THE POWER OF SYNTENY: DEEP EVOLUTIONARY INSIGHTS FROM COMPARATIVE GENOMICS

A Dissertation

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfilment

of the Requirements for the Degree

Doctor of Philosophy

---

by

RICHARD SHAWN ABRAHAMS

Dr. J. Chris Pires, Dissertation Supervisor

JULY 2021

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled:

THE POWER OF SYNTENY: DEEP EVOLUTIONARY INSIGHTS FROM COMPARATIVE GENOMICS

Presented by R. Shawn Abrahams, a candidate for the degree of doctor of philosophy, and hereby certify that, in their opinion, it is worthy of acceptance.

_____

Dr. J. Chris Pires

_____

Dr. James A. Birchler

_____

Dr. Ruthie Angelovici

_____

Dr. Gavin C. Conant

DEDICATION

To those enslaved Black people who built a foundation for the University of Missouri, Columbia and whose dreams I hope I am a part.

ACKNOWLEDGEMENTS

I want to thank my advisor, J. Chris Pires, whose mentorship has been invaluable, starting from the first moments we met. I've learned so much about being a scientist, an academic, a hustler, and an ally from you. Thank you for always encouraging my curiosity and providing the slack I've needed to pull myself up, time and time again. I additionally want to thank my committee members Ruthie Angelovici, Jim Birchler, and Gavin Conant for their support and confidence in my abilities.

I want to thank my lab mates, who have supported me over this time as we have shared and commiserated this unique part of our lives. Jacob Washburn, Hong An, Andrea Revelo, Shawn Thomas, Michael Pisias, Sarah Turner-Hissong, Daniel Westfall, Wade Dismukes, and Kevin Bird, I know our paths will continue to cross, and I am excited to share those moments with you as well. Of course, Makenzie Mabry, who started graduate school with me, the impact you have had on my time here can not be understated. You are indeed an inspiration, and I can't wait to see what the world has in store for you in the future.

Thank you to my fellow graduate students who have helped to form a community and support structure for me here in the Midwest. Patricka Williams-Simon, Candice King, Emi Asante, Deise Cruz, Maya Parker-Smith, Nadia Patterson Stapelton, Sherryll Henderson, Rana Farber Kennedy, Levi Storks (and Kevin too!), Michael Vierling, Zack Miller, and Nat Graham. Your friendship has held me up over these years.

To my constellation of mentors who have helped me lay the groundwork for a bright future, I cannot repay the impact you've made in my life, though I hope to pay in

forward. Pamela Krauss, Krisitine Callis-Duehl, Eric Schranz, Libby King, Lauren Sullivan, David Schulz, Stephanie Shonekan, Emily Sessa, Jeremy Yoder, Dan Kliebenstein, and Ihsan Al-Shebaz. Your influences will carry me through to the next adventure in wisdom, care, patience, and curiosity.

To my chosen and gifted family, thank you for supporting me even when you couldn't always understand. Thank you for believing in me even when I could not, and for sharing that light. My late father Richie, I wish you could have seen me reach this point; I know you would have been proud. My stepfather, Esteban Rodriguez, thank you for the guidance and the extended kitchen talks. To my siblings, Marquis and Mia, here is the start of our many successes. May we always get a chance to celebrate together. To my loved ones Shai Andrew Garcia, Cortland Russel, Alice Patrice Simone Rian, and Darien Williams thank you for keeping my heart in one piece. To my mother, Ingrid Abrahams, thank you for telling me that I would have to go to graduate school, no matter the topic. I know your drive for education came from having been overlooked and under-supported in your dreams. It is a lesson I will keep with me always, and your tenacity is something I can only hope to emulate. I love you.

Finally, I want to acknowledge that the University of Missouri - Columbia, where I obtain my graduate degree, is located on the ancestral lands of Native peoples who were removed unjustly and that my academic community is the beneficiary of that injustice. To the Missouria, Illini, Osage, Ioway, Oto, Quapaw, Chickasaw, and other first nations people who call these lands home, I am sorry, and I acknowledge your continued connection to these lands.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

DISSERTATION ABSTRACT

Synteny, or the order of genes in a given genome, is an emergent property of individuals and species that has only, with the implementation of next gen-sequencing, become available for evolutionary consideration. In this dissertation, I leverage syntenic information in concert with sequence data to draw connections between evolutionary mechanisms, species divergence, and trait innovation. In Chapter I, I review the major themes that ties my dissertation research together, highlighting important mechanisms at work in evolutionary complexity and introducing the system of which it will be a part. In Chapter II, I use a phylogenomic approach to better understand species relationships within the tribe. I utilize transcriptome sequences and genome derived synteny information to improve orthology detection over standard sequence similarity approaches and gain greater insight into the relationships of the tribe. I also implement differential fractionation rate orthology inference information to address gene tree-species tree incongruence. In Chapter III, as published in Abrahams et al., 2020, I utilize a micro-synteny network and phylogenetic inference to investigate the origin and diversification of the MAM/IPMS gene family. I uncover unique MAM-like genes found at the orthologous locus in the Cleomaceae that shed light on the transition from IPMS to MAM. In the Brassicaceae, I identify six distinct MAM clades across Lineages I, II, and III. I characterize the evolutionary impact and consequences of local duplications, transpositions, whole genome duplications, and gene fusion events, generating several new hypotheses on the function and diversity of the MAM locus.

CHAPTER 1: IMPLEMENTING GENOMIC COMPARISON FOR THE STUDY OF KEY INNOVATIONS AND GENE DUPLICATION

R. Shawn Abrahams[1]

1 Division of Biological Sciences and Bond Life Sciences Center, University of Missouri, Columbia, Missouri 65211, USA

BACKGROUND

All extant lineages of life have ancestors with traits that allowed them to maintain evolutionary fitness in the context of their environment. Some of these ancestors weren't just fit. They flourished in their environments, these adaptations allowed them to become some of the most speciose and dominant lineages on Earth. These adaptations, called key innovations, are novel phenotypic traits that have resulted in evolutionary radiations, where lineages experience increased levels of speciation and diversification (Miller 1949; Hunter 1998; Soltis & Soltis, 2016). Identifying a trait as a true "key innovation" can be confounded by various factors, including difficulty in connecting specific traits to increased diversification, issues in connecting causal genotype to the trait phenotype, and difficulty in defining the bounds of a trait in biological terms. Wherein trait as defined may be a series of biological innovations that form a fitness advantage when primarily when occurring together (Cracraft, 1990; Galis, 2001; Donoghue, 2005; Soltis & Soltis, 2016). For example, the evolution of the flower is considered a key innovation of the land plant's angiosperm clade. Still, it could also be looked at as a series of structural changes at different nodes of the angiosperm phylogeny that led to the development of the perianth, stamens, carpels, and the complex feature of double fertilization. Stepwise innovations may have occurred millions of years apart but still, come together at critical points of species radiation in the phylogeny. To this end, it's essential to reconstruct the origin of these traits in evolutionary time and describe and place the underlying genomic mechanisms that may have led to the origin of these traits. Studying these trait

innovations can give us insight into the mechanisms that have underwritten biological complexity.

Polyploidy as a mechanism for innovation

Whole genome duplication (WGD) has been identified as an important mechanism in the origin and diversification of angiosperm species (Buzgo et al. 2004; de Bodt et al. 2005; Zahn et al. 2005; Jiao et al. 2011). WGD can occur due to mistakes in meiosis where one or both parent's gametes fail to reduce and can still form a viable offspring whose ploidy level (n value) is increased relative to their progenitors. Some have suggested that polyploidy can limit diversification in flowering plants and that polyploids are evolutionary dead-ends (Stebbins 1950; Arrigo and Barker 2012; Mayrose et al., 2011). In contrast, others have shown that WGD can be associated with upticks in the species diversification rates across various Eukaryotic lineages (Schranz et al. 2012; Tank et al. 2015; Wood et al. 2009; Vamosi and Dickinson 2006; Barker et al. 2008; Barker et al. 2016; Moriyama and Koshiba-Takeuchi 2018; Robertson and Gundappa 2017; Soltis et al. 2009; Soltis et al. 2014; Soltis and Soltis 2016; Tate and Simpson 2003; Van de Peer et al. 2017; Freeling & Thomas 2006).

Over generations, the processes of diploidization, a series of chromosomal rearrangements, translocations, fractionation, and biased gene retention, return a polyploid lineage to a functional diploid genetic system (2n) having removed generally disadvantageous or neutral gene duplicates (Freeling et al., 2015). Genes can be biasedly retained based on the function of the gene and gene dosage effects (Barker et al., 2008; Birchler & Veitia, 2012). Gene duplicates that are retained can collect mutations over time through relaxed selection and can shift function. Subfunctionalization is when

paralogous genes divide the function of an ancestral task over both duplicates, potentially resulting in tissue specification or the decoupling of enzymatic processes. Neofunctionalization is when a paralog gains a novel role after a period of neutral mutation or subfunctionalization. (Conant & Wolfe, 2008). These processes can lay the groundwork for the development of new traits that may offer a selective advantage. Drawing connections between these traits, WGD, and increased rates of diversification can provide us with greater insight into the origins of species diversity.

The Brassicales as a system

The order Brassicales, made up of 17 families with about 4700 species (Magallon et al. 1999), is one of the most economically significant plant lineages studied today. Its diversity of desirable traits, both in terms of human consumption and adaptability, make it an ideal system for understanding the connections between genomic mechanisms and key innovations. The order is well known for a series of characterized WGD events occurring across vast time scales, encompassing paleo-, meso-, and neo-polyploid events. The major paleopolyploid events as described in the genome of the model system *Arabidopsis thaliana*, are At- α, shared with all members of the Brassicaceae (Vision et al. 2000, Haudry et al. 2013; Edger et al. 2015), At-β, near the base of the Brassicales (Edger et al., 2015, 2018a), and At-γ an older event shared by all angiosperms [Figure 1.1].

An important key innovation associated with WGD is a group of specialized metabolites called mustard oils or glucosinolates (Ehrlich & Raven 1964; Edger et al. 2015). These are a class of defense compounds found in plants of the order Brassicales (Fahey et al. 2001; Daxenbichler et al. 1991). They are also found in the genera *Drypetes*

and *Putranjiva* of the family Putranjivaceae, formerly of the Euphorbiaceae (Rodman et al. 1998; Soltis & Soltis, 2004), and putatively present in the genus *Rinorea* of the Violaceae (Montaut et al. 2017), though these latter occurrences have more limited diversity. Physical damage from an herbivore, such as chewing, causes compartments of the plant cell to rupture and release a myrosinase enzyme that hydrolyzes the glucosinolates to create an isothiocyanate anion, damaging the attacker (Agrawal & Kurashige 2003). Following WGD events at At-β and At-α, there are increases in species radiation, and innovations in the biosynthesis pathway that resulted in novel classes of glucosinolate compounds (Edger et al., 2015). Although these events cover a wide distribution of taxa, we know many of the most diverse clades in this group have a recurrent history of WGD. The effects of these more recent events on the evolution of the glucosinolate biosynthesis pathway have yet to be clearly described. Thus, an open question is whether subsequent gene and genome duplication events in the speciose clades of the Brassicaceae, Cleomaceae, and Capparaceae are associated with glucosinolate novelty.

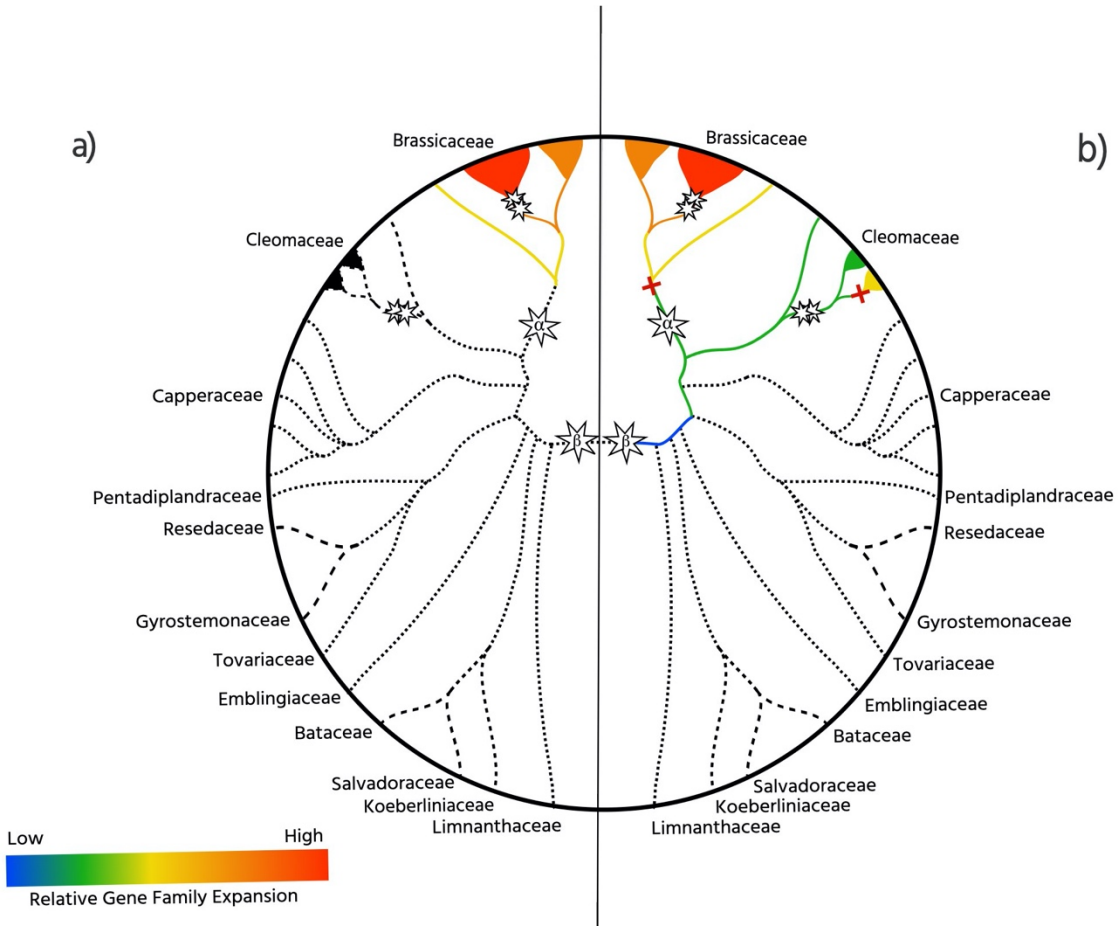


Figure 1.1 The impact of syntenic approaches on gene family understandings. As summarized by Abrahams et al. 2020 A), Gene family expansion of the glucosinolate biosynthesis pathway was primarily understood through the Brassicaceae family. This expansion was limited to the MAM-ancestral locus. Estimates of diversity reflected primarily the role tandem duplications played in gene family expansion without any estimation for when key events occurred in the transition from IPMS identity into what we recognize as MAM identity. B) With the inclusion of the Cleomaceae in the phylogenic estimation of the plant family, key "missing link" states were identified, and critical loss parallel domain loss events were essential to the function of MAM in the context of phylogeny. With an updated understanding of gene family expansion many, more questions can be asked regarding the role intermediate states of MAM play in glucosinolate diversity and the relative roles different gene duplication types play in the evolutionary trajectory of specialized metabolism. With the origin of MAM being traced to the IPMS duplication event, we can assume all post Beta event Brassicales families have an orthologous MAM locus, but currently without the genomic resources for the comparison and characterization.

Synteny as a tool

As a part of this dissertation, I explore the role WGD, and other gene duplication types, have had on species evolution and trait innovation throughout the Brassicales. These studies are made possible by advancements in genomic sequencing and technology, allowing for syntenic or gene address information to supplement phylogenetic analyses and improve our estimation of past biological events. In Chapter II, I utilize genomic comparison to improve phylogenomic analysis of the tribe Brassiceae of the family Brassicaceae. This analysis emphasizes the methodological difficulties that are created by WGD when trying to reconstruct lineage divergence. I also show how, in mesopolyploid lineages, genomic differences between parental sub-genomes can be uncovered millions of years after the parental species have gone extinct. In Chapter III, I leverage a syntenic network approach to characterize the MAM/IPMS gene family's evolutionary dynamics in relation to the glucosinolate biosynthesis pathway. The combination of phylogenetic and syntenic methods allowed for the unique identification of a mid-state or "missing link" in the transition from a primary metabolic role to a specialized metabolic function [Figure 1.1]. Biology is at the very beginning of what we will understand is the full capability of this quickly developing technology. Hopefully, the research detailed in the coming chapters is a small part of what science can gain from broadening our sampling of diverse plant genomes beyond standard model species.

REFERENCES

Agrawal, Anurag A., and Nile S. Kurashige. 2003. "A Role for Isothiocyanates in Plant Resistance against the Specialist Herbivore Pieris Rapae." *Journal of Chemical Ecology* 29 (6): 1403–15. https://doi.org/10.1023/a:1024265420375.

Arrigo, Nils, and Michael S. Barker. 2012. "Rarely Successful Polyploids and Their Legacy in Plant Genomes." *Current Opinion in Plant Biology* 15 (2): 140–46. https://doi.org/10.1016/j.pbi.2012.03.010.

Barker, M. S., N. C. Kane, M. Matvienko, A. Kozik, R. W. Michelmore, S. J. Knapp, and L. H. Rieseberg. 2008. "Multiple Paleopolyploidizations during the Evolution of the Compositae Reveal Parallel Patterns of Duplicate Gene Retention after Millions of Years." *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msn187.

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. https://doi.org/10.3732/ajb.1600272.

Birchler, James A., and Reiner A. Veitia. 2012. "Gene Balance Hypothesis: Connecting Issues of Dosage Sensitivity across Biological Disciplines." *Proceedings of the National Academy of Sciences of the United States of America* 109 (37): 14746–53. https://doi.org/10.1073/pnas.1207726109.

Buzgo, Matyas, Douglas E. Soltis, Pamela S. Soltis, and Hong Ma. 2004. "Towards a Comprehensive Integration of Morphological and Genetic Studies of Floral

Development." *Trends in Plant Science*. https://doi.org/10.1016/j.tplants.2004.02.003.

Conant, Gavin C., and Kenneth H. Wolfe. 2008. "Probabilistic Cross-Species Inference of Orthologous Genomic Regions Created by Whole-Genome Duplication in Yeast." *Genetics* 179 (3): 1681–92. https://doi.org/10.1534/genetics.107.074450.

Daxenbichler, Melvin E., Gayland F. Spencer, Diana G. Carlson, Gertrude B. Rose, Anita M. Brinker, and Richard G. Powell. 1991. "Glucosinolate Composition of Seeds from 297 Species of Wild Plants." *Phytochemistry* 30 (8): 2623–38. https://doi.org/10.1016/0031-9422(91)85112-D.

De Bodt, Stefanie, Steven Maere, and Yves Van de Peer. 2005. "Genome Duplication and the Origin of Angiosperms." *Trends in Ecology & Evolution* 20 (11): 591–97. https://doi.org/10.1016/j.tree.2005.07.008.

Donoghue, Michael J. 2005. "Key Innovations, Convergence, and Success: Macroevolutionary Lessons from Plant Phylogeny." *Paleobiology* 31 (sp5): 77–93. https://doi.org/10.1666/0094-8373(2005)031[0077:KICASM]2.0.CO;2.

Edger, Patrick P., Hanna M. Heidel-Fischer, Michaël Bekaert, Jadranka Rota, Gernot Glöckner, Adrian E. Platts, David G. Heckel, et al. 2015. "The Butterfly Plant Arms-Race Escalated by Gene and Genome Duplications." *Proceedings of the National Academy of Sciences of the United States of America* 112 (27): 8362–66. https://doi.org/10.1073/pnas.1503926112.

Ehrlich, Paul R., and Peter H. Raven. 1964. "Butterflies and Plants: A Study in Coevolution." *Evolution; International Journal of Organic Evolution* 18 (4): 586–608. https://doi.org/10.2307/2406212.

Freeling, Michael, and Brian C. Thomas. 2006. “Gene-Balanced Duplications, like Tetraploidy, Provide Predictable Drive to Increase Morphological Complexity.” *Genome Research* 16 (7): 805–14. https://doi.org/10.1101/gr.3681406.

Galis, F. 2001. “Key Innovations and Radiations In: Wagner GP, Editor., Editor. The Character Concept in Evolutionary Biology.” San Diego: Academic Press.

Haudry, Annabelle, Adrian E. Platts, Emilio Vello, Douglas R. Hoen, Mickael Leclercq, Robert J. Williamson, Ewa Forczek, et al. 2013. “An Atlas of over 90,000 Conserved Noncoding Sequences Provides Insight into Crucifer Regulatory Regions.” *Nature Genetics* 45 (8): 891–98. https://doi.org/10.1038/ng.2684.

Hunter, J. P. 1998. “Key Innovations and the Ecology of Macroevolution.” *Trends in Ecology & Evolution* 13 (1): 31–36. https://doi.org/10.1016/s0169-5347(97)01273-1.

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*. https://doi.org/10.1038/nature09916.

Magallon, Susana, Peter R. Crane, and Patrick S. Herendeen. 1999. “Phylogenetic Pattern, Diversity, and Diversification of Eudicots.” *Annals of the Missouri Botanical Garden. Missouri Botanical Garden* 86 (2): 297–372. https://doi.org/10.2307/2666180.

Mayrose, Itay, Shing H. Zhan, Carl J. Rothfels, Karen Magnuson-Ford, Michael S. Barker, Loren H. Rieseberg, and Sarah P. Otto. 2011. “Recently Formed

Polyploid Plants Diversify at Lower Rates." *Science* 333 (6047): 1257. https://doi.org/10.1126/science.1207205.

Miller, Alden H. 1949. "Some Ecologic and Morphologic Considerations in the Evolution of Higher Taxonomic Categories." *Ornithologie Als Biologische Wissenschaft* 28: 84–88.

Montaut, S., G. R. De Nicola, and H. Agnaniet. 2017. "Probing for the Presence of Glucosinolates in Three Drypetes spp.(Drypetes Euryodes (Hiern) Hutch., Drypetes Gossweileri S. Moore, Drypetes Laciniata Hutch.) and …." *Natural Product*. https://www.tandfonline.com/doi/abs/10.1080/14786419.2016.1236099.

Moriyama, Yuuta, and Kazuko Koshiba-Takeuchi. 2018. "Significance of Whole-Genome Duplications on the Emergence of Evolutionary Novelties." *Briefings in Functional Genomics* 17 (5): 329–38. https://doi.org/10.1093/bfgp/ely007.

Peer, Yves Van de, Yves Van de Peer, Eshchar Mizrachi, and Kathleen Marchal. 2017. "The Evolutionary Significance of Polyploidy." *Nature Reviews Genetics*. https://doi.org/10.1038/nrg.2017.26.

Robertson, Fiona M., Manu Kumar Gundappa, Fabian Grammes, Torgeir R. Hvidsten, Anthony K. Redmond, Sigbjørn Lien, Samuel A. M. Martin, Peter W. H. Holland, Simen R. Sandve, and Daniel J. Macqueen. 2017. "Lineage-Specific Rediploidization Is a Mechanism to Explain Time-Lags between Genome Duplication and Evolutionary Diversification." *Genome Biology*. https://doi.org/10.1186/s13059-017-1241-z.

Rodman, J., P. Soltis, D. Soltis, K. Sytsma, and K. Karol. 1998. "Parallel Evolution of Glucosinolate Biosynthesis Inferred from Congruent Nuclear and Plastid Gene

Phylogenies." *American Journal of Botany* 85 (7): 997. https://doi.org/10.2307/2446366.

Schranz, M. Eric, M. Eric Schranz, Setareh Mohammadin, and Patrick P. Edger. 2012. "Ancient Whole Genome Duplications, Novelty and Diversification: The WGD Radiation Lag-Time Model." *Current Opinion in Plant Biology*. https://doi.org/10.1016/j.pbi.2012.03.011.

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. https://doi.org/10.3732/ajb.0800079.

Soltis, Douglas E., María Claudia Segovia-Salcedo, Ingrid Jordon-Thaden, Lucas Majure, Nicolas M. Miles, Evgeny V. Mavrodiev, Wenbin Mei, María Beatriz Cortez, Pamela S. Soltis, and Matthew A. Gitzendanner. 2014. "Are Polyploids Really Evolutionary Dead-ends (again)? A Critical Reappraisal of Mayrose et Al. (2011)." *New Phytologist*. https://doi.org/10.1111/nph.12756.

Soltis, Pamela S., and Douglas E. Soltis. 2004. "The Origin and Diversification of Angiosperms." *American Journal of Botany* 91 (10): 1614–26. https://doi.org/10.3732/ajb.91.10.1614.

———. 2016. "Ancient WGD Events as Drivers of Key Innovations in Angiosperms." *Current Opinion in Plant Biology* 30 (April): 159–65. https://doi.org/10.1016/j.pbi.2016.03.015.

Stebbins, G. Ledyard, and G. Ledyard Stebbins. 1950. "Variation and Evolution in Plants." https://doi.org/10.7312/steb94536.

Tank, David C., Jonathan M. Eastman, Matthew W. Pennell, Pamela S. Soltis, Douglas E. Soltis, Cody E. Hinchliff, Joseph W. Brown, Emily B. Sessa, and Luke J. Harmon. 2015. "Nested Radiations and the Pulse of Angiosperm Diversification: Increased Diversification Rates Often Follow Whole Genome Duplications." *The New Phytologist* 207 (2): 454–67. https://doi.org/10.1111/nph.13491.

Tate, Jennifer A., and Beryl B. Simpson. 2003. "Paraphyly of Tarasa (Malvaceae) and Diverse Origins of the Polyploid Species." *Systematic Botany* 28 (4): 723–37. https://doi.org/10.1043/02-64.1.

Vamosi, Jana C., and Timothy A. Dickinson. 2006. "Polyploidy and Diversification: A Phylogenetic Investigation in Rosaceae." *International Journal of Plant Sciences* 167 (2): 349–58. https://doi.org/10.1086/499251.

Vision, T. J., D. G. Brown, and S. D. Tanksley. 2000. "The Origins of Genomic Duplications in Arabidopsis." *Science* 290 (5499): 2114–17. https://doi.org/10.1126/science.290.5499.2114.

Wood, Troy E., Naoki Takebayashi, Michael S. Barker, Itay Mayrose, Philip B. Greenspoon, and Loren H. Rieseberg. 2009. "The Frequency of Polyploid Speciation in Vascular Plants." *Proceedings of the National Academy of Sciences of the United States of America* 106 (33): 13875–79. https://doi.org/10.1073/pnas.0811575106.

Zahn, Laura M., Hongzhi Kong, James H. Leebens-Mack, Sangtae Kim, Pamela S. Soltis, Lena L. Landherr, Douglas E. Soltis, Claude W. Depamphilis, and Hong Ma. 2005. "The Evolution of the SEPALLATA Subfamily of MADS-Box Genes: A Preangiosperm Origin with Multiple Duplications throughout Angiosperm

History.” *Genetics* 169 (4): 2209–23.

https://doi.org/10.1534/genetics.104.037770.

CHAPTER 2: GENOME GUIDED PHYLOTRANSCRIPTOMICS OF THE TRIBE BRASSICEAE OF THE MUSTARD FAMILY (BRASSICACEAE)

R. Shawn Abrahams[1], Shawn Thomas[1], Tatiana Arias[1], Jacob Washburn[1,2], J. Chris Pires[1]

1 Division of Biological Sciences and Bond Life Sciences Center, University of Missouri, Columbia, Missouri 65211, USA

2 USDA Plant Genetics Research Unit – Curtis Hall, University of Missouri, Columbia Missouri 65211, USA

ABSTRACT

The tribe Brassiceae of the Mustard Family has a complex evolutionary history of whole genome duplication and hybridization. These mechanisms contribute to phenotypic plasticity and adaptability of the tribe, while also confounding attempts to understand the relationships among the species of the tribe. To account for these complications, we utilize a genome guided phylotranscriptomic method that leverages genomic synteny to inform phylogenomic inference. With this method we infer a nuclear tree with a novel topology that places the *Crambe* clade as sister to both the *Nigra* and *Rapa/Oleracea* group species. Further investigation of single copy genes from ancestral sub-genome of the tribal hexaploidy identified differences in topology among genes derived from different parental genomes. These findings help explain why we see various topologies based on the data type method of inference when attempting to reconstruct clade relationships.

INTRODUCTION

The tribe Brassiceae of the Mustard family (Brassicaceae) is the most economically important lineage in the family containing crop plants critical to global human nutrition. It contains many of the members of the genus *Brassica* as well as other species (e.g., *Raphanus sativa,* Radishes; *Eruca vesicaria*, Arugala). That said, generic relationships within the tribe show major polyphyletic breakdown, especially in the three largest genera, Brassica, Diplotaxis, and Erucastrum (Yanagino et al. 1987; Song et al. 1990). The more significant clade designations show distinct incongruence between different molecular data sources (Hall et al., 2011; Arias & Pires, 2012). We see this acutely in the "Core Brassiceae" which houses most of the species diversity and all major crops.

This taxonomic confusion is due in part to the variable morphology of plants in the tribe. Key features used to classify species (e.g., fruit and leaf shape) display patterns of trait convergence and therefore are uninformative for taxonomic classification (Al-Shehbaz 2012). The tribal clades were assigned as seven groups based on morphology (Schultz 1919, 1957), and then, based on chloroplast markers, were understood as eight clades: *Oleracea/Rapa, Nigra, Savignya, Cakile, Crambe, Henophyton, Vella,* and *Zilla* (Arias & Pires 2012). When compared to nuclear data, there is significant incongruence, both in clade branching order and in the integrity of the *Rapa/Oleracea* and *Nigra* clades [Supplemental 2.1] (Warwick & Saunder 2005; Warwick & Hall 2009). These conflicts are critical to why researchers have had difficulty re-classifying the tribe's genera.

The chloroplast locus shows incongruence with nuclear loci in cases of introgression, incomplete lineage sorting, and difficult orthology detection caused by gene

duplication (Wendel & Doyle 1998; van der Niet et al. 2008; Rokas et al. 2003; Linder et al. 2004; Mendes & Hahn 2016; Lysak & Lexer 2006; Schranz et al. 2006). The Brassiceae is well known for its cases of introgression, occurring within genera (Nagaharu 1935), between genera (Mizushima, 1950; Dolstra 1982), and even with species placed outside of the tribe (Li et al. 1995; Li et al. 1998) although researchers have yet to describe the full natural history of hybridization within the tribe. The Brassiceae's complex genomic history of WGD is another cause for the incongruence seen between molecular data (Beilstein et al. 2008; Lysak et al. 2005; Hall et al. 2011; Arias and Pires 2012; Lysak et al. 2007).

While gene duplicates may confound any nuclear phylogenetic inference, ITS markers have shown specific sensitivity when true orthology cannot be determined (Álvarez & Wendel 2003). The history of recurrent whole-genome duplication (WGD) events in the tribe has culminated in much of the extant genomic diversity. This legacy of WGD includes the paleopolyploid events At-α & At-β shared with the rest of the Brassicaceae and a hexaploidy event specific to the tribe. Reconstructing the evolutionary relationships in the tribe with this history of hybridization and polyploidy is difficult when considering sequence similarity (Bastide et al., 2017; Mendes and Hahn 2016; Solis-Lemus et al. 2017; Szollosi et al. 2015).

As a part of diploidization, fractionation can result in nuclear phylogeny gene trees that do not match the inferred species tree (Mayfield-Jones et al. 2013). Gene dosage effects (Birchler and Veitia 2012; Conant et al. 2014), transposon load (Bird et al. 2018; Bird et al. 2021), and random chance govern the loss pattern for these genes. The most common method of accounting for these effects is the use of single-copy gene lists. However, this leaves many potentially informative genes out of the analysis and biases genomic selection

to that of genes kept at single copy based on their functional dosage balance. Syntenic comparison has been used to improve our understanding of lineage divergence and account for the difficulties of using sequence comparison alone.

The majority of synteny methods require quality genomes for all taxon comparisons (Soderlund et al., 2011; Jun et al., 2009; Daniels et al., 2010). Because of this, studies are often limited by resource availability and are unable to utilize syntenic methods. Some methods utilize genomic comparisons to inform the selection of genes from transcriptome sequencing, effectively lowering the available resources necessary to analyze non-model species (Washburn et al., 2017). These methods leverage outgroup and in-group genomic comparison to build a candidate gene list of true orthologs based on synteny information (i.e., the physical position of genes within the genome). Added syntenic information has the potential to provide clear insight into otherwise obscured clades. However, such methods have yet to be used to infer lineage relationships in the context of recent WGT.

MATERIALS AND METHODS

Taxon Sampling & Extraction

Thirty in-group and four outgroup species were selected for analysis. Plants were grown in a greenhouse environment. Upon reaching sufficient maturity, leaf samples were taken for RNA, DNA, and genome size. Where necessary, accessions were grown to full maturity to serve as voucher specimens to be submitted to the herbarium at Missouri Botanical Garden. RNA & DNA leaf samples were flash-frozen in liquid nitrogen and RNA was extracted using the PureLink RNA Mini Kit from Ambion and stored at -80 degrees C. DNA was extracted from some samples using the DNeasy Mini

Kit from Quiagen or Urea extraction protocol and stored at -80 degrees C. Genome size was determined through flow cytometry.

Sequencing mRNA library preparation was performed using Truseq RNA kit (NonStranded) and samples were sequenced using a read size 2X100 on an Illumina Hi-seq (6 samples per lane) at the University of Missouri Sequencing Core. Genome survey sequencing (GSS) library prep was conducted using either a DNA PCR-Free library prep (Truseq) or a Nextera Genomic kit. A read size of 2X100 was obtained on an Illumina Hi-seq (24 samples per lane) at the University of Missouri Sequencing Core.

Genome Survey Sequencing and Chloroplast Phylogeny

Chloroplast DNA was processed following a GSS analysis pipeline (https://bit.ly/3xjfCTk). Sequence quality was checked using FastQC ver. 0.11.5 (Andrews 2010) and filtered with Prinseq ver 0.20.4 (Schmieder and Edwards 2011). A reference database was generated using *Arabidopsis thaliana* and *Brassica napus* chloroplast protein sequences from NCBI Organelle Genome Resource Database. Assembly was performed using SPADES ver. 3.10.0 and combined overlapping reads using CAP3 software. Assemblies were annotated using BLAST (Camacho et al. 2009) and aligned using MAFFT ver. 7.299 (Katoh et al. 2002; Katoh & Standley 2013). Alignments were cleaned using Mesquite and concatenated gene trees were inferred in RAxML ver. 8.2.11(Stamitakis 2014) for 1000 bootstraps.

