

UC Riverside

UC Riverside Electronic Theses and Dissertations

Title

Comparative Genomic and Transcriptomic Analyses in Plants Elucidate the Genetic Mechanisms Underlying of Diverse Evolutionary Processes

Permalink

<https://escholarship.org/uc/item/0hf7f7n1>

Author

Ostovar, Talieh

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
RIVERSIDE

AND

SAN DIEGO STATE UNIVERSITY

Comparative Genomic and Transcriptomic Analyses in Plants Elucidate the Genetic
Mechanisms Underlying of Diverse Evolutionary Processes

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Evolutionary Biology

by

Talieh Ostovar

September 2024

Dissertation Committee:

Dr. Elizabeth R. Waters, Co-Chairperson

Dr. Jason E. Stajich, Co-Chairperson

Dr. Amy Litt

Dr. Robert Zeller

Copyright by
Talieh Ostovar
2024

The Dissertation of Talieh Ostovar is approved:

Committee Co-Chairperson

Committee Co-Chairperson

University of California, Riverside
San Diego State University

Acknowledgement

Undertaking this PhD has been a transformative experience, made possible by the support and guidance of many individuals, to whom I am deeply grateful.

First and foremost, I extend my sincere gratitude to my PhD advisor, Dr. Elizabeth Waters, my co-advisors, Dr. Amy Litt, & Dr. Jason Stajich. I consider myself fortunate to have spent several years under their supervision, benefiting from their extensive personal and professional guidance, and their invaluable scientific knowledge. Their unwavering support and advice helped me navigate various challenges. Their leadership and continuous feedback were essential to the completion of this PhD.

I am grateful for the advice and support from my supervising committee members, Dr. Elizabeth Waters, Dr. Amy Litt, Dr. Jason Stajich, and Dr. Robert Zeller.

I am thankful to my lab-mates and fellow graduate students, including Cassie, Tania, Mark, Rudy, Sharon and others, for their feedback, cooperation, and friendship. Special thanks to my research collaborators Dr. Jacob Landis and Dr. Elizabeth McCarthy.

I would like to thank my father and mother, who supported me emotionally and financially throughout these challenging years. It was never possible to earn this PhD without the endless love, inspirations, and sacrifices that my parents made for me from my first day on earth until now. They advocated me through thick and thin during all the hardships of immigration and enabled me to pursue my dream.

I am forever thankful to Bahareh, my beloved sister, whose love and care have made me not to be complacent with life; the impossible becomes possible. I am truly obliged to her for all the love and encouragement she gave me during the tough times.

Without her endurance and unconditional love, I would not have had the courage to embark on this journey in the first place.

Additionally, I would like to thank Martin for his support, encouragement, and understanding throughout this journey. Your belief in me and your constant motivation have been invaluable, and I could not have accomplished this without you.

To my parents, Enayat and Maryam, my sister, Bahareh,
For their endless love & support

ABSTRACT OF THE DISSERTATION

Comparative Genomic and Transcriptomic Analyses in Plants Elucidate the Genetic Mechanisms Underlying of Diverse Evolutionary Processes

by

Talieh Ostovar

Doctor of Philosophy, Graduate Program in Evolutionary Biology
University of California, Riverside and San Diego State University, September 2024
Dr. Elizabeth R. Waters and Dr. Jason E. Stajich, Co-Chairpersons

This dissertation advances our knowledge of allopolyploidy, genomic adaptations, and heat stress responses in plants, focusing on *Nicotiana* section *Repandae* and two *Boechea* species, *Boechea arcuata* and *Boechea perennans*. Through comparative genomic and transcriptomic analyses, it elucidates the genetic mechanisms underlying these processes and their evolutionary implications. In *Nicotiana* section *Repandae*, I investigate global gene expression and subgenome dominance among three allotetraploid species derived from hybridization between *N. sylvestris* and *N. obtusifolia* about 4.5 million years ago. My findings reveal consistent subgenome dominance of *N. obtusifolia* across various developmental stages, impacting flower size and morphology. This dominance plays a significant role in shaping the genetic architecture and evolutionary trajectory of these species. In *Boechea*, I examine genomic adaptations to extreme environments, focusing on *B. arcuata* and *B. perennans*, which inhabit distinct climatic regions. Through comparative genomics and functional analysis, I identify unique and expanded gene families associated with environmental resilience. Gene Ontology enrichment reveals

significant terms related to photosynthesis and stress response, particularly in *B. perennans*. The results highlight the evolutionary processes enabling these species to thrive in diverse habitats. I further investigate gene expression profiles and evolutionary patterns of heat stress responses in *B. arcuata* and *B. perennans*. Using RNA sequencing data, I identify differentially expressed genes under control and basal heat stress conditions, with the top 50 upregulated genes predominantly being heat shock proteins. GO enrichment analysis reveals distinct metabolic and enzyme activity responses, indicating species-specific adaptations. Positive selection tests identify amino acid sites under evolutionary pressure, suggesting adaptive changes contributing to thermotolerance. By integrating comparative genomic and transcriptomic analyses, here I provide insights into the genetic and evolutionary mechanisms driving these processes in *Nicotiana* and *Boecheira* species, offering valuable information for future research and potential applications in improving crop resilience.

Table of Contents

Introduction	1
References.....	10
 Chapter 1: Differential gene expression and subgenome dominance in 4 million year old allopolyploids of <i>Nicotiana</i> section <i>Repandae</i>	
Abstract	17
Introduction	18
Materials and Methods.....	23
Results	32
References	64
 Chapter 2: Comparative genomics of two <i>Boecheira species</i> : identification of environmental adaptations	
Abstract	74
Introduction	75
Materials and Methods.....	80
Results	84
References	128
 Chapter 3: Gene expression profiles of the heat stress response in two <i>Boecheira</i> species reveals both conserved and divergent heat stress	
Abstract	135
Introduction	136
Materials and Methods.....	140
Results	143
References	186
 Conclusion	193
References.....	196

List of Figures

Chapter 1:

Figure 1. Floral morphology of <i>Nicotiana</i> section <i>Repandae</i> allopolyploids & diploid progenitors.....	22
Figure 2. Transcript expression profiles of <i>Repandae</i> are intermediate between the two progenitors.....	35
Figure 3. Hierarchical clustering of differentially expressed genes across the corolla developmental stages in each species.....	37
Figure 4. Homeolog bias and subgenome dominance in <i>Nicotiana</i> section <i>Repandae</i> allopolyploids.....	41
Figure 5. Homeolog expression bias across corolla developmentstages.....	44
Figure 6. Results of Gene Ontology (GO) enrichment analysis of upregulated genes in short-flowered species at 60% of anthesis length.....	46
Figure 7. Results of Gene Ontology (GO) enrichment analysis of upregulated genes in short-flowered species at 85% of anthesis length.....	47
Figure 8. Results of Gene Ontology (GO) enrichment analysis of upregulated genes in short-flowered species at 95% of anthesis length.....	48
Figure 9. Results of Gene Ontology (GO) enrichment analysis of upregulated genes in long-flowered species at 60% of anthesis length.	49
Figure 10. Results of Gene Ontology (GO) enrichment analysis of upregulated genes in long-flowered species at 85% of anthesis length.....	50
Figure 11. Results of Gene Ontology (GO) enrichment analysis of upregulated genes in long-flowered species at 95% of anthesis length.	51
Figure 12. Expression of known candidate genes involved in floral organ size across corolla development stages.....	54

Chapter 2:

Figure 1. Distribution of <i>Boechnera arcuata</i> and <i>Boechnera perennans</i>	79
Figure 2. Genomic features of <i>B. arcuata</i> and <i>B. perennans</i>	87
Figure 3. Synteny analysis of the <i>B. arcuata</i> and <i>B. perennans</i> genome compared to each other and <i>Arabidopsis thaliana</i>	89
Figure 4. Ks analysis identifies possible WGD events.....	91
Figure 5. Interspersed repeat landscape.....	94
Figure 6. Identification of orthologous groups shared across genomes and those that are unique to each genome.....	96
Figure 7. Results of Gene Ontology enrichment analysis of species-specific genes.....	107
Figure 8. Gene Ontology enrichment analysis of orthologous genes expanded in either <i>B. arcuata</i> or <i>B. perennans</i>	111
Figure 9. Number of genes expanded in either <i>B. arcuata</i> or <i>B. perennans</i> that are associated with photosynthesis and abiotic stress responses... ..	114
Figure 10. Gene families with expansion and/or contraction in the genomes of <i>B. arcuata</i> , <i>B. perennans</i> , and selected representative of Brassicaceae predicted by CAFÉ.....	115

Chapter 3:

Figure 1. Differential gene expression across control and basal heat stress (BHS) conditions....	144
Figure 2. Comparison of upregulated and downregulated genes in <i>B. arcuata</i> and <i>B. perennans</i> exposed to heat stress versus control conditions.	156
Figure 3. Gene Ontology (GO) enrichment analysis of upregulated and downregulated genes under basal heat stress conditions.	163
Figure 4. Shared GO terms enriched in upregulated and down regulated genes under heat stress conditions in <i>B. arcuata</i> and <i>B. perennans</i>	166
Figure 5. AT1g57630 alignment.	174
Figure 6. AT5G54160 alignment.....	175
Figure 7. AT5G55440 alignment.	176
Figure 8. AT1g57630 phylogenetic tree.....	177
Figure 9. AT5G54160 phylogenetic tree.....	178
Figure 10. AT5G55440 phylogenetic tree.....	179

List of Tables

Chapter 1:

Table 1. Number of differentially expressed genes within a taxon through three developmental stages.....	34
Table 2. Goodness of fit χ^2 test results to test for subgenome dominance	40

Chapter 2:

Table 1. Genome assembly and annotation statistics	86
Table 2. Repeat content for <i>B. arcuata</i> and <i>B. perennans</i> genomes	92
Table 3. Enriched GO terms for species-specific <i>B. arcuata</i> genes	100
Table 4. Enriched GO terms for genes that are species-specific in <i>B. perennans</i>	102
Table 5. Orthogroup gene expansions in <i>B. arcuata</i> for genes associated with photosynthesis and abiotic stress response.....	117
Table 6. Orthogroup gene expansions in <i>B. perennans</i> for genes associated with photosynthesis and abiotic stress response.....	119

Chapter 3:

Table 1. Differential gene expression analysis summary.....	143
Table 2. Top 50 upregulated genes in <i>B. arcuata</i> under basal heat stress condition.....	146
Table 3. Top 50 downregulated genes in <i>B. arcuata</i> under basal heat stress condition.....	148
Table 4. Top 50 upregulated genes in <i>B. perennans</i> under basal heat stress condition.....	152
Table 5. Top 50 downregulated genes in <i>B. perennans</i> under basal heat stress condition.....	154
Table 6. The list of genes that are upregulated and expanded in <i>B. arcuata</i>	168
Table 7. The list of genes that are upregulated and expanded in <i>B. perennans</i>	169
Table 8. Selection analysis of upregulated and expanded genes in <i>B. arcuata</i> and <i>B. perennans</i>	172

Introduction

The analysis of plant genomes provides invaluable insights into plant evolutionary history and the adaptive mechanisms that have enabled plants to inhabit a diverse range of climates. This rich history aids in understanding both past evolutionary events and predicting future trends in plant science, such as crop improvement, medicinal discovery, and conservation strategies (P. S. Soltis and Soltis 2021). The diversity of plant genomes is remarkable, spanning a wide range of sizes, compositions, and complexities. For example, genome sizes in angiosperms vary nearly 2,500-fold, from ~60 megabase pairs (Mb) in the carnivorous plant *Genlisea* to 149 gigabase pairs (Gb) in the monocot *Paris japonica* (Pellicer and Leitch 2020). Similarly, ferns and gymnosperms exhibit substantial genome sizes, with average genome sizes of ~12-14 Gb and ~18 Gb, respectively (Marchant et al. 2019; Pellicer et al. 2018). This variation is driven by several evolutionary processes, including the extensive proliferation of cycles of duplication events (De La Torre et al. 2014; Wendel 2015).

Gene duplication is a pivotal mechanism in the evolution of all genomes (Ohno 2013) contributing significantly to genetic diversity and complexity. Ancient duplication events and the high rate of retention of duplicate genes have led to an abundance of duplicate genes in plant genomes, which have played crucial roles in the development of novel functions and traits (Panchy, Lehti-Shiu, and Shiu 2016). Gene duplication occurs through several mechanisms, including whole-genome duplications (WGDs), tandem duplications, and transposon-mediated duplications (Ohta 1994; Long and Thornton 2001; Raes and Van de Peer 2003; Cui et al. 2006; Jiao et al. 2011) WGDs, which involve the

duplication of the entire genome, have been particularly influential, occurring multiple times throughout the evolutionary history of angiosperms and contributing to major innovations such as the development of new floral structures and stress responses (Lyons et al. 2008; D. E. Soltis et al. 2009). Tandem duplications result from unequal crossing-over events, leading to the duplication of genes within localized regions of the genome (J. Zhang 2003). Transposon-mediated duplications involve the movement of transposable elements, which can capture and duplicate gene sequences, further contributing to genetic diversity (Jiang et al. 2004). Gene duplication has had profound impacts on the evolution of plant genomes, leading to the development of novel functions and adaptive traits (Demuth and Hahn 2009); (Assis et al. 2023). Duplicated genes can diverge in function, contributing to increased complexity and the evolution of new traits. For example, the duplication of MADS-box genes has been instrumental in the diversification of floral structures, while the duplication of resistance genes has enhanced disease resistance in various plant species (Hu et al. 2015; S. Zhang et al. 2015).

Whole genome duplication (WGD), or polyploidy, plays a major role in the evolution of angiosperms (D. E. Soltis et al. 2016; Jiao et al. 2011; Stebbins 1985). The most important contribution of polyploidy to evolution is producing new genetic material for mutation, drift and selection to act upon which can pave the way for making new gene functions and consequently affecting genome evolution (Lynch and Conery 2000; Blanc and Wolfe 2004; Z. Li et al. 2016; Alix et al. 2017). Polyploids are divided into two different classes of auto- and allopolyploidy depending on their manner of formation as well as their chromosomal composition (Lloyd and Bomblies 2016). Autopolyploids have

arisen from failure of chromosomal segregation during gamete formation or from somatic doubling resulting in an individual, which has two complete sets of chromosomes, both of which from the same parental species. Allopolyploidy are those polyploids that have emerged through the processes of hybridization between different genome sets (Barker, Husband, and Pires 2016). The hybrid products may undergo meiosis resulting in chromosome doubling through which there are chances to create a genetically stable polyploid (D. E. Soltis, Visger, and Soltis 2014). However, the distinction of autopolyploids and allopolyploids is not absolute. In allopolyploids, the chromosome sets difference is proportional to the divergence of their progenitors' genomes (Paun et al. 2009). In other words, the less divergence of the chromosome sets of their parental genome, the more similar the resulting allopolyploid is to an autopolyploid. This ambiguity between two types of polyploidy leads a third class to be suggested, segmental allopolyploidy as an intermediate form between true autopolyploidy and allopolyploidy, but the utility of this category is still unanswered (Leal-Bertioli et al. 2018). Within the past few years, studies on polyploidy of angiosperms at the molecular level have undergone enormous progress due to the application of various genome and transcriptome sequencing data, novel methods for studying gene expression and using different bioinformatic tools (you need some citations here for example: (Fox et al. 2020; Van de Peer et al. 2021; Z. Li et al. 2016). These approaches have been used to improve our knowledge about the timing and frequency of polyploidy and to provide an interesting insight into the processes like duplicate gene retention and loss, gene silencing, subfunctionalization and neofunctionalization (Freeling, Scanlon, and Fowler 2015). Despite the big progress in the

polyploidy research area, there still are challenges of conflicting the research data on varied polyploidy species. These challenges are due to the lack of a single hypothesis to be able to illustrate the evolution of polyploidy in angiosperms (Z. Li et al. 2016). Moreover, in lots of species I still don't have sufficient information to evaluate the impact of polyploidy on fitness under particular environmental conditions or specific time (Sandve, Rohlfs, and Hvidsten 2018). Another important difficulty is the challenge of distinguishing between novelties that have arisen after or before polyploid formation, because sometimes exact diploid progenitors of allopolyploids, especially for ancient allopolyploids, cannot be studied directly (Steige and Slotte 2016). As I gain more molecular information from different plant systems and multiple synthetic polyploids and at particular times or environments, in various tissues of the plants or under varied genetic contexts, I will improve my knowledge about the genetic and epigenetic effects of polyploidy and improve more insight into gene expression modifications on a genomewide level following polyploidy (Adams and Wendel 2005).

Plant genomes are dynamic entities characterized by significant expansion and contraction over evolutionary time. Polyploidy is a key mechanism in plant genome expansions. Transposition is another major cause of genome expansion (Fedoroff 2012). Retrotransposons, which transpose via RNA intermediates, generate additional DNA by inserting reverse-transcribed elements into new genomic locations (Feschotte 2023). The accumulation of retrotransposon blocks between genes significantly contributes to genome size differences among plants. For instance, retrotransposon blocks occupy 74% of the recently sequenced 240-kilobase (kb) maize *Adh* region, highlighting their impact on

genome expansion (Tikhonov et al. 1999). While plant genomes primarily expand, there are mechanisms for genome contraction. The maize genome predominantly contains intact retrotransposons rather than solo long terminal repeats (LTRs), indicating limited unequal crossovers and deletion events. Despite this, mechanisms that eliminate sequences do exist. For instance, the maize transposons *Ac* and *Spm* frequently give rise to internally deleted elements, suggesting deletional decay of transposon sequences (Masson et al. 1987). Furthermore, the preferential loss of non-redundant sequences early after polyploidization has been detected in wheat, providing some evidence for genome contraction mechanisms (Feldman et al. 1997).

Molecular genetics, and high-throughput sequencing technology have greatly advanced our understanding of plant genomes (Wendel et al. 2016). Most major food and fiber crops now have publicly available genome assemblies, although their quality varies. This variation is due to the shift from map-based Sanger sequencing to second-generation, low-cost, short-read, whole-genome shotgun sequencing, which often results in 'gene space' assemblies. Sequencing large genomes or those with recent polyploidy is challenging due to the high sequence similarity among recently merged or doubled genomes, as seen in the complex allopolyploid genome of wheat (*Triticum aestivum*, ~15 Gbp), which still lacks a high-quality reference genome. Highly similar repetitive elements in these genomes are frequently omitted from whole-genome assemblies. This omission impacts not only genome completeness but also the understanding of epigenetic/chromatin remodeling pathways that influence gene expression or structure (Bennetzen and Wang 2014; Hollister et al. 2011). Third-generation, long-read sequencing technologies (Pacific

BioSciences, Oxford Nanopore) promise high-quality, gap-free genome sequences, crucial for comprehensively understanding genome structure and function (Levene et al. 2003; Howorka, Cheley, and Bayley 2001). In the next two to three years, many current assemblies are expected to be upgraded or replaced with these advanced technologies.

In the past two decades, gene expression profiling has revolutionized, driven by advancements in transcriptome sequencing and analysis technologies (Nejat, Ramalingam, and Mantri 2018). Traditional methods like northern blotting, in situ hybridization, and RT-PCR allowed the analysis of single or small groups of transcripts (Princiville et al. 2005; Weis et al. 1992). Real-time RT-PCR, a sensitive technique for detecting low-abundance mRNA, has been widely used for gene expression quantification (Deepak et al. 2007). The advent of microarrays in the mid-1990s transformed gene expression studies, enabling large-scale analysis of thousands of genes simultaneously (Schena et al. 1995). Microarrays have been extensively used to understand plant development and stress response mechanisms (Harmer et al. 2000; Mantri et al. 2010). More recently, next-generation sequencing (NGS) has brought another revolution in transcriptomics, enhancing our understanding of gene regulatory mechanisms and epigenetics (Mardis 2008). NGS-based RNA sequencing (RNA-seq) allows for the detection and quantification of known, novel, and rare transcripts, genome annotation, and non-coding RNA discovery, providing deeper insights into biological pathways and molecular mechanisms (Nejat et al. 2015; Mortazavi et al. 2008). Next-generation sequencing (NGS) has transformed transcriptomics, greatly enhancing our capacity to study the full array of RNA molecules in an organism or specific cell population. NGS technologies offer high-throughput and

cost-effective methods for profiling and analyzing RNA molecules, providing profound insights into gene expression, alternative splicing, non-coding RNA regulation, and various biological processes and diseases (Satam et al. 2023; J. Li and Liu 2019; Lucchinetti and Zaugg 2020; Choi, Kim, and Nam 2019). RNA-seq, a prominent NGS application in transcriptomics, involves sequencing and quantifying mRNA molecules. This method generates millions of short sequencing reads, enabling precise identification and quantification of gene expression levels. RNA-seq data can be analyzed to detect differential gene expression between conditions, discover novel transcripts, evaluate alternative splicing events, and study gene expression dynamics over time or across different tissues or cell types (Chen et al. 2023). NGS data also facilitate the reconstruction and annotation of an organism's transcriptome. Aligning RNA-seq reads to a reference genome or employing de novo assembly methods identifies novel transcripts, splice variants, untranslated regions, and other transcript features. This enhances our understanding of transcriptome complexity and improves the annotation of reference genomes, allowing the discovery of previously unknown genes and regulatory elements (Raghavan et al. 2022).

Plant genomics has revolutionized the study of plant adaptation to environmental conditions (Rellstab et al. 2020; Xu et al. 2020; Kress et al. 2022). The analysis of complete genomes combined with analysis of gene expression in light of organismal features such as growth under varying conditions has recently provided important insights into how plants can adapt to extreme conditions. It is noteworthy that many of these new exciting studies have been conducted with members of the Brassicaceae family. A study of the

genome of *Microthlaspi erraticum* has provided insight into adaptations to soils with high calcium concentrations. Other studies have examined the adaptation of Brassicaceae to high salt levels (Yuan et al. 2022; Dassanayake et al. 2011) and adaptation to the extreme cold of the arctic (Rellstab et al. 2020). The combination of high endemism and species ranges that are frequently limited to one climatic region makes the *Boechera* genus and excellent system with which to examine adaptation to climatic variables. (Rushworth et al. 2011; Stitt 2014; Windham and Al-Shehbaz 2006, 2007; Alexander et al. 2013; Hay et al. 2023). Previous studies of *Boechera arcuata* and *Boechera perennans* have shown that these species have differing species distributions and differing responses to heat-stress (Gallas and Waters 2015; Halter et al. 2017; Kannan et al. 2018; Pierce, Halter, and Waters 2020). Thus, they make an interesting pair of species with which to examine genomic adaptations to heat stress.

In this dissertation, the integration of comparative genomics and transcriptomics is employed to explore the complexities of plant evolution and adaptation. Comparative genomics identifies conserved and divergent genomic features across diverse plant species, elucidating evolutionary processes and genetic diversity. Transcriptomics complements this by analyzing gene expression patterns and regulatory mechanisms, revealing plant responses to environmental stresses and adaptations over time. Furthermore, the evolution of gene families and the role of polyploidy are examined. Polyploidy contributes significantly to genetic variation and the emergence of novel traits, playing a pivotal role in plant evolution. This dual approach provides a comprehensive understanding of plant genome dynamics, crucial for advancing crop improvement, stress resilience, uncovering

novel plant functions, and morphological variation across different plant species. By combining these methodologies, a deeper, more holistic understanding of plant biology is achieved, paving the way for innovative strategies in agriculture and conservation.

References

- Adams, Keith L., and Jonathan F. Wendel. 2005. "Polyploidy and Genome Evolution in Plants." *Current Opinion in Plant Biology* 8 (2): 135–41.
- Alexander, Patrick J., Michael D. Windham, James B. Beck, Ihsan A. Al-Shehbaz, Loreen Allphin, and C. Donovan Bailey. 2013. "Molecular Phylogenetics and Taxonomy of the Genus *Boechera* and Related Genera (Brassicaceae: Boechereae)." *Systematic Botany* 38 (1): 192–209.
- Alix, Karine, Pierre R. Gérard, Trude Schwarzacher, and J. S. Pat Heslop-Harrison. 2017. "Polyploidy and Interspecific Hybridization: Partners for Adaptation, Speciation and Evolution in Plants." *Annals of Botany* 120 (2): 183–94.
- Assis, Raquel, Gavin Conant, Barbara Holland, David A. Liberles, Malgorzata M. O'Reilly, and Amanda E. Wilson. 2023. "Models for the Retention of Duplicate Genes and Their Biological Underpinnings." *F1000Research* 12: 1400.
- Barker, Michael S., Brian C. Husband, and J. Chris Pires. 2016. "Spreading Winge and Flying High: The Evolutionary Importance of Polyploidy after a Century of Study." *American Journal of Botany* 103 (7): 1139–45.
- Bennetzen, Jeffrey L., and Hao Wang. 2014. "The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes." *Annual Review of Plant Biology* 65 (February): 505–30.4
- Blanc, Guillaume, and Kenneth H. Wolfe. 2004. "Functional Divergence of Duplicated Genes Formed by Polyploidy during Arabidopsis Evolution." *The Plant Cell* 16 (7): 1679–91.
- Chen, Jiung-Wen, Lisa Shrestha, George Green, André Leier, and Tatiana T. Marquez-Lago. 2023. "The Hitchhikers' Guide to RNA Sequencing and Functional Analysis." *Briefings in Bioinformatics* 24 (1). <https://doi.org/10.1093/bib/bbac529>.
- Choi, Seo-Won, Hyun-Woo Kim, and Jin-Wu Nam. 2019. "The Small Peptide World in Long Noncoding RNAs." *Briefings in Bioinformatics* 20 (5): 1853–64.
- Cui, Liying, P. Kerr Wall, James H. Leebens-Mack, Bruce G. Lindsay, Douglas E. Soltis, Jeff J. Doyle, Pamela S. Soltis, et al. 2006. "Widespread Genome Duplications throughout the History of Flowering Plants." *Genome Research* 16 (6): 738–49.'
- Dassanayake, Maheshi, Dong-Ha Oh, Jeffrey S. Haas, Alvaro Hernandez, Hyewon Hong, Shahjahan Ali, Dae-Jin Yun, et al. 2011. "The Genome of the Extremophile Crucifer *Thellungiella parvula*." *Nature Genetics* 43 (9): 913–18.

- Deepak, Sa, Kr Kottapalli, R. Rakwal, G. Oros, Ks Rangappa, H. Iwahashi, Y. Masuo, and Gk Agrawal. 2007. "Real-Time PCR: Revolutionizing Detection and Expression Analysis of Genes." *Current Genomics* 8 (4): 234–51.
- De La Torre, Amanda R., Inanc Birol, Jean Bousquet, Pär K. Ingvarsson, Stefan Jansson, Steven J. M. Jones, Christopher I. Keeling, et al. 2014. "Insights into Conifer Giga-Genomes." *Plant Physiology* 166 (4): 1724–32.
- Demuth, Jeffery P., and Matthew W. Hahn. 2009. "The Life and Death of Gene Families." *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 31 (1): 29–39.
- Fedoroff, Nina V. 2012. "Transposable Elements, Epigenetics, and Genome Evolution." *Science* 338 (6108): 758–67.
- Feldman, M., B. Liu, G. Segal, S. Abbo, A. A. Levy, and J. M. Vega. 1997. "Rapid Elimination of Low-Copy DNA Sequences in Polyploid Wheat: A Possible Mechanism for Differentiation of Homoeologous Chromosomes." *Genetics* 147 (3): 1381–87.
- Feschotte, Cédric. 2023. "Transposable Elements: McClintock's Legacy Revisited." *Nature Reviews. Genetics* 24 (11): 797–800.
- Fox, Donald T., Douglas E. Soltis, Pamela S. Soltis, Tia-Lynn Ashman, and Yves Van de Peer. 2020. "Polyploidy: A Biological Force From Cells to Ecosystems." *Trends in Cell Biology* 30 (9): 688–94.
- Freeling, Michael, Michael J. Scanlon, and John E. Fowler. 2015. "Fractionation and Subfunctionalization Following Genome Duplications: Mechanisms That Drive Gene Content and Their Consequences." *Current Opinion in Genetics & Development* 35 (December): 110–18.
- Gallas, Genna, and Elizabeth R. Waters. 2015. "Boechera Species Exhibit Species-Specific Responses to Combined Heat and High Light Stress." *PloS One* 10 (6): e0129041.
- Halter, Gillian, Nicole Simonetti, Cristy Suguitan, Kenneth Helm, Jessica Soroksky, and Elizabeth R. Waters. 2017. "Patterns of Thermotolerance, Chlorophyll Fluorescence, and Heat Shock Gene Expression Vary among four *Boechera* species and *Arabidopsis Thaliana*." *Botany* 95 (1): 9–27.
- Harmer, S. L., J. B. Hogenesch, M. Straume, H. S. Chang, B. Han, T. Zhu, X. Wang, J. A. Kreps, and S. A. Kay. 2000. "Orchestrated Transcription of Key Pathways in *Arabidopsis* by the Circadian Clock." *Science* 290 (5499): 2110–13.

- Hay, Nikolai M., Michael D. Windham, Terezie Mandáková, Martin A. Lysak, Kasper P. Hendriks, Klaus Mummenhoff, Frederic Lens, Kathleen M. Pryer, and C. Donovan Bailey. 2023. “A Hyb-Seq Phylogeny of *Boechera* and Related Genera Using a Combination of Angiosperms353 and Brassicaceae-Specific Bait Sets.” *American Journal of Botany* 110 (10): e16226.
- Hollister, Jesse D., Lisa M. Smith, Ya-Long Guo, Felix Ott, Detlef Weigel, and Brandon S. Gaut. 2011. “Transposable Elements and Small RNAs Contribute to Gene Expression Divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (6): 2322–27.
- Howorka, S., S. Cheley, and H. Bayley. 2001. “Sequence-Specific Detection of Individual DNA Strands Using Engineered Nanopores.” *Nature Biotechnology* 19 (7): 636–39.
- Hu, Yun, Wanqi Liang, Changsong Yin, Xuelian Yang, Baozhe Ping, Anxue Li, Ru Jia, et al. 2015. “Interactions of *OsMADS1* with Floral Homeotic Genes in Rice Flower Development.” *Molecular Plant* 8 (9): 1366–84.
- Jiang, Ning, Zhirong Bao, Xiaoyu Zhang, Sean R. Eddy, and Susan R. Wessler. 2004. “Pack-MULE Transposable Elements Mediate Gene Evolution in Plants.” *Nature* 431 (7008): 569–73.
- Jiao, Yuannian, Norman J. Wickett, Saravanaraj Ayyampalayam, André S. Chanderbali, Lena Landherr, Paula E. Ralph, Lynn P. Tomsho, et al. 2011. “Ancestral Polyploidy in Seed Plants and Angiosperms.” *Nature* 473 (7345): 97–100.
- Kannan, Sumetha, Gillian Halter, Tanya Renner, and Elizabeth R. Waters. 2018. “Patterns of Alternative Splicing Vary between Species during Heat Stress.” *AoB Plants* 10 (2): ly013.
- Kress, W. John, Douglas E. Soltis, Paul J. Kersey, Jill L. Wegrzyn, James H. Leebens-Mack, Morgan R. Gostel, Xin Liu, and Pamela S. Soltis. 2022. “Green Plant Genomes: What We Know in an Era of Rapidly Expanding Opportunities.” *Proceedings of the National Academy of Sciences of the United States of America* 119 (4). <https://doi.org/10.1073/pnas.2115640118>.
- Leal-Bertioli, Soraya C. M., Ignácio J. Godoy, João F. Santos, Jeff J. Doyle, Patrícia M. Guimarães, Brian L. Abernathy, Scott A. Jackson, Márcio C. Moretzsohn, and David J. Bertioli. 2018. “Segmental Allopolyploidy in Action: Increasing Diversity through Polyploid Hybridization and Homoeologous Recombination.” *American Journal of Botany* 105 (6): 1053–66.

- Levene, M. J., J. Korlach, S. W. Turner, M. Foquet, H. G. Craighead, and W. W. Webb. 2003. "Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations." *Science* 299 (5607): 682–86.
- Li, Jing, and Changning Liu. 2019. "Coding or Noncoding, the Converging Concepts of RNAs." *Frontiers in Genetics* 10 (May): 496.
- Li, Zhen, Jonas Defoort, Setareh Tasdighian, Steven Maere, Yves Van de Peer, and Riet De Smet. 2016. "Gene Duplicability of Core Genes Is Highly Consistent across All Angiosperms." *The Plant Cell* 28 (2): 326–44.
- Lloyd, Andrew, and Kirsten Bomblies. 2016. "Meiosis in Autopolyploid and Allopolyploid Arabidopsis." *Current Opinion in Plant Biology* 30 (April): 116–22.
- Long, Manyuan, and Kevin Thornton. 2001. "Gene Duplication and Evolution." *Science* 293 (5535): 1551–1551.
- Lucchinetti, E., and M. Zaugg. 2020. "RNA Sequencing: The Twenty-First Century Stethoscope in Pulmonary Medicine." *Anesthesiology*, August. <https://doi.org/10.1097/ALN.0000000000003524>.
- Lynch, M., and J. S. Conery. 2000. "The Evolutionary Fate and Consequences of Duplicate Genes." *Science*. American Association for the Advancement of Science (AAAS).
- Lyons, Eric, Brent Pedersen, Josh Kane, Maqsoodul Alam, Ray Ming, Haibao Tang, Xiyin Wang, et al. 2008. "Finding and Comparing Syntenic Regions among Arabidopsis and the Outgroups Papaya, Poplar, and Grape: CoGe with Rosids." *Plant Physiology* 148 (4): 1772–81.
- Mantri, Nitin L., Rebecca Ford, Tristan E. Coram, and Edwin C. K. Pang. 2010. "Evidence of Unique and Shared Responses to Major Biotic and Abiotic Stresses in Chickpea." *Environmental and Experimental Botany* 69 (3): 286–92.
- Marchant, D. Blaine, Emily B. Sessa, Paul G. Wolf, Kweon Heo, W. Brad Barbazuk, Pamela S. Soltis, and Douglas E. Soltis. 2019. "The C-Fern (*Ceratopteris richardii*) Genome: Insights into Plant Genome Evolution with the First Partial Homosporous Fern Genome Assembly." *Scientific Reports* 9 (1): 18181.
- Mardis, Elaine R. 2008. "Next-Generation DNA Sequencing Methods." *Annual Review of Genomics and Human Genetics* 9: 387–402.
- Masson, P., R. Surosky, J. A. Kingsbury, and N. V. Fedoroff. 1987. "Genetic and Molecular Analysis of the Spm-Dependent a-m2 Alleles of the Maize a Locus." *Genetics* 117 (1): 117–37.

- Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods* 5 (7): 621–28.
- Nejat, Naghmeh, David M. Cahill, Ganesan Vadamalai, Mark Ziemann, James Rookes, and Neda Naderali. 2015. "Transcriptomics-Based Analysis Using RNA-Seq of the Coconut (*Cocos Nucifera*) Leaf in Response to Yellow Decline Phytoplasma Infection." *Molecular Genetics and Genomics: MGG* 290 (5): 1899–1910.
- Nejat, Naghmeh, Abirami Ramalingam, and Nitin Mantri. 2018. "Advances in Transcriptomics of Plants." *Advances in Biochemical Engineering/biotechnology* 164: 161–85.
- Ohno, Susumu. 2013. *Evolution by Gene Duplication*. Springer Science & Business Media.
- Ohta, T. 1994. "Further Examples of Evolution by Gene Duplication Revealed through DNA Sequence Comparisons." *Genetics* 138 (4): 1331–37.
- Panchy, Nicholas, Melissa Lehti-Shiu, and Shin-Han Shiu. 2016. "Evolution of Gene Duplication in Plants." *Plant Physiology* 171 (4): 2294–2316.
- Paun, Ovidiu, Félix Forest, Michael F. Fay, and Mark W. Chase. 2009. "Hybrid Speciation in Angiosperms: Parental Divergence Drives Ploidy." *The New Phytologist* 182 (2): 507–18.
- Pellicer, Jaume, Oriane Hidalgo, Steven Dodsworth, and Ilia J. Leitch. 2018. "Genome Size Diversity and Its Impact on the Evolution of Land Plants." *Genes* 9 (2). <https://doi.org/10.3390/genes9020088>.
- Pellicer, Jaume, and Ilia J. Leitch. 2020. "An Updated Online Repository of Plant Genome Size Data for Comparative Studies." *The New Phytologist* 226 (2): 301–5.
- Pierce, Ian, Gillian Halter, and Elizabeth R. Waters. 2020. "Transcriptomic Analysis of the Heat-Stress Response in *Boechera Depauperata* and *Arabidopsis* Reveals a Distinct and Unusual Heat-Stress Response in *Boechera*." *Botany* 98 (10): 589–602.
- Princivalle, Alessandra P., Rachel M. C. Parker, Terri J. Dover, and Nicholas M. Barnes. 2005. "mRNA: Detection by in Situ and Northern Hybridization." *Methods in Molecular Biology* 306: 51–91.
- Raes, Jeroen, and Yves Van de Peer. 2003. "Gene Duplication, the Evolution of Novel Gene Functions, and Detecting Functional Divergence of Duplicates in Silico." *Applied Bioinformatics* 2 (2): 91–101.

- Raghavan, Venket, Louis Kraft, Fantin Mesny, and Linda Rigerte. 2022. "A Simple Guide to de Novo Transcriptome Assembly and Annotation." *Briefings in Bioinformatics* 23 (2). <https://doi.org/10.1093/bib/bbab563>.
- Rellstab, Christian, Stefan Zoller, Christian Sailer, Andrew Tedder, Felix Gugerli, Kentaro K. Shimizu, Rolf Holderegger, Alex Widmer, and Martin C. Fischer. 2020. "Genomic Signatures of Convergent Adaptation to Alpine Environments in Three Brassicaceae Species." *Molecular Ecology* 29 (22): 4350–65.
- Rushworth, Catherine A., Bao-Hua Song, Cheng-Ruei Lee, and Thomas Mitchell-Olds. 2011. "Boechera, a Model System for Ecological Genomics." *Molecular Ecology* 20 (23): 4843–57.
- Sandve, Simen R., Rori V. Rohlf, and Torgeir R. Hvidsten. 2018. "Subfunctionalization versus Neofunctionalization after Whole-Genome Duplication." *Nature Genetics*.
- Satam, Heena, Kandarp Joshi, Upasana Mangrolia, Sanobar Waghoo, Gulnaz Zaidi, Shravani Rawool, Ritesh P. Thakare, et al. 2023. "Next-Generation Sequencing Technology: Current Trends and Advancements." *Biology* 12 (7). <https://doi.org/10.3390/biology12070997>.
- Schena, M., D. Shalon, R. W. Davis, and P. O. Brown. 1995. "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray." *Science* 270 (5235): 467–70.
- Soltis, Douglas E., Victor A. Albert, Jim Leebens-Mack, Charles D. Bell, Andrew H. Paterson, Chunfang Zheng, David Sankoff, Claude W. Depamphilis, P. Kerr Wall, and Pamela S. Soltis. 2009. "Polyploidy and Angiosperm Diversification." *American Journal of Botany* 96 (1): 336–48.
- Soltis, Douglas E., Clayton J. Visger, D. Blaine Marchant, and Pamela S. Soltis. 2016. "Polyploidy: Pitfalls and Paths to a Paradigm." *American Journal of Botany* 103 (7): 1146–66.
- Soltis, Douglas E., Clayton J. Visger, and Pamela S. Soltis. 2014. "The Polyploidy Revolution Then...and Now: Stebbins Revisited." *American Journal of Botany* 101 (7): 1057–78.
- Soltis, Pamela S., and Douglas E. Soltis. 2021. "Plant Genomes: Markers of Evolutionary History and Drivers of Evolutionary Change." *Plants, People, Planet* 3 (1): 74–82.
- Stebbins, G. Ledyard. 1985. "Polyploidy, Hybridization, and the Invasion of New Habitats." *Annals of the Missouri Botanical Garden. Missouri Botanical Garden* 72 (4): 824–32.

- Steige, Kim A., and Tanja Slotte. 2016. "Genomic Legacies of the Progenitors and the Evolutionary Consequences of Allopolyploidy." *Current Opinion in Plant Biology* 30 (April): 88–93.
- Stitt, Mark. 2014. "Chill out with Rockcress: Quantitative Genetics of Frost Tolerance in the North American Wild Perennial *Boechera Stricta*." *Plant, Cell & Environment*.
- Tikhonov, A. P., P. J. SanMiguel, Y. Nakajima, N. M. Gorenstein, J. L. Bennetzen, and Z. Avramova. 1999. "Colinearity and Its Exceptions in Orthologous Adh Regions of Maize and Sorghum." *Proceedings of the National Academy of Sciences of the United States of America* 96 (13): 7409–14.
- Van de Peer, Yves, Tia-Lynn Ashman, Pamela S. Soltis, and Douglas E. Soltis. 2021. "Erratum to: Polyploidy: An Evolutionary and Ecological Force in Stressful Times." *The Plant Cell* 33 (8): 2899.
- Weis, J. H., S. S. Tan, B. K. Martin, and C. T. Wittwer. 1992. "Detection of Rare mRNAs via Quantitative RT-PCR." *Trends in Genetics: TIG* 8 (8): 263–64.
- Wendel, Jonathan F. 2015. "The Wondrous Cycles of Polyploidy in Plants." *American Journal of Botany* 102 (11): 1753–56.
- Wendel, Jonathan F., Scott A. Jackson, Blake C. Meyers, and Rod A. Wing. 2016. "Evolution of Plant Genome Architecture." *Genome Biology* 17 (March): 37.
- Windham, Michael D., and Ihsan A. Al-Shehbaz. 2006. "NEW AND NOTEWORTHY SPECIES OF BOECHERA (BRASSICACEAE) I: SEXUAL DIPLOIDS." *Harvard Papers in Botany* 11 (1): 61–88. *Harvard Papers in Botany* 11 (2): 257–74.
- Xu, Shaohua, Jiayan Wang, Zixiao Guo, Ziwen He, and Suhua Shi. 2020. "Genomic Convergence in the Adaptation to Extreme Environments." *Plant Communications* 1 (6): 100117.
- Yuan, Fang, Xi Wang, Boqing Zhao, Xiaojing Xu, Miao Shi, Bingying Leng, Xinxiu Dong, et al. 2022. "The Genome of the Recretohalophyte *Limonium Bicolor* Provides Insights into Salt Gland Development and Salinity Adaptation during Terrestrial Evolution." *Molecular Plant* 15 (6): 1024–44.
- Zhang, Jianzhi. 2003. "Evolution by Gene Duplication: An Update." *Trends in Ecology & Evolution* 18 (6): 292–98.
- Zhang, Shaohua, Ji-Si Zhang, Jing Zhao, and Chaoying He. 2015. "Distinct Subfunctionalization and Neofunctionalization of the B-Class MADS-Box Genes in *Physalis Florida*." *Planta* 241 (2): 387–402.

Chapter 1: Differential gene expression and subgenome dominance in 4 million year old allopolyploids of *Nicotiana* section *Repandae*

Abstract

Allopolyploidy, a phenomenon prevalent in angiosperms involving hybridization and whole-genome duplication, results in species with multiple subgenomes, altering genome structure and gene expression, leading to novel phenotypes. Allopolyploids often experience subgenome dominance, the preferential expression of homeologs from one of the two progenitor genomes (subgenomes). To explore the consequences of allopolyploidy and subgenome dominance, I investigate global gene expression architecture and the fate of homeologs in *Nicotiana* (Solanaceae). I focus on *Nicotiana* section *Repandae*, including three allotetraploid species, *N. nudicaulis*, *N. repanda*, and *N. stocktonii*, derived from hybridization between *N. sylvestris* and *N. obtusifolia* ~4.3 million years ago. I identify genes with differential expression across the three species and the progenitor species. Moreover, I examine subgenome dominance to trace expression bias in allopolyploids, and finally, investigate expression of candidate genes for flower size variation. The results show expression differences between the two progenitor species, with the allopolyploids intermediate between them, with a slight bias toward *N. obtusifolia*. Moreover, I demonstrate the consistent dominance of the *N. obtusifolia* subgenome across various developmental stages of the corolla in the allopolyploids, with a stronger bias in *N. nudicaulis*. I also see differential expression of several known flower size candidate genes across the corolla developmental stages, providing insights into the genetic mechanisms

driving flower morphology. The results highlight the role of subgenome dominance in shaping the genetic architecture and evolutionary trajectory of *Nicotiana* species and provide a foundation for future research into the ecological and evolutionary implications of allopolyploidy in flowering plants.

Introduction

Polyploidy is having more than two complete sets of chromosomes in an organism, fundamentally altering the genomic architecture (Van de Peer et al. 2017; Soltis et al. 2015). Polyploidy is a major driving force in the evolution of angiosperms, leading to new gene functions, through mechanisms like subfunctionalization and neofunctionalization, and increasing species diversification (Soltis et al. 2009; Landis et al. 2018). There are two fundamental forms of polyploidy: autopolyploidy and allopolyploidy (Grant 1981; Stebbins 1947). Autopolyploids are generated by whole genome duplication (WGD) within a species, whereas allopolyploids are produced by WGD events involving inter-specific hybridization (Stebbins 1947; Barker et al. 2016). Allopolyploids inherit two copies of every gene from both their progenitors. Thus, tracing the evolutionary trajectory and expression patterns of duplicated genes inherited from two different progenitor species over time is pivotal to understanding the immediate and long-term consequences of allopolyploidy (Buggs et al. 2014).

One important trend of allopolyploidy is that one of the progenitors' genomes (subgenomes) is often less fractionated and displays more gene retention (Cheng et al. 2018), a greater number of tandem gene duplications (Edger et al. 2019), and higher levels

of gene expression (Wendel et al. 2018) compared to the other subgenome. This genome is referred to as the dominant subgenome (Wang et al. 2022). The expression bias associated with subgenome dominance has been identified through genome-wide analyses in various plant species including in a synthetic allotetraploid of *Arabidopsis thaliana* and *Arabidopsis arenosa* (Wang et al. 2006), and in *Zea mays* (maize) (Schnable et al. 2011). Homeologous genes, which are orthologous genes from different progenitor subgenomes, display unequal expression levels that can differ among allopolyploid offspring from the same progenitors (Yoo et al. 2014). Moreover, the dominance of a subgenome may vary across tissues or developmental stages. In the tetraploid blueberry (*Vaccinium corymbosum*), one subgenome is dominant and has higher expression in many tissues at different developmental stages, including flowers at anthesis and post-fertilization, whereas the other subgenome is more highly expressed during fruit development (Colle et al. 2019). In *Gossypium hirsutum* (cotton), gene expression analysis in distinct organs such as petals and ovular tissue revealed different dominant subgenomes (Flagel et al. 2008).

The dominant subgenome in natural *Erythranthe* (monkeyflower, formerly *Mimulus*) (Edger et al. 2017) and *G. hirsutum* (Flagel et al. 2008) allopolyploids has been shown also to be dominant in synthesized allopolyploids. Natural allopolyploids of *E. peregrina* that are less than 140 years old, and a resynthesized allopolyploid, indicate that subgenome expression dominance was likely established from the initial F1 hybrid generation (Edger et al. 2017). This is based on significant expression bias observed in the resynthesized hybrid, which increased in magnitude over subsequent generations to the greatest bias in the natural allopolyploid (Edger et al. 2017; Flagel et al. 2008). The

dominance of one subgenome results in differences in trait heritability and phenotypic variation (Pfeifer et al. 2014), as well as adaptability and diversification among the allopolyploid species (Alger & Edger 2020; Ma et al. 2024). Furthermore, morphological differences, including flower organ size variation between allopolyploids and their progenitors, has been shown to be the product of subgenome expression bias in *Senecio mohavensis* (Mojave ragwort) (Alexander-Webber et al. 2016).

Floral size, specifically corolla size, plays a crucial role in plant reproduction by influencing visual attractiveness to pollinators and pollination efficiency (Solís-Montero & Vallejo-Marín 2017; Valadão-Mendes et al. 2022). While there is a proposed pathway regulating organ size (Krizek & Anderson 2013), the involvement of candidate genes associated with flower size has largely been studied only in model species (Du et al. 2014; Roxrud et al. 2007; Feng et al. 2011; Landis et al. 2020). In *A. thaliana* and other model organisms, GA-stimulated transcript (*GAST/GASA*) (Herzog et al. 1995), *LONGIFOLIA* (*LNG*) (Lee et al. 2006), *AUXIN-REGULATED GENE INVOLVED IN ORGAN SIZE* (*ARGOS*) (Hu et al. 2003) and *ARGOS-LIKE* (*ARL*) (Hu et al. 2006), have been identified as promoters or activators of elongation. Moreover, *BIG BROTHER* (*BB*) (Disch et al. 2006) and *DAI RELATED* (*DAR*) (Li et al. 2008; Peng et al. 2015) have been identified as repressors of growth or cell division. There has been an increase in floral transcriptomic studies over recent years; however, only a few have examined the genes related to differences in flower size in non-model species such as *Saltugilia* (Landis et al. 2017), *Brassica napus* (rapeseed) (Wang et al. 2017), and *Nicotiana tabacum* (Landis et al. 2020).

Thus, the role of these genes and their expression patterns in non-model species are poorly understood.

