectional relationships in the genus. As I investigated these questions, I also evaluated the pros and cons of different polyploid inference methods and found that multiple lines of evidence are needed to infer a polyploidy event.

In Chapter 1, I modeled the rate of polyploidy and single chromosome gains and losses across *Drosera* in a phylogenetic framework. While there was no evidence for different rates of polyploidy among *Drosera* subgenera*, D.* subg. *Ergaleium* had a significantly higher rate of single chromosome change than the other *Drosera* subgenera. Contrary to the expectation that single chromosome variation, like polyploidy, is more deleterious to and less common in self-incompatible species (Husband et al., 2013; Van Drunen and Husband, 2019), *D.* subg. *Ergaleium* also had a higher percentage of self-incompatible species than the rest of *Drosera*. In the ancestral state reconstruction analysis, I inferred eight polyploid events, four of which include one or more sections of *Drosera.*

Chapter 2 focused on the circumboreal allopolyploid species *D. anglica* that has been well studied cytologically. Only one other species, *D. rotundifolia,* has an equally wide range of distribution. Cytological pairing in artificial hybridization supported *D. rotundifolia* and *D. intermedia* as potential parents (Kondo and Segawa, 1988; Gervais and Gauthier, 1999). Through field collection and phylogenomic study using transcriptomic data, I found that *D. rotundifolia* and *D. linearis* are the paternal and maternal parents of *D. anglica*, respectively. European *D. anglica* appears more closely related to other North American *D. anglica* populations than the North American populations are related to each other. Interestingly, even in a recent polyploid like *D. anglica*, chromosome pairing in hybrids does not necessarily reflect parentage.

In Chapter 3, I sequenced transcriptomes from *Drosera* species across the genus and reconstructed the history of polyploidy and reticulate evolution among sections. I found evidence for six polyploidy events across *Drosera,* two of which were allopolyploid events resulting in the polyphyly of *D.* sect. *Drosera* and *Drosera*, respectively. Of the four polyploidy events inferred in Chapter 1 that included one or more *Drosera* sections, two were supported, and two were rejected by phylogenomic evidence. The two polyploid events that were rejected by phylogenomic data both occurred in *D.* subg. *Ergaleium,* suggesting that the rapid increases of chromosome number were due to chromosome fissions instead. This further supported the highly elevated rate of single chromosome evolution in *D.* subg. *Ergaleium*.

Across chapters, I found that both polyploidy and reticulate evolution have been recurring themes throughout *Drosera* though limited in *D.* subg. *Ergaleium*. Instead, single chromosome evolution was prevalent in *D.* subg. *Ergaleium.* Overall, different methods were better at detecting different types of chromosome evolution and inferring the history of polyploid events with different ages.

Modeling chromosome evolution in a phylogenetic framework was able to estimate rates of both polyploidy and single chromosome changes. However, in a lineage with highly elevated single chromosome evolution rates, the model over-estimated polyploid events. A second caveat of the current modeling approach is that it cannot take allopolyploidy into consideration.

In the recent polyploid, *Drosera anglica*, previous studies of its chromosome number, and chromosome pairing had led to the inference of its allopolyploid origin. Despite its relatively recent divergence from *D. linearis,* chromosome pairing in hybrids between *D. anglica* and *D. linearis* did not produce 10 univalent and 10 bivalent chromatids, suggesting chromosome rearrangements. Due to its relatively recent origin, neither a Ks peak nor an increased number of paralogs were detected from the HybPiper assembly. Instead, nucleotide diversity statistics, reference-based phasing, and haplotype phasing helped to detect and disentangle the subgenomes. In allopolyploids where parental lineages are not available for reference-based phasing and subgenome divergence is minimal, nucleotide diversity statistics and haplotype phasing provide means of detecting and beginning to explore the species origin.

For older polyploid lineages like those found along the backbone of *Drosera*, short-read assembly algorithms are able to correctly separate paralogs, but different tools are needed to detect the phylogenetic location of polyploid events and disentangle reticulation history. Modeling chromosome number evolution was able to correctly infer polyploidy events when the rate of single chromosome changes was relatively low. Nucleotide diversity statistics were informative for inferring recent polyploidy events where the Ks plot was ambiguous. Ks plots, on the other hand, were informative on which species shared more distant polyploid events, like that of *D.* subg. *Regiae* and *D.* subg. *Arcturia.* The combination of Ks plots and the most recent common ancestor of gene duplications mapped onto a species tree, pointed to the parental lineages of allopolyploid species when at least one parental lineage had been sampled. Cloudograms, in which gene trees were aligned at both root and tips, provide a useful visual validation for these inferred allopolyploid events.

In summary, the utility of any given tool will depend on the age and type of polyploidy event. Nucleotide diversity statistics and reference-based or haplotype phasing are all very useful for more recent polyploid events while Ks plots and mapping the MRCA of gene duplications are useful for events where the alleles are more diverged. Chromosome count modeling is useful for inferring polyploid events when single chromosome number change is low and also provides hypotheses of chromosome number change to explore with other data. Moving forward, the transcriptome data I generated in this dissertation has huge potential for further analyses for molecular evolution. These include overall Dn/Ds rates across the genome or in particular genes, especially those involved in centromere evolution between *D.* subg. *Ergaleium* and the other subgenera of *Drosera.* Because of the correlation between self-incompatibility and single chromosome number changes, future work should explore the presence and role of meiotic drive, either due to maternal drive or holocentric drive, in *D.* subg. *Ergaleium* versus *D.* subg. *Drosera.* Lastly, because of the global distribution and morphological diversity of *Drosera*, future research should include biogeography and morphological evolution in the context of polyploidy, chromosome rearrangement, and the reticulate history in this charismatic carnivorous plant lineage.

REFERNCES:

Gervais, C., and R. Gauthier. 1999. Etude cytotaxonomique des espèces et des hybrides naturels du genre *Drosera* (Droseraceae) au Québec. *Acta Botanica Gallica* 146: 387–401.

Husband, B. C., S. J. Baldwin, and J. Suda. 2013. The incidence of polyploidy in natural plant populations: major patterns and evolutionary processes. *In* I. J. Leitch, J. Greilhuber, J. Dolezel, and J. F. Wendel [eds.], Plant Genome Diversity, 255–276. Springer.

Kondo, K., and P. S. Lavarack. 1984. A cytotaxonomic study of some Australian species of *Drosera* Droseraceae. *Bot. J. Linn. Soc.* 88: 317–334.

Kondo, K., and M. Segawa. 1988. A cytotaxonomic study in artificial hybrids between *Drosera anglica* Huds. and its certain closely related species in series *Drosera*, section *Drosera*, subgenus *Drosera*, *Drosera*. *La Kromosomo II*: 1697–1709.

Kress, A. 1970. Zytotaxonomische Untersuchungen an einigen Insektenfängern (Droseraceae, Byblidaceae, Cephalotaceae, Roridulaceae, SarraceniaceaeDroseraceae, Insektenfängerneniaceae). *Berichte der Deutschen Botanischen Gesellschaft* 83: 55–62.

Rosenberg, O. 1903. Das Verhalten der Chromosomen in einer hybriden Pflanze. *Berichte der Deutschen Botanischen Gesellschaft* 21: 110–119.

Van Drunen, W. E., and B. C. Husband. 2019. Evolutionary associations between polyploidy, clonal reproduction, and perenniality in the angiosperms. *New Phytologist* 224: 1266–1277.