Transcriptome Assembly and Nuclear Phylogeny A genome-guided phylotranscriptomic method was followed for phylogenomic inference (Washburn et al. 2017) (https://bit.ly/2TsUD1U). RNA-seq data were quality checked using custom scripts

(Yang and Smith 2014). Trinity ver. 2.3.2 was used for further processing and de novo assembly of transcriptomes (Grabherr et al. 2011; Haas et al. 2013) and converted into peptide files using Transdecoder ver. 3.0.1. The sequenced genomes *Brassica rapa* (id32114) and *Thellungiella halophila* (id38350), as available on CoGe (https://genomevolution.org/CoGe), represented in-group and out-group taxa, respectively, and were used for the initial syntenic ortholog determination. They were selected among various genomes based on the genome quality, relative ploidy level, and phylogenetic placement. Syntenic orthologs between *B. rapa* and *T. halophila* were inferred using the SynMap tool in CoGe with the QuotaAlign set to filter out syntenic paralogous regions using a quota setting of 1:3 (Lyons et al. 2008; Tang et al. 2011). Protein sequences of the *B. rapa* representative orthologs were used as references for the analysis. The assembled transcripts were then mapped to the *B. rapa* reference orthologs using BLAST. Sequences were then grouped into orthologous sets for each gene, and multiple alignments was created using MAFFT ver. 7.299 (Katoh et al. 2002; Katoh & Standley 2013). They were then filtered using phyutility and custom scripts (Smith & Dunn 2008; Yang and Smith 2014). A coalescent species tree was created using RAxML ver. 82.11 (Stamitakis 2014) to generate gene trees and ASTRAL III v. 5.6.1 to generate the species tree (Mirarab et al. 2014a; Mirarab et al. 2014b). The program Phyparts (Smith et al. 2015) was used for calculation of gene tree discordance and visualized using Phyparts Piecharts (https://github.com/mossmatters/phyloscripts/tree/master/phypartspiecharts).

RESULTS

Chloroplast Tree Inference

The chloroplast [Figure 2.1] tree shows the *Zilla* clade (represented by *Schowia thebiaca* and *Zilla macroptera* with a bootstrap score of 95) as sister to the rest of the tribe. *Vella anherermica* and *Psychine stylosa* are sister to each other, although with low bootstrap support of 42. The *Nigra & Crambe* clades also fall within the Core Brassiceae, sister to one another. *Cakile* is sister to all four previous groups as the outermost members of the Core Brassiceae.

Genome Guided Phylotranscriptomics

Vella anhemerica is sister to the rest of the tribe in the nuclear tree [Figure 2.2], with the fully supported Zilla clade branching from the next node. *Psychine stylosa* is sister to what was considered the *Savignya* clade, though with low support values. Again, the *Cakile* clade makes up the first branch of the core Brassiceae. As defined by the chloroplast tree, the *Rapa/Oleracea* and *Nigr*a groups do not represent monophyletic clades in the nuclear tree. The *Crambe* clade falls out as a sister to that grouping with complete support. The lowest tree support occurs within the core Brassiceae, particularly between members of the *Nigra* chloroplast clade. We see a significant signal of incongruent topologies at several nodes throughout the tribe when looking at gene-tree species-tree incongruence [Supplemental Figure 2.2]. However, there is a primarily supported topology and no significant secondary topology.

Sub-Genome Analysis

Least Fractionated Sub-Genome

In this tree [Supplemental Figure 2.3]*, Vella anhermeica* is inferred as a sister to the rest of the tribe. The *Savignya* clade presents with low local posterior probability (LPP) of 0.61 and is sister to the core Brassiceae. The *Rapa/Oleracea* clade, as defined by the chloroplast tree, does not represent a monophyletic grouping. However, most species in that clade show a close relationship. In the nuclear tree, *Coincya longerorstra* from the *Nigra* chloroplast clade is nested within *Rapa/Oleracea* grouping while *Erucastrum nasturtrifolliums* and *Raphanus raphanistrum* of the *Rapa/Oleracea* chloroplast clade occur within the majority of *Nigra* grouping. The majority of species found in the *Nigra* clades cluster together, though with low support at some internal noes. The *Crambe* clade maintains strong support and is sister to the majority of *Nigra* clade species.

Most Fractionated Sub-Genome I

In this tree [Supplemental Figure 2.4] *Vella anhmerica and Psychine stylosa* form a clade that is sister to the rest of the tribe, with a lower support value of 0.63. The *Savignya* clade also displays some lower support between *Fezia pterocarpa* & *Savigyra parviflora*. The *Rapa/Oleracea* clade & *Nigra* clade display a similar breakdown to that of the nuclear tree topology but with *Crambe* as sister to all *Rapa/Oleracea* & *Nigra* species. The *Nigra* group shows several internal nodes with low support.

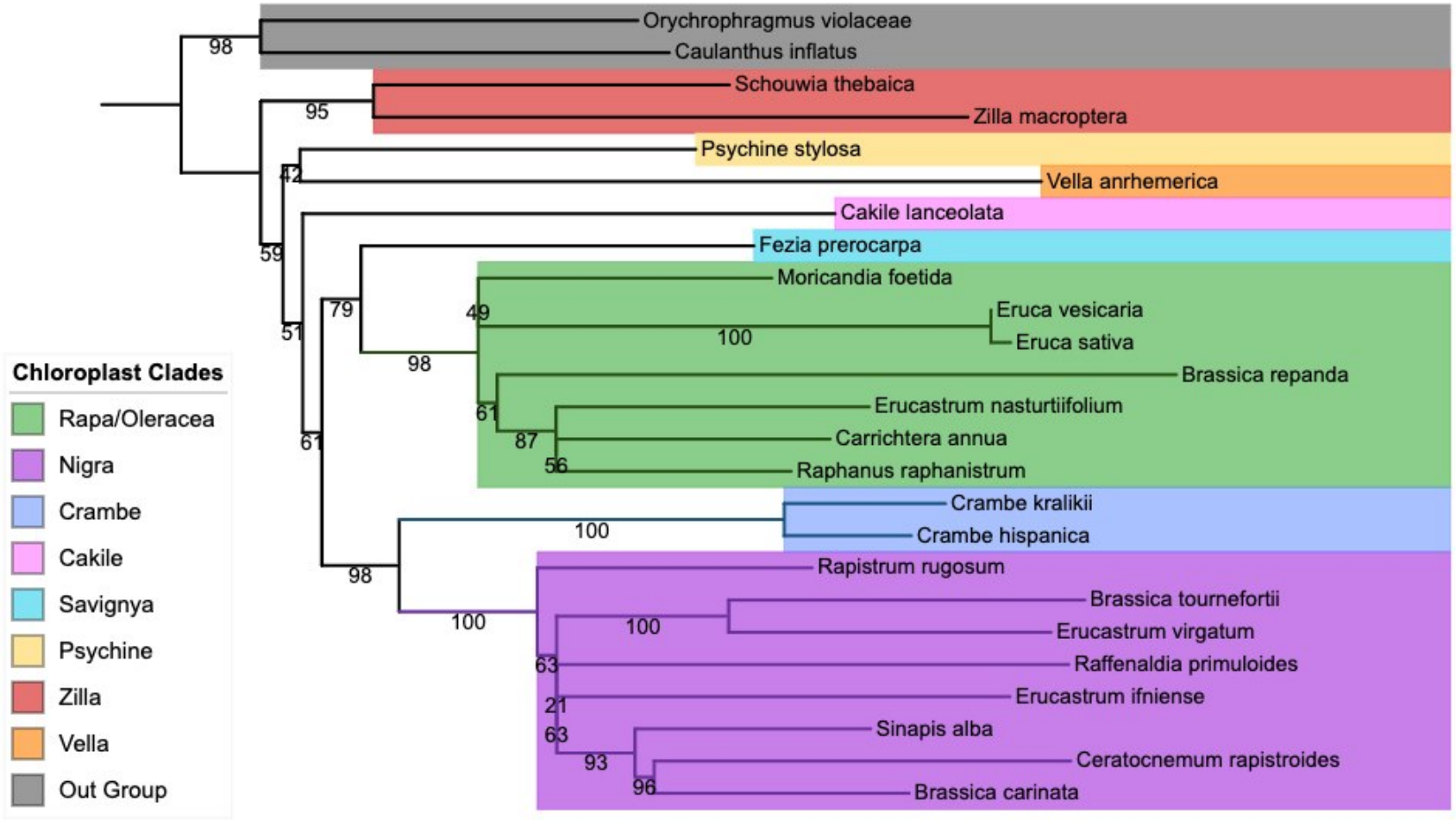


Figure 2.1: Chloroplast tree concatenated of 25 representative species of the tribe Brassiceae with 1000 bootstraps. This tree recovers most of the clades outlined in Arias 2012, but does not include the Henophyton clade which has gone unsampled, and places *Psychine stylosa* as its own grouping as it does not fall with the expected Savignya clade.

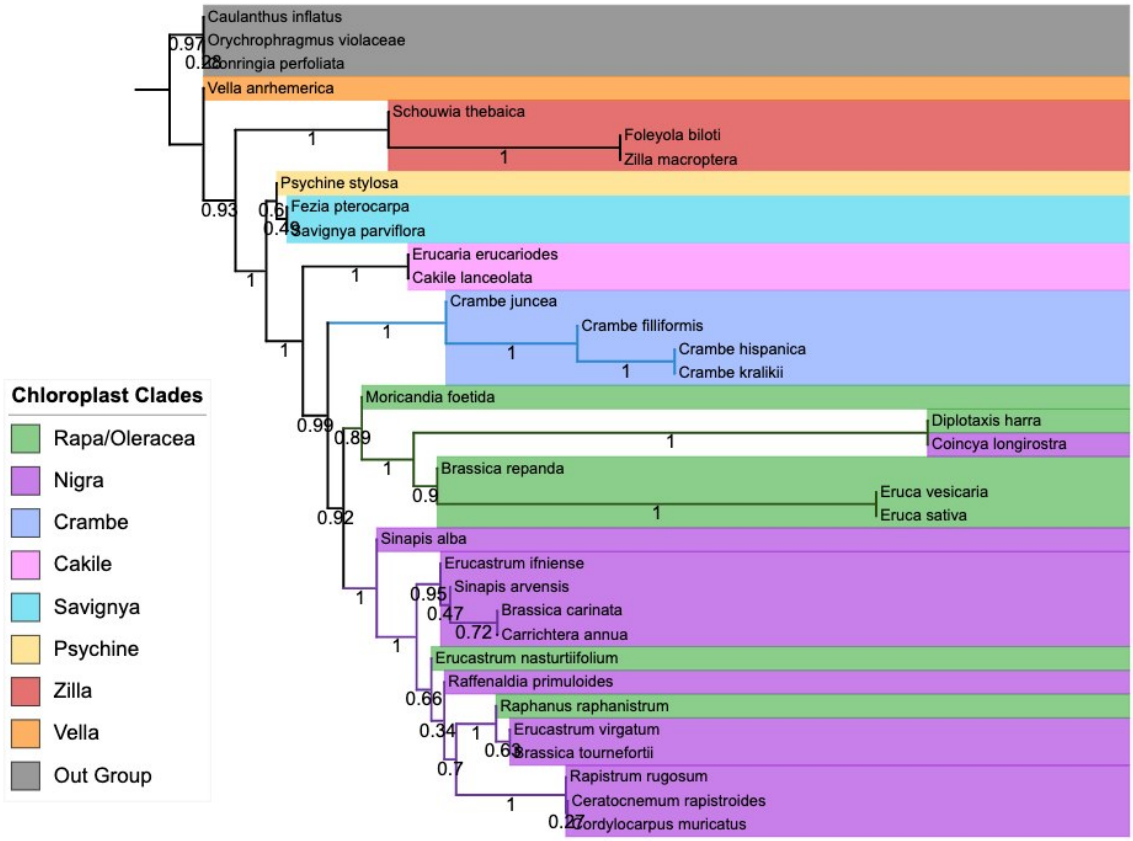


Figure 2.2: Genome-Guided Phylotranscriptomic Nuclear Tree of representative species from the tribe Brassiceae. Species have been colored based on expected chloroplast phylogeny groupings. Local Posterior Probability scores as implemented in Astral are used as the support for branch relationships. Obvious break down of the Nigra and Rapa/Oleracea clades is seen as a part of the core Brassiceae.

Most Fractionated Sub-Genome II

The most fractionated sub-genome two topology [Supplemental Figure 2.4] has the lowest overall support values out of all sub-genome trees. The *Nigra* and *Rapa/Oleracea* clades are further broken down, with *Sinapis alba* & *Moricandia foetida* falling outside of their internal placement as seen in other nuclear trees. *Crambe* clade comes out as sister to all *Rapa/Oleracea* & *Nigra* group species with the support value of 0.85.

DISCUSSION

Gene-tree species-tree relationships underly our understandings of lineage divergence in all eukaryotic lineages. Many methodological approaches have been used to simplify complex ancestries to accommodate the limits of methodology (e.g., single-copy gene lists, orthogroup analyses, and genomic comparison). Methods that utilize single-copy gene lists have been amongst the most prevalent, especially for comparison across deep phylogenetic depths (Aguileta et al. 2008). These nuclear approaches are a less biased approach when compared to using chloroplast or mitochondrial DNA. These data types are inherited uniparentally, and therefore may fail to represent hybridization/introgression events that can occur between species. In this study, we demonstrate that single-copy gene lists are not free from potential biases associated with complex genetic histories and the evolutionary history of the tribe Brassiceae makes for a compelling model to explore phylogenomic complexity.

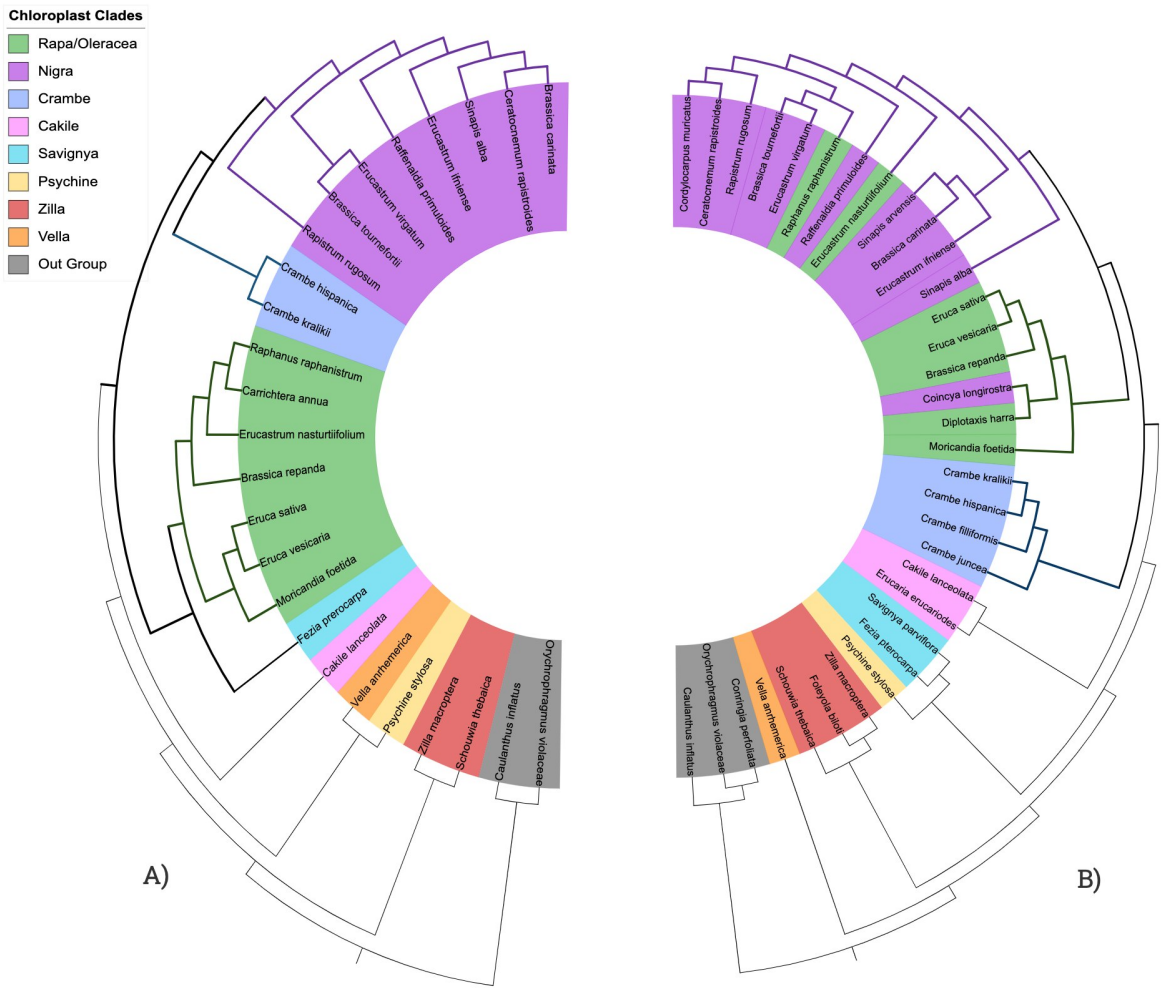


Figure 2.3: Incongruence between the Chloroplast and Nuclear Trees. A) The chloroplast tree meets most expectations for clade relationships though with Psychine displaying variable positions. The core Brassiceae topology shows the Nigra clade as sister to the Crambe clade and the Rapa/Oleracea clade as sister to the Fezia pterocarpa. B) The nuclear tree displays a breakdown of the Rapa/Oleracea and Nigra clades, with the Crambe clade presenting as sister to all members.

The tribe Brassiceae, as a hexaploid lineage, represents several levels of genomic complexity and technological resources for interrogating the mechanisms that underlie that complexity. The chloroplast and nuclear trees from this study recover an incongruence that has been seen in previous studies of the tribe [Figure 2.3] [Supplemental Figure 2.1] (Warwick & Saunder 2005; Warwick & Hall 2009, Arias et al. 2012). This is seen with the placement of the *Zilla* & *Vella* clades of the tribe as the groups sister to the rest of the tribe and in the breakdown of the *Rapa/Oleracea* & *Nigra* clades [Figure 2.3]. In the nuclear topology, *Conciya logistrostra* of the *Nigra* clade falls within the majority of the *Rapa/Oleracea* clades species with high support. *Erucastrum nasturtifolluium* & *Raphans raphanistrom* of the *Rapa/Oleracea* clade are found within the majority of *Nigra* species though *Erucastrum nasturtifolum* retains low support for its closest relationship. These incongruences are putatively a result of introgression and chloroplast capture within the tribe.

Interestingly, unlike in previous nuclear phylogenies of the tribe, we see a significant shift in the placement of the Crambe clade. Typically, this monophyletic group is found sister to the *Nigra* clade species in the chloroplast tree as well as nuclear phylogenies [Supplemental Figure 2.1] (Warwick & Saunder 2005; Warwick & Hall 2009, Arias et al. 2012).

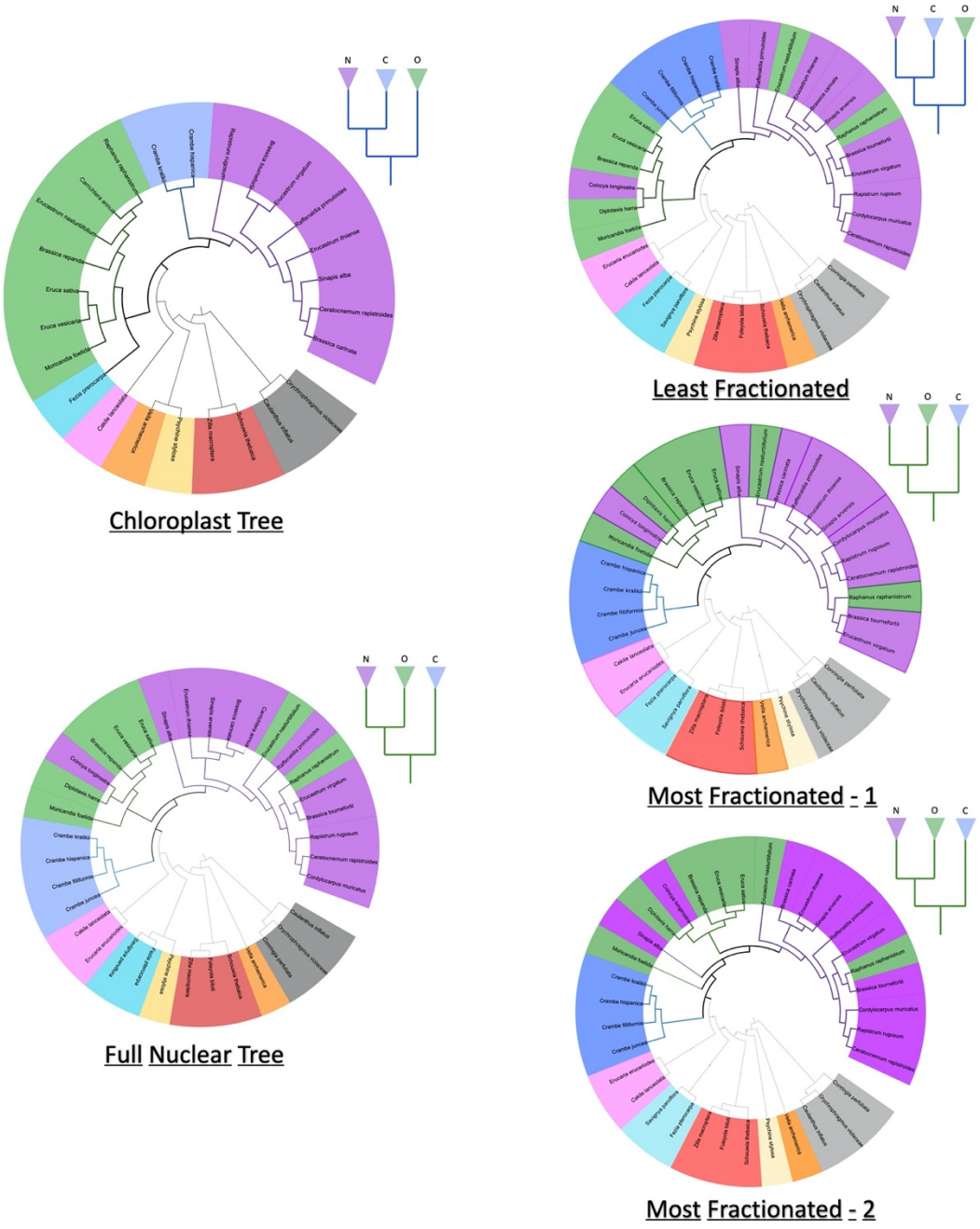


Figure 2.4: Tree topologies of sub-genome, full-nuclear, and chloroplast trees. The least fractionated sub-genome core Brassiceae topology is most similar to that of the chloroplast tree, while that of the MF sub-genomes show a similar topology to what we see in the Full Nuclear tree, of which they represent a subset of genes. Full trees for the sub-genome datasets are found in Supplemental Figures 2.3-5.

Here we find the *Crambe* clade as sister to both the *Nigra & Rapa/Oleracea* clades, a placement that could potentially be leveraged to unite major genera of the tribe based on morphological characters.

To explore this finding further, we leverage the syntenic relationships offered to us by the genome-guided phylotranscriptomics method and gene lists derived from generic comparisons between members of the tribe (Hao et al. 2021). We grouped the homoeologs into three sub-genomes, a least fractionated genome (LF) and two more fractionated genomes (MF1 & MF2). We then isolated these single-copy genes from each grouping and analyzed them separately. We found that the LF genome represented a topology that was most congruent with the chloroplast tree and the expectation of the nuclear tree topology concerning the *Crambe* clade. Both MF1 and MF2 had a *Crambe* placement similar to that of the total nuclear tree [Figure 2.4].

Several factors could be producing these patterns. Firstly, our method for generating the nuclear tree in Figure 2.2 may over-represent genes from the LF sub-genome compared to other phylogenomic that don't require syntenic signal. Biologically we might expect this if the higher percentage of LF sub-genome loci have lost syntenic signal due to rearrangements or local duplications. This bias may also affect what "kinds" of genes are retained from the LF sub-genome instead of the other sub-genomes

The biological reasons as to why these sub-genomes produced alternative topologies are also factors to consider. It may have to do with the relationships between parental genomes before the hexaploidy. The two MF sub-genomes appear to be more closely related, a finding that supports other hypotheses about the genomic history of the

tribe. Extrapolating more than that from these differences may be dubious, without understanding which branch or branches carry the hexaploid event. This pattern of mixed gene-tree topologies mirrors hypotheses around the hybrid origin of individual species (Folk et al. 2018) and, if applied here, would suggest that the *Nigra* clade represents the most "hybrid" Brassiceae clade. That hypothesis in the context of this dataset could be taken to mean that the genes sampled of the Nigra clade most represent an intermediate sub-genome distribution between the *Crambe* and *Rapa/Oleracea* clades. That said, we would expect this pattern to be influenced heavily by species sampling.

CONCLUSION

To fully answer the questions uncovered by this research questions, increased sampling and further exploration of the dataset will be crucial. Sampling of the outer Brassiceae clades, as well as a sampling of *Henophyton*, will give us a more specific understanding of how hybridization is affecting inconsistent clade relationships between *Zilla, Vella, Savignya* and *Psychines* clades. Complete sampling of the core Brassiceae will allow us to better account for sampling biases in our assumptions. Further examination of the duplicated and triplicated gene sets to see how they differ from single-copy genes will provide insight into how our selection of certain gene groups biases our tree inferences.

REFERENCES

Aguileta, G., S. Marthey, H. Chiapello, M-H Lebrun, F. Rodolphe, E. Fournier, A. Gendrault-Jacquemard, and T. Giraud. 2008. "Assessing the Performance of Single-Copy Genes for Recovering Robust Phylogenies." *Systematic Biology* 57 (4): 613–27. https://doi.org/10.1080/10635150802306527.

Al-Shehbaz, Ihsan A. 2012. "A Generic and Tribal Synopsis of the Brassicaceae (Cruciferae)." *Taxon* 61 (5): 931–54. https://doi.org/10.1002/tax.615002.

Alvarez, I., and J. F. Wendel. 2003. "Ribosomal ITS Sequences and Plant Phylogenetic Inference." *Molecular Phylogenetics and Evolution* 29 (3): 417–34. https://doi.org/10.1016/s1055-7903(03)00208-2.

Andrews, Simon, and Others. 2010. "FastQC: A Quality Control Tool for High Throughput Sequence Data." Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Ané, Cécile, Bret Larget, David A. Baum, Stacey D. Smith, and Antonis Rokas. 2007. "Bayesian Estimation of Concordance among Gene Trees." *Molecular Biology and Evolution* 24 (2): 412–26. https://doi.org/10.1093/molbev/msl170.

Arias, Tatiana, and J. Chris Pires. 2012. "A Fully Resolved Chloroplast Phylogeny of the Brassica Crops and Wild Relatives (Brassicaceae: Brassiceae): Novel Clades and Potential Taxonomic Implications." *Taxon* 61 (5): 980–88. https://doi.org/10.1002/tax.615005.

Bastide, Paul, Mahendra Mariadassou, and Stéphane Robin. 2017. "Detection of Adaptive Shifts on Phylogenies by Using Shifted Stochastic Processes on a Tree."

Journal of the Royal Statistical Society: Series B (Statistical Methodology). https://doi.org/10.1111/rssb.12206.

Beilstein, Mark A., Ihsan A. Al-Shehbaz, Sarah Mathews, and Elizabeth A. Kellogg. 2008. "Brassicaceae Phylogeny Inferred from Phytochrome A and ndhF Sequence Data: Tribes and Trichomes Revisited." *American Journal of Botany* 95 (10): 1307–27. https://doi.org/10.3732/ajb.0800065.

Birchler, James A., and Reiner A. Veitia. 2012. "Gene Balance Hypothesis: Connecting Issues of Dosage Sensitivity across Biological Disciplines." *Proceedings of the National Academy of Sciences of the United States of America* 109 (37): 14746–53. https://doi.org/10.1073/pnas.1207726109.

Bird, Kevin A., Chad E. Niederhuth, Shujun Ou, Malia Gehan, J. Chris Pires, Zhiyong Xiong, Robert VanBuren, and Patrick P. Edger. 2021. "Replaying the Evolutionary Tape to Investigate Subgenome Dominance in Allopolyploid Brassica Napus." *The New Phytologist* 230 (1): 354–71. https://doi.org/10.1111/nph.17137.

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. https://doi.org/10.1111/nph.15256.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST : Architecture and Applications." *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-10-421.

Conant, Gavin C., James A. Birchler, and J. Chris Pires. 2014. "Dosage, Duplication, and Diploidization: Clarifying the Interplay of Multiple Models for Duplicate Gene Evolution over Time." *Current Opinion in Plant Biology* 19 (June): 91–98. https://doi.org/10.1016/j.pbi.2014.05.008.

Daniels, Jan-Peter, Keith Gull, and Bill Wickstead. 2010. "Cell Biology of the Trypanosome Genome." *Microbiology and Molecular Biology Reviews: MMBR* 74 (4): 552–69. https://doi.org/10.1128/MMBR.00024-10.

Dolstra, O. 1982. "Synthesis and Fertility of Brassicoraphanus and Ways of Transferring Raphanus Characters to Brassica." https://library.wur.nl/WebQuery/wurpubs/75936.

Fahey, J. W., A. T. Zalcmann, and P. Talalay. 2001. "The Chemical Diversity and Distribution of Glucosinolates and Isothiocyanates among Plants." *Phytochemistry* 56 (1): 5–51. https://doi.org/10.1016/s0031-9422(00)00316-2.

Folk, Ryan A., Pamela S. Soltis, Douglas E. Soltis, and Robert Guralnick. 2018. "New Prospects in the Detection and Comparative Analysis of Hybridization in the Tree of Life." *American Journal of Botany* 105 (3): 364–75. https://doi.org/10.1002/ajb2.1018.

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. https://doi.org/10.1038/nbt.1883.

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. https://doi.org/10.1038/nprot.2013.084.

Hall, Jocelyn C., Tracy E. Tisdale, Kathleen Donohue, Andrew Wheeler, Mohammed A. Al-Yahya, and Elena M. Kramer. 2011. "Convergent Evolution of a Complex Fruit Structure in the Tribe Brassiceae (Brassicaceae)." *American Journal of Botany* 98 (12): 1989–2003. https://doi.org/10.3732/ajb.1100203.

Hao, Yue, Makenzie E. Mabry, Patrick P. Edger, Michael Freeling, Chunfang Zheng, Lingling Jin, Robert VanBuren, et al. 2020. "The Contributions of the Allopolyploid Parents of the Mesopolyploid Brassiceae Are Evolutionarily Distinct but Functionally Compatible." *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2020.08.10.245258.

Hénocq, Laura, Sophie Gallina, Eric Schmitt, Vincent Castric, Xavier Vekemans, and Céline Poux. 2020. "A New Tree-Based Methodological Framework to Infer the Evolutionary History of Mesopolyploid Lineages: An Application to the Brassiceae Tribe (Brassicaceae)." *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2020.01.09.900571.

Jones, Graham, Serik Sagitov, and Bengt Oxelman. 2013. "Statistical Inference of Allopolyploid Species Networks in the Presence of Incomplete Lineage Sorting." *Systematic Biology* 62 (3): 467–78. https://doi.org/10.1093/sysbio/syt012.

Jun, Jin, Ion I. Mandoiu, and Craig E. Nelson. 2009. "Identification of Mammalian Orthologs Using Local Synteny." *BMC Genomics* 10 (December): 630. https://doi.org/10.1186/1471-2164-10-630.

Katoh, Kazutaka, Kazuharu Misawa, Kei-Ichi Kuma, and Takashi Miyata. 2002. "MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform." *Nucleic Acids Research* 30 (14): 3059–66. https://doi.org/10.1093/nar/gkf436.

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. https://doi.org/10.1093/molbev/mst010.

Li, Z., H. L. Liu, and P. Luo. 1995. "Production and Cytogenetics of Intergeneric Hybrids between Brassica Napus and Orychophragmus Violaceus." *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik* 91 (1): 131–36. https://doi.org/10.1007/BF00220869.

Li, Z., J. G. Wu, Y. Liu, H. L. Liu, and W. K. Heneen. 1998. "Production and Cytogenetics of the Intergeneric Hybrids Brassica juncea×Orychophragmus Violaceus and B. carinata×O. Violaceus." *Theoretical and Applied Genetics* 96 (2): 251–65. https://doi.org/10.1007/s001220050734.

Linder, C. Randal, C. Randal Linder, and Loren H. Rieseberg. 2004. "Reconstructing Patterns of Reticulate Evolution in Plants." *American Journal of Botany*. https://doi.org/10.3732/ajb.91.10.1700.

Lyons, Eric, Brent Pedersen, Josh Kane, and Michael Freeling. 2008. "The Value of Nonmodel Genomes and an Example Using SynMap within CoGe to Dissect the Hexaploidy That Predates the Rosids." *Tropical Plant Biology* 1 (3-4): 181–90. https://doi.org/10.1007/s12042-008-9017-y.

Lysak, M. A., and C. Lexer. 2006. "Towards the Era of Comparative Evolutionary Genomics in Brassicaceae." *Osterreichische Botanische Zeitschrift* 259 (2-4): 175–98. https://doi.org/10.1007/s00606-006-0418-9.

Lysak, Martin A., Kwok Cheung, Michaela Kitschke, and Petr Bures. 2007. "Ancestral Chromosomal Blocks Are Triplicated in Brassiceae Species with Varying Chromosome Number and Genome Size." *Plant Physiology* 145 (2): 402–10. https://doi.org/10.1104/pp.107.104380.

Lysak, Martin A., Marcus A. Koch, Ales Pecinka, and Ingo Schubert. 2005. "Chromosome Triplication Found across the Tribe Brassiceae." *Genome Research* 15 (4): 516–25. https://doi.org/10.1101/gr.3531105.

Mayfield-Jones, Dustin, Jacob D. Washburn, Tatiana Arias, Patrick P. Edger, J. Chris Pires, and Gavin C. Conant. 2013. "Watching the Grin Fade: Tracing the Effects of Polyploidy on Different Evolutionary Time Scales." *Seminars in Cell & Developmental Biology* 24 (4): 320–31. https://doi.org/10.1016/j.semcdb.2013.02.002.

Mayrose, Itay, and Martin A. Lysak. 2021. "The Evolution of Chromosome Numbers: Mechanistic Models and Experimental Approaches." *Genome Biology and Evolution* 13 (2). https://doi.org/10.1093/gbe/evaa220.

Mendes, Fábio K., and Matthew W. Hahn. 2016. "Gene Tree Discordance Causes Apparent Substitution Rate Variation." *Systematic Biology* 65 (4): 711–21. https://doi.org/10.1093/sysbio/syw018.

Mirarab, S., R. Reaz, Md S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. 2014. "ASTRAL: Genome-Scale Coalescent-Based Species Tree Estimation." *Bioinformatics* 30 (17): i541–48. https://doi.org/10.1093/bioinformatics/btu462.

Mirarab, Siavash, Md Shamsuzzoha Bayzid, and Tandy Warnow. 2016. "Evaluating Summary Methods for Multilocus Species Tree Estimation in the Presence of Incomplete Lineage Sorting." *Systematic Biology* 65 (3): 366–80. https://doi.org/10.1093/sysbio/syu063.

Mizushima, Usaburo, and Others. 1950. "Karyogenetic Studies of Species and Genus Hybrids in the Tribe Brassiceae of Cruciferae." *Tohoku Journal of Agricultural Research* 1: 1–14. https://www.cabdirect.org/cabdirect/abstract/19511601931.

Nagaharu, U. 1935. "Genome Analysis in Brassica with Special Reference to the Experimental Formation of B. Napus and Peculiar Mode of Fertilization." *Journal of Japanese Botany* 7 (7): 389–452.