Nicotiana, a genus in the Solanaceae that includes 90 species (D'Andrea et al. 2023), is an excellent system for studying the evolutionary effects of allopolyploidy since approximately half of the species are allotetraploids that arose at different timepoints (Chase et al. 2003; Leitch et al. 2008; Clarkson et al. 2017). Although studies have provided valuable information about genomic and transcriptomic effects driven by allopolyploidy in *Nicotiana* (Renny-Byfield et al. 2012; Sierro et al. 2018; Landis et al. 2020; McCarthy et al. 2024), little is known about the long-term evolutionary consequences of allopolyploidy on genomes and on subgenome dominance in ancient species. *Nicotiana* section *Repandae* consists of four allopolyploid species: *Nicotiana nudicaulis*, *Nicotiana repanda*, *Nicotiana stocktonii*, and *Nicotiana nesophila*, derived from a maternal (*Nicotiana sylvestris*) and paternal (*Nicotiana obtusifolia*) diploid progenitor approximately 4.5 MYA, and the evolutionary relationships of the species are well known (Clarkson et al. 2004, 2005, 2017). *Nicotiana stocktonii*, *N. repanda*, and *N. nesophila* have similar floral morphology with long floral tubes (2.1-6.2 cm), similar to the maternal progenitor *N. sylvestris* (McCarthy et al. 2016). *Nicotiana nudicaulis* has divergent floral morphology with short corolla tubes (1.3-1.6 cm), similar to the paternal progenitor *N. obtusifolia* (McCarthy et al. 2016) (Figure 1.1). Overall, *Nicotiana* section *Repandae* offers a unique opportunity to explore how allopolyploidy impacts long-term evolutionary dynamics and morphological variation.

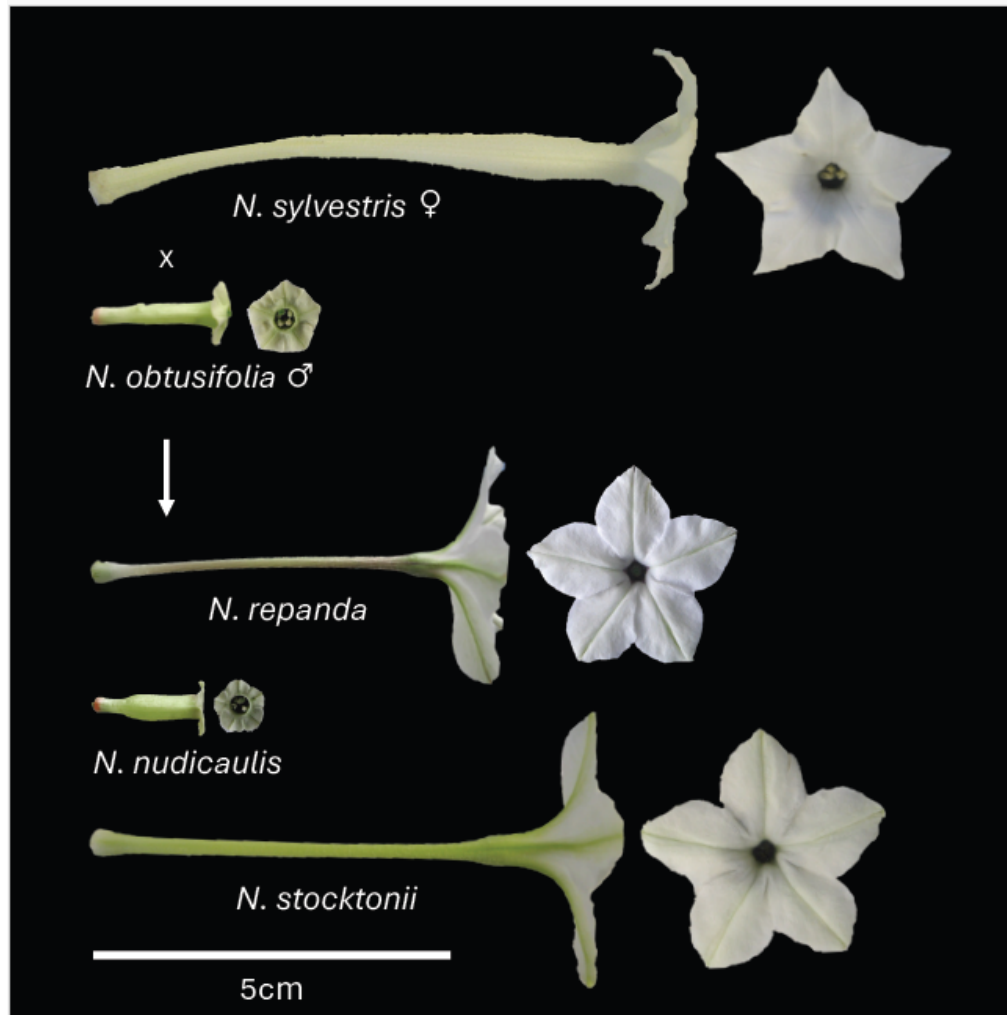


Figure 1.1. Floral morphology of *Nicotiana* section *Repandae* allopolyploids and their diploid progenitors. Maternal progenitor *N. sylvestris* and paternal progenitor *N. obtusifolia* produced allopolyploid offspring ~4 million years ago, which have since speciated into *N. nudicaulis*, *N. repanda*, and *N. stocktonii* allopolyploid species. *Nicotiana sylvestris*, *N. repanda*, and *N. stocktonii* have long-tubed flowers with broad floral limbs (the portion of the corolla that opens) whereas *N. obtusifolia* and *N. nudicaulis* have short-tubed flowers and smaller floral limbs. These photographs represent the species I investigated here, but some are from different accessions than were used in this study. The accessions in the photographs are *N. sylvestris* A04750326, *N. obtusifolia* TW143, *N. nudicaulis* 964750114, *N. repanda* TW110, and *N. stocktonii* TW126.

Here, I conduct comparative transcriptome analyses using *Nicotiana* section *Repandae* and their progenitor species to investigate subgenome dominance and associated changes in floral morphologies. Specifically, I examine 1) gene expression architecture across the allotetraploids and their progenitors at different corolla developmental stages and 2) the extent to which subgenome dominance is a legacy of the progenitors' gene expression. I further explore differential gene expression of known floral size candidate genes to enhance the understanding of how the expression of these specific genes is associated with the floral phenotypic differences in the tetraploid species.

Materials and Methods

Plant material and RNA isolation

Seeds for three polyploids were obtained from the USDA GRIN: *N. repanda* (TW110), *N. stocktonii* (TW86), and *N. nudicaulis* (TW90), as well as the two progenitor species, *N. sylvestris* (TW136) and *N. obtusifolia* (TW143). *N. nesophila* was not used because it is unclear whether it is a different species from *N. stocktonii* or not. All five species were grown in a plant growth room with twice a week watering, 16 hours daylight, and temperature of 22°C at the University of California, Riverside. I collected young emerging leaves from three biological replicates of all five species. To evaluate gene expression over the course of corolla development, I measured corolla length at anthesis including the floral limb, the part of the flower that opens, and calculated the mean from a minimum of five flowers per accession. I then collected dissected corolla tissue from three developmental stages (60%, 85%, and 95% of mean anthesis corolla length) from three

biological replicates for all taxa. Freshly collected tissues were flash frozen in liquid nitrogen and stored at -80°C to be used for RNA extraction. Frozen samples were ground using a micropestle attached to a cordless drill, and total RNA was extracted using E.Z.N.A. Plant RNA kit (Omega Bio-tek Inc., Norcross, Georgia, USA) following the manufacturer's protocol. The concentration of isolated RNA was measured using a Qubit 2.0 Fluorometer with the RNA BR assay kit (Invitrogen, Carlsbad, California, USA) and integrity was assessed with an Agilent 2100 Bioanalyzer Nano assay at the University of California, Riverside Institute for Integrative Genome Biology (IIGB) Genomics Core. High-quality samples with a minimum of 200 ng and RIN of at least 7 were shipped to Philip Morris Products S.A. (Neuchâtel, Switzerland) for library preparation and sequencing. mRNA sequencing libraries were prepared using the Universal Plus mRNA-Seq with NuQuant library preparation kit (Tecan Genomics, Redwood City, CA, USA) and sequenced on an Illumina Novaseq 6000 instrument (Illumina, San Diego, CA, USA) as 2 x 151 bp dual indexed paired-end reads.

Overall transcriptome analyses

Quality of generated raw reads was checked using FastQC v0.11.9. I used fastp v0.20.0 (Chen et al. 2018) to remove adaptor sequence and low quality reads with the following options: automatically detect adapters, minimum quality of 20, and minimum length of 75 bp. Raw reads of all the samples were submitted to the NCBI SRA database with accession numbers available in Table S1.

Initially I used a reference guided approach for analyzing differential gene expression by creating an *in silico* tetraploid genome reference to represent the hybrid genomic structure of an allopolyploid by concatenating the available genomes of the progenitor species, *N. sylvestris* (Sierro et al. 2013) and *N. obtusifolia* (<http://nadh.ice.mpg.de/NaDH/download/overview>). The annotation files corresponding to each progenitor genome were concatenated. The cleaned RNA-seq reads were mapped to the tetraploid reference genome using STAR v2.7.9a (Dobin et al. 2013). Following mapping, featureCounts (Liao et al. 2014) in the Subread v2.0.2 (Liao et al. 2013) package was used to quantify gene expression levels. Subsequent analyses revealed significant inconsistencies in mapping efficiency and expression quantification. Specifically, I observed skewed progenitor ratios heavily favoring *N. sylvestris*, along with disproportionately high median counts for *N. sylvestris* samples for *N. sylvestris* homeologs, suggesting underlying issues with the reference-genome approach. These discrepancies, highlighted by abnormal distributions of progenitor ratios and uniquely mapped read percentages across samples, indicated potential biases that would confound subgenome dominance analyses. Given these challenges, it became apparent that the current genomic references might not accurately reflect the transcriptomes of the specific accessions used in this study. Therefore, to ensure a more reliable and consistent mapping, I implemented a *de novo* assembly approach for the transcriptome data.

A reference transcriptome assembly for *N. obtusifolia* was generated with the newly sequenced Illumina data using Trinity v2.15 (Grabherr et al. 2011; Haas et al. 2013) with the following settings: jaccard clip, minimum contig length of 500 bp, and no salmon. The

N. sylvestris assembly was performed in the same way with the addition of previously generated Oxford Nanopore long-reads from Landis et al. (2020) using the `--long_read` option. A super transcript assembly was made from each *de novo* assembly using the `Trinity_gene_splice_modeler.py` script in the Trinity 2.15 package to collapse isoforms of the same gene into one contig. Quality of the two assemblies was determined using BUSCO v5.5.0 (Manni et al. 2021) using the embryophyta version 10 database. An *in silico* tetraploid transcriptome was generated similar to previous studies in *Nicotiana* (Landis et al. 2020; McCarthy et al. 2024) by merging high quality assemblies of the two progenitor species, *N. obtusifolia* and *N. sylvestris*. The *in silico* tetraploid transcriptome was annotated with Trinotate v3.2.2 (Bryant et al. 2017) and TransDecoder v5.7.0 (Haas et al. 2013) by searching the SwissProt database with BLASTx and BLASTp (Camacho et al. 2009).

Cleaned forward reads of the three polyploid species and the two progenitor species were mapped to the *in silico* transcriptome using bowtie2 v2.4.2 (Langmead et al. 2019) and quantified with RSEM v1.3.3 (Li & Dewey 2011) using the `align_and_estimate_abundance.pl` script in Trinity 2.15.0. Only the forward reads were used for mapping due to collapsing isoforms down to super transcripts, which could result in mismapping of some reads due to the forward and reverse reads being in opposite directions from each other. Differential gene expression across three flower developmental stages, as well as leaf samples, was conducted with DESeq2 1.4.2.0 (Love et al. 2014) after read counts were TMM normalized (McCarthy et al. 2012) using edgeR v3.42.2 (Robinson et al. 2010) with the `Trinity run_DE_analysis.pl` script. Genes were considered

differentially expressed using a Likelihood Ratio Test (LRT) with an adjusted p-value = 0.05 and a $|\log_2 \text{ fold change}| = 2$. General expression trends across all samples were visualized using variance-stabilized count data.

To investigate sequence divergence between single copy orthologs, OrthoFinder v2.5.4 (Emms & Kelly 2019), with the diamond v2.0.5 (Buchfink et al. 2021) and MAFFT v7.310 (Katoh & Standley 2013) options was run using the nucleotide SuperTrans files from both progenitors. The resulting 13,143 single copy orthologs were aligned with MAFFT v7.505 (Katoh & Standley 2013) using the automatic setting with 5000 iterations and adjusting directionality of the reads prior to calculating pairwise sequence divergence using mothur v1.47.0 (Schloss et al. 2009) using the dist.seq command with ignoring gaps and not penalizing gaps at the ends of the sequences. The mean sequence divergence of the single copy genes was 10.2%, which was used as the comparison for sequence divergence in candidate genes.

Subgenome dominance

I investigated subgenome dominance in *Nicotiana* section *Repandae* allopolyploids, using the single copy orthologs from the OrthoFinder results (see above) and TMM values at each developmental stage (60% corolla, 85% corolla, 95% corolla, and leaf) separately. I calculated homeolog bias for each sample by finding the proportion of *N. sylvestris* reads in the summed *N. sylvestris* and *N. obtusifolia* reads across a homeolog pair within a biological replicate; I called this measurement ‘proportion sylv.’ To ensure that I only included homeolog pairs with minimal off target mapping (fewer than 10% reads

mapped to the wrong subgenome), I used a threshold of > 0.9 proportion sylv for *N. sylvestris* samples and < 0.1 proportion sylv for *N. obtusifolia* samples. However, this threshold also eliminates homeolog pairs which are not expressed in *N. sylvestris*, but are expressed in *N. obtusifolia* because this scenario results in proportion sylv values of 0 for *N. sylvestris* samples. Because the allotetraploids may express homeologs not expressed in *N. sylvestris*, but expressed in *N. obtusifolia*, I adjusted methods to also retain homeolog pairs not expressed in *N. sylvestris*.

Each homeolog pair fit into one of four possible scenarios based on the median TMM values for *N. sylvestris* and *N. obtusifolia* samples. I altered the retention methods for each of these four scenarios to retain homeolog pairs where both progenitors expressed the homeolog and where only one progenitor expressed the homeolog (as long as there was minimal off target mapping). These four scenarios were calculated for each developmental stage separately, as a given homeolog pair may fall into a different scenario in a different developmental stage if its expression patterns changed. In Scenario 1, the median TMM value for both *N. sylvestris* and *N. obtusifolia* samples was zero (meaning that the homeolog was not expressed in either two or all three of the replicates). I eliminated all homeolog pairs in this scenario because neither homeolog was expressed in the progenitors, which prevented us from determining the degree of off target mapping. In Scenario 2, the median TMM value for both *N. sylvestris* and *N. obtusifolia* was greater than zero (meaning that the homeolog was not expressed in zero or only one of the replicates). I only retained homeolog pairs from this scenario where the samples that expressed the homeolog had proportion sylv values of > 0.9 for *N. sylvestris* and < 0.1 for *N. obtusifolia*. In Scenario 3,

the median TMM value for *N. sylvestris* was zero (meaning that the homeolog was not expressed in either two or all three replicates) and the median TMM value for *N. obtusifolia* was greater than zero (meaning that the homeolog was not expressed in either zero or only one of the replicates). I only retained homeolog pairs from this scenario where either none of the *N. sylvestris* samples expressed the homeolog or the one *N. sylvestris* sample that did express the homeolog had a proportion sylv value of > 0.9 and the *N. obtusifolia* samples that expressed the homeolog had proportion sylv values of < 0.1 . In Scenario 4, the median TMM value for *N. sylvestris* was greater than zero (meaning that the homeolog was not expressed in either zero or only one of the replicates) and the median TMM value for *N. obtusifolia* was zero (meaning that the homeolog was not expressed in either two or all three of the replicates). I only retained homeolog pairs from this scenario where the *N. sylvestris* samples that expressed the homeolog had proportion sylv values of > 0.9 and either none of the *N. obtusifolia* samples expressed the homeolog or the one *N. obtusifolia* sample that did express the homeolog had a proportion sylv value of < 0.1 .

I also calculated the progenitor ratio of expression as the proportion sylv for each homeolog pair by dividing the median TMM value from *N. sylvestris* samples on the *N. sylvestris* homeolog by the sum of the median TMM value from *N. sylvestris* samples on the *N. sylvestris* homeolog and the median TMM value from *N. obtusifolia* samples on the *N. obtusifolia* homeolog (median *N. sylvestris*/(median *N. sylvestris* + median *N. obtusifolia*)).

I evaluated subgenome dominance by using goodness of fit χ^2 tests in R v4.2.1 (R Core team 2021) to compare the number of *N. sylvestris*-biased and *N. obtusifolia*-biased

homeologs at each developmental stage (60% corolla, 85% corolla, 95% corolla, and leaf) for the progenitors and each allotetraploid species independently. I used two different estimates for equal homeolog bias: 0.45-0.55 proportion sylv, in which homeologs with < 0.45 proportion sylv were considered *N. obtusifolia*-biased whereas homeologs with > 0.55 proportion sylv were considered *N. sylvestris*-biased, and 0.4-0.6 proportion sylv, in which homeologs with < 0.4 proportion sylv were considered *N. obtusifolia*-biased and homeologs with > 0.6 proportion sylv were considered *N. sylvestris*-biased.

I plotted the proportion sylv of the progenitors and the homeologs for each allotetraploid species at each developmental stage (60% corolla, 85% corolla, 95% corolla, and leaf) using ggplot2 v3.3.6 (Wickham 2016). I also plotted the degree of subgenome dominance (*e.g.* 1.5x more *N. obtusifolia*-biased homeologs than *N. sylvestris*-biased homeologs) for the progenitors and each allotetraploid species at each developmental stage.

Hierarchical clustering of expression patterns across corolla development stages

To identify genes showing differential expression within each accession at 60%, 85%, and 95% of corolla length, I conducted a Likelihood Ratio Test (LRT) in DESeq2. From the LRT analysis, I selected significant differentially expressed genes based on an adjusted p-value of 0.05 and a $|\log_2 \text{fold change}| = 2$ as the cut-off values. Subsequently, I employed the ‘DEGreport’ R package v1.28.0 (<http://lpantano.github.io/DEGreport/>) to group genes with similar expression patterns based on the LRT results across different developmental stages of the corolla in each species.

Comparisons of gene expression in short- and long-flowered species

To compare gene expression across short- and long-flowered polyploid species, I focused on shared differentially expressed genes among species at 60%, 85%, and 95% of corolla length. Pairwise comparisons between *N. nudicaulis* with short flowers and *N. repanda* and *N. stocktonii* with long flowers were conducted, and I extracted shared differentially expressed genes using gplots (Warnes et al. 2015). GO enrichment was performed on the shared differentially expressed genes across short- vs long-flower comparisons to identify enriched GO terms, using the R package GOstats v2.58.0 (Falcon & Gentleman 2007).

Overall expression analysis of floral organ size candidate genes

To compare expression patterns of genes associated with flower size variation across the short- and long-flowered species, I identified known floral size candidate genes via the Trinotate annotation file. To ensure the comprehensiveness and accuracy of the gene annotations, I conducted a local BLAST v2.14.1 (Camacho et al. 2009) analysis with a minimum e-value 1e-5. I filtered the genes to include only single-copy orthologs in each progenitor. I calculated the median TMM values of the three replicates within each developmental stage for visualization. Plots were generated to represent expression patterns over the three corolla developmental stages of organ size genes of interest using ggplot2 v3.4.4 (Wickham 2016) in R.

Results

Transcriptome assemblies

The assembled *N. sylvestris* transcriptome, after collapsing isoforms to super transcripts, resulted in 61,877 contigs with an N50 of 2,578 bp. In terms of completeness as measured by BUSCO, the *N. sylvestris* assembly had 97.3% complete (1566 complete and single copy and five complete and duplicated), 1.9% (30) fragmented, and 0.8% (13) missing BUSCO sequences out of the 1,614 contigs comprising the embryophyta version 10 library. The assembled *N. obtusifolia* transcriptome resulted in 50,299 contigs with an N50 of 2,903 bp. In terms of completeness as measured by BUSCO, the *N. obtusifolia* assembly had 97.6% complete (1565 complete and single copy and 10 complete and duplicated), 1.8% (29) fragmented, and 0.6% (10) missing BUSCO sequences out of the 1,614 contigs comprising the embryophyta version 10 library. The resulting *in silico* tetraploid transcriptome contained 112,176 contigs. The overall mapping rate of the cleaned reads to the *in silico* reference transcriptome ranged from 28.04% to 86.62%, with 69.30% to 93.90% of the mapped reads being uniquely mapped (Table S1). All but one progenitor sample (Obtus7_leaf) had over 85% uniquely mapped reads (of the total reads that mapped at all), with *N. nudicaulis* averaging 77.71% of reads uniquely mapped, *N. repanda* having 81.47% uniquely mapped, and *N. stocktonii* with 82.27% uniquely mapped. In contrast, the reference guided approach using STAR resulted in the following uneven average mapping percentages across species: 29.52% for *N. obtusifolia*, 60.34%

for *N. sylvestris*, 30.81% for *N. nudicaulis*, 38.61% for *N. repanda*, and 43.69% for *N. stocktonii*.

Comparison of differentially expressed genes

I identified 33,290 differentially expressed genes across all taxa using an adjusted p-value = 0.05 and a $|\log_2 \text{fold change}| = 2$. There was a distinct division between the two progenitors, *N. obtusifolia* and *N. sylvestris*, with the allopolyploids being intermediate with a slight bias towards *N. obtusifolia*, particularly in the leaf tissue (Figure 1.2).

Across the three stages of corolla development (60%, 85%, and 95% of corolla length) in *N. obtusifolia*, I identified 406 significantly differentially expressed genes. In *N. sylvestris*, 3,683 genes were differentially expressed. In polyploids, I identified 327, 5452, and 6333 differentially expressed genes across the three developmental stages in *N. repanda*, *N. stocktonii*, and *N. nudicaulis*, respectively. The largest number of differentially expressed genes was found in comparisons between 60% and 95% in all the species except for *N. obtusifolia*, in which the largest number of differentially expressed genes occurred between 60% and 85%. I also observed a pattern in which the 85% and 95% stages were most similar in all the species except for *N. nudicaulis*, in which 60% and 85% were most similar. Within *N. obtusifolia*, there were 1107 differentially expressed genes between 60% and 85%, 237 differentially expressed genes between 60% and 95%, and 76 differentially expressed genes between 85% and 95%. Within *N. sylvestris*, there were 202 differentially expressed genes between 60% and 85%, 1225 differentially expressed genes between 60% and 95%, and 49 differentially expressed genes between 85% and 95%. Within *N.*

nudicaulis, I identified 231 differentially expressed genes between 60% and 85%, 3248 differentially expressed genes between 60% and 95%, and 2873 differentially expressed genes between 85% and 95%. Within *N. repanda*, there were 310 differentially expressed genes between 60% and 85%, 513 differentially expressed genes between 60% and 95%, and 95 differentially expressed genes between 85% and 95%. Within *N. stocktonii*, I identified 447 differentially expressed genes between 60% and 85%, 3009 differentially expressed genes between 60% and 95%, and 125 differentially expressed genes between 85% and 95% (Table 1).

Table 1.1. Number of differentially expressed genes within a taxon through the three developmental stages: 60%, 85%, and 95% of anthesis. The number of differentially expressed genes with a 4-fold change in expression and FDR less than 0.05 are reported, along with the total number of transcripts present in the given comparison.

Taxon	60% vs. 85%	60% vs. 95%	85% vs. 95%
<i>N. obtusifolia</i>	1107/44,764	237/45,032	76/44,356
<i>N. sylvestris</i>	202/40,512	1225/41,252	49/41,154
<i>N. nudicaulis</i>	231/43,228	3248/44,524	2873/43,227
<i>N. repanda</i>	310/37,426	513/37,496	95/38,109
<i>N. stocktonii</i>	447/41,221	3009/41,609	125/39,759

Hierarchical clustering of the significantly differentially expressed genes across the corolla developmental stages within each accession yielded four groups of differentially expressed genes. Within any accession, the genes in each group shared a common expression pattern, although the four patterns were not identical across all accessions. (Figure 1.3). I identified 265, 73, 65, and 3 genes respectively in *N. obtusifolia* and 2065, 1495, 65, and 58 genes in *N. sylvestris*. Performing the same analysis within each of the *Repandae* accessions, I identified 161, 107, 44 and 15 clustered genes in *N. repanda*, 3619, 2556, 106, and 52 clustered genes in *N. nudicaulis*, and 2721, 2551, 126, and 54 clustered genes in *N. stocktonii*. In *N. sylvestris* and the three polyploid species, group 1 genes showed an increasing linear expression pattern across the developmental stages with the highest expression in the 95% developmental stage. Group 2 genes displayed a linear decreasing pattern, with the lowest expression in the 95% stage. In *N. obtusifolia*, the observed expression patterns were different from those dominant in *N. sylvestris* and the *Nicotiana* section *Repandae* species. The patterns in the first two clusters were unique and non-linear, meaning that the 85% expression was not between 60% and 95%, with the highest expression in the 85% stage in group 1 and the lowest expression in the 85% stage in group 2. The patterns in groups 3 and 4 were non-linear in all taxa except for group 3 in *N. obtusifolia* which displayed a linear increasing pattern. In most of the non-linear groups, expression in the 85% stage was higher than 60% and 95%. This pattern was opposite in group 3 of *N. stocktonii* and group 4 in *N. nudicaulis* and *N. obtusifolia* with lowest expression in the 85% developmental stage.

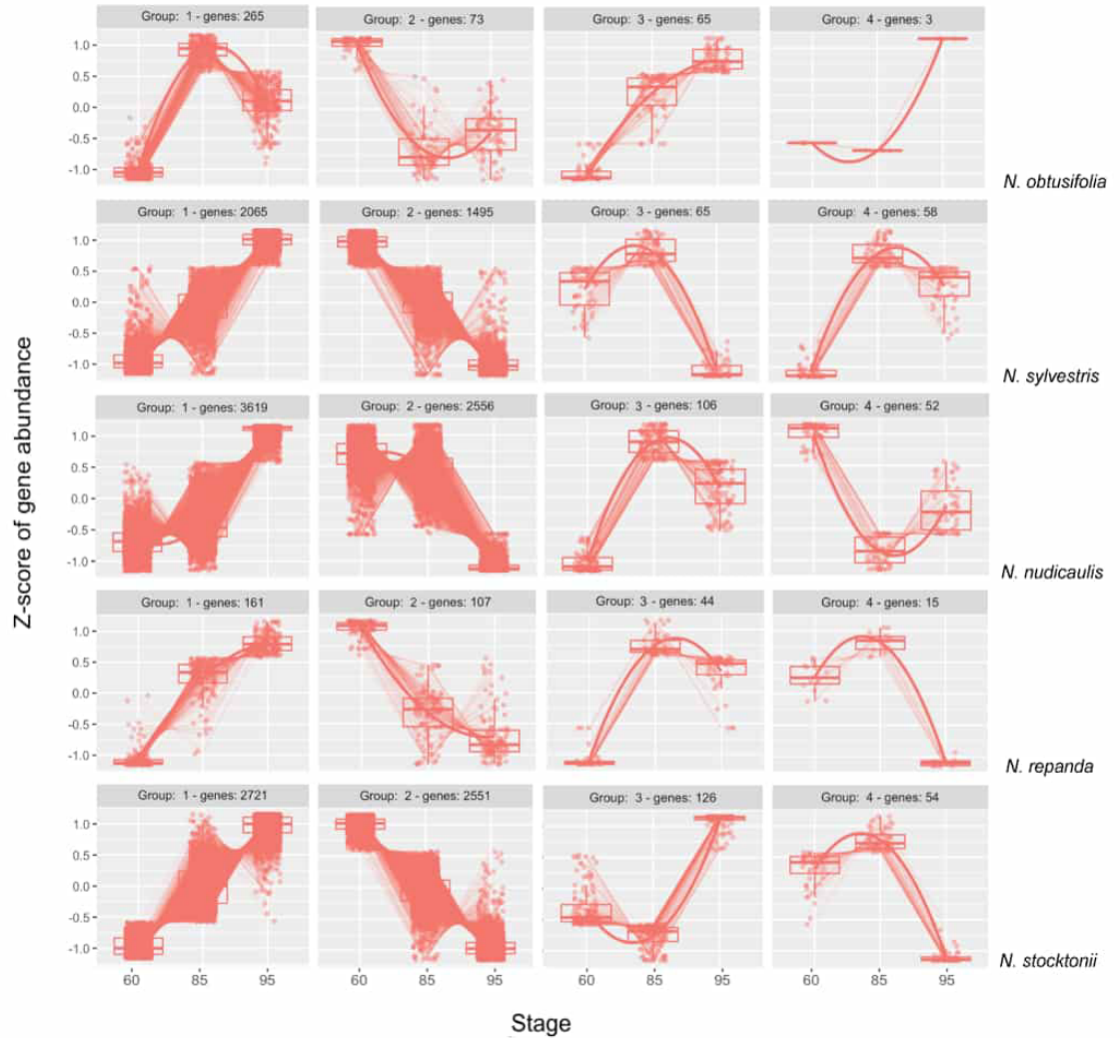


Figure 1.3. Hierarchical clustering of differentially expressed genes across the corolla developmental stages in each species. Box plots for four distinct groups of genes exhibiting similar expression patterns across 60%, 85%, and 95% of corolla developmental stages within each species and number of genes in each group. Each row in the figure represents the four groups identified in a specific species.

Subgenome dominance and homeolog bias

To test for subgenome dominance in *Nicotiana* section *Repandae* allopolyploids, I retained single copy orthologs when both *N. sylvestris* and *N. obtusifolia* expressed their respective homeologs, and when a gene was only expressed in one progenitor and not the other. All expressed homeologs had proportion sylv values of > 0.9 for *N. sylvestris* samples and < 0.1 for *N. obtusifolia* samples, showing that they had minimal off target mapping to the incorrect subgenome. I retained 7511 homeolog pairs for the 60% corolla dataset, 7652 homeolog pairs for the 85% corolla dataset, 7595 homeolog pairs for the 95% corolla dataset, and 7165 homeolog pairs for the leaf dataset. To estimate the extent of expression bias between the progenitor species, I calculated proportion sylv values across the retained homeolog pairs for the progenitors at each developmental time point. Goodness of fit χ^2 tests show that between the progenitors, there are more genes upregulated in *N. sylvestris* compared to genes upregulated in *N. obtusifolia* for both 60% corolla (1.12-1.18x) and leaf (1.32-1.38x) datasets at both 45-55 and 40-60 equal homeolog expression thresholds (Figure 1.4, Table 1.2). However, there are more genes upregulated in *N. obtusifolia* than upregulated in *N. sylvestris* for both 85% (1.53-1.58x) and 95% (1.46-1.47x) corolla datasets at both 45-55 and 40-60 equal homeolog expression thresholds (Figure 1.4, Table 1.2). The pattern of bias within progenitor expression provides a baseline to compare to subgenome dominance in the polyploid species. Within the allotetraploid species, there are more *N. obtusifolia*-biased homeologs than *N. sylvestris*-biased homeologs across all developmental stages and equal homeolog expression thresholds,

with *N. nudicaulis* having a stronger bias (1.35-1.43x) than either *N. repanda* (1.11-1.17x) or *N. stocktonii* (1.12-1.17x) allotetraploids (Figure 1.4, Table 1.2).

Table 1.2. Goodness of fit χ^2 test results to test for subgenome dominance, using different thresholds for equal homeolog expression. All χ^2 tests have d.f.=1.

Dev. time point	Equal homeolog threshold	Progenitor ratio	<i>N. nudicaulis</i> homeolog ratio	<i>N. repanda</i> homeolog ratio	<i>N. stocktonii</i> homeolog ratio
60	45-55	1.12x sylv-biased ($\chi^2=22.654$, $p=1.94 \times 10^{-6}$)	1.39x obtus-biased ($\chi^2=155.91$, $p<2.2 \times 10^{-16}$)	1.13x obtus-biased ($\chi^2=18.856$, $p=1.41 \times 10^{-5}$)	1.14x obtus-biased ($\chi^2=26.419$, $p=2.749 \times 10^{-7}$)
	40-60	1.18x sylv-biased ($\chi^2=41.396$, $p=1.243 \times 10^{-16}$)	1.42x obtus-biased ($\chi^2=160.17$, $p<2.2 \times 10^{-16}$)	1.11x obtus-biased ($\chi^2=12.393$, $p=4.309 \times 10^{-4}$)	1.15x obtus-biased ($\chi^2=25.874$, $p=3.644 \times 10^{-7}$)
85	45-55	1.53x obtus-biased ($\chi^2=293.97$, $p<2.2 \times 10^{-16}$)	1.36x obtus-biased ($\chi^2=136.35$, $p<2.2 \times 10^{-16}$)	1.17x obtus-biased ($\chi^2=36.789$, $p=1.316 \times 10^{-9}$)	1.16x obtus-biased ($\chi^2=32.768$, $p=1.038 \times 10^{-8}$)
	40-60	1.58x obtus-biased ($\chi^2=293.19$, $p<2.2 \times 10^{-16}$)	1.39x obtus-biased ($\chi^2=143.4$, $p<2.2 \times 10^{-16}$)	1.17x obtus-biased ($\chi^2=34.791$, $p=3.67 \times 10^{-9}$)	1.16x obtus-biased ($\chi^2=29.458$, $p=5.716 \times 10^{-8}$)
95	45-55	1.46x obtus-biased ($\chi^2=230.05$, $p<2.2 \times 10^{-16}$)	1.39x obtus-biased ($\chi^2=157$, $p<2.2 \times 10^{-16}$)	1.13x obtus-biased ($\chi^2=20.755$, $p=5.219 \times 10^{-6}$)	1.17x obtus-biased ($\chi^2=37.007$, $p=1.177 \times 10^{-9}$)
	40-60	1.47x obtus-biased ($\chi^2=214.4$, $p<2.2 \times 10^{-16}$)	1.43x obtus-biased ($\chi^2=164.43$, $p<2.2 \times 10^{-16}$)	1.15x obtus-biased ($\chi^2=25.695$, $p=3.999 \times 10^{-7}$)	1.17x obtus-biased ($\chi^2=33.246$, $p=8.121 \times 10^{-9}$)
leaf	45-55	1.32x sylv-biased ($\chi^2=125.48$, $p<2.2 \times 10^{-16}$)	1.35x obtus-biased ($\chi^2=113.15$, $p<2.2 \times 10^{-16}$)	1.13x obtus-biased ($\chi^2=18.216$, $p=1.972 \times 10^{-5}$)	1.13x obtus-biased ($\chi^2=20.564$, $p=5.767 \times 10^{-6}$)
	40-60	1.38x sylv-biased ($\chi^2=151.3$, $p<2.2 \times 10^{-16}$)	1.38x obtus-biased ($\chi^2=115.92$, $p<2.2 \times 10^{-16}$)	1.14x obtus-biased ($\chi^2=18.08$, $p=2.118 \times 10^{-5}$)	1.12x obtus-biased ($\chi^2=16.865$, $p=4.013 \times 10^{-5}$)

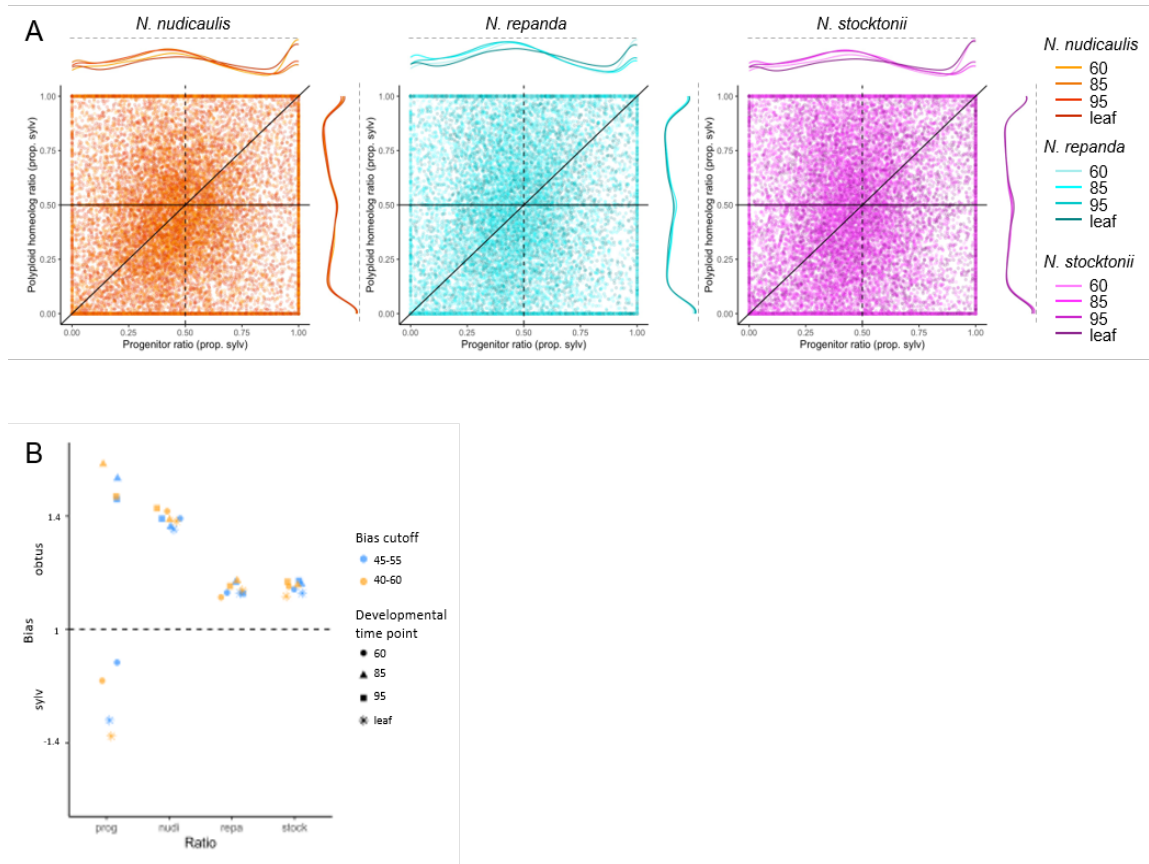


Figure 1.4. Homeolog bias and subgenome dominance in *Nicotiana* section *Repandae* allopolyploids. (A) Progenitor expression ratio plotted against polyploid homeolog ratio in each allopolyploid species. Included transcripts have minimal mismatching to the incorrect progenitor genome (>0.9 proportion sylv for *N. sylvestris* and <0.1 proportion sylv for *N. obtusifolia*). Each developmental time point (60%, 85%, and 95% corolla and leaf) was plotted in a different color. Curves at the top and right of the plots represent the distribution of points in each developmental time point for the progenitor ratio and polyploid homeolog ratio, respectively. Gray dashed lines above the histogram curves provide a reference to better observe changes in height across the curves. (B) Progenitor ratio and subgenome dominance results at different development time points. Bias values of 1 represent equal homeolog expression whereas positive bias values show a bias toward *N. obtusifolia* and negative bias values show a bias toward *N. sylvestris*. For example, 1.4 means that there are 1.4 times more *N. obtusifolia*-biased homeologs than *N. sylvestris*-biased homeologs. Colors represent different thresholds for calculating bias. For 45-55 cutoff, transcripts were counted as *N. sylvestris*-biased if the proportion sylv was above 0.55 and *N. obtusifolia*-biased if the proportion sylv was below 0.45. For the 40-60 cutoff, transcripts were counted as *N. sylvestris*-biased if the proportion sylv was above 0.6 and *N. obtusifolia*-biased if the proportion sylv was below 0.4. Both cutoffs yielded similar results. prog=progenitors, nud= *N. nudicaulis*, repa= *N. repanda*, stock= *N. Stocktonii*.

Subgenome dominance of the floral size candidate genes was not explicitly tested because nine of the thirteen single copy candidate genes did not meet the requirement of less than 10% of reads mapping to the other homeolog. Sequence divergence in the progenitor gene sequences showed that the average sequence divergence was 7.8% while the average sequence divergence of 13,143 single copy orthologs was 10.2%. The lowest sequence divergence was found in *LNG1* which showed 3.1% divergence, while *ARGOS* showed the highest at 15.9% (one of two out of 13 genes that showed sequence divergence greater than the average). Although not all homeologs with high sequence divergence met the read mapping criteria, the four genes that did tended to have higher sequence divergence than the average for floral size genes; these included the two copies of *LNG2* (divergence of 7.3% and 8.7%), *ARL* (9.7%), and *GASA9* (10.3%).

While I could not assess subgenome dominance on only four candidate genes, I did calculate homeolog bias for these four genes, measured as proportion *sylv* (Figure 1.5). Each gene, all of which likely promote cell expansion, displayed a different homeolog bias pattern. In *ARL*, all polyploid species were strongly biased toward *N. sylvestris* (Figure 1.5A). In *GASA9*, *N. nudicaulis* shows a very strong bias toward *N. sylvestris* whereas *N. repanda* and *N. stocktonii* exhibit homeolog expression levels similar to the progenitor ratio, which is biased toward *N. obtusifolia* (Figure 1.5B). In *LNG2*, all polyploid species are similar to the progenitor ratio in at least one developmental time point, but *N. nudicaulis* has equal homeolog expression whereas *N. repanda* and *N. stocktonii* are biased toward *N. obtusifolia* (Figure 1.5C). In *LNG2-2*, all polyploid species are similar to the progenitor ratio, which is biased toward *N. obtusifolia* (Figure 1.5D). The observed differences in the

direction of homeolog bias in *GASA9* and *LNG2* in related allopolyploid species are intriguing, but do not seem to correlate to the flower length differences observed in these species. Short-flowered *N. nudicaulis* is biased toward *N. sylvestris* or equal homeolog expression, whereas long-flowered *N. repanda* and *N. stocktonii* are biased toward *N. obtusifolia*, which is opposite what I might expect based on the flower length of the diploid progenitors.

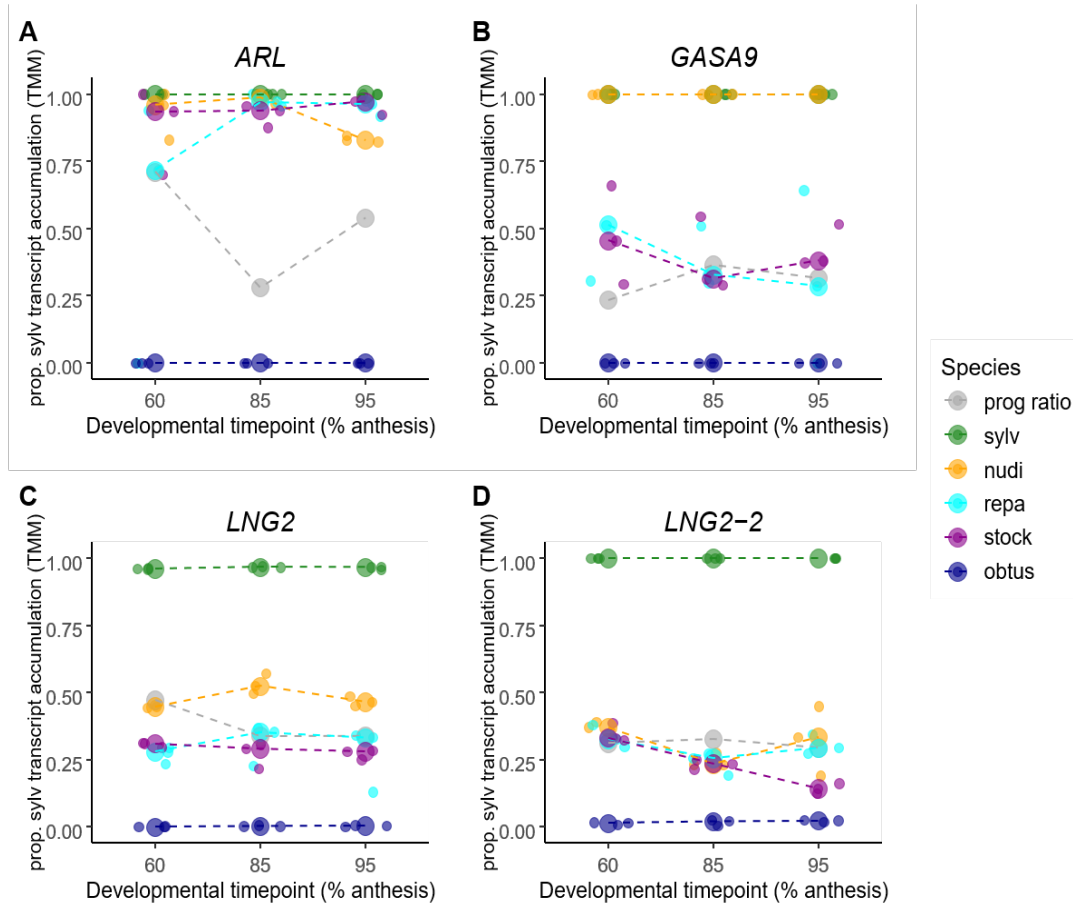


Figure 1.5. Homeolog expression bias across corolla development stages in candidate genes involved in floral organ size that met the criteria for minimal off target mapping to the incorrect progenitor genome. Proportion sylv is plotted for the *Nicotiana* section *Repandae* species and their diploid progenitors as well as the progenitor ratio in flower size genes (A) *ARL*, (B) *GASA9*, (C) *LNG2*, and (D) *LNG2-2*, which all promote cell elongation. Large points represent median proportion sylv values and are connected by dashed lines. Small points represent proportion sylv values for each biological replicate. prog ratio=progenitor ratio, nudi=*N. nudicaulis*, obtus=*N. obtusifolia*, repa=*N. repanda*, stock=*N. stocktonii*, sylv=*N. sylvestris*.

Pairwise transcriptome comparisons in short- and long-flowered species

Comparisons of short- and long-flowered species within *Nicotiana* section *Repandae* identified shared differentially expressed genes across all comparisons. At 60% of corolla anthesis size, I found 3374 up-regulated genes in short-flowered *N. nudicaulis* compared to both long-flowered species, *N. repanda* and *N. stocktonii*, and 3801 and 3103 at the 85% and 95% stages. I found 969, 2034, and 1405 up-regulated shared genes in long-flowered *N. repanda* and *N. stocktonii* at 60%, 85% and 95% stages respectively, compared to short-flowered *N. nudicaulis*.

For genes upregulated in short-flowered *N. nudicaulis*, GO enrichment analysis at the 60% stage revealed 86 GO terms related to biological processes (BP), 18 cellular components (CC), and 50 molecular functions (MF) (Table S2). The top 20 significantly enriched GO terms upregulated in *N. nudicaulis* at the 60% stage are found in Figure 1.6. At the 85% stage, 168 GO terms from BP, 18 from CC, and 61 from MF were found; the top 20 are found in Figure 1.7. At the 95% stage, I identified 148 GO terms from BP, 19 from CC, and 72 from MF; the top 20 are found in Figure 1.8. The most significantly enriched GO terms were related to photosynthesis, and this was consistent across all the developmental stages. For instance, I found photosynthesis electron transport chain, photosynthetic light reaction, chloroplast thylakoid, photosystem, and chlorophyll binding GO terms. GO enrichment results of upregulated genes in long-flowered species, *N. repanda* and *N. stocktonii*, demonstrated 120 GO terms from BP, 12 from CC, and 62 from MF at 60%; the top 20 are found in Figure 1.9. Moreover, 202 GO terms from BP, 12 from

CC, and 97 from MF were found at 85%; the top 20 are found in Figure 1.10. At 95%, I identified GO terms from 148 BP, 10 from CC, and 80 from MF; the top 20 are found in Figure 1.11. The most significantly enriched GO terms were related to response to stress and response to stimulus at the 60% stage. At 85% and 95%, the GO terms were related to various metabolic and biosynthetic processes such as phenylpropanoid metabolic and biosynthetic process, lignin metabolic and biosynthetic process, carbohydrate metabolic process, and polysaccharide metabolic process.



Figure 1.6. Results of Gene Ontology (GO) enrichment analysis of upregulated genes in short-flowered species at 60% of anthesis length. The y-axis lists enriched GO terms and the x-axis and the bubble size indicate the number of genes associated with a GO term. Bubbles are colored by the adjusted p-value < 0.01. The analysis was performed at each of three GO classes: biological processes (BP), molecular functions (MF), and cellular components (CC). Each GO class (BP, CC, MF) is represented by a different color (pink, yellow and purple, respectively).