Niet, Timotheüs van der, and H. Peter Linder. 2008. "Dealing with Incongruence in the Quest for the Species Tree: A Case Study from the Orchid Genus Satyrium." *Molecular Phylogenetics and Evolution* 47 (1): 154–74. https://doi.org/10.1016/j.ympev.2007.12.008.

Rokas, Antonis, Barry L. Williams, Nicole King, and Sean B. Carroll. 2003. "Genome-Scale Approaches to Resolving Incongruence in Molecular Phylogenies." *Nature* 425 (6960): 798–804. https://doi.org/10.1038/nature02053.

Schmieder, Robert, and Robert Edwards. 2011. "Quality Control and Preprocessing of Metagenomic Datasets." *Bioinformatics* 27 (6): 863–64. https://doi.org/10.1093/bioinformatics/btr026.

Schranz, M. Eric, Martin A. Lysak, and Thomas Mitchell-Olds. 2006. "The ABC's of Comparative Genomics in the Brassicaceae: Building Blocks of Crucifer Genomes." *Trends in Plant Science* 11 (11): 535–42. https://doi.org/10.1016/j.tplants.2006.09.002.

Schulz, and O. E. 1919. "Part I : Brassicinae and Raphaninae." *Cruciferae-Brassiceae*, 194–210. https://ci.nii.ac.jp/naid/10025025895/.

Schulz, Otto Eugen. 1957. *Cruciferae-Brassiceae*. Vol. 70. HR Engelmann.

Soderlund, Carol, Matthew Bomhoff, and William M. Nelson. 2011. "SyMAP v3.4: A Turnkey Synteny System with Application to Plant Genomes." *Nucleic Acids Research* 39 (10): e68. https://doi.org/10.1093/nar/gkr123.

Solís-Lemus, Claudia, Paul Bastide, and Cécile Ané. 2017. "PhyloNetworks: A Package for Phylogenetic Networks." *Molecular Biology and Evolution* 34 (12): 3292–98. https://doi.org/10.1093/molbev/msx235.

Song, K., T. C. Osborn, and P. H. Williams. 1990. "Brassica Taxonomy Based on Nuclear Restriction Fragment Length Polymorphisms (RFLPs)." *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik* 79 (4): 497–506. https://doi.org/10.1007/BF00226159.

Stamatakis, Alexandros. 2014. "RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies." *Bioinformatics* 30 (9): 1312–13. https://doi.org/10.1093/bioinformatics/btu033.

Szöllősi, Gergely J., Adrián Arellano Davín, Eric Tannier, Vincent Daubin, and Bastien Boussau. 2015. "Genome-Scale Phylogenetic Analysis Finds Extensive Gene Transfer among Fungi." *Philosophical Transactions of the Royal Society of*

London. Series B, Biological Sciences 370 (1678): 20140335. https://doi.org/10.1098/rstb.2014.0335.

Tang, Haibao, Eric Lyons, Brent Pedersen, James C. Schnable, Andrew H. Paterson, and Michael Freeling. 2011. "Screening Synteny Blocks in Pairwise Genome Comparisons through Integer Programming." *BMC Bioinformatics* 12 (April): 102. https://doi.org/10.1186/1471-2105-12-102.

Warwick, Suzanne I., and Jocelyn C. Hall. 2009. "Phylogeny of Brassica and Wild Relatives." *Biology and Breeding of Crucifers* 19: 36.

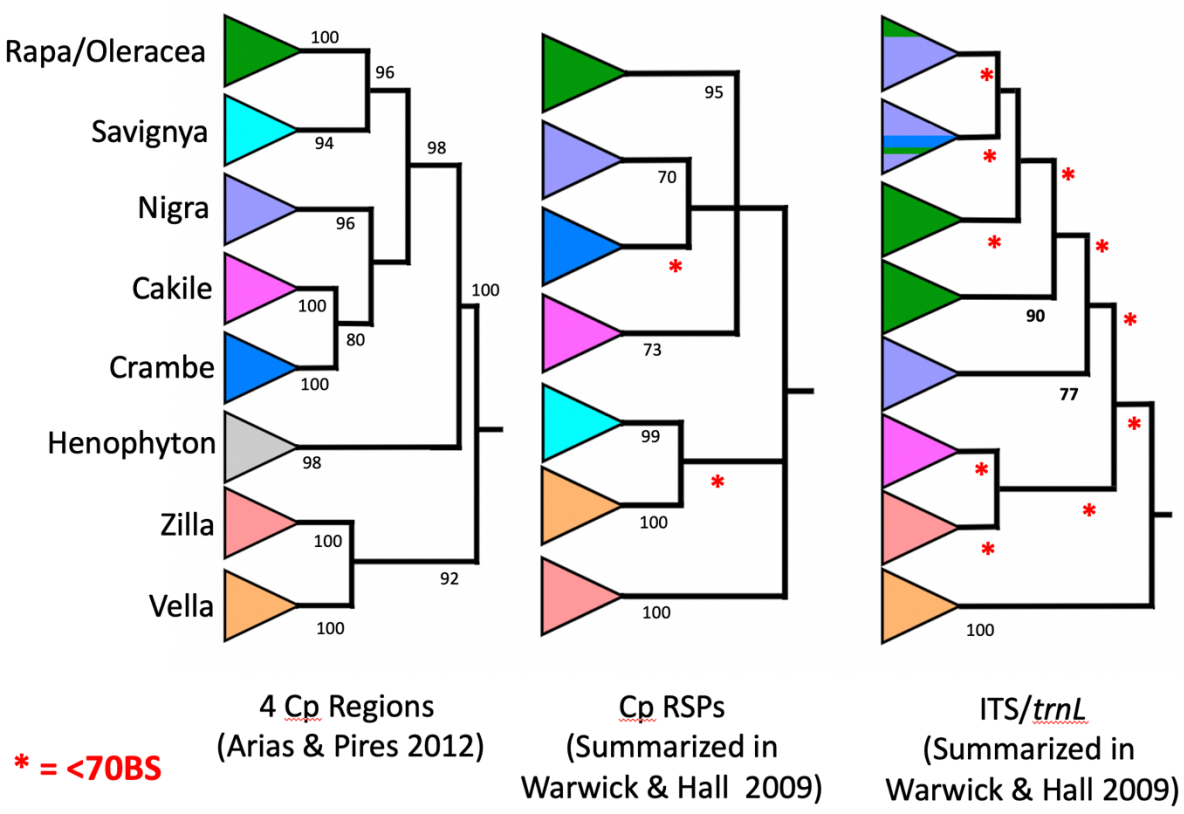
Warwick, Suzanne I., and Connie A. Sauder. 2005. "Phylogeny of Tribe Brassiceae (Brassicaceae) Based on Chloroplast Restriction Site Polymorphisms and Nuclear Ribosomal Internal Transcribed Spacer and Chloroplast trnL Intron Sequences." *Canadian Journal of Botany*. https://doi.org/10.1139/b05-021.

Washburn, Jacob D., James C. Schnable, Gavin C. Conant, Thomas P. Brutnell, Ying Shao, Yang Zhang, Martha Ludwig, Gerrit Davidse, and J. Chris Pires. 2017. "Genome-Guided Phylo-Transcriptomic Methods and the Nuclear Phylogentic Tree of the Paniceae Grasses." *Scientific Reports* 7 (1): 13528. https://doi.org/10.1038/s41598-017-13236-z.

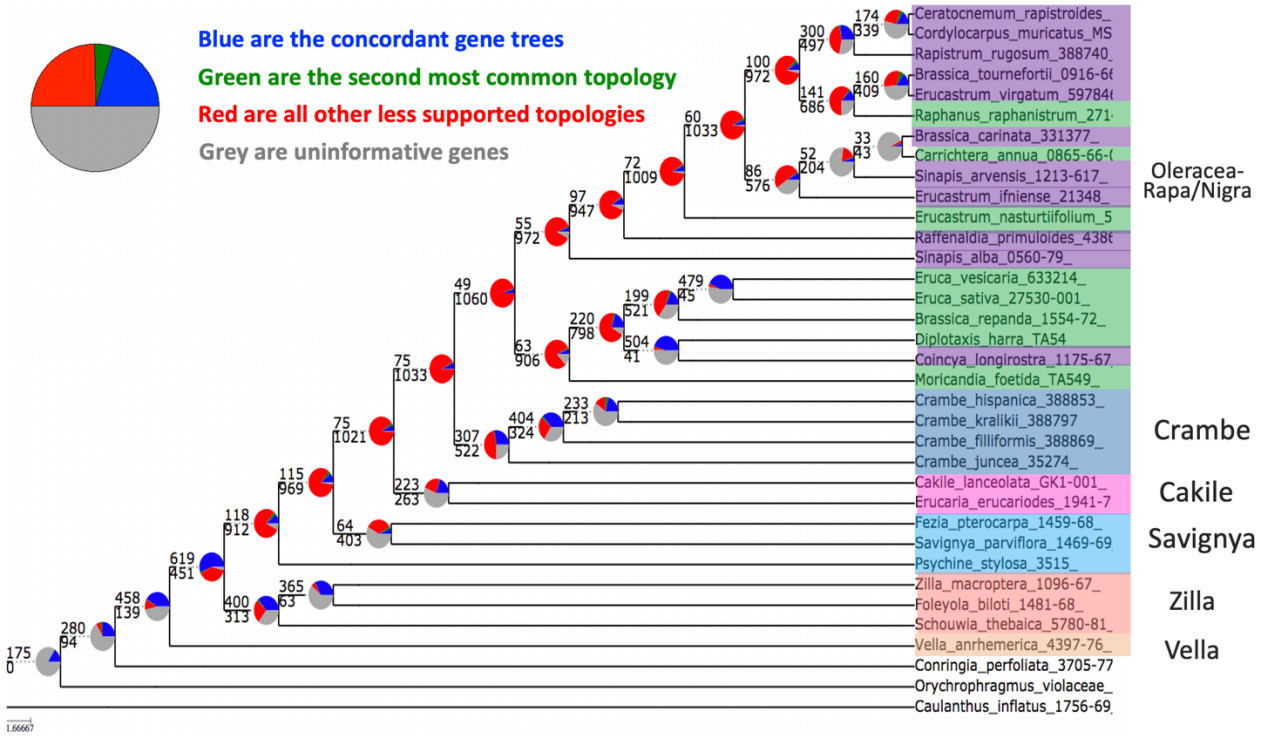
Wendel, Jonathan F., and Jeff J. Doyle. 1998. "Phylogenetic Incongruence: Window into Genome History and Molecular Evolution." In *Molecular Systematics of Plants II: DNA Sequencing*, edited by Douglas E. Soltis, Pamela S. Soltis, and Jeff J. Doyle, 265–96. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4615-5419-6_10.

Yanagino, Toshiya, Yoshihito Takahata, and Kokichi Hinata. 1987. “Chloroplast DNA Variation among Diploid Species in Brassica and Allied Genera.” 遺伝學雜誌 62 (2): 119–25. https://doi.org/10.1266/jjg.62.119.

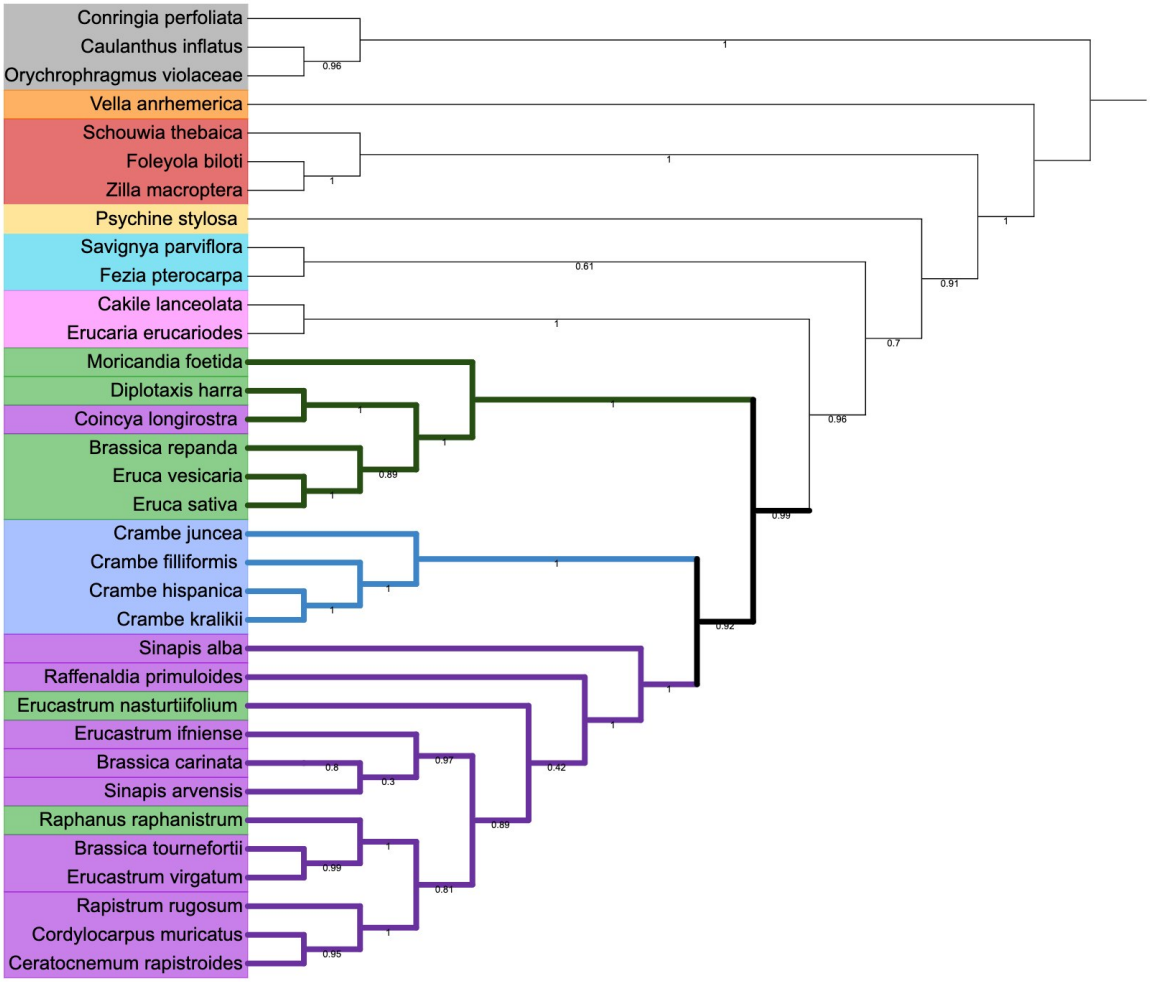
Yang, Ya, and Stephen A. Smith. 2014. “Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics.” *Molecular Biology and Evolution* 31 (11): 3081–92. https://doi.org/10.1093/molbev/msu245.



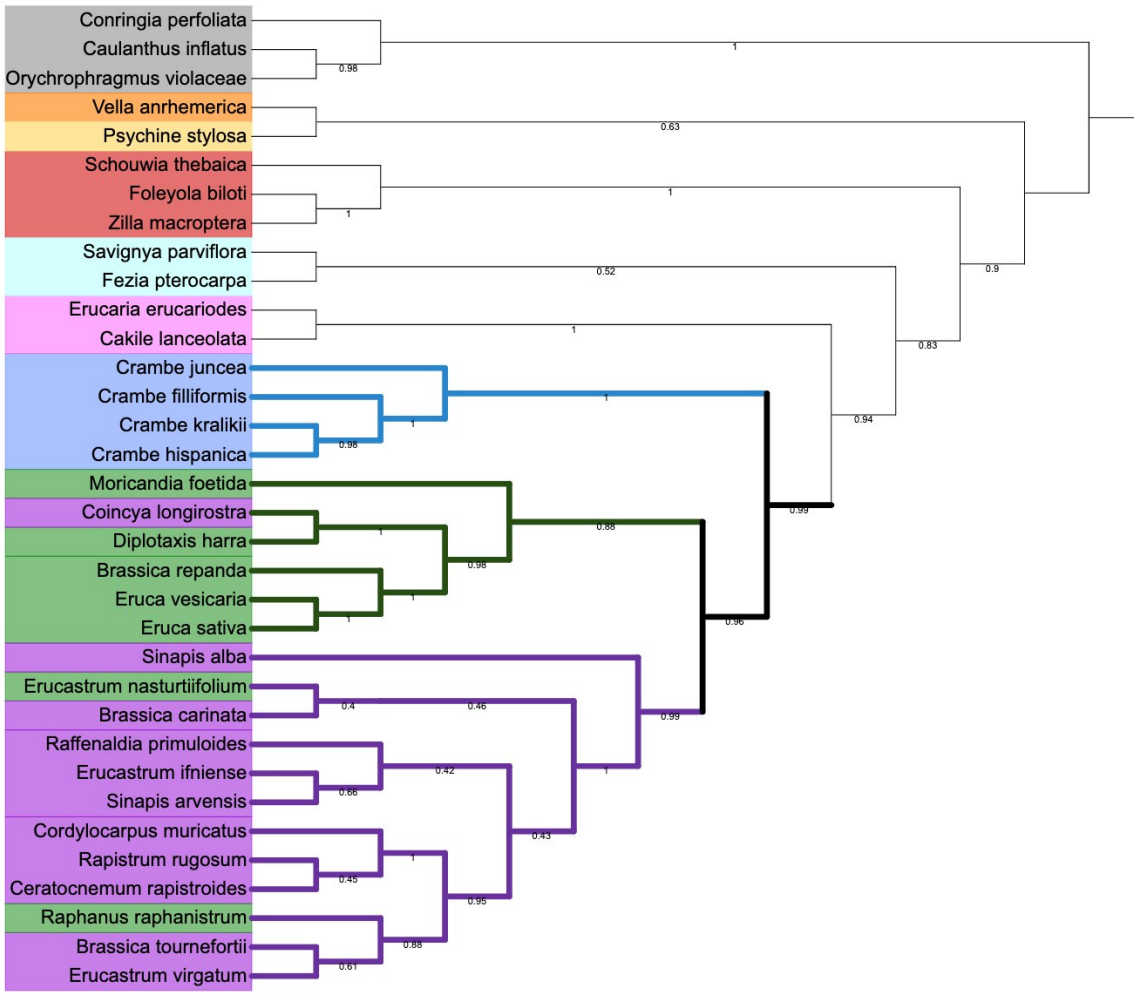
Supplementary Figure 2.1. Summary of chloroplast and ITS trees of the tribe. The chloroplast trees show similar clade groupings and recover some similar relationships but not all, with many displaying weak support. The ITS trees of the time support a topology in which the *Crambe* clade is nested within the *Nigra* and *Rapa/Oleracea* group.



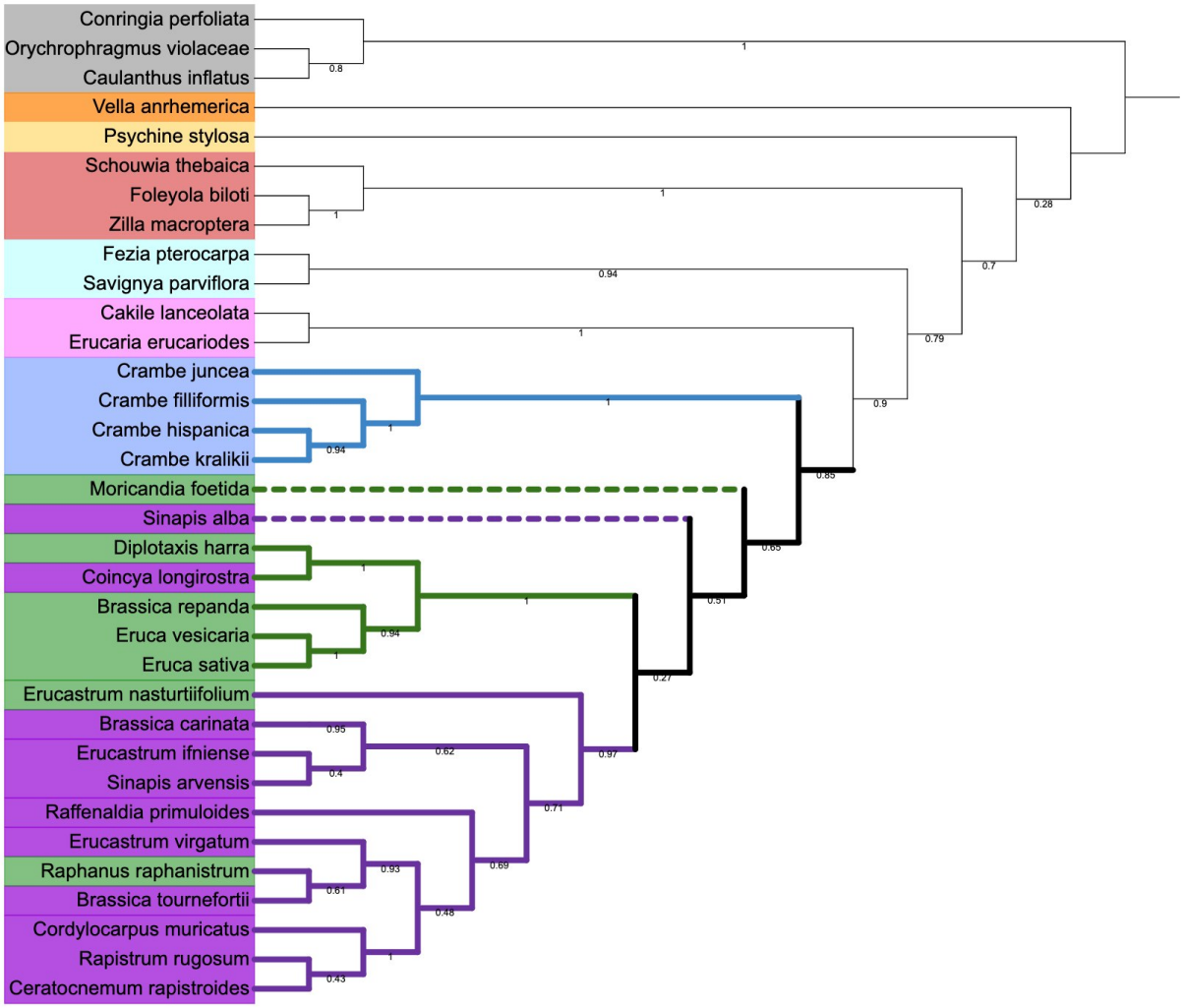
Supplementary Figure 2.2. Gene tree incongruence metrics for the nuclear tree from Figure 1.2. The above number at each node represents the number of gene trees that agree with the concordant topology and the below for each node represents the total of gene trees that support all other informed topologies.



Supplementary Figure 2.3. The inferred nuclear tree from the Least Fractionated Sub-genome. Generated via the same methods as the full gene set nuclear tree.



Supplementary Figure 2.4. The inferred nuclear tree from the Most Fractionated Sub-genome 1. Generated via the same methods as the full gene set nuclear tree.



Supplementary Figure 2.5. The inferred nuclear tree from the Most Fractionated Sub-genome 1. Generated via the same methods as the full gene set nuclear tree.

CHAPTER 3: GENOMIC ORIGIN AND DIVERSIFICATION OF THE GLUCOSINOLATE MAM LOCUS

R. Shawn Abrahams[1,2], J. Chris Pires[1] and M. Eric Schranz[2]

1 Division of Biological Sciences and Bond Life Sciences Center, University of Missouri, Columbia, Missouri 65211, USA

2 Biosystematics Group, Plant Sciences, Wageningen University, Wageningen 6700 AA, Netherlands

Please cite the published work here:

Abrahams, R. Shawn, J. Chris Pires[1], and M. Eric Schranz[2]. 2020. "Genomic Origin and Diversification of the Glucosinolate MAM Locus." *Frontiers in Plant Science* 11 (June): 711.

ABSTRACT

Glucosinolates are a diverse group of plant metabolites that characterize the order Brassicales. The *MAM* locus is one of the most significant QTLs for glucosinolate diversity. However, most of what we understand about evolution at the locus is focused on only a few species and not within a phylogenetic context. In this study, we utilize a micro-synteny network and phylogenetic inference to investigate the origin and diversification of the *MAM*/IPMS gene family. We uncover unique *MAM*-like genes found at the orthologous locus in the Cleomaceae that shed light on the transition from *IPMS* to *MAM*. In the Brassicaceae, we identify six distinct *MAM* clades across Lineages I, II, and III. We characterize the evolutionary impact and consequences of local duplications, transpositions, whole genome duplications, and gene fusion events, generating several new hypothesizes on the function and diversity of the *MAM* locus.

INTRODUCTION

Glucosinolates (GSL) are a diverse class of amino-acid derived sulphur containing metabolites characteristic of plants of the order Brassicales (Rodman et al., 1998; Borpatragohain et al., 2016; Kliebenstein and Cacho, 2016; Olsen et al., 2016; Chhajed et al., 2019; Blazevic et al., 2020;). When the plant experiences physical damage, such as chewing by herbivores, compartments of the cell rupture and release myrosinase enzymes that hydrolyze the GSLs to create an isothiocyanate anion, damaging the attacker (Rodman et al., 1998). Besides their roles in direct defense, GSLs have also been shown to play important roles such as nutrient transport and physiological signaling (del Carmen et al., 2013). They are considered a key innovation of the Brassicales, as adaptations in the biosynthesis pathway have been shown to correlate with increased rates of speciation (Edger et al., 2015). The GSL pathway is a model for investigating processes underlying natural variation within and among species; including the roles of genome and gene duplication (Kliebenstein, 2008; Bekaert et al., 2012; Hofberger et al., 2013; Edger et al., 2015; Van den Bergh et al., 2016; Wisecaver et al., 2017.) Aliphatic GSLs, the largest sub-group of compounds, are especially implicated in this rate of speciation as they are only found in the most species-rich groups such as the family Brassicaceae.

The often multi-gene *methylthioalkylmalate* (*MAM*) locus, also called the Elong locus, accounts for much of the natural variation observed in aliphatic GSLs (Kliebenstein et al., 2001b, c; Textor et al., 2004, 2007; Keurentjes et al., 2006; de Kraker et al., 2007; Wentzell et al., 2008; Benderoth et al., 2006, 2008, 2009; de Kraker and

Gershenzon, 2011; Kliebenstein and Cacho, 2016; Kumar et al., 2019; Petersen et al., 2019; Zhang et al. 2015.). *MAM* enzymes catalyze the condensation reaction that extends the carbon chain in amino acid derived GSL precursors (Benderoth et al. 2006). The extended amino acid expands the types (Kliebenstein & Cacho, 2016). Most of what we understand about the evolution of *MAM* has been learned from studying just a handful of species, without a broad phylogenetic context (Kleibenstien & Cacho, 2016). MAM diversification in the Brassicaceae is thought to have occurred independently in separate lineages. Specifically, *MAM* diversity has been largely examined in Lineage I of the family (*Arabidopsis* and relatives) and to a lesser extent in Lineage II (*Brassica* and relatives). This work has been supported by large gene datasets, though with differing gene tree topologies (Zhang et al. 2015, Supplemental Figure 3.1).

In *Arabidopsis thaliana*, phenotypic variation of the *MAM* locus is characterized by the accumulation of different majority carbon chain-length GSL profiles (Kleibenstein & Cacho 2016). The most common profiles have majority three carbon (3C) or four carbon (4C) molecules, but can extend up to 8C majority profiles, with variability at the population level (Benderoth et al 2009; Kleibenstein & Cacho 2016). Copy number variation and allelic diversity/presence-absence drive these differences, as one *MAM* gene may mask the phenotype of another at the same locus (Benderoth et al. 2008, 2009). This plays out in the interactions between *MAM1* and *MAM2* in *A. thaliana* populations, where variation is well understood. The 4C majority phenotype is seen in populations where *MAM1* and *MAM2* are both present and intact or when *MAM2* is absent. In populations lacking a *MAM1* gene, the GSL profile exhibits a 3C majority phenotype. In some cases, *MAM1* and *MAM2* genes have been fused (e.g. gene chimerism) wherein they are

reformed into a *MAM1-like* functional gene with partial *MAM2* sequences, or vice versa (Benderoth et al. 2008). Crop Brassicas most commonly accumulate 3C, 4C, or a mix of 3C and 4C majority profiles, the latter displaying a seemingly unmasked phenotype, unlike what we see in *A. thaliana* (Benderoth et al 2009; Klibenstein & Cacho 2016).

Naming conventions for *MAM* orthologs are either directly based on *A. thaliana* (*MAM1*, *MAM2*, and *MAM3*) or based on *A. lyrata MAM* (*MAMa, MAMb*, and *MAMc*) (Benderoth et al., 2009). The *Arabidopsis* centered model of *MAM* diversity is vulnerable to miss-characterization as *Arabidopsis* genes may be highly derived, and thus not generalizable. We also see that the number of genes at the *MAM* locus can vary between populations as well as species, potentially misleading ancestral state estimations with poor sampling. To accurately understand *MAM* diversification, it is necessary for gene selection across a broader species phylogeny with comparisons to their primary metabolic ancestor, isopropylmalate synthase (*IPMS*).

Though diverged, *IPMS* and *MAM* share a high sequence similarity and similar enzymatic function (Moghe & Last 2015). I*PMS* contains two conserved protein domains: a pyruvate carboxylase (HMGL-like), that is involved in the carbon condensation reaction, and a leucine allosteric domain (LeuA), that commits the protein to the leucine biosynthesis pathway forming a homodimer (Koon et al., 2004). *MAM* genes only retain the HMGL-like domain, the loss of LeuA being considered a key step in the transition of *MAM* from an *IPMS-like* gene (de Kraker et al., 2007). To our knowledge, no previous work has investigated when the loss of this domain occurred in the evolution of the locus.

In this study, we examine the evolutionary history and diversity of the *MAM/IPMS* gene family, uncovering critical steps in the origin of *MAM* and identifying patterns of domain-specific diversity across the Brassicaceae and its sister-family the Cleomaceae. We utilize a genomic networking methodology to analyze the wealth of newly available genome sequences (Zhao et al. 2017; Zhao and Schranz 2019). The method analyzes the conserved physical location of gene family members across queried genomes, known as synteny, to characterize the impact of different gene duplication types in the expansion of the *MAM/IPMS* gene family (Zhao et al. 2017; Zhao and Schranz 2019). Ultimately we show that a mix of gene duplication types and domain changes played important roles in the evolution and innovation of the *MAM* locus.

METHODS

Genomic Network Construction

The genomic network analysis included 40 complete plant genomes representing 38 different species. This included 34 Brassicaceae species from Lineages I, II, III and *Aethionema arabicum* as sister to the rest of the family, three genomes from the sister-family Cleomaceae, and three outgroup species (*Theobroma cacao, Citrus sinensis,* and *Vitis vinifera*) [Supplemental Table 1]. For each genome, we utilized protein sequences in FASTA format and a BED/GFF file. One of two *Capsella rubella* genomes was excluded from downstream analysis due to insufficient quality. The *Thellungiella halophila* and *Thellungiella salsuginea* are two different sequencing efforts of the same species, now under the name *Eutrema salsugineum*. The genome sequenced as *Alyssum linifolium* has since been identified as *Descurainia pinnata*. Network analyses were performed as described in Zhao et al. 2017. Reciprocal all-against-all whole genome protein sequence

comparison were made using RAPSeach2 (Zhao et al., 2012). MCScanX (Tange et al. 2008; Wang et al 2012) was used to calculate generic collinearity between genomes and all comparisons were saved to generate the full genomic network.

Gene Family Network

We identified candidate *IPMS/MAM* genes using HMMER (Fin et al., 2011), cross-referencing the Pfam, PDBe, and GO databases with domain signature HMGL-like PF00682, and filtered by an inclusion threshold e-value of 0.007. Selected genes were later filtered by relative branch lengths as compared to known *IPMS* and *MAM* genes and then queried against the overall syntenic network with a 25 gene window to extract the gene family network. We visualized the resulting network in Cytoscape version 3.3.0 (Shannon et al. 2003). We then pruned the network of gene nodes that did not contain an HMGL-like domain but were dragged in by potential domain fusions. Clique percolation, as implemented in CFindier (Derenyi et al. 2005; Palla et al. 2005; Fortueto 2010), was used to locate all K-clique comments to identify communities or clusters of gene nodes.

Phylogenetic inference

Full amino acid sequences for all gene family members were aligned using MAFFT (Kuraku et al., 2013; Katoh et al. 2017) and cleaned using Phyutility at a 50% occupancy threshold (Smith & Dunn, 2008). We used RAxML (Stamatkis, 2014) for phylogenetic inference with the GTRCAT model (Boostrap = 1000). The same procedure was repeated for the HMGL-like domain region of each gene FASTA file as estimated by HMMER. Supplemental sequence comparisons were made using MView (Maderia et al., 2019) and analyzed using R.

RESULTS

Synteny and Domain analysis

Micro-synteny network analysis identified three major syntenic clusters [Figure 3.1], two of which encompass many genes of the known *MAM* gene clade (orange and green clusters) and one encompassing the known *IPMS* gene clade (blue cluster). Of the syntenic clusters found in the *MAM* clade, the green cluster identifies the ancestral *MAM* position, what we will call the *MAM*-Ancestral locus, and is equivalent to the Elong locus. The orange cluster represents a transposed and retained *MAM* locus-specific to Lineage II of the Brassicaceae, which we will

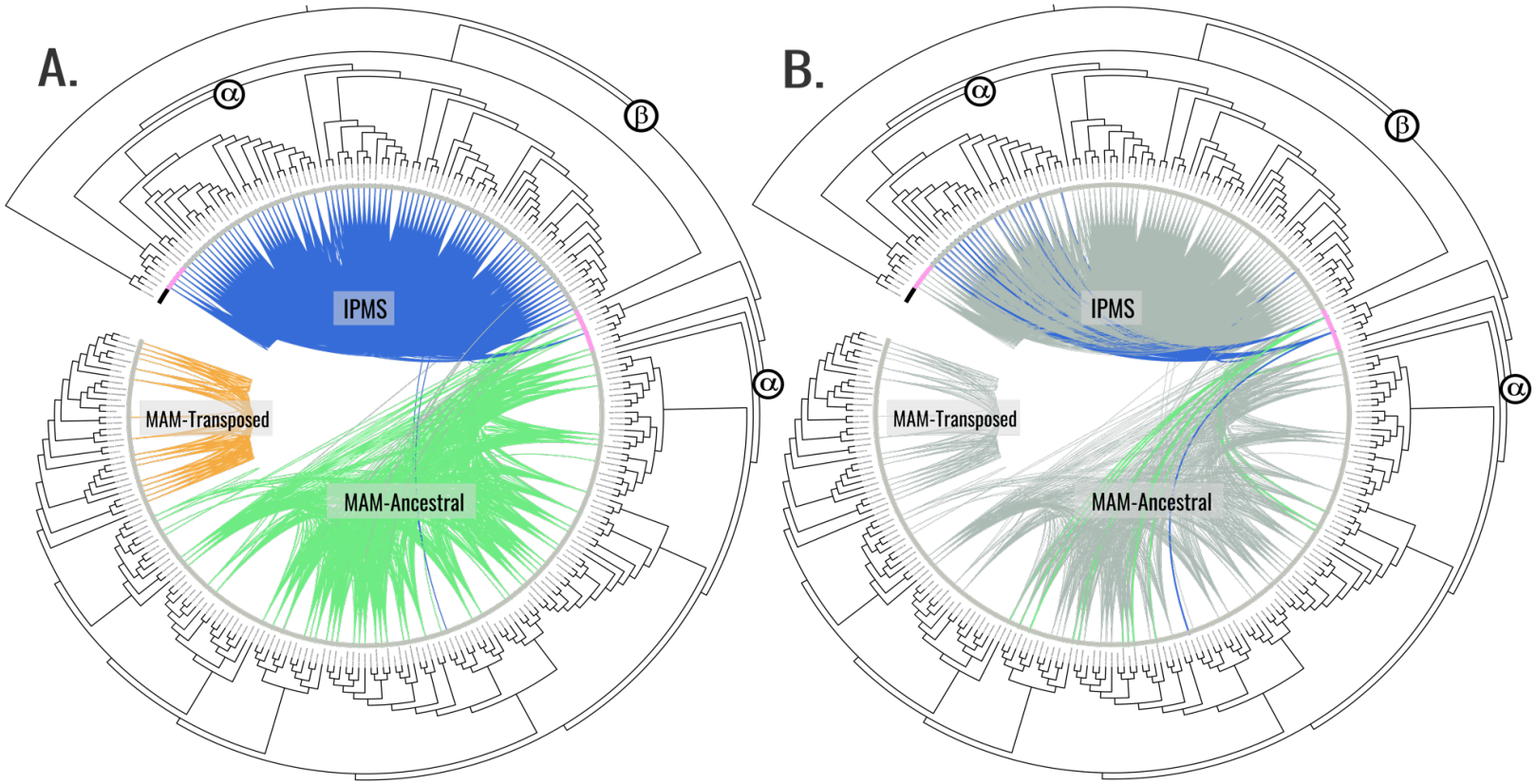


Figure 3.1: Synteny Clusters and Gene Tree Phylogeny of identified *IPMS* and *MAM* genes consisting of 262 total. For (A) and (B) the bar along the tips represent species lineage where black bars indicate genes from out group genomes, the pink bar indicates genes from Cleomaceae genomes, and the grey bar indicates genes from Brassicaceae genomes. (A) Syntenic cluster analysis identified three distinct gene clusters, each representing a different conserved genomic location. The IPMS cluster in blue, the MAM-Ancestral cluster in green, and the novel lineage specific MAM-Transposed cluster in orange. Grey lines here indicate connections between the IPMS and MAM clusters. (B) Emphasizes those connections between MAM-like genes in the Cleomaceae that exhibit both *IPMS* & *MAM* cluster membership (Clevi.0004s0713 and tha_Th2v24105) despite being physically located at the MAM-Ancestral locus in their respective genomes. [For Bootstrap scores: Supplemental Figure 3.6; Online interactive trees: (A) - http://bit.ly/2tHVgYK; (B) - http://bit.ly/2Svu8Vf]

call the *MAM*-Transposed locus. The analysis also recovered the 4th cluster of an unnamed lineage of genes that have retained only a single HMGL-like domain and are found in both our outgroup and in-group genomes. The *A. thaliana* representative gene of this clade (AT2G26800) has been shown to play a role in seed amino acid concentration (Peng et al., 2015). Relative branch lengths showed this gene clade as highly diverged from both MAM and IPMS sequences. Because of this, all genes of this clade were filtered from downstream analyses.

95.7% of IPMS genes identified by sequence were also found in the IPMS syntenic cluster. 39.6% of MAM genes, not associated with the conserved Lineage II transposition, were found in the MAM-Ancestral syntenic cluster. 51.6% of genes found in the Lineage II transposed sub-clade were found in the syntenic cluster. Differences in percent synteny are tied to increased rates of tandem duplications, as the local duplicate syntenic signal was often masked, and transposed duplication events, which remove syntenic context. It is expected that many new transposed duplicates are in the process of pseudogenization and are not active MAM genes.

All genes at the Cleomaceae *MAM*-Ancestral locus have retained their LeuA domain from their time as *IPMS* duplicates, with some showing syntenic connections to both the *MAM*-Ancestral and *IPMS* syntenic cluster [Figure 3.1]. For example, Th2v2405 from *Tarenaya hassleriana* has more syntenic connections with *IPMS* cluster members than with genes of the *MAM*-Ancestral locus, despite belonging to the direct orthologous chromosomal region of the MAM-Ancestral locus in the Brassicaceae [Figure 3.1; Figure 3.2]. Genes of the Cleomaceae *MAM*-Ancestral locus and the IPMS locus also appear to have a shared pattern of gene dosage. A duplication of the IPMS locus following WGD,

brings the total *IPMS* gene number to two, followed by a compensatory reduction in MAM gene number at the *MAM*-Ancestral locus [Figure 3.2C]. An exception to this is found in the *Tarenaya hassleriana* genome, where a novel transposed *MAM*-like gene has lost the LeuA domain. This allows for three *MAM*-like genes to co-occur with two *IPMS* genes [Supplemental Figure 3.3].

Phylogenetic inference

The HMGL-like domain and full protein sequence gene trees identified distinct IPMS and MAM clades [Figure 3.1]. In both cases, Cleomaceae genes are sister to a larger Brassicaceae clade, and *Aethionema arabicum* is sister to the rest of the Brassicaceae, which agrees with the species tree topology. Within the core Brassicaceae, the domain and full sequence trees display topological incongruence to each other [Figure 3.3] and neither perfectly match the species tree.

The domain tree divides MAM into six supported clades [Figure 3.3]. Though the branching order could not be determined, the supported clades were assigned *MAMa-f*. These domain clade designations are based on the *Arabidopsis lyrata* MAM gene-tree clades. Given the branch length, a measure of sequence divergence, of the genes found at the MAM-Transposed locus [Figure 3.3], the sub-clade of *MAMe* was designated *MAMet*. The closest non-*MAMet* domain sequence to the group was a *MAMe* sequence from the *Lunaria annua* genome.

Summary amino acid comparison at 80% similarity threshold shows *MAMa* is the most conserved domain, *MAMe* is the most variable domain, and *MAMet* and *MAMc* are the most diverged [Supplemental Figure 3.5]. Exon/Intron comparisons of full *MAMet* genes

show the expected number of domains for a functional MAM gene but with differences in exon size. When plotted on the species tree, *MAMa-b* and *MAMe* are ancestral to Lineage I, *MAMa-b* and *MAMd-f* are ancestral to Lineage II, and *MAMb* and *MAMd* are ancestral to Lineage III [Figure 3.4; Supplemental Figure 3.3].

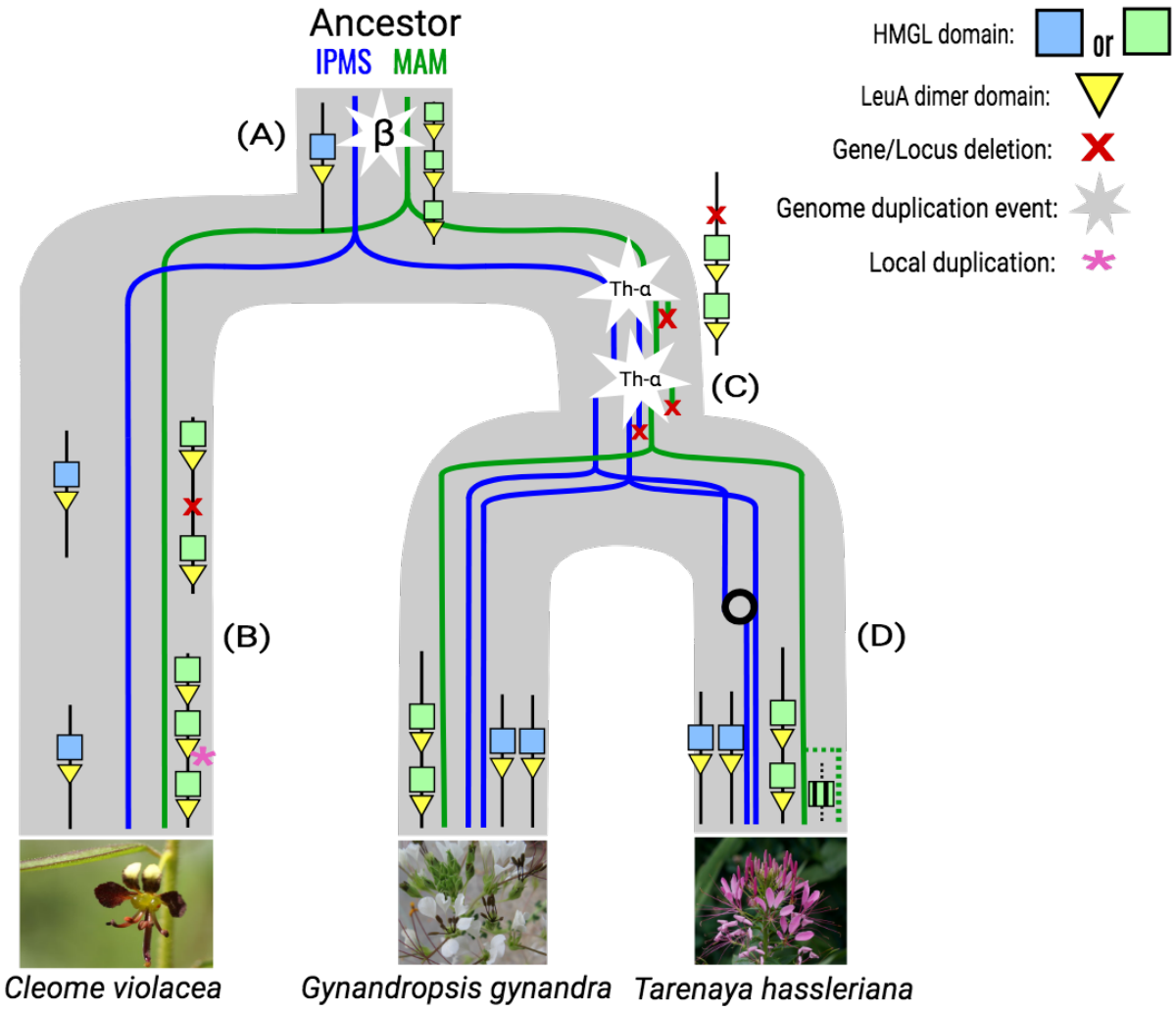


Figure 3.2: Inferred Evolutionary Trajectory of *MAM* and *IPMS* loci in the Cleomaceae based on genomic synteny and phylogenetic information. (A) The MAM-Ancestral locus originated from the β whole genome duplication event and is characterized by *MAM*–like genes that experience local duplication and have retained their LeuA domain. (B) In the genome of *Cleome violacea* there is a gene deletion followed by a novel tandem duplication at the MAM locus. (C) Following the Th-α whole genome triplication the *IPMS* locus is duplicated and the *MAM* locus experiences compensatory gene loss and is reflected in the *Gyanandropsis gynandra* genome. (D) In the *Tarenaya hassleriana* genome, the IPMS locus experiences a gene conversion event that maintains sequence similarity between the two copies. There is also a novel transposition of the MAM-like gene from the MAM-Ancestral locus that does not maintain the LeuA domain. As the placement of the Th-α whole genome duplication event is not confirmed to be fully shared by both lineages, an alternative reconstruction is also possible.

The MAM full-sequence tree shows bootstrap support between clades, but also a breakdown of some domain clades as well as clade nesting [Figure 3.3]. *MAMa* and *MAMb* separate by species lineage, while MAMc is unique to a small subset of Lineage I species and appears closely related to *MAMb* and *MAMe*. MAMd and MAMe are primarily the same as in the domain tree, but with other domains nested within. MAMf is consistent with the domain tree and sister to Lineage II *MAMa*.

To test for potential gene fusion events, full sequences of *MAMa* and *MAMb* Lineage I genes were broken up into "before the domain," "domain," and "after domain" sequences [Supplemental Figure 3.4]. Pairwise sequence comparisons were made between the Lineage I gene segments and corresponding segments of Lineage I *MAMe* genes, and Lineage II genes for *MAMa* or Lineage III genes for *MAMb*. In both cases, the domain portion best matches the corresponding domain regardless of Lineage. For Lineage I *MAMb*, the region before the domain is more similar to Lineage I MAMe than it is to Lineage II *MAMb*. For Lineage I *MAMa*, the region before the domain is more similar on average to Lineage I MAMe but was not significantly different from Lineage II *MAMa*.

DISCUSSION

The origin of all specialized metabolic pathways is primary metabolic genes, often with similar enzymatic chemistry (Moghe & Last 2015). This transition is mediated by the process of gene duplication and subsequent neutral mutation and neo/subfunctionalization (Conant & Wolf 2008; Moghe & Last, 2015). For the MAM locus of the glucosinolate (GSL) biosynthesis pathway, the role of tandem duplication

events in the evolution of the locus has been well characterized at the population level. The majority of work has only looked at Arabidopsis and its close relatives, and

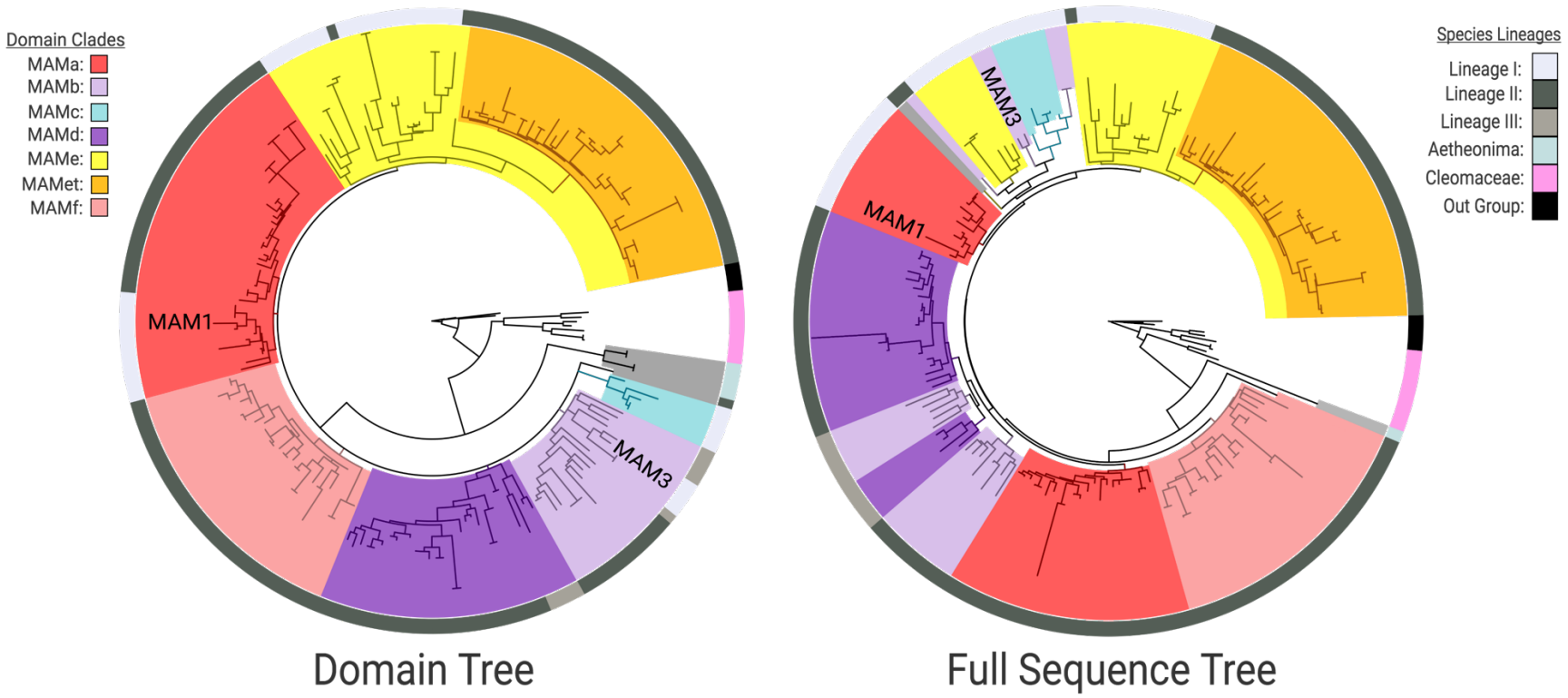


Figure 3.3: Clade comparison between Brassicaceae MAM domain and full sequence gene trees highlighting incongruence. HMGL-like sequences were used for the domain tree and resolved six clades of MAM (MAMa-f) but could not infer branching order. In the full sequence tree there is a breakdown of MAMa and MAMb that is correlated with species lineage. In both trees, the placement of MAM1 and MAM3 from the *Arabidopsis thaliana* Col genome are indicated. [For Bootstrap scores: Supplemental Figure 3.7 – Domain Tree Supplemental Figure 3.8 - Full Sequence; Online interactive trees: Domain - http://bit.ly/2Hb5jIS; Full Sequence - http://bit.ly/37btHEZ]

to a lesser extent, in the crop Brassicas (Kleibenstein & Cacho 2016). Much of what we understand about the MAM locus function has not been understood in the context of phylogeny, except to say that based on gene tree relationships, Lineage II and Lineage II have independently diversified from some initial gene substrate (Benderoth et al. 2008; Zhang et al. 2015). In this study, we utilized a micro-synteny network of genomes and phylogenetic inference to elucidate the evolutionary history of the MAM locus.

MAM in the Cleomaceae

The inclusion of Cleomaceae genomes in our analysis has provided novel insight into the origin of the *MAM* locus, following the whole genome duplication (WGD) event β, the hypothesized origin of *MAM* from *IPMS* (Van de Bergh et al. 2016). We estimate through micro-synteny and gene tree information that the Ancestral-*MAM* locus at the formation of the Cleomaceae was characterized by multiple MAM-like gene duplicates, the result of tandem duplications or local transposition [Figure 3.2]. These genes are different from what has been characterized in the Brassicaceae orthologous Ancestral-*MAM* locus, the Elong locus. They have retained their LeuA domain, the loss of which has been considered a critical step in the evolution of Brassicaceae MAM (de Kraker et al. 2007). Within the Cleomaceae, some genes of the Ancestral-*MAM* locus exhibit both Ancestral-*MAM* and *IPMS* syntenic cluster identity [Figure 3.1]. The syntenic window for these intermediates is shifted in comparison to other analyzed neighboring MAM-like genes. This allows for the inclusion of neighboring non-MAM genes that are more characteristic of the IPMS genomic context. This evidence supports the hypothesis that

the Ancestral-MAM locus was once a full context duplicate of the IPMS locus, and in the process of specialization over millions of years, degraded in collinearity.

How these MAM-like genes interact with GSL biosynthesis is unknown, but they have shown levels of expression in the leaf, seed, and roots in *Tarenaya hassleriana* (van den Bergh et al. 2016). The retention of the LeuA domain suggests that MAM-like proteins may have some continued interaction with IPMS or leucine biosynthesis. The ways in which genes respond to duplication events are constrained by their biochemical interactions, and therefore may shed insight into enzyme behavior (Birchler and Veitia, 2012; Bekaert et al., 2012; Conant et al., 2014; McLysaght et al. 2014). For example, given that IPMS experiences purifying selection of local gene duplicates and that *MAM-like* Cleomaceae genes found at the *MAM*-Ancestral locus do exhibit some local duplication, it is likely that these MAM-like genes have significantly sub- or neofunctionalized from their IPMS ancestor in terms of biochemical role. With that said, the dosage effects of *IPMS* are broader than only limiting local duplication, and through stoichiometric effects constrain most duplication types. Only after the β WGD event, is IPMS able to be retained and reduced in multiples of two. A pattern we see recapitulated after subsequent WGD events, with a few potential exceptions [Supplemental Figure 3.2]. Following Th-α, the Cleomaceae whole-genome triplication (WGT) or hexaploidy, there is an expected full context duplication of the IPMS locus, but with no context duplication of the Ancestral-*MAM* locus [Figure 3.2C]. In fact, we see a compensatory loss of a *MAM-like* gene following the increase in IPMS copy number. The presence of stoichiometric conflict between IPMS and these *MAM*-like genes would support the

hypothesis that they have retained some IPMS role and constraint. Further sampling across the Cleomaceae will be necessary to see if these patterns hold.

In the *Tarenaya hassleriana* genome, there is a novel a transposition of *MAM* [Figure 3.2D]. This transposed gene does not have a LeuA domain, bringing the overall *MAM/IPMS* gene number beyond what would be expected under an IPMS dosage constraint [Supplementary Figure 3.2]. This transposed locus has been shown to express in several tissues and to a greater extent in the leaf when compared to *MAM*-like counterparts at the Ancestral-*MAM* locus (van den Bergh et al. 2016). Increased species sampling, as well as an understanding of population-level variation in Cleomaceae *MAM*, is necessary for any conclusions on the dosage to be explored further using these methods. Direct biochemical assays of these *MAM*-like proteins will also be critical for characterizing any role they may play in glucosinolate biosynthesis and how that may differ from what is seen in the Brassicaceae. The Cleomaceae, and potentially the Capparaceae, which also shares the β duplication event (Edger et al. 2015), could serve as a powerful window into the evolution of early Brassicaceae *MAM* and a model for how gene families transition from primary to specialized metabolism.

MAM in the Brassicaceae

Between Lineages I, II, and III of the Brassicaceae, we have identified six distinct clades of *MAM*, *MAMa-f*, based on conserved HMGL-like domain sequences [Figure 3.3; Supplementary Figure 3.3]. Based on occurrence patterns across the family, we can say that *MAMb* and *MAMd* clades are ancestral to all three lineages, and *MAMa* and *MAMe* may be ancestral to only Linages I and Lineage II. The latter conclusion could not be confirmed by gene tree information and may be vulnerable to

sampling bias. The dispute between the chloroplast and nuclear species tree topologies could also affect the evolutionary relationships between the MAM clades and hamper our ability to predict (Nikolov et al., 2019). Improved sampling across the Brassicaceae is necessary before a robust estimation of the ancestral type can be made. That said, we are confident that *MAMc, MAMet,* and *MAMf* domain types are more recent innovations

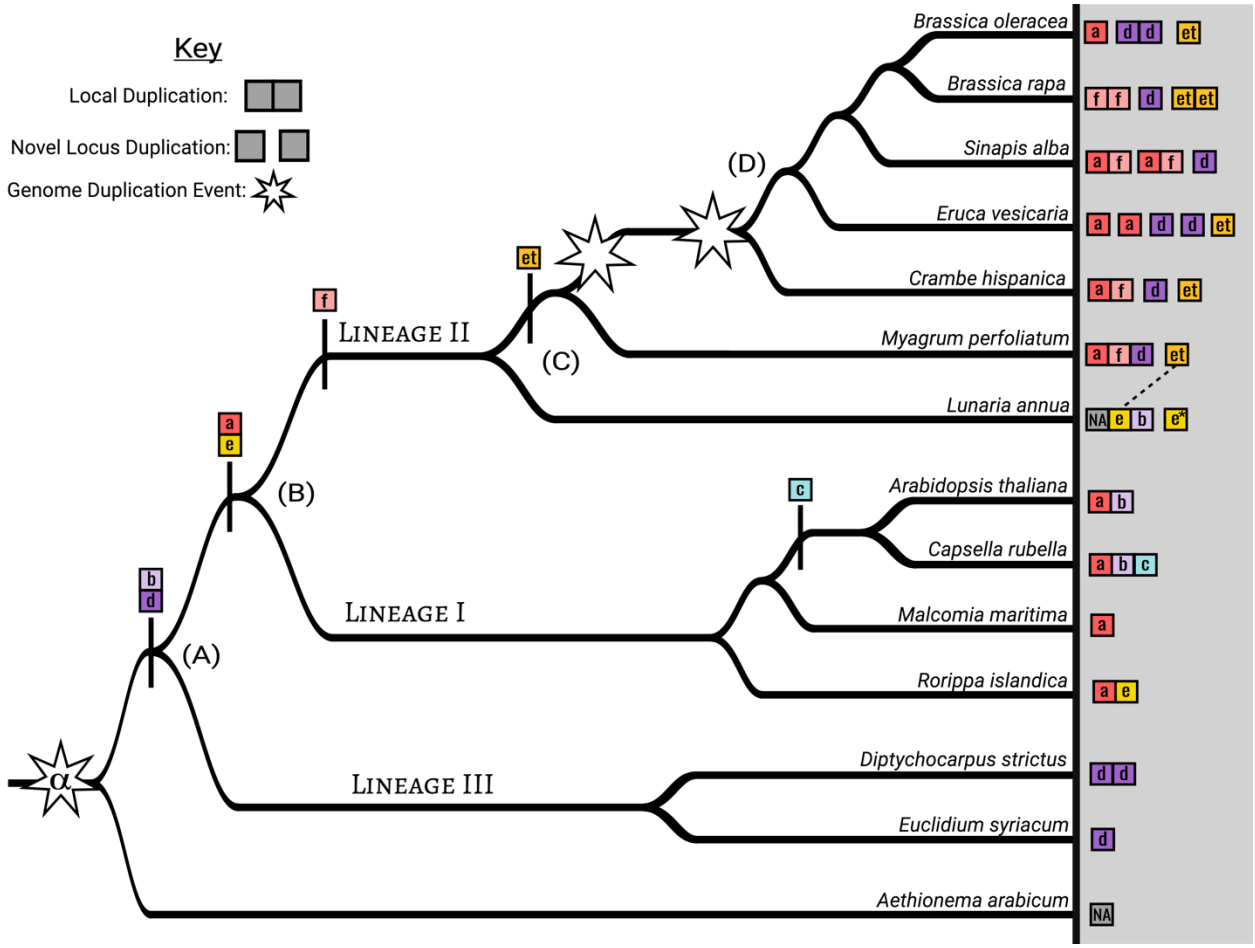


Figure 3.4: MAM clade and genomic context diversity within the Brassicaceae based on a subsample of analyzed genomes. Each square represents a MAM gene with an indicated HMGL-like domain type. Connected squares are found at the same physical location in the genome and not connected squares represent separate MAM loci (i.e. the MAM-Ancestral locus, a syntenic duplicate of the of the MAM-Ancestral locus, or MAM-Transposed). Non-syntenic gene transpositions were not included. (A) We estimate that the shared ancestor of Lineages I, II, and III maintained both MAMb and MAMd domain types. In the Lineage III genomes sampled MAMb genes were not located at the MAM-Ancestral locus, but at transposed loci. (B) At the ancestor of Lineage I and II, MAMa and MAMe appear, while the MAMc innovation occurs within a sub clade of Lineage I. (C) MAMf originates at the ancestor of Lineage II. The MAMet transposition that creates the MAM-Transposed locus occurs following the split from Lunaria annua, with all MAMet genes being closely related to a MAMe gene at the MAM-Ancestral locus. Lunaria annua also contains a context duplication of the MAM-Ancestral locus* that does not appear to be associated with whole genome duplication. (D) The unnamed whole genome duplication found in the tribe Brassiceae of Lineage II has resulted in multiple context duplications of the MAM-Ancestral locus. Full comparison is found in Supplemental Figure 3.3.

occurring in Lineage I and Lineage II, with specific branch placements [Figure 3.4; Supplementary Figure 3.3]

Given the functional role this domain plays in MAM biochemistry, we expect amino acid differences between domain types to be associated with generalizable patterns in *MAM* function. *MAMa* is the most conserved of the domains [Supplementary Figure 3.5B], suggesting that *MAMa* genes may contribute a necessary function to GSL biosynthesis, as compared to other MAM types. *MAMc* & *MAMet* are the most diverged, each having several unique amino acid substitutions when compared to other domain types [Supplementary Figure 3.5B]. Across all the domains, some sites were characterized by amino acid variability within and between domain types. Based on the characterization of *MAM* proteins in *Brassica juncea* (Kumar et al. 2019), we identified that oxo-acid binding sites were most often found at flexible amino acid positions followed by COA binding sites [Supplementary Figure 3.5A]. A better understanding of these patterns can give us insight into the forces driving the adaptation of *MAM*.

The domain and full-sequence gene trees conflict most significantly within the core Brassicaceae [Figure 3.3]. In the full-sequence tree Lineage I *MAMa* and *MAMb* genes appear more closely related to *MAMe* genes than to other genes of their shared domain. Sequence comparison reveals split-sequence similarities in both *MAMa* and *MAMb* domain clade groups. This pattern suggests two possibilities: 1) *MAM* genes experienced convergent evolution of their amino acid sequences, or 2) a gene fusion event of separate *MAM* types occurred sometime during the divergence of Lineage I MAM. The latter scenario is both the more parsimonious conclusion, and it is supported by the previous characterization of population-level gene fusion events at the

Ancestral-*MAM* locus (Benderoth et al. 2009). Given that Lineage I *MAMa* and *MAMb* genes show a close phylogenetic relationship to Lineage I *MAMe*, in conflict with the domain tree, it is the most likely donor gene. Both fusion events would have occurred at separate nodes of the Lineage I species tree, *MAMa/MAMe* fusion happening earlier than the *MAMb/MAMe* event. Improved sampling of Lineage I is necessary to identify the specific species branch points at which the events occurred. The fusion of MAM genes at the MAM-Ancestral locus, though largely studied from only a population level, may have been a critical driver of MAM diversity and innovation within Lineage I in the Brassicaceae.

Most of the genes in each domain clade exist at the *MAM*-Ancestral locus. This is true for genes of the *MAMe* group except for a nested clade of transposed genes, *MAMet*, that form the unique syntenic cluster MAM-Transposed [Figure 3.1; Figure 3.4]. There are subsequent transpositions from the MAM-Transposed locus, many of which show signs of degradation. The initial transposition occurred sometime following the split from the ancestor of *Lunaria annua* to the common ancestor of *Thellungiella (Eutrema*) and the rest of Lineage II [Supplementary Figure 3.3]. Following the transposition event, there is a loss of all *MAMe* domain type genes. Of our dataset, *L. annua* is the only member of Lineage II to retain any copies of *MAMe*. Of those *MAMe* genes, most appear closely related to Lineage I *MAMe* genes, while one copy is most closely related to *MAMet* in both the domain and full sequence trees [Figure 3.3]. This transposition event is the earliest conserved instance of a novel *MAM* context, which allows for an escape from cis-regulatory effects that may be experienced at the *MAM*-Ancestral locus (Chen & Ni 2006; Conant & wolf 2008). The possibilities exist

that these genes are performing some yet to be characterized function or potentially may represent the GSL-PRO locus characterized in Brassica species. With this current analysis, we cannot further speculate on the role *MAMet* genes may be playing in GSL biosynthesis, except to say that experimental analysis of these genes will be necessary to understand their place in metabolic innovation.

Polyploidy offers another mechanism for *MAM diversification, by escaping* potential cis-regulatory effects of other *MAM* genes or sub- and neofunctionalization of resulting duplicates. In the Cleomaceae, the *MAM*-Ancestral locus duplicates are not retained following genome doubling, putatively due to the presence of their LeuA domain and restrictions under gene dosage. Without such dosage constraints in Brassicaceae *MAM*, most genomes sampled show retention of a duplicated *MAM*-Ancestral locus following known WGD events. For example, the WGT event in the tribe Brassiceae of Lineage II resulted in three homoeologous *MAM*-Ancestral loci in subsequently diploidized genomes [Figure 3.4; Supplementary Figure 3.2]. In *Brassica rapa*, *Brassica oleracea,* and *Eruca vesicaria*, the *MAM*-Ancestral loci maintain a single MAM domain type (*MAMa, MAMd, or MAMf*) at each. Whereas in other genomes, like S*inapis alba*, *MAMa* and *MAMf* genes remain paired although duplicated at separate loci. We propose that phenotypic differences between Brassica and Arabidopsis, such as the ability to co-synthesize different carbon chain majority phenotypes, are facilitated by the physical separation of MAM genes within the genome. By influencing the rate of diversification for MAM genes at the different MAM-Ancestral loci and allowing for novel genomic interactions, the WGT may have been a

critical step in driving the specialized metabolic innovation we see in this dynamic crop lineage.

CONCLUSION

The *MAM*/IPMS gene family serves as an excellent example of how a primary metabolic gene can, over millions of years and leveraging any source of novelty, give rise to a diverse lineage of highly adaptive specialized metabolic genes. Utilizing micro-synteny gene networks and broad phylogenetic sampling, we find that multiple modes of gene duplication have significantly influenced the evolutionary trajectory of the *MAM* locus and thereby diversity of aliphatic GSL profiles. By exploring some of the evolutionary consequences of whole-genome duplication, gene transposition, local duplication, and gene fusion, we have generated several new testable hypotheses as to the nature of MAM and GSL diversity. In the future, new experimental approaches and broad phylogenetically informed sampling will be critical to continue developing a robust understanding of this important gene family.

REFERENCES

Bekaert, Michaël, Patrick P. Edger, Corey M. Hudson, J. Chris Pires, and Gavin C. Conant. 2012. "Metabolic and Evolutionary Costs of Herbivory Defense: Systems Biology of Glucosinolate Synthesis." *The New Phytologist* 196 (2): 596–605. https://doi.org/10.1111/j.1469-8137.2012.04302.x.

Benderoth, Markus, Marina Pfalz, and Juergen Kroymann. 2008. "Methylthioalkylmalate Synthases: Genetics, Ecology and Evolution." *Phytochemistry Reviews* 8 (1): 255. https://doi.org/10.1007/s11101-008-9097-1.

Benderoth, Markus, Susanne Textor, Aaron J. Windsor, Thomas Mitchell-Olds, Jonathan Gershenzon, and Juergen Kroymann. 2006. "Positive Selection Driving Diversification in Plant Secondary Metabolism." *Proceedings of the National Academy of Sciences of the United States of America* 103 (24): 9118–23. https://doi.org/10.1073/pnas.0601738103.

Bergh, Erik van den, Johannes A. Hofberger, and M. Eric Schranz. 2016. "Flower Power and the Mustard Bomb: Comparative Analysis of Gene and Genome Duplications in Glucosinolate Biosynthetic Pathway Evolution in Cleomaceae and Brassicaceae." *American Journal of Botany* 103 (7): 1212–22. https://doi.org/10.3732/ajb.1500445.

Birchler, James A., and Reiner A. Veitia. 2012. "Gene Balance Hypothesis: Connecting Issues of Dosage Sensitivity across Biological Disciplines." *Proceedings of the National Academy of Sciences of the United States of America* 109 (37): 14746–53. https://doi.org/10.1073/pnas.1207726109.

Blažević, Ivica, Sabine Montaut, Franko Burčul, Carl Erik Olsen, Meike Burow, Patrick Rollin, and Niels Agerbirk. 2020. "Glucosinolate Structural Diversity, Identification, Chemical Synthesis and Metabolism in Plants." *Phytochemistry* 169 (January): 112100. https://doi.org/10.1016/j.phytochem.2019.112100.

Borpatragohain, Priyakshee, Terry J. Rose, and Graham J. King. 2016. "Fire and Brimstone: Molecular Interactions between Sulfur and Glucosinolate Biosynthesis in Model and Crop Brassicaceae." *Frontiers in Plant Science* 7 (November): 1735. https://doi.org/10.3389/fpls.2016.01735.