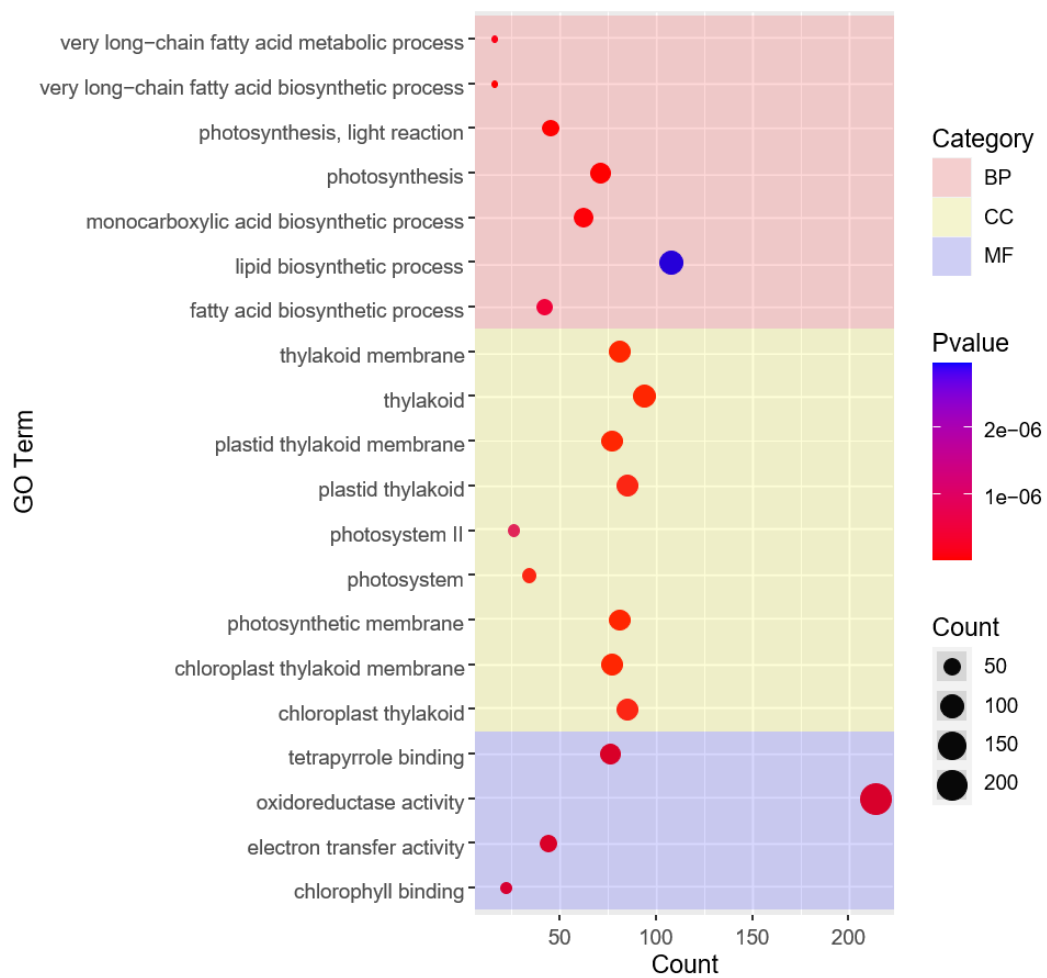


Figure 1.7. Results of Gene Ontology (GO) enrichment analysis of upregulated genes in short-flowered species at 85% of anthesis length. The y-axis lists enriched GO terms and the x-axis and the bubble size indicate the number of genes associated with a GO term. Bubbles are colored by the adjusted p-value < 0.01. The analysis was performed at each of three GO classes: biological processes (BP), molecular functions (MF), and cellular components (CC). Each GO class (BP, CC, MF) is represented by a different color (pink, yellow and purple, respectively).

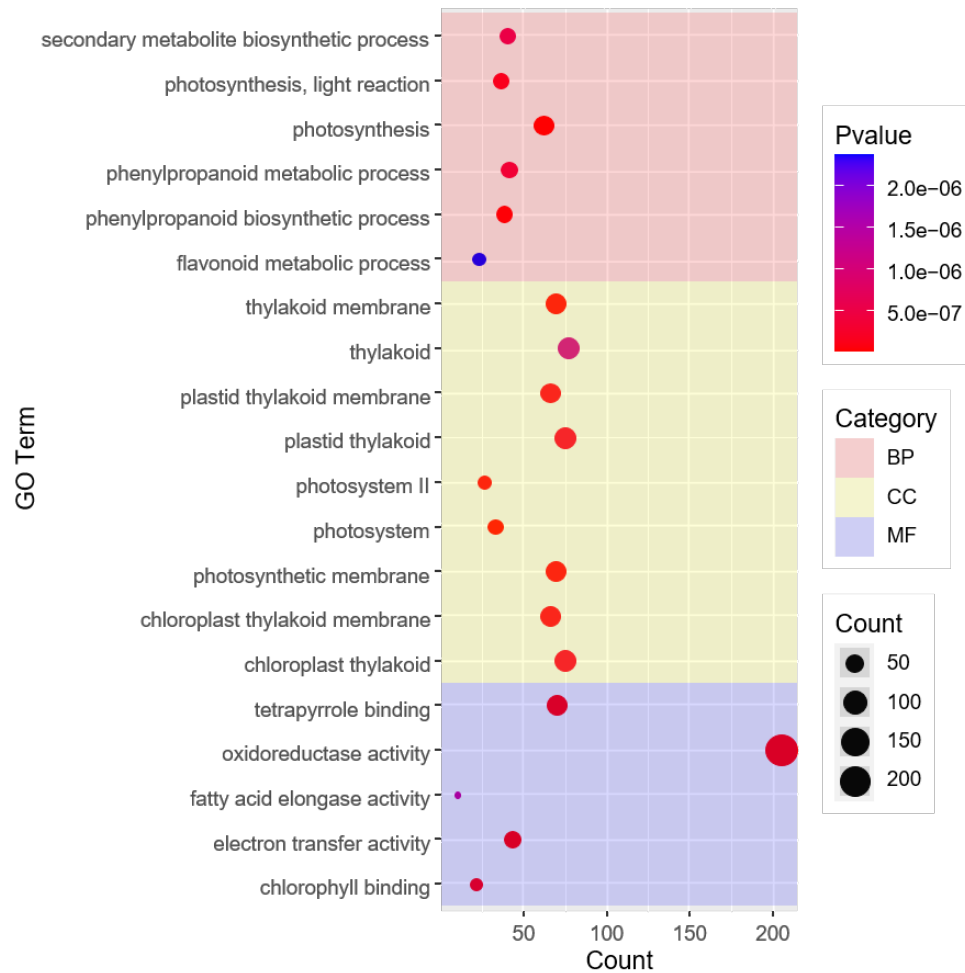


Figure 1.8. Results of Gene Ontology (GO) enrichment analysis of upregulated genes in short-flowered species at 95% of anthesis length. The y-axis lists enriched GO terms and the x-axis and the bubble size indicate the number of genes associated with a GO term. Bubbles are colored by the adjusted p-value < 0.01. The analysis was performed at each of three GO classes: biological processes (BP), molecular functions (MF), and cellular components (CC). Each GO class (BP, CC, MF) is represented by a different color (pink, yellow and purple, respectively).

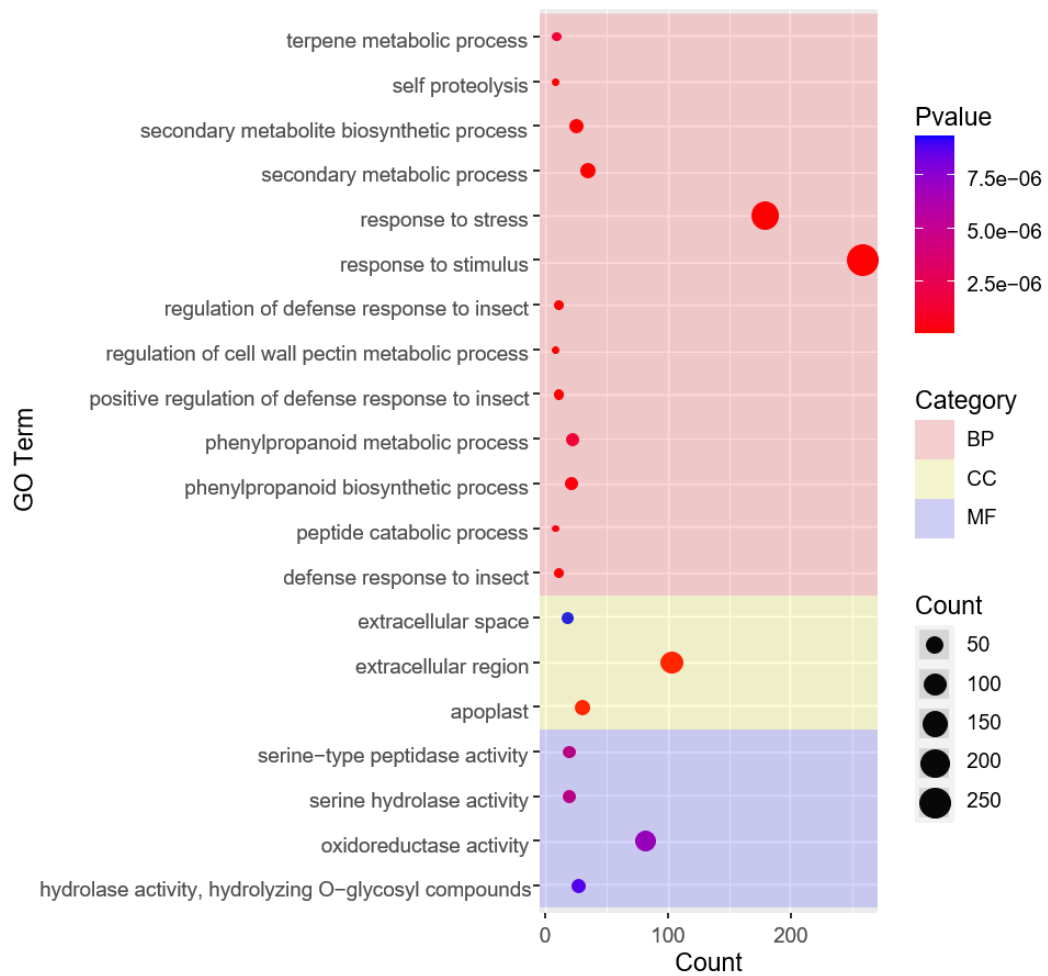


Figure 1.9. Results of Gene Ontology (GO) enrichment analysis of upregulated genes in long-flowered species at 60% of anthesis length. The y-axis lists enriched GO terms and the x-axis and the bubble size indicate the number of genes associated with a GO term. Bubbles are colored by the adjusted p-value < 0.01. The analysis was performed at each of three GO classes: biological processes (BP), molecular functions (MF), and cellular components (CC). Each GO class (BP, CC, MF) is represented by a different color (pink, yellow and purple, respectively).

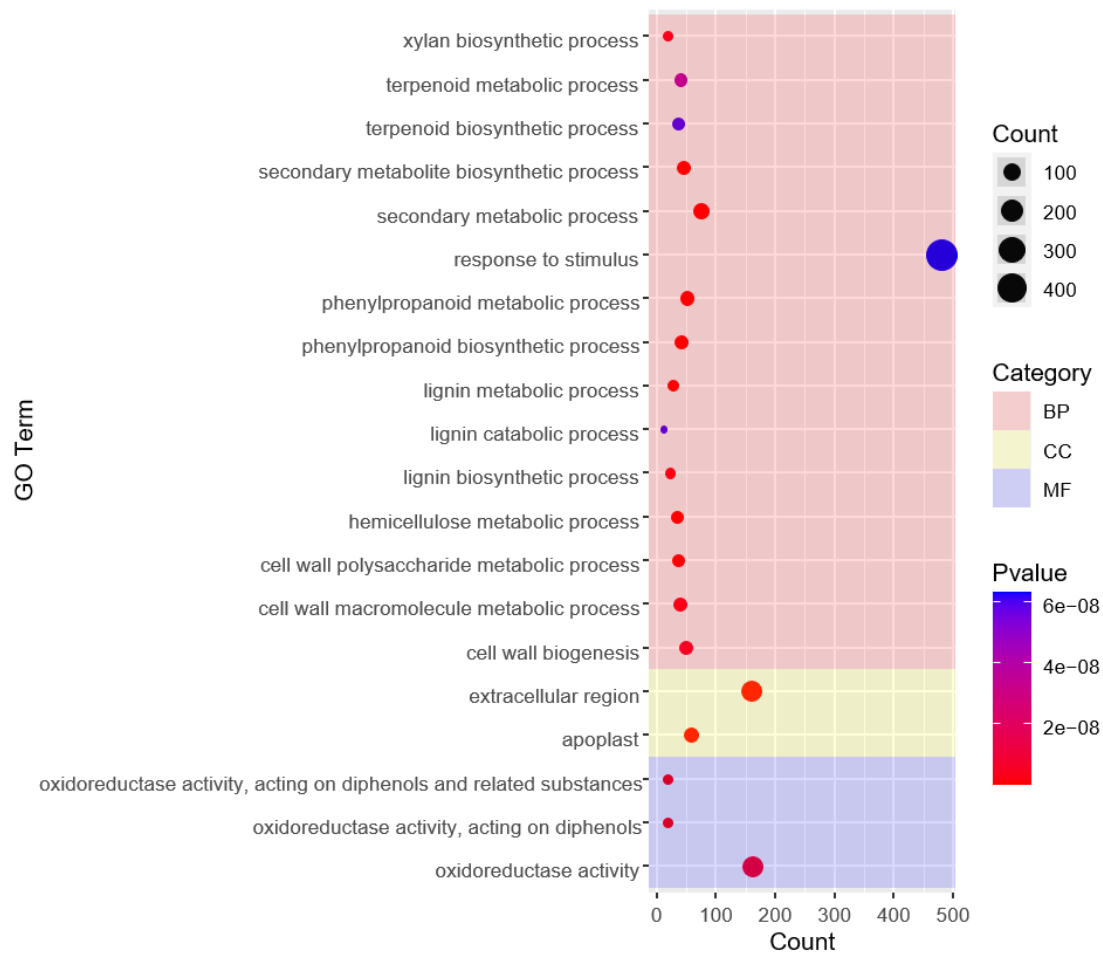


Figure 1.10. Results of Gene Ontology (GO) enrichment analysis of upregulated genes in long-flowered species at 85% of anthesis length. The y-axis lists enriched GO terms and the x-axis and the bubble size indicate the number of genes associated with a GO term. Bubbles are colored by the adjusted p-value < 0.01 . The analysis was performed at each of three GO classes: biological processes (BP), molecular functions (MF), and cellular components (CC). Each GO class (BP, CC, MF) is represented by a different color (pink, yellow and purple, respectively).

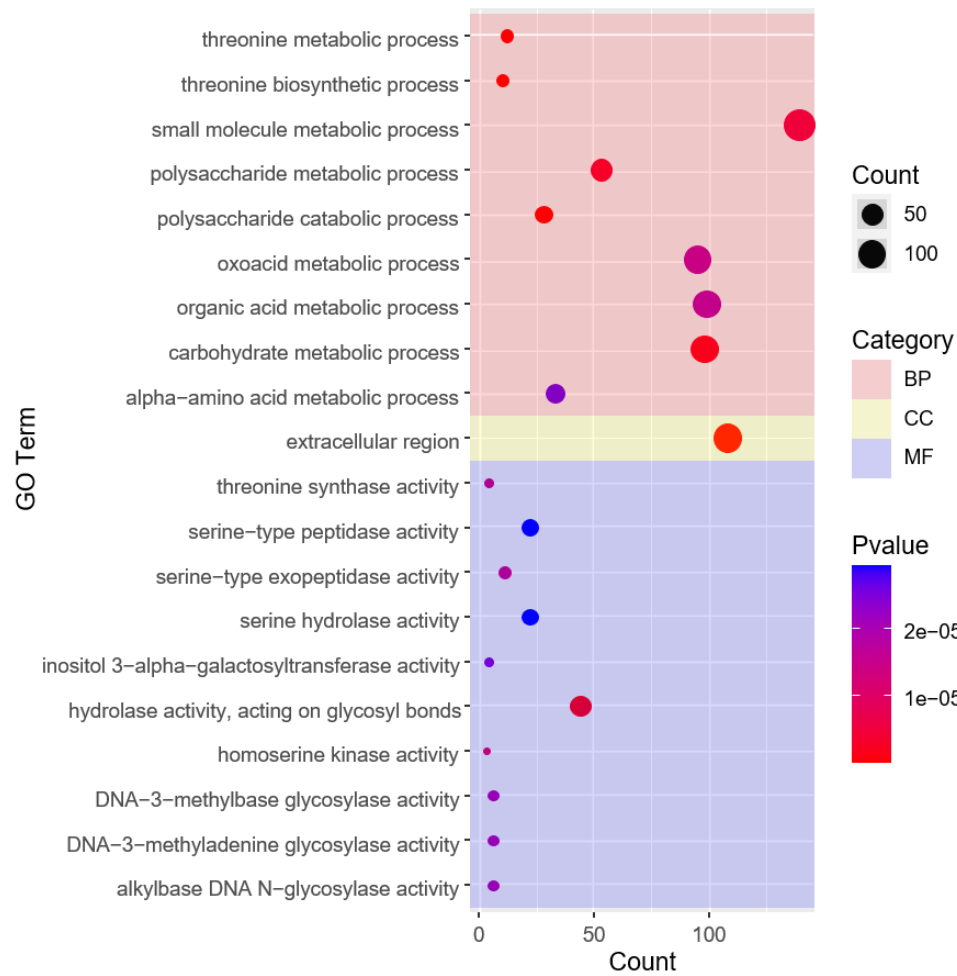


Figure 1.11. Results of Gene Ontology (GO) enrichment analysis of upregulated genes in long-flowered species at 95% of anthesis length. The y-axis lists enriched GO terms and the x-axis and the bubble size indicate the number of genes associated with a GO term. Bubbles are colored by the adjusted p-value < 0.01. The analysis was performed at each of three GO classes: biological processes (BP), molecular functions (MF), and cellular components (CC). Each GO class (BP, CC, MF) is represented by a different color (pink, yellow and purple, respectively).

Short- and long-flowered species show different expression patterns of genes involved in floral organ size

Comparison of expression patterns of flower size candidate genes among *Nicotiana* section *Repandae* polyploids and their diploid progenitors showed varying results, including distinction between short- and long-flowered taxa as well as different patterns of expression across polyploid taxa (Figure 1.12). Shorter-flowered species had higher expression of genes that putatively repress cell division. Specifically, *N. obtusifolia* showed higher expression levels of *DAR1* across development compared to other species (Figure 1.12K) whereas *N. nudicaulis* had higher levels of *BB* at 95% of anthesis length (Figure 1.12J). The long-flowered species, *N. sylvestris*, *N. repanda*, and *N. stocktonii*, showed higher expression levels in *GASA6*, which likely promotes cell elongation (Figure 1.12D). Additionally, expression of both *LNG1* and *LNG2*, which also promote cell elongation, decreases drastically at 95% of anthesis length in *N. nudicaulis*, which has short flowers (Figure 1.12G-H). *Nicotiana sylvestris*, which has the longest flowers, also has higher expression of *GASAE*, which also likely promotes cell elongation, than the others across development (Figure 1.12F). Expression patterns of *ARF2a*, *ARGOS*, *ARL*, *GASA9*, and *LNG2-2* do not seem to correlate with flower size differences (Figure 1.12A-C,E,I). Expression levels for *ANT* and *DAR2* were below 2.5 TMM counts across development, so these results are not presented in figures.

In addition to gene expression patterns correlating with short- and long-flowered species, I also found transgressive expression (outside the range of both progenitors) of

some genes in the polyploids. For *ARL*, all *Nicotiana* section *Repandae* polyploids showed transgressively higher expression at 95% of anthesis length, and *N. nudicaulis* also showed transgressive expression at 85% of anthesis (Figure 1.12C). For *GASA9*, *N. nudicaulis* and *N. stocktonii* showed transgressive downregulation across developmental stages whereas *N. repanda* only showed this pattern at 60% and 95% of anthesis length (Figure 1.12E). For *LNG1*, transgressive expression varied across development; all polyploids display transgressive upregulation at 60% and 85% of anthesis length, but this pattern is only retained by *N. repanda* at 95% of anthesis length whereas *N. nudicaulis* is transgressively downregulated at the 95% stage (Figure 1.12G). For *LNG2*, *N. repanda* and *N. stocktonii* are transgressively upregulated across development (Figure 1.12H). For *BB*, *N. repanda* and *N. stocktonii* are transgressively downregulated across development whereas *N. nudicaulis* is transgressively upregulated at 95% of anthesis length (Figure 1.12J). The reciprocal transgressive expression patterns in *N. nudicaulis* versus *N. repanda* and *N. stocktonii*, especially in *BB* and *LNG1*, may also play a role in determining the phenotypic differences in corolla tube length observed across these polyploid species.

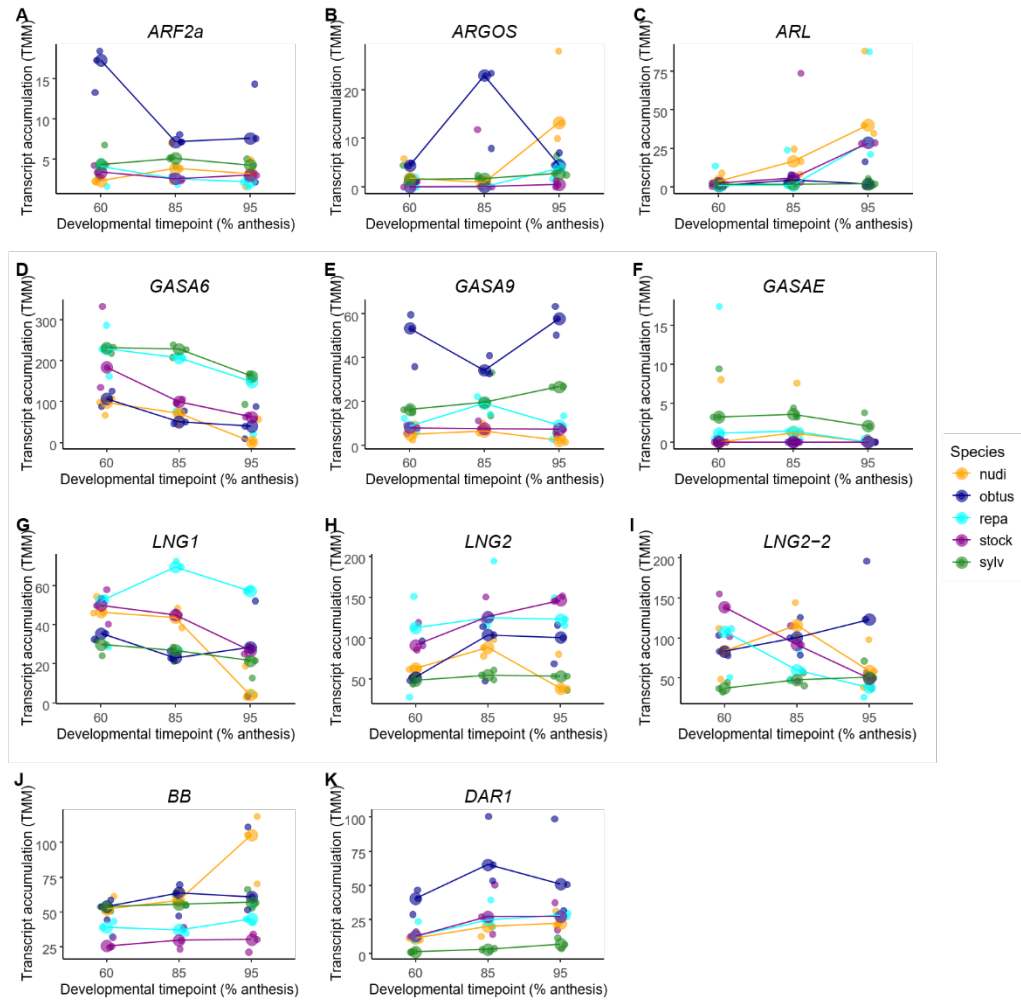


Figure 1.12. Expression of known candidate genes involved in floral organ size across corolla development stages. Here I depict TMM counts of flower size related orthologs across 60%, 85%, and 95% corolla development stages. Large points represent the median TMM values of three biological replicates and are connected via a line to facilitate observation of patterns of expression across development. Small points represent TMM values of each biological replicate. Flower size genes that promote growth are (A) *ARF2a*, (B) *ARGOS*, (C) *ARL* (*ARGOS*-like), (D) *GASA6*, (E) *GASA9*, (F) *GASAE*, (G) *LNG1*, (H) *LNG2*, and (I) *LNG2-2*. Flower size genes that repress growth are (J) *BB* and (K) *DAR1*. TMM values for *ANT*, which promotes growth, and *DAR2*, which represses growth, were not plotted as TMM values were below 2.5 across development for both of those genes. nudi=*N. nudicaulis*, obtus=*N. obtusifolia*, repa=*N. repanda*, stock=*N. stocktonii*, sylv=*N. Sylvestris*.

Discussion

Polyploidy has an important role in the evolution of organisms because it can lead to species diversification, new gene functions (Soltis et al. 2014), and the emergence of novel traits (Salman-Minkov et al. 2016). In allopolyploids, one of the two subgenomes often becomes dominant and undergoes less fragmentation; homeologs within the dominant subgenome typically exhibit higher expression levels and may play a more substantial role in determining morphology compared to those in the more fractionated subgenome (Alger & Edger 2020). Here I investigated subgenome dominance and expression of candidate genes involved in floral organ size in *Nicotiana* section *Repandae* allopolyploids and their progenitors. Notably, all allotetraploids exhibited intermediate overall gene expression profiles between those of the progenitors, with a slight but consistent bias towards *N. obtusifolia* (Figure 1.2). Moreover, *Nicotiana* section *Repandae* species demonstrated subgenome dominance toward the *N. obtusifolia* subgenome, which was stronger in *N. nudicaulis* compared to *N. repanda* and *N. stocktonii* (Figure 1.4). Further analysis into the expression of floral size candidate genes indicated differential expression that correlates with variations in flower size among the species (Figure 1.12).

Unbalanced expression bias in the corolla and leaf transcriptome

Previous studies show that in allopolyploids, many homeologs exhibit a phenomenon known as expression level dominance, in which the total level of expression is similar to one of the two progenitors (Sigel et al. 2019; Li et al. 2014). Global differential gene expression analysis showed an overall slight bias of the allopolyploids towards their

paternal progenitor, *N. obtusifolia*. When a reference-based approach was taken, the opposite trend (strong bias towards *N. sylvestris*) was observed, but this spurious result was driven by low mapping percentages of *N. obtusifolia* reads. This illustrates how important comparable mapping percentages across both subgenomes can be for reliable results in subgenome dominance analyses.

I observe consistent bias towards the paternal progenitor in *Nicotiana* section *Repandae* species across all developmental stages of corolla and leaf samples, although studies in other polyploid species have shown that expression patterns can vary among tissues. For instance, in allotetraploid *Capsella bursa-pastoris*, expression patterns in flowers were closer to that of *Capsella orientalis*, the maternal progenitor, and expression in root and leaf to that of *Capsella grandiflora*, the paternal progenitor (Kryvokhyzha et al. 2019). In the current study, leaves and corollas at three developmental stages were sampled; other tissues, such as roots or fruits, may show other patterns of bias.

When I extracted shared differentially expressed genes from comparisons of short- vs long-flowered polyploid species, I found that among genes upregulated in the short-flowered *N. nudicaulis*, genes involved in photosynthesis were significantly enriched throughout flower development (Figure 1.6-1.9). This enrichment may be due to the increased chlorophyll content in *N. nudicaulis* flowers compared to flowers of *N. repanda*, which lack chlorophyll, and *N. stocktonii*, which contains less chlorophyll, based on the magnitude of the chlorophyll absorption peak in floral reflectance spectra (McCarthy et al. 2015).

Subgenome dominance

Subgenome dominance can be influenced by genetic and epigenetic factors leading to different phenotypic outcomes: a single subgenome may dominate and control the phenotype, both subgenomes may contribute equally, or each subgenome may influence distinct traits (Bird et al. 2018). Understanding the influences of gene expression patterns on phenotypic evolution is essential for elucidating the origin and evolutionary trajectories of divergent or convergent phenotypes following allopolyploidy (Alexander-Webber et al. 2016).

I observed that *Nicotiana* section *Repandae* species exhibit subgenome dominance towards the *N. obtusifolia* subgenome, which is stronger in *N. nudicaulis* than in *N. repanda* and *N. stocktonii*, at all developmental time points (Figure 1.4, Table 1.2). Subgenome dominance has been observed in polyploids of *Arabidopsis* (Wang et al. 2004), *Brassica* (Cheng et al. 2016), *Erythranthe/Mimulus* (Edger et al. 2017), *Gossypium* (Renny-Byfield et al. 2015), *Tragopogon* (Buggs et al. 2010), *Triticum* (Li et al. 2014), and *Zea* (Schnable et al. 2011). Subgenome dominance is often associated with the level of transposable elements (TEs) found within subgenomes, with dominance correlated to the subgenome with fewer TEs (Bird et al. 2018). Although several classes of TEs show equal prevalence across the *N. sylvestris* and *N. obtusifolia* genomes, *N. sylvestris* has more TEs overall compared to *N. obtusifolia* (Parisod et al. 2012), which may explain why *Nicotiana* section *Repandae* polyploids are biased toward the *N. obtusifolia* subgenome. *Nicotiana* section *Repandae* polyploids also tend to lose TEs from the *N. obtusifolia* subgenome more

frequently than from the *N. sylvestris* subgenome, and *N. nudicaulis* had fewer new TE bands than *N. repanda* and *N. stocktonii* (Parisod et al. 2012), consistent with the stronger subgenome dominance observed in that species. In addition, *N. nudicaulis* has less genome turnover than *N. nesophila*, another species from *Nicotiana* section *Repandae* (Clarkson et al. 2005), and slower genome turnover may maintain stronger subgenome dominance. Subgenome bias tends to increase with polyploid age (Flagel et al. 2008; Flagel & Wendel 2010; Edger et al. 2017; McCarthy et al. 2024), and ~4 million year old *Nicotiana* section *Repandae* polyploids exhibit consistent subgenome dominance toward *N. obtusifolia* whereas 0.6 million year old *N. tabacum* natural polyploids and first generation synthetic lines do not show subgenome dominance across their transcriptomes as a whole (McCarthy et al. 2024).

Differential gene expression of floral organ size related genes

The vast morphological differences in floral morphology that are observed across the angiosperms are tied to the interactions between flowers and their pollinators (Crepet & Niklas 2009). This relationship drives the evolution of various floral traits, including flower size, which co-evolves with pollinator preferences to enhance reproductive success (Goodwillie et al. 2010). Larger flowers have been shown to provide more nectar and to be more visually prominent, making them more attractive to pollinators than smaller flowers (Fenster et al. 2006). These findings show differential expression of genes regulating floral organ size, including *BB*, *DAR1*, *GASA* and *LNG*. Subgenome dominance of the candidate genes was not explicitly tested because most of the single copy candidate genes did not

meet the criterion of fewer than 10% of reads mapping to the other homeolog. However, overall expression of the candidate genes indicated that *N. nudicaulis*, possessing short flowers, exhibited higher expression levels of *BB* compared to long-flowered species at 95% of anthesis length (Figure 1.12J). Shorter flowers of *Saltugilia* also had upregulated *BB* at anthesis compared to longer flowers (Landis et al. 2017). These observations are consistent with Disch et al. (2006) which showed that *BB* acts as a negative regulator of organ growth, limiting organ size in a dosage-dependent manner. Lack of *BB* activity resulted in larger organs, while overexpression of *BB* led to the formation of smaller organs (Disch et al. 2006). In addition, the *N. tomentosiformis* homeolog of the *BB* gene was upregulated in synthetic *N. tabacum* lines compared to natural *N. tabacum* accessions (Landis et al. 2020). The synthetic lines have larger cells than the natural *N. tabacum* accessions (Landis et al. 2020), which suggests that increased *BB* in synthetic lines may counteract the increased organ growth in polyploids, presumably by decreasing cell proliferation. Similarly, overexpression of *BB* in *Arabidopsis* reduced the cell number in leaves, but increased cell expansion, resulting in partial organ size compensation (Vanhaeren et al. 2017).

Nicotiana obtusifolia showed higher expression of *DAR1* across all the developmental stages of the corolla compared to the other taxa (Figure 1.12K). Peng et al. (2015) showed that *DAR1* represses cell proliferation, activates endoreduplication, and therefore coordinates organ growth in *Arabidopsis thaliana*. In maize, overexpression of mutated *ZmDA1* and *ZmDAR1*, which produces a phenotype similar to the loss-of-function allele, resulted in larger seeds and increased grain yield compared to wild-type plants (Xie

et al. 2018). *DARI*, which works redundantly with *DAI* in *Arabidopsis* (Peng et al. 2015), and *BB* seem to influence organ size independently as double *dal/bb* mutants have a more prolonged period of cell proliferation than either single mutant (Vanhaeren et al. 2017). This suggests that higher expression of either of these genes in short-flowered species may be sufficient to limit flower size.

I observed that *N. sylvestris*, *N. repanda*, and *N. stocktonii*, the long-flowered species, showed higher expression levels of genes identified as floral growth promoters, including *GASA6*, across development compared to short-flowered species (Figure 1.12D). *Nicotiana sylvestris*, which has the longest flowers, also had higher expression of *GASAE* than the others (Figure 1.12F). Similarly, Landis et al. (2020) identified high expression levels of *GASA/GASTI* genes, particularly the *N. sylvestris* homeolog, in the synthetic *N. tabacum* line QM25, which exhibits the longest flowers of the synthetic and natural *N. tabacum* accessions in the study. Recent studies have found 18 *GASA* genes in the *N. tabacum* genome distributed across 13 chromosomes (Li et al. 2021) and 17 *GASA* genes across eight chromosomes in the related *Solanum lycopersicum* (Su et al. 2023), highlighting the large and potentially variable number of genes in this gene family. I found differential expression in three different *GASA* copies, each with different expression patterns (Figure 1.12). *GASA6* showed upregulation in *N. sylvestris* and *N. repanda* compared to all other species (Figure 1.12D), while *GASA9* was upregulated in *N. obtusifolia* (Figure 1.12E), and *GASAE* was upregulated in only *N. sylvestris* (Figure 1.12F). Two of the three *GASA* genes exhibiting differential expression also showed lower sequence divergence than the average (10.2%) across all single copy genes, with *GASAE*

being 7.3% divergent and *GASA6* 8.1% divergent, and with *GASA9* showing the most divergence at 10.3%. The low sequence divergence is a consequence of selection on phenotypes, where pressure to retain specific gene functions limits divergence, thereby conserving gene function across phenotypically diverse species (*e.g.*, large vs small flowers), however functional work in multiple *Nicotiana* species is necessary. This would involve comparing the expression levels and phenotypic effects of different alleles, specifically looking at how these variations influence flower size. Such studies would provide insights into how conserved gene functions contribute to the phenotypic diversity observed in these species. In *A. thaliana*, *GASA6* was shown to promote elongation of the embryonic axis (Zhong et al. 2015). In *Saltugilia* (Polemoniaceae), *GASA* expression levels were highest in flowers at anthesis, corresponding to cell elongation in the petal tube (Landis et al. 2017, 2015). However, some *GASA* genes have been shown to repress cell elongation in ray floret petals of *Gerbera* (Kotilainen et al. 1999) and *Chrysanthemum* (He et al. 2024), suggesting that further functional work is necessary to determine the role that individual *GASA* genes play in determining flower size in *Nicotiana*.

I also observe that expression of *LNG1* and *LNG2*, which promote cell elongation, decreases drastically at 95% of anthesis length in short-flowered *N. nudicaulis* (Figure 1.12G-H). Overexpression of *Arabidopsis LNG1* in tobacco plants resulted in increased leaf petiole and blade length compared with wild type tobacco, suggesting functional conservation of *Arabidopsis LNG1* in tobacco (Lee & Kim 2018). A quantitative trait locus for petal length in roses was found to include an ortholog of *LNG* (Schulz et al. 2021), further suggesting the shared function of these genes. The hypothesis that *LNG* genes have

a conserved function of promoting cell elongation and increasing floral organ size in *Nicotiana* section *Repandae* needs to be tested in the future.

Other candidate genes showed variable expression patterns. Both *ARF2a* and *ARGOS*, which have been shown to promote cell proliferation (Hu et al. 2003, 2006), showed little differences in expression aside from *N. obtusifolia* (Figure 1.12A-B). The *ARL* (*ARGOS*-like) gene controls cell expansion (Hu et al. 2006) and showed a unique pattern of differential expression with all three allopolyploids having transgressively higher gene expression, especially later in flower development, than either progenitor species (Figure 1.12C), which seems to be driven by increased expression of the *N. sylvestris* homeolog, based on strong bias toward the *N. sylvestris* copy in these species (Figure 1.5A). *ARL* has been shown to promote organ growth via cell expansion (Hu et al. 2006), however *N. sylvestris*, which has the longest flowers at 4.7-7.6cm, has the lowest expression along with the much smaller *N. obtusifolia* at 1.6-1.9 cm (McCarthy et al. 2016).

I also observe different patterns of homeolog bias in certain genes (*GASA9* and *LNG2*) among *Nicotiana* section *Repandae* polyploids, with *N. nudicaulis* being biased toward *N. sylvestris* or having equal expression whereas *N. repanda* and *N. stocktonii* are biased toward *N. obtusifolia* (Figure 1.5B-C). These results are in contrast with those from *N. tabacum*, in which McCarthy et al. (2024) observed that the homeolog bias pattern was always consistent between related polyploids of the same age (within synthetic lines and within natural accessions), even though homeolog bias patterns varied from gene to gene. This suggests that greater divergence times are necessary to develop different patterns of

homeolog bias in related allopolyploid species. *Nicotiana nudicaulis* diverged from the lineage that led to *N. repanda* and *N. stocktonii* ~2-3 million years ago whereas *N. repanda* and *N. stocktonii* diverged only ~one million years ago, and the polyploidization event that resulted in the formation of *N. tabacum* occurred ~0.6 million years ago (Clarkson et al. 2017), suggesting that 2-3 million years of divergence between polyploid species may be necessary for different patterns of homeolog bias to evolve in related polyploid species, at least in *Nicotiana*.

Conclusion

My transcriptome analyses show that subgenome dominance is a dynamic evolutionary force, varying across tissues and developmental stages. Specifically, we see the dominance of the *N. obtusifolia* subgenome in different developmental stages in *N. nudicaulis*, and to a lesser extent in *N. repanda* and *N. stocktonii*, highlighting the evolutionary importance of this progenitor in shaping phenotypic diversity in the allopolyploids. Furthermore, I see differential expression of candidate genes associated with floral organ size. The expression patterns of flower size related genes in short- versus long-flowered species illuminate the genetic underpinnings that contribute to morphological differences, emphasizing the evolutionary implications of allopolyploidy. This relationship is crucial for future work understanding the ecological interactions of these species, particularly with pollinators, and their reproductive strategies.

References

- Alexander-Webber, Douglas, Richard J. Abbott, and Mark A. Chapman. 2016. "Morphological Convergence Between an Allopolyploid and One of Its Parental Species Correlates with Biased Gene Expression and DNA Loss." *The Journal of Heredity* 107 (5): 445–54.
- Alger, Elizabeth I., and Patrick P. Edger. 2020. "One Subgenome to Rule Them All: Underlying Mechanisms of Subgenome Dominance." *Current Opinion in Plant Biology* 54 (April):108–13.
- Barker, Michael S., Nils Arrigo, Anthony E. Baniaga, Zheng Li, and Donald A. Levin. 2016. "On the Relative Abundance of Autopolyploids and Allopolyploids." *The New Phytologist* 210 (2): 391–98.
- Bird, Kevin A., Robert VanBuren, Joshua R. Puzey, and Patrick P. Edger. 2018. "The Causes and Consequences of Subgenome Dominance in Hybrids and Recent Polyploids." *The New Phytologist* 220 (1): 87–93.
- Bryant, Donald M., Kimberly Johnson, Tia DiTommaso, Timothy Tickle, Matthew Brian Couger, Duygu Payzin-Dogru, Tae J. Lee, et al. 2017. "A Tissue-Mapped Axolotl DE Novo Transcriptome Enables Identification of Limb Regeneration Factors." *Cell Reports* 18 (3): 762–76.
- Buchfink, Benjamin, Klaus Reuter, and Hajk-Georg Drost. 2021. "Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND." *Nature Methods* 18 (4): 366–68.
- Buggs, Richard J. A., Natalie M. Elliott, Linjing Zhang, Jin Koh, Lyderson F. Viccini, Douglas E. Soltis, and Pamela S. Soltis. 2010. "Tissue-Specific Silencing of Homoeologs in Natural Populations of the Recent Allopolyploid *Tragopogon Mirus*." *The New Phytologist* 186 (1): 175–83.
- Buggs, Richard J. A., Jonathan F. Wendel, Jeffrey J. Doyle, Douglas E. Soltis, Pamela S. Soltis, and Jeremy E. Coate. 2014. "The Legacy of Diploid Progenitors in Allopolyploid Gene Expression Patterns." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 369 (1648). <https://doi.org/10.1098/rstb.2013.0354>.
- Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (1): 421.

- Chase, Mark W., Sandra Knapp, Antony V. Cox, James J. Clarkson, Yelena Butsko, Jeffrey Joseph, Vincent Savolainen, and Alex S. Parokonny. 2003. "Molecular Systematics, GISH and the Origin of Hybrid Taxa in *Nicotiana* (Solanaceae)." *Annals of Botany* 92 (1): 107–27.
- Cheng, Feng, Chao Sun, Jian Wu, James Schnable, Margaret R. Woodhouse, Jianli Liang, Chengcheng Cai, Michael Freeling, and Xiaowu Wang. 2016. "Epigenetic Regulation of Subgenome Dominance Following Whole Genome Triplication in *Brassica Rapa*." *The New Phytologist* 211 (1): 288–99.
- Cheng, Feng, Jian Wu, Xu Cai, Jianli Liang, Michael Freeling, and Xiaowu Wang. 2018. "Gene Retention, Fractionation and Subgenome Differences in Polyploid Plants." *Nature Plants* 4 (5): 258–68.
- Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. "Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor." *Bioinformatics (Oxford, England)* 34 (17): i884–90.
- Clarkson, James J., Steven Dodsworth, and Mark W. Chase. 2017. "Time-Calibrated Phylogenetic Trees Establish a Lag between Polyploidisation and Diversification in *Nicotiana* (Solanaceae)." *Plant Systematics and Evolution* 303 (8): 1001–12.
- Clarkson, James J., Sandra Knapp, Vicente F. Garcia, Richard G. Olmstead, Andrew R. Leitch, and Mark W. Chase. 2004. "Phylogenetic Relationships in *Nicotiana* (Solanaceae) Inferred from Multiple Plastid DNA Regions." *Molecular Phylogenetics and Evolution* 33 (1): 75–90.
- Clarkson, James J., K. Yoong Lim, Ales Kovarik, Mark W. Chase, Sandra Knapp, and Andrew R. Leitch. 2005. "Long-Term Genome Diploidization in Allopolyploid *Nicotiana* Section *Repandae* (Solanaceae)." *The New Phytologist* 168 (1): 241–52.
- Colle, Marivi, Courtney P. Leisner, Ching Man Wai, Shujun Ou, Kevin A. Bird, Jie Wang, Jennifer H. Wisecaver, et al. 2019. "Haplotype-Phased Genome and Evolution of Phytonutrient Pathways of Tetraploid Blueberry." *GigaScience* 8 (3). <https://doi.org/10.1093/gigascience/giz012>.
- Crepet, William L., and Karl J. Niklas. 2009. "Darwin's Second 'Abominable Mystery': Why Are There so Many Angiosperm Species?" *American Journal of Botany* 96 (1): 366–81.
- D'Andrea, Lucio, Nicolas Sierro, Sonia Ouadi, Tomas Hasing, Elijah Rinaldi, Nikolai V. Ivanov, and Aureliano Bombarely. 2023. "Polyploid *Nicotiana* Section *Suaveolentes* Originated by Hybridization of Two Ancestral *Nicotiana* Clades." *Frontiers in Plant Science* 14 (May):999887.

- Disch, Sabine, Elena Anastasiou, Vijay K. Sharma, Thomas Laux, Jennifer C. Fletcher, and Michael Lenhard. 2006. "The E3 Ubiquitin Ligase BIG BROTHER Controls Arabidopsis Organ Size in a Dosage-Dependent Manner." *Current Biology: CB* 16 (3): 272–79.
- Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21.
- Du, Liang, Na Li, Liangliang Chen, Yingxiu Xu, Yu Li, Yueying Zhang, Chuanyou Li, and Yunhai Li. 2014. "The Ubiquitin Receptor DA1 Regulates Seed and Organ Size by Modulating the Stability of the Ubiquitin-Specific Protease UBP15/SOD2 in Arabidopsis." *The Plant Cell* 26 (2): 665–77.
- Edger, Patrick P., Thomas J. Poorten, Robert VanBuren, Michael A. Hardigan, Marivi Colle, Michael R. McKain, Ronald D. Smith, et al. 2019. "Origin and Evolution of the Octoploid Strawberry Genome." *Nature Genetics* 51 (3): 541–47.
- Edger, Patrick P., Ronald Smith, Michael R. McKain, Arielle M. Cooley, Mario Vallejo-Marin, Yaowu Yuan, Adam J. Bewick, et al. 2017. "Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower." *The Plant Cell* 29 (9): 2150–67.
- Emms, David M., and Steven Kelly. 2019. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology* 20 (1): 238.
- Falcon, S., and R. Gentleman. 2007. "Using GOstats to Test Gene Lists for GO Term Association." *Bioinformatics* 23 (2): 257–58.
- Feng, Guanping, Zhixiang Qin, Jingzhou Yan, Xiaoran Zhang, and Yuxin Hu. 2011. "Arabidopsis ORGAN SIZE RELATED1 Regulates Organ Growth and Final Organ Size in Orchestration with ARGOS and ARL." *The New Phytologist* 191 (3): 635–46.
- Fenster, Charles B., George Cheely, Michele R. Dudash, and Richard J. Reynolds. 2006. "Nectar Reward and Advertisement in Hummingbird-Pollinated *Silene virginica* (Caryophyllaceae)." *American Journal of Botany* 93 (12): 1800–1807.
- Flagel, Lex E., and Jonathan F. Wendel. 2010. "Evolutionary Rate Variation, Genomic Dominance and Duplicate Gene Expression Evolution during Allotetraploid Cotton Speciation." *The New Phytologist* 186 (1): 184–93.
- Flagel, Lex, Joshua Udall, Dan Nettleton, and Jonathan Wendel. 2008. "Duplicate Gene Expression in Allopolyploid *Gossypium* Reveals Two Temporally Distinct Phases of Expression Evolution." *BMC Biology* 6 (April): 16.

- Goodwillie, Carol, Risa D. Sargent, Christopher G. Eckert, Elizabeth Elle, Monica A. Geber, Mark O. Johnston, Susan Kalisz, et al. 2010. "Correlated Evolution of Mating System and Floral Display Traits in Flowering Plants and Its Implications for the Distribution of Mating System Variation." *The New Phytologist* 185 (1): 311–21.
- Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, et al. 2011. "Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome." *Nature Biotechnology* 29 (7): 644–52.
- Grant, Verne. 1981. *Plant Speciation*. Columbia University Press.
- Haas, Brian J., Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D. Blood, Joshua Bowden, Matthew Brian Couger, et al. 2013. "De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis." *Nature Protocols* 8 (8): 1494–1512.
- Herzog, M., A. M. Dorne, and F. Grellet. 1995. "GASA, a Gibberellin-Regulated Gene Family from Arabidopsis Thaliana Related to the Tomato GAST1 Gene." *Plant Molecular Biology* 27 (4): 743–52.
- He, Ziyang, Rui Jiang, Xiaojing Wang, and Yaqin Wang. 2024. "A GASA Protein Family Gene, CmGEG, Inhibits Petal Growth in Chrysanthemum." *International Journal of Molecular Sciences* 25 (6). <https://doi.org/10.3390/ijms25063367>.
- Hu, Yuxin, Huay Mei Poh, and Nam-Hai Chua. 2006. "The Arabidopsis ARGOS-LIKE Gene Regulates Cell Expansion during Organ Growth." *The Plant Journal: For Cell and Molecular Biology* 47 (1): 1–9.
- Hu, Yuxin, Qi Xie, and Nam-Hai Chua. 2003. "The Arabidopsis Auxin-Inducible Gene ARGOS Controls Lateral Organ Size." *The Plant Cell* 15 (9): 1951–61.
- Katoh, Kazutaka, and Daron M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30 (4): 772–80.
- Kotilainen, M., Y. Helariutta, M. Mehto, E. Pollanen, V. A. Albert, P. Elomaa, and T. H. Teeri. 1999. "GEG Participates in the Regulation of Cell and Organ Shape during Corolla and Carpel Development in Gerbera Hybrida." *The Plant Cell* 11 (6): 1093–1104.
- Krizek, Beth A., and Jill T. Anderson. 2013. "Control of Flower Size." *Journal of Experimental Botany* 64 (6): 1427–37.

- Kryvokhyzha, Dmytro, Pascal Milesi, Tianlin Duan, Marion Orsucci, Stephen I. Wright, Sylvain Glémin, and Martin Lascoux. 2019. “Towards the New Normal: Transcriptomic Convergence and Genomic Legacy of the Two Subgenomes of an Allopolyploid Weed (*Capsella Bursa-Pastoris*).” *PLoS Genetics* 15 (5): e1008131.
- Landis, Jacob B., Amelda Kurti, Amber J. Lawhorn, Amy Litt, and Elizabeth W. McCarthy. 2020. “Differential Gene Expression with an Emphasis on Floral Organ Size Differences in Natural and Synthetic Polyploids of *Nicotiana Tabacum* (Solanaceae).” *Genes* 11 (9): 1097.
- Landis, Jacob B., Rebecca D. O’Toole, Kayla L. Ventura, Matthew A. Gitzendanner, David G. Oppenheimer, Douglas E. Soltis, and Pamela S. Soltis. 2015. “The Phenotypic and Genetic Underpinnings of Flower Size in Polemoniaceae.” *Frontiers in Plant Science* 6:1144.
- Landis, Jacob B., Douglas E. Soltis, Zheng Li, Hannah E. Marx, Michael S. Barker, David C. Tank, and Pamela S. Soltis. 2018. “Impact of Whole-Genome Duplication Events on Diversification Rates in Angiosperms.” *American Journal of Botany* 105 (3): 348–63.
- Landis, Jacob B., Douglas E. Soltis, and Pamela S. Soltis. 2017. “Comparative Transcriptomic Analysis of the Evolution and Development of Flower Size in *Saltugilia* (Polemoniaceae).” *BMC Genomics* 18 (1): 475.
- Langmead, Ben, Christopher Wilks, Valentin Antonescu, and Rone Charles. 2019. “Scaling Read Aligners to Hundreds of Threads on General-Purpose Processors.” *Bioinformatics (Oxford, England)* 35 (3): 421–32.
- Lee, Young Koungh, Gyung-Tae Kim, In-Jung Kim, Jeongmoo Park, Sang-Soo Kwak, Giltso Choi, and Won-Il Chung. 2006. “LONGIFOLIA1 and LONGIFOLIA2, Two Homologous Genes, Regulate Longitudinal Cell Elongation in Arabidopsis.” *Development* 133 (21): 4305–14.
- Lee, Young Koungh, and In-Jung Kim. 2018. “Functional Conservation of Arabidopsis LNG1 in Tobacco Relating to Leaf Shape Change by Increasing Longitudinal Cell Elongation by Overexpression.” *Genes & Genomics* 40 (10): 1053–62.
- Leitch, I. J., L. Hanson, K. Y. Lim, A. Kovarik, M. W. Chase, J. J. Clarkson, and A. R. Leitch. 2008. “The Ups and Downs of Genome Size Evolution in Polyploid Species of *Nicotiana* (Solanaceae).” *Annals of Botany* 101 (6): 805–14.
- Li, Aili, Dengcai Liu, Jun Wu, Xubo Zhao, Ming Hao, Shuaifeng Geng, Jun Yan, et al. 2014. “mRNA and Small RNA Transcriptomes Reveal Insights into Dynamic Homoeolog Regulation of Allopolyploid Heterosis in Nascent Hexaploid Wheat.” *The Plant Cell* 26 (5): 1878–1900.