Chen, Z. Jeffrey, and Zhongfu Ni. 2006. "Mechanisms of Genomic Rearrangements and Gene Expression Changes in Plant Polyploids." *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 28 (3): 240–52. https://doi.org/10.1002/bies.20374.

Chhajed, Shweta, Biswapriya B. Misra, Nathalia Tello, and Sixue Chen. 2019. "Chemodiversity of the Glucosinolate-Myrosinase System at the Single Cell Type Resolution." *Frontiers in Plant Science*. https://doi.org/10.3389/fpls.2019.00618.

Conant, Gavin C., James A. Birchler, and J. Chris Pires. 2014. "Dosage, Duplication, and Diploidization: Clarifying the Interplay of Multiple Models for Duplicate Gene Evolution over Time." *Current Opinion in Plant Biology* 19 (June): 91–98. https://doi.org/10.1016/j.pbi.2014.05.008.

Conant, Gavin C., and Kenneth H. Wolfe. 2008. "Probabilistic Cross-Species Inference of Orthologous Genomic Regions Created by Whole-Genome Duplication in Yeast." *Genetics* 179 (3): 1681–92. https://doi.org/10.1534/genetics.107.074450.

Del Carmen Martínez-Ballesta, María, Diego A. Moreno, and Micaela Carvajal. 2013. "The Physiological Importance of Glucosinolates on Plant Response to Abiotic Stress in Brassica." *International Journal of Molecular Sciences* 14 (6): 11607–25. https://doi.org/10.3390/ijms140611607.

Derényi, Imre, Gergely Palla, and Tamás Vicsek. 2005. "Clique Percolation in Random Networks." *Physical Review Letters* 94 (16): 160202. https://doi.org/10.1103/PhysRevLett.94.160202.

Edger, Patrick P., Hanna M. Heidel-Fischer, Michaël Bekaert, Jadranka Rota, Gernot Glöckner, Adrian E. Platts, David G. Heckel, et al. 2015. "The Butterfly Plant Arms-Race Escalated by Gene and Genome Duplications." *Proceedings of the National Academy of Sciences of the United States of America* 112 (27): 8362–66. https://doi.org/10.1073/pnas.1503926112.

Finn, R. D., J. Clements, and S. R. Eddy. 2011. "HMMER Web Server: Interactive Sequence Similarity Searching." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkr367.

Fortunato, Santo. 2010. "Community Detection in Graphs." *Physics Reports* 486 (3): 75–174. https://doi.org/10.1016/j.physrep.2009.11.002.

Hofberger, Johannes A., Eric Lyons, Patrick P. Edger, J. Chris Pires, and M. Eric Schranz. 2013. "Whole Genome and Tandem Duplicate Retention Facilitated Glucosinolate Pathway Diversification in the Mustard Family." *Genome Biology and Evolution* 5 (11): 2155–73. https://doi.org/10.1093/gbe/evt162.

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. https://doi.org/10.1093/bib/bbx108.

Keurentjes, Joost J. B., Jingyuan Fu, C. H. Ric de Vos, Arjen Lommen, Robert D. Hall, Raoul J. Bino, Linus H. W. van der Plas, Ritsert C. Jansen, Dick Vreugdenhil, and Maarten Koornneef. 2006. "The Genetics of Plant Metabolism." *Nature Genetics* 38 (7): 842–49. https://doi.org/10.1038/ng1815.

Kliebenstein, D. J., and N. I. Cacho. 2016. "Chapter Three - Nonlinear Selection and a Blend of Convergent, Divergent and Parallel Evolution Shapes Natural Variation in Glucosinolates." In *Advances in Botanical Research*, edited by Stanislav Kopriva, 80:31–55. Academic Press. https://doi.org/10.1016/bs.abr.2016.06.002.

Kliebenstein, D. J., J. Gershenzon, and T. Mitchell-Olds. 2001. "Comparative Quantitative Trait Loci Mapping of Aliphatic, Indolic and Benzylic Glucosinolate Production in Arabidopsis Thaliana Leaves and Seeds." *Genetics* 159 (1): 359–70. https://www.ncbi.nlm.nih.gov/pubmed/11560911.

Kliebenstein, D. J., V. M. Lambrix, M. Reichelt, J. Gershenzon, and T. Mitchell-Olds. 2001. "Gene Duplication in the Diversification of Secondary Metabolism: Tandem 2-Oxoglutarate-Dependent Dioxygenases Control Glucosinolate Biosynthesis in Arabidopsis." *The Plant Cell* 13 (3): 681–93. https://doi.org/10.1105/tpc.13.3.681.

Kliebenstein, Daniel J. 2008. "A Role for Gene Duplication and Natural Variation of Gene Expression in the Evolution of Metabolism." *PloS One* 3 (3): e1838. https://doi.org/10.1371/journal.pone.0001838.

Koon, Nayden, Christopher J. Squire, and Edward N. Baker. 2004. "Crystal Structure of LeuA from Mycobacterium Tuberculosis, a Key Enzyme in Leucine Biosynthesis." *Proceedings of the National Academy of Sciences of the United States of America* 101 (22): 8295–8300. https://doi.org/10.1073/pnas.0400820101.

Kraker, Jan-Willem de, and Jonathan Gershenzon. 2011. "From Amino Acid to Glucosinolate Biosynthesis: Protein Sequence Changes in the Evolution of Methylthioalkylmalate Synthase in Arabidopsis." *The Plant Cell* 23 (1): 38–53. https://doi.org/10.1105/tpc.110.079269.

Kraker, Jan-Willem de, Katrin Luck, Susanne Textor, James G. Tokuhisa, and Jonathan Gershenzon. 2007. "Two Arabidopsis Genes (IPMS1 and IPMS2) Encode Isopropylmalate Synthase, the Branchpoint Step in the Biosynthesis of Leucine." *Plant Physiology* 143 (2): 970–86. https://doi.org/10.1104/pp.106.085555.

Kroymann, Juergen, and Thomas Mitchell-Olds. 2005. "Epistasis and Balanced Polymorphism Influencing Complex Trait Variation." *Nature* 435 (7038): 95–98. https://doi.org/10.1038/nature03480.

Kumar, Roshan, Soon Goo Lee, Rehna Augustine, Micheal Reichelt, Daniel G. Vassão, Manoj H. Palavalli, Aron Allen, Jonathan Gershenzon, Joseph M. Jez, and Naveen C. Bisht. 2019. "Molecular Basis of the Evolution of Methylthioalkylmalate Synthase and the Diversity of Methionine-Derived Glucosinolates." *The Plant Cell* 31 (7): 1633–47. https://doi.org/10.1105/tpc.19.00046.

Kuraku, Shigehiro, Christian M. Zmasek, Osamu Nishimura, and Kazutaka Katoh. 2013. "aLeaves Facilitates on-Demand Exploration of Metazoan Gene Family Trees on MAFFT Sequence Alignment Server with Enhanced Interactivity." *Nucleic Acids Research* 41 (Web Server issue): W22–28. https://doi.org/10.1093/nar/gkt389.

Madeira, Fábio, Young Mi Park, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, et al. 2019. "The EMBL-EBI Search and Sequence Analysis Tools APIs in 2019." *Nucleic Acids Research* 47 (W1): W636–41. https://doi.org/10.1093/nar/gkz268.

McLysaght, Aoife, Takashi Makino, Hannah M. Grayton, Maria Tropeano, Kevin J. Mitchell, Evangelos Vassos, and David A. Collier. 2014. "Ohnologs Are Overrepresented in Pathogenic Copy Number Mutations." *Proceedings of the National Academy of Sciences of the United States of America* 111 (1): 361–66. https://doi.org/10.1073/pnas.1309324111.

Moghe, Gaurav D., and Robert L. Last. 2015. "Something Old, Something New: Conserved Enzymes and the Evolution of Novelty in Plant Specialized Metabolism." *Plant Physiology* 169 (3): 1512–23. https://doi.org/10.1104/pp.15.00994.

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. https://doi.org/10.1111/nph.15732.

Olsen, Carl Erik, Xiao-Chen Huang, Cecilie I. C. Hansen, Don Cipollini, Marian Ørgaard, Annemarie Matthes, Fernando Geu-Flores, Marcus A. Koch, and Niels Agerbirk. 2016. "Glucosinolate Diversity within a Phylogenetic Framework of the Tribe Cardamineae (Brassicaceae) Unraveled with HPLC-MS/MS and NMR-Based Analytical Distinction of 70 Desulfoglucosinolates." *Phytochemistry* 132 (December): 33–56. https://doi.org/10.1016/j.phytochem.2016.09.013.

Palla, Gergely, Imre Derényi, Illés Farkas, and Tamás Vicsek. 2005. "Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society." *Nature* 435 (7043): 814–18. https://doi.org/10.1038/nature03607.

Peng, Cheng, Sahra Uygun, Shin-Han Shiu, and Robert L. Last. 2015. "The Impact of the Branched-Chain Ketoacid Dehydrogenase Complex on Amino Acid Homeostasis in Arabidopsis." *Plant Physiology* 169 (3): 1807–20. https://doi.org/10.1104/pp.15.00461.

Petersen, Annette, Lea Gram Hansen, Nadia Mirza, Christoph Crocoll, Osman Mirza, and Barbara Ann Halkier. 2019. "Changing Substrate Specificity and Iteration of Amino Acid Chain Elongation in Glucosinolate Biosynthesis through Targeted Mutagenesis of Arabidopsis Methylthioalkylmalate Synthase 1." *Bioscience Reports*. https://doi.org/10.1042/bsr20190446.

Rodman, J., P. Soltis, D. Soltis, K. Sytsma, and K. Karol. 1998. "Parallel Evolution of Glucosinolate Biosynthesis Inferred from Congruent Nuclear and Plastid Gene Phylogenies." *American Journal of Botany* 85 (7): 997. https://doi.org/10.2307/2446366.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research* 13 (11): 2498–2504. https://doi.org/10.1101/gr.1239303.

Smith, Stephen A., and Casey W. Dunn. 2008. "Phyutility: A Phyloinformatics Tool for Trees, Alignments and Molecular Data." *Bioinformatics* 24 (5): 715–16. https://doi.org/10.1093/bioinformatics/btm619.

Stamatakis, Alexandros. 2014. "RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies." *Bioinformatics* 30 (9): 1312–13. https://doi.org/10.1093/bioinformatics/btu033.

Tang, Haibao, John E. Bowers, Xiyin Wang, Ray Ming, Maqsudul Alam, and Andrew H. Paterson. 2008. "Synteny and Collinearity in Plant Genomes." *Science* 320 (5875): 486–88. https://doi.org/10.1126/science.1153917.

Textor, Susanne, Stefan Bartram, Jürgen Kroymann, Kimberly L. Falk, Alastair Hick, John A. Pickett, and Jonathan Gershenzon. 2004. "Biosynthesis of Methionine-Derived Glucosinolates in Arabidopsis Thaliana: Recombinant Expression and Characterization of Methylthioalkylmalate Synthase, the Condensing Enzyme of the Chain-Elongation Cycle." *Planta* 218 (6): 1026–35. https://doi.org/10.1007/s00425-003-1184-3.

Textor, Susanne, Jan-Willem de Kraker, Bettina Hause, Jonathan Gershenzon, and James G. Tokuhisa. 2007. "MAM3 Catalyzes the Formation of All Aliphatic

Glucosinolate Chain Lengths in Arabidopsis." *Plant Physiology* 144 (1): 60–71. https://doi.org/10.1104/pp.106.091579.

Wang, Yupeng, Haibao Tang, Jeremy D. Debarry, Xu Tan, Jingping Li, Xiyin Wang, Tae-Ho Lee, et al. 2012. "MCScanX: A Toolkit for Detection and Evolutionary Analysis of Gene Synteny and Collinearity." *Nucleic Acids Research* 40 (7): e49. https://doi.org/10.1093/nar/gkr1293.

Wentzell, Adam M., Heather C. Rowe, Bjarne Gram Hansen, Carla Ticconi, Barbara Ann Halkier, and Daniel J. Kliebenstein. 2007. "Linking Metabolic QTLs with Network and Cis-eQTLs Controlling Biosynthetic Pathways." *PLoS Genetics* 3 (9): 1687–1701. https://doi.org/10.1371/journal.pgen.0030162.

Wisecaver, Jennifer H., Alexander T. Borowsky, Vered Tzin, Georg Jander, Daniel J. Kliebenstein, and Antonis Rokas. 2017. "A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants." *The Plant Cell* 29 (5): 944–59. https://doi.org/10.1105/tpc.17.00009.

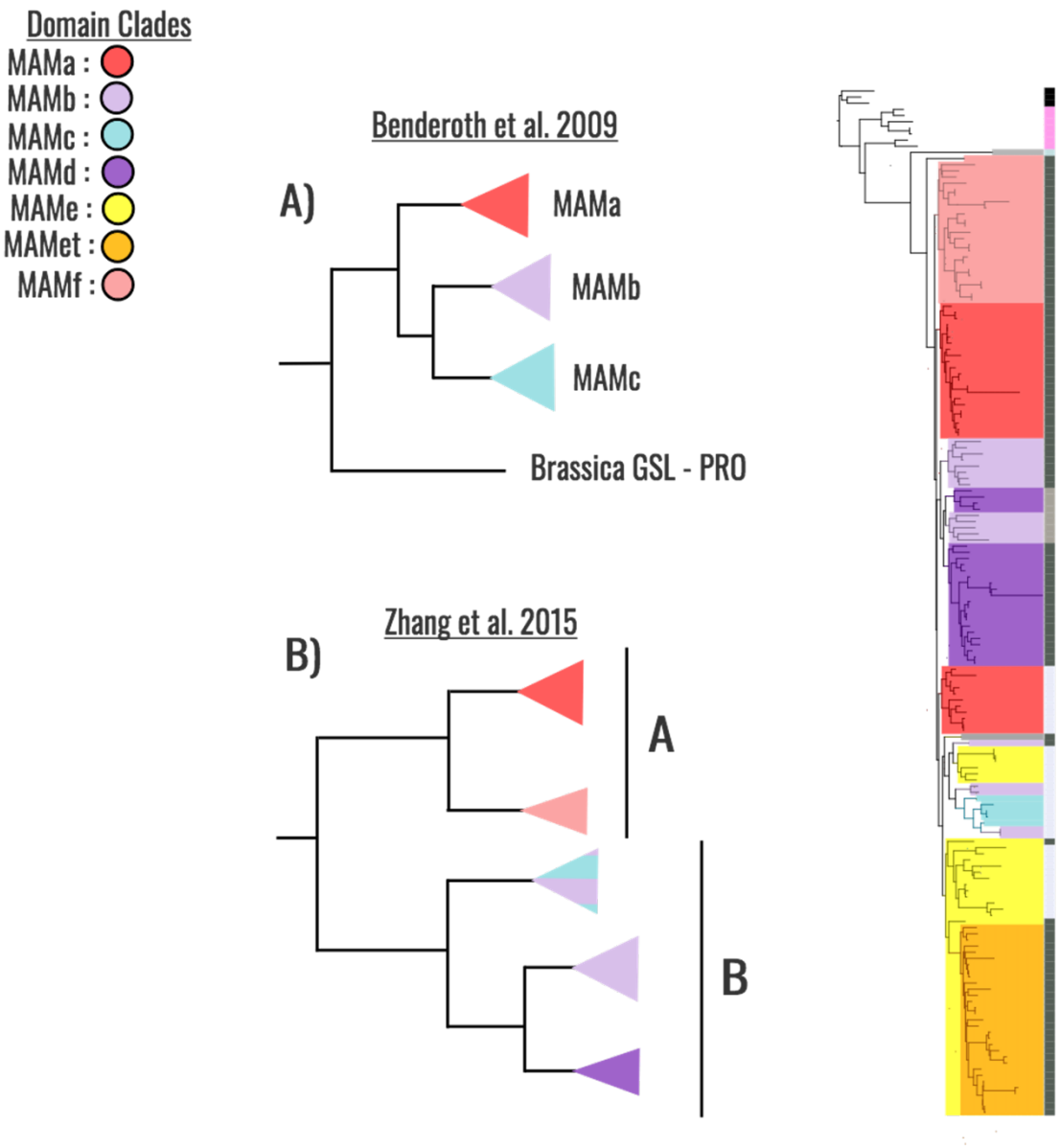
Zhang, Jifang, Xiaobo Wang, Feng Cheng, Jian Wu, Jianli Liang, Wencai Yang, and Xiaowu Wang. 2015. "Lineage-Specific Evolution of Methylthioalkylmalate Synthases (MAMs) Involved in Glucosinolates Biosynthesis." *Frontiers in Plant Science* 6 (February): 18. https://doi.org/10.3389/fpls.2015.00018.

Zhao, Tao, Rens Holmer, Suzanne de Bruijn, Gerco C. Angenent, Harrold A. van den Burg, and M. Eric Schranz. 2017. "Phylogenomic Synteny Network Analysis of MADS-Box Transcription Factor Genes Reveals Lineage-Specific Transpositions, Ancient Tandem Duplications, and Deep Positional Conservation." *The Plant Cell* 29 (6): 1278–92. https://doi.org/10.1105/tpc.17.00312.

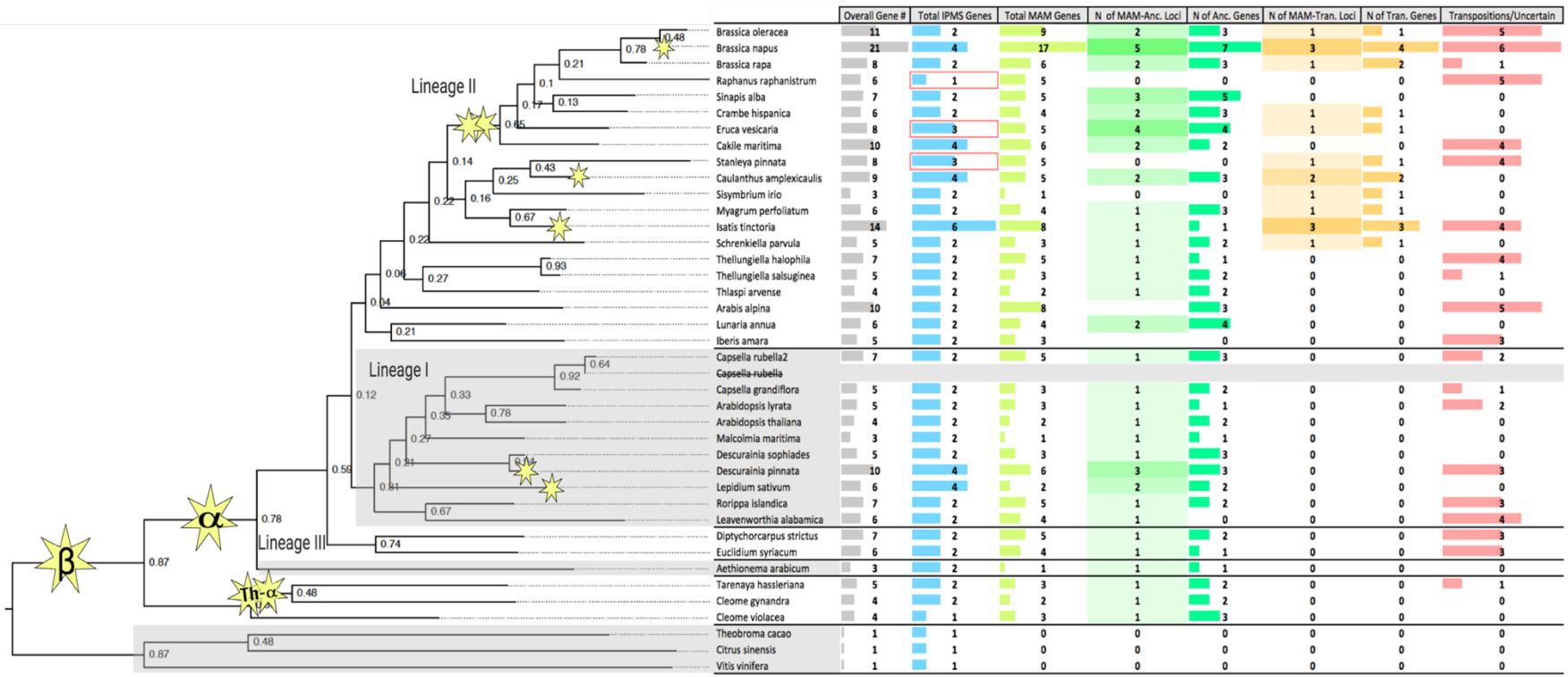
Zhao, Tao, Dong Liang, Ping Wang, Jingying Liu, and Fengwang Ma. 2012. "Genome-Wide Analysis and Expression Profiling of the DREB Transcription Factor Gene Family in Malus under Abiotic Stress." *Molecular Genetics and Genomics: MGG* 287 (5): 423–36. https://doi.org/10.1007/s00438-012-0687-7

Zhao, Tao, and M. Eric Schranz. 2019. "Network-Based Microsynteny Analysis Identifies Major Differences and Genomic Outliers in Mammalian and Angiosperm Genomes." *Proceedings of the National Academy of Sciences of the United States of America* 116 (6): 2165–74. https://doi.org/10.1073/pnas.1801757116.

Zhao, Yongan, Haixu Tang, and Yuzhen Ye. 2012. "RAPSearch2: A Fast and Memory-Efficient Protein Similarity Search Tool for next-Generation Sequencing Data." *Bioinformatics* 28 (1): 125–26. https://doi.org/10.1093/bioinformatics/btr595.

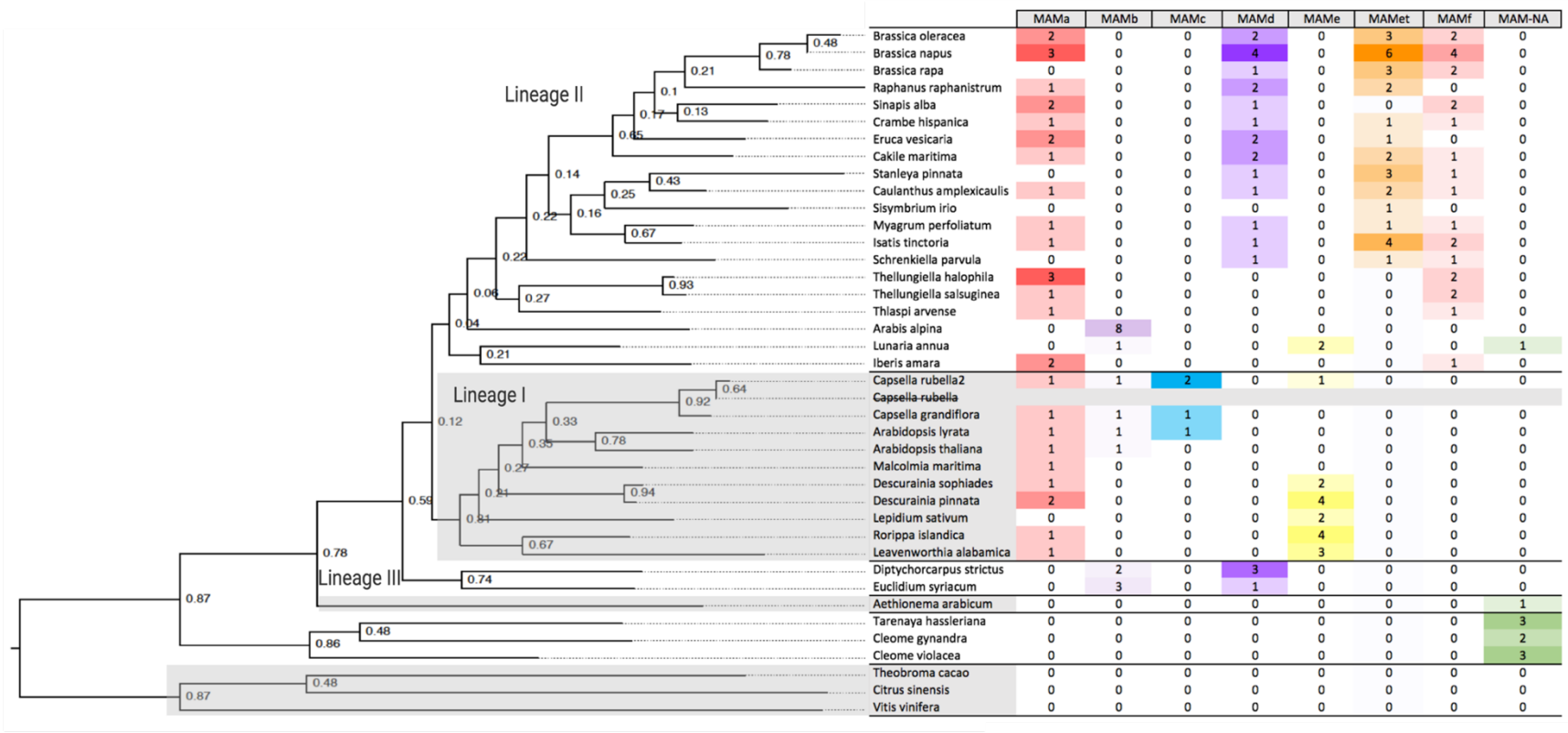


Supplementary Figure 3.2. (A) Benderoth et al. 2009 describes the MAM lineage in terms of orthology to *Arabidopsis lyrata* gene tree clades. While the topology generally agrees with our tree, the emphasis on Arabidopsis and close relatives gives a limited picture of MAM diversity. This tree also supported the hypothesis that MAM has evolved separately in the Lineage I and II. (B) Zhang et al. 2015 generally agrees with this hypothesis though they do show shared clades not solely informed by the species tree. Some of their topology conflicts with our full sequence tree and yet agrees with the domain specific tree. This may be due to how their alignment was cleaned and their species sampling.



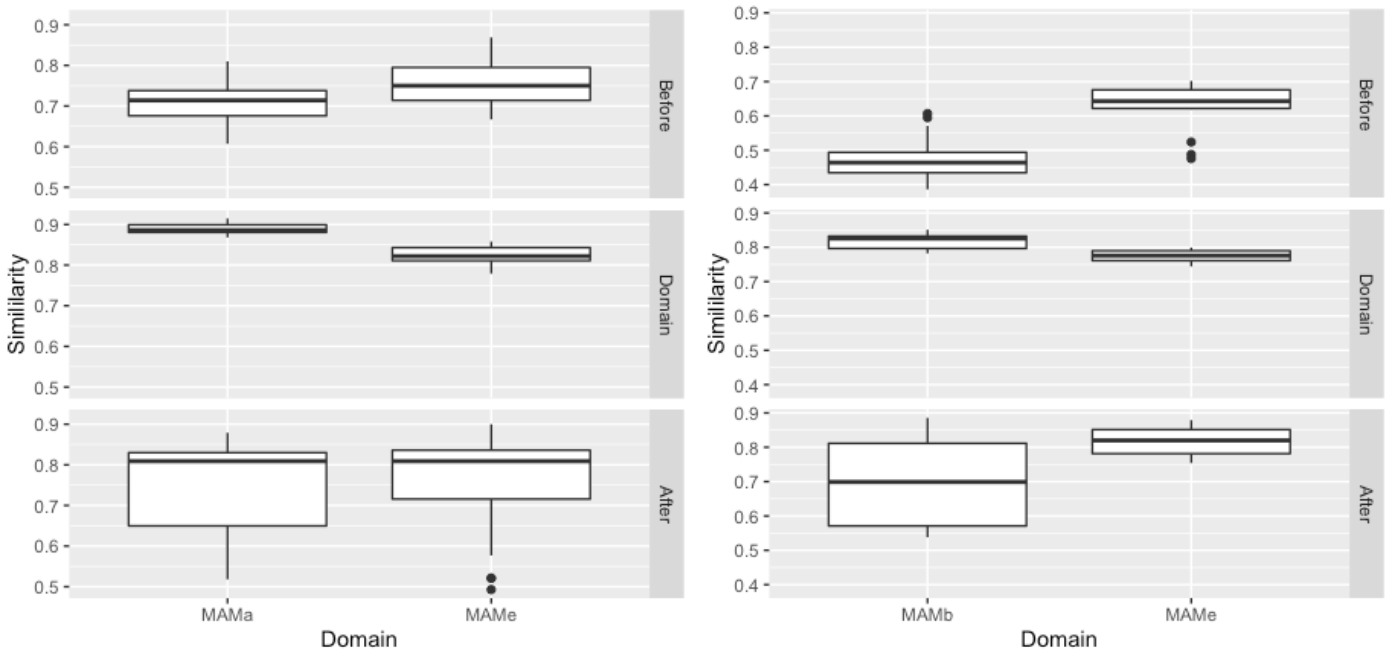
	Overall Gene #	Total IPMS Genes	Total MAM Genes	N of MAM-Anc. Loci	N of Anc. Genes	N of MAM-Tran. Loci	N of Tran. Genes	Transpositions/Uncertain
Brassica oleracea	11	2	9	2	3	1	1	5
Brassica napus	21	4	17	5	7	3	4	6
Brassica rapa	8	2	6	2	3	1	2	1
Raphanus raphanistrum	6	1	5	0	0	0	0	5
Sinapis alba	7	2	5	3	5	0	0	0
Crambe hispanica	6	2	4	2	3	1	1	0
Eruca vesicaria	8	3	5	4	4	1	1	0
Cakile maritima	10	4	6	2	2	0	0	4
Stanleya pinnata	8	3	5	0	0	1	1	4
Caulanthus amplexicaulis	9	4	5	2	3	2	2	0
Sisymbrium irio	3	2	1	0	0	1	1	0
Myagrum perfoliatum	6	2	4	1	3	1	1	0
Isatis tinctoria	14	6	8	1	1	3	3	4
Schrenkiella parvula	5	2	3	1	2	1	1	0
Thellungiella halophila	7	2	5	1	1	0	0	4
Thellungiella salsuginea	5	2	3	1	2	0	0	1
Thlaspi arvense	4	2	2	1	2	0	0	0
Arabis alpina	10	2	8		3	0	0	5
Lunaria annua	6	2	4	2	4	0	0	0
Iberis amara	5	2	3		0	0	0	3
Capsella rubella2	7	2	5	1	3	0	0	2
~~Capsella rubella~~								
Capsella grandiflora	5	2	3	1	2	0	0	1
Arabidopsis lyrata	5	2	3	1	1	0	0	2
Arabidopsis thaliana	4	2	2	1	2	0	0	0
Malcolmia maritima	3	2	1	1	1	0	0	0
Descurainia sophiades	5	2	3	1	3	0	0	0
Descurainia pinnata	10	4	6	3	3	0	0	3
Lepidium sativum	6	4	2	2	2	0	0	0
Rorippa islandica	7	2	5	1	2	0	0	3
Leavenworthia alabamica	6	2	4	1	0	0	0	4
Diptychorcarpus strictus	7	2	5	1	2	0	0	3
Euclidium syriacum	6	2	4	1	1	0	0	3
Aethionema arabicum	3	2	1	1	1	0	0	0
Tarenaya hassleriana	5	2	3	1	2	0	0	1
Cleome gynandra	4	2	2	1	2	0	0	0
Cleome violacea	4	1	3	1	3	0	0	0
Theobroma cacao	1	1	0	0	0	0	0	0
Citrus sinensis	1	1	0	0	0	0	0	0
Vitis vinifera	1	1	0	0	0	0	0	0

Supplementary Figure 3.3. The overall gene counts per genome for the *MAM/IPMS* gene family. Gene numbers, especially in IPMS, are correlated with recent polyploidy. Three genomes conflict with the expected *IPMS* dosage expectation of multiples of two. The *Raphanus raphanistrum* and *Stanleya pinnata* IPMS deviations may be an artifact of lower quality genomes, but the Eruca vesicaria retention appears to be a newly sub-functionalized *IPMS* copy, exhibiting an intermediate syntenic relationships to that of some *MAM*-Ancestral genes in the Cleomaceae. For MAM, the number of Loci indicates whether MAM-Ancestral or MAM-Transposed has experienced a context duplication. The number of genes at that locus is the overall total of genes across all syntenic loci of that type.

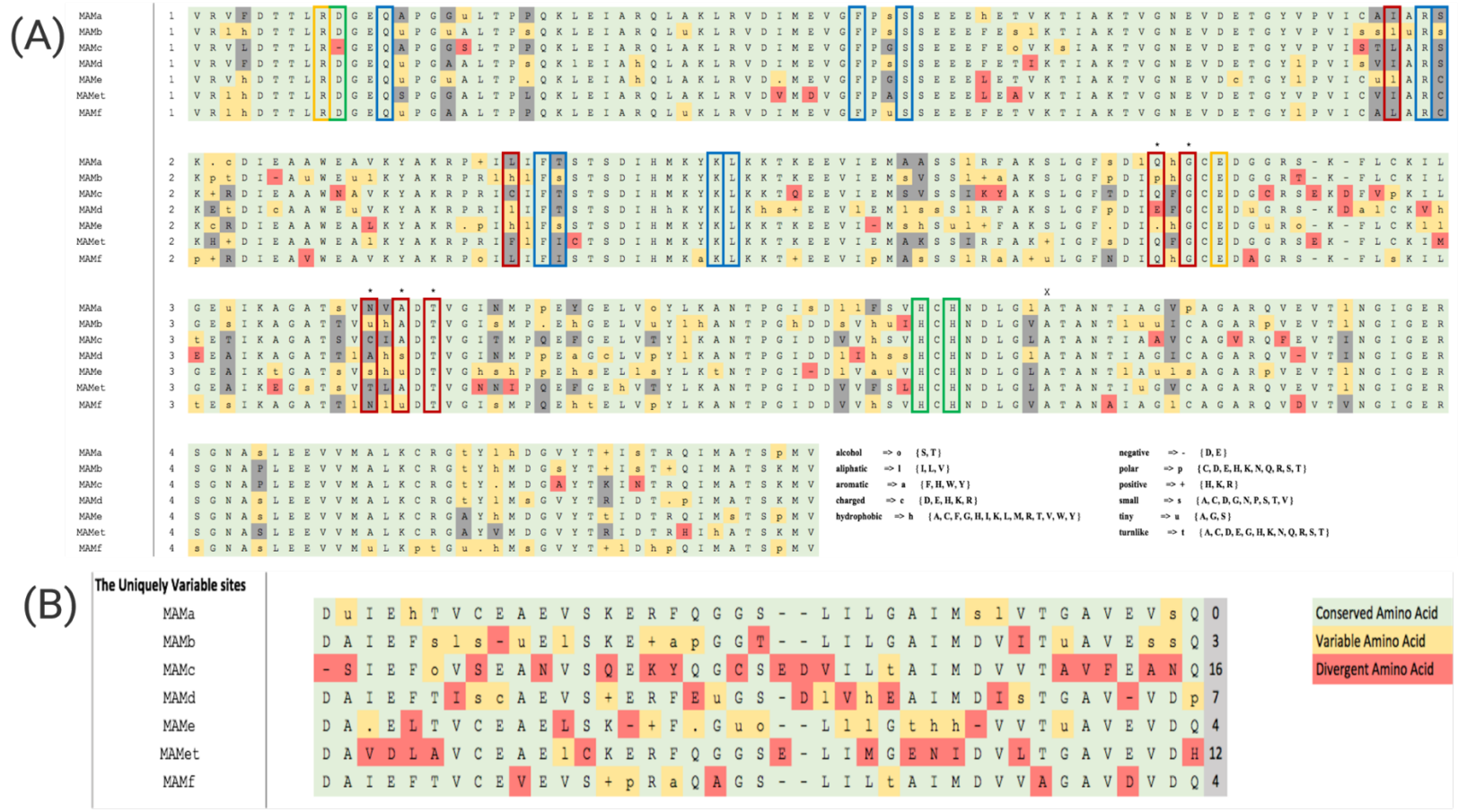


	MAMa	MAMb	MAMc	MAMd	MAMe	MAMet	MAMf	MAM-NA
Brassica oleracea	2	0	0	2	0	3	2	0
Brassica napus	3	0	0	4	0	6	4	0
Brassica rapa	0	0	0	1	0	3	2	0
Raphanus raphanistrum	1	0	0	2	0	2	0	0
Sinapis alba	2	0	0	1	0	0	2	0
Crambe hispanica	1	0	0	1	0	1	1	0
Eruca vesicaria	2	0	0	2	0	1	0	0
Cakile maritima	1	0	0	2	0	2	1	0
Stanleya pinnata	0	0	0	1	0	3	1	0
Caulanthus amplexicaulis	1	0	0	1	0	2	1	0
Sisymbrium irio	0	0	0	0	0	1	0	0
Myagrum perfoliatum	1	0	0	1	0	1	1	0
Isatis tinctoria	1	0	0	1	0	4	2	0
Schrenkiella parvula	0	0	0	1	0	1	1	0
Thellungiella halophila	3	0	0	0	0	0	2	0
Thellungiella salsuginea	1	0	0	0	0	0	2	0
Thlaspi arvense	1	0	0	0	0	0	1	0
Arabis alpina	0	8	0	0	0	0	0	0
Lunaria annua	0	1	0	0	2	0	0	1
Iberis amara	2	0	0	0	0	0	1	0
Capsella rubella2	1	1	2	0	1	0	0	0
~~Capsella rubella~~								
Capsella grandiflora	1	1	1	0	0	0	0	0
Arabidopsis lyrata	1	1	1	0	0	0	0	0
Arabidopsis thaliana	1	1	0	0	0	0	0	0
Malcolmia maritima	1	0	0	0	0	0	0	0
Descurainia sophiades	1	0	0	0	2	0	0	0
Descurainia pinnata	2	0	0	0	4	0	0	0
Lepidium sativum	0	0	0	0	2	0	0	0
Rorippa islandica	1	0	0	0	4	0	0	0
Leavenworthia alabamica	1	0	0	0	3	0	0	0
Diptychorcarpus strictus	0	2	0	3	0	0	0	0
Euclidium syriacum	0	3	0	1	0	0	0	0
Aethionema arabicum	0	0	0	0	0	0	0	1
Tarenaya hassleriana	0	0	0	0	0	0	0	3
Cleome gynandra	0	0	0	0	0	0	0	2
Cleome violacea	0	0	0	0	0	0	0	3
Theobroma cacao	0	0	0	0	0	0	0	0
Citrus sinensis	0	0	0	0	0	0	0	0
Vitis vinifera	0	0	0	0	0	0	0	0

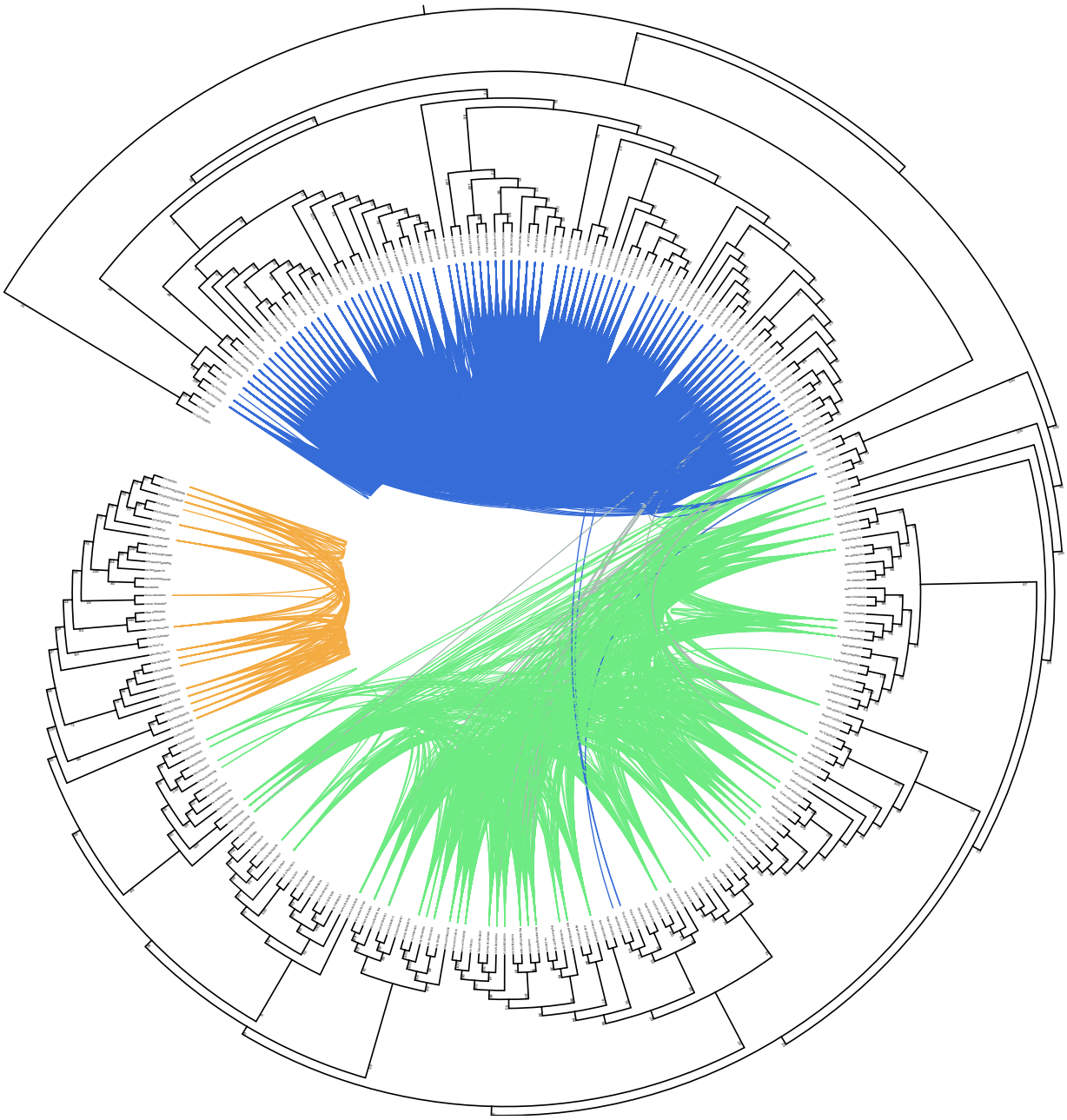
Supplementary Figure 3.4. Here we show the full domain clade distribution of MAM genes across the genomes, regardless of synteny or genomic position. This data was used ultimately to place the points of innovation for different MAM types in Figure 3.4.



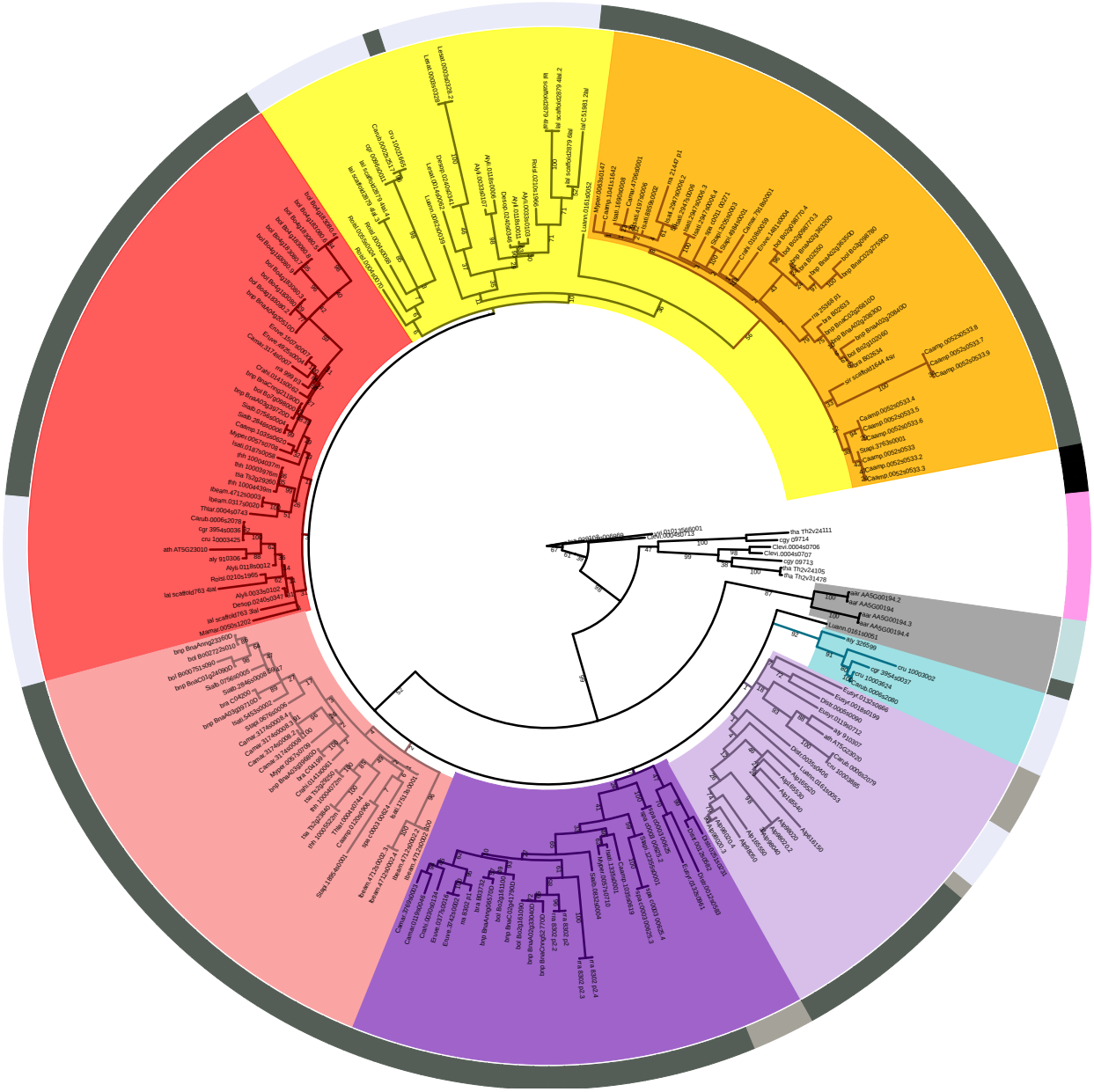
Supplementary Figure 3.5. MAM protein sequences were divided into before domain, domain, and after domain segments and each significantly different section of the MAMa or MAMb genes from lineage I were compared to corresponding MAMe sections.



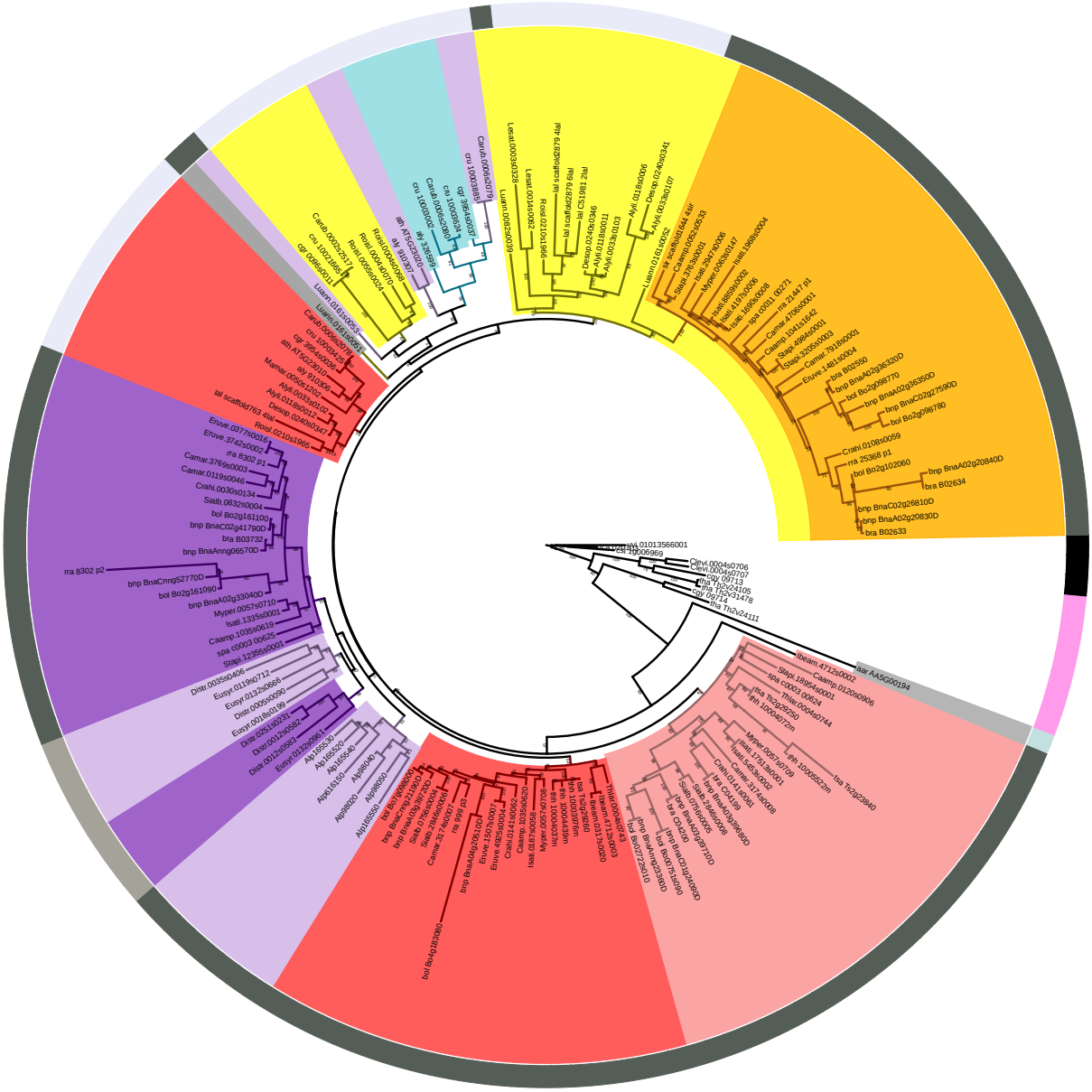
Supplementary Figure 3.6. Amino acid sequence comparisons at 80% sequence similarity. (A) Colored rectangles indicate specific biochemical functions as described by Kumar et al. 2019 in *Brassica juncea*. Green - metal binding sites; Yellow - catalytic sites; Red - 2-oxo acid binding sites; Blue - CoA binding sites. (B) Summarizes all sites with a uniquely divergent amino acid to quantify the significance of domain divergence.



Supplementary Figure 3.6. Full gene family phylogeny with bootstrap scores at 1000 bootstraps with syntenic clusters mapped. Used in Figure 3.1. May also be accessed via: http://bit.ly/2tHVgYK.



Supplementary Figure 3.7. Domain tree phylogeny with clades colored and bootstrap scores at 1000 bootstraps. Used in Figure 3.3. May also be accessed via: http://bit.ly/2Hb5jIS.



Supplementary Figure 3.8. Full gene family phylogeny with bootstrap scores at 1000 bootstraps with clades colored. Used in Figure 3.3. May also be accessed via: http://bit.ly/37btHEZ.

Supplemental Data Set 1

Downloaded datasets

Scientific Name	Gene Prefix	N50	Genome Source
Aethionema arabicum	aar	10.1M	CoGe_ID:34234
Alyssum linifolium (Descurainia pinnata)*	Aly	0.8M	JGI - DOI:10.25585/1488060 v1
Arabidopsis lyrata	aly	24.5M	phytozome v2.1
Arabidopsis thaliana	ath	23.5M	phytozome TAIR10
Arabis alpina	Alp	28.3M	CoGe_ID:34227
Brassica napus	bnp	45.9M	Genoscope v5
Brassica oleracea	bol	0.85M	Ensemble v2.1
Brassica rapa	bra	28.5M	phytozome v1.3
Cakile maritima	Cam	0.085M	JGI - DOI:10.25585/1488060 v1
Capsella grandiflora	cgr	0.1M	phytozome v1.1
Capsella rubella	Car	15M	JGI - DOI:10.25585/1488060 v1
Capsella rubella2	cru	15.1M	phytozome v1.0
Caulanthus amplexicaulis var. barbarae	Caa	3.8M	JGI - DOI:10.25585/1488060 v1
Citrus sinesis	csi	1.69M	phytozome v1.1
Cleome gynandra	cgy	0.45M	CoGe_ID:23319
Cleome violacea	Cle	2.15M	JGI - DOI:10.25585/1488060 v1
Crambe hipsanica	Cra	0.3M	JGI - DOI:10.25585/1488060 v1
Descurainia sophioides	Des	1.83M	JGI - DOI:10.25585/1488060 v1
Diptychocarpus strictus	Dis	3.97M	JGI - DOI:10.25585/1488060 v1
Eruca vesicaria	Eru	0.15M	JGI - DOI:10.25585/1488060 v1
Euclidium syriacum	Eus	5.65M	JGI - DOI:10.25585/1488060 v1
Iberis amara	Ibe	0.1M	JGI - DOI:10.25585/1488060 v1
Isatis tinctoria	Ise	0.085M	JGI - DOI:10.25585/1488060 v1
Leavenworthia alabamica	lal	0.07M	brassicadb.org
Lepidium sativum	Les	2.4M	JGI - DOI:10.25585/1488060 v1
Lunaria annua	Lua	0.58M	JGI - DOI:10.25585/1488060 v1
Malcomia maritima	Mam	0.83M	JGI - DOI:10.25585/1488060 v1
Myagrum perfoliatum	Myp	1.3M	JGI - DOI:10.25585/1488060 v1
Raphanus raphanistrum	rra	0.01M	CoGe_ID:25862
Rorippa islandica	Roi	3.1M	JGI - DOI:10.25585/1488060 v1
Schrenkiella parvula	spa	9.6M	brassicadb.org v7.0
Sinapis alba	Sia	0.18M	JGI - DOI:10.25585/1488060 v1
Sisymbrium irio	sir	0.14M	brassicadb.org
Stanleya pinnata	Sta	0.087M	JGI - DOI:10.25585/1488060 v1
Tarenaya hassleriana	tha	1.26M	CoGe_ID:23455
Thellungiella halophila	thh	8M	brassicadb.org
Thellungiella salsuginea	tsa	0.4M	brassicadb.org
Theobroma cacao	tca	34.4M	phytozome v1.1
Thlaspi arvense	Thl	1.1M	JGI - DOI:10.25585/1488060 v1
Vitis vinifera	vvi	3.43M	phytozome Genoscope12X

Appendix A: Genetic variation, environment and demography intersect to shape Arabidopsis defense metabolite variation across Europe

Ella Katz[1], Jia-Jie Li[1], Benjamin Jaegle[2], Haim Ashkenazy[3], **Shawn R Abrahams**[4], Clement Bagaza[5], Samuel Holden[5], Chris J Pires[4], Ruthie Angelovici[5], Daniel J Kliebenstein[1,6]

1 Department of Plant Sciences, University of California, Davis, Davis, United States; 2 Gregor Mendel Institute, Austrian Academy of Sciences, Vienna Biocenter (VBC), Vienna, Austria; 3 Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany; 4 Division of Biological Sciences, Bond Life Sciences Center, University of Missouri, Columbia, United States; 5 Division of Biological Sciences, Interdisciplinary Plant Group, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, United States; 6 DynaMo Center of Excellence, University of Copenhagen, Frederiksberg, Denmark

Please cite the published work here:

Katz, Ella, Jia-Jie Li, Benjamin Jaegle, Haim Ashkenazy, **R. Shawn Abrahams**, Clement Bagaza, Samuel Holden, J. Chris Pires, Ruthie Angelovici, and Daniel J. Kliebenstein. "Genetic variation, environment and demography intersect to shape Arabidopsis defense metabolite variation across Europe." *Elife* 10 (2021): e67784.

Authorial Contribution: Analysis of the MAM/IPMS Gene family with the inclusion of non- "Columbia" *Arabidopsis thaliana* MAM alleles.

ABSTRACT

Plants produce diverse metabolites to cope with the challenges presented by complex and ever-changing environments. These challenges drive the diversification of specialized metabolites within and between plant species. However, we are just beginning to understand how frequently new alleles arise controlling specialized metabolite diversity and how the geographic distribution of these alleles may be structured by ecological and demographic pressures. Here, we measure the variation in specialized metabolites across a population of 797 natural *Arabidopsis thaliana* accessions. We show that a combination of geography, environmental parameters, demography and different genetic processes all combine to influence the specific chemotypes and their distribution. This showed that causal loci in specialized metabolism contain frequent independently generated alleles with patterns suggesting potential within-species convergence. This provides a new perspective about the complexity of the selective forces and mechanisms that shape the generation and distribution of allelic variation that may influence local adaptation.

INTRODUCTION

Continuous and dynamic change in a plant's habitat/environment creates a complex system to which a plant must adapt. Central to this adaptation are the production and accumulation of different metabolites ranging from signaling hormones, primary metabolites to a wide array of multi-functional specialized metabolites (Erb and Kliebenstein, 2020; Hanower and Brzozowska, 1975; Hayat et al., 2012; Kim et al., 2012; Kliebenstein, 2004; Malcolm, 1994; Thakur and Rai, 1982; Wolters and Jürgens, 2009; Yang et al., 2000). The complete blend, chemotype, of these metabolites helps to determine the plants' survival and development, but the creation of any blend is complicated by the fact that individual specialized metabolites can have contrasting effects on the plant. For example, individual specialized metabolites can defend the plant against some stressors while simultaneously making the plant more sensitive to other biotic or abiotic stresses (Agrawal, 2000; Bialy et al., 1990; Erb and Kliebenstein, 2020; Futuyma and Agrawal, 2009; Hu et al., 2018; Lankau, 2007; Opitz and Müller, 2009; Uremıs et al., 2009; Züst and Agrawal, 2017). These opposing effects create offsetting ecological benefits and costs for individual metabolites. Integrating these offsetting effects across dynamic environments involves multiple selective pressures that might contribute to shaping the genetic and metabolic variation within a species (Fan et al., 2019; Kerwin et al., 2015; Malcolm, 1994; Sønderby et al., 2010; Szakiel et al., 2011; Wentzell and Kliebenstein, 2008; Züst et al., 2012).

Significant advances have been made in recent decades to identify genetic sources contributing to metabolic variation. A common finding of these studies is that the

metabolic variation within and between species is the result of structural variation at the enzymes responsible for the chemical structures, or variation at the expression levels of these enzymes, which contributes to the quantitative variation in specialized metabolism (Chan et al., 2011; Chan et al., 2010; Fan et al., 2019; Kroymann et al., 2003; Moore et al., 2019; Schilmiller et al., 2012). These structural and regulatory variants and the resulting chemical variation strongly influence plant fitness in response to a broad range of biotic interactions including herbivores and other plant species (Bednarek and Osbourn, 2009; Brachi et al., 2015; Kerwin et al., 2017; Kerwin et al., 2015; Lankau and Kliebenstein, 2009; Lankau and Strauss, 2007; Lankau, 2007). The potential for these genetic variants influencing plant chemical variation is derived from the enhanced proportion of gene duplication in enzyme encoding genes for specialized metabolism, both at the local and whole genome level (Kliebenstein et al., 2001c; Moghe and Last, 2015). Many mechanistic studies of natural variation in specialized metabolism have focused on biallelic phenotypic variation linked to loss-of-function variants. However, it is not clear if biallelic phenotypic variation is created by biallelic genetic causation when investigating a large collection of individuals from wide-ranging populations within a species.

If selective pressures are sufficiently non-linear, it is possible to have repeated and independent generation of structural variants creating the same metabolic variation in processes that are akin to parallel and convergent evolution are used to describe interspecific variation. Specifically, parallel and convergent evolution describe independent evolution of the same trait that differs depending on the beginning state of the organisms. In parallel evolution, the lineages begin from the same state and in parallel

evolve to the same new state, while in convergent evolution the lineages start at different states and independently converge on the same new state (Figure 1). In this context, we are focusing the analogy on the fact that the two processes differ where they begin, same or different state. This raises the possibility for chemical variation within a species to exhibit parallel evolution, wherein independent new haplotypes with identical metabolic consequences arise multiple times from single-core haplotype. Equally it may be possible to find within-species convergent evolution, where genotypes with the same metabolic profile actually contain completely different haplotypes that themselves arose from distinct haplotypic lineages. These genetic processes and interplay between genetics and selection overlap with neutral demographic processes like gene flow. Thus, it is necessary to understand how the intersection of environmental pressure, demography and genomic complexity gives rise to the pattern of metabolic variation across a plant species.

To better understand how genomic variation, demography and environmental pressure shape the variation of specialized metabolism within a species, we used the Arabidopsis glucosinolate (GSL) pathway as a model. GSLs are a diverse class of specialized metabolites that display extensive variation across the order Brassicales, which includes the model plant Arabidopsis (*Arabidopsis thaliana*) (Bakker et al., 2008; Benderoth et al., 2006; Brachi et al., 2015; Chan et al., 2010; Daxenbichler et al., 1991; Halkier and Gershenzon, 2006; Kerwin et al., 2015; Kliebenstein et al., 2001a; Kliebenstein et al., 2001b; Kliebenstein et al., 2001c; Rodman et al., 1981; Rodman, 1980; Sønderby et al., 2010; Wright et al., 2002). GSLs consist of a common core structure with a diverse side chain that determines biological activity in defense, growth, development and abiotic stress resistance (Beekwilder et al.,

2008; Hansen et al., 2008; Hasegawa et al., 2000; Katz et al., 2020; Katz et al., 2015; Malinovsky et al., 2017; Salehin et al., 2019; Yamada et al., 2003). The Arabidopsis-GSL system is an optimal model to study the species-wide processes driving specialized metabolite variation because the identity of the whole biosynthetic pathway is known, including the major causal loci for natural variation (Benderoth et al., 2006; Brachi et al., 2015; Chan et al., 2011; Chan et al., 2010; Hansen et al., 2007; Kliebenstein et al., 2001a; Kliebenstein et al., 2002b; Kliebenstein et al., 2002a; Kroymann and Mitchell-Olds, 2005; Pfalz et al., 2007; Sønderby et al., 2010; Wentzell et al., 2007). These major loci have been proven to influence Arabidopsis fitness and can be linked to herbivore pressure (Brachi et al., 2015; Hansen et al., 2008; Jander et al., 2001; Kerwin et al., 2017; Kerwin et al., 2015; Züst et al., 2012). Beyond the major causal loci, there is also evidence from genome-wide association (GWA) studies for highly polygenic variation in the genetic background that contributes to modulating GSL variation (Chan et al., 2011). The public availability of over 1000 widely distributed accessions with genomic sequences facilitates phenotyping GSL variation across a large spatial scale and analyses of causal haplotypes at the major GSL causal loci.

In Arabidopsis and other Brassicas, the main GSLs are methionine-derived, aliphatic, GSLs. Variation in the structure of aliphatic GSL is controlled by natural genetic variation at three loci: *GS-Elong, GS-AOP* and *GS-OH*. The specific alleles at these three loci combine to determine a predominant chemical structure and define chemically distinct aliphatic GSL chemotypes. In addition to these large-effect loci, there is a large suite of loci that can quantitatively alter the total accumulation and relative

concentrations of GSLs within each chemotype (Brachi et al., 2015; Chan et al., 2011; Chan et al., 2010). *GS-Elong* differentially elongates the methionine side chain by the methylthioalkylmalate synthase enzymes (MAM). The elongation of the side chain by one methylene group is the result of one cycle that includes three steps: deamination of the methionine to create a ω-methylthio-2-oxoalkanoic-acid, condensation of the ω-methylthio-2-oxoalkanoic-acid with acetyl-CoA, and then isomerization and oxidative decarboxylation. The one carbon longer outcome can then undergo additional cycles of elongation (Benderoth et al., 2006; Graser et al., 2000; Kroymann et al., 2001; Textor et al., 2007). In Arabidopsis, MAM2 catalyzes the addition of one carbon to the side chain, creating GSLs with three carbon side chains. MAM1 catalyzes the addition of two carbons to make GSLs with four carbon side chains (Figure 2). MAM3 (also known as MAM-L) catalyzes the additions up to six carbons (Kliebenstein et al., 2001c; Kroymann et al., 2003; Mithen et al., 1995; Textor et al., 2007). The core pathway leads to the creation of methylthio GSL (MT). Then, the MT is converted to a methylsulfinyl (MSO) with a matching number of carbons (Giamoustaris and Mithen, 1996; Hansen et al., 2007). Structural variation at the *GS-AOP* locus leads to differential modification of the MSO by differential expression of a family of 2-oxoacid-dependent dioxygenases (2ODD). The AOP2 enzyme removes the MSO moiety leaving an alkenyl sidechain, while AOP3 leaves a hydroxyl moiety. Previous work has suggested three alleles of *GS-AOP*: the *OHP* allele that expresses only *AOP3* and accumulates terminal OH containing GLS in the leaves and seeds; an alkenyl allele expressing *AOP2* in the leaf and *AOP2* and *AOP3* in the seed leading to solely alkenyl GLS in the leaf and both alkenyl and OH aliphatic GLS in the seed; and a final allele containing a null mutation in

the *AOP2* gene that accumulates MSO GLS in the leaf and enhanced MSO and OH GLS in the seed. (Figure 2; Chan et al., 2010; Kliebenstein et al., 2001b; Kliebenstein et al., 2001c; Mithen et al., 1995). The C4 alkenyl side chain can be further modified by adding a hydroxyl group at the 2C via the GS-OH 2-ODD (Figure 2; Hansen et al., 2008). In spite of the evolutionary distance, independent variation at the same three loci influences the structural diversity in aliphatic-GSLs within Brassica, Streptanthus and Arabidopsis (Kliebenstein and Cacho, 2016; Lankau and Kliebenstein, 2009). For example, the MAMs responsible for C3 GSLs in Arabidopsis and Brassica represent two independent lineages, same as the MAMs responsible for C4 GSLs; in fact, the *MAM* locus contains at least three independent lineages that recreate the same length variation (Abrahams et al., 2020). This indicates repeated evolution across species, but it is not clear how frequently these loci are changing within a single species or how ecological or demographic processes may shape within-species variation at these loci.

In this work, we described GSL variation in seeds of a collection of 797 *A. thaliana* natural accessions collected from different locations mainly in and around Europe. The amounts of GSLs can vary across different tissues and life stages, but there is a strong correlation in the type of aliphatic GSL produced across tissues (Brown et al., 2003; Kliebenstein et al., 2001a; Kliebenstein et al., 2001b; Petersen et al., 2002). Thus, in most cases the chemotype of the seeds is the same as the leaves. The seeds have the highest level of GSLs in Arabidopsis and are stable at room temperature until germination, which makes the seeds a perfect tissue to survey variation. Further, GSLs are known to be important for seed defenses against herbivores and pathogens (Raybould and Moyes, 2001). By combining GSL seed measurements with prior whole-genome

sequencing in a European collection of accessions, we show that all three major causal loci controlling GSL metabolic diversity contain multiple independently derived alleles that recreate the same phenotypes using a combination of single nucleotide polimorphism (SNPs) and structural variation. Using these causal genotypes and chemotypes in combination with their geographic distribution provided evidence that the distribution of GSL metabolic diversity across Europe is influenced by a combination of demography and ecological factors. The ecological relationships to chemotype suggested a potential for variation in selective processes across the geographic regions studied. Future work will be needed to identify the specific biotic and/or abiotic factors shaping this distribution.

RESULTS

GSL variation across Europe

To investigate the genetic, environmental and demographic parameters influencing the distribution of Arabidopsis GSL chemotypes, we measured GSLs from seeds of a collection of 797 *A. thaliana* natural accessions (The 1001 Genomes Consortium, 2016). These Arabidopsis accessions were collected from different geographical locations, mainly in and around Europe (Figure 3A). 23 different GSLs were detected and quantified, identifying a wide diversity in composition and amount among the natural accessions with a median heritability of 83%, ranging from 34% to 93% (Supplementary file 1). To summarize the GSL variation among the accessions, we

performed principal component analyses (PCA) on the accumulation of all the individual GSLs across the accessions as an unbiased first step. The first two PCs only captured 33% of the total variation with PC1 describing GSLs with four and seven carbons and PC2 mainly capturing GSLs with eight carbons in their side chain (Figure 3—figure supplement 1). Previous work using a collection of predominantly central European accessions had suggested a simple continental gradient chain-elongation variation from the south-west (SW) (that is enriched with alkenyl and hydroxyalkenyl GSLs) to the north-east (NE) (Brachi et al., 2015; Züst et al., 2012). To assess if this was still apparent in this larger collection, we plotted the accessions based on their geographical locations and colored them based on their PC1 and PC2 scores that are linked to chain elongation variation (Figure 3A and Figure 3—figure supplement 2A, respectively). This larger collection shows that there is not a single gradient shaping GSL diversity across Europe (Figure 3A). Instead, the extended sampling of accessions around the Mediterranean Basin in this collection shows that the SW to NE pattern reiterates within the Iberian Peninsula. In each of these areas (Iberian Peninsula and Central Europe), the SW is enriched with C4 GSLs, and the NE with C3 GSLs (Figure 3—figure supplement 1).

To test which of the major causal loci are detectable in this collection and to identify new genomic regions that are associated with the observed GSL variation, we performed GWA (with EMMAX algorithms) analyses using the PC1 and PC2 values. This collection of natural accessions presents a dense variant map that is 3× larger than previous GSL GWA mapping populations and includes 6,973,565 SNPs. In spite of the large population size, both PC1- and PC2-based analyses identified the same two major peaks covering two of the known causal gene clusters controlling GSL diversity (Figure

3B for PC1 GWA analyses, Figure 3—figure supplement 2B for PC2 GWA analyses) (Brachi et al., 2015; Chan et al., 2011; Chan et al., 2010). The largest peak in both cases is the *GS-Elong* locus on chromosome 5, containing the *MAM1* (AT5G23010), *MAM2* (that is not present in Col-0 plants) and *MAM3* (AT5G23020) genes.

The peak on chromosome 4 is the *GS-AOP* locus containing the *AOP2* and *AOP3* genes (AT4G03060 and AT4G03050, respectively). Applying a more permissive cutoff did not result in the detection of any other related genes (Supplementary file 2). Previous QTL mapping and molecular experiments have shown that the genes within *GS-AOP* and *GS-Elong* loci are the causal genes for GSL variation within these regions (Benderoth et al., 2006; Brachi et al., 2015; Chan et al., 2011; Chan et al., 2010; Kliebenstein et al., 2001a; Kliebenstein et al., 2002a; Kliebenstein et al., 2002a; Kroymann and Mitchell-Olds, 2005; Pfalz et al., 2007; Wentzell et al., 2007). Surprisingly, none of the other known natural variants within the GSL biosynthetic pathway (listed in Supplementary file 2) were identified by GWA including three that were found with 96 accessions and three that were found with 595 accessions using PC1 and 2 (Brachi et al., 2015; Chan et al., 2011; Chan et al., 2010; Kliebenstein, 2009). Performing GWA studies using the accumulation of each of the 23 individual GSL detected in this collection resulted in an identical result, no additional known GSL-related genes were detected, while a few additional unknown genes were found (Figure 3—figure supplement 3 and Supplementary file 2). One explanation for that is that the dense sampling in this collection is available for mainly the Iberian Peninsula, the southern coast of Sweden and the south-western coast of Italy, and is still insufficient for Central

Europe. Another possibility is that allelic heterogeneity for the other loci, and more complex patterns of interaction, may hamper their detection and influenced this high false-negative error rate where ~80% of prior validated natural variants found using multiple RIL populations were missed.

Complex GSL chemotypic variation

One potential complicating factor is that GSL chemotypic variation is best described as a discrete multimodal distribution involving the epistatic interaction of multiple genes which PCA's linear decomposition cannot accurately capture (Figure 2). To test if PCA was inaccurately describing GSL chemotypic variation, we directly called the specific GSL chemotypes in each accession. Using Arabidopsis QTL mapping populations and GWA, we have shown that the *GS-AOP*, *Elong* and *OH* loci determine seven discrete chemotypes, 3MSO, 4MSO, 3OHP, 4OHB, Allyl, 3-Butenyl, 2-OH-3-Butenyl (Figure 2), that can be readily assigned from GSLs' phenotypic data (Brachi et al., 2015; Chan et al., 2011; Chan et al., 2010; Kliebenstein et al., 2001a). The presence and amounts of these seven chemotypes provide a reliable indication about the existence and activity of each of the major GSL loci. Using accessions with previously known chemotypes and genotypes, we developed a phenotypic classification scheme to assign the chemotype for each accession (Figure 4; for details, see Methods and Figure 4—figure supplements 1–3; for structures, see Figure 2 and Supplementary file 1). Since the aliphatic GSLs' composition in the seeds reliably indicates the GSL structural composition in the other plant's life stages and tissues, assigning a chemotype for each accession based on the seeds' composition is expected to be highly stable across tissues

of the same accession (Brown et al., 2003; Chan et al., 2011; Chan et al., 2010; Kliebenstein et al., 2001a; Kliebenstein et al., 2001b). Most accessions were classified as 2-OH-3-Butenyl (27%) or Allyl (47%) with lower frequencies for the other chemotypes. Mapping the chemotypes onto Europe showed that the PCA decomposition was missing substantial information on GSL chemotype variation (Figure 4). Instead of a continuous distribution across Europe, the chemotype classifications revealed specific geographic patterns. Central and parts of Northern Europe (like north Germany and Poland) were characterized by a high variability involving the co-occurrence of individuals from all chemotypes. In contrast, southern Europe, which presents a dense sampling, including the Iberian Peninsula, Italy and the Balkan, has two predominant chemotypes, Allyl or 2-OH-3-Butenyl, that are separated from each other by a clear and sharp geographic partitioning (Figure 4 and Figure 4—figure supplement 4). Uniquely, Swedish accessions displayed a striking presence of almost solely Allyl chemotypes. Deeper sampling is required to test if this is or is not mirrored on the eastern side of the Baltic Sea as the few accessions from that region are almost solely 3OHP chemotypes (Finnish, Lithuanian, Latvian or Estonian accessions). Directly assigning GSL variation by discrete chemotypes provided a more detailed image not revealed by PCA decomposition. Further, the different chemotypic to geographic patterns suggest that there may be different pressures shaping GSL variation particularly when comparing Central and Southern Europe.

Geography and environmental parameters affect GSL variation

Because GSL chemotypes may be more reflective of local environment, we proceed to test if they are associated with weather parameters and landscape conditions. Further, given the difference in chemotype occurrence in Central and Southern Europe, we hypothesized that these environmental connections may change between Central and Southern Europe. For these tests, we chose environmental parameters that capture a majority of the environmental variance and by that may describe the type of ecosystem (Ferrero-Serrano and Assmann, 2019). We assigned each accession the environmental value based on its location. These environmental parameters include geographic proximity (distance to the coast), precipitation descriptors (precipitation of wettest and driest month) and temperature descriptors (maximal temperature of warmest month and minimal temperature of coldest month) and capture major abiotic pressures as well as provide information about the type of ecosystem in which each accession exists. Because demography and environment can be confounded, we included demography in our models using the previously assigned genomic groupings as components of the model (The 1001 Genomes Consortium, 2016). Further, we included specific geographic information by assigning the accessions to a northern or a southern collection, based on their location in relation to the following chain of mountains: the Pyrenees, the Alps and the Carpathians (Figure 4—figure supplement 4). We then ran a linear model for each geographic area separately (north and central vs. south) to check if the environmental parameters and the genomic population group associate with specific chemotypes. To directly test for an interaction of environment and geography, we ran the model with all accessions and incorporated the geography parameters and genomic population group. As the most frequent chemotypes in the collection are Allyl and 2-OH-3-Butenyl (Figure 4—

figure supplement 4B), we focused the models on these chemotypes. The models showed that the environmental conditions have different relationships to the chemotypes that shift by geographic areas. Moreover, two of the parameters (min temp of coldest month and precipitation of wettest month) have a significant interaction with geography, suggesting that the relationship of these environmental parameters to specific GSL chemotypes is different between Northern and Southern Europe (Table 1; for details on the models, see Methods). This suggests that the relationship of GSL chemotype to environmental parameters varies across geographic regions of Europe rather than fitting a simple linear model.

As the two main chemotypes in the collection differ by the length of the carbon chain (C3 for Allyl, C4 for 2-OH-3-Butenyl), we created a linear model to further check the interaction between each environmental condition to geography in respect to the carbon chain length. As was shown by the chemotypes models, most of the environmental parameters (min temp of coldest month, precipitation of wettest month and distance to the coast) significantly interacted with geography, showing again that the relationship of environment to GSL alleles changes across Europe (Figure 4—source data 1; for details on the models, see Methods). Conducting this analysis for each of the geographic areas separately highlighted this by showing that these parameters have different effects on the carbon chain length in each of the areas (Figure 4—source data 1).

The genetic architecture of GSL variation

The presence of different GSL chemotype to environmental relationships across Europe raises the question of how these chemotypes are generated. Are these chemotypes from locally derived alleles or obtained by the intermixing of widely distributed causal

alleles? Further, if there are multiple alleles, do they display within-species convergent or parallel signatures? We focus on the *GS-AOP*, *GS-Elong* and *GS-OH* loci, the causal genes creating Arabidopsis GSL chemotypes, and use the available genomic sequences in all these accessions to investigate the allelic variation in these genes to map the allelic distribution and test the potential for convergent and/or parallel evolution within each locus.

GS-Elong: Because the variation in the *GS-Elong* locus is caused by complex structural variation in *MAM1* and *MAM2* that is not resolvable using the available data from short-read genomic sequence, we used the *MAM3* sequence within this locus to ascertain the genomic relationship of accessions at the causal *GS-Elong* locus (Kroymann et al., 2003). We aligned the *MAM3* sequence from each of the accessions, rooted the tree with the *Arabidopsis lyrata* orthologue (*MAMb*) and colored the tree tips based on the accessions-dominant chemotype.

The accessions were distributed across eight distinctive clades with each clade clustering accessions having either a C3 or C4 phenotype (Figure 5A and Figure 5—figure supplement 1 for bootstrap support). The clades C3/C4 status altered across the tree with three of the clades expressing C3 (*MAM2*) and five clades expressing the C4 (*MAM1*). The use of *MAM3* clades as proxy for C3/C4 GLS chemotypes is supported by prior genomic sequencing of the *GS-Elong* region from 15 accessions (Figure 5B; Kroymann et al., 2003). To test for potential within locus recombination that may influence the overarching patterns, we compared the *MAM3* tree to a tree obtained using *MYB37*, which is on the opposite end of the *MAM* locus from *MAM3* (Figure 5—figure supplement 1F). We found that while the order of the clades in the *MYB37* tree is

different than their order in the *MAM3* tree, the accessions' classification to clades was similar among the two trees. This suggests that while there are potentially individual instances of within-locus recombination they are not influencing the overall genotype to chemotype linkage from *MAM3*, and *MAM3* can be used as a reliable reflection of the structural variation in this locus.

Six of the clades in the *MAM3* tree include accession/s with a previously sequenced *MAM* locus (Figure 5B; Kroymann et al., 2003), while two clades (clades 6 and 7) did not include any accession with a previously determined structure. We obtained long-read-based sequencing of 11 additional accessions from the 1001 Genome project for the *MAM* locus that included accessions in all clades including clades 6 and 7 (Figure 5—figure supplement 2—source data 1 for sequences). This showed that clade 6 are accessions that have a haplotype that contains a previously described chimeric *MAM* gene that combines the 5′ of *MAM2* with the 3′ of *MAM1* (Figure 5B and Figure 5—figure supplement 2; Benderoth et al., 2006; Kroymann et al., 2003). In these accessions, the chimeric gene leads to predominantly C3 GSLs. Clade 7 has a haplotype that is highly similar to clade 2 with a single copy of *MAM1* leading to C4 GSLs. Comparing transposable elements in the two clades shows that they are different configurations.

The new sequenced accessions present in the clades with existing genomic haplotypes predominantly agreed with these previously published haplotypes. There were only three accessions with differences, two with a local duplication of a truncated *MAM1* pseudogene in clades 1 and 2 (PHW-34 and TAL 07, respectively), and a second with a local duplication of a *MAM1* pseudogene in clade 2 (Qar-8a, Figure 5—figure supplement 2).

The bootstrap support and smaller trees raised the possibility that clade 2 could be considered as two distinct clades (Figure 5—figure supplement 1, clades 2a and 2b). The chemotypes and haplotype in the accessions do not provide a clear mechanistic basis for separating this clade into two (Kroymann et al., 2003; Figure 5). Comparing the accessions across the main split in this clade suggested that one group of accessions (clade 2b) has lower total GSLs and a higher fraction of short-chain GSLs in comparison to the longer chain structures. Future work involving populations solely focused on this question would be needed to resolve the mechanistic basis of this difference and if this represents two distinct *MAM* loci.

One complication in interpreting the potential for parallel vs. convergent evolution in this locus is that the relationship between the major chemotype/haplotype groups is not resolvable with very low bootstraps (Figure 5—figure supplement 1). Functional parsimony would suggest that clade 4, by having both *MAM1* and *MAM2*, may represent a single haplotype that can give rise to the other functional haplotypes via independent mutations akin to parallel evolution. Supporting this potential is the observation that *MAM1* and *MAM2* are likely derived via a tandem duplication with ensuing divergence since the separation from *A. lyrata* (Figure 5—figure supplement 3; Benderoth et al., 2009; Benderoth et al., 2006). Fully resolving this would require collecting more accessions to identify additional alleles that may contain the information necessary to better resolve the relationships amongst the haplotypes.

Using this phylogeny, we investigated the presence of the different *GS-Elong* haplotypes across Europe to ask if each region has a specific allele/clade or if the alleles are distributed across the continent. Specifically, we were interested if the strong

C3/C4 partitioning in Southern Europe was driven by the creation of local alleles or if this partitioning might contain a wide range of alleles. If the latter is true, this can argue for a selective pressure shaping this C3/C4 divide. To understand the patterning of the C3/C4 haplotypes and chemotypes in Iberia, we plotted the accessions on the map and colored them based on their *GS-Elong* clade (Figure 5C and Figure 5—figure supplement 4). As expected given that genetic variation in Iberia results from a series of range expansions from Central Europe and Africa (Lee et al., 2017; Durvasula et al., 2017), there is extensive mixing of nearly all major European *GS-Elong* haplotypes in Iberia, except of clade 3 that is not present. In contrast, there is a sharp partitioning between the C3/C4 chemotypes created by these haplotypes. The strong geographic separation between the two chemotypes involving nearly all causal haplotypes (Figure 5—figure supplement 4) raises the possibility that the strong geographic partitioning of the C3/C4 chemotypes in Iberia may be driven by selective pressure enhancing the partitioning of the chemotypes, and not solely neutral demographic processes. The presence of a few accessions in Iberia that disagree with the sharp C3/C4 partition (Figure 5—figure supplement 4A) suggests that a new configuration of this loci arose in this area and is reflected in a few accessions. However, this requires further assessment.

Shifting focus to all of Europe showed that while most clades were widely distributed across Europe there were a couple of over-arching patterns (Figure 5C and Figure 5—figure supplement 5). *GS-Elong* clades 1 and 6 provide an example of potential gene flow between Iberia and Central Europe. In contrast, the absence of clade 3 in Iberia is more parsimonious with this haplotype having a glacial refugium in the Balkans followed by a northward flow wherein it mixed with the other clades. Other

clades do not present evidence of a gene flow to the north as they are exclusive to the south as shown by clades 5 and 7. While these are both C4 clades, other C4 clades like clades 2 and 8 present a case of a gene flow to the north (Figure 5—figure supplement 5). This suggests that there are either differences in their GSL chemotype influencing their distribution or there are neighboring genes known to be under selection in Arabidopsis like *FLC* (AT5G10140) that may have influenced their distribution. In combination, this suggests that a complex demography is involved in shaping the chemotype's identity with some regions, Iberia, showing evidence of local selection while other regions, Central Europe, possibly showing a blend requiring further work to delineate (Figure 5—figure supplement 5).

GS-AOP: Side chain modification of the core MSO GSL is determined by the *GS-AOP* locus. Most of the accessions contain a copy of *AOP2* and a copy of *AOP3*, but only one of them will be functionally expressed (Chan et al., 2010), while in some cases both will be non-functional. To better understand the demography and evolution of the *GS-AOP* locus, we separately aligned the *AOP2* and *AOP3* sequences, rooted each tree with the *A. lyrata* orthologue and colored the trees tips based on the accessions-dominant chemotype (Figure 6—figure supplement 1).

The phylogenetic trees shared a very similar topology, yielding a clear separation between alkenyl (*AOP2* expressed) and hydroxyalkyl (*AOP3* expressed) accessions. Alkenyl expressing accessions like Cvi-0 with an expressed copy of the AOP2 enzyme formed a single continuous cluster (Figure 6A and Figure 6—figure supplement 1). In contrast, hydroxyalkyl (*AOP3* expressed) accessions clustered into two separate groups with one group of 3OHP-dominant accessions partitioning from the rest of the accessions

at the most basal split in the tree (Figure 6—figure supplement 1A, *AOP2* tree). This haplotype is marked by having an inversion swapping the *AOP2* and *AOP3* promoters as shown in bacterial artificial chromosome sequencing of the Ler-0 accession (Figure 6D; Chan et al., 2010). The *AOP3* tree also identified a second group of 3OHP-dominant accessions located among the alkenyl accessions. Analyzing the sequences of these accessions reveals that this small group of 3OHP accessions has a complete deletion of *AOP2* and contains only *AOP3* (Figure 6E). Thus, there are at least two independent transitions from alkenyl to hydroxyalkyl GSLs within Arabidopsis, neither of which are related to the alkenyl to hydroxyalkyl conversion within *A. lyrata*. This indicates that there are multiple alkenyl to hydroxyalkyl GSL conversions both within and between Arabidopsis species.

The null accessions (MSO-dominant chemotypes) were identifiable in all the major clades on the tree (Figure 6—figure supplement 1, middle column of heatmap), suggesting that there are independent LOF mutations that abolish either *AOP2* or *AOP3*. Deeper examination of the sequences of these accessions identified three convergent LOF alleles leading to the MSO chemotype. Most of the null accessions harbor a 5 bps deletion in their *AOP2* sequence, which causes a frameshift mutation. This mutation arose within the alkenyl haplotype and was first reported in the Col-0 reference genome (Figure 6B; Kliebenstein et al., 2001c). In addition, there are additional independent LOF events arising in both the alkenyl haplotype (e.g., Sp-0, Figure 6C) and within the Ler-0 inversion haplotype (e.g., Fr-2, Figure 6F). Thus, *GS-AOP* has repeated LOF alleles arising within several of the major *AOP* haplotypes, suggesting convergent evolution of the MSO chemotype out of both the alkenyl and hydroxyalkyl chemotypes.

Using the combined chemotype/genotype assignments at *GS-AOP*, we investigated the distribution of the alleles across Europe. The alkenyl haplotype is spread across the entire continent. In contrast, the hydroxyalkyl haplotypes are geographically more restricted. The Ler-like 3OHP haplotype is present in only Central and North Europe (Figure 6D), while the other 3OHP haplotype, possessing only *AOP3*, is limited to Azerbaijan, along the Caspian Sea (Figure 6E). In contrast to the distinct hydroxyalkyl locations, the distribution of the independent LOF null haplotypes overlaps with all of them being located within Central and North Europe (Figure 6B, C and F). The fact that these independently derived LOF alleles are all contiguous suggests that they may be beneficial or neutral in Central Europe.

GS-OH: The final major determinant of natural variation in Arabidopsis GSL chemotype is the GS-OH enzyme that adds a hydroxyl group to the carbon 2 on 3-butentyl GSL to create 2-OH-3-Butenyl GSL. Previous work had suggested two *GS-OH* alleles measurable in the seed, a functional allele in almost all accessions and a non-functional allele caused by active site mutations represented by the Cvi-0 accession (Hansen et al., 2008). Because of functional epistasis, we can only obtain functional phenotypic information from accessions that accumulate the GS-OH substrate, 3-Butenyl GLS. This identified 11 accessions with a non-functional GS-OH. Surveying these 11 accessions in the polymorph database (The 1001 Genomes Consortium, 2016) identified multiple independent LOF events. One of these 11 accessions has the Cvi active site mutations, two accessions have a shared nonsense SNP that introduces premature stop codons and two accessions have a complete loss of this gene (Table 2). The other six

accessions with a loss of enzyme activity had an unidentified lesion due to sequence quality for this locus.

All these independent *GS-OH* LOF alleles are found in accessions that do not accumulate 3-Butenyl GSL, for example, three carbon or non-alkenyl accessions, suggesting that functional epistasis may be influencing the maintenance of these alleles in nature. Thus, we searched for the accessions that do not accumulate 3-Butenyl GLS and carry *GS-OH* LOF events (Table 2). In all cases, the LOF allele is more frequent in the non four carbon-alkenyl accessions than expected by random chance. This suggests that there is selection against 3-Butenyl GSL synthesis since LOF alleles are more frequent when the *GS-OH* gene is cryptic by functional epistasis. This agrees with the fact that the 3-Butenyl chemotype is the most sensitive to generalist lepidopteran herbivory (Hansen et al., 2008). Thus, these mutations may represent ongoing pseudogenization of the *GS-OH* gene when it is functionally hidden by epistasis at the *GS-AOP* and *GS-Elong* loci. These LOF events would then only be displayed upon rare admixture with 2-OH-3-Butenyl accessions.

DISCUSSION

Understanding the genetic, demographic and environmental factors that shape variation within a trait in a population is key to understanding trait evolution. In this work, we used a family of specialized metabolites, aliphatic GSLs, and measured their amounts in seeds of *A. thaliana* to query how genetics, geography, environment and demography intersect to shape chemotypic variation across Europe. We found that environmental conditions, together with geography, affect the presence and distribution

of chemotypes within the accessions. This was demonstrated by specific traits that were associated with specific environmental conditions, and this association was shifted across the continent. Comparing the associations of traits to specific environmental conditions in Central Europe versus the south revealed different behaviors. This demonstrated that chemotypic variation across Europe is created by a blend of all these processes that differ at the individual loci. This implies that a simultaneous analysis of both genotype and phenotype is required to fully interpret these processes and relationships. The above analysis is extensively using abiotic factors because of their availability while aliphatic GSLs have mainly been linked to biotic interactions. GSLs have been mainly linked to influencing biotic interaction with herbivores considered the primary drivers shaping GSL genetic diversity. However, because climate drives the distribution of biotic factors like herbivores, it is likely that climatic factors might appear as indirectly associated with GSLs. Interestingly, an aliphatic GSL was recently mechanistically linked to drought resistance in Arabidopsis, suggesting a potential role for abiotic factors to influence GSL diversity (Salehin et al., 2019). More work is needed to dissect all the potential components of an environment that influence selection on GSL chemotypes.

All three major aliphatic GSL loci display extensive allelic heterogeneity that is shaped by a blend of evolutionary events at each locus reminiscent of either parallel or convergent evolution. For this analogy, we are defining parallel evolution to be when a new chemotype arises two or more independent times by independent mutations from shared ancestral haplotype. Conversely, we are considering convergent evolution of a chemotype to occur when independent mutations in independent ancestral haplotypes derive the same chemotype. *GS-OH* provided clear evidence of parallel evolution where a

single functional haplotype gave rise to at least four independent LOF alleles all with similar phenotypic consequence. *GS-AOP* suggested the potential for convergent evolution-style events leading to MSO chemotype arising from LOF events at the *GS-AOP* locus. The first characterized LOF event was in the Col-0 accession that has a 5 bp frameshift indel in the *AOP2* gene arising within the alkenyl *AOP2*-dominant *GS-AOP* haplotype. In this study, we identified additional parallel *AOP2* LOF events in this haplotype. More critical, we could identify multiple independent LOF events arising in the *AOP3* gene within the *AOP3*-dominant inversion *GS-AOP* haplotype. Thus, the same GSL chemotype, MSO, arises from independent LOF alleles in two different genes, *AOP2* and *AOP3*, that represent two different ancestral haplotypes (Figure 6). Thus, it appears that parallel-style events occur at all the loci and at least at the *GS-AOP* locus there is potential for a convergent-style evolutionary event leading to a single chemotype.

In addition to independent evolutionary events, three-way epistasis is shaping the allelic heterogeneity at these loci and their evolutionary potential. For example, the multiple independent *GS-OH* LOF variants all appear to have arisen in lineages where the *GS-AOP* and *GS-Elong* loci had haplotypes that epistatically combined to block the formation of the but-3-enyl GSL precursor for *GS-OH* (Figure 2). Thus, the parallel evolution of *GS-OH* LOF alleles is epistatically conditioned by *GS-AOP* and *GS-Elong*. A similar epistatic contingency also exists between the two independent *AOP3* alleles of *GS-AOP* (Figure 6D and E) and GS-Elong. Both of *AOP3* alleles of *GS-AOP* are coordinated with C3 haplotypes in GS-Elong. The *AZE* allele of *GS-AOP* is associated with a novel geographically limited *GS-Elong* allele in clade 8 (Figure 5—figure

supplement 5) while the Central European *AOP3* allele is limited to accessions containing the clade 1, 3 or 6 C3 haplotypes of *GS-Elong*. There is also within-locus epistasis in the *GS-Elong* locus wherein a functional MAM1 leads to the creation of C4 GSLs regardless of the functional state of the *MAM2* gene and C3 haplotypes are marked by the loss of *MAM1*. All of these within- and between-loci interactions create a directional arrow for most loci where the haplotypes do not have equal evolutionary potential. For example, the clade 4 haplotype of *GS-Elong* can equally mutate to create a C3 or C4 chemotype because it has both *MAM1* and *MAM2*. However, the remaining clades like clade 3 have lost one or the other gene limiting their ability to create alternative chemotypes. In this case, the loss of *MAM1* likely prevents the ability for this C3 lineage to recreate the C4 chemotype. Thus, the potential evolutionary trajectory of a haplotype/allele at one or even within one of these GSL loci may be epistatically conditioned by the allelic state all the loci within a specific lineage.

This level of allelic diversity at these loci raises a question of how do pathways with this level of diversity and structural variation pass through speciation boundaries (Figure 5—figure supplement 4). Confounding this further is the observation that Brassica ssp. have genetic variation at *GS-Elong* and *GS-AOP* creating the exact same chemotypic variation found in Arabidopsis (Heidel et al., 2006; Ramos-Onsins et al., 2004; Windsor et al., 2005). There is similar within-species (*GS-AOP*) and between-species (*GS-AOP* and *GS-Elong*) variation between the closely related Arabidopsis sister species, *A. lyrata, A. petraea* and *A. halleri*. However, the underlying genetic basis is independent events at the *AOP* and *MAM* loci showing that the variation did not go through the speciation boundary. Instead, this suggests that this variation has been

recreated repeatedly in these species This raises the possibility that there may be a class of loci that are being repeatedly sampled by pangenomic variation across species within a family. To test this possibility would need a deeper phylogenetic sampling within and between species, particularly for understanding the intersection of ecology and evolution (Göktay et al., 2021; Durvasula et al., 2017).

Previous work on other biotic interactions genes like pathogen resistance gene-for-gene loci had indicated a predominant model of having two moderate-frequency ancient alleles creating the phenotypic variation within the species (Atwell et al., 2010; Corrion and Day, 2001; MacQueen et al., 2016). In contrast to R-gene loci characterized by old/stable biallelic variation, the GSL loci are characterized by a blend of structural and SNP-based variation with numerous alleles that appear young. In other cases, alleles of genes involved in biotic defense can present more complex patterns, for example, natural variation in the immune gene *ACCELERATED CELL DEATH 6* (*ACD6*) is caused by a rare allele causing an extreme lesion phenotype. It is not yet clear what selective pressures influence *ACD6* genetic variation (Todesco et al., 2010; Zhu et al., 2018). Thus, loci controlling resistance to diverse biotic traits under natural conditions have diverse genetic architectures and further work is needed to assess the range of allelic heterogeneity in these adaptive loci.

The allelic diversity at the GSL loci illustrates the benefit of simultaneously tracking the phenotype and genotype when working to understand the distribution of trait variation. For example, the Iberian Peninsula and the Mediterranean Basin had low variability in aliphatic GSL chemotypes, which show strong geographic structure. By contrast, Central/North Europe had high aliphatic GSL diversity with chemotypes

showing overlapping geographic distributions. At first glance, this contrasts with previous work showing that the Iberian Peninsula and the Mediterranean Basin are genetically diverse. However, this discrepancy was caused by one of the causal loci. Specifically, the *GS-AOP* locus was largely fixed as the Alkenyl allele in Iberia/Mediterranean Basin with the alternative *GS-AOP* alleles enriched in Central Europe. In contrast to *GS-AOP*, Iberia and the Mediterranean Basin were highly genetically diverse for the *GS-Elong* locus and appear to contain almost all the variation in *GS-Elong* found throughout Europe (The 1001 Genomes Consortium, 2016). Thus, the chemotypic divergence from genomic variation expectations was driven by just the *GS-AOP* locus. This indicates that the high level of chemotypic variation in Central Europe is a blend of alleles that emerged in the south (*GS-Elong*) and alleles that possibly arose locally (*GS-AOP*, both nulls and *AOP3*). Further, the chemotypes found in any one region appear to be created by a combination of alleles as a result of a gene flow across the continent, local generation of new polymorphisms and local selective pressures.

Another challenge potentially caused by allelic heterogeneity and differential selective pressures, as displayed within this system, is detecting the known and validated causal natural variants within a population. Specifically, the GWA with this collection of 797 accessions was unable to find 80% of the known causal loci including one of the three major effect loci, *GS-OH*. Maximizing the number of genotypes and the SNP marker density was unable to overcome the complications imposed by the complex pressures shaping the distribution of these traits, potentially due to unequal dense sampling from the different areas. In this system, the optimal path to identifying the causal polymorphisms has instead been a small number of Recombinant Inbred Line

populations derived from randomly chosen parents. In complex adaptive systems, the optimal solution to identifying causal variants is likely a blend of structured mapping populations and then translating the causal genes from this system to the GWA results and tracking the causal loci directly.

In this work, we combined different approaches to uncover some of the parameters shaping the aliphatic GSL content across Europe. Widening the size of the population will enable us to deepen our understanding on the evolutionary mechanisms shaping a phenotype in a population.

MATERIALS AND METHODS

Plant material

Seeds for 1135 Arabidopsis (*A. thaliana*) genotypes were obtained from the 1001 genomes catalog of *A. thaliana* genetic variation (https://1001genomes.org/). All Arabidopsis genotypes were grown at 22°C/24°C (day/night) under long-day conditions (16 hr of light/8 hr of dark). Two independent replicates were performed, each of them included the full set of genotypes. The replicates obtained from independent maternal plants were grown in randomized fashion. In the analyses, only accessions from Europe and around Europe were included (Figure 3A), resulting in an analysis of 797 accessions. A list of the accessions can be found in Supplementary file 1.

GSL extractions and analyses

GSLs were measured as previously described (Kliebenstein et al., 2001a; Kliebenstein et al., 2001b; Kliebenstein et al., 2001c). Briefly, ~3 mg of seeds

were harvested in 200 µL of 90% methanol. Samples were homogenized for 3 min in a paint shaker, centrifuged, and the supernatants were transferred to a 96-well filter plate with DEAE sephadex. The filter plate with DEAE sephadex was washed with water, 90% methanol and water again. The sephadex-bound GSLs were eluted after an overnight incubation with 110 µL of sulfatase. Individual desulfo-GSLs within each sample were separated and detected by HPLC-DAD, identified, quantified by comparison to standard curves from purified compounds and further normalized to the weight. A list of GSLs and their structure is given in Supplementary file 1A. Raw GSLs data are given in Supplementary file 1B.

Statistics, heritability and data visualization

Statistical analyses were conducted using R software (https://www.R-project.org/) with the RStudio interface (http://www.rstudio.com/). For each independent GLS, a linear model followed by ANOVA was utilized to analyze the effect of accession, replicate and location in the experiment plate upon the measured GLS amount. Broad-sense heritability (Supplementary file 1C) for the different metabolites was estimated from this model by taking the variance due to accession and dividing it by the total variance. Estimated marginal means (emmeans) for each accession were calculated for each metabolite from the same model using the package emmeans (CRAN, 2021a; Supplementary file 1D). PCAs were done with FactoMineR and factoextra packages (Abdi and Williams, 2010). Data analyses and visualization were done using R software with tidyverse (Wickham et al., 2019) and ggplot2 (Kahle and Wickham, 2013) packages.
Maps were generated using ggmap package (Kahle and Wickham, 2013).

Phenotypic classification based on GSL content

For each accession, the expressed enzyme in each of the following families was determined based on the content (presence and amounts) of short-chained aliphatic GSLs. *MAM enzymes*: The total amount of three carbon GSLs and four carbon GSLs was calculated for each accession. Three carbon GSLs include 3MT, 3MSO, 3OHP and Allyl GSL. Four carbon GSLs include 4MT, 4MSO, 4OHB, 3-Butenyl and 2-OH-3-Butenyl GSL (for structures and details, see Supplementary file 1). Accessions that the majority of aliphatic short-chained GSL contained three carbons in their side chains were classified as *MAM2* expressed (Figure 4—figure supplement 1). Accessions that the majority of aliphatic short-chained GSL contained four carbons in their side chains were classified as *MAM1* expressed (Figure 4—figure supplement 1). The accessions were plotted on a map based on their original collection sites (Figure 4—figure supplement 1).

AOP enzymes: The relative amount of alkenyl GSL, alkyl GSL and MSO GSL was calculated in respect to the total short-chained aliphatic GSL as follows:

AlkenylGSL(AOP2expressed)=Allyl +2−OH−3−butenyl + 3−butenylTotal short c hained GSLAlkenylGSL(AOP2expressed)=Allyl +2−OH−3−butenyl + 3−butenylTotal sh ort chained GSL

AlkylGSL(AOP3expressed)=3OHP + 4OHBTotal short chained GSLAlkylGSL(AOP3expressed)=3OHP + 4OHBTotal short chained GSL

MSOGSL(AOPnull)=3MSO + 4MSOTotal short chained GSLMSOGSL(AOPnul l)=3MSO + 4MSOTotal short chained GSL

The expressed AOP enzyme was determined based on those ratios: Accessions with majority alkenyl GSL were classified as *AOP2* expressed. Accessions with majority of alkyl GSL were classified as *AOP3* expressed. Accessions with majority of MSO GSL were classified as *AOP* null. The accessions were plotted on a map based on their original collection sites (Figure 4—figure supplement 2).

GS-OH enzyme: The ratio between 2-OH-3-Butenyl GSL to 3-Butenyl GSL was calculated only for *MAM1*-expressed accessions (accessions that the majority of GSLs contain four carbons in their side chain). Accessions with high amounts of 2-OH-3-Butenyl GSL were classified as *GS-OH* functional. Accessions with high amounts of 3-Butenyl GSL were classified as *GS-OH* non-functional. The accessions were plotted on a map based on their original collection sites (Figure 4—figure supplement 3).

Each accession was classified to one of seven aliphatic short-chained GSLs based on the combination of the dominancy of the enzymes as follows: MAM2, AOP null: classified as 3MSO dominant. MAM1, AOP null: classified as 4MSO dominant. MAM2, AOP3: classified as 3OHP dominant. MAM1, AOP3: classified as 4OHB dominant. MAM2, AOP2: classified as Allyl dominant. MAM1, AOP2, GS-OH non-functional: classified as 3-Butenyl dominant. MAM1, AOP2, GS-OH functional: classified as 2-OH-3-Butenyl dominant. The accessions were plotted on a map based on their original collection sites and colored based on their dominant chemotype (Figure 4).

Environmental and demographic data

Environmental and demographic data (referred to as ‘genomic group’) were obtained from the 1001 genomes website (https://1001genomes.org/, for geographical

and demographic data) and from the Arabidopsis CLIMtools (http://www.personal.psu.edu/sma3/CLIMtools.html, Ferrero-Serrano and Assmann, 2019) for environmental data. We chose the five variables that captured a majority of the variance in this dataset based on PCA using different combinations of variables. The chosen variables are maximal temperature of warmest month (WC2_BIO5), minimal temperature of coldest month (WC2_BIO6), precipitation of wettest month (WC2_BIO13), precipitation of driest month (WC2_BIO14) and distance to the coast (in km). Each one of the above variables (including genomic group) was assigned to each one of the accessions.

Environmental models

Linear models to test the effect of geographical and environmental parameters (Figure 3—figure supplement 1 and Figure 4—source data 1) were conducted using dplyr package (CRAN, 2021b) and included the following parameters:

Figure 3—figure supplement 1 linear models for collection sites: PC score ~ Latitude + Longitude + Latitude * Longitude.

Table 1 and Figure 4—source data 1 for all the data: C length (C3 and C4) or the chemotypes (Allyl and 2-OH-3Butenyl) ~ Genomic group + Geography (north versus south) + Max temperature of warmest month + Min temperature of coldest month + Precipitation of wettest month + Precipitation of driest month + Distance to the coast + Geography * Genomic group + Geography * Max temperature of warmest month + Geography * Min temperature of coldest month + Geography * Precipitation of driest

month + Geography * Precipitation of wettest month + Geography * Distance to the coast.

For the north and the south: C length (C3 and C4) or the chemotypes (Allyl and 2-OH-3Butenyl) ~ Genomic group + Geography (north versus south)+ Max temperature of warmest month + Min temperature of coldest month + Precipitation of wettest month + Precipitation of driest month + Distance to the coast.

Genome-wide association studies

The phenotypes for GWA studies were each accession value for PC1 and 2. GWA was implemented with the easyGWAS tool (Grimm et al., 2017) using the EMMAX algorithms (Kang et al., 2010) and a minor allele frequency (MAF) cutoff of 5%. The results were visualized as Manhattan plots using the qqman package in R (Turner, 2014).

Phylogeny

Genomic sequences from the accessions for *MAM3* – AT5G23020, *AOP2* – Chr4, 1351568 until 1354216, *AOP3* – AT4G03050.2, *GS-OH* – AT2G25450 and *MYB37* – AT5G23000 were obtained using the Pseudogenomes tool (https://tools.1001genomes.org/pseudogenomes/#select_strains).

Multiple sequence alignment was done with the msa package (default settings) in R using the ClustalW, ClustalOmega and Muscle algorithms (Bodenhofer et al., 2015). Phylogenetic trees were generated with the 'ape' package (neighbor-joining tree) (Paradis and Schliep, 2019) and were visualized with ggtree package in R (Yu, 2020). Each tree was rooted by the genes matching *A. lyrata's* functional orthologue or closest homologue.

Bootstrap analyses (Bootstrap = 100) was done with 'ape' package in R (Paradis and Schliep, 2019), with the same tree inference method as described before.
For *MAM3* bootstrap analysis, the accessions with low-quality sequencing were excluded.

Amino acid phylogenies: Sequences were taken from Abrahams et al., 2020, which uses *A. thaliana* Col-0 genome and the *MAM2* amino acid sequence 1006452109 from the Arabidopsis Information Resource (TAIR) database. Alignments were run using MAFFT (Katoh et al., 2017; Kuraku et al., 2013) and cleaned using Phyutility at a 50% occupancy threshold (Smith and Dunn, 2008). RAxML was used for phylogenetic inference (Stamatakis, 2014) with the PROTCATWAG model (Bootstrap = 1000).

Sequencing

PacBio long read-based de novo genome assemblies of the relevant accession were generated as part of the 1001 Genomes Plus project. The genomes were assembled with Canu (v1.71) (Koren et al., 2017) and polished using the long reads followed by a second polishing step with PCR-free short reads.

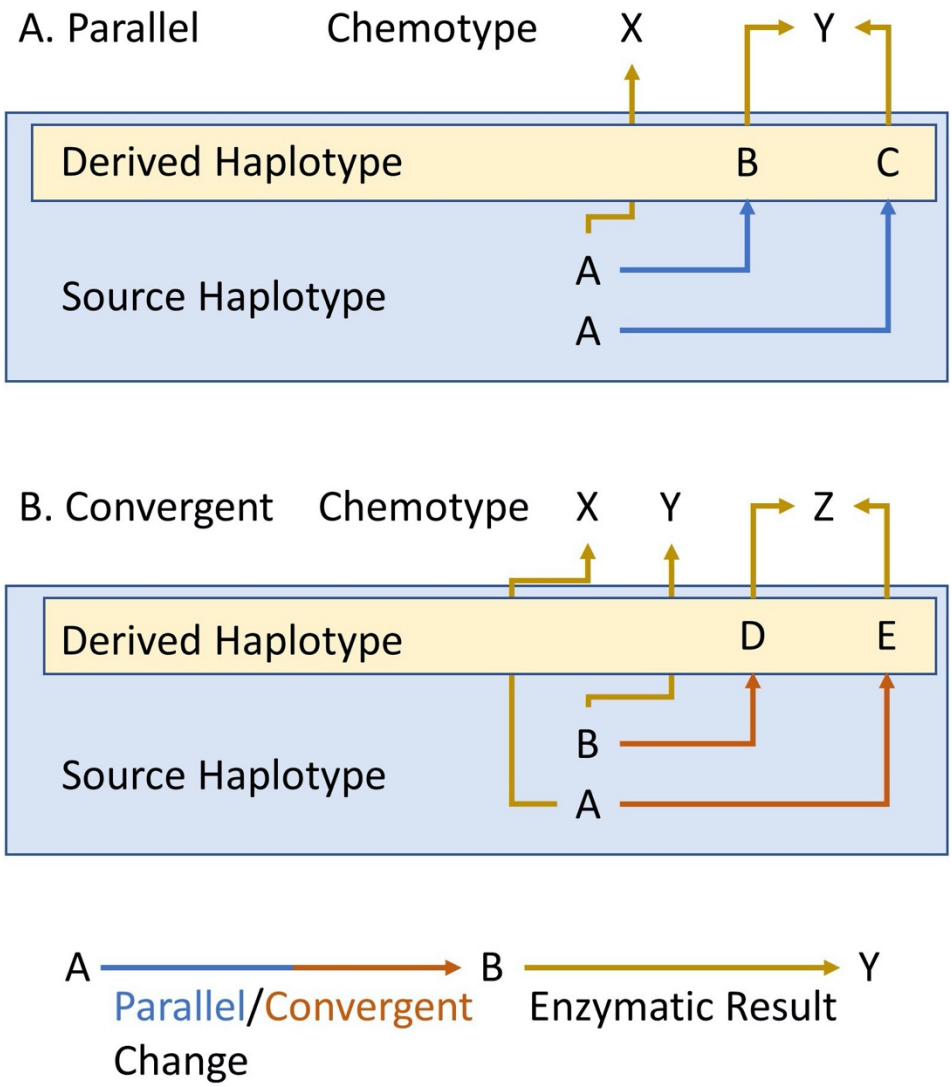


Figure A-1: Parallel and convergent evolution.
The schema describes our use of parallel (**A**) and convergent (**B**) evolution for within-species chemotypic variation. The letters in the blue box represent the state of the source/ancestral haplotypes. The letters within the yellow box represent the newly derived haplotypes that arose by genetic mutation in the source haplotype. Finally, X, Y and Z show the chemotypes that arise from each haplotype. Blue and red arrows represent parallel or convergent genetic changes (respectively), while mustard arrows represent the enzymatic result.

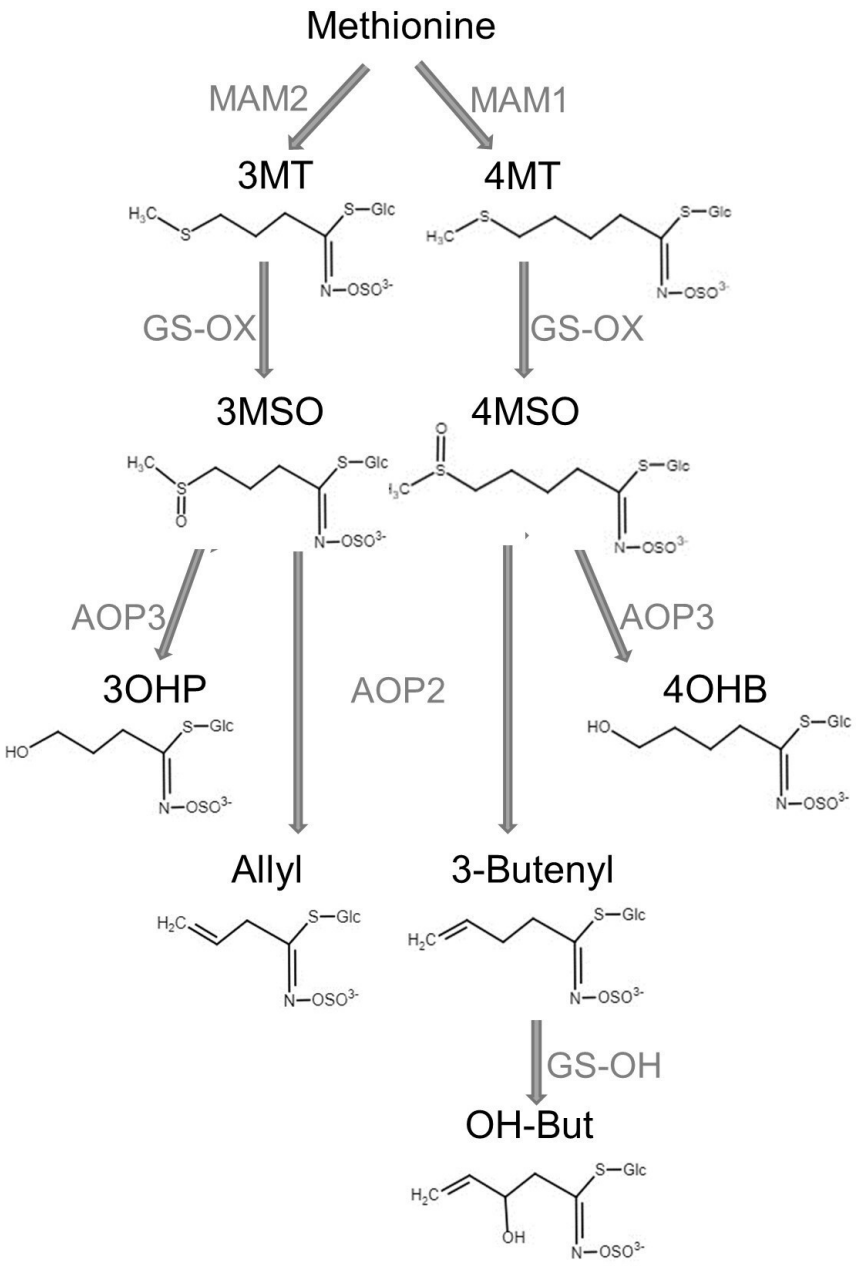


Figure A-2: Aliphatic glucosinolate (GSL) biosynthesis pathway.
Short names and structures of the GSLs are in black. Genes encoding the causal enzyme for each reaction (arrow) are in gray. *GS-OX* is a gene family of five or more genes. OH-But: 2-OH-3-Butenyl.

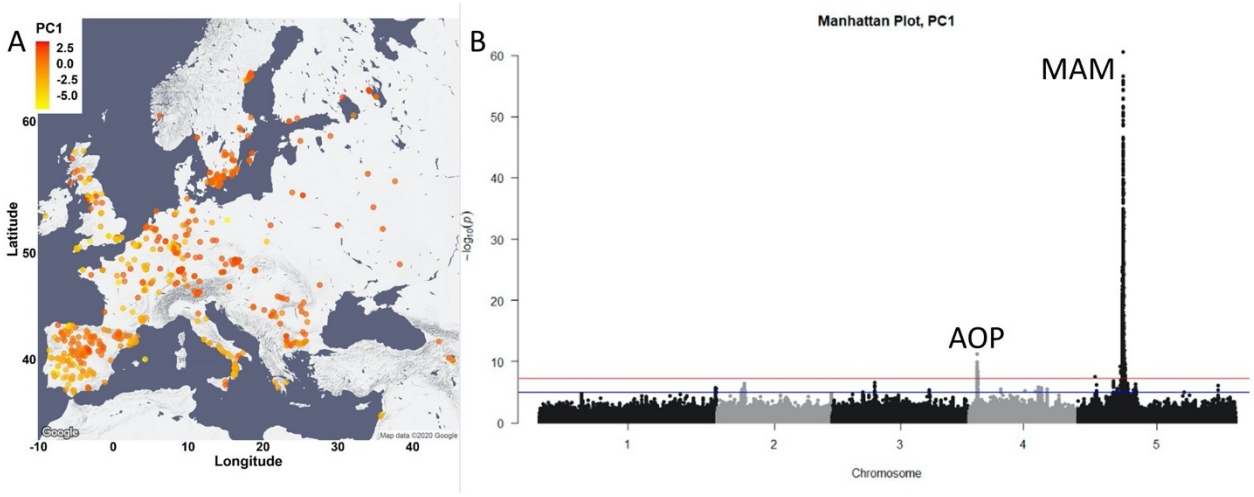


Figure A-3: Glucosinolate variation across Europe is dominated by two loci.
(**A**) The accessions are plotted on the map based on their collection site and colored based on their principal component (PC)1 score. (**B**) Manhattan plot of genome-wideassociation analyses using PC1. Horizontal lines represent 5% significance thresholds using Bonferroni (red) and Benjamini–Hochberg (blue).

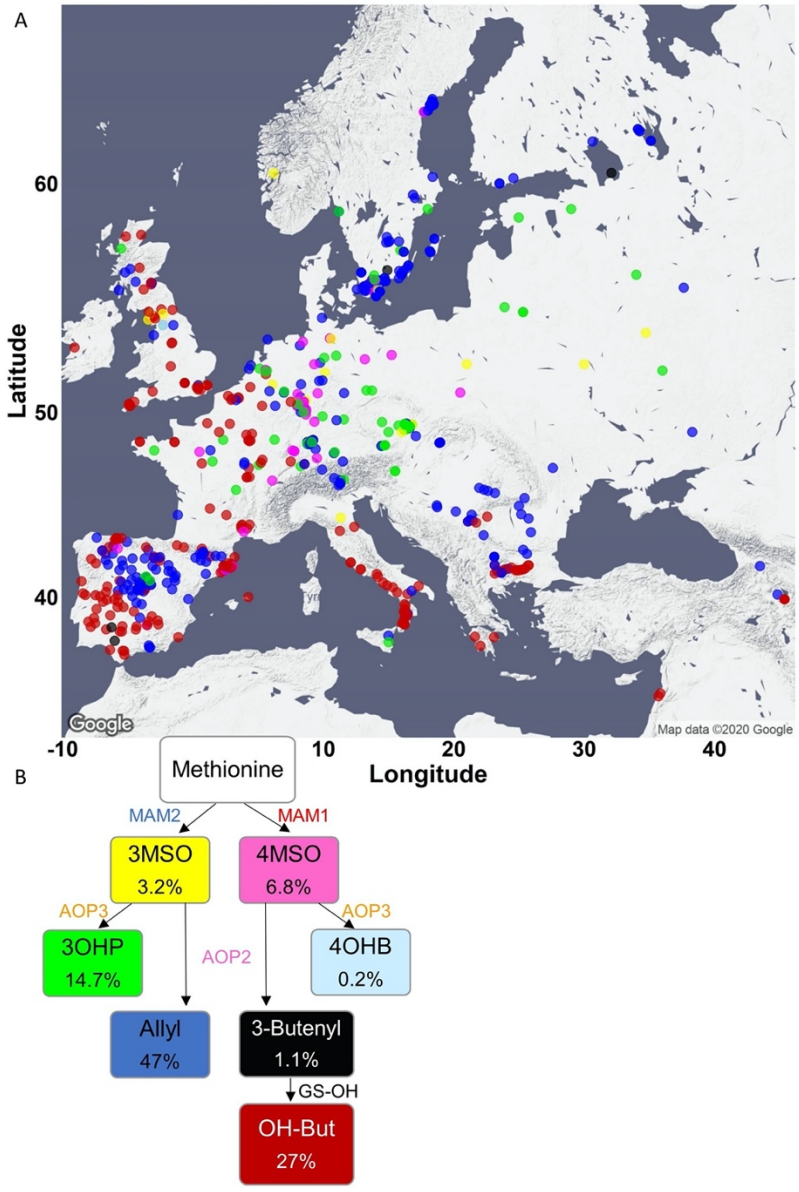


Figure A-4: Phenotypic classification based on glucosinolate (GSL) content.
(**A**) Using the GSL accumulation, each accession was classified to one of seven aliphatic short-chained GSL chemotypes based on the enzyme functions as follows: *MAM2, AOP* null: classified as 3MSO dominant, colored in yellow. *MAM1, AOP* null: classified as 4MSO dominant, colored in pink. MAM2, AOP3: classified as 3OHP dominant, colored in green. *MAM1, AOP3*: classified as 4OHB dominant, colored in light blue. *MAM2, AOP2*: classified as Allyl dominant, colored in blue. *MAM1, AOP2, GS-OH* non-functional: classified as 3-Butenyl dominant, colored in black. *MAM1, AOP2, GS-OH* functional: classified as 2-OH-3-Butenyl dominant, colored in red. The accessions were plotted on a map based on their collection sites and colored based on their dominant chemotype. (**B**) The coloring scheme with functional GSL enzymes in the aliphatic GSL pathway is shown with the percentage of accessions in each chemotypes (out of the total 797 accessions) shown in each box.

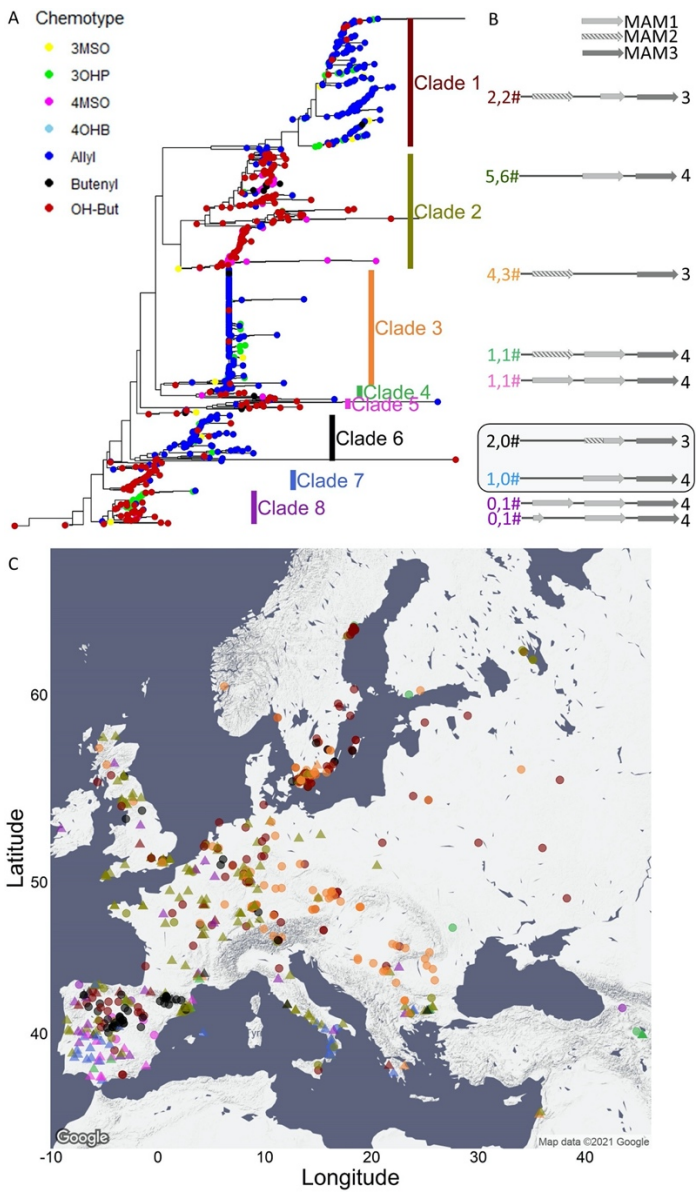


Figure A-5*: MAM3* phylogeny.
(**A**) *MAM3* phylogeny of *Arabidopsis thaliana* accessions, rooted by *Arabidopsis lyrata MAMb*, which is not shown because of distance. Tree tips are colored based on the accession chemotype. (**B**) The genomic structure of the *GS-Elong* regions in the previously sequenced accessions is shown based on Kroymann et al., 2003. The structures in the box are based on sequences obtained in this work. The numbers left to the structures indicate the number of sequenced accessions in this work (left) or by Kroymann et al., 2003 (right). The numbers are colored based on their clades. Bright gray arrows represent *MAM1* sequences, and dashed arrows represent *MAM2* sequences. Dark gray arrows represent *MAM3* sequences. The number to the right of the genomic cartoon represents the number of carbons in the side chain. (**C**) Collection sites of the accessions colored by their clade classification (from section A) and shaped based on the side chain length of the aliphatic short-chained glucosinolates (circles for C3, triangles for C4).

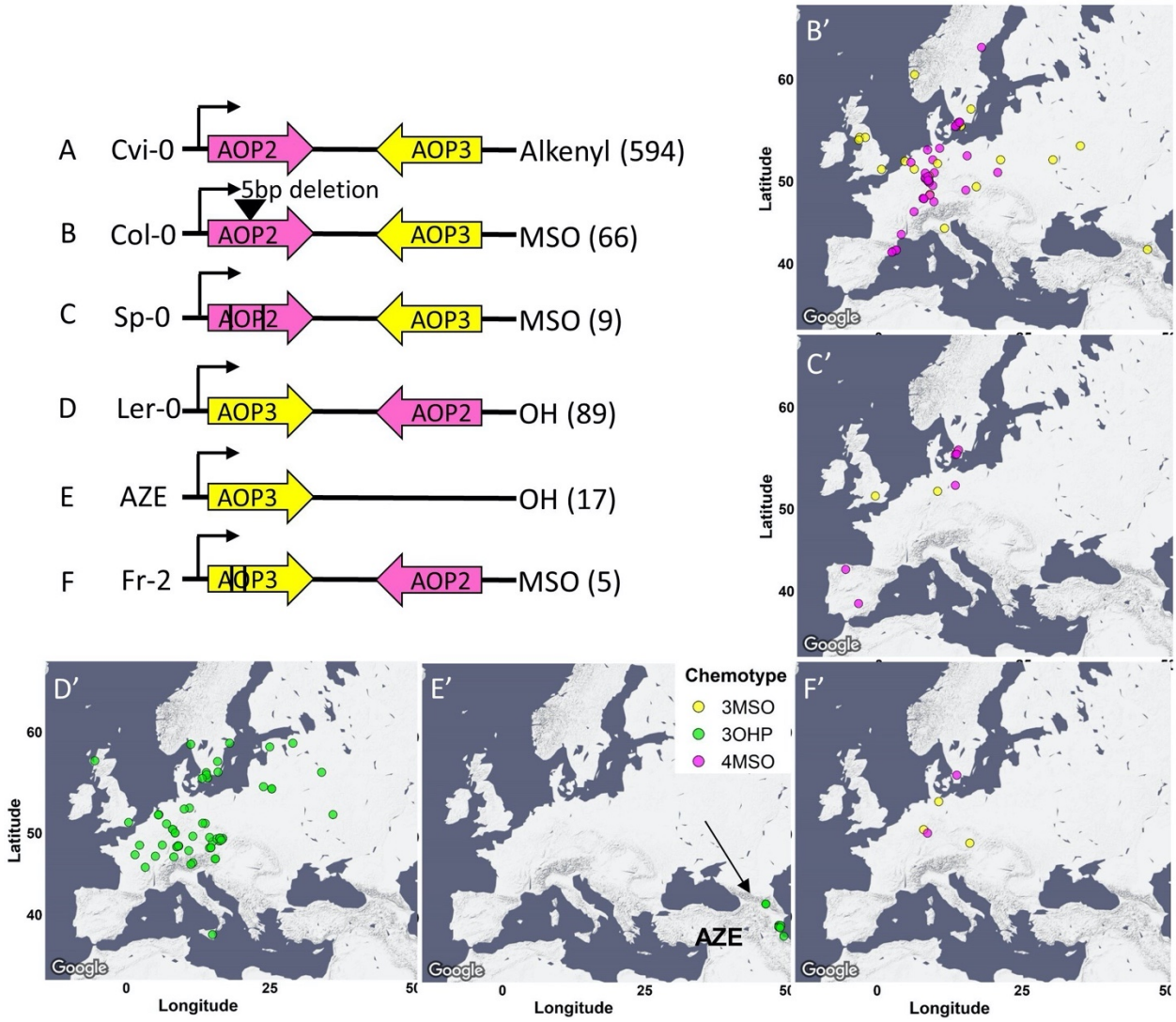


***Figure A-6: AOP* genomic structure.**
The genomic structure and causality of the major *AOP2/AOP3* haplotypes are illustrated. Pink arrows show the *AOP2* gene while yellow arrows represent *AOP3*. The black arrows represent the direction of transcription from the *AOP2* promoter as defined in the Col-0 reference genome. Its position does not change in any of the regions. **A-F** represent the different structures. The black lines in **C** and **F** represent theoretical positions of independent variants creating premature stop codons. The GSL chemotype for each haplotype is listed to the right with the number of the accessions in brackets. The maps show the geographic distribution of the accessions from each structure.

REFERENCES

1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at, and 1001 Genomes Consortium. 2016. "1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis Thaliana." *Cell* 166 (2): 481–91. https://doi.org/10.1016/j.cell.2016.05.063.

Abdi, Hervé, and Lynne J. Williams. 2010. "Principal Component Analysis." *Wiley Interdisciplinary Reviews. Computational Statistics* 2 (4): 433–59. https://doi.org/10.1002/wics.101

Abrahams, R. Shawn, J. Chris Pires, and M. Eric Schranz. 2020. "Genomic Origin and Diversification of the Glucosinolate MAM Locus." *Frontiers in Plant Science* 11 (June): 711. https://doi.org/10.3389/fpls.2020.00711.

Agrawal, A. A. 2000. "Overcompensation of Plants in Response to Herbivory and the by-Product Benefits of Mutualism." *Trends in Plant Science* 5 (7): 309–13. https://doi.org/10.1016/s1360-1385(00)01679-4.

Atwell, Susanna, Yu S. Huang, Bjarni J. Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, et al. 2010. "Genome-Wide Association Study of 107 Phenotypes in Arabidopsis Thaliana Inbred Lines." *Nature* 465 (7298): 627–31. https://doi.org/10.1038/nature08800.

Bakker, Erica G., M. Brian Traw, Christopher Toomajian, Martin Kreitman, and Joy Bergelson. 2008. "Low Levels of Polymorphism in Genes That Control the Activation of Defense Response in Arabidopsis Thaliana." *Genetics* 178 (4): 2031–43. https://doi.org/10.1534/genetics.107.083279.

Bednarek, Pawel, and Anne Osbourn. 2009. “Plant-Microbe Interactions: Chemical Diversity in Plant Defense.” *Science* 324 (5928): 746–48. https://doi.org/10.1126/science.1171661.

Beekwilder, Jules, Wessel van Leeuwen, Nicole M. van Dam, Monica Bertossi, Valentina Grandi, Luca Mizzi, Mikhail Soloviev, et al. 2008. “The Impact of the Absence of Aliphatic Glucosinolates on Insect Herbivory in Arabidopsis.” *PloS One* 3 (4): e2068. https://doi.org/10.1371/journal.pone.0002068.

Benderoth, Markus, Marina Pfalz, and Juergen Kroymann. 2008. “Methylthioalkylmalate Synthases: Genetics, Ecology and Evolution.” *Phytochemistry Reviews* 8 (1): 255. https://doi.org/10.1007/s11101-008-9097-1.

Benderoth, Markus, Susanne Textor, Aaron J. Windsor, Thomas Mitchell-Olds, Jonathan Gershenzon, and Juergen Kroymann. 2006. “Positive Selection Driving Diversification in Plant Secondary Metabolism.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (24): 9118–23. https://doi.org/10.1073/pnas.0601738103.

Bialy, Z., W. Oleszek, J. Lewis, and G. R. Fenwick. 1990. “Allelopathic Potential Ofgluco Sinolates (mustard Oil Glycosides) and Their Degradation Products.”

Bodenhofer, Ulrich, Enrico Bonatesta, Christoph Horejš-Kainrath, and Sepp Hochreiter. 2015. “Msa: An R Package for Multiple Sequence Alignment.” *Bioinformatics* 31 (24): 3997–99. https://doi.org/10.1093/bioinformatics/btv494.

Brachi, Benjamin, Christopher G. Meyer, Romain Villoutreix, Alexander Platt, Timothy C. Morton, Fabrice Roux, and Joy Bergelson. 2015. “Coselected Genes Determine Adaptive Variation in Herbivore Resistance throughout the Native

Range of Arabidopsis Thaliana." *Proceedings of the National Academy of Sciences of the United States of America* 112 (13): 4032–37. https://doi.org/10.1073/pnas.1421416112.

Brown, Paul D., Jim G. Tokuhisa, Michael Reichelt, and Jonathan Gershenzon. 2003. "Variation of Glucosinolate Accumulation among Different Organs and Developmental Stages of Arabidopsis Thaliana." *Phytochemistry* 62 (3): 471–81. https://doi.org/10.1016/s0031-9422(02)00549-6.

Chan, Eva K. F., Heather C. Rowe, Jason A. Corwin, Bindu Joseph, and Daniel J. Kliebenstein. 2011. "Combining Genome-Wide Association Mapping and Transcriptional Networks to Identify Novel Genes Controlling Glucosinolates in Arabidopsis Thaliana." *PLoS Biology* 9 (8): e1001125. https://doi.org/10.1371/journal.pbio.1001125.

Chan, Eva K. F., Heather C. Rowe, and Daniel J. Kliebenstein. 2010. "Understanding the Evolution of Defense Metabolites in Arabidopsis Thaliana Using Genome-Wide Association Mapping." *Genetics* 185 (3): 991–1007. https://doi.org/10.1534/genetics.109.108522.

Corrion, Alex, and Brad Day. 2015. "Pathogen Resistance Signalling in Plants." In *eLS*, 1–14. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470015902.a0020119.pub2.

Daxenbichler, Melvin E., Gayland F. Spencer, Diana G. Carlson, Gertrude B. Rose, Anita M. Brinker, and Richard G. Powell. 1991. "Glucosinolate Composition of Seeds from 297 Species of Wild Plants." *Phytochemistry* 30 (8): 2623–38. https://doi.org/10.1016/0031-9422(91)85112-D.

Durvasula, Arun, Andrea Fulgione, Rafal M. Gutaker, Selen Irez Alacakaptan, Pádraic J. Flood, Célia Neto, Takashi Tsuchimatsu, et al. 2017. "African Genomes Illuminate the Early History and Transition to Selfing in Arabidopsis Thaliana." *Proceedings of the National Academy of Sciences of the United States of America* 114 (20): 5213–18. https://doi.org/10.1073/pnas.1616736114.

Erb, Matthias, and Daniel J. Kliebenstein. 2020. "Plant Secondary Metabolites as Defenses, Regulators, and Primary Metabolites: The Blurred Functional Trichotomy." *Plant Physiology* 184 (1): 39–52. https://doi.org/10.1104/pp.20.00433.

Fan, Pengxiang, Bryan J. Leong, and Robert L. Last. 2019. "Tip of the Trichome: Evolution of Acylsugar Metabolic Diversity in Solanaceae." *Current Opinion in Plant Biology* 49 (June): 8–16. https://doi.org/10.1016/j.pbi.2019.03.005.

Ferrero-Serrano, Ángel, and Sarah M. Assmann. 2019. "Phenotypic and Genome-Wide Association with the Local Environment of Arabidopsis." *Nature Ecology & Evolution* 3 (2): 274–85. https://doi.org/10.1038/s41559-018-0754-5.

Futuyma, Douglas J., and Anurag A. Agrawal. 2009. "Macroevolution and the Biological Diversity of Plants and Herbivores." *Proceedings of the National Academy of Sciences of the United States of America* 106 (43): 18054–61. https://doi.org/10.1073/pnas.0904106106.

Giamoustaris, A., and R. Mithen. 1996. "Genetics of Aliphatic Glucosinolates. IV. Side-Chain Modification in Brassica Oleracea." *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik* 93 (5-6): 1006–10. https://doi.org/10.1007/BF00224105.

Göktay, Mehmet, Andrea Fulgione, and Angela M. Hancock. 2021. "A New Catalog of Structural Variants in 1,301 A. Thaliana Lines from Africa, Eurasia, and North America Reveals a Signature of Balancing Selection at Defense Response Genes." *Molecular Biology and Evolution* 38 (4): 1498–1511. https://academic.oup.com/mbe/article-abstract/38/4/1498/6008718.

Graser, G., B. Schneider, N. J. Oldham, and J. Gershenzon. 2000. "The Methionine Chain Elongation Pathway in the Biosynthesis of Glucosinolates in Eruca Sativa (Brassicaceae)." *Archives of Biochemistry and Biophysics* 378 (2): 411–19. https://doi.org/10.1006/abbi.2000.1812.

Grimm, Dominik G., Damian Roqueiro, Patrice A. Salomé, Stefan Kleeberger, Bastian Greshake, Wangsheng Zhu, Chang Liu, et al. 2017. "easyGWAS: A Cloud-Based Platform for Comparing the Results of Genome-Wide Association Studies." *The Plant Cell* 29 (1): 5–19. https://doi.org/10.1105/tpc.16.00551.

Halkier, Barbara Ann, and Jonathan Gershenzon. 2006. "Biology and Biochemistry of Glucosinolates." *Annual Review of Plant Biology* 57: 303–33. https://doi.org/10.1146/annurev.arplant.57.032905.105228.

Hanower, Pawel, and Janina Brzozowska. 1975. "Influence D'un Choc Osmotique Sur La Composition Des Feuilles de Cotonnier En Acides Amines Libres." *Phytochemistry* 14 (8): 1691–94. https://doi.org/10.1016/0031-9422(75)85275-7.

Hansen, B. G., D. J. Kliebenstein, and B. A. Halkier. 2007. "Identification of a Flavin-monooxygenase as the S-oxygenating Enzyme in Aliphatic Glucosinolate Biosynthesis in Arabidopsis." *The Plant Journal: For Cell and Molecular*

Biology. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-313X.2007.03101.x.

Hansen, Bjarne G., Rachel E. Kerwin, James A. Ober, Virginia M. Lambrix, Thomas Mitchell-Olds, Jonathan Gershenzon, Barbara A. Halkier, and Daniel J. Kliebenstein. 2008. "A Novel 2-Oxoacid-Dependent Dioxygenase Involved in the Formation of the Goiterogenic 2-Hydroxybut-3-Enyl Glucosinolate and Generalist Insect Resistance in Arabidopsis." *Plant Physiology* 148 (4): 2096–2108. https://academic.oup.com/plphys/article-abstract/148/4/2096/6107613.

Hasegawa, T., K. Yamada, S. Kosemura, S. Yamamura, and K. Hasegawa. 2000. "Phototropic Stimulation Induces the Conversion of Glucosinolate to Phototropism-Regulating Substances of Radish Hypocotyls." *Phytochemistry* 54 (3): 275–79. https://doi.org/10.1016/s0031-9422(00)00080-7.

Hayat, S., Q. Hayat, M. N. Alyemeni, A. S. Wani, J. Pichtel, A. Ahmad, and Others. 2012. "Role of Proline under Changing Environments: A Review. Plant Signal Behav 7: 1456--1466."

Heidel, Andrew J., Maria J. Clauss, Juergen Kroymann, Outi Savolainen, and Thomas Mitchell-Olds. 2006. "Natural Variation in MAM within and between Populations of Arabidopsis Lyrata Determines Glucosinolate Phenotype." *Genetics* 173 (3): 1629–36. https://doi.org/10.1534/genetics.106.056986.

Hu, L., P. Mateo, M. Ye, X. Zhang, J. D. Berset, V. Handrick, D. Radisch, et al. 2018. "Plant Iron Acquisition Strategy Exploited by an Insect Herbivore." *Science* 361 (6403): 694–97. https://doi.org/10.1126/science.aat4082.

Jander, G., J. Cui, B. Nhan, N. E. Pierce, and F. M. Ausubel. 2001. “The TASTY Locus on Chromosome 1 of Arabidopsis Affects Feeding of the Insect Herbivore Trichoplusia Ni.” *Plant Physiology* 126 (2): 890–98. https://doi.org/10.1104/pp.126.2.890.

Kahle, David, and Hadley Wickham. 2013. “Ggmap: Spatial Visualization with ggplot2.” *The R Journal* 5 (1): 144. https://doi.org/10.32614/rj-2013-014.

Kang, Hyun Min, Jae Hoon Sul, Susan K. Service, Noah A. Zaitlen, Sit-Yee Kong, Nelson B. Freimer, Chiara Sabatti, and Eleazar Eskin. 2010. “Variance