- Liao, Yang, Gordon K. Smyth, and Wei Shi. 2013. “The Subread Aligner: Fast, Accurate and Scalable Read Mapping by Seed-and-Vote.” *Nucleic Acids Research* 41 (10): e108.
- Li, Bo, and Colin N. Dewey. 2011. “RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome.” *BMC Bioinformatics* 12 (1): 323.
- Li, Yunhai, Leiying Zheng, Fiona Corke, Caroline Smith, and Michael W. Bevan. 2008. “Control of Final Seed and Organ Size by the DA1 Gene Family in *Arabidopsis Thaliana*.” *Genes & Development* 22 (10): 1331–36.
- Li, Zhaowu, Junping Gao, Genhong Wang, Shuaibin Wang, Kai Chen, Wenxuan Pu, Yaofu Wang, Qingyou Xia, and Xiaorong Fan. 2021. “Genome-Wide Identification and Characterization of GASA Gene Family in *Nicotiana Tabacum*.” *Frontiers in Genetics* 12:768942.
- Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.
- Manni, Mosè, Matthew R. Berkeley, Mathieu Seppey, Felipe A. Simão, and Evgeny M. Zdobnov. 2021. “BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes.” *Molecular Biology and Evolution* 38 (10): 4647–54.
- Ma, Peng-Fei, Yun-Long Liu, Cen Guo, Guihua Jin, Zhen-Hua Guo, Ling Mao, Yi-Zhou Yang, et al. 2024. “Genome Assemblies of 11 Bamboo Species Highlight Diversification Induced by Dynamic Subgenome Dominance.” *Nature Genetics*, March. <https://doi.org/10.1038/s41588-024-01683-0>.
- McCarthy, Davis J., Yunshun Chen, and Gordon K. Smyth. 2012. “Differential Expression Analysis of Multifactor RNA-Seq Experiments with Respect to Biological Variation.” *Nucleic Acids Research* 40 (10): 4288–97.
- McCarthy, Elizabeth W., Sarah E. J. Arnold, Lars Chittka, Steven C. Le Comber, Robert Verity, Steven Dodsworth, Sandra Knapp, et al. 2015. “The Effect of Polyploidy and Hybridization on the Evolution of Floral Colour in *Nicotiana* (Solanaceae).” *Annals of Botany* 115 (7): 1117–31.
- McCarthy, Elizabeth W., Mark W. Chase, Sandra Knapp, Amy Litt, Andrew R. Leitch, and Steven C. Le Comber. 2016. “Transgressive Phenotypes and Generalist Pollination in the Floral Evolution of *Nicotiana* Polyploids.” *Nature Plants* 2 (August):16119.

- McCarthy, Elizabeth W., Jacob B. Landis, Abigail G. McCoy, Amber J. Lawhorn, Amelda Kurti, Yimin Xu, James J. Giovannoni, and Amy Litt. 2023. "Homeologue Differential Expression in the Flavonoid Biosynthetic Pathway Underlies Flower Colour Variation in Natural and Synthetic Polyploids of *Nicotiana Tabacum* (Solanaceae)." *Botanical Journal of the Linnean Society. Linnean Society of London* 204 (October):283–303.
- Parisod, Christian, Corinne Mhiri, K. Yoong Lim, James J. Clarkson, Mark W. Chase, Andrew R. Leitch, and Marie-Angèle Grandbastien. 2012. "Differential Dynamics of Transposable Elements during Long-Term Diploidization of *Nicotiana* Section Repandae (Solanaceae) Allopolyploid Genomes." *PloS One* 7 (11): e50352.
- Peng, Yuancheng, Liangliang Chen, Yaru Lu, Yingbao Wu, Jack Dumenil, Zhengge Zhu, Michael W. Bevan, and Yunhai Li. 2015. "The Ubiquitin Receptors DA1, DAR1, and DAR2 Redundantly Regulate Endoreduplication by Modulating the Stability of TCP14/15 in Arabidopsis." *The Plant Cell* 27 (3): 649–62.
- Pfeifer, Matthias, Karl G. Kugler, Simen R. Sandve, Bujie Zhan, Heidi Rudi, Torgeir R. Hvidsten, International Wheat Genome Sequencing Consortium, Klaus F. X. Mayer, and Odd-Arne Olsen. 2014. "Genome Interplay in the Grain Transcriptome of Hexaploid Bread Wheat." *Science* 345 (6194): 1250091.
- R Core team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org>.
- Renny-Byfield, Simon, Lei Gong, Joseph P. Gallagher, and Jonathan F. Wendel. 2015. "Persistence of Subgenomes in Paleopolyploid Cotton after 60 My of Evolution." *Molecular Biology and Evolution* 32 (4): 1063–71.
- Renny-Byfield, Simon, Ales Kovařík, Michael Chester, Richard A. Nichols, Jiri Macas, Petr Novák, and Andrew R. Leitch. 2012. "Independent, Rapid and Targeted Loss of Highly Repetitive DNA in Natural and Synthetic Allopolyploids of *Nicotiana Tabacum*." *PloS One* 7 (5): e36963.
- Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics (Oxford, England)* 26 (1): 139–40.
- Roxrud, Ingrid, Stein Erik Lid, Jennifer C. Fletcher, Ed D. L. Schmidt, and Hilde-Gunn Opsahl-Sorteberg. 2007. "GASA4, One of the 14-Member Arabidopsis GASA Family of Small Polypeptides, Regulates Flowering and Seed Development." *Plant & Cell Physiology* 48 (3): 471–83.
- Salman-Minkov, Ayelet, Niv Sabath, and Itay Mayrose. 2016. "Whole-Genome Duplication as a Key Factor in Crop Domestication." *Nature Plants* 2 (August):16115.

- Schloss, Patrick D., Sarah L. Westcott, Thomas Ryabin, Justine R. Hall, Martin Hartmann, Emily B. Hollister, Ryan A. Lesniewski, et al. 2009. "Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities." *Applied and Environmental Microbiology* 75 (23): 7537–41.
- Schnable, James C., Nathan M. Springer, and Michael Freeling. 2011. "Differentiation of the Maize Subgenomes by Genome Dominance and Both Ancient and Ongoing Gene Loss." *Proceedings of the National Academy of Sciences of the United States of America* 108 (10): 4069–74.
- Schulz, Dietmar, Marcus Linde, and Thomas Debener. 2021. "Detection of Reproducible Major Effect QTL for Petal Traits in Garden Roses." *Plants* 10 (5). <https://doi.org/10.3390/plants10050897>.
- Sierro, N., J. N. D. Battey, L. Bovet, V. Liedschulte, S. Ouadi, J. Thomas, H. Broye, et al. 2018. "The Impact of Genome Evolution on the Allotetraploid *Nicotiana Rustica* - an Intriguing Story of Enhanced Alkaloid Production." *BMC Genomics* 19 (1): 855.
- Sierro, Nicolas, James N. D. Battey, Sonia Ouadi, Lucien Bovet, Simon Goepfert, Nicolas Bakaher, Manuel C. Peitsch, and Nikolai V. Ivanov. 2013. "Reference Genomes and Transcriptomes of *Nicotiana Sylvestris* and *Nicotiana Tomentosiformis*." *Genome Biology* 14 (6): R60.
- Sigel, Erin M., Joshua P. Der, Michael D. Windham, and Kathleen M. Pryer. 2019. "Expression Level Dominance and Homeolog Expression Bias in Recurrent Origins of the Allopolyploid Fern *Polypodium Hesperium*." *American Fern Journal* 109 (3): 224–47.
- Solis-Montero, Lislie, and Mario Vallejo-Marín. 2017. "Does the Morphological Fit between Flowers and Pollinators Affect Pollen Deposition? An Experimental Test in a Buzz-Pollinated Species with Anther Dimorphism." *Ecology and Evolution* 7 (8): 2706–15.
- Soltis, Douglas E., Victor A. Albert, Jim Leebens-Mack, Charles D. Bell, Andrew H. Paterson, Chunfang Zheng, David Sankoff, Claude W. Depamphilis, P. Kerr Wall, and Pamela S. Soltis. 2009. "Polyploidy and Angiosperm Diversification." *American Journal of Botany* 96 (1): 336–48.
- Soltis, Pamela S., Xiaoxian Liu, D. Blaine Marchant, Clayton J. Visger, and Douglas E. Soltis. 2014. "Polyploidy and Novelty: Gottlieb's Legacy." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 369 (1648): 20130351.

- Soltis, Pamela S., D. Blaine Marchant, Yves Van de Peer, and Douglas E. Soltis. 2015. "Polyploidy and Genome Evolution in Plants." *Current Opinion in Genetics & Development* 35 (December):119–25.
- Stebbins, G. L., Jr. 1947. "Types of Polyploids; Their Classification and Significance." *Advances in Genetics* 1:403–29.
- Su, Dan, Kaidong Liu, Zhuoshu Yu, Ying Li, Yaixin Zhang, Yunqi Zhu, Yi Wu, et al. 2023. "Genome-Wide Characterization of the Tomato GASA Family Identifies SlGASA1 as a Repressor of Fruit Ripening." *Horticulture Research* 10 (1): uhac222.
- Valadão-Mendes, L. B., I. Rocha, D. A. L. Meireles, F. B. Leite, M. Sazima, P. K. Maruyama, and V. L. G. Brito. 2022. "Flower Morphology and Plant-Bee Pollinator Interactions Are Related to Stamen Dimorphism in Melastomataceae." *Plant Biology* 24 (2): 240–48.
- Van de Peer, Yves, Eshchar Mizrahi, and Kathleen Marchal. 2017. "The Evolutionary Significance of Polyploidy." *Nature Reviews. Genetics* 18 (7): 411–24.
- Vanhaeren, Hannes, Youn-Jeong Nam, Liesbeth De Milde, Eunyoung Chae, Veronique Storme, Detlef Weigel, Nathalie Gonzalez, and Dirk Inzé. 2017. "Forever Young: The Role of Ubiquitin Receptor DA1 and E3 Ligase BIG BROTHER in Controlling Leaf Growth and Development." *Plant Physiology* 173 (2): 1269–82.
- Wang, Jianlin, Lu Tian, Hyeon-Se Lee, Ning E. Wei, Hongmei Jiang, Brian Watson, Andreas Madlung, et al. 2006. "Genomewide Nonadditive Gene Regulation in Arabidopsis Allotetraploids." *Genetics* 172 (1): 507–17.
- Wang, Jianlin, Lu Tian, Andreas Madlung, Hyeon-Se Lee, Meng Chen, Jinsuk J. Lee, Brian Watson, Trevor Kagochi, Luca Comai, and Z. Jeffrey Chen. 2004. "Stochastic and Epigenetic Changes of Gene Expression in Arabidopsis Polyploids." *Genetics* 167 (4): 1961–73.
- Wang, Jie-Li, Min-Qiang Tang, Sheng Chen, Xiang-Feng Zheng, Hui-Xian Mo, Sheng-Jun Li, Zheng Wang, et al. 2017. "Down-Regulation of BnDA1, Whose Gene Locus Is Associated with the Seeds Weight, Improves the Seeds Weight and Organ Size in Brassica Napus." *Plant Biotechnology Journal* 15 (8): 1024–33.
- Wang, Zheng, Jinghua Yang, Feng Cheng, Peirong Li, Xiaoyun Xin, Weihong Wang, Yangjun Yu, et al. 2022. "Subgenome Dominance and Its Evolutionary Implications in Crop Domestication and Breeding." *Horticulture Research* 9 (April):uhac090.
- Warnes, G. R., B. Bolker, L. Bonebakker, R. Gentleman, W. H. A. Liaw, T. Lumley, and B. Venables. 2015. "Gplots: Various R Programming Tools for Plotting Data. R Package Version 2.17. 0 [computer Software]."

- Wendel, Jonathan F., Damon Lisch, Guanjing Hu, and Annaliese S. Mason. 2018. “The Long and Short of Doubling down: Polyploidy, Epigenetics, and the Temporal Dynamics of Genome Fractionation.” *Current Opinion in Genetics & Development* 49 (April):1–7.
- Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer International Publishing.
- Xie, Guangning, Zhaoxia Li, Qijun Ran, Hui Wang, and Juren Zhang. 2018. “Over-Expression of Mutated ZmDA1 or ZmDAR1 Gene Improves Maize Kernel Yield by Enhancing Starch Synthesis.” *Plant Biotechnology Journal* 16 (1): 234–44.
- Yoo, Mi-Jeong, Xiaoxian Liu, J. Chris Pires, Pamela S. Soltis, and Douglas E. Soltis. 2014. “Nonadditive Gene Expression in Polyploids.” *Annual Review of Genetics* 48:485–517.
- Zhong, Chunmei, Hao Xu, Siting Ye, Shiyi Wang, Lingfei Li, Shengchun Zhang, and Xiaojing Wang. 2015. “Gibberellic Acid-Stimulated Arabidopsis6 Serves as an Integrator of Gibberellin, Absciscic Acid, and Glucose Signaling during Seed Germination in Arabidopsis.” *Plant Physiology* 169 (3): 2288–2303.

Chapter 2: Comparative genomics of two *Boechnera* species: identification of environmental adaptations

Abstract

As global temperatures increase, understanding the genomic basis of plant adaptability to extreme environments and tolerance to abiotic stresses, especially heat and drought, has a major role to optimize productivity and survival of crops and other important plant species. Prior studies have shown that several species in a group of native California plants in the genus *Boechnera* can survive longer than their closely related model species, *Arabidopsis thaliana*, under environmental extremes. The two species of interest in this study include *Boechnera arcuata*, found in low-elevation coastal areas, and *Boechnera perennans*, found in desert and inland chaparral regions. To elucidate the molecular level of their adaptations to diverse environments and more specifically to desert conditions, I use generated chromosome-level genome assemblies and annotations of the two *Boechnera* species and then integrate comparative and functional genomics approaches. Synteny analysis across the seven homologous chromosomes of *Boechnera* reveal extensive conservation between the two species. Synteny comparison of *Boechnera* with *A. thaliana* (n=5) demonstrates notable recombination and rearrangement events since the divergence of these genera. I identify 639 and 920 unique genes in *B. arcuata* and *B. perennans*, respectively. GO enrichment analysis of these species-specific genes uncover a significant enrichment of terms related to photosynthesis and stress response in *B. perennans*. Moreover, I find expansions of gene families associated with survival under environmental

extremes providing the basis for future analyses of their evolutionary trajectories. Through comparative genomics, this chapter elucidates the evolutionary processes that enable *Boechera* species to thrive in diverse and challenging habitats, contributing to broader knowledge of plant adaptation and speciation.

Introduction

The Brassicaceae is a large family with approximately 4,000 species. There are three major lineages (I, II and II) six clades (A-F) and 52 monophyletic tribes (Franzke et al. 2011; Hohmann et al. 2015; Franzke, Koch, and Mummenhoff 2016; Huang et al. 2016; Nikolov et al. 2019; German et al. 2023; Guo et al. 2017; Hay et al. 2023; Hendriks et al. 2023). These diverse lineages include several economically important species including the members of the *Brassica* genus. In addition, the model plant species *Arabidopsis thaliana* is a member of this family (Koenig and Weigel 2015). This has facilitated the molecular and genomic studies of other members of this family. The small genomes, wide distribution, economic importance, and ecological diversity of the Brassicaceae have made this an excellent system for a wide range of research studies.

The tribe Boechereae is composed of nine genera and approximately 130 species. The vast majority of Boechereae taxa occur only in North America. The vast majority of Boechereae genera (seven out of nine: Anelsonia, Cusickiella, Nevada, Phoenicaulis, Polytanium, Sandbergia, and Yosemitea) have only one or two identified species. With over 100 species the genus *Boechera* is by far the most diverse and speciose genus within this tribe (Alexander et al. 2013; Rushworth et al. 2011; Hay et al. 2023). As currently

defined the *Boechea* is composed of 83 sexual diploid species and more than 400 documented apomictic hybrids (F.-W. Li et al. 2017). These hybrids can be diploid or polyploid, depending on the specific hybrid combination and reproductive mechanisms involved. Some are widespread and may represent new species, others are rare and may not represent long-lived lineages. Due to the large number of hybrids and few morphological features that are useful for identification there is much uncertainty concerning *Boechea* systematics. Definitive species identifications must be confirmed with genetic analysis (F.-W. Li et al. 2017).

Despite these challenges, *Boechea* presents a highly valuable system for ecological and genetic studies due to its wide geographic distribution and diverse habitats across North America, ranging from deserts to alpine meadows (Rushworth et al. 2011). This genus has become an excellent model for ecological and evolutionary functional genomics due to its many advantageous traits. These include its relationship to *Arabidopsis thaliana* which enhances genomic resources (Alexander et al. 2013; Schranz et al. 2007), the presence of both sexual and asexual reproductive strategies within the species (Schranz et al. 2005, 2006), and the adaptability to inbreeding and self-fertilization, which aids genetic experiments such as the development of recombinant inbred lines (Rushworth et al. 2011). Research on *Boechea* also extends into the realm of conservation biology and climate change, as its native habitats are often sensitive to environmental disruptions. Studies have focused on how populations of *Boechea* respond to factors such as drought, temperature fluctuations, and other climatic variables, providing insights into the resilience and adaptability of plant species under global change scenarios (Lin, Mitchell-Olds, and

Lee 2021; Pierce, Halter, and Waters 2020; Gallas and Waters 2015; Bemmels and Anderson 2019). Despite being less well-studied than the model organism *Arabidopsis*, *Boechera* offers a unique perspective on evolutionary dynamics due to its complex genetic structure and the ecological diversity of its habitats. This makes it especially useful for studying evolutionary and ecological genetics, adaptation strategies, and speciation mechanisms (Song and Mitchell-Olds 2007; Lee and Mitchell-Olds 2013). Additionally, recent genomic advances have positioned *Boechera* as a model in addressing both fundamental and complex biological questions in natural and experimental settings, thus bridging gaps left by traditional model organisms which lack ecological and evolutionary variability (Aliyu, Schranz, and Sharbel 2010; Knight et al. 2006; Sharbel et al. 2010).

In the context of understanding adaptive mechanisms in plants, particularly under the stress of increasing temperatures due to climate change, the species *Boechera arcuata* and *Boechera perennans* offer significant insights. These two *Boechera* species are of considerable interest due to their distinct ecological niches and remarkable heat tolerance. Relying on older identifications of *Boechera* species and habitat distributions are challenging because of the high rate of false species identifications within the literature. The seed sources for this study were genetically identified using genetic analysis and have both been confirmed to both be diploid species (F.-W. Li et al. 2017). *Boechera arcuata* is found from the coast up to the summits of the Peninsular ranges in Southern California (Figure 2.1). The *B. arcuata* seed source sequenced here was collected in the Angeles National Forest. *Boechera perennans* is found on the east-side of the Peninsular Ranges into the deserts of California and Arizona (Figure 2.1). The *B. perennans* seed source

sequenced in this study was collected in the Anza Borrego Desert. Studies have elucidated that both *B. arcuata* and *B. perennans* exhibit significantly higher tolerance to heat stress compared to *A. thaliana* (Halter et al. 2017). Notably, *B. arcuata* has been identified as the more thermotolerant species of the two, suggesting a highly refined set of evolutionary adaptations that allow it to maintain cellular stability and continue metabolic functions even at elevated temperatures. This enhanced thermotolerance could be attributed to a variety of physiological and molecular mechanisms, such as increased expression of heat-shock proteins, improved efficiency of photosynthetic processes under thermal stress, or more robust cellular membrane stability which prevents denaturation of cellular components.

Here I elucidate the adaptive genomic features of *B. arcuata* and *B. perennans* by comparing their genomes to each other and to the well-studied model plant *A. thaliana*. I aim to identify specific gene family variations that confer environmental resilience, allowing these species to thrive in distinct ecological niches—low-elevation coastal areas for *B. arcuata* and desert and inland chaparral regions for *B. perennans*. Through synteny analysis, gene ontology (GO) enrichment, and phylogenetic relationships within the Brassicaceae family, this research seeks to understand the genetic basis of speciation, diversification, and adaptation. By integrating diverse genomic analyses, I establish a broad understanding of the similarities and differences in the genomes of these two *Boechera* species and unravel the patterns of genomic evolution that have influenced their divergence from *A. thaliana*.

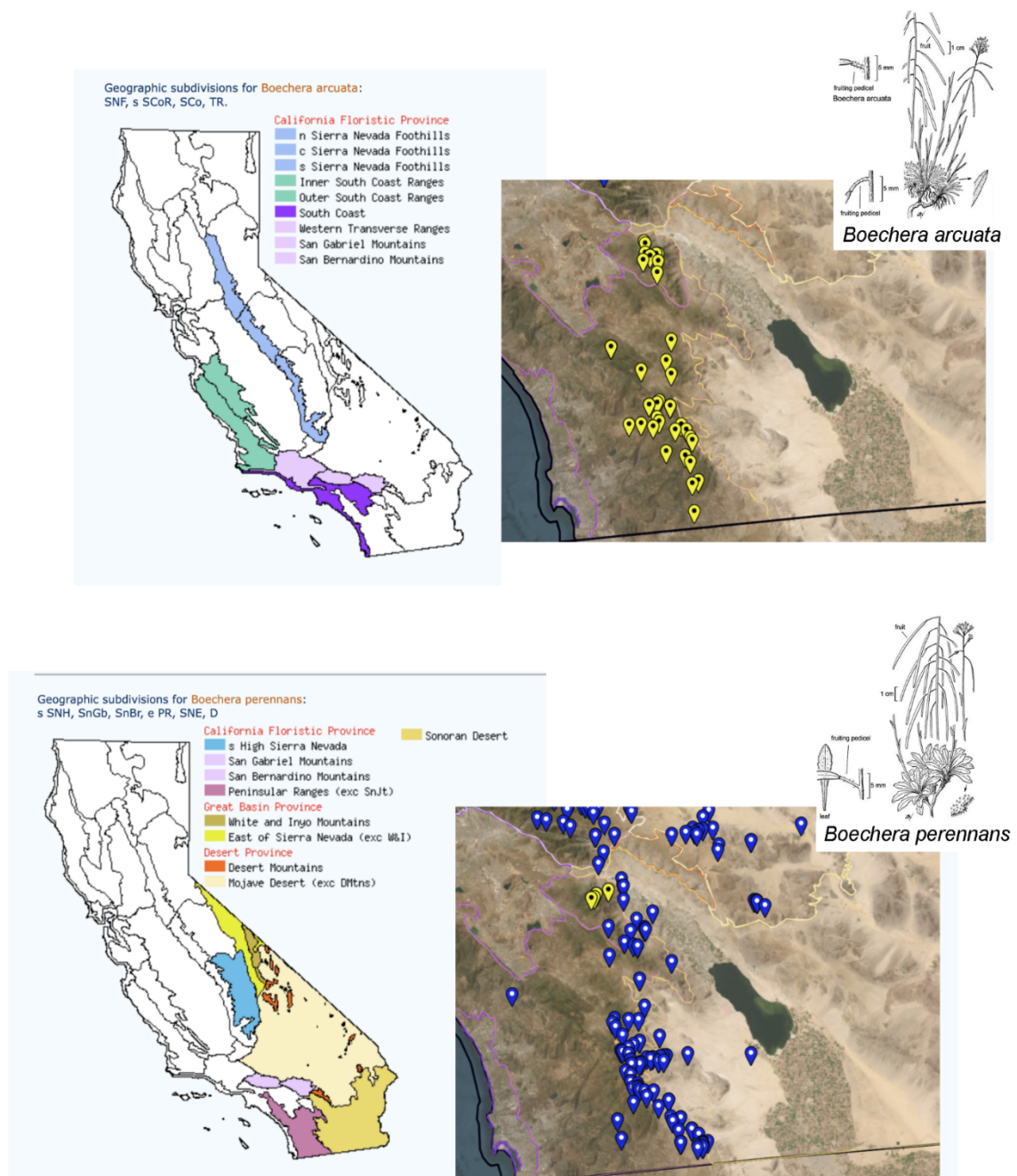


Figure 2.1. Distribution of *Boechera arcuata* and *Boechera perennans*.

Materials and Methods

Genomic data

The genome sequences of *Boechnera arcuata* and *Boechnera perennans* were obtained from the Joint Genome Institute (JGI), which provided chromosome-level assemblies and annotations. Assembly evaluation for each genome was performed using QUAST v5.0.0 (Gurevich et al. 2013). Genome annotation evaluation was performed using Funnannotate v1.7.4 (Palmer and Stajich 2017). The annotation results were summarized in custom code written in R v4.0.3 using the tidyverse v1.3.0 package (Wickham et al. 2019).

Transposable element content

The program EDTA v2.1.0 (Ou et al. 2022) was used to generate a curated library of transposable element sequences from the *Boechnera* genome assemblies using default parameters. This library was used to soft-mask these elements in the genome with RepeatMasker v4.1.4 (Smit 2004) (Table 2.2). The “calcDivergenceFromAlign.pl script” that is available in the util directory of RepeatMasker package was used to calculate and summarize the divergence from the consensus for each predicted repeat. This analysis was conducted using the Kimura 2-Parameter distance with an adjustment for higher mutation rates at CpG sites. The “createRepeatLandscape.pl” script was then used to create a repeat landscape plot, visualize the repeat elements diversity, and estimate their relative age and divergence using the divergence summary data (Figure 2.5).

Synteny analysis

To explore the syntenic relationships and genomic conservation among *B. arcuata*, *B. perennans*, and *A. thaliana*, synteny analysis was conducted using GENESPACE v1.2.0 (Lovell et al. 2022). GENESPACE utilizes genes as anchors to identify syntenic blocks, allowing for precise mapping of conserved genomic regions. After identifying syntenic blocks, I constructed the GENESPACE riparian plot to visually depict the genomic synteny across the *Boechera* species, using *A. thaliana* as the reference genome. This visual representation aids in understanding the conservation of gene order and content, as well as revealing evolutionary patterns. Further investigation of these syntenic blocks involves examining conserved gene functions and their potential evolutionary implications within and between the *Boechera* species.

Identification of gene family variation in *Boechera* and related species

Orthofinder v2.5.4 (Emms and Kelly 2019) was used as an initial step toward comparative genomics by identifying gene families unique to each *Boechera* species. *A. thaliana* was used as the reference in this analysis to find the shared and unique genes in the two genera. Orthofinder visualizations were rendered with UpsetR v1.4.0 (Conway, Lex, and Gehlenborg 2017). Gene Ontology (GO) enrichment analysis was performed on sets of genes including the *B. arcuata* and *B. perennans* species-specific genes and the identified sets of expanded genes in each species to test for enriched GO terms, using the R package GOSTats v2.58.0 (Falcon & Gentleman 2007). GO assignments were derived from the annotation data specific to each *B. arcuata* and *B. perennans* species. A *p*-value

= 0.01 cutoff was used to only report significant enrichment GO categories. This analysis was performed for each of the three GO classes: biological process (BP), molecular functions (MF), and cellular components (CC).

I filtered the Orthofinder results to orthogroups that were at least $|\log_2 \text{fold change}| > 1$ higher in copy number in *B. arcuata* than in *B. perennans* or *A. thaliana*. A similar approach was applied to identify orthogroups that were at least $|\log_2 \text{fold change}| > 1$ higher in copy number in *B. perennans* compared to either *B. arcuata* or *A. thaliana*, ensuring *A. thaliana* had non-zero counts. I then investigated the functions of the *A. thaliana* genes present in each relevant orthogroup on *Arabidopsis* Information Resource (TAIR) to determine their roles, focusing specifically on genes related to photosynthesis and abiotic stress response. Through this targeted approach, I narrowed down the list of genes of interest. The final selection of genes was then visualized using the ggplot2 package v3.4.4 (Wickham 2016) in R.

Orthogroup expansions of interest were confirmed through phylogenetic analyses. Briefly, I utilized genomic data from *Arabidopsis lyrata*, *Arabidopsis halleri*, *Carica papaya*, *A. thaliana*, and the two *Boechera* species. Orthofinder was employed to identify relevant homologous genes. Subsequently, I performed multiple sequence alignments of the genes within these orthogroups using MUSCLE v5.1.1 (Edgar 2004).

To examine gene family contraction and expansion, I applied CAFE v5.0.0 (Mendes et al. 2021). This approach estimates the likelihood of observed gene family sizes by accounting for average rates of gene gain and loss across the phylogenetic tree, thus

identifying significant expansions or contractions of gene families at each node. For this analysis, I used Orthofinder to determine the number of genes in each gene family. Orthofinder also provided a rooted species tree in Newick format, depicting the phylogenetic relationships of the genes within each orthogroup. To ensure the tree was suitable for CAFE analysis, I converted the Orthofinder species tree into an ultrametric tree using ETE3. An ultrametric tree assumes that all tips are equidistant from the root, which is necessary for the accurate estimation of evolutionary rates. This transformation involved setting an appropriate outgroup (*C. papaya*) and adjusting branch lengths to ensure all terminal nodes were equidistant from the root. The resulting ultrametric tree, along with gene family counts from Orthofinder, was then used as input for CAFE v5.0.0. By analyzing gene family sizes within this phylogenetic framework, CAFE identified significant expansions and contractions of gene families, providing insights into the evolutionary dynamics within the *Boechera* species and their relatives.

Identification of ancient Whole Genome Duplication events

For examining the history and molecular dating of whole-genome duplications (WGD) in the *Boechera* species and *A. thaliana*, I used wgd v.2.0.23 (Zwaenepoel and Van de Peer 2018; Zwaenepoel, n.d.) using CDS sequences.. Initially, I identified the whole paranome, which involves clustering genes across the entire genome in each species based on sequence similarity. This step ensures that all potential duplicated genes are included, providing a comprehensive overview of the genomic landscape. Following this, I conducted pairwise comparisons of genes within each cluster to estimate synonymous

substitution rates (Ks) for these gene pairs. Thus the Ks distribution are as a marker for identifying and dating past WGD events, with peaks in the distribution indicating potential ancient duplication events. To further enhance this analysis, I performed synteny analysis, which involved aligning genes and identifying collinear blocks. Synteny analysis is crucial for confirming Ks-based evidence of WGDs by revealing the physical arrangement of duplicated gene blocks across the genome. By identifying collinear blocks, it is possible to distinguish between segmental duplications and whole-genome duplications, as well as to infer the evolutionary history and the timing of these events.

Results

***B. arcuata* and *B. perennans* genomes are similar in size, number of chromosomes, and number of genes**

The assembly of the two *Boechera* genomes indicated they are highly similar in genome organization. The genomic analyses of *B. arcuata* and *B. perennans*, using data generated by the Joint Genome Institute (JGI), yielded comprehensive genome statistics summarized in Table 2.1. The *B. arcuata* genome assembly consists of 7 contigs which represents the seven chromosomes. The largest contig is approximately 36.195 Mb. The *B. perennans* genome was composed of 157 contigs, including 7 chromosomes and 152 scaffolds, with the largest being about 33.584 Mb. The total length of the *B. arcuata* genome was around 231.7 Mb, with a GC content of 35.67%, while *B. perennans* had a slightly smaller genome of approximately 219.87 Mb with a GC content of 35.78%. The

N50 statistic was 32.689 Mb for *B. arcuata* and 30.248 Mb for *B. perennans*, indicating robust assemblies for both species.

The similarity among these two genomes was also clearly observed when repeats and gene models were examined. Repeat analysis showed that repetitive regions make up 44.76% of *B. arcuata* genome and 42.3% of *B. perennans* genome. Genome annotation evaluation revealed a total of 27,156 and 28,159 gene models for *B. arcuata* and *B. perennans*, respectively. The number of mRNA genes identified was 45,943 for *B. arcuata* and 49,237 for *B. perennans* (Table 2.1). A higher number of mRNA transcripts compared to gene models indicates alternative splicing mRNA variants. I also provided a visual comparison of genome size, intergenic regions, genic regions, and repeats (Figure 2.2). *B. arcuata* was shown to have larger genome size, repeat, and intergenic regions compared to *B. perennans*, while *B. perennans* had marginally higher genic region.

Table 2.1. Genome assembly and annotation statistics. Genome statistics were calculated using QUAST, RepeatMasker, and Funannotate.

Software	Statistic	<i>B. arcuata</i>	<i>B. perennans</i>
QUAST	# contigs (≥ 0 bp)	7	157
	# chromosomes	7	7
	Largest contig	36195003	33584199
	Total length (≥ 0 bp)	231707436	219876689
	GC (%)	35.67	35.78
	N50	32689921	30248120
	N90	31050160	27715230
	L50	4	4
	L90	7	7
	# N's per 100 kbp	4.32	27.29
RepeatMasker	Repetitive regions (%)	44.76	42.3
Funannotate	Total number of gene models	27156	28159
	Number of mRNA genes	45943	49237
	Mean gene length	2535.55	2510.15
	Total exons	393180	433533
	Mean exon length	266.5	267.64
	Mean protein length	458.61	468.93

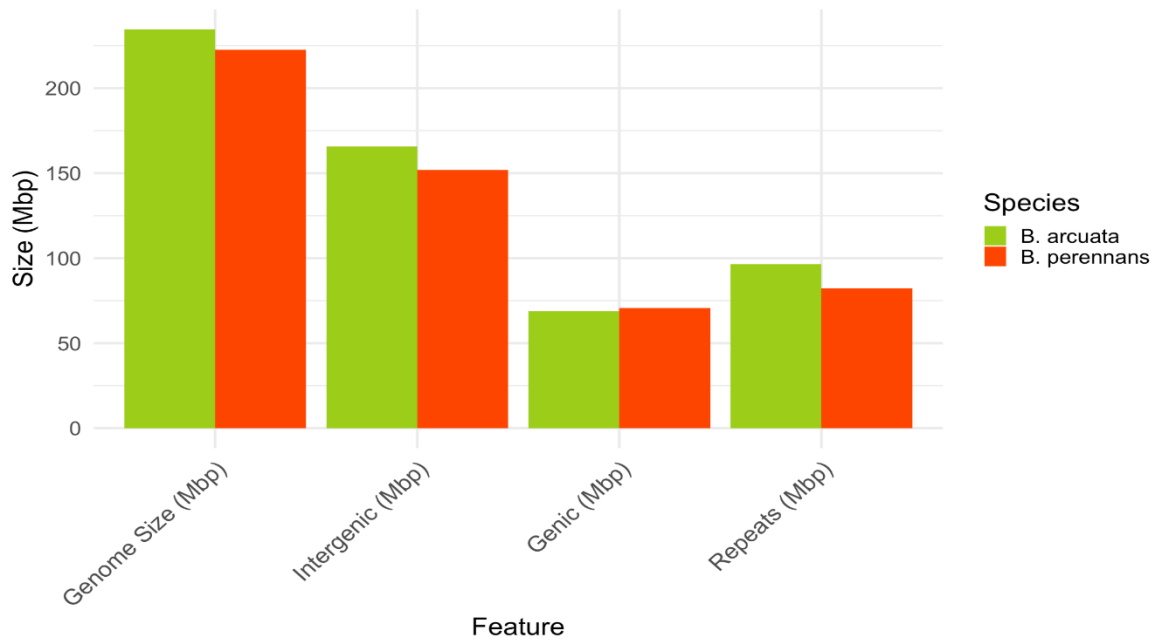


Figure 2.2. Genomic features of *B. arcuata* and *B. perennans*. A barplot is shown depicting the length in Mbp of overall length, intergenic, genic, and repeat content between *B. arcuata* (green) and *B. perennans* (orange).

***Boechera* species share extensive synteny across homologous chromosomes**

The synteny analysis results revealed the chromosomal relationships and genomic linkages within and between the two *Boechera* species, and between the *Boechera* species and *A. thaliana*. The synteny between *B. arcuata* and *B. perennans* show extensive conservation of genomic segments between the two species (Figure 2.3A). The syntenic blocks between the chromosomes of the two *Boechera* species studied here and those of *A. thaliana*, revealed significant chromosomal rearrangements (Figure 2.3B). While some *Boechera* chromosomes showed a one-to-one correspondence with *A. thaliana* chromosomes, others exhibit a more complex pattern, with segments scattered across multiple *A. thaliana* chromosomes, indicative of significant recombination, duplication, and rearrangement events over time. Both *Boechera* chromosome 1 and chromosome 2

share syntenic segments with chromosome 1 in *A. thaliana*. Notably these relationships seem more complex than a simple fission event. *Boechera* chromosome 4 was entirely composed of sequences corresponding to *A. thaliana* chromosome 2, although not all segments of *A. thaliana* chromosome 2 were located on *Boechera* chromosome 4, and some *A. thaliana* chromosome 2 fragments were also found on *Boechera* chromosomes 3 and 5. *Arabidopsis thaliana* chromosome 3 was split between *Boechera* chromosomes 3 and 5. Similarly, segments of *A. thaliana* chromosome 4 were distributed across *Boechera* chromosomes 6 and 7. Finally, segments of *A. thaliana* chromosome 5 were distributed across *Boechera* chromosomes 3, 5, 6, and 7, demonstrating the complex and dynamic evolutionary relationships among these species.

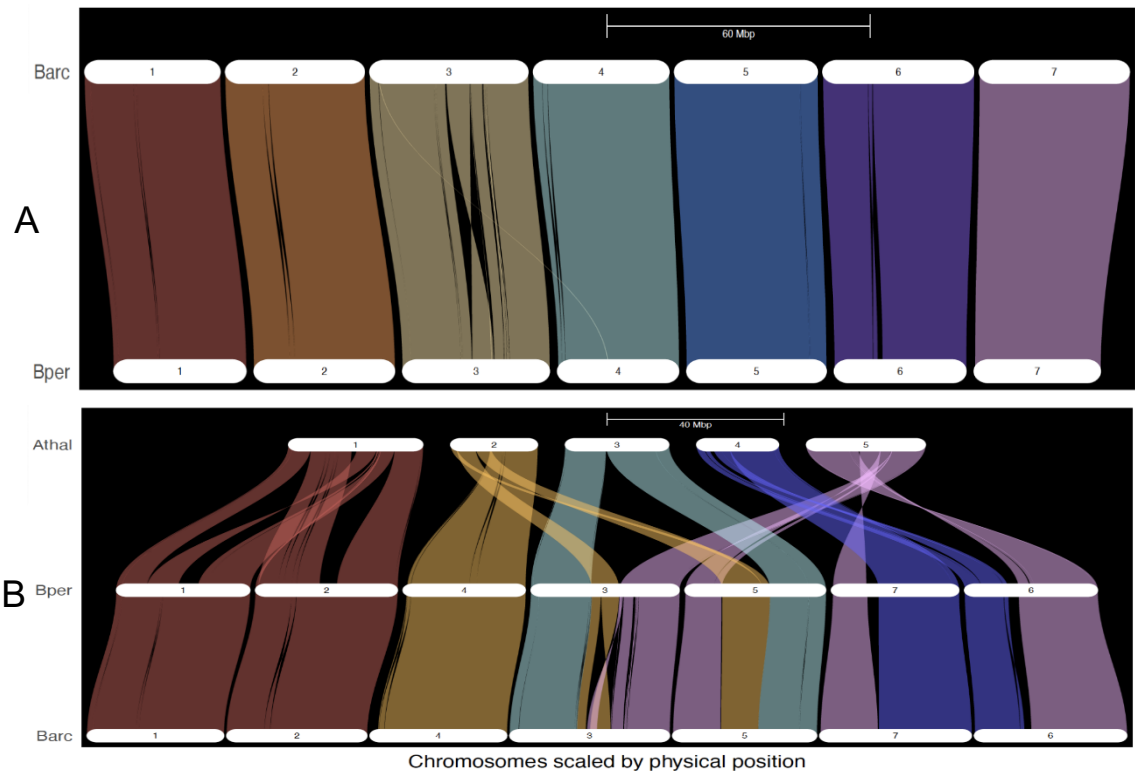


Figure 2.3. Synteny analysis of the *B. arcuata* and *B. perennans* genome compared to each other and *A. thaliana*. The Riparian plot generated by GENESPACE v1.2.3 (A) *B. arcuata* and *B. perennans*, (B) *A. thaliana*, *B. perennans*, and *B. arcuata*. Genomes are organized by row and chromosomes are organized by column. Linkages are color-coded by their respective chromosomes in the other taxa.

Divergent duplication histories and segmental duplication patterns in *A. thaliana* and *Boechera* Species

Analysis of whole-genome duplication (WGD) analysis revealed distinct Ks distributions for *B. arcuata*, *B. perennans*, and *A. thaliana* which reflected the duplication history within each genome (Figure 2.4A). The Ks distribution based on paralogous genes results for *B. arcuata* and *B. perennans*, as compared to *A. thaliana*, demonstrated the

frequency of gene duplications across different synonymous substitution rate categories. In *A. thaliana*, I observed two peaks in the Ks distribution. The first peak, representing a larger number of duplications, was located at a lower Ks value, indicative of more recent duplication events. The second, smaller peak occurred at a higher Ks value, suggesting older duplication events. In *B. arcuata* and *B. perennans*, the Ks histogram also revealed a notable number of duplications, with two distinct peaks that suggested an ancient WGD event (evidenced by the peak with higher Ks values) and a more recent one within each species. The more recent duplication events in both *Boechea* species, indicated by the peaks at lower Ks values, were smaller compared to the recent duplication event peak in *A. thaliana*. Both *Boechea* species displayed a similar bimodal distribution with slightly different Ks values, suggesting that both *Boechea* species share a common WGD history. The presence of very recent duplication events, as indicated by the low Ks peaks, may seem unexpected. However, similar patterns have been reported in other plant species, suggesting that these recent duplications are part of a broader phenomenon observed in plant genomes. For instance, a recent study on *Huperzia asiatica* and *Diphasiastrum complanatum* also identified very recent duplications in the form of low Ks values (Li et al. 2024). I also demonstrated the frequency of segmental duplications categorized by their collinearity ratios for each genome (Figure 2.4B). I observed a higher frequency at lower collinearity ratios (1:1, 2:2) in all species, indicating a prevalence of common smaller-scale duplication events. The frequency of duplications decreased with higher collinearity ratios, possibly due to the complexity and rarity of maintaining larger duplicated segments through evolutionary history. While the overall trend was similar, the specific frequencies and extent of the decrease in duplication events at higher collinearity ratios were varied. For instance, *A.*

thaliana extended to a higher range (up to 9:9) compared to *B. arcuata* (up to 8:8), and *B. perennans* (7:7), possibly indicating a greater complexity or a longer evolutionary history of duplications that were maintained within its genome.

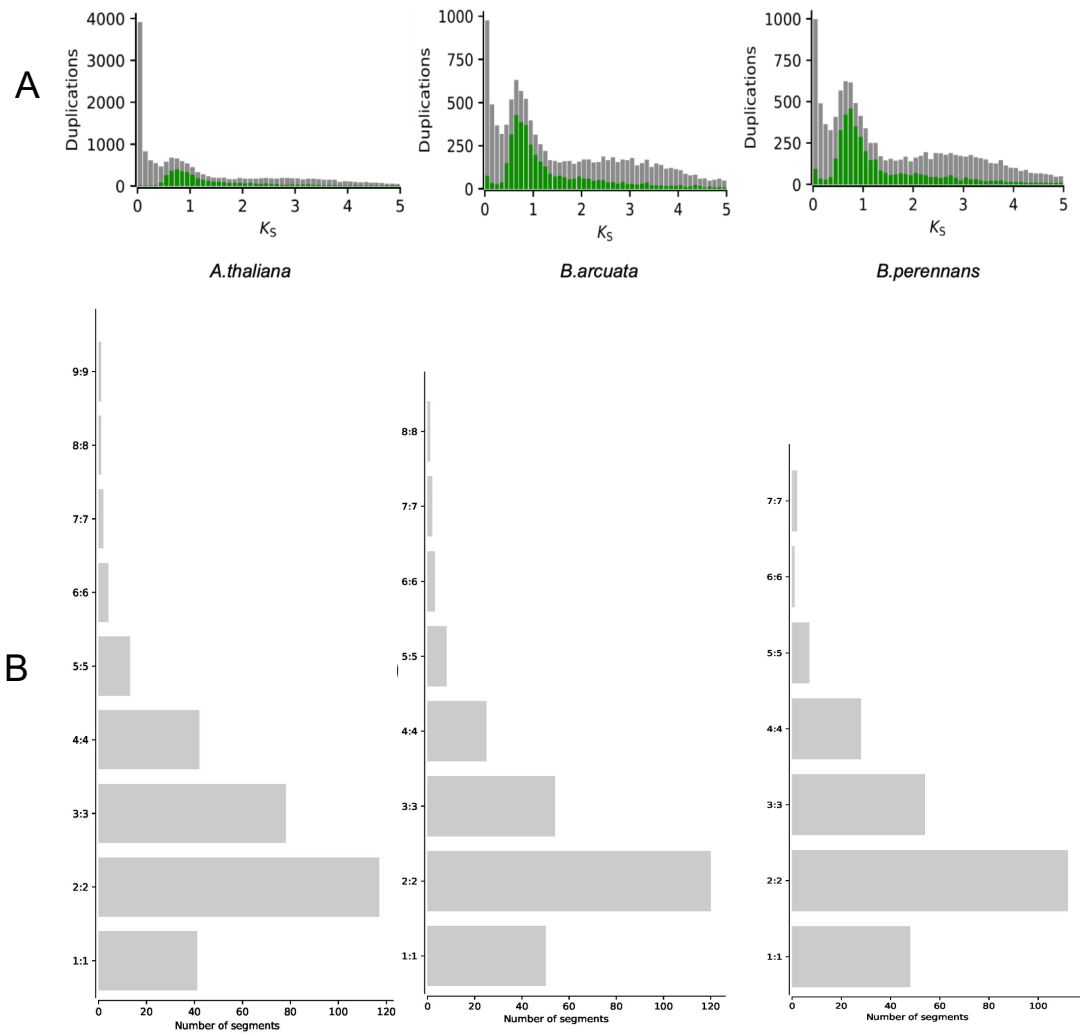


Figure 2.4. Ks analysis identifies possible WGD events (A) The constructed whole paranome Ks age distribution of *A. thaliana*, *B. arcuata*, and *B. perennans*. The plots show an observation for potentially conceivable WGD in each species. The retained anchor pairs with green color validate a WGD event. (B) Syndepth plots showing the number of duplicated segments in each species.

Repeat landscape of *Boechera* genomes demonstrates multiple possible repeat expansion events

Analysis of repeats indicated that both *Boechera* genomes have similar levels of repeated elements (Table 2.2). In *B. arcuata*, 44.76 % of the genome was identified and masked as repetitive elements, and in *B. perennans*, 42.3% of the genome was made up of repeats. In *B. arcuata*, LTR elements comprised 23.22% of the genome, with *Copia* and *Gypsy* elements representing 6.78% and 13.38%, respectively, and DNA elements accounted for 11.27%. In comparison, *B. perennans* had a slightly lower percentage of repeats. The percentage of LTR elements was 18.73%, with *Copia* at 5.58% and *Gypsy* elements at 10.54%. DNA elements were more abundant in *B. perennans*, making up 13.81% of the genome.

Table 2.2. Repeat content for *B. arcuata* and *B. perennans* genomes. Repeat type and content as determined by RepeatMasker.

Type of repeat element	Percent of Genome	
	<i>B. arcuata</i>	<i>B. perennans</i>
SINES	0	0
LINES	0.17	0.2
LTR elements	23.22	18.73
<i>Ty1/Copia</i>	6.78	5.58
<i>Gypsy</i>	13.38	10.54
DNA elements	11.27	13.81
Unclassified	8.8	8.24
Total interspersed repeats	44.76	42.3
Simple repeats	1.04	1.06
Low complexity	0.26	0.27

Examining the repeat landscapes of both *Boechera* species depicted the diversity of repeats and their evolutionary dynamics, indicating the presence of both ancient and recent repeat insertions in these genomes (Figure 2.5). The landscapes revealed patterns that suggest multiple expansion events. In *B. arcuata*, the landscape pointed to four possible expansion events, three that appeared to be ancient and one that seems to be more recent. This suggests an ongoing and dynamic repeat accumulation process. In *B. perennans*, the repeat landscape indicated at least three distinct expansion events over time. Notably, in addition to expansions of the LTR retrotransposons, there was clear evidence of DNA transposon activity, suggesting bursts of both Type I and Type II transposition. I also observed a recent expansion of repeats designated as “unknown” (i.e., these repeats are not members of previously characterized TE families) in *B. perennans* that were not shared with *B. arcuata*. The study found that the expansion of TEs in *Rhizophagus irregularis* strains varies, with some strains showing significant recent bursts of TE activity. These findings are similar to those seen in *Boechera* species, suggesting that recent TE activity could be a broader trend observed in other organisms as well (Oliveira and Corradi 2024).

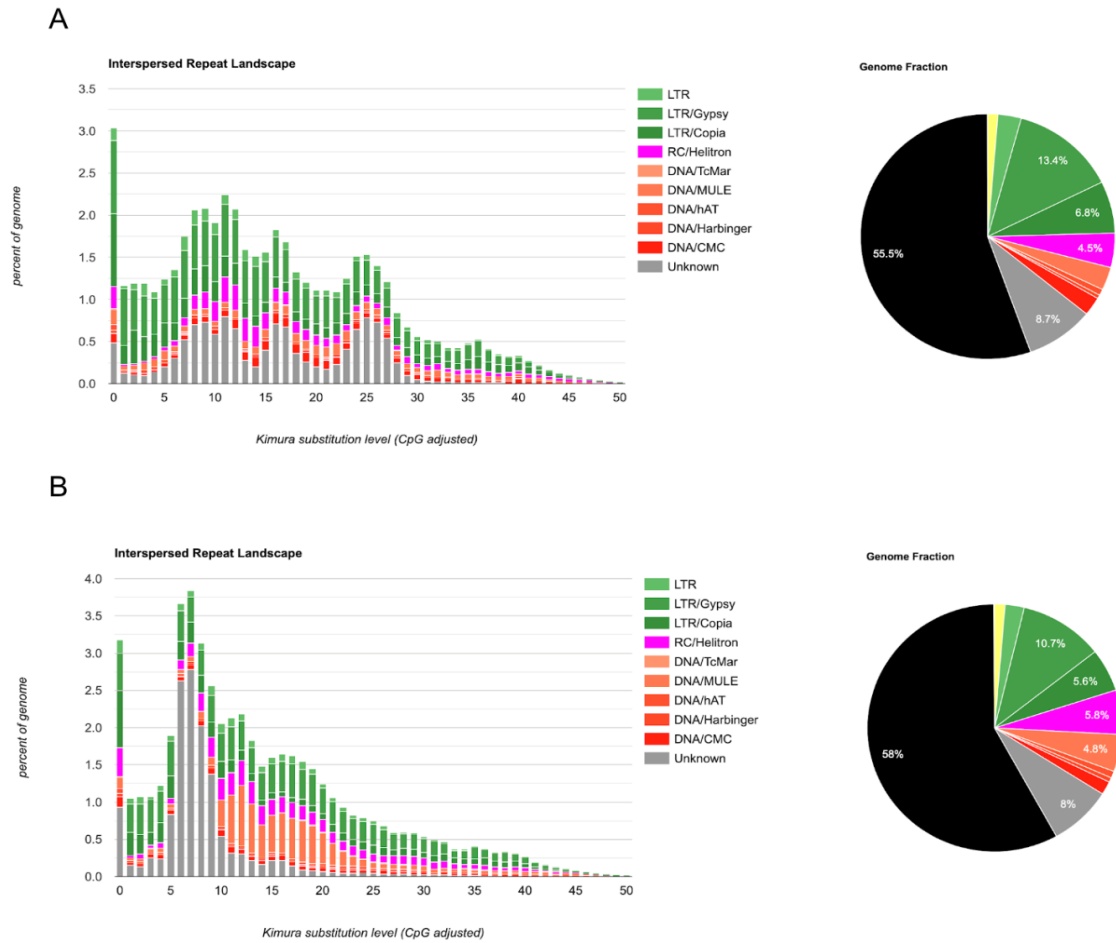


Figure 2.5. Interspersed repeat landscape. The interspersed repeat landscape, revealing the distribution and divergence of repeat elements across the genomes of (A) *B. arcuata*, (B) *B. perennans*. The percentages of the genome composed of various repeat elements (Y-axis) are categorized according to their Kimura substitution level (X-axis; CpG adjusted values ranging from 0 to 50), each bar in the plot represents a 1% sequence divergence bin, adjusted for CpG Kimura divergence. The landscape is represented by a stacked bar plot, where each color corresponds to a specific class of repeats. The plot illustrates the copy-divergence analysis of these elements, with older copies indicated by higher Kimura values towards the right, and more recent copies towards the left, indicating the evolutionary dynamics of the repetitive elements within the genome.

Identification of orthologous groups and gene families reveal both shared and species-specific patterns of gene family evolution

Analysis of *Boechera* orthologous groups with *A. thaliana* indicated the presence of unique orthogroups in each species and unsurprisingly a large number of shared orthogroups across *Arabidopsis* and *Boechera* species (Figure 2.7). I identified 19,208 orthogroups shared among all three species. Furthermore, I found 1,375 orthogroups shared between the two *Boechera* species which were absent in *A. thaliana*. Further analysis revealed that *A. thaliana* shares 171 orthogroups with *B. arcuata* and 310 orthogroups with *B. perennans*. Additionally, *A. thaliana* exhibited a total of 228 specific orthogroups that were not found in either of the two *Boechera* species. I identified 109 unique orthogroups in *B. arcuata*, and 138 unique orthogroups in *B. perennans*.

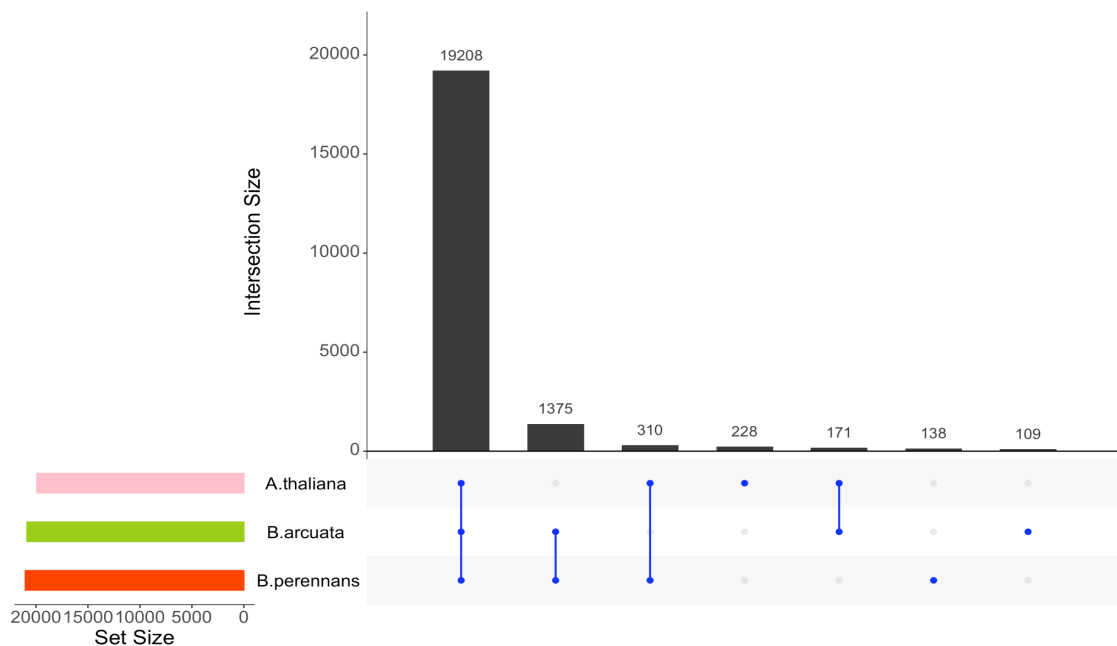


Figure 2.6. Identification of orthologous groups shared across genomes and those that are unique to each genome. Dots and lines in bottom right indicate unique or shared membership of orthogroups among species; vertical columns indicate total orthogroup counts for those categories. Colored lateral columns indicate total counts of orthogroups in each set. Numbers above bars represent the number of orthogroups shared by a given intersection of species.

GO enrichment analysis of the species-unique orthogroups reveals species specific genome evolution

GO enrichment analysis of *B. arcuata* specific genes, identified 34 GO terms. Of these 34 GO terms 14 are related to biological processes (BP), five are cellular compartments terms (CC), and 15 molecular functions (MF) GO terms (Table 2.3, Figure 2.7A). Six GO terms were related to protein assembly and biosynthesis in the BP category, involving processes such as protein homooligomerization (GO:0051260), protein complex oligomerization (GO:0051259), molybdopterin cofactor biosynthetic process

(GO:0032324), cellular component biogenesis (GO:0044085), DNA integration (GO:0015074), and tubulin complex assembly (GO:0007021). Additionally, four GO terms were associated with various transport mechanisms (GO:0006865: Amino acid transport, GO:0034220: Ion transmembrane transport, GO:0015696: Ammonium transport, GO:0072488: Ammonium transmembrane transport), and two GO terms highlighted critical energy-related processes (GO:0022900: Electron transport chain, GO:0006430: Lysyl-tRNA aminoacylation). In the CC category, five GO terms were related to essential cellular structures (GO:0005739: Mitochondrion, GO:0009521: Photosystem, GO:0034357: Photosynthetic membrane, GO:0009579: Thylakoid, GO:0015935: Small ribosomal subunit) reflecting the significant involvement of these components in cellular functionality. The MF category included 13 GO terms describing specific enzymatic and binding activities crucial for cellular metabolism and regulation. These included sulfotransferase activity (GO:0008146), iron ion binding (GO:0005506), various transferase activities (GO:0016782: Transferring sulfur-containing groups, GO:0008883: Glutamyl-tRNA reductase activity), and binding activities related to genetic material and molecules critical for electron transport (GO:0020037: Heme binding, GO:0046906: Tetrapyrrole binding, GO:0003690: Double-stranded DNA binding, GO:0016705: Oxidoreductase activity). Additional functionalities such as amino acid and organic compound transport activities were also identified (GO:0015171: Amino acid transmembrane transporter activity, GO:0046943: Carboxylic acid transmembrane transporter activity, GO:0005342: Organic acid transmembrane transporter activity), along with pre-mRNA processing functions (GO:0045131: Pre-mRNA branch point binding,

GO:0036002: Pre-mRNA binding) and a specific activity related to tRNA (GO:0004824: Lysine-tRNA ligase activity).

In *B. perennans* species-specific genes GO terms from 23 biological processes (BP), 34 cellular compartments (CC), and 33 molecular functions (MF) were identified (Table 4, Figure 2.7B). In the BP category, five GO terms were related to energy production and photosynthetic processes (GO:0015979: Photosynthesis, GO:0019684: Photosynthesis, light reaction, GO:0006091: Generation of precursor metabolites and energy, GO:0042775: Mitochondrial ATP synthesis coupled electron transport, GO:0009765: Photosynthesis, light harvesting). Moreover, seven GO terms were associated with response mechanisms (GO:0006952: Defense response, GO:0050896: Response to stimulus, GO:0006950: Response to stress, GO:0009733: Response to auxin, GO:0007165: Signal transduction, GO:0023052: Signaling, GO:0031647: Regulation of protein stability). In the CC category, 10 GO terms were associated with photosynthetic and respiratory components (GO:0009579: Thylakoid, GO:0009521: Photosystem, GO:0009523: Photosystem II, GO:0034357: Photosynthetic membrane, GO:0005746: Mitochondrial respirasome, GO:0005750: Mitochondrial respiratory chain complex III, GO:0005743: Mitochondrial inner membrane, GO:0098800: Inner mitochondrial membrane protein complex, GO:0045275: Respiratory chain complex III, GO:0009512: Cytochrome b6f complex), six GO terms related to various mitochondrial structures (GO:0005739: Mitochondrion, GO:0005740: Mitochondrial envelope, GO:0031966: Mitochondrial membrane, GO:0031967: Organelle envelope, GO:0031975: Envelope, GO:0098798: Mitochondrial protein-containing complex). Finally, in the MF category, six

GO terms described activities related to oxidoreductase and transferase processes (GO:0009055: Electron transfer activity, GO:0015036: Disulfide oxidoreductase activity, GO:0016651: Oxidoreductase activity, acting on NAD(P)H, GO:0016668: Oxidoreductase activity, acting on a sulfur group of donors, NAD(P) as acceptor, GO:0047134: Protein-disulfide reductase activity, GO:0016667: Oxidoreductase activity, acting on a sulfur group of donors). Also, four GO terms highlighted activities associated with protein and DNA interactions (GO:0003917: DNA topoisomerase type I (single strand cut, ATP-independent) activity, GO:0003916: DNA topoisomerase activity, GO:0019843: rRNA binding, GO:0042301: Phosphate ion binding).

Table 2.3. Enriched GO terms for species-specific *B. arcuata* genes. Enrichment analysis identified GO terms that were significantly enriched in *B. arcuata*. ($p < 0.05$). These terms were summarized and filtered with Revigo. Here I provide the GO ID, P-value, count, GO term, and GO category (BP: biological process, CC: cellular component, MF: molecular function) for each enriched term.

	GOID	P value	Count	Term	Category
1	GO:0051260	0.003	2	protein homooligomerization	BP
2	GO:0051259	0.0036	2	protein complex oligomerization	BP
3	GO:0032324	0.0042	1	molybdopterin cofactor biosynthetic process	BP
4	GO:0044085	0.012	6	cellular component biogenesis	BP
5	GO:0015074	0.013	1	DNA integration	BP
6	GO:0007021	0.017	1	tubulin complex assembly	BP
7	GO:0006865	0.018	2	amino acid transport	BP
8	GO:0034220	0.019	4	ion transmembrane transport	BP
9	GO:0015696	0.021	1	ammonium transport	BP
10	GO:0072488	0.021	1	ammonium transmembrane transport	BP
11	GO:0022900	0.022	2	electron transport chain	BP
12	GO:0006430	0.03	1	lysyl-tRNA aminoacylation	BP
13	GO:0006397	0.046	3	mRNA processing	BP
14	GO:0010467	0.048	21	gene expression	BP
15	GO:0005739	0.0015	5	mitochondrion	CC
16	GO:0009521	0.0027	3	photosystem	CC
17	GO:0034357	0.003	3	photosynthetic membrane	CC

18	GO:0009579	0.0034	3	thylakoid	CC
19	GO:0015935	0.016	2	small ribosomal subunit	CC
20	GO:0008146	0.00021	3	sulfotransferase activity	MF
21	GO:0005506	0.00037	10	iron ion binding	MF
22	GO:0016782	0.0008	3	transferase activity, transferring sulfur-containing groups	MF
23	GO:0020037	0.0014	9	heme binding	MF
24	GO:0046906	0.0015	9	tetrapyrrole binding	MF
25	GO:0003690	0.0036	4	double-stranded DNA binding	MF
26	GO:0016705	0.012	7	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	MF
27	GO:0015171	0.015	2	amino acid transmembrane transporter activity	MF
28	GO:0046943	0.016	2	carboxylic acid transmembrane transporter activity	MF
29	GO:0005342	0.016	2	organic acid transmembrane transporter activity	MF
30	GO:0008883	0.023	1	glutamyl-tRNA reductase activity	MF
31	GO:0008519	0.029	1	ammonium transmembrane transporter activity	MF
32	GO:0045131	0.035	1	pre-mRNA branch point binding	MF
33	GO:0036002	0.035	1	pre-mRNA binding	MF
34	GO:0004824	0.04	1	lysine-tRNA ligase activity	MF

Table 2.4. Enriched GO terms for genes that are species-specific in *B. perennans*. Enrichment analysis identified GO terms that were significantly enriched in *B. perennans*. ($p < 0.01$). These terms were summarized and filtered with Revigo. Here I provide the GO ID, P-value, count, GO term, and GO category (BP: biological process, CC: cellular component, MF: molecular function) for each enriched term.

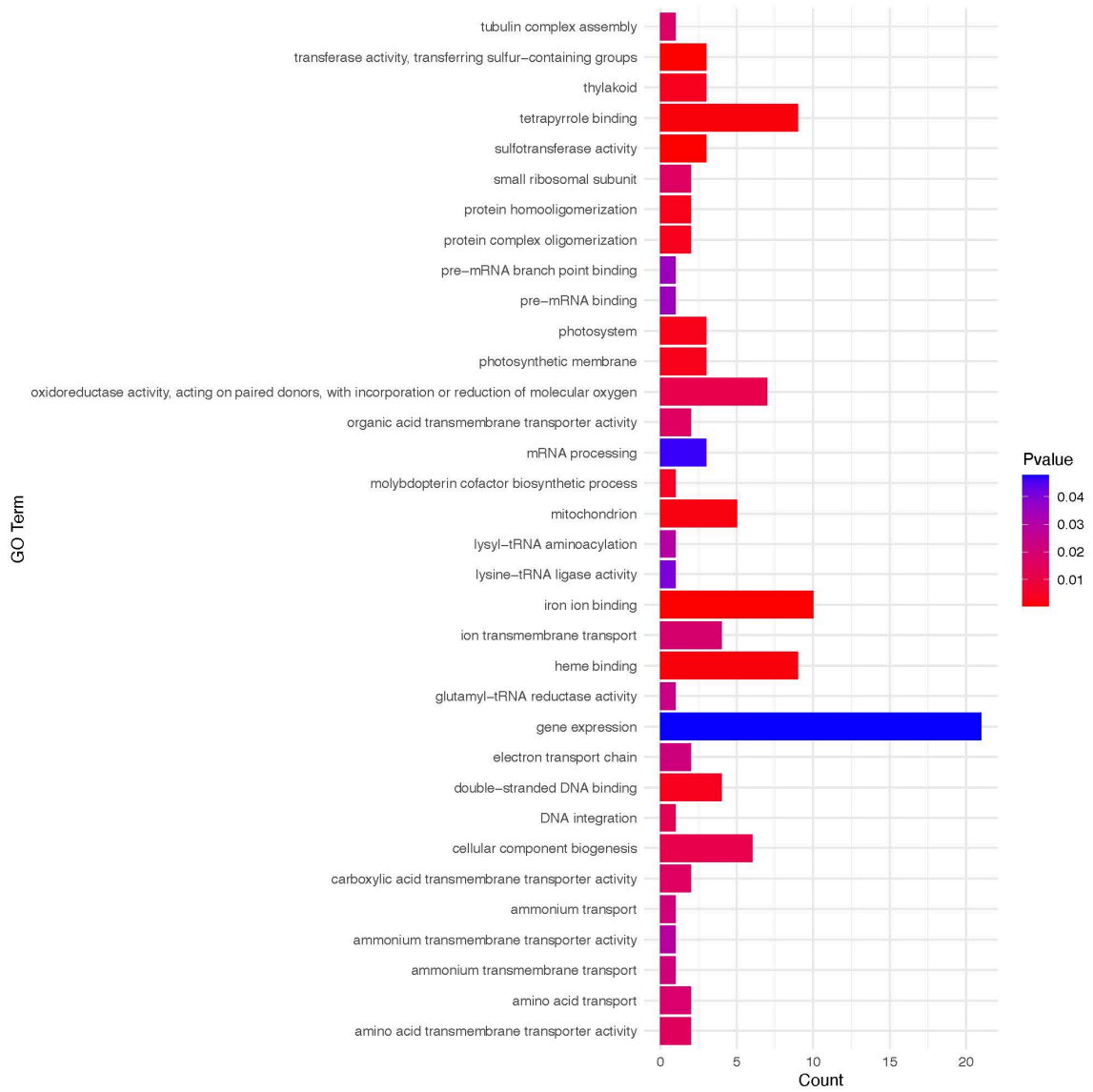
	GOID	P value	Count	Term	Category
1	GO:0006952	4.42 e-10	15	defense response	BP
2	GO:0006412	1.19 e-06	21	translation	BP
3	GO:0043603	1.63e-05	21	cellular amide metabolic process	BP
4	GO:0015979	7.86 e-05	7	photosynthesis	BP
5	GO:0050896	0.00017	35	response to stimulus	BP
6	GO:0042775	0.0014	3	mitochondrial ATP synthesis coupled electron transport	BP
7	GO:1901566	0.0021	25	organonitrogen compound biosynthetic process	BP
8	GO:0006950	0.0032	16	response to stress	BP
9	GO:0051260	0.0058	2	protein homooligomerization	BP
10	GO:0051259	0.0066	2	protein complex oligomerization	BP
11	GO:0019684	0.0072	3	photosynthesis, light reaction	BP
12	GO:0006004	0.0077	1	fucose metabolic process	BP
13	GO:0050821	0.0077	1	protein stabilization	BP
14	GO:0031647	0.0077	1	regulation of protein stability	BP
15	GO:0006091	0.0097	9	generation of precursor metabolites and energy	BP
16	GO:0009765	0.017	2	photosynthesis, light harvesting	BP
17	GO:0009164	0.023	1	nucleoside catabolic process	BP
18	GO:0072529	0.023	1	pyrimidine-containing compound catabolic process	BP
19	GO:0006367	0.03	2	transcription initiation from RNA polymerase II promoter	BP
20	GO:0009733	0.032	3	response to auxin	BP
21	GO:0007165	0.044	15	signal transduction	BP
22	GO:0023052	0.044	15	signaling	BP
23	GO:0034227	0.045	1	tRNA thio-modification	BP
24	GO:0005840	2.61 e-09	20	ribosome	CC
25	GO:0070069	5.86 e-06	5	cytochrome complex	CC
26	GO:0009579	6.35 e-05	6	thylakoid	CC

27	GO:0005750	0.00012	3	mitochondrial respiratory chain complex III	CC
28	GO:0070469	0.00017	4	respirasome	CC
29	GO:0031966	0.00018	6	mitochondrial membrane	CC
30	GO:0043228	0.00022	22	non-membrane-bounded organelle	CC
31	GO:0043232	0.00022	22	intracellular non-membrane-bounded organelle	CC
32	GO:0005740	0.00027	6	mitochondrial envelope	CC
33	GO:0005743	0.00032	5	mitochondrial inner membrane	CC
34	GO:0098800	0.00033	4	inner mitochondrial membrane protein complex	CC
35	GO:0019866	0.00034	5	organelle inner membrane	CC
36	GO:0098798	0.00043	4	mitochondrial protein-containing complex	CC
37	GO:0034357	0.00045	5	photosynthetic membrane	CC
38	GO:0005739	0.00051	8	mitochondrion	CC
39	GO:0005746	0.00055	3	mitochondrial respirasome	CC
40	GO:0045275	0.001	3	respiratory chain complex III	CC
41	GO:0009512	0.0012	2	cytochrome b6f complex	CC
42	GO:0031967	0.0014	6	organelle envelope	CC
43	GO:0031975	0.0014	6	envelope	CC
44	GO:0098803	0.0016	3	respiratory chain complex	CC
45	GO:0098796	0.0028	10	membrane protein complex	CC
46	GO:0009521	0.0027	4	photosystem	CC
47	GO:0005622	0.0037	49	intracellular anatomical structure	CC
48	GO:0009523	0.0051	3	photosystem II	CC
49	GO:0016602	0.0078	1	CCAAT-binding factor complex	CC
50	GO:0045261	0.011	2	proton-transporting ATP synthase complex, catalytic core F(1)	CC
51	GO:0031090	0.013	7	organelle membrane	CC
52	GO:0031360	0.015	1	intrinsic component of thylakoid membrane	CC
53	GO:0031361	0.015	1	integral component of thylakoid membrane	CC
54	GO:0043229	0.02	36	intracellular organelle	CC
55	GO:0043226	0.026	36	organelle	CC
56	GO:0009539	0.038	1	photosystem II reaction center	CC
57	GO:1990204	0.041	3	oxidoreductase complex	CC

58	GO:0005198	3.96 e-11	24	structural molecule activity	MF
59	GO:0003735	1.66 e-10	21	structural constituent of ribosome	MF
60	GO:0009055	4.96 e-05	12	electron transfer activity	MF
61	GO:0004185	0.001	4	serine-type carboxypeptidase activity	MF
62	GO:0003917	0.001	3	DNA topoisomerase type I (single strand cut, ATP-independent) activity	MF
63	GO:0070008	0.0014	4	serine-type exopeptidase activity	MF
64	GO:0004180	0.0016	4	carboxypeptidase activity	MF
65	GO:0008238	0.0048	4	exopeptidase activity	MF
66	GO:0015036	0.006	6	disulfide oxidoreductase activity	MF
67	GO:0003916	0.0061	3	DNA topoisomerase activity	MF
68	GO:0016651	0.0063	7	oxidoreductase activity, acting on NAD(P)H	MF
69	GO:0042301	0.0067	1	phosphate ion binding	MF
70	GO:0008736	0.0067	1	L-fucose isomerase activity	MF
71	GO:0004857	0.0077	5	enzyme inhibitor activity	MF
72	GO:0047134	0.01	4	protein-disulfide reductase activity	MF
73	GO:0016667	0.014	6	oxidoreductase activity, acting on a sulfur group of donors	MF
74	GO:0020037	0.014	9	heme binding	MF
75	GO:0015252	0.014	2	proton channel activity	MF
76	GO:0046933	0.014	2	proton-transporting ATP synthase activity, rotational mechanism	MF
77	GO:0046906	0.015	9	tetrapyrrole binding	MF
78	GO:0016668	0.015	4	oxidoreductase activity, acting on a sulfur group of donors, NAD(P) as acceptor	MF
79	GO:0098772	0.019	7	molecular function regulator	MF
80	GO:0004126	0.02	1	cytidine deaminase activity	MF
81	GO:0046961	0.025	2	proton-transporting ATPase activity, rotational mechanism	MF
82	GO:0044769	0.025	2	ATPase activity, coupled to transmembrane movement of ions, rotational mechanism	MF
83	GO:0019843	0.027	2	rRNA binding	MF

84	GO:0005199	0.034	2	structural constituent of cell wall	MF
85	GO:0050136	0.044	3	NADH dehydrogenase (quinone) activity	MF
86	GO:0003954	0.044	3	NADH dehydrogenase activity	MF
87	GO:0008137	0.044	3	NADH dehydrogenase (ubiquinone) activity	MF
88	GO:0003955	0.044	3	NAD(P)H dehydrogenase (quinone) activity	MF
89	GO:0030234	0.046	5	enzyme regulator activity	MF
90	GO:0016655	0.045	3	oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor	MF

A



B



Figure 2.7. Results of Gene Ontology enrichment analysis of species-specific genes. (A) *B. arcuata*, (B) *B. perennans*. The Y-axis lists enriched GO terms and the X-axis indicates the number of genes associated with a GO term. The bars in plots are colored by the p-value (FDR < 5%).

GO enrichment analysis of gene families that while present in both genomes were expanded in one or the other further elucidated the species-specific evolutionary patterns (Figure 2.8). I observed some overlap GO terms of species- specific gene expansions across the two species, including GO:0007154 (cell communication), GO:0007165 (signal transduction), GO:0023052 (signaling), GO:0051716 (cellular response to stimulus), GO:0050896 (response to stimulus), GO:0065007 (biological regulation), GO:0050789 (regulation of biological process), and GO:0050794 (regulation of cellular process). I also identified specific GO terms in each species. In *B. arcuata*, GO terms from 83 biological processes (BP), 18 cellular compartments (CC), and 71 molecular functions (MF) were identified. Of the GO terms in BP category, ten GO terms were related to response mechanisms (e.g., GO:0050896: response to stimulus, GO:0051716: cellular response to stimulus, GO:0006952: defense response), and 22 terms were related to biosynthetic and metabolic processes (e.g., GO:0006526: arginine biosynthetic process, GO:0009084: glutamine family amino acid biosynthetic process, GO:0009116: nucleoside metabolic process). Of the GO terms in the CC category, four terms were associated with ribosomal components (e.g.,GO:0015934: large ribosomal subunit, GO:0044391: ribosomal subunit, GO:1990904: ribonucleoprotein complex), and four terms were related to photosynthesis (GO:0009538: photosystem I reaction center, GO:0009522: photosystem I, GO:0034357: photosynthetic membrane). Of the GO terms in the MF category, 17 terms were related to molecular binding (e.g., GO:0043531: ADP binding, GO:0032559: adenylyl ribonucleotide binding, GO:0032555: purine ribonucleotide binding), and 14 terms were related to enzyme and protein interactions (e.g., GO:0047134: protein-disulfide reductase activity,

GO:0016668: oxidoreductase activity, acting on a sulfur group of donors, NAD(P) as acceptor, GO:0016651: oxidoreductase activity, acting on NAD(P)H).

In *B. perennans*, GO terms from 90 biological processes (BP), 10 cellular compartments (CC), and 63 molecular functions (MF) were identified. Of the GO terms in BP category, four GO terms were related to signal transduction and cell communication (e.g., GO:0007154: cell communication, GO:0007165: signal transduction, GO:0023052: signaling), three terms were associated with energy production and transport (GO:0015985: energy coupled proton transport, down electrochemical gradient, GO:0015986: ATP synthesis coupled proton transport, GO:1902600: proton transmembrane transport), and four terms related to biosynthetic and metabolic processes ((e.g., GO:0006754: ATP biosynthetic process, GO:0009142: nucleoside triphosphate biosynthetic process, GO:0009199: ribonucleoside triphosphate metabolic process. Of the GO terms in CC category, four terms were associated with proton-transporting ATP synthase complex (e.g., GO:0045263: Proton-transporting ATP synthase complex, coupling factor F(o), GO:0033177: proton-transporting two-sector ATPase complex, proton-transporting domain, GO:0016469: proton-transporting two-sector ATPase complex). Finally, of the GO terms in MF category, 11 terms were related to binding activities (e.g., GO:0043531: ADP binding, GO:0032559: adenylyl ribonucleotide binding, GO:0032555: purine ribonucleotide binding) and six terms related to transporter activities (e.g., GO:0015078: proton transmembrane transporter activity, GO:0022890: inorganic cation transmembrane transporter activity, GO:0008324: cation transmembrane transporter activity).





Figure 2.8. Gene Ontology enrichment analysis of orthologous genes expanded in either *B. arcuata* or *B. perennans*. (A) *B. arcuata*, (B) *B. perennans*. The Y-axis lists the top 50 most enriched GO terms and the X-axis indicates the number of genes associated with a GO term. The bars in plots are colored by the p-value (FDR < 5%).

Comparative analysis of individual gene expansions in *B. arcuata* and *B. perennans* reveals gene duplications in distinct stress pathways

Through a targeted analysis of gene family expansions within *B. arcuata* and *B. perennans*, I identified several genes potentially pivotal for stress responses to their unique habitats (Figure 2.9). I refined the gene selection by focusing on those with more than a two-fold expansion in copy number in *B. arcuata* compared to either *B. perennans* or *A. thaliana* as well as the genes with more than a two-fold expansion in copy number in *B. perennans* relative to either *B. arcuata* or *A. thaliana*. Expansion of genes associated with photosynthetic processes and responses to abiotic stress, suggests these genes play a vital role in each species responses to their distinct habitats (Figure 2.9). In *B. arcuata* I identified eight genes with significantly increased copy number compared to *B. perennans*.

- 1) *Beta Glucosidase 34* gene (*BGLU34*), which encodes a myrosinase, participates in carbohydrate metabolism, glucosinolate catabolic processes, and salt stress response,
- 2) *Activated Disease Resistance 2* gene (*ADR2*) which has roles in cellular UV response and signal transduction pathways,
- 3) the genes for Toll-Interleukin-Resistance (TIR) domain family protein which are linked to the cellular response to hypoxia,
- 4) the *DARK INDUCIBLE 11* gene (*DIN11*) which is responsive to sugar levels and is involved in the cellular response to starvation,
- 5) the *REDUCED EPIDERMAL FLUORESCENCE 2* gene (*REF2*), a cytochrome p450 enzyme with functions in glucosinolate biosynthesis and UV response,
- 6) Proteins from the NAD(P)-binding Rossmann-fold superfamily and FAD/NAD(P)-binding oxidoreductase family that are crucial for energy production and stress responses,
- 7) the gene *ANKYRIN REPEAT-CONTAINING 2B* (*AKR2B*) which is

crucial for chloroplast biogenesis, 8) the gene *ATACER*, whose produce is involved in sphingolipid metabolism, and plays a significant role in defense response to other organisms, autophagy, and response to salt stress.

In *B. perennans*, I found eight gene expansions not found in *B. arcuata*, many of which are associated with the proper functioning of chloroplast: 1) the gene for SOC3, a TIR-NB-leucine-rich repeat protein, which is crucial for modulating auto-immunity and thermally activated cell death, 2) the gene for CAFFEATE O-METHYLTRANSFERASE 1 (COMT1), a protein that is central in wide range of processes from aromatic compound synthesis to osmotic stress response, 3) the gene *FRUCTOKINASE-LIKE 1 (FLN1)*, whose product is critical for chloroplast development and transcription, indicating adaptation in energy metabolism, 4) the gene for PALE CRESS (PAC), this protein impacts chloroplast assembly and leaf development, and notably is important for adapting to light conditions, 5) the gene for MSL1, which is crucial for cellular response to oxidative stress, 6) the gene for NEP-INTERACTING PROTEIN 1 (NIP1), which is important for light response, 7) the gene for PLASTID TRANSCRIPTIONALLY ACTIVE 14 (PTAC14), and 8) the gene for ARABIDOPSIS THALIANA SULFOTRANSFERASE 1 (ATSOT1), both of which are involved in chloroplast organization and stress response.



Figure 2.9. Number of genes expanded in either *B. arcuata* or *B. perennans* that are associated with photosynthesis and abiotic stress responses. Here I depict a heatmap, showing the copy number of orthologous genes with expanded gene counts in *B. arcuata* relative to *B. perennans* and *A. thaliana* as well as expanded gene counts in *B. perennans* relative to *B. arcuata* and *A. thaliana*. Red color indicates a high gene count and yellow indicates a low count.

Gene family evolution across *Boechea* and some members of Brassicaceae

Analyzing the evolutionary trajectory of gene families across *Boechea* species and several species of Brassicaceae, I found the overall numbers of expansions and contractions in these species (Figure 2.10). In the phylogenetic tree analysis through CAFE5, a common ancestor diverged into two primary lineages. In the first lineage, *A. halleri* expanded 561 gene families while while showing evidence of having reduced counts in 1053 families. In contrast, *A. lyrata* showed an expansion in 982 gene families but a contraction in 320,

which was a less dramatic loss compared to *A. halleri*. *Arabidopsis thaliana*, on a separate branch, exhibited a moderate increase in gene families with 363 expansions and a higher contraction of 461 gene families. The second lineage diverged into two distinct paths: *B. perennans* showed a significant number of gene family expansions at 492, and a contraction of 208 gene families. *B. arcuata*, similarly displayed an expansion in 466 gene families and a contraction in 261. Notably, *B. perennans* and *B. arcuata* show a higher ratio of expansions to contractions. *C. papaya*, as the outgroup on a distinct branch, was characterized by a substantial loss of 1933 gene families and an expansion of 1597 (Figure 2.10).

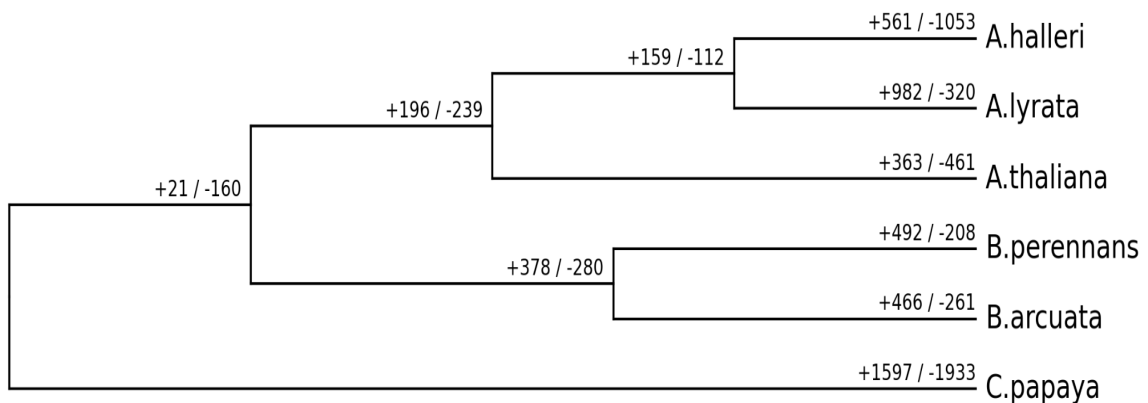


Figure 2.10. Gene families with expansion and/or contraction in the genomes of *B. arcuata*, *B. perennans*, and selected representative of Brassicaceae predicted by CAFE. The number of gene families that have undergone expansion and contraction are indicated by + and -, respectively. *C. papaya* is used as the outgroup species.

To enhance the understanding of gene family evolution within the context of environmental adaptation, I identified orthologs of the candidate expanded genes previously found in *B. arcuata* and *B. perennans*, in additional Brassicaceae species, *A. halleri* and *A. lyrata*, as well as in *C. papaya* (Table 2.5, 2.6). The gene families previously found to be expanded in *B. arcuata* and *B. perennans* continued to exhibit expansion relative to the additional species as well, highlighting a consistent pattern of gene family expansion in these species.

Table 2.5. Orthogroup gene expansions in *B. arcuata* for genes associated with photosynthesis and abiotic stress response.
The table shows the expansion of selected orthogroup genes within *B. arcuata* in comparison to *B. perennans*, *A. halleri*, *A. lyrata*, *A. thaliana*, and *C. papaya*. Orthogroups are shown with gene identifiers specific to each species.

	Orthogroup	<i>A. halleri</i>	<i>A. lyrata</i>	<i>A. thaliana</i>	<i>B. arcuata</i>	<i>B. perennans</i>	<i>C. papaya</i>
1	OG00000198	Ah7G19840 Ah7G19930	AL1G52250 AL1G59360 AL1G59380	AT1G47600 AT1G51470 AT1G51490	Boarc.1G173200.1 Boarc.1G173300.1 Boarc.1G173400.1 Boarc.1G173500.1 Boarc.1G175000.1 Boarc.1G175100.1 Boarc.1G175200.1 Boarc.1G175300.1	Boper.1G170700.1 Boper.1G172200.1 Boper.1G172300.1 Boper.1G172400.1 Boper.1G172500.1 Boper.1G172600.1 Boper.H010200.1	28829.1PACid_A_16429637 29746.1PACid_16429727 45941.1PACid_16432238 17.152PACid_16410647 84.113PACid_16427275
2	OG00000201	Ah2G09910. 1 Ah2G09950. 1	AL1G66180.tl AL1G66260.tl AL1G66350.t2 AL1G66380.tl AL2G23700.tl	AT1G56510.1 AT1G56520.2 AT1G56540.1 AT1G66090.1	Boarc.1G242300.1 Boarc.1G242400.1 Boarc.1G242500.1 Boarc.1G243000.1 Boarc.1G243100.1 Boarc.1G246600.1 Boarc.1G246800.1 Boarc.2G108500.1 Boarc.2G194200.1	Boper.1G242900.1 Boper.1G243100.1 Boper.1G243300.1 Boper.1G243400.1 Boper.1G243500.1 Boper.1G246400.1 Boper.2G199800.1 Boper.H026200.1	
3	OG00000223	Ah2G09010. 1		AT1G57630.1 AT5G38344.1	Boarc.1G029300.1 Boarc.1G248700.1 Boarc.1G249600.1 Boarc.1G260900.1 Boarc.1G270400.1 Boarc.2G253200.1 Boarc.3G035100.1 Boarc.3G039500.1 Boarc.3G276600.1 Boarc.4G065100.1 Boarc.5G031900.1 Boarc.5G032100.1 Boarc.5G032400.1 Boarc.5G212800.1 Boarc.5G235700.1 Boarc.5G237500.1	Boper.5G240300.1	

						Boarc.5G238200.1 Boarc.6G100500.1 Boarc.6G102400.1 Boarc.7G031700.1 Boarc.7G032400.1 Boarc.7G060200.1 Boarc.7G063300.1				
4	OG0000528	Ah5G18720. 1Ah5G1939 0.3	AL5G29480.tl AL5G30340.tl AL8G22460.tl	AT3G49620.1 AT3G49630.1 AT3G50210.1	Boarc.1G133600.1 Boarc.5G273400.1 Boarc.5G274000.1 Boarc.5G274300.1 Boarc.5G274600.1 Boarc.5G275000.1 Boarc.5G275300.1 Boarc.5G280800.1	Boper.5G278200.1 Boper.5G284200.1	evm.model.supercontig_55.109PACid_16422494			
5	OG0001395	Ah7G31670. 1	AL7G42750.tl AL8G10030.tl 1	AT4G13770.1	Boarc.7G040500.1 Boarc.7G040600.1 Boarc.7G101300.1 Boarc.7G101400.1 Boarc.7G101500.1 Boarc.7G101600.1	Boper.7G101100.1 Boper.7G101200.1 Boper.7G101400.1				
6	OG0001423	Ah2G11140. 1Ah3G4480 0.1 Ah8G13770. 1	AL3G50270.tl	AT2G17390.1	Boarc.1G235000.1 Boarc.3G308300.1 Boarc.3G314200.1 Boarc.5G106000.1	Boper.3G300800.1 Boper.3G303300.1 Boper.5G108900.1	evm.model.supercontig_63.43PACid_16424133			
7	OG0002449	Ah6G03840. 1	AL4G42800.tl AL6G13700.tl	AT5G04070.1	Boarc.1G111100.1 Boarc.1G111700.1 Boarc.3G000700.1 Boarc.6G346200.1	Boper.1G109400.1 Boper.6G351600.1	evm.model.supercontig_87.4PACid_16427707			
8	OG0002828	Ah7G39830. 4	AL7G50200.tl	AT5G39940.1	Boarc.7G039500.1 Boarc.7G039900.1 Boarc.7G041100.1 Boarc.7G041500.1	Boper.7G038700.1 Boper.7G039300.1 Boper.7G039700.1	evm.model.supercontig_14.49PACid_16408796			
9	OG0003473	Ah7G19330. 1	AL7G31280.tl	AT4G22330.1	Boarc.6G060100.1 Boarc.6G060300.1 Boarc.6G060900.1 Boarc.7G199200.1	Boper.6G063300.1 Boper.7G196600.1	evm.TU.contig_36104.1PACid_16430773			

Table 2.6. Orthogroup gene expansions in *B. perennans* for genes associated with photosynthesis and abiotic stress response.
The table shows the expansion of selected orthogroup genes within *B. perennans* in comparison to *B. arcuata*, *A. halleri*, *A. lyrata*, *A. thaliana*, and *C. papaya*. Orthogroups are shown with gene identifiers specific to each species.

	Orthogroup	<i>A. halleri</i>	<i>A. lyrata</i>	<i>A. thaliana</i>	<i>B. arcuata</i>	<i>B. perennans</i>	<i>C. papaya</i>
1	OG0000196	Ah1G38660.1 Ah2G23940.1	AL1G29920.t1 AL2G32530.t1 AL2G32570.t1 AL7G49910.t1	AT1G17600.1 AT1G72840.2 AT1G72860.2 AT4G09360.1 AT4G09430.1 AT4G09435.1 AT5G40100.1 AT5G48770.1	Boarc.2G291200.1 Boarc.2G291300.1 Boarc.6G132000.1 Boarc.7G035500.1	Boper.2G294100.1 Boper.2G294600.1 Boper.7G033600.1 Boper.7G033800.1 Boper.7G034100.1 Boper.7G034500.1 Boper.7G034600.1 Boper.7G034900.1 Boper.7G035100.1 Boper.7G035200.1	
2	OG0000420	Ah7G12910.1 Ah8G19350.1 Ah8G19360.1	AL7G24850.t1 AL8G29310.t1 AL8G29370.t1	AT5G54160.1	Boarc.3G354200.1 Boarc.3G354700.1 Boarc.3G409200.1 Boarc.6G139600.1 Boarc.7G258200.1 Boarc.7G259600.1	Boper.3G347500.1 Boper.3G347600.1 Boper.3G348100.1 Boper.3G403900.1 Boper.6G146900.1 Boper.7G255500.1 Boper.7G256900.1	evm.model.supercontig_3.161 PACid_16416437
3	OG0001918	Ah5G24270.1	AL5G35100.t1 AL5G35110.t1	AT3G54090.1	Boarc.5G320200.1 Boarc.5G320700.1 Boarc.5G321100.1	Boper.5G324800.1 Boper.5G324900.1 Boper.5G325700.1 Boper.5G330600.1	evm.model.supercontig_56.93 PACid_16422806
4	OG0002596	Ah4G37980.2	AL4G47720.t1	AT2G48120.1	Boarc.2G377800.1 Boarc.4G302200.1 Boarc.5G001200.1	Boper.4G307400.1 Boper.5G001300.1 Boper.5G002200.1 Boper.6G000900.1	evm.model.supercontig_485.5 PACid_16420985
5	OG0003843	Ah6G44410.1	AL6G53230.t1	AT4G00234.1 AT4G00290.1	Boarc.6G004400.1	Boper.6G004800.1 Boper.6G005400.1 Boper.6G005800.1	evm.model.supercontig_11.45 PACid_16405711
6	OG0004094	Ah3G44330.1	AL3G50750.t1	AT2G17750.1	Boarc.1G214900.1 Boarc.2G187500.1	Boper.1G214400.1 Boper.1G214500.1 Boper.1G214600.1	

							Boper.2G192200.1	
7	OG0004646	Ah7G22290.1	AL7G33690.t1	AT4G20130.1	Boarc.7G178400.1	Boper.7G000700.1 Boper.7G000800.1 Boper.7G175100.1	evm.model.supercontig_30.93 PACid_16417047	
8	OG0004756	Ah5G03010.1	AL5G13550.t1	AT2G03760.1	Boarc.5G171300.1	Boper.5G171600.1 Boper.5G171700.1 Boper.5G171800.1 Boper.5G209000.1		

Discussion

The genus *Boechnera* offers a unique opportunity to study the evolutionary dynamics of plant genomes in natural environments. Unlike *Arabidopsis thaliana*, *Boechnera* species are predominantly perennial and occupy diverse habitats across western North America. The perennial nature of certain *Boechnera* species allows researchers to study temporal variations in trait expression and natural selection (Wadgymar, Daws, and Anderson 2017; Rushworth et al. 2022). Moreover, the ecological diversity of *Boechnera* species makes them an excellent system for investigating how plants adapt to varying environmental conditions (Rushworth et al. 2022). For instance, population genetic studies have shown that balancing selection maintains genetic variation in approximately 10% of the *Boechnera stricta* genome, highlighting the importance of selection in shaping genomic diversity (Wang et al. 2019).

The results presented here reveal significant similarities in genome size, chromosome number, and gene content in *B. arcuata* and *B. perennans*. *Boechnera arcuata* and *B. perennans*, each have a total of 27,156 and 28,159 gene models, respectively. With genome sizes of 231.7 Mb for *B. arcuata* and 219.87 Mb for *B. perennans*. The slightly higher gene count in *B. perennans* compared to *B. arcuata*, despite *B. perennans* having a smaller genome, might be due to single or repeated gene duplication events in *B. perennans*. *Boechnera arcuata* and *B. perennans* have similar gene numbers as those reported for other *Boechnera* species. For instance, *Boechnera retrofracta* and *Boechnera stricta* have 27,048 and 27,416 protein-coding loci, respectively. The genome sizes of *B. arcuata* (231.7 Mb) and *B. perennans* (219.87 Mb) are comparable to those of other

Boechea species, such as *B. retrofracta* (227 Mb) and *B. stricta* (184 Mb). It is important to note here the other *Boechea* genomes are not assembled into chromosomes and thus were not included in my comparative syntenic analysis. The genomes sizes reported for the *Boechea* species are slightly larger than those of *A. lyrata* (207 Mb) and significantly larger than *A. thaliana* (135 Mb) (Kliver et al. 2018). Moreover, repetitive elements in the genomes of *B. arcuata* and *B. perennans* accounted for almost 40% of the total genome space. Nearly half of these were long terminal repeats (LTRs), aligning with the findings in *B. retrofracta* (Kliver et al. 2018). These are higher than the approximately 20% of the genome being composed of repetitive elements reported for *Arabidopsis* (Wright, Agrawal, and Bureau 2003; Tenaillon, Hollister, and Gaut 2010; Quadrana et al. 2016; Legrand et al. 2019).

The considerable syntenic conservation between *B. arcuata* and *B. perennans* indicates genomic stability and preservation of gene order between these two species. In contrast, the syntenic analysis with *Arabidopsis thaliana* revealed more complex patterns, indicative of significant recombination, duplication, and rearrangement events. In a recent paper on chromosomal evolution (Mandáková et al. 2020) examined chromosomal organization in three *Boechea* species and in six other closely related genera (*Borodinia*, *Cusickiella*, *Phoenicaulis*, *Polycytenium*, *Nevada*, and *Sandbergia*) all within the tribe *Boecheae* (*Brassicaceae*). All were found to have $n=7$. These researchers also reported that even among distinct genera there was significant chromosomal conservation. This study was based on cytogenetics and chromosomal painting and was not based on genome sequence data. While this means that a direct comparison of synteny with these other

Boecheae species is not possible, it does suggest that the high level of syntenic conservation between *B. arcuata* and *B. perennans* is consistent with other findings. Surprisingly in a study of *Boechea stricta* in the Rocky Mountains researchers found chromosomal inversions within populations of this species. The inversion contains QTL loci implicated in traits associated with flowering and ecological adaptation (Lee et al. 2017). The large size of the *Boechea* genus suggests I could expect variation across its constituent members. The data presented here and that of Mandakova et al 2020 suggests that chromosomal conservation may be the rule within *Boechea* and the findings within *B. stricta* may be the outlier.

Despite the similarities in genome size, total gene number and synteny between *B. arcuata* and *B. perennans*, there are differences in gene family or orthogroup content and these differences provide hypotheses into the adaptive mechanisms acting in these species. While the high level of synteny between these two *Boechea* species and *A. thaliana* underpins their shared evolutionary history, yet their unique gene family (orthogroup) expansions suggest species-specific genomic adaptations. As mentioned in the Introduction above, numerous studies have found that lineage or species-specific expansions may underlie adaptations to environmental conditions in other species (see recent review Das Laha et al 2020). The GO enrichment analysis (Figures 2.7 and 2.8) provided here gives further insights into the functional roles of species-specific and expanded gene groups. In *B. arcuata*, unique genes were enriched in terms related to protein assembly, transport mechanisms, and energy-related processes, indicating a robust system for maintaining cellular function under thermal stress. *Boechea perennans*, on the other hand, showed

enrichment in terms associated with photosynthesis, stress response, and mitochondrial function, suggesting its adaptation to high-stress, low-water environments.

Lineage-specific expansions of genes and/or gene families have been associated with genomic and organismal adaptation in a wide range of organisms (Nikolaidis and Nei 2004; Gibbs et al. 2004; Thomas 2006, 2007; Lehti-Shiu et al. 2017; Kim et al. 2022). Here, I examined the differences in the functional genomics of the two *Boechera* species by identifying genes that are expanded in one species compared to the other and *A. thaliana*. For instance, the enrichment of genes related to abiotic stress responses in *B. arcuata* (Figure 2.9, Tables 2.5 and 2.6) suggests its adaptation to variable coastal climates, where efficient stress tolerance are critical for survival. One such gene is *BGLU34*, which encodes beta-glucosidase 34. *BGLU34* is involved in the regulation of active phytohormone levels through the hydrolysis of hormone-glucoside conjugates. This activity is essential for modulating hormone availability and thus, influencing various developmental and stress response processes in plants (Xu et al. 2004). The presence and expansion of *BGLU34* in *B. arcuata*, compared to *A. thaliana*, suggest that this species may have developed enhanced metabolic and defensive capabilities, possibly contributing to its thermotolerance and adaptability to low-elevation coastal areas with fluctuating temperature and humidity. Another instance of species-specific gene expansions in *B. arcuata* is the gene for *REDUCED EPIDERMAL FLUORESCENCE (REF)*. The product of this gene is involved in the biosynthesis and regulation of secondary metabolites (Stout et al. 2008; Franke et al. 2002), and plays critical roles in plant structure, defense, and stress response. The expansion of the *REF* gene in *B. arcuata* and its critical role in the

biosynthesis of phenylpropanoids and other secondary metabolites suggests that it may enhance its ability to respond to environmental stresses. Further detailed studies of these genes and their role in stress tolerance will be needed to test this hypothesis.

Boechera perennans displays an expansion of gene families associated with chloroplast function and energy metabolism, which may reflect its adaptation to arid desert environments. One of the challenges of seeking the genetic or biochemical underpinning of organismal traits such as tolerance to the extreme conditions in deserts is that it can be difficult to choose which genes and proteins should be examined in detail for extensive in vivo and in vitro studies. Whole genome comparisons such as that presented here are highly valuable for their ability to identify possible candidates for future studies. For instance, the species -specific expansion of the genes *COMT1*, which is involved in osmotic stress response (Vincent et al. 2005; Y. Li et al. 2022) is intriguing. Direct comparisons of the osmotic stress responses of a variety of related species with and without these expanded *COMT1* genes was outside the scope of this project but future detailed physiological and biochemical studies should prove highly informative. The *FRUCTOKINASE-LIKE 1* (*FLN1*) gene, a member of the pfkB-type carbohydrate kinase family, is another expanded gene, indicating its potential significance in *B. perennans*. According to a study in *A. thaliana*, *FLN1* plays a crucial role in plant growth, development, and chloroplast function. *FLN1* is part of the transcriptionally active chromosome (TAC) complex in chloroplasts, specifically associated with the plastid-encoded RNA polymerase (PEP) complex. This complex is crucial for the transcription of chloroplast genes, which are essential for chloroplast development and function (Gilkerson et al. 2012). This highlights the vital role

of *FLNI* in supporting normal plant growth and photosynthetic competency. Also, *FLNI* is involved in metabolic pathways related to sugar phosphorylation, which is essential for energy production and metabolic regulation within the plant cells. The expansion of the *FLNI* gene in *B. perennans* suggests its crucial role in maintaining efficient chloroplast function and supporting robust plant growth and development. By ensuring proper transcription of chloroplast genes and contributing to metabolic processes, *FLNI* might help optimize the energy production and adaptability of *B. perennans* to its environment. These findings suggest that the genomic adaptations in *B. perennans* are geared towards optimizing photosynthesis and maintaining cellular stability in extreme environments. To fully understand the adaptive importance of gene family in *Boechera*, future biochemical or functional studies will be necessary.

Conclusion

Understanding the genomic basis of environmental adaptation in *Boechera* species has important implications for conservation and climate change studies. The identification of key genes involved in stress responses and metabolic processes can inform strategies to enhance the resilience of other plant species to environmental stressors. This work provides important insight into these little-studied *Boechera* species that inhabit geographically close but distinct environments. Here, I present the analysis of the first chromosomal-level assemblies within the diverse *Boechera* genus. These findings provide important comparative genomic data for additional genomic analyses within the *Boechera* species. Perhaps most importantly, this research identified important differences in the gene content, i.e., species-specific gene expansions that illuminate adaptations to diverse

environments. These lineage-specific expansions will be the basis for many future studies of these two species.

References

- Alexander, Patrick J., Michael D. Windham, James B. Beck, Ihsan A. Al-Shehbaz, Loreen Allphin, and C. Donovan Bailey. 2013. "Molecular Phylogenetics and Taxonomy of the Genus *Boechera* and Related Genera (Brassicaceae: Boechereae)." *Systematic Botany* 38 (1): 192–209.
- Aliyu, Olawale M., M. Eric Schranz, and Timothy F. Sharbel. 2010. "Quantitative Variation for Apomictic Reproduction in the Genus *Boechera* (Brassicaceae)." *American Journal of Botany* 97 (10): 1719–31.
- Bemmels, Jordan B., and Jill T. Anderson. 2019. "Climate Change Shifts Natural Selection and the Adaptive Potential of the Perennial Forb *Boechera Stricta* in the Rocky Mountains." *Evolution; International Journal of Organic Evolution* 73 (11): 2247–62.
- Conway, Jake R., Alexander Lex, and Nils Gehlenborg. 2017. "UpSetR: An R Package for the Visualization of Intersecting Sets and Their Properties." *Bioinformatics* 33 (18): 2938–40.
- Edgar, Robert C. 2004. "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput." *Nucleic Acids Research* 32 (5): 1792–97.
- Emms, David M., and Steven Kelly. 2019. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology* 20 (1): 238.
- Franke, Rochus, Matthew R. Hemm, Jeff W. Denault, Max O. Ruegger, John M. Humphreys, and Clint Chapple. 2002. "Changes in Secondary Metabolism and Deposition of an Unusual Lignin in the ref8 Mutant of *Arabidopsis*." *The Plant Journal: For Cell and Molecular Biology* 30 (1): 47–59.
- Franzke, Andreas, Marcus A. Koch, and Klaus Mummenhoff. 2016. "Turnip Time Travels: Age Estimates in Brassicaceae." *Trends in Plant Science* 21 (7): 554–61.
- Franzke, Andreas, Martin A. Lysak, Ihsan A. Al-Shehbaz, Marcus A. Koch, and Klaus Mummenhoff. 2011. "Cabbage Family Affairs: The Evolutionary History of Brassicaceae." *Trends in Plant Science* 16 (2): 108–16.
- Gallas, Genna, and Elizabeth R. Waters. 2015. "Boechera Species Exhibit Species-Specific Responses to Combined Heat and High Light Stress." *PloS One* 10 (6): e0129041.

- German, Dmitry A., Kasper P. Hendriks, Marcus A. Koch, Frederic Lens, Martin A. Lysak, C. Donovan Bailey, Klaus Mummenhoff, and Ihsan A. Al-Shehbaz. 2023. “An Updated Classification of the Brassicaceae (Cruciferae).” *PhytoKeys* 220 (March): 127–44.
- Gibbs, Richard A., George M. Weinstock, Michael L. Metzker, Donna M. Muzny, Erica J. Sodergren, Steven Scherer, Graham Scott, et al. 2004. “Genome Sequence of the Brown Norway Rat Yields Insights into Mammalian Evolution.” *Nature* 428 (6982): 493–521.
- Gilkerson, Jonathan, Juan Manuel Perez-Ruiz, Joanne Chory, and Judy Callis. 2012. “The Plastid-Localized pfkB-Type Carbohydrate Kinases FRUCTOKINASE-LIKE 1 and 2 Are Essential for Growth and Development of *Arabidopsis Thaliana*.” *BMC Plant Biology* 12 (July): 102.
- Guo, Xinyi, Jianquan Liu, Guoqian Hao, Lei Zhang, Kangshan Mao, Xiaojuan Wang, Dan Zhang, et al. 2017. “Plastome Phylogeny and Early Diversification of Brassicaceae.” *BMC Genomics* 18 (1): 176.
- Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. “QUAST: Quality Assessment Tool for Genome Assemblies.” *Bioinformatics* 29 (8): 1072–75.
- Halter, Gillian, Nicole Simonetti, Cristy Suguitan, Kenneth Helm, Jessica Soroksky, and Elizabeth R. Waters. 2017. “Patterns of Thermotolerance, Chlorophyll Fluorescence, and Heat Shock Gene Expression Vary among four *Boechera* species and *Arabidopsis Thaliana*.” *Botany* 95 (1): 9–27.
- Hay, Nikolai M., Michael D. Windham, Terezie Mandáková, Martin A. Lysak, Kasper P. Hendriks, Klaus Mummenhoff, Frederic Lens, Kathleen M. Pryer, and C. Donovan Bailey. 2023. “A Hyb-Seq Phylogeny of *Boechera* and Related Genera Using a Combination of Angiosperms353 and Brassicaceae-Specific Bait Sets.” *American Journal of Botany* 110 (10): e16226.
- Hendriks, Kasper P., Christiane Kiefer, Ihsan A. Al-Shehbaz, C. Donovan Bailey, Alex Hooft van Huysduynen, Lachezar A. Nikolov, Lars Nauheimer, et al. 2023. “Global Brassicaceae Phylogeny Based on Filtering of 1,000-Gene Dataset.” *Current Biology: CB* 33 (19): 4052–68.e6.
- Hohmann, Nora, Eva M. Wolf, Martin A. Lysak, and Marcus A. Koch. 2015. “A Time-Calibrated Road Map of Brassicaceae Species Radiation and Evolutionary History.” *The Plant Cell* 27 (10): 2770–84.
- Huang, Chien-Hsun, Renran Sun, Yi Hu, Liping Zeng, Ning Zhang, Liming Cai, Qiang Zhang, et al. 2016. “Resolution of Brassicaceae Phylogeny Using Nuclear Genes

- Uncovers Nested Radiations and Supports Convergent Morphological Evolution.” *Molecular Biology and Evolution* 33 (2): 394–412.
- Kim, Heesoo, Han-Woo Kim, Jun Hyuck Lee, Joonho Park, Hyoungseok Lee, Sanghee Kim, and Seung Chul Shin. 2022. “Gene Family Expansions in Antarctic Winged Midge as a Strategy for Adaptation to Cold Environments.” *Scientific Reports* 12 (1): 18263.
- Kliver, Sergei, Mike Rayko, Alexey Komissarov, Evgeny Bakin, Daria Zhernakova, Kasavajhala Prasad, Catherine Rushworth, et al. 2018. “Assembly of the *Boechera Retrofracta* Genome and Evolutionary Analysis of Apomixis-Associated Genes.” *Genes* 9 (4). <https://doi.org/10.3390/genes9040185>.
- Knight, Charles A., Heiko Vogel, Juergen Kroymann, Alice Shumate, Hanneke Witsenboer, and Thomas Mitchell-Olds. 2006. “Expression Profiling and Local Adaptation of *Boechera Holboellii* Populations for Water Use Efficiency across a Naturally Occurring Water Stress Gradient.” *Molecular Ecology* 15 (5): 1229–37.
- Koenig, Daniel, and Detlef Weigel. 2015. “Beyond the Thale: Comparative Genomics and Genetics of *Arabidopsis* Relatives.” *Nature Reviews. Genetics* 16 (5): 285–98.
- Lee, Cheng-Ruei, and Thomas Mitchell-Olds. 2013. “Complex Trait Divergence Contributes to Environmental Niche Differentiation in Ecological Speciation of *Boechera Stricta*.” *Molecular Ecology* 22 (8): 2204–17.
- Lee, Cheng-Ruei, Baosheng Wang, Julius P. Mojica, Terezie Mandáková, Kasavajhala V. S. K. Prasad, Jose Luis Goicoechea, Nadeesha Perera, et al. 2017. “Publisher Correction: Young Inversion with Multiple Linked QTLs under Selection in a Hybrid Zone.” *Nature Ecology & Evolution* 1 (10): 1585.
- Legrand, Sylvain, Thibault Caron, Florian Maumus, Sol Schvartzman, Leandro Quadrana, Eléonore Durand, Sophie Gallina, et al. 2019. “Differential Retention of Transposable Element-Derived Sequences in Outcrossing *Arabidopsis* Genomes.” *Mobile DNA* 10 (July): 30.
- Lehti-Shiu, Melissa D., Nicholas Panchy, Peipei Wang, Sahra Uygun, and Shin-Han Shiu. 2017. “Diversity, Expansion, and Evolutionary Novelty of Plant DNA-Binding Transcription Factor Families.” *Biochimica et Biophysica Acta, Gene Regulatory Mechanisms* 1860 (1): 3–20.
- Li, Cheng, David Wickell, Li-Yaung Kuo, Xueqing Chen, Bao Nie, Xuezhu Liao, Dan Peng, et al. 2024. “Extraordinary Preservation of Gene Collinearity over Three Hundred Million Years Revealed in Homosporous Lycophytes.” *Proceedings of the National Academy of Sciences of the United States of America* 121 (4): e2312607121.

- Li, Fay-Wei, Catherine A. Rushworth, James B. Beck, and Michael D. Windham. 2017. "Boechera Microsatellite Website: An Online Portal for Species Identification and Determination of Hybrid Parentage." *Database: The Journal of Biological Databases and Curation* 2017 (1). <https://doi.org/10.1093/database/baw169>.
- Lin, Ya-Ping, Thomas Mitchell-Olds, and Cheng-Ruei Lee. 2021. "The Ecological, Genetic and Genomic Architecture of Local Adaptation and Population Differentiation in *Boechera stricta*." *Proceedings. Biological Sciences / The Royal Society* 288 (1949): 20202472.
- Li, Yan, Yan Sun, Huiting Cui, Mingna Li, Guofeng Yang, Zengyu Wang, and Kun Zhang. 2022. "Carex rigescens Caffeic Acid O-Methyltransferase Gene CrCOMT Confer Melatonin-Mediated Drought Tolerance in Transgenic Tobacco." *Frontiers in Plant Science* 13 (August): 971431.
- Lovell, John T., Avinash Sreedasyam, M. Eric Schranz, Melissa Wilson, Joseph W. Carlson, Alex Harkess, David Emms, David M. Goodstein, and Jeremy Schmutz. 2022. "GENESPACE Tracks Regions of Interest and Gene Copy Number Variation across Multiple Genomes." *eLife* 11 (September). <https://doi.org/10.7554/eLife.78526>.
- Mandáková, Terezie, Petra Hloušková, Michael D. Windham, Thomas Mitchell-Olds, Kaylynn Ashby, Bo Price, John Carman, and Martin A. Lysak. 2020. "Chromosomal Evolution and Apomixis in the Cruciferous Tribe Boechereae." *Frontiers in Plant Science* 11 (May): 514.
- Mendes, Fábio K., Dan Vanderpool, Ben Fulton, and Matthew W. Hahn. 2021. "CAFE 5 Models Variation in Evolutionary Rates among Gene Families." *Bioinformatics* 36 (22-23): 5516–18.
- Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era." *Molecular Biology and Evolution* 37 (5): 1530–34.
- Nikolaidis, Nikolas, and Masatoshi Nei. 2004. "Concerted and Nonconcerted Evolution of the Hsp70 Gene Superfamily in Two Sibling Species of Nematodes." *Molecular Biology and Evolution* 21 (3): 498–505.
- Nikolov, Lachezar A., Philip Shushkov, Bruno Nevado, Xiangchao Gan, Ihsan A. Al-Shehbaz, Dmitry Filatov, C. Donovan Bailey, and Miltos Tsiantis. 2019. "Resolving the Backbone of the Brassicaceae Phylogeny for Investigating Trait Diversity." *The New Phytologist* 222 (3): 1638–51.

- Oliveira, Jordana Inácio Nascimento, and Nicolas Corradi. 2024. “Strain-Specific Evolution and Host-Specific Regulation of Transposable Elements in the Model Plant Symbiont *Rhizophagus Irregularis*.” *G3* 14 (5). <https://doi.org/10.1093/g3journal/jkae055>.
- Ou, Shujun, Weijia Su, Yi Liao, Kapeel Chougule, Jireh R. A. Agda, Adam J. Hellings, Carlos Santiago Blanco Lugo, et al. 2022. “Author Correction: Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline.” *Genome Biology* 23 (1): 76.
- Palmer, J., and J. E. Stajich. 2017. “Funannotate: Eukaryotic Genome Annotation Pipeline.” *Zenodo*. *Doi* 10.
- Pierce, Ian, Gillian Halter, and Elizabeth R. Waters. 2020. “Transcriptomic Analysis of the Heat-Stress Response in *Boechera Depauperata* and *Arabidopsis* Reveals a Distinct and Unusual Heat-Stress Response in *Boechera*.” *Botany* 98 (10): 589–602.
- Quadrana, Leandro, Amanda Bortolini Silveira, George F. Mayhew, Chantal LeBlanc, Robert A. Martienssen, Jeffrey A. Jeddloh, and Vincent Colot. 2016. “The *Arabidopsis* *Thaliana* Mobilome and Its Impact at the Species Level.” *eLife* 5 (June). <https://doi.org/10.7554/eLife.15716>.
- Rushworth, Catherine A., Bao-Hua Song, Cheng-Ruei Lee, and Thomas Mitchell-Olds. 2011. “*Boechera*, a Model System for Ecological Genomics.” *Molecular Ecology* 20 (23): 4843–57.
- Rushworth, Catherine A., Maggie R. Wagner, Thomas Mitchell-Olds, and Jill T. Anderson. 2022. “The *Boechera* Model System for Evolutionary Ecology.” *American Journal of Botany* 109 (11): 1939–61.
- Schranz, M. Eric, Christoph Dobes, Marcus A. Koch, and Thomas Mitchell-Olds. 2005. “Sexual Reproduction, Hybridization, Apomixis, and Polyploidization in the Genus *Boechera* (Brassicaceae).” *American Journal of Botany* 92 (11): 1797–1810.
- Schranz, M. Eric, Laksana Kantama, Hans de Jong, and Thomas Mitchell-Olds. 2006. “Asexual Reproduction in a Close Relative of *Arabidopsis*: A Genetic Investigation of Apomixis in *Boechera* (Brassicaceae).” *The New Phytologist* 171 (2): 425–38.
- Schranz, M. Eric, Aaron J. Windsor, Bao-Hua Song, Amy Lawton-Rauh, and Thomas Mitchell-Olds. 2007. “Comparative Genetic Mapping in *Boechera Stricta*, a Close Relative of *Arabidopsis*.” *Plant Physiology* 144 (1): 286–98.
- Sharbel, Timothy F., Marie-Luise Voigt, José M. Corral, Giulio Galla, Jochen Kumlehn, Christian Klukas, Falk Schreiber, Heiko Vogel, and Björn Rotter. 2010. “Apomictic

- and Sexual Ovules of *Boechera* Display Heterochronic Global Gene Expression Patterns.” *The Plant Cell* 22 (3): 655–71.
- Smit, Arian F. A. 2004. “Repeat-Masker Open-3.0.” [Http://www. Repeatmasker. Org](http://www.Repeatmasker.Org). <https://cir.nii.ac.jp/crid/1571135649974754688>.
- Song, Bao-Hua, and Thomas Mitchell-Olds. 2007. “High Genetic Diversity and Population Differentiation in *Boechera Fecunda*, a Rare Relative of *Arabidopsis*.” *Molecular Ecology* 16 (19): 4079–88.
- Stout, Jake, Ethan Romero-Severson, Max O. Ruegger, and Clint Chapple. 2008. “Semidominant Mutations in Reduced Epidermal Fluorescence 4 Reduce Phenylpropanoid Content in *Arabidopsis*.” *Genetics* 178 (4): 2237–51.
- Tenaillon, Maud I., Jesse D. Hollister, and Brandon S. Gaut. 2010. “A Triptych of the Evolution of Plant Transposable Elements.” *Trends in Plant Science* 15 (8): 471–78.
- Thomas, James H. 2006. “Concerted Evolution of Two Novel Protein Families in *Caenorhabditis* Species.” *Genetics* 172 (4): 2269–81.
PLoS Genetics 3 (5): e67.
- Vincent, Delphine, Catherine Lapierre, Brigitte Pollet, Gabriel Cornic, Luc Negroni, and Michel Zivy. 2005. “Water Deficits Affect Caffeate O-Methyltransferase, Lignification, and Related Enzymes in Maize Leaves. A Proteomic Investigation.” *Plant Physiology* 137 (3): 949–60.
- Wadgyamar, Susana M., S. Caroline Daws, and Jill T. Anderson. 2017. “Integrating Viability and Fecundity Selection to Illuminate the Adaptive Nature of Genetic Clines.” *Evolution Letters* 1 (1): 26–39.
- Wang, Baosheng, Julius P. Mojica, Nadeesha Perera, Cheng-Ruei Lee, John T. Lovell, Aditi Sharma, Catherine Adam, et al. 2019. “Correction to: Ancient Polymorphisms Contribute to Genome-Wide Variation by Long-Term Balancing Selection and Divergent Sorting in *Boechera Stricta*.” *Genome Biology* 20 (1): 161.
- Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer International Publishing.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemond, et al. 2019. “Welcome to the Tidyverse.” *Journal of Open Source Software* 4 (43): 1686.
- Wright, Stephen I., Newton Agrawal, and Thomas E. Bureau. 2003. “Effects of Recombination Rate and Gene Density on Transposable Element Distributions in *Arabidopsis Thaliana*.” *Genome Research* 13 (8): 1897–1903.

- Xu, Zhiwei, Luis Escamilla-Treviño, Lihui Zeng, Mallikarjun Lalgondar, David Bevan, Brenda Winkel, Ali Mohamed, et al. 2004. “Functional Genomic Analysis of Arabidopsis Thaliana Glycoside Hydrolase Family 1.” *Plant Molecular Biology* 55 (3): 343–67.
- Zwaenepoel, Arthur. n.d. *Wgd: Python Package and CLI for Whole-Genome Duplication Related Analyses. This Package Is Deprecated in Favor of <https://github.com/heche-psb/wgd>*. Github. Accessed April 6, 2024. <https://github.com/arzwa/wgd>.
- Zwaenepoel, Arthur, and Yves Van de Peer. 2018. “Wgd—simple Command Line Tools for the Analysis of Ancient Whole-Genome Duplications.” *Bioinformatics* 35 (12): 2153–55.

Chapter 3: Gene expression profiles of the heat stress response in two *Boechnera* species reveals both conserved and divergent heat stress responses

Abstract

Understanding plant responses to environmental stresses is crucial for improving crop resilience and productivity. Among various stressors, heat stress significantly affects plant growth and survival. This study investigates the gene expression profiles and evolutionary patterns of heat stress responses in two *Boechnera* species, *B. arcuata* and *B. perennans*, which inhabit distinct climatic regions in California. Using RNA sequencing data, I identified differentially expressed genes under control and basal heat stress conditions. The top 50 upregulated genes in both species were predominantly heat shock proteins, highlighting their crucial role in stress responses. GO enrichment analysis revealed distinct and overlapping responses to heat stress, with *B. arcuata* showing enrichment in metabolic processes and *B. perennans* in enzyme activity and nucleic acid binding. Shared GO terms included stress response and protein folding, indicating common adaptive mechanisms. I also focused on genes both upregulated and expanded in each species, listing them along with their *A. thaliana* orthologs and functional descriptions to reveal species-specific adaptations. Positive selection tests identified amino acid sites under positive selection, indicating evolutionary changes. Phylogenetic trees showed branches under positive selection, suggesting potential neofunctionalization. This study provides insights into the adaptive evolution of heat stress responses in *Boechnera* species,

enhancing our understanding of the molecular mechanisms underlying thermotolerance and offering a foundation for future research on stress adaptation in plants.

Introduction

Plants are incessantly exposed to a variety of environmental stresses that adversely affect their growth, development, and distribution. Among these abiotic stressors, drought, extreme temperatures, and soil salinity present significant challenges, directly impacting agricultural productivity and global food security (Zhu 2016). The adaptive responses of plants to abiotic stresses play a crucial role for plant survival and are key targets for improving crop resilience (H. Zhang et al. 2022). These responses vary, and involve a range of physiological, metabolic, and genetic changes (Atkinson and Urwin 2012; Nakashima et al. 2012; Zhou et al. 2009).

From the discovery of the “puffs” in *Drosophila* polytene chromosomes in response to heat by Ritossa in 1962 (Ritossa 1962, 1996) to the present day, the heat shock or stress response has generated considerable interest (Lindquist and Craig 1988; E. Vierling 1991; E. Vierling and Kimpel 1992; Boston, Viitanen, and Vierling 1996; Waters, Lee, and Vierling 1996; Elizabeth Vierling 1997; Kotak et al. 2007; Wahid et al. 2007; Richter, Haslbeck, and Buchner 2010; Mittler, Finka, and Goloubinoff 2012; Hilton et al. 2013; Waters and Vierling 2020; Guihur, Rebeaud, and Goloubinoff 2022). We now know that heat stress and other abiotic stressors can have dramatic effects on gene expression and that the ability to change gene expression in response to heat stress can provide organismal thermotolerance (Gurley 2000; Sharkey 2005; Sorte and Hofmann 2005; Yeh et al. 2012).

All organisms possess the heat shock response, and both the heat-shock proteins (Lindquist and Craig 1988; Becker and Craig 1994; Schöffl, Prändl, and Reindl 1998) and heat-shock transcription factors (Nover et al. 1996; Pirkkala, Nykänen, and Sistonen 2001; Scharf, Siddique, and Vierling 2001; Mittler, Finka, and Goloubinoff 2012) are highly conserved across organismal diversity. We now know that most of the HSPs (heat-shock proteins) are chaperones that assist in the proper folding of other proteins (Bukau and Horwich 1998; Haslbeck 2002; W. Wang et al. 2004; Liberek, Lewandowska, and Zietkiewicz 2008; Eyles and Gierasch 2010; Kim et al. 2013). Some HSPs are expressed only during heat-stress, some are expressed during heat stress and other abiotic stressors, a few are developmentally expressed at specific developmental stages or in specific structures and still others are constitutively expressed (Waters et al. 1996; Waters and Vierling 2020). Extensive studies in plants and other organisms have revealed that while the HSPs play a key and essential role in tolerance to heat stress they are not the only component of the plant HS (heat stress) response (Larkindale et al. 2005; Mittler 2006; Kotak et al. 2007; Larkindale and Vierling 2008; Mittler, Finka, and Goloubinoff 2012; Baxter, Mittler, and Suzuki 2014).

Responses to abiotic stressors in plants has been shown to involve a complex network of signals involving reactive oxygen species (ROS), plant growth regulators, alterations in lipids, changes in flavonols and other phenolics (Cheong et al. 2002; Hazen, Wu, and Kreps 2003; Atkinson and Urwin 2012; Suzuki et al. 2012; Pintó-Marijuan and Munné-Bosch 2014; You and Chan 2015; Hou, Ufer, and Bartels 2016; Pandey et al. 2017; Zhu 2016; Mishra et al. 2023). There can also be significant differences among plants in their constitutive or basal response to stress and acquired or induced response to stress

(Larkindale and Vierling 2008; Larkindale et al. 2005). While the HS response is perhaps the best-studied abiotic stressor, most of our knowledge of this important response has been conducted in the model species, *Arabidopsis*, or in important crop species such as maize or rice. There is a general lack of understanding of the levels and types of variation of HS responses in native (non-crop) plant species (French, Robinson, and Lia 2017; Amano, Iida, and Kosuge 2012).

Over the last ten years the Waters lab has conducted extensive studies on the comparative thermotolerance of the *Boecheira* genus. *Boecheira* is a large genus of over 100 species within the Brassicaceae (Hay et al. 2023). The *Boecheira* have undergone a relatively recent expansion and are found throughout the Western US and have their center of diversity within California (Windham per com). In a number of studies, the Waters lab has examined the heat stress response and thermotolerance of a group of *Boecheira* species that inhabit a variety of climatic regions within California. *Boecheira arcuata* is found from the peninsular ranges in San Diego to the coast. *Boecheira californica* is found sympatrically to *B. arcuata* and is a polyploid hybrid of *B. arcuata* and other *Boecheira*. A *Boecheira* triploid that was previously identified as *Boecheira depauperata* is found only at high elevations within the Sierra Nevada range and is known only from a few populations (Hay et al. 2023). *Boecheira johnstonii* (hirshbergiae) is a rare and endangered species found only in or near Cuyamaca State Park. *Boecheira perennans* is found in the eastern (drier) slopes of the peninsular range and into the Anza Borrego and Sonoran deserts. These distinct habitats within the same genus provide an excellent natural experiment with which to examine the evolution of thermotolerance and the heat shock response.

All of the *Boechea* species are more tolerant to heat stress than is the model plant species, *Arabidopsis* ecotype Columbia (Gallas and Waters 2015; Halter et al. 2017; Kannan et al. 2018; Pierce, Halter, and Waters 2020). This is not too surprising as *Arabidopsis* is not known for its high tolerance to abiotic stress. However, what was surprising was that the *Boechea* species all had distinct HS responses and patterns of thermotolerance. Among the *Boechea* species studied, the triploid *Boechea* sp (formerly *depauperata*) has the highest level of tolerance to high heat, high light, and the combination of high light and high heat (Gallas and Waters 2015; Halter et al. 2017; Kannan et al. 2018; Pierce, Halter, and Waters 2020). It was somewhat surprising to find that *B. perennans*, that is native to the deserts in San Diego and Arizona, was not the most heat-tolerant of the *Boechea* species. However, it is well known that UV stress is also a potent abiotic stressor and that HS and UV stress share common gene expression pathways (Jenkins, Suzuki, and Mount 1997). The high elevation of the *Boechea* triploid (above 8,000 ft) may be the primary driving force in its tolerance to HS. The unusual nature of the HS response of the *Boechea* triploid was further confirmed when its gene expression response to HS was examined (Kannan et al. 2018; Pierce, Halter, and Waters 2020). The triploid expresses a large number of genes not found in other species and maybe these species-specific genes that provide its exceptional thermotolerance.

A comparison of the heat-stress responses of the two diploid San Diego *Boechea* species merits a closer examination. These two species share a small region of overlapping habitat at the peaks of the peninsular range near Julian, CA. However, *B. arcuata* is found on the western (cooler, wetter sides) slopes and *B. perennans* is found on the eastern (hotter,

drier) slopes. These two species differed in their responses to basal stress and acquired stress (Yeh et al. 2012; Kotak et al. 2007). These differences suggest that there could be changes in gene expression or gene content that underlie these organismal phenotypes. One of the major motivations for these studies was to examine the complete genomes of these related species which have adapted to distinct climates.

Advances in whole genome sequencing have significantly expanded our understanding of plant responses to abiotic stresses, which crucially impact agricultural productivity. Employing transcriptomic analyses in particular, have highlighted the dynamic nature of a plethora of genes responsive to various abiotic stresses and the expression of these genes in response to environmental challenges as well as the plants' ability to enhance survival and growth under such conditions (X. Wang et al. 2020; Sahoo et al. 2020; Azad et al. 2024). Here, I examine gene expression of the heat stress response in *B. arcuata* and *B. perennans* species and identify evolutionary patterns of genes responsive to heat stress.

Materials and Methods

Transcriptome analyses

The plants were grown and the samples were collected as described by Halter et al. (Halter et al. 2017). To perform a comprehensive transcriptome analysis of *B. arcuata* and *B. perennans*, RNA sequencing data were obtained from the Joint Genome Institute (JGI). These data were from shoot tissue samples under both basal heat stress and control conditions for a total of three biological replicates per species. Transcriptome read counts

were quantified against the reference transcriptomes of the respective *Boechera* species using Kallisto v0.48.0 (Bray et al. 2016). To visualize general expression trends, the variance-stabilized count data were used in R.

For the identification of differentially expressed transcripts across heat stress and normal conditions, I applied DESeq2 v1.4.2.0 in R (Love, Huber, and Anders 2014). Significant genes were defined as transcripts with a false discovery rate adjusted p -value of < 0.05 and a $|\log_2 \text{fold change}| > 2$. The plots were generated using ggplot2 in R. Any genes with a $\log_2\text{FoldChange} > 0$ were considered “upregulated” and genes with $\log_2\text{FoldChange} < 0$ were considered “downregulated”. To identify orthologs of the upregulated and downregulated genes in another *Boechera* species and determine their uniqueness or shared status, the previously generated OrthoFinder results from chapter 2 were utilized.

Gene Ontology enrichment analysis

Gene Ontology (GO) enrichment analysis was performed on the shared and unique *B. arcuata* and *B. perennans* upregulated and downregulated genes using the corresponding genome annotations to identify enriched GO terms and a $p\text{-value} = 0.01$ was considered as the cut-off value. This analysis was performed at each of three GO classes: biological process (BP), molecular functions (MF), and cellular components (CC).

Analysis of positive selection

To investigate the evolutionary patterns and selection pressures on the upregulated and expanded genes in *B. arcuata* and *B. perennans*, I used the Datamonkey server to

perform selection tests (Weaver et al. 2018; Kosakovsky Pond et al. 2019). The genes of interest were first identified based on their upregulation under basal heat stress conditions and their expansion in each species compared to the other species and *A. thaliana*.

Peptide sequences of the orthologous copies of selected genes were aligned using MAFFT (Katoh, Rozewicki, and Yamada 2019) at the RevTrans (Wernersson and Pedersen 2003) site to ensure accurate codon-based alignment. The alignment images were generated in JALVIEW (Waterhouse et al. 2009) using CLUSTAL colors for visual clarity. To detect positive selection, I employed two different methodologies: MEME (Mixed Effects Model of Evolution) and BUSTED (Branch-Site Unrestricted Statistical Test for Episodic Diversification) (Murrell et al. 2015). These tools identify amino acid sites under positive selection, with MEME focusing on site-specific selection pressures and BUSTED evaluating episodic selection across entire branches. MEME (Murrell et al. 2012) was used to detect individual sites under positive selection by testing for site-to-site variation in the selective pressure. Sites with a p-value < 0.05 were considered to be evolving under positive selection. BUSTED was applied to test for evidence of episodic positive selection at the gene level, identifying branches and sites that have experienced positive selection. Sites with a p-value < 0.05 were considered to be under positive selection.

Phylogenetic trees were constructed using aBSREL (adaptive Branch-Site Random Effects Likelihood) (Smith et al. 2015) on the Datamonkey server. This method evaluates the selective pressures on specific branches of the phylogeny, allowing for the identification of branches undergoing positive selection. The aligned gene sequences were

then analyzed using IQTREE2 v2.2.2.6 (Minh et al. 2020) to generate maximum likelihood phylogenetic trees using the default parameters.

Results

Transcript expression differs between normal and basal heat stress conditions

Differential gene expression analysis between control and basal heat stress (BHS) conditions within *B. arcuata* revealed 1269 differentially expressed genes, using an adjusted p -value = 0.05 and a $|\log_2 \text{fold change}| = 2$. Of these, 584 genes were upregulated and 685 genes were downregulated. In *B. perennans*, 1364 differentially expressed genes were identified, 553 genes were upregulated and 811 genes were downregulated (Table 3.1). Moreover, there was a distinct division of gene expression between the control and heat stress conditions within each species (Figure 3.1A, B). The data indicated a robust upregulation of specific genes under heat stress in both species.

Table 3.1. Differential gene expression analysis summary. DE (differentially expressed) genes were identified as those with $|\log_2 \text{fold change}| > 2$ and an adjusted p -value of < 0.05 . Upregulated genes were defined as having a log fold change > 0 . Downregulated genes were defined as having a log fold change < 0 .

Description	<i>B. arcuata</i>	<i>B. perennans</i>
Total DE genes	1269	1364
Total upregulated genes	584	553
Total downregulated genes	685	811

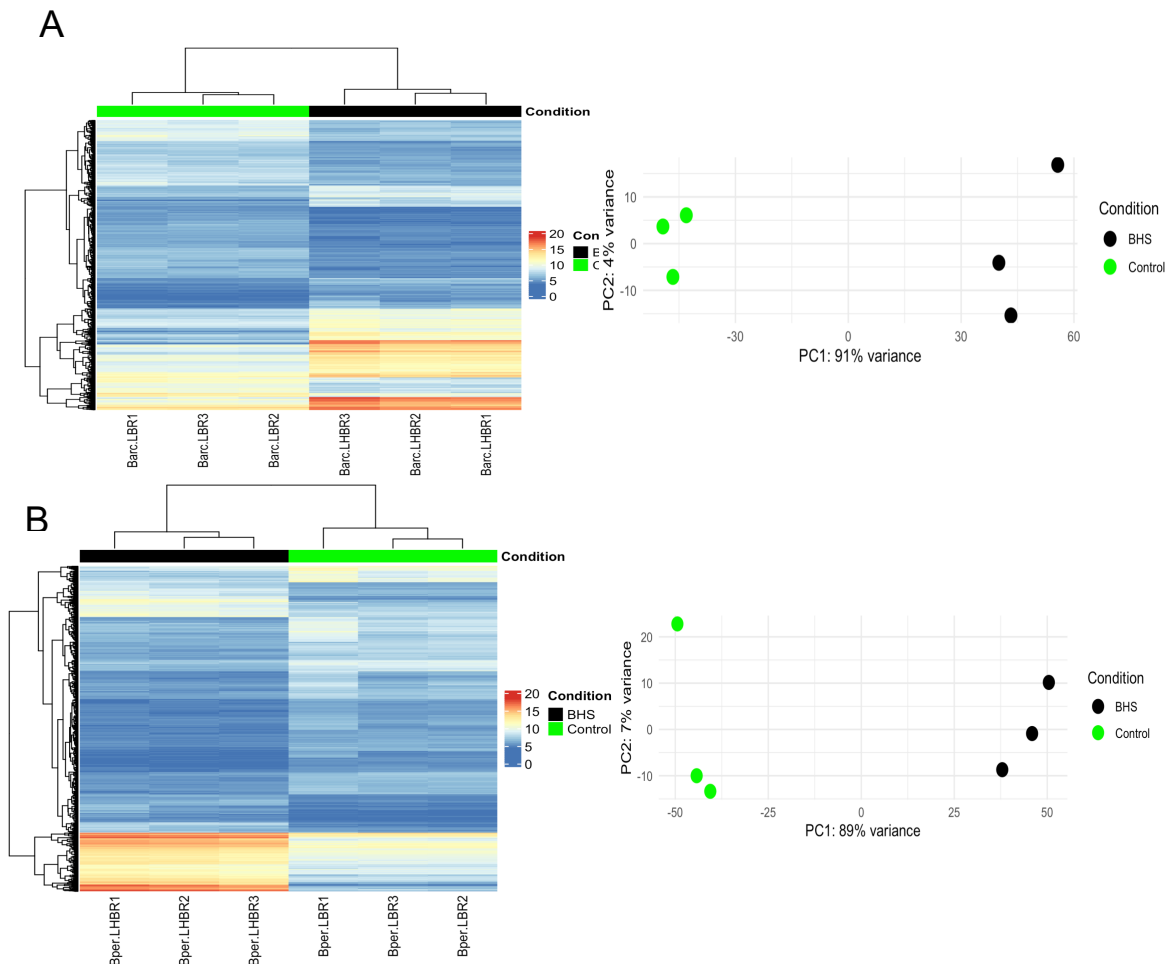


Figure 3.1. Differential gene expression across control and basal heat stress (BHS) conditions. Heatmaps and Principal component Analysis (PCA) showing the variance-stabilized counts of the transcripts with differential expression across control and BHS in (A) *B. arcuata*, (B) *B. perennans*.

Focusing on the top 50 most upregulated and downregulated genes, I identified orthologs in *A. thaliana* and determined their corresponding PFAM IDs and functional annotations, which helped classify these genes into specific protein families (Table 3.2-3.5). The top 50 upregulated genes in *B. arcuata* included genes involved in the heat shock response and protein homeostasis under basal heat stress conditions. For example, genes Boarc.1G275300.1, Boarc.4G124100.1, and Boarc.2G030000.1, all orthologous to AT1G16030.1 from *A. thaliana*, are associated with PFAM-ID PF00012, identifying them

as HSP70 proteins known for their role in assisting protein folding and protection against stress-induced damage. Additionally, I found multiple instances of HSP20 gene upregulation, such as Boarc.5G219500.1, Boarc.5G241600.1, and Boarc.5G241700.1, which are linked to PFAM-ID PF00011 and are orthologous to AT1G53540.1 from *A. thaliana*. Other upregulated genes included Boarc.4G281900.1 (PF02179, BAG family), which plays a role in apoptosis regulation, and Boarc.6G063800.1 (PF13414, TPR_11 domain), indicating involvement in protein-protein interactions, possibly related to stress signaling pathways. Genes such as Boarc.3G329700.1 and Boarc.4G047500.1, with PFAM-IDs PF01435 and PF00004 respectively, represent peptidase and AAA domain-containing proteins, which are typically involved in protein degradation and cellular processes essential during stress. The top 50 downregulated genes encompassed a variety of biological functions, indicating a repression of these pathways under heat stress. The gene Boarc.1G254600.1, orthologous to AT1G57943.1 in *A. thaliana* with PFAM ID PF16913, represents the PUNUT family, involved in biological processes. Other genes like Boarc.2G118900.1 and Boarc.3G389700.1, linked to the B3 and glycoside hydrolase families respectively, showed a downregulation of developmental and metabolic processes under heat stress. Additionally, genes with functions in signal transduction, such as Boarc.2G040200.1 (PF05678, VQ motif) and Boarc.2G365300.1 (PF03106, WRKY transcription factor), indicated a broad suppression of gene regulatory networks under BHS conditions. The diversity of PFAM domains represented, including Glyco_hydro_16 (PF00722), LEA_2 (PF03168), and PP2C (PF00481), reflected the downregulation of genes involved in various metabolic and regulatory pathways.

Table 3.2. Top 50 upregulated genes in *B. arcuata* under basal heat stress condition

	Gene-ID	<i>A. thaliana</i> Ortholog	PFAM-ID	Name
1	Boarc.1G275300.1	AT1G16030.1	PF00012	HSP70
2	Boarc.4G124100.1	AT2G32120.1	PF00012	HSP70
3	Boarc.5G219500.1	AT1G53540.1	PF00011	HSP20
4	Boarc.3G174600.1	AT3G16050.1	PF01680	SOR_SNZ
5	Boarc.5G241600.1	AT1G53540.1	PF00011	HSP20
6	Boarc.4G281900.1	AT2G46240.1	PF02179	BAG
7	Boarc.2G308400.1	AT1G74310.1	PF07728	AAA_5
8	Boarc.5G241700.1	AT1G53540.1	PF00011	HSP20
9	Boarc.6G063800.1	AT4G12400.2	PF13414	TPR_11
10	Boarc.7G258300.1	AT4G27670.1	PF00011	HSP20
11	Boarc.6G151800.1	AT5G23950.2	PF00168	C2
12	Boarc.1G153900.1	AT1G53540.1	PF00011	HSP20
13	Boarc.5G414700.1	AT3G63310.1	PF01027	UPF0005
14	Boarc.2G183400.1		PF00011	HSP20
15	Boarc.3G134400.1	AT1G16030.1	PF00012	HSP70
16	Boarc.3G329700.1	AT5G51740.2	PF01435	Peptidase_M48
17	Boarc.6G272300.1	AT5G11590.1	PF00847	AP2
18	Boarc.3G339200.1	AT5G52640.1	PF02518	HATPase_c
19	Boarc.4G047500.1	AT2G25140.1	PF00004	AAA
20	Boarc.2G030000.1	AT1G16030.1	PF00012	HSP70
21	Boarc.1G068700.1	AT1G07400.1	PF00011	HSP20
22	Boarc.6G292000.1	AT4G32208.1	PF00012	HSP70
23	Boarc.6G311500.1	AT5G07330.1		
24	Boarc.5G140900.1	AT2G20560.1	PF00226	DnaJ
25	Boarc.6G237300.1	AT5G15250.1	PF00004	AAA
26	Boarc.6G333700.1	AT5G05410.1	PF00847	AP2
27	Boarc.2G015800.1	AT1G54050.1	PF00011	HSP20
28	Boarc.7G230600.1	AT4G25200.1	PF00011	HSP20
29	Boarc.1G221900.1	AT1G44414.1	PF17032	zinc_ribbon_15
30	Boarc.7G083400.1	AT5G35320.1		

31	Boarc.3G285500.1	AT1G58450.1	PF00254	FKBP_C
32	Boarc.6G086600.1	AT4G10250.1	PF00011	HSP20
33	Boarc.3G094400.1	AT3G08970.1	PF00226	DnaJ
34	Boarc.7G002600.1	AT1G09080.1	PF00012	HSP70
35	Boarc.7G248300.1	AT4G26770.1	PF01148	CTP_transf_1
36	Boarc.3G189900.1	AT3G17611.1	PF00641	zf-RanBP
37	Boarc.6G292200.1	AT5G09570.1	PF06747	CHCH
38	Boarc.3G129300.1	AT3G12050.1	PF09229	Aha1_N
39	Boarc.2G216000.1	AT1G66080.1	PF05603	DUF775
40	Boarc.4G291700.1		PF00481	PP2C
41	Boarc.7G070100.1	AT5G22480.1	PF03367	zf-ZPR1
42	Boarc.7G121000.1	AT4G15420.2	PF03152	UFD1
43	Boarc.2G198500.1	AT1G35660.1		
44	Boarc.6G266900.1	AT5G12020.1	PF00011	HSP20
45	Boarc.2G030100.1	AT1G16040.1	PF06699	PIG-F
46	Boarc.6G235400.1	AT5G15450.1	PF07728	AAA_5
47	Boarc.3G185000.1			
48	Boarc.2G152700.1	AT1G30070.2	PF09032	Siah-Interact_N
49	Boarc.1G032500.1	AT1G03970.1	PF00170	bZIP_1
50	Boarc.5G008300.1	AT5G47120.1	PF01027	UPF0005

Table 3.3. Top 50 downregulated genes in *B. arcuata* under basal heat stress condition

	Gene-ID	A. thaliana Ortholog	PFAM-ID	Name
1	Boarc.1G254600.1	AT1G57943.1	PF16913	PUNUT
2	Boarc.2G118900.1	AT1G25560.1	PF02362	B3
3	Boarc.3G389700.1	AT5G57685.1		
4	Boarc.7G003000.1	AT2G06265.1	PF06839	zf-GRF
5	Boarc.3G388600.1	AT4G25810.1	PF00722	Glyco_hydro_16
6	Boarc.3G318600.1	AT5G50570.1	PF03110	SBP
7	Boarc.6G034500.1	AT1G15890.1	PF00931	NB-ARC
8	Boarc.7G162400.1	AT4G18540.1		
9	Boarc.7G264800.1	AT4G28240.1	PF12609	DUF3774
10	Boarc.3G078800.1	AT3G07040.1	PF00931	NB-ARC
11	Boarc.2G258900.1	AT1G69920.1	PF13417	GST_N_3
12	Boarc.2G040200.1	AT1G17147.1	PF05678	VQ
13	Boarc.7G017200.1	AT5G41170.1	PF13041	PPR_2
14	Boarc.3G423300.1	AT5G60900.1	PF08276	PAN_2
15	Boarc.2G323100.1	AT1G03130.1	PF02531	PsaD
16	Boarc.1G031400.1	AT1G03870.1	PF02469	Fasciclin
17	Boarc.1G314100.1		PF00931	NB-ARC
18	Boarc.7G345800.1	AT4G35940.1		
19	Boarc.7G315500.1	AT4G33050.5	PF00612	IQ
20	Boarc.7G156900.1	AT4G18210.1	PF16913	PUNUT
21	Boarc.3G224000.1	AT3G20590.1	PF03168	LEA_2
22	Boarc.3G018600.1	AT3G02550.1	PF03195	DUF260
23	Boarc.4G008900.1	AT2G02955.1	PF11781	RRN7
24	Boarc.5G193200.1		PF13676	TIR_2
25	Boarc.1G154800.1	AT1G53430.1	PF13855	LRR_8
26	Boarc.2G193300.1	AT1G35140.1	PF04674	Phi_1
27	Boarc.5G339500.1	AT3G55840.1	PF07231	Hs1pro-1_N
28	Boarc.1G318200.1	AT1G06580.1	PF13041	PPR_2
29	Boarc.3G324200.1	AT1G12210.1		
30	Boarc.2G338700.1	AT1G77210.1	PF07690	MFS_1

31	Boarc.1G042600.1		PF00069	Pkinase
32	Boarc.2G365300.1	AT1G80840.1	PF03106	WRKY
33	Boarc.2G288500.1	AT1G72610.1	PF07883	Cupin_2
34	Boarc.1G335000.1	AT1G63855.4	PF10294	Methyltransf_16
35	Boarc.2G354800.1	AT1G78820.1	PF08276	PAN_2
36	Boarc.6G003700.1			
37	Boarc.5G239500.1	AT3G45960.2	PF03330	DPBB_1
38	Boarc.3G181000.1	AT3G16710.1	PF13041	PPR_2
39	Boarc.6G031000.1		PF03242	LEA_3
40	Boarc.3G423200.1		PF07714	Pkinase_Tyr
41	Boarc.5G054000.1	AT1G12615.1	PF13041	PPR_2
42	Boarc.1G098700.1	AT1G10100.1	PF00010	HLH
43	Boarc.2G208600.1		PF00069	Pkinase
44	Boarc.3G255900.1	AT3G23730.1	PF00722	Glyco_hydro_16
45	Boarc.7G068000.1	AT5G37490.1	PF04564	U-box
46	Boarc.5G357700.1	AT3G57450.1		
47	Boarc.4G054600.1	AT2G25735.1		
48	Boarc.1G339800.1	AT1G64100.2	PF13041	PPR_2
49	Boarc.5G269300.1	AT1G15890.1	PF13855	LRR_8
50	Boarc.5G331000.1	AT3G55050.1	PF00481	PP2C

The top 50 upregulated genes in *B. perennans* under basal heat stress conditions, indicated that a significant number of heat shock proteins (HSPs) were highly expressed. For instance, Gene-ID Boper.2G028900.1, Boper.1G271900.1, and Boper.4G134100.1, corresponding to the *A. thaliana* ortholog AT1G16030.1, were identified with PFAM-ID PF00012, categorizing them as HSP70 proteins, which are known for their role in aiding protein folding and stress response. Another upregulated gene, Boper.4G059800.1, orthologous to AT2G26150.1 with PFAM-ID PF00447, is characterized as an HSF DNA-binding protein, indicating a role in transcriptional regulation under heat stress. Proteins with TPR domains, such as those encoded by Boper.6G066300.1 (PF00515) and Boper.3G285300.1 (PF13414), were also notable among the upregulated genes, which are typically involved in protein-protein interactions and are critical in stress response mechanisms. I also identified genes such as Boper.5G141100.1 and Boper.3G322100.1, which are associated with PFAM-IDs corresponding to DnaJ and Peptidase_M48 protein families, respectively, involved in protein stabilization and degradation. Several genes like Boper.4G149700.1 were associated with PFAM-ID PF00249, denoting Myb DNA-binding proteins that are involved in gene regulatory processes. Additionally, a number of genes in the list, such as Boper.3G238600.1 and Boper.2G187200.1, contain the PPR motif (PFAM-ID PF13041 and PF13041), which is typically associated with RNA metabolism within organelles. genes such as Boper.6G367100.1, which encodes a protein with the PFAM-ID PF00112, known as Peptidase_C1, suggesting roles in proteolysis. Several genes associated with the NB-ARC domain (PF00931), including Boper.1G304900.1 and Boper.6G054400.1, are involved in plant defense mechanisms and were notably

downregulated. The top 50 down regulated genes in *B. perennans*, showed genes encoding for proteins with LRR domains (PF13855), like Boper.1G035200.1 and Boper.1G303900.1, which are often involved in protein-protein interactions and signaling. The SBP-box protein encoded by Boper.3G310200.1 (PF03110) is a plant-specific transcription factor that was also downregulated, as was Boper.6G005000.1 with the NB-ARC domain. Several genes without an *A. thaliana* ortholog listed, such as Boper.3G051400.1, still had functional annotations like the LRR_8 domain, suggesting they may play a role in signal transduction or immune response. Boper.2G365700.1, associated with the PF13947 GUB_WAK_bind domain, is potentially involved in cell wall-associated kinase activity and signaling. Other notable genes included Boper.5G392200.1 (PF03514, GRAS family transcription factor), Boper.3G145900.1 (PF02045, CBFB_NFYA, a component of a transcription factor complex), and Boper.5G054300.1 (PF02309, AUX_IAA, an auxin-responsive protein), all of which are typically important for plant growth and development.

Table 3.4. Top 50 upregulated genes in *B. perennans* under basal heat stress condition

	Gene-ID	<i>A. thaliana</i> Ortholog	PFAM-ID	Name
1	Boper.4G059800.1	AT2G26150.1	PF00447	HSF_DNA-bind
2	Boper.2G028900.1	AT1G16030.1	PF00012	HSP70
3	Boper.2G310400.1	AT1G74310.1	PF07728	AAA_5
4	Boper.1G271900.1	AT1G16030.1	PF00012	HSP70
5	Boper.6G066300.1	AT4G12400.2	PF00515	TPR_1
6	Boper.4G134100.1	AT2G32120.1	PF00012	HSP70
7	Boper.3G278300.1			
8	Boper.4G286400.1	AT2G46240.1	PF02179	BAG
9	Boper.5G141100.1	AT2G20560.1	PF00226	DnaJ
10	Boper.3G322100.1	AT5G51740.2	PF01435	Peptidase_M48
11	Boper.3G285300.1	AT1G58450.1	PF13414	TPR_11
12	Boper.H057900.1	AT5G12020.1	PF00011	HSP20
13	Boper.3G168300.1	AT3G16050.1	PF01680	SOR_SNZ
14	Boper.3G332200.1	AT5G52640.1	PF00183	HSP90
15	Boper.5G246300.1	AT1G53540.1	PF00011	HSP20
16	Boper.2G225400.1	AT5G12020.1	PF00011	HSP20
17	Boper.1G068500.1	AT1G07400.1	PF00011	HSP20
18	Boper.7G065500.1	AT5G37670.1	PF00011	HSP20
19	Boper.7G083500.1	AT5G35320.1		
20	Boper.2G014800.1	AT1G54050.1	PF00011	HSP20
21	Boper.3G127700.1	AT1G16030.1	PF00012	HSP70
22	Boper.H028700.1	AT3G12050.1	PF08327	AHSA1
23	Boper.2G187100.1		PF00011	HSP20
24	Boper.6G244900.1	AT5G15250.1	PF00004	AAA
25	Boper.6G159200.1	AT5G23950.2	PF00168	C2
26	Boper.3G405900.1	AT5G59720.1	PF00011	HSP20
27	Boper.5G219800.1	AT1G53540.1	PF00011	HSP20
28	Boper.7G220000.1			
29	Boper.1G152000.1	AT1G53540.1	PF00011	HSP20
30	Boper.5G004300.1	AT5G47610.1	PF13639	zf-RING_2

31	Boper.5G300700.1	AT3G51910.1	PF00447	HSF_DNA-bind
32	Boper.7G255600.1	AT4G27670.1	PF00011	HSP20
33	Boper.2G046500.1	AT1G17870.1		
34	Boper.7G120800.1	AT4G15420.2	PF03152	UFD1
35	Boper.3G238600.1	AT3G23330.1	PF13041	PPR_2
36	Boper.4G106400.1	AT2G29500.1	PF00011	HSP20
37	Boper.5G044100.1	AT2G01050.1	PF14392	zf-CCHC_4
38	Boper.3G098100.1	AT3G10020.1		
39	Boper.7G228000.1	AT4G25200.1	PF00011	HSP20
40	Boper.4G149700.1	AT2G33610.1	PF00249	Myb_DNA-binding
41	Boper.1G162100.1	AT1G52560.1	PF00011	HSP20
42	Boper.4G048300.1	AT2G25140.1	PF07728	AAA_5
43	Boper.6G090100.1	AT4G10250.1	PF00011	HSP20
44	Boper.4G221000.1	AT2G40240.1	PF01535	PPR
45	Boper.2G187200.1	AT1G33350.1	PF13041	PPR_2
46	Boper.6G274700.1	AT5G12020.1	PF00011	HSP20
47	Boper.7G210000.1	AT4G23493.1		
48	Boper.7G184700.1	AT4G21320.1	PF02679	ComA
49	Boper.3G254000.1	AT3G24500.1	PF01381	HTH_3
50	Boper.3G184100.1	AT3G17611.1	PF01694	Rhomboid

Table 3.5. Top 50 downregulated genes in *B. perennans* under basal heat stress condition

	Gene-ID	A. thaliana Ortholog	PFAM-ID	Name
1	Boper.6G367100.1		PF00112	Peptidase_C1
2	Boper.1G304900.1		PF00931	NB-ARC
3	Boper.3G310200.1	AT5G50570.1	PF03110	SBP
4	Boper.7G013700.1		PF00004	AAA
5	Boper.6G054400.1		PF00931	NB-ARC
6	Boper.1G035200.1	AT1G12210.1	PF13855	LRR_8
7	Boper.1G303900.1		PF13855	LRR_8
8	Boper.1G274900.1		PF00931	NB-ARC
9	Boper.3G051400.1		PF13855	LRR_8
10	Boper.1G290900.1	AT1G06580.1	PF13041	PPR_2
11	Boper.1G116500.1			
12	Boper.2G298700.1	AT1G03400.2	PF14226	DIOX_N
13	Boper.2G365700.1	AT1G17910.1	PF13947	GUB_WAK_bind
14	Boper.6G005000.1	AT1G69550.1	PF00931	NB-ARC
15	Boper.6G016300.1		PF00931	NB-ARC
16	Boper.3G423900.1	AT5G61440.1	PF00085	Thioredoxin
17	Boper.1G307400.1		PF00931	NB-ARC
18	Boper.6G039100.1	AT1G15890.1	PF13855	LRR_8
19	Boper.5G033300.1		PF07649	C1_3
20	Boper.1G323400.1		PF00931	NB-ARC
21	Boper.5G161200.1		PF03107	C1_2
22	Boper.5G392200.1	AT3G60630.1	PF03514	GRAS
23	Boper.3G232300.1		PF00560	LRR_1
24	Boper.1G325100.1		PF13676	TIR_2
25	Boper.2G294400.1	AT1G17590.1	PF02045	CBFB_NFYA
26	Boper.3G145900.1	AT3G14020.1	PF02045	CBFB_NFYA
27	Boper.5G054300.1	AT5G43700.1	PF02309	AUX_IAA
28	Boper.5G229000.1	AT1G65850.2	PF00560	LRR_1
29	Boper.1G097100.1	AT1G10100.1	PF00010	HLH
30	Boper.5G158900.1		PF07649	C1_3

31	Boper.6G333000.1	AT5G06100.2	PF00249	Myb_DNA-binding
32	Boper.2G069000.1	AT1G20190.1	PF03330	DPBB_1
33	Boper.7G281400.1	AT4G30080.1	PF06507	Auxin_resp
34	Boper.7G071000.1	AT4G00930.1		
35	Boper.1G304100.1		PF13855	LRR_8
36	Boper.2G187600.1		PF12609	DUF3774
37	Boper.4G014000.1	AT4G26190.1	PF03031	NIF
38	Boper.1G293500.1	AT1G06580.1	PF13041	PPR_2
39	Boper.1G341900.1	AT1G06580.1	PF13041	PPR_2
40	Boper.1G035400.1	AT1G12210.1	PF13855	LRR_8
41	Boper.3G191300.1		PF00560	LRR_1
42	Boper.4G258300.1	AT2G43580.1	PF00182	Glyco_hydro_19
43	Boper.7G069100.1			
44	Boper.1G325400.1	AT1G63740.2	PF00931	NB-ARC
45	Boper.4G025600.1	AT2G23200.1	PF07714	Pkinase_Tyr
46	Boper.1G193000.1	AT4G03440.1	PF12796	Ank_2
47	Boper.2G229600.1	AT4G23130.2	PF07714	Pkinase_Tyr
48	Boper.3G008300.1	AT3G01350.1	PF00854	PTR2
49	Boper.1G303700.1		PF00931	NB-ARC
50	Boper.2G365600.1		PF07714	Pkinase_Tyr

Building upon the orthogroup categorizations identified through OrthoFinder analysis between *A. thaliana*, *B. arcuata*, and *B. perennans*, the differential gene expression analysis under heat stress conditions revealed a distinct response within each species (Figure 3.2). Specifically, in *B. arcuata*, 244 genes showed upregulation and 392 showed downregulation uniquely, while *B. perennans* demonstrated a unique upregulation in 206 genes and downregulation in 402 genes.

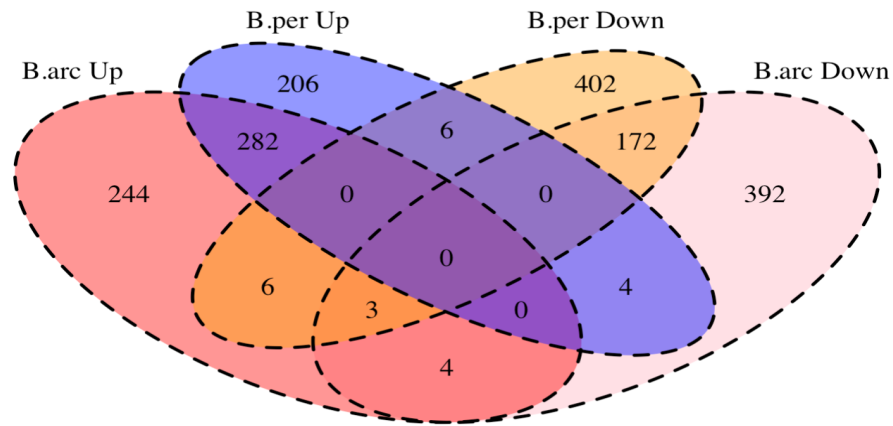


Figure 3.2. Comparison of upregulated and downregulated genes in *B. arcuata* and *B. perennans* exposed to heat stress versus control conditions. The numbers in the venn diagram indicate gene counts unique to each species or shared between them. Data derived from OrthoFinder analyses enabled correlation with previously established orthogroups.

GO enrichment analysis of unique and shared differentially expressed genes

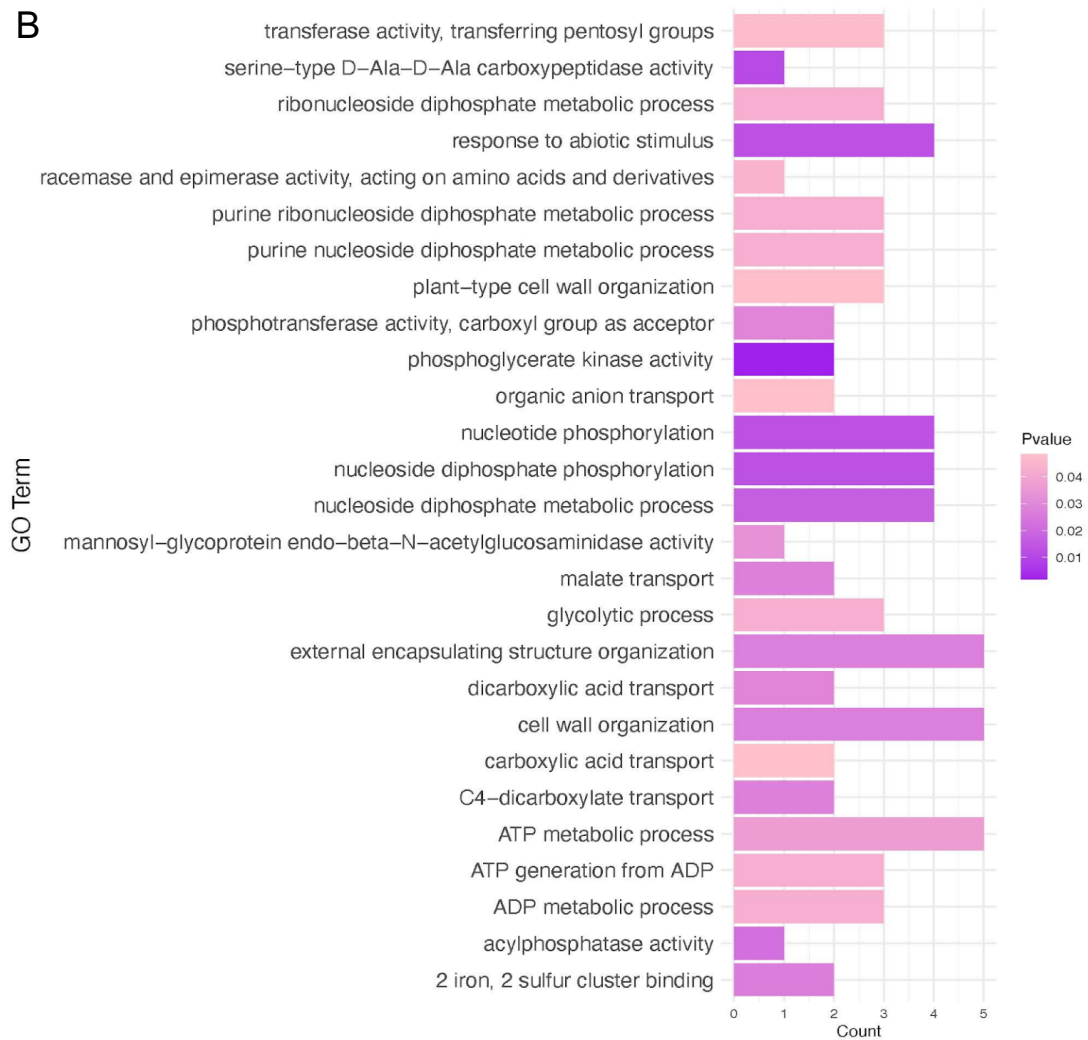
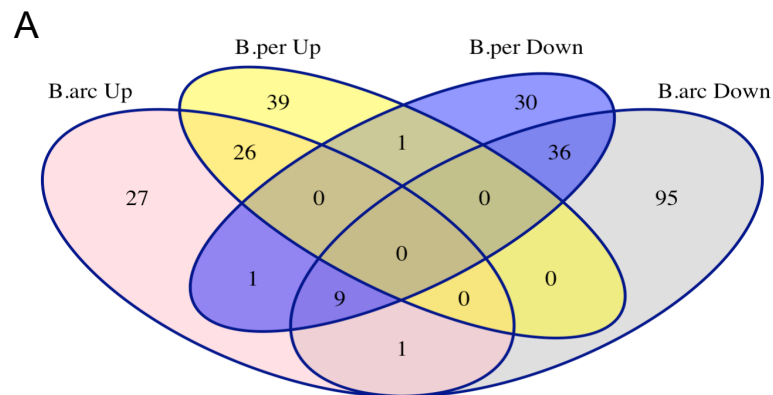
Gene Ontology (GO) enrichment was performed for genes upregulated and downregulated under heat stress conditions within *B. arcuata* and *B. perennans*. The analysis aimed to identify the shared and unique biological processes, molecular functions, and cellular components assigned to each species' response to heat stress. I then found both shared and unique GO terms across species, reflective of their potentially distinct adaptive mechanisms (Figure 3.2A). There were 27 unique GO terms associated with genes upregulated exclusively in *B. arcuata* and 95 unique GO terms associated with genes downregulated in this species. In *B. perennans*, 39 GO terms were unique to genes upregulated, and 30 GO terms were unique to genes downregulated in response to heat stress. The number of shared GO terms among upregulated genes in both species was 26 and there were 36 GO terms of downregulated genes shared between the two species.

In *B. arcuata*, the unique GO terms associated with upregulated genes, revealed significant enrichments across several biological processes (BP) and molecular functions (MF) categories. Within metabolic processes, genes associated with nucleoside diphosphate phosphorylation (GO:0006165) and ATP metabolic process (GO:0046034) were prominently enriched. In terms of transport processes, the enrichments included C4-dicarboxylate transport (GO:0015740) and malate transport (GO:0015743). I also found some enrichments related to response to stimuli, such as response to abiotic stimulus (GO:0009628), and cellular organization, such as cell wall organization (GO:0071555). Additionally, significant enrichments were observed in enzymatic activities, with examples including acylphosphatase activity (GO:0003998) and phosphoglycerate kinase activity

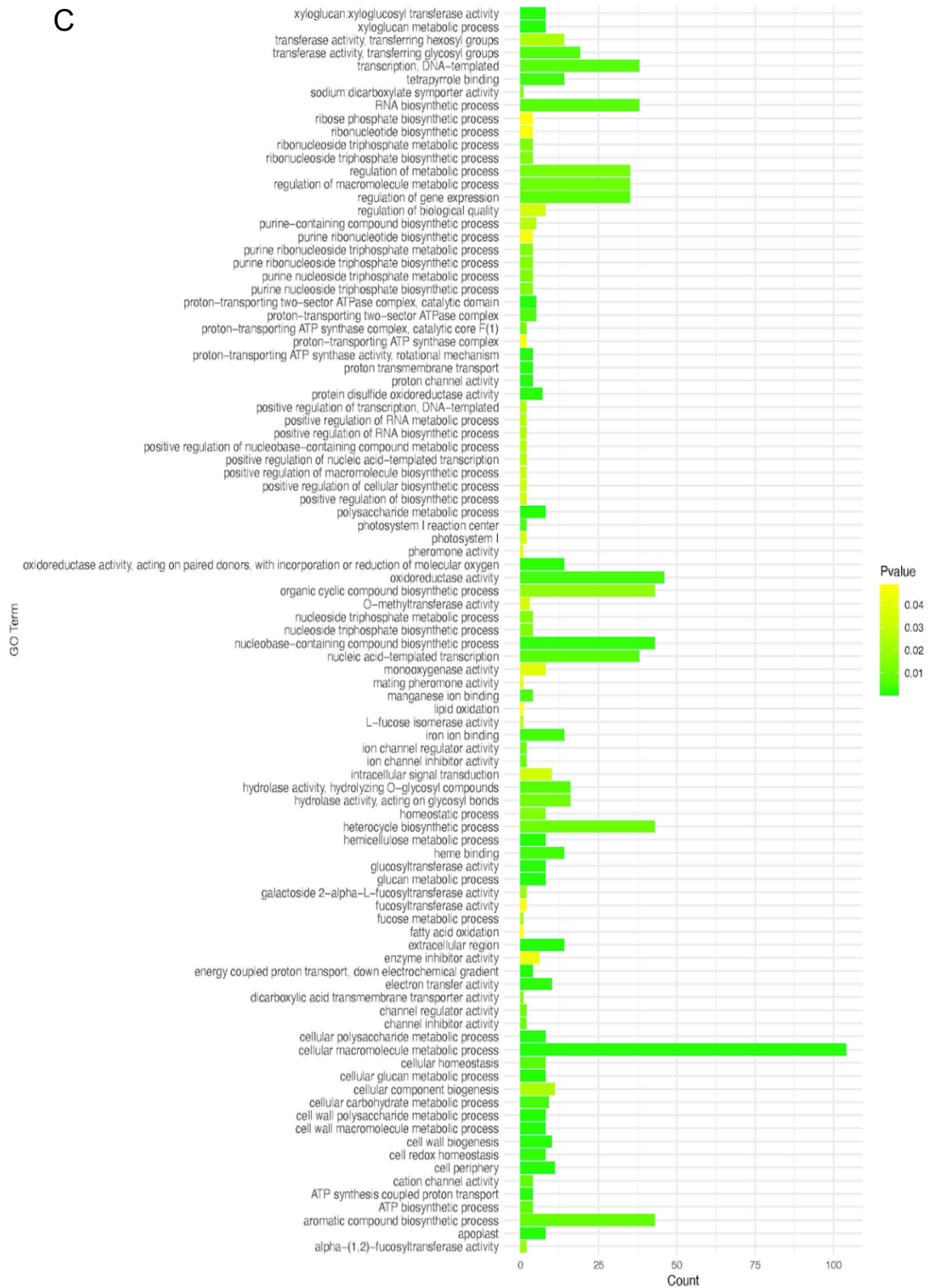
(GO:0004618). The unique GO terms associated with downregulated genes in *B. arcuata* were related to different functional categories. Significant enrichments including cell wall macromolecule metabolic process (GO:0044036) and cell wall polysaccharide metabolic process (GO:0010383), were related to cell wall associated processes. Moreover, enrichments related to energy metabolism were observed, such as ATP synthesis coupled proton transport (GO:0015986) and ATP biosynthetic process (GO:0006754). I also identified enrichments in photosystem related processes, including photosystem I reaction center (GO:0009538) and photosystem I (GO:0009522). Significant enrichments were observed in regulatory processes, such as regulation of gene expression (GO:0010468) and regulation of metabolic process (GO:0019222).

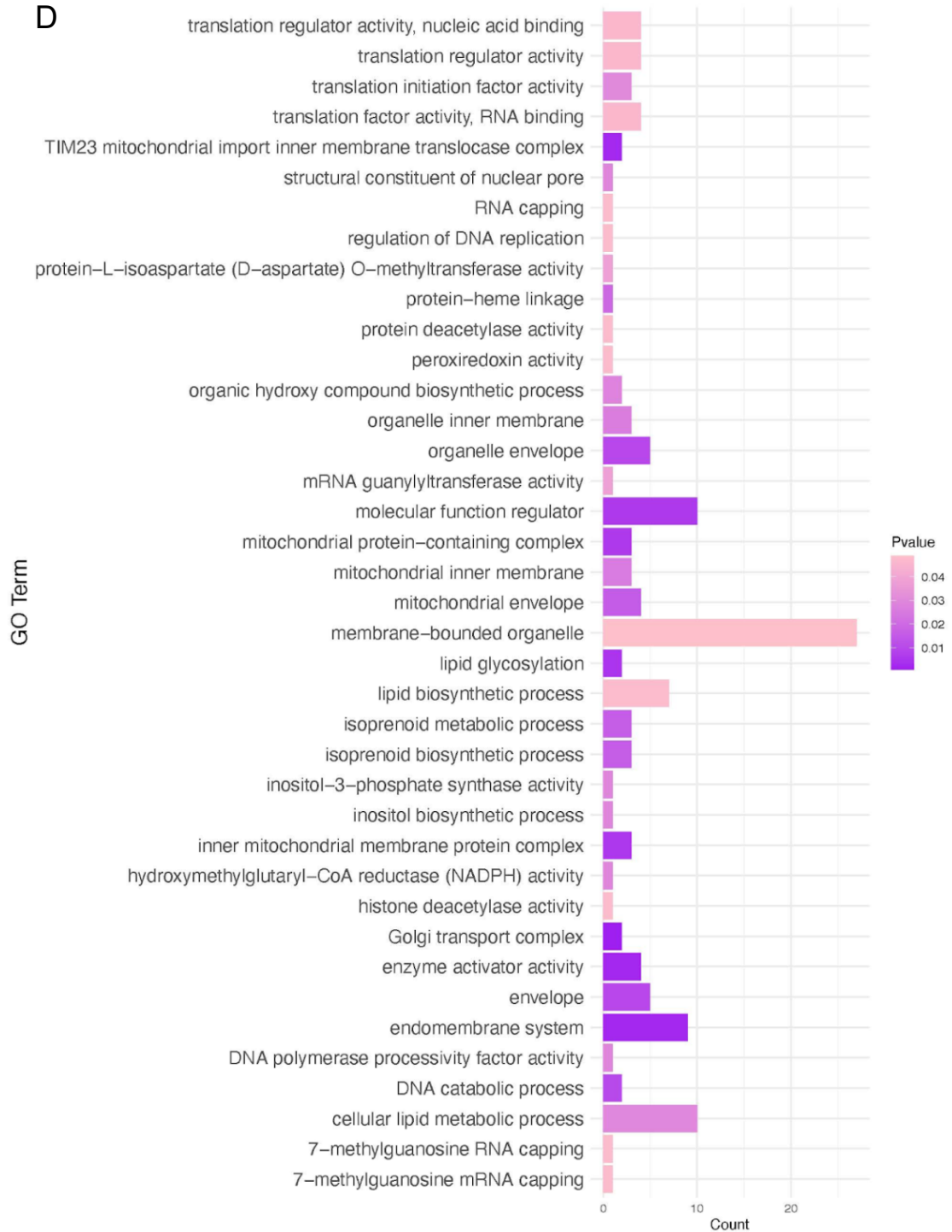
In *B. perennans*, the unique GO terms of the upregulated genes were related to enzyme activity such as enzyme activator activity (GO:0008047), hydroxymethylglutaryl-CoA reductase (NADPH) activity (GO:0004420), and protein-L-isoaspartate (D-aspartate) O-methyltransferase activity (GO:0004719), nucleic acid binding including translation regulator activity, nucleic acid binding (GO:0090079), translation factor activity, RNA binding (GO:0008135), membrane-bound structures including Golgi transport complex (GO:0017119) and endomembrane system (GO:0012505), and inner mitochondrial membrane protein complex (GO:0098800), and metabolic processes such as isoprenoid metabolic process (GO:0006720) and cellular lipid metabolic process (GO:0044255). The unique GO terms associated with downregulated genes in *B. perennans* were related to different functional categories. I found significant enrichments including signal transduction (GO:0007165), signaling (GO:0023052), cell communication (GO:0007154),

and cellular response to stimulus (GO:0051716) which were related to cellular communication and signaling. Moreover, significant enrichments were observed in transport activity including transferase activity, transferring phosphorus-containing groups (GO:0016772), gated channel activity (GO:0022836), and channel activity (GO:0015267).



C





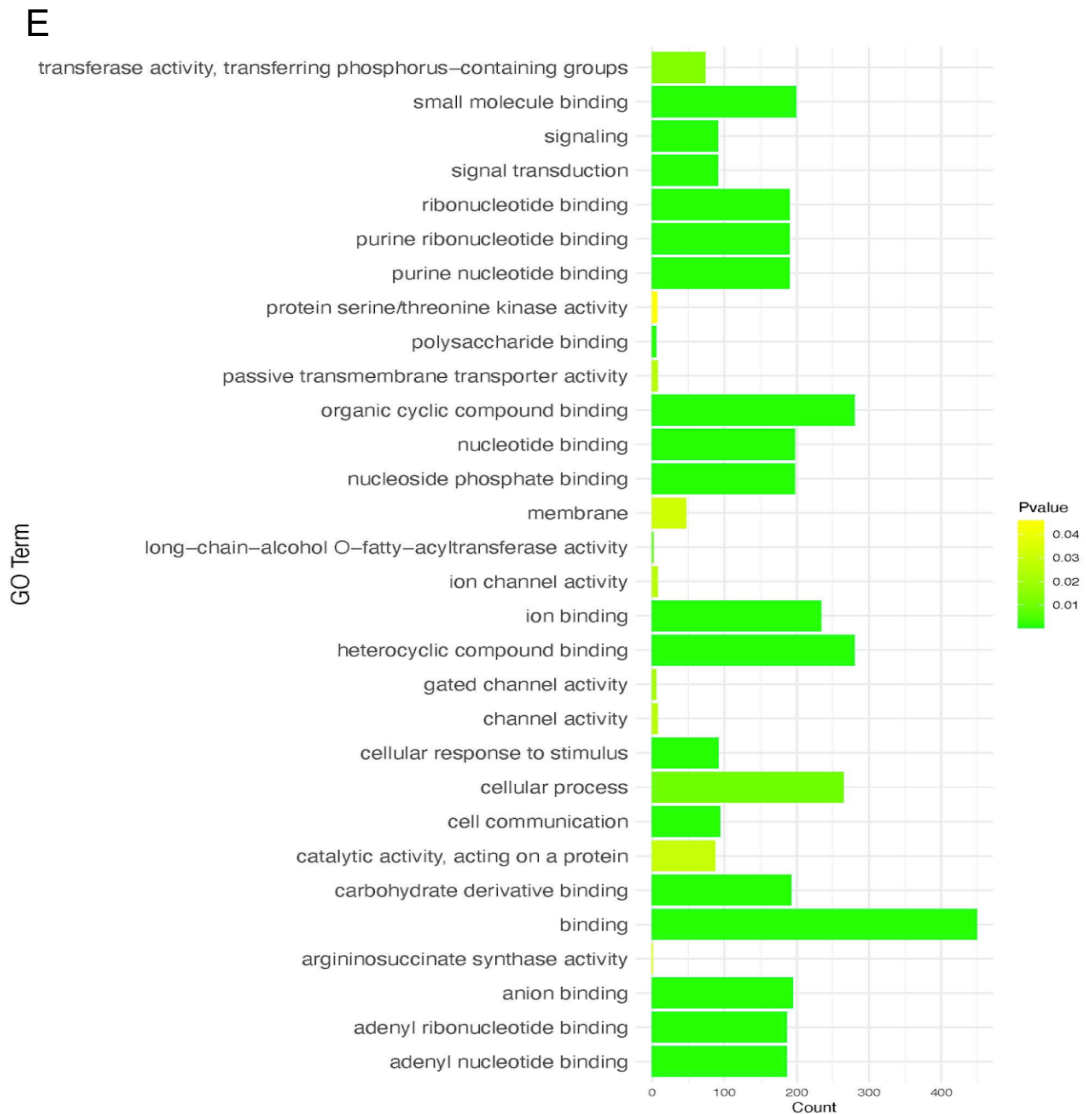


Figure 3.3. Gene Ontology (GO) enrichment analysis of upregulated and downregulated genes under basal heat stress conditions. (A) Venn diagram indicates the number of GO terms of upregulated and downregulated genes that are unique or shared across the two *Boechera* species. (B) Unique GO terms of upregulated (C) and downregulated genes in *B. arcuata*. (D) Unique GO terms of upregulated (E) and downregulated genes in *B. perennans*.

I also found the GO terms shared between upregulated genes in *B. arcuata* and *B. perennans* under heat stress conditions. I found GO terms include adenylyl-nucleotide exchange factor activity (GO:0000774), ATPase activator activity (GO:0001671), response to stress (GO:0006950), response to heat (GO:0009408), protein folding (GO:0006457), and cytoplasm (GO:0005737). These GO terms were common molecular functions, biological processes, and cellular components enriched in both species' response to heat stress, reflecting shared mechanisms. Other enriched GO terms include response to temperature stimulus (GO:0009266), protein processing (GO:0016485), and chaperone binding (GO:0051087), highlighting potential regulatory and protective roles of these genes in response to heat stress across both species. Additionally, enzyme regulator activity (GO:0030234) and unfolded protein binding (GO:0051082) indicated involvement in regulatory processes and protein quality control mechanisms. These shared GO terms provide insights into the conserved molecular responses underlying heat stress adaptation in both *B. arcuata* and *B. perennans*. The GO enrichment analysis also revealed several highly enriched GO terms shared between downregulated genes in *B. arcuata* and *B. perennans*. These included responses to endogenous stimulus (GO:0009719), hormone (GO:0009725), and auxin (GO:0009733), indicating intricate regulatory responses to various stimuli. Additionally, biological regulation (GO:0065007), regulation of cellular processes (GO:0050794), and regulation of biological processes (GO:0050789) were enriched, highlighting the importance of regulatory mechanisms in coordinating responses to heat stress. Furthermore, protein kinase activity (GO:0004672) and protein phosphorylation (GO:0006468) were enriched. Molecular functions such as protein-

disulfide reductase activity (GO:0047134), ADP binding (GO:0043531), and carbohydrate binding (GO:0030246) were also enriched, indicating alterations in redox homeostasis and nucleotide metabolism under heat stress conditions.

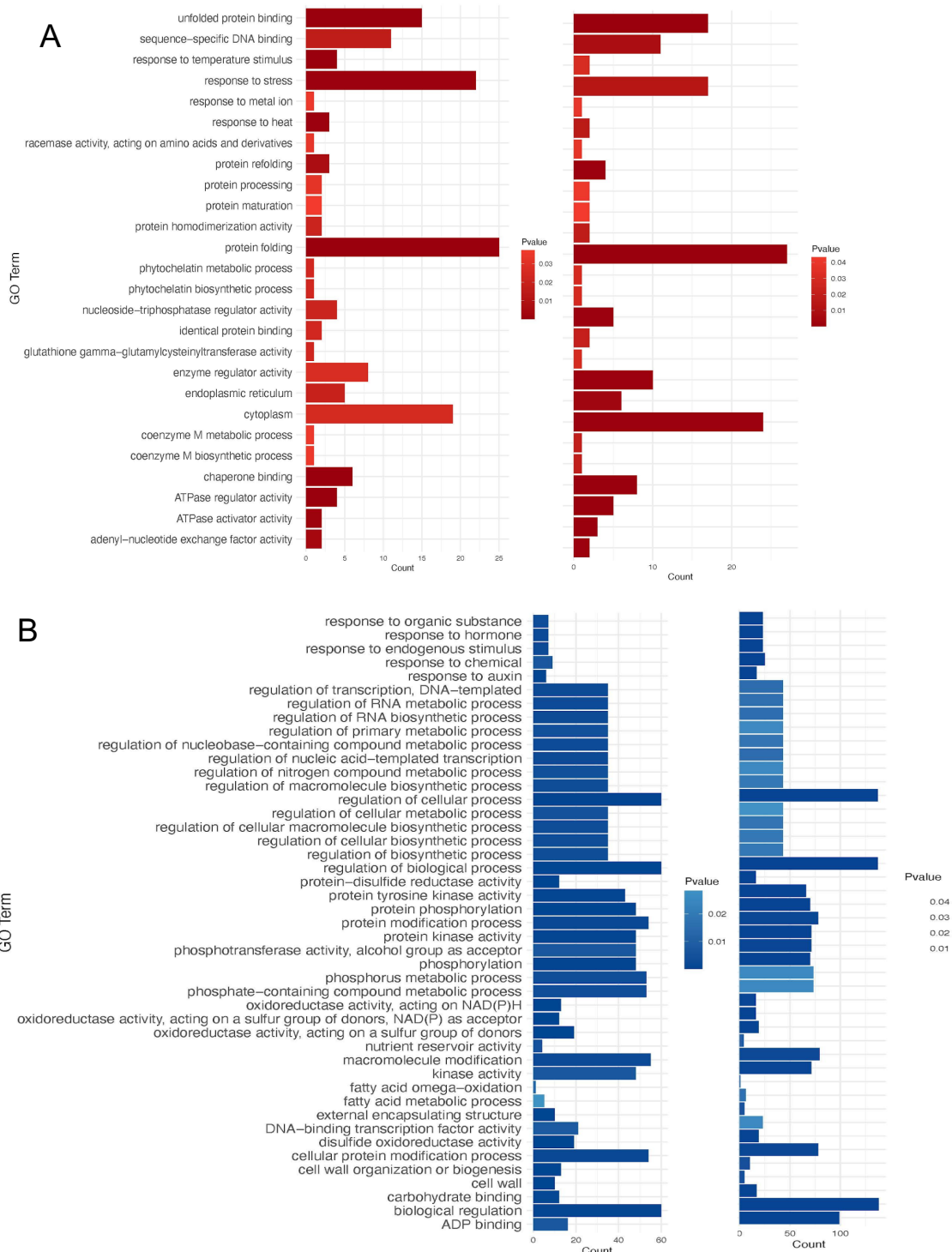


Figure 3.4. Shared GO terms enriched in upregulated and down regulated genes under heat stress conditions in *B. arcuata* and *B. perennans* (A) Shared GO terms of upregulated genes (B) down regulated genes.

To investigate the unique gene expression patterns in *B. arcuata* and *B. perennans*, I focused on upregulated genes that were also expanded within each species. I then created separate tables for each species, listing the unique upregulated genes along with their *A. thaliana* orthologs and functional descriptions. The results demonstrated that both *B. arcuata* and *B. perennans* possess unique sets of upregulated genes associated with various biological functions. In *B. arcuata*, unique upregulated genes included several involved in disease resistance, protein kinase activity, and responses to stress. On the other hand, *B. perennans* exhibited upregulated genes involved in protein metabolism, developmental regulation, and stress responses (Tables 3.6 and 3.7). These findings suggested species-specific adaptations and functional specializations that may contribute to their differentiation and evolutionary trajectories.

Table 3.6. The list of genes that are upregulated and expanded in *B. arcuata*. The orthologs and functional descriptions from *A. thaliana* provide insights into the potential roles of these genes.

	Gene-ID	<i>A. thaliana</i> Ortholog	Description
1	Boarc.1G341600.1	AT4G16990	TIR-NBS class
2	Boarc.7G033300.1	AT5G55440	F-box protein, putative
3	Boarc.3G051900.1	AT3G57660	Encodes a subunit of RNA polymerase I
4	Boarc.5G158300.1	AT2G02590	molecular_function unknown
5	Boarc.1G298900.1	AT1G61860	Protein kinase superfamily protein
6	Boarc.7G157300.1	AT4G18253	Serine/Threonine kinase-like
7	Boarc.7G008100.1	AT5G41800	Transmembrane amino acid transporter family protein
8	Boarc.2G263500.1	AT1G70510	Class I knotted1-like homeobox
9	Boarc.4G102100.1	AT2G31320	Encodes a poly(ADP-ribose) polymerase. response to abscisic acid, response to oxidative stress
10	Boarc.1G232800.1	AT4G14590	embryo defective 2739
11	Boarc.1G029300.1	AT1G57630	(TIR) domain family protein

Table 3.7. The list of genes that are upregulated and expanded in *B. perenans*. The orthologs and functional descriptions from *A. thaliana* provide insights into the potential roles of these genes.

	Gene-ID	<i>A. thaliana</i> Ortholog	Description
1	Boper.3G293800.1	AT2G07689.1	NADH-Ubiquinone/plastoquinone (complex I) protein
2	Boper.2G194800.1	AT1G70250.1	Protease inhibitor/seed storage/LTP family protein
3	Boper.5G008800.1	AT3G45670.1	Protein kinase superfamily protein
4	Boper.1G238300.1	AT5G66450.3	Plastid Phosphatidic acid phosphatase
5	Boper.1G320000.1	AT1G63550.1	Receptor-like protein kinase-related family protein
6	Boper.2G022600.1	AT1G68300.1	Adenine nucleotide alpha hydrolases-like superfamily protein
7	Boper.4G161500.1	AT2G34780.1	Novel protein of unknown function embryo defective
8	Boper.2G309100.1	AT1G48720.1	Copia-like polyprotein/retrotransposon
9	Boper.7G320300.1	AT4G37250.1	Leucine-rich repeat protein kinase family protein
10	Boper.1G262600.1	AT5G47260.1	NLR (nucleotide-binding and leucine-rich repeat)
11	Boper.3G347500.1	AT5G54160.1	Caffeic acid/5-hydroxyferulic acid O-methyltransferase

Positive selection acts on expanded, upregulated genes

I employed various methods to determine whether the expanded upregulated genes have undergone positive selection, revealing different evolutionary trajectories. Initially, I selected several genes expanded and upregulated in either *B. arcuata* or *B. perennans* from tables 3.6 and 3.7. These genes' orthologs in *A. thaliana* include AT5G55440.1, AT4G18253.1, AT2G31320.1, and AT1G57630.1 which were expanded upregulated genes in *B. arcuata*, and AT2G07689.1, AT1G63550.1, and AT5G54160.1 in *B. perennans*. These genes were then subjected to tests for positive selection to understand their evolutionary patterns (Table 3.8). For the gene AT1G57630 (TIR/Hypoxia), which includes 25 sequences and 964 sites, significant evidence of positive selection was found. aBSREL identified 7 sites under positive selection ($p=0.0001$), and BUSTED detected 21 sites (19 sites, $p=0.01$). Conversely, the AT1G63550 (Cysteine-rich repeat secreted) gene family, consisting of 10 sequences and 376 sites, showed no evidence of positive selection (ne). This indicates that while the gene family expanded, the sequences have not been under selective pressure to alter amino acids, possibly retaining their original functions. In fact, although the gene family has experienced mutations, there has not been a consistent selective pressure to favor amino acid changes. Instead, the mutations that do occur are either neutral or deleterious mutations that are selectively neutral or removed, leading to the retention of the gene family's original functions. Similarly, the AT2G07689 (NADH-Ubiquinone/plastoquinone complex I) gene family (6 sequences, 457 sites) showed no positive selection (ne), suggesting expansion without functional diversification under selective pressures. The AT2G31320 (oxidative stress nucleolus DNA repair) gene family

(8 sequences, 1585 sites) displayed significant positive selection. aBSREL detected 2 sites ($p=6.07e-8$), and BUSTED identified 20 sites (11 sites, $p=0.05$). The AT4G18253 (Kinase) gene family, comprising 7 sequences and 337 sites, exhibited positive selection with 3 sites detected by aBSREL ($p=0.0003$) and 8 sites by BUSTED (7 sites, $p=0.05$). The AT5G54160 (Flavonol biosynthesis) gene family (18 sequences, 375 sites) showed positive selection with 5 sites identified by aBSREL ($p=0.01747$) and 17 sites by BUSTED (10 sites, $p=0.05$). The AT5g55440 (F-Box- DUF295) gene family, with 22 sequences and 487 sites, displayed significant positive selection. aBSREL detected 7 sites ($p=2.457e-10$), and BUSTED identified 16 sites (7 sites, $p=0.05$). My findings reveal two distinct evolutionary patterns among the upregulated and expanded genes in *B. arcuata* and *B. perennans*. Some genes exhibit retention of duplicate copies without significant amino acid changes, while others show an increase in copy number along with adaptive changes in amino acid sequences, potentially leading to neofunctionalization.

Table 3.8. Selection analysis of upregulated and expanded genes in *B. arcuata* and *B. perennans*. The table presents the results of selection tests (aBSREL, BUSTED, and MEME) conducted on the upregulated and expanded genes in *B. arcuata* and *B. perennans*. The analysis identifies the number of sequences, the number of sites, and the evidence of positive selection detected by each method.

Gene family	# of sequences	# of sites	Methods		
			aBSREL	BUSTED	MEME
AT1G57630 (TIR) / Hypoxia	25	964	7 ^a	p=0.0001 21 sites	19 p=0.01
AT1G63550 Cysteine-rich repeat secreted	10	376	ne	ne	ne
AT2G07689 NADH-Ubiquinone/plastoquinone (complex I)	6	457	ne	ne	ne
AT2G31320 oxidative stress nucleolus DNA repair	8	1585	2 ^b	p=6.07e-8 20 sites	11 p=0.05
AT4G18253 Kinase	7	337	3 ^c	p=0.0003 8 sites	7 p=0.05
AT5G54160 Flavonol biosynthesis	18	375	5 ^d	p=0.01747 17 sites	10 p=0.05
AT5g55440 F-Box-DUF295	22	487	7 ^e	p=2.457e-10 16 sites	7 p=0.05

^a BOARC_5G032400; BOARC_1G029300; BOARC_5G212800; BOARC_7G032400; AH2G09010; BOARC_1G248700; BOARC_7G063300

^b BOARC_4G102100; BOARC_4G103000; BOPER_7G255500_1

^c BOARC_7G157300; BOARC_7G210100_1; BOPER_7G207100_1

^d BOARC_7G258200_1; BOARC_6G139600_1; BOPER_7G255500_1; BOARC_6G139600_1; BOPER_7G256900_1

^e BOARC_7G033500; BOARC_7G039400; BOARC_7G033300; BOARC_7G096600; BOARC_7G159300; BOPER_7G156700; BOARC_7G141200

Figures 3.5 through 3.7 show the alignment results for AT1G57630, AT5G54160, and AT5G55440, respectively. The figures display the amino acid sites evolving under positive selection ($p < 0.05$) as indicated by both MEME and BUSTED. This identification of conserved sites under positive selection across different methodologies provides robust evidence of their evolutionary significance. Figures 3.8.A through 3.10.A illustrate the fitted aBSREL model trees for AT1G57630, AT5G54160, and AT5G55440, respectively. In these trees, branches showing evidence of positive selection are indicated in green, with the intensity of the green color representing the extent of positive selection (dN/dS). These analyses reveal that many of the expanded and upregulated genes in both *Boechera* species have undergone adaptive evolution, which is likely contributing to their functional roles under basal heat stress conditions. I also constructed the maximum likelihood phylogenies for each gene family to identify their evolutionary history (3.8.B-3.10.B). The phylogenetic trees and alignments together provide comprehensive insights into the evolutionary trajectories and selective pressures shaping these gene families.

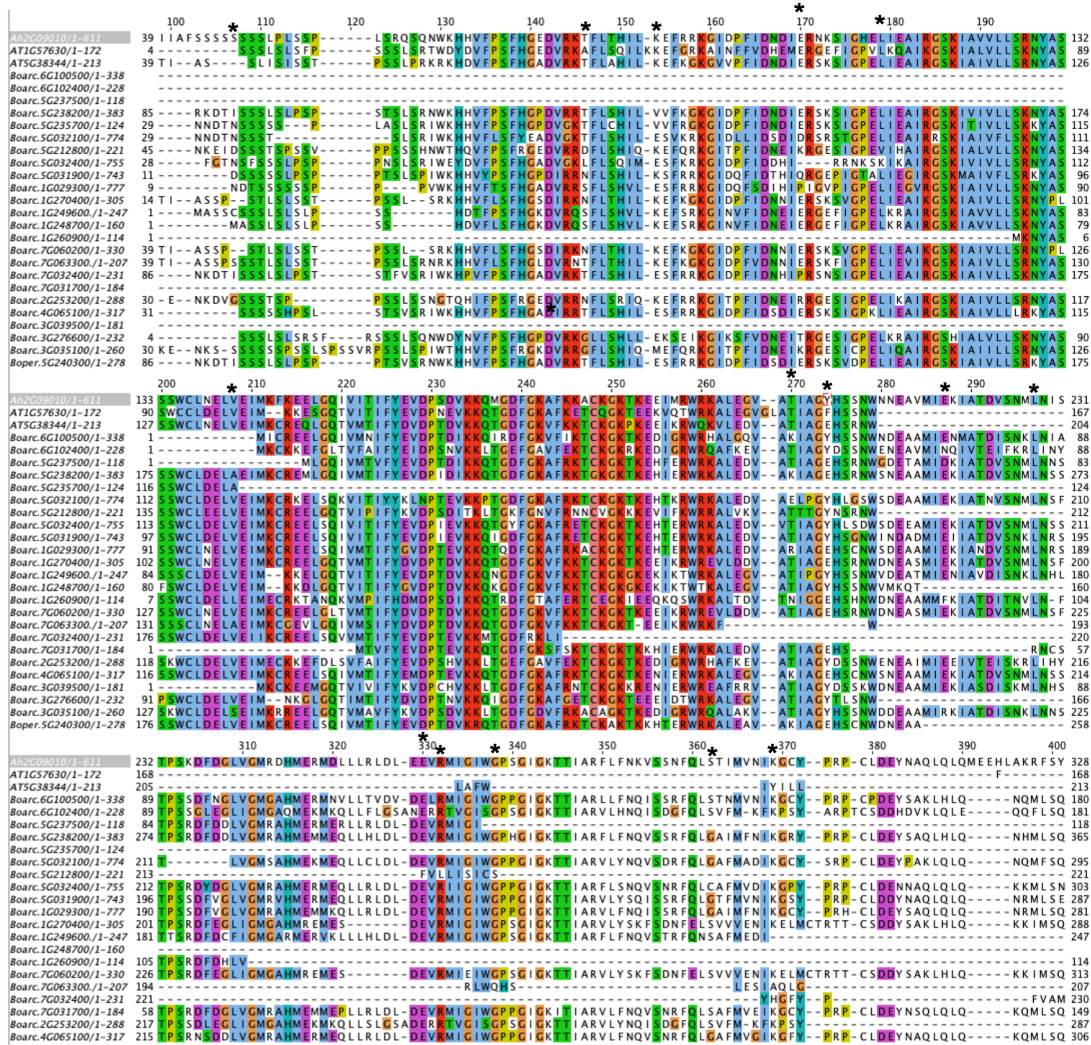


Figure 3.5. AT1g57630 alignment. Alignment of peptide sequences showing amino acid sites identified by both MEME and BUSTED as being under positive selection. Sites under positive selection identified by both tools are marked with asterisks.

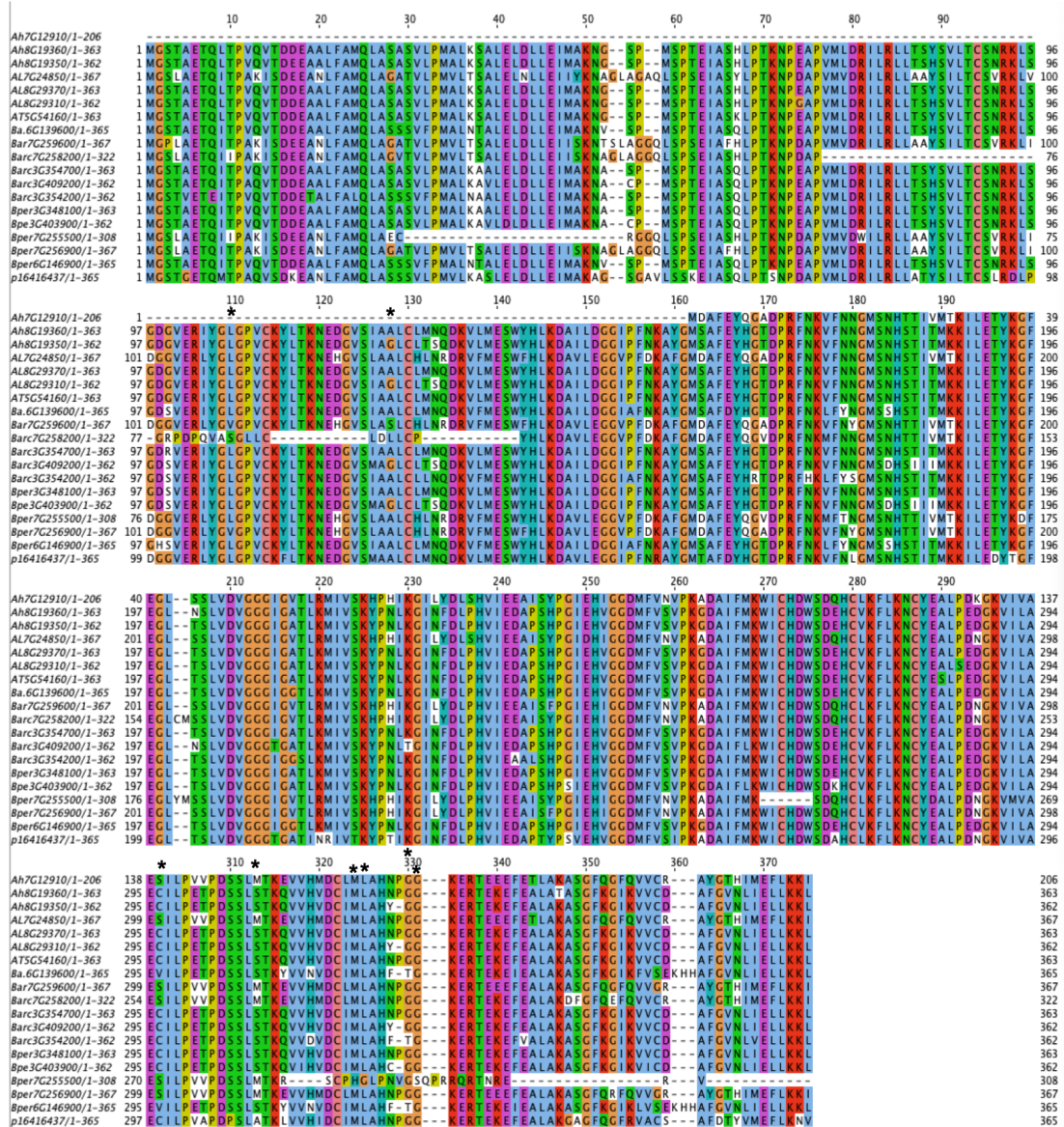


Figure 3.6. AT5G54160 alignment. Alignment of peptide sequences showing amino acid sites identified by both MEME and BUSTED as being under positive selection. Sites under positive selection identified by both tools are marked with asterisks.



Figure 3.7. AT5G55440 alignment. Alignment of peptide sequences showing amino acid sites identified by both MEME and BUSTED as being under positive selection. Sites under positive selection identified by both tools are marked with asterisks.

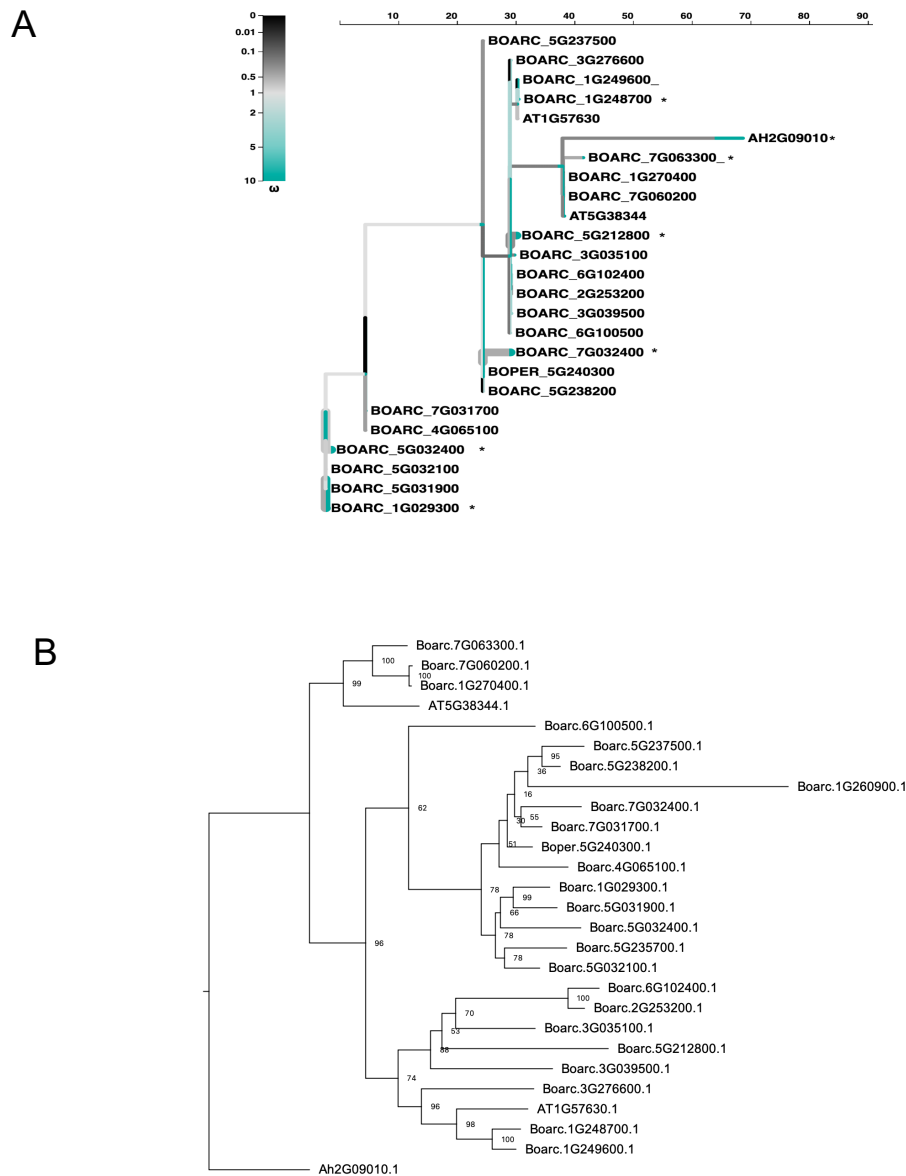


Figure 3.8. AT1g57630 phylogenetic tree. (A) Branches that show evidence of positive selection are indicated in green, with the intensity of the green color representing the extent of positive selection (dN/dS). Darker green branches indicate a higher probability of having a dN/dS ratio greater than one. The thickness of the branches represents the strength of evidence for positive selection, with thicker branches indicating stronger support. The branch lengths represent the expected number of substitutions per site, reflecting the evolutionary distance between sequences. Asterisks (*) at the branches denote statistical significance for dN/dS being over one, indicating strong positive selection. The scale provided is generated by aBSREL. Branches marked by * are those with the highest support for positive selection (dN/dS << 0). (B) Maximum likelihood phylogeny inferred by IQTREE.

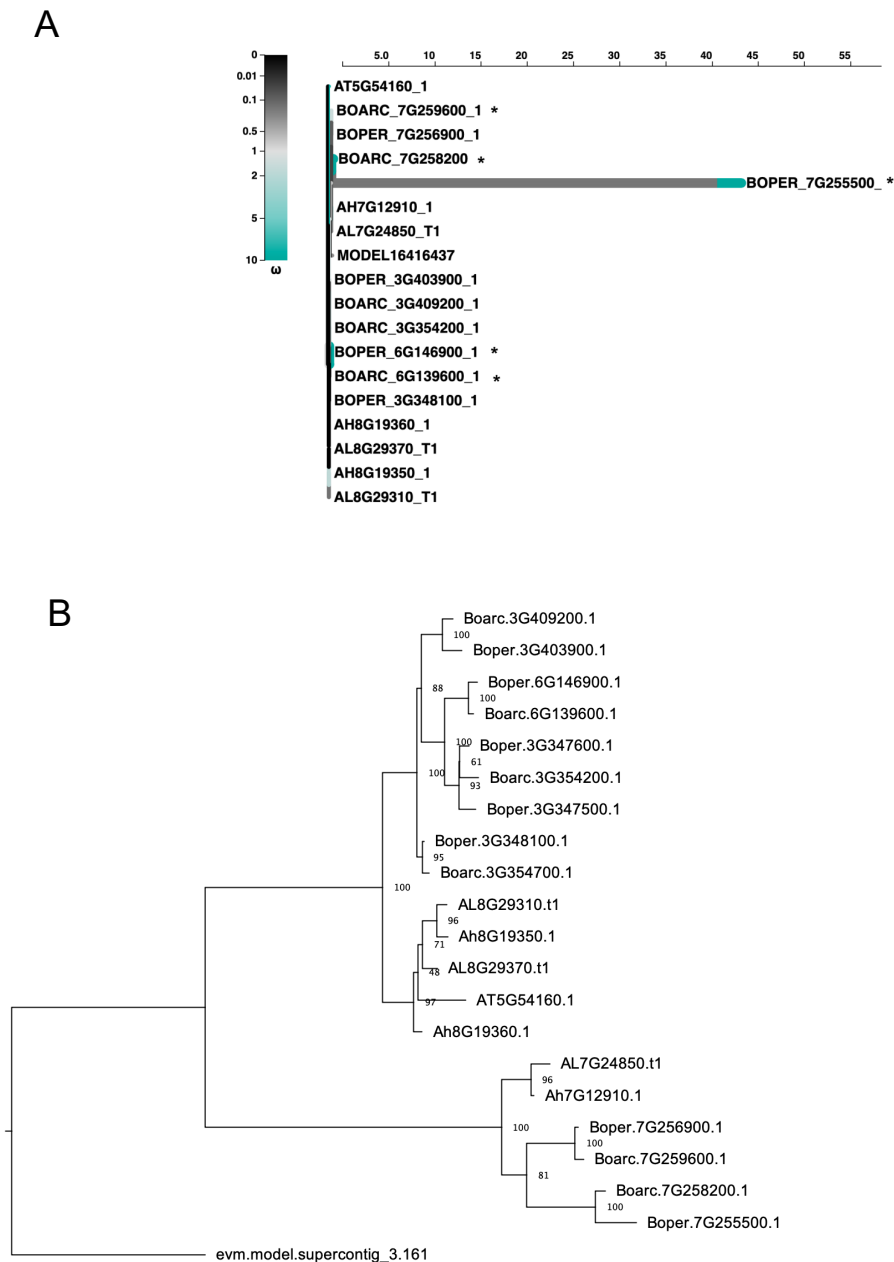


Figure 3.9. AT5G54160 phylogenetic tree. (A) Branches that show evidence of positive selection are indicated in green, with the intensity of the green color representing the extent of positive selection (dN/dS). Darker green branches indicate a higher probability of having a dN/dS ratio greater than one. Asterisks (*) at the branches denote statistical significance for dN/dS being over one, indicating strong positive selection. The scale provided is generated by aBSREL. Branches marked by * are those with the highest support for positive selection (dN/dS << 0). (B) Maximum likelihood phylogeny inferred by IQTREE.

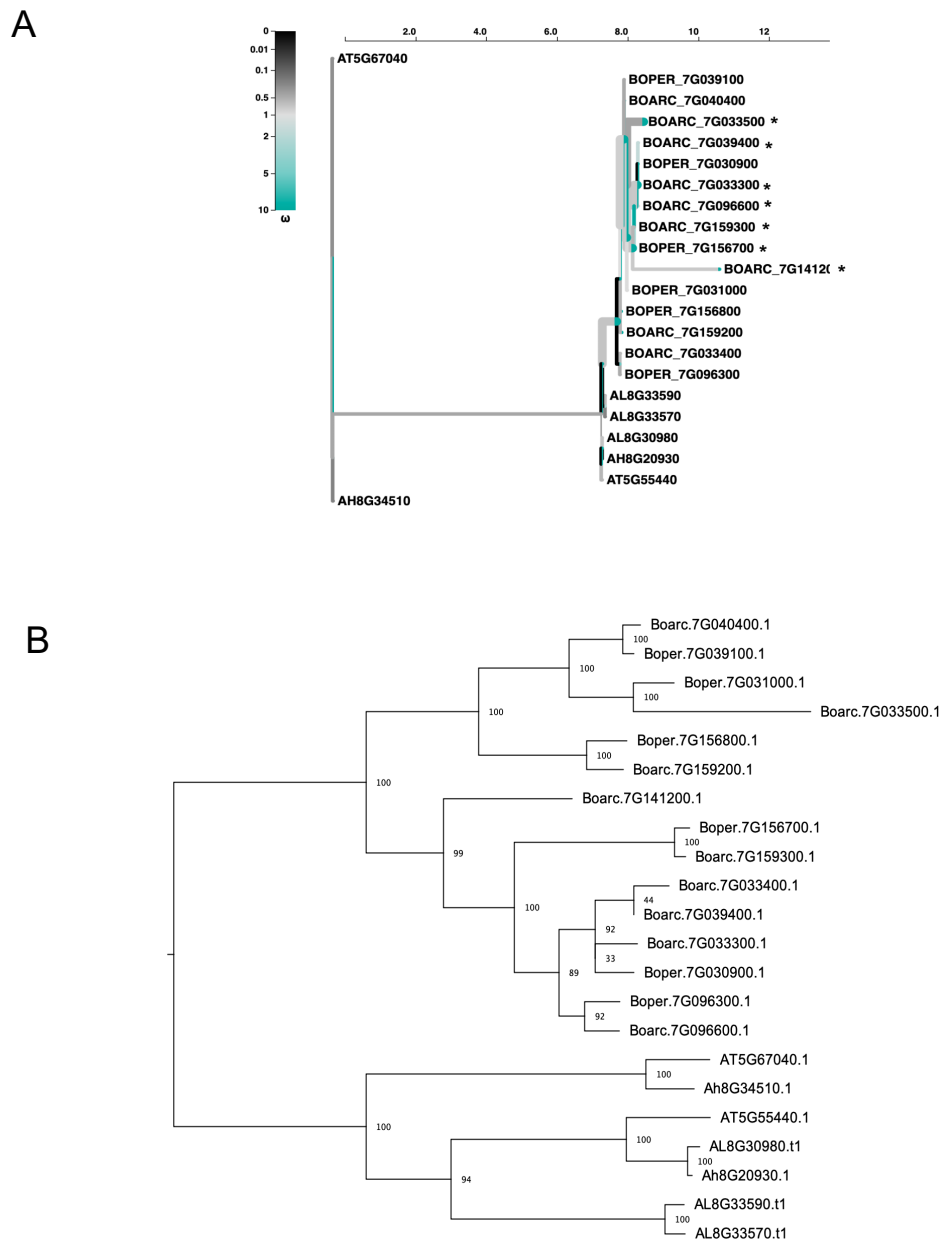


Figure 3.10. AT5G55440 phylogenetic tree. (A) Branches that show evidence of positive selection are indicated in green, with the intensity of the green color representing the extent of positive selection (dN/dS). Darker green branches indicate a higher probability of having a dN/dS ratio greater than one. Asterisks (*) at the branches denote statistical significance for dN/dS being over one, indicating strong positive selection. The scale provided is generated by aBSREL. Branches marked by * are those with the highest support for positive selection (dN/dS << 0). (B) Maximum likelihood phylogeny inferred by IQTREE.

Discussion

The heat shock or stress response is found in all organisms. During heat-stress the protective heat shock proteins are expressed and these proteins via their chaperone function maintain protein function, and cellular homeostasis and by doing so provide organismal thermotolerance. Over the last 30 to 40 years our understanding of the heat shock response and HSP protein function has greatly expanded. At the same time researchers have also seen evidence that some aspects of the heat-stress response can and do vary among organisms. *Boechera arcuata* and *B. perennans* were chosen for this study because of both their close phylogenetic relationships, the close proximity of their habitats, the differences in climate within their habitats, and their demonstrated differences in organismal thermotolerance. Here I demonstrate that these two species both display the conserved aspects of the HSP response mentioned above and also have species-specific patterns of heat-stress response.

Both species upregulated the heat shock transcription factors as well as HSP20s, HSP70s, and AAA+ homologs. The heat shock response of both *Boechera* species was dominated by the HSP20. This is not surprising as HSP20 has long been shown to be a crucial component of plant thermotolerance. Many other components of the HS response are upregulated in both species including modulators of HSP70, calcium dependent signaling molecules and the ubiquitin degradation pathways. These findings are consistent with previous studies (Larkindale et al. 2005; Mittler, Finka, and Goloubinoff 2012; Guihur, Rebeaud, and Goloubinoff 2022; Schramm et al. 2008; Larkindale and Vierling 2008; Lee et al. 2007; W. Wang et al. 2004).

Not all genes that are upregulated in one species are upregulated in the other. This could be due to stochastic differences in gene expression that do not reflect differences in organismal thermotolerance (Bohrer and Larson 2021). However, the previous findings that *B. arcuata* and *B. perennans* do have significantly different HS response and patterns of thermotolerance (Gallas and Waters 2015; Halter et al. 2017) suggests that not all of the gene expression differences between these two species are stochastic and some may underlie the organismal differences in thermotolerance. It is interesting that the differences between these two species include differential upregulation of distinct sets of protein kinases. An important question for future study would include determining if these differentially expressed protein kinases are involved in distinct or overlapping signalling networks.

Another important difference between these two species are the distinct gene families that have undergone lineage-specific expansions within each species, and the unique evolutionary history each gene family displays. The gene family that includes AT1g57630 as the orthologous gene in *A. thaliana* is differentially regulated and has a lineage-specific expansion in *B. arcuata*. Based on Gene Ontology annotation the *Arabidopsis* homolog localizes to the chloroplast is involved in signal transduction and is related to cellular responses to abiotic stress (hypoxia) in *A. thaliana*. Sequence comparisons indicate that this gene has a TIR-like domain. Here I found that this gene is differentially regulated in response to heat stress. There have been other reports of “immunity” genes in plants also having important roles in abiotic stress responses (MacQueen and Bergelson 2016; Ren et al. 2020). *Boechera perennans* has one homolog

in this gene family and *B. arcuata* has 23. The *B. arcuata* gene duplicates are found on all seven chromosomes. The dispersed chromosomal positions suggest that these are not the result of tandem duplications. The positions in the alignment that were found to have undergone positive selection are also dispersed along the primary sequence of the proteins but are more common in the variable N and C terminal regions than in the core TIR domain. Future studies on the gene family should prove highly informative. Is this lineage-specific expansion evolved specifically to respond to abiotic stressor such as heat, or has it evolved to respond to biotic stressors that are specific to the Southern California region? Much more detailed studies of the expression patterns of these genes over a range of abiotic and biotic stressors would be needed to distinguish between these possibilities. Immune genes are known to be subjects of significant levels of natural selection both for gene duplication and amino acid changes as the pathogens and hosts continually evolve (Clay and Kover 1996). This is a large complex gene family; some genes are specific for biotic stress and others seem to be specialized for abiotic stress. This sets up a hypothesis that needs to be tested through experimental analysis.

Another differentially-regulated and expanded gene includes the homolog of *A. thaliana*, AT5G54160. This gene is an important part of the flavonol biosynthetic pathway and has been associated with osmotic stress resistance (Z. Wang et al. 2022). The other *Arabidopsis* species have three genes within this super family, however *B. arcuata* has six and *B. perennans* has five. While this gene expansion is not as dramatic as that seen for the TIR-like proteins discussed above these expanded genes also show evidence of positive selection on amino acid residues. It is also interesting that based on chromosomal location

and phylogenetic analysis there is strong evidence that this gene expansion occurred prior to the divergence of the common ancestor of these two *Boechea* species. The flavonols, like the TIR proteins, are known to have roles both in abiotic and biotic stress responses. It is notable that in this study, these genes were identified in a screen for expanded and differentially regulated gene families. This suggests that this expansion is related to the role of flavonols in responses to heat stress and the associated rise in reactive oxygen species that accompanies heat stress. The role of these genes in heat-tolerance will need to be examined in further experiments.

Perhaps the most intriguing gene family that was found to be both differentially regulated and expanded in this study are the DUF295-putative F-box proteins (AT5G55440). As the DUF designation (Domain of Unknown Function) signifies there is very little known about the function of these proteins during stress or during normal plant growth and development. This gene family also contains a putative F-box domain; however, its GO ontologies provide no information on this gene family's biological role. These proteins are predicted (based on sequence analysis) to be mitochondrial-localized. For my analysis a study on the larger gene family that contains DUF295 gene family was conducted recently (Lama et al. 2019). This study examined 94 DUF295-related genes in the *Arabidopsis thaliana* genome and in a range of Brassicaceae genomes. These authors conclude that the DUF295 genes are products of recent gene expansions within the Brassicaceae. In their analysis of the complete genomes of a wide range of plant species they found that this specific gene family (AT5G55440) was present in only one copy in the available Brassicaceae genomes. This analysis, along with that presented here, indicates

that the lineage specific expansion found in both *B. arcuata* and *B. perennans* of the (AT5G55440) gene family is at least *Boechea*-specific. The pattern of gene duplications for this gene family is distinct from that seen in the two gene expansions above. Here all the *Boechea* genes are found on chromosome seven. The phylogenetic relationships seen in Figures 3.9-3.11 show four orthologous pairs with other the expanded *B. arcuata* and *B. perennans* genes indicating either more recent gene duplications or gene conversion within genomes. Of the four orthologous pairs, only one pair has evidence of positive selection on amino acid residues. Intriguingly, all but one of the other expanded genes (i.e. those not showing orthologous relationships) have evidence of selection. Thus, the orthologous pairs are much less likely to have undergone positive selection on amino acids. It is important to note that the amino acid residues evolving under positive selection are found within the DUF295 C-terminal domain and not in the F-box region. While there is little known of the DUF295 proteins, much more is known about the F-Box proteins (Kipreos and Pagano 2000). There are as many as 1,000 different F-box proteins in plant genomes. What these proteins share is the 40-50 amino acid F-box motif. However, these proteins are functionally very diverse. The presence of the F-box does not help define function, but we do know that these proteins can play a role in abiotic stress responses in plants (X. Zhang et al. 2019).

Conclusion

This study elucidates the gene expression profiles and evolutionary patterns of heat stress responses in two *Boechera* species, *B. arcuata* and *B. perennans*. The findings demonstrate that both species exhibit distinct and overlapping patterns of heat-induced gene expression changes. These differences may represent distinct adaptations to heat stress. The predominance of the heat shock proteins among the top upregulated genes underscores their essential role in stress tolerance. The evolutionary analysis revealed that many of the upregulated and expanded genes have undergone positive selection, suggesting adaptive changes that may contribute to their environmental adaptation. This research provides a comprehensive understanding of the molecular basis of heat stress responses in these species and offers valuable insights into the broader context of plant adaptation to environmental stresses. Future studies should focus on further characterizing these genes and exploring their functional roles in various stress conditions to develop strategies for enhancing crop resilience.

References

- Amano, Momoe, Satoko Iida, and Keiko Kosuge. 2012. "Comparative Studies of Thermotolerance: Different Modes of Heat Acclimation between Tolerant and Intolerant Aquatic Plants of the Genus *Potamogeton*." *Annals of Botany* 109 (2): 443–52.
- Atkinson, Nicky J., and Peter E. Urwin. 2012. "The Interaction of Plant Biotic and Abiotic Stresses: From Genes to the Field." *Journal of Experimental Botany* 63 (10): 3523–43.
- Azad, Mahnaz, Masoud Tohidfar, Rahele Ghanbari Moheb Seraj, Mohammad Mehralian, and Keyvan Esmaeilzadeh-Salestani. 2024. "Identification of Responsive Genes to Multiple Abiotic Stresses in Rice (*Oryza Sativa*): A Meta-Analysis of Transcriptomics Data." *Scientific Reports* 14 (1): 5463.
- Baxter, Aaron, Ron Mittler, and Nobuhiro Suzuki. 2014. "ROS as Key Players in Plant Stress Signalling." *Journal of Experimental Botany* 65 (5): 1229–40.
- Becker, Jörg, and Elizabeth A. Craig. 1994. "Heat-shock Proteins as Molecular Chaperones." *European Journal of Biochemistry / FEBS* 219 (1-2): 11–23.
- Bohrer, Christopher H., and Daniel R. Larson. 2021. "The Stochastic Genome and Its Role in Gene Expression." *Cold Spring Harbor Perspectives in Biology* 13 (10). <https://doi.org/10.1101/cshperspect.a040386>.
- Boston, R. S., P. V. Viitanen, and E. Vierling. 1996. "Molecular Chaperones and Protein Folding in Plants." *Plant Molecular Biology* 32 (1-2): 191–222.
- Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. "Erratum: Near-Optimal Probabilistic RNA-Seq Quantification." *Nature Biotechnology* 34 (8): 888.
- Bukau, B., and A. L. Horwich. 1998. "The Hsp70 and Hsp60 Chaperone Machines." *Cell* 92 (3): 351–66.
- Cheong, Yong Hwa, Hur-Song Chang, Rajeev Gupta, Xun Wang, Tong Zhu, and Sheng Luan. 2002. "Transcriptional Profiling Reveals Novel Interactions between Wounding, Pathogen, Abiotic Stress, and Hormonal Responses in *Arabidopsis*." *Plant Physiology* 129 (2): 661–77.
- Clay, K., and P. X. Kover. 1996. "The Red Queen Hypothesis and Plant/pathogen Interactions." *Annual Review of Phytopathology* 34: 29–50.

- Eyles, Stephen J., and Lila M. Gierasch. 2010. "Nature's Molecular Sponges: Small Heat Shock Proteins Grow into Their Chaperone Roles." *Proceedings of the National Academy of Sciences of the United States of America*.
- French, Kris, Sharon A. Robinson, and Jodie Lia. 2017. "Thermotolerance Capacities of Native and Exotic Coastal Plants Will Lead to Changes in Species Composition under Increased Heat Waves." *Conservation Physiology* 5 (1): cox029.
- Gallas, Genna, and Elizabeth R. Waters. 2015. "Boechera Species Exhibit Species-Specific Responses to Combined Heat and High Light Stress." *PloS One* 10 (6): e0129041.
- Guihur, Anthony, Mathieu E. Rebeaud, and Pierre Goloubinoff. 2022. "How Do Plants Feel the Heat and Survive?" *Trends in Biochemical Sciences* 47 (10): 824–38.
- Gurley, W. B. 2000. "HSP101: A Key Component for the Acquisition of Thermotolerance in Plants." *The Plant Cell* 12 (4): 457–60.
- Halter, Gillian, Nicole Simonetti, Cristy Suguitan, Kenneth Helm, Jessica Soroksky, and Elizabeth R. Waters. 2017. "Patterns of Thermotolerance, Chlorophyll Fluorescence, and Heat Shock Gene Expression Vary among four *Boechera* species and *Arabidopsis Thaliana*." *Botany* 95 (1): 9–27.
- Haslbeck, M. 2002. "sHsps and Their Role in the Chaperone Network." *Cellular and Molecular Life Sciences: CMLS* 59 (10): 1649–57.
- Hay, Nikolai M., Michael D. Windham, Terezie Mandáková, Martin A. Lysak, Kasper P. Hendriks, Klaus Mummenhoff, Frederic Lens, Kathleen M. Pryer, and C. Donovan Bailey. 2023. "A Hyb-Seq Phylogeny of *Boechera* and Related Genera Using a Combination of Angiosperms353 and Brassicaceae-Specific Bait Sets." *American Journal of Botany* 110 (10): e16226.
- Hazen, Samuel P., Yajun Wu, and Joel A. Kreps. 2003. "Gene Expression Profiling of Plant Responses to Abiotic Stress." *Functional & Integrative Genomics* 3 (3): 105–11.
- Hilton, Gillian R., Hadi Lioe, Florian Stengel, Andrew J. Baldwin, and Justin L. P. Benesch. 2013. "Small Heat-Shock Proteins: Paramedics of the Cell." *Topics in Current Chemistry* 328: 69–98.
- Hou, Quancan, Guido Ufer, and Dorothea Bartels. 2016. "Lipid Signalling in Plant Responses to Abiotic Stress." *Plant, Cell & Environment* 39 (5): 1029–48.
- Jenkins, M. E., T. C. Suzuki, and D. W. Mount. 1997. "Evidence That Heat and Ultraviolet Radiation Activate a Common Stress-Response Program in Plants That Is Altered in the *uvh6* Mutant of *Arabidopsis Thaliana*." *Plant Physiology* 115 (4): 1351–58.

- Kannan, Sumetha, Gillian Halter, Tanya Renner, and Elizabeth R. Waters. 2018. "Patterns of Alternative Splicing Vary between Species during Heat Stress." *AoB Plants* 10 (2): ly013.
- Katoh, Kazutaka, John Rozewicki, and Kazunori D. Yamada. 2019. "MAFFT Online Service: Multiple Sequence Alignment, Interactive Sequence Choice and Visualization." *Briefings in Bioinformatics* 20 (4): 1160–66.
- Kim, Yujin E., Mark S. Hipp, Andreas Bracher, Manajit Hayer-Hartl, and F. Ulrich Hartl. 2013. "Molecular Chaperone Functions in Protein Folding and Proteostasis." *Annual Review of Biochemistry* 82: 323–55.
- Kosakovsky Pond, Sergei L., Art F. Y. Poon, Ryan Velazquez, Steven Weaver, N. Lance Hepler, Ben Murrell, Stephen D. Shank, et al. 2019. "HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies." *Molecular Biology and Evolution* 37 (1): 295–99.
- Kotak, Sachin, Elizabeth Vierling, Helmut Bäumlein, and Pascal von Koskull-Döring. 2007. "A Novel Transcriptional Cascade Regulating Expression of Heat Stress Proteins during Seed Development of Arabidopsis." *The Plant Cell* 19 (1): 182–95.
- Lama, Sbatie, Martyna Broda, Zahra Abbas, Dries Vaneechoutte, Katharina Belt, Torbjörn Säll, Klaas Vandepoele, and Olivier Van Aken. 2019. "Neofunctionalization of Mitochondrial Proteins and Incorporation into Signaling Networks in Plants." *Molecular Biology and Evolution* 36 (5): 974–89.
- Larkindale, Jane, Jennifer D. Hall, Marc R. Knight, and Elizabeth Vierling. 2005. "Heat Stress Phenotypes of Arabidopsis Mutants Implicate Multiple Signaling Pathways in the Acquisition of Thermotolerance." *Plant Physiology* 138 (2): 882–97.
- Larkindale, Jane, and Elizabeth Vierling. 2008. "Core Genome Responses Involved in Acclimation to High Temperature." *Plant Physiology* 146 (2): 748–61.
- Lee, Ung, Ignatius Rioflorido, Suk-Whan Hong, Jane Larkindale, Elizabeth R. Waters, and Elizabeth Vierling. 2007. "The Arabidopsis ClpB/Hsp100 Family of Proteins: Chaperones for Stress and Chloroplast Development." *The Plant Journal: For Cell and Molecular Biology* 49 (1): 115–27.
- Liberek, Krzysztof, Agnieszka Lewandowska, and Szymon Zietkiewicz. 2008. "Chaperones in Control of Protein Disaggregation." *The EMBO Journal* 27 (2): 328–35.
- Lindquist, S., and E. A. Craig. 1988. "The Heat-Shock Proteins." *Annual Review of Genetics* 22: 631–77.

- Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550.
- MacQueen, Alice, and Joy Bergelson. 2016. "Modulation of R-Gene Expression across Environments." *Journal of Experimental Botany* 67 (7): 2093–2105.
- Mishra, Neelam, Chenkai Jiang, Lin Chen, Abhirup Paul, Archita Chatterjee, and Guoxin Shen. 2023. "Achieving Abiotic Stress Tolerance in Plants through Antioxidative Defense Mechanisms." *Frontiers in Plant Science* 14 (June): 1110622.
- Mittler, Ron. 2006. "Abiotic Stress, the Field Environment and Stress Combination." *Trends in Plant Science* 11 (1): 15–19.
- Mittler, Ron, Andrija Finka, and Pierre Goloubinoff. 2012. "How Do Plants Feel the Heat?" *Trends in Biochemical Sciences* 37 (3): 118–25.
- Murrell, Ben, Steven Weaver, Martin D. Smith, Joel O. Wertheim, Sasha Murrell, Anthony Aylward, Kemal Eren, et al. 2015. "Gene-Wide Identification of Episodic Selection." *Molecular Biology and Evolution* 32 (5): 1365–71.
- Murrell, Ben, Joel O. Wertheim, Sasha Moola, Thomas Weighill, Konrad Scheffler, and Sergei L. Kosakovsky Pond. 2012. "Detecting Individual Sites Subject to Episodic Diversifying Selection." *PLoS Genetics* 8 (7): e1002764.
- Nakashima, Kazuo, Hironori Takasaki, Junya Mizoi, Kazuo Shinozaki, and Kazuko Yamaguchi-Shinozaki. 2012. "NAC Transcription Factors in Plant Abiotic Stress Responses." *Biochimica et Biophysica Acta* 1819 (2): 97–103.
- Nover, L., K. D. Scharf, D. Gagliardi, P. Vergne, E. Czarnecka-Verner, and W. B. Gurley. 1996. "The Hsf World: Classification and Properties of Plant Heat Stress Transcription Factors." *Cell Stress & Chaperones* 1 (4): 215–23.
- Pandey, Saurabh, Dharendra Fartyal, Aakrati Agarwal, Tushita Shukla, Donald James, Tanushri Kaul, Yogesh K. Negi, Sandeep Arora, and Malireddy K. Reddy. 2017. "Abiotic Stress Tolerance in Plants: Myriad Roles of Ascorbate Peroxidase." *Frontiers in Plant Science* 8 (April): 581.
- Pierce, Ian, Gillian Halter, and Elizabeth R. Waters. 2020. "Transcriptomic Analysis of the Heat-Stress Response in *Boechera Depauperata* and *Arabidopsis* Reveals a Distinct and Unusual Heat-Stress Response in *Boechera*." *Botany* 98 (10): 589–602.
- Pintó-Marijuan, Marta, and Sergi Munné-Bosch. 2014. "Photo-Oxidative Stress Markers as a Measure of Abiotic Stress-Induced Leaf Senescence: Advantages and Limitations." *Journal of Experimental Botany* 65 (14): 3845–57.

- Pirkkala, L., P. Nykänen, and L. Sistonen. 2001. "Roles of the Heat Shock Transcription Factors in Regulation of the Heat Shock Response and beyond." *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology* 15 (7): 1118–31.
- Ren, Shuping, Min Sun, Haidong Yan, Bingchao Wu, Tingting Jing, Linkai Huang, and Bing Zeng. 2020. "Identification and Distribution of NBS-Encoding Resistance Genes of *Dactylis Glomerata* L. and Its Expression Under Abiotic and Biotic Stress." *Biochemical Genetics* 58 (6): 824–47.
- Richter, Klaus, Martin Haslbeck, and Johannes Buchner. 2010. "The Heat Shock Response: Life on the Verge of Death." *Molecular Cell* 40 (2): 253–66.
- Ritossa, F. 1962. "A New Puffing Pattern Induced by Temperature Shock and DNP in *Drosophila*." *Experientia* 18 (12): 571–73.
- . 1996. "Discovery of the Heat Shock Response." *Cell Stress & Chaperones* 1 (2): 97–98.
- Sahoo, Jyoti Prakash, Laxmipreeya Behera, Siddhartha Shankar Sharma, Jannila Praveena, Suman Kumari Nayak, and Kailash Chandra Samal. 2020. "Omics Studies and Systems Biology Perspective towards Abiotic Stress Response in Plants." *American Journal of Plant Sciences* 11 (12): 2172–94.
- Scharf, K. D., M. Siddique, and E. Vierling. 2001. "The Expanding Family of Arabidopsis Thaliana Small Heat Stress Proteins and a New Family of Proteins Containing Alpha-Crystallin Domains (Acd Proteins)." *Cell Stress & Chaperones* 6 (3): 225–37.
- Schöffl, F., R. Prändl, and A. Reindl. 1998. "Regulation of the Heat-Shock Response." *Plant Physiology* 117 (4): 1135–41.
- Schramm, Franziska, Jane Larkindale, Elke Kiehlmann, Arnab Ganguli, Gisela Englich, Elizabeth Vierling, and Pascal von Koskull-Döring. 2008. "A Cascade of Transcription Factor DREB2A and Heat Stress Transcription Factor HsfA3 Regulates the Heat Stress Response of Arabidopsis." *The Plant Journal: For Cell and Molecular Biology* 53 (2): 264–74.
- Sharkey, Thomas D. 2005. "Effects of Moderate Heat Stress on Photosynthesis: Importance of Thylakoid Reactions, Rubisco Deactivation, Reactive Oxygen Species, and Thermotolerance Provided by Isoprene." *Plant, Cell & Environment* 28 (3): 269–77.

- Smith, Martin D., Joel O. Wertheim, Steven Weaver, Ben Murrell, Konrad Scheffler, and Sergei L. Kosakovsky Pond. 2015. "Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection." *Molecular Biology and Evolution* 32 (5): 1342–53.
- Sorte, Cascade J. B., and Gretchen E. Hofmann. 2005. "Thermotolerance and Heat-Shock Protein Expression in Northeastern Pacific *Nucella* Species with Different Biogeographical Ranges." *Marine Biology* 146 (5): 985–93.
- Suzuki, Nobuhiro, Shai Koussevitzky, Ron Mittler, and Gad Miller. 2012. "ROS and Redox Signalling in the Response of Plants to Abiotic Stress." *Plant, Cell & Environment* 35 (2): 259–70.
- Vierling, E. 1991. "The Roles of Heat Shock Proteins in Plants." *Annual Review of Plant Biology* 42 (Volume 42, 1991): 579–620.
- Vierling, E., and J. A. Kimpel. 1992. "Plant Responses to Environmental Stress." *Current Opinion in Biotechnology* 3 (2): 164–70.
- Vierling, Elizabeth. 1997. "The Small Heat Shock Proteins in Plants Are Members of an Ancient Family of Heat Induced Proteins." *Acta Physiologiae Plantarum / Polish Academy of Sciences, Committee of Plant Physiology Genetics and Breeding* 19 (4): 539–47.
- Wahid, A., S. Gelani, M. Ashraf, and M. R. Foolad. 2007. "Heat Tolerance in Plants: An Overview." *Environmental and Experimental Botany* 61 (3): 199–223.
- Wang, Wangxia, Basia Vinocur, Oded Shoseyov, and Arie Altman. 2004. "Role of Plant Heat-Shock Proteins and Molecular Chaperones in the Abiotic Stress Response." *Trends in Plant Science* 9 (5): 244–52.
- Wang, Xiaojuan, Na Li, Wei Li, Xinlei Gao, Muha Cha, Lijin Qin, and Lihong Liu. 2020. "Advances in Transcriptomics in the Response to Stress in Plants." *Global Medical Genetics* 7 (2): 30–34.
- Wang, Zhiqing, Yao Mu, Xuefeng Hao, Jinbao Yang, Daixuan Zhang, Zhuping Jin, and Yanxi Pei. 2022. "H₂S Aids Osmotic Stress Resistance by S-Sulfhydration of Melatonin Production-Related Enzymes in *Arabidopsis Thaliana*." *Plant Cell Reports* 41 (2): 365–76.
- Waterhouse, Andrew M., James B. Procter, David M. A. Martin, Michèle Clamp, and Geoffrey J. Barton. 2009. "Jalview Version 2—a Multiple Sequence Alignment Editor and Analysis Workbench." *Bioinformatics* 25 (9): 1189–91.

- Waters, Elizabeth R., Garrett J. Lee, and Elizabeth Vierling. 1996. "Evolution, Structure and Function of the Small Heat Shock Proteins in Plants." *Journal of Experimental Botany* 47 (3): 325–38.
- Waters, Elizabeth R., and Elizabeth Vierling. 2020. "Plant Small Heat Shock Proteins - Evolutionary and Functional Diversity." *The New Phytologist* 227 (1): 24–37.
- Weaver, Steven, Stephen D. Shank, Stephanie J. Spielman, Michael Li, Spencer V. Muse, and Sergei L. Kosakovsky Pond. 2018. "Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes." *Molecular Biology and Evolution* 35 (3): 773–77.
- Wernersson, Rasmus, and Anders Gorm Pedersen. 2003. "RevTrans: Multiple Alignment of Coding DNA from Aligned Amino Acid Sequences." *Nucleic Acids Research* 31 (13): 3537–39.
- Yeh, Ching-Hui, Nicholas J. Kaplinsky, Catherine Hu, and Yee-Yung Charng. 2012. "Some like It Hot, Some like It Warm: Phenotyping to Explore Thermotolerance Diversity." *Plant Science: An International Journal of Experimental Plant Biology* 195 (October): 10–23.
- You, Jun, and Zhulong Chan. 2015. "ROS Regulation During Abiotic Stress Responses in Crop Plants." *Frontiers in Plant Science* 6 (December): 1092.
- Zhang, Huiming, Jianhua Zhu, Zhizhong Gong, and Jian-Kang Zhu. 2022. "Abiotic Stress Responses in Plants." *Nature Reviews. Genetics* 23 (2): 104–19.
- Zhang, Xuebin, Zinnia H. González-Carranza, Shulin Zhang, Y. Miao, Chang-jun Liu, and Jeremy A. Roberts. 2019. "F -Box Proteins in Plants." *Annual Plant Reviews Online*, February. <https://doi.org/10.1002/9781119312994.APR0701>.
- Zhou, Rengang, Bing Li, Hongtao Liu, and Daye Sun. 2009. "Progress in the Participation of Ca²⁺–calmodulin in Heat Shock Signal Transduction." *Progress in Natural Science: Communication of State Key Laboratories of China* 19 (10): 1201–8.
- Zhu, Jian-Kang. 2016. "Abiotic Stress Signaling and Responses in Plants." *Cell* 167 (2): 313–24.

Conclusion

This dissertation advances our knowledge of allopolyploidy, genomic adaptations, and heat stress responses in plants, focusing on *Nicotiana* section *Repandae* and two *Boechea* species, *Boechea arcuata* and *Boechea perennans*. Through comparative genomic and transcriptomic analyses, it elucidates the genetic mechanisms underlying these processes and their evolutionary implications.

My research on *Nicotiana* section *Repandae* allopolyploids revealed significant insights into subgenome dominance and gene expression patterns. I found that subgenome dominance in these species is biased towards the *N. obtusifolia* subgenome, which exhibits less fragmentation and higher expression levels compared to the *N. sylvestris* subgenome (Alger & Edger, 2020). This bias is particularly pronounced in *N. nudicaulis*, reflecting a stronger evolutionary influence of *N. obtusifolia* in shaping the phenotypic traits of these allopolyploids. The differential expression of candidate genes involved in floral organ size provides a clear example of how subgenome dominance and gene expression can influence phenotypic diversity. Genes such as *BB*, *DARI*, *GASA*, and *LNG* showed varying expression patterns that correlate with flower size differences among *Nicotiana* species. For instance, the higher expression of *BB* in short-flowered *N. nudicaulis* is consistent with its role as a negative regulator of organ growth (Disch et al., 2006). These findings highlight the evolutionary significance of gene expression regulation in the context of allopolyploidy and its impact on morphological traits.

The genomic analysis of *Boechea arcuata* and *Boechea perennans* provided valuable insights into the evolutionary dynamics of these species. The genomes sizes

reported for the *Boehera* species are slightly larger than those of *A. lyrata* and *A. thaliana* (Kliver et al. 2018). Nearly half of repeats were long terminal repeats (LTRs), aligning with the findings in *B. retrofracta* (Kliver et al. 2018). Despite their similar genome sizes and gene counts, I observed notable differences in gene family expansions and functional adaptations. The high level of synteny between *B. arcuata* and *B. perennans* suggests genomic stability and preservation of gene order, while the differences in gene family content reflect species-specific adaptations to their respective environments. GO enrichment analysis revealed that *B. arcuata* has a unique enrichment of genes involved in protein assembly, transport mechanisms, and energy-related processes, indicating robust cellular function under thermal stress. In contrast, *B. perennans* exhibited an enrichment of genes associated with photosynthesis, stress response, and mitochondrial function, highlighting its adaptation to arid environments. These lineage-specific gene expansions underscore the role of genomic adaptations in the ecological diversity and resilience of *Boehera* species.

The heat-stress response in *B. arcuata* and *B. perennans* demonstrates both conserved and species-specific patterns of gene expression. Both species upregulated key heat shock proteins (HSPs) such as HSP20, HSP70, and AAA+ homologs, which are essential for thermotolerance (Waters, 2013). However, the differential expression of certain genes, including distinct sets of protein kinases, suggests unique adaptations to heat stress in each species. I identified several gene families with lineage-specific expansions, such as those related to chloroplast function and osmotic stress response, which may contribute to the distinct thermotolerance observed in these species. The expansion of the

TIR-like domain-containing gene family in *B. arcuata*, for example, highlights its potential role in abiotic stress response and environmental adaptation (MacQueen & Bergelson, 2016). These findings provide a comprehensive understanding of the molecular mechanisms underlying heat stress responses and suggest avenues for future research to enhance crop resilience.

This dissertation contributes to our understanding of the genetic and evolutionary mechanisms underlying polyploidy, gene expression regulation, and environmental adaptation in plants. The insights gained from the study of *Nicotiana* and *Boechea* species emphasize the importance of subgenome dominance, gene family expansion, and stress response pathways in shaping the evolutionary trajectories of polyploid organisms. By identifying key genes involved in stress responses and metabolic processes, this work lays the foundation for developing strategies to enhance the resilience of other plant species to environmental stressors. The comparative genomic data presented here will serve as a valuable resource for future studies on the genomic and phenotypic diversity of *Boechea* and other polyploid species.

References

- Alger, Elizabeth I., and Patrick P. Edger. 2020. "One Subgenome to Rule Them All: Underlying Mechanisms of Subgenome Dominance." *Current Opinion in Plant Biology* 54 (April): 108–13.
- Disch, Sabine, Elena Anastasiou, Vijay K. Sharma, Thomas Laux, Jennifer C. Fletcher, and Michael Lenhard. 2006. "The E3 Ubiquitin Ligase BIG BROTHER Controls Arabidopsis Organ Size in a Dosage-Dependent Manner." *Current Biology: CB* 16 (3): 272–79.
- Kliver, Sergei, Mike Rayko, Alexey Komissarov, Evgeny Bakin, Daria Zhernakova, Kasavajhala Prasad, Catherine Rushworth, et al. 2018. "Assembly of the Boechera Retrofracta Genome and Evolutionary Analysis of Apomixis-Associated Genes." *Genes* 9 (4). <https://doi.org/10.3390/genes9040185>.
- MacQueen, Alice, and Joy Bergelson. 2016. "Modulation of R-Gene Expression across Environments." *Journal of Experimental Botany* 67 (7): 2093–2105.
- Waters, Elizabeth R. 2013. "The Evolution, Function, Structure, and Expression of the Plant sHSPs." *Journal of Experimental Botany* 64 (2): 391–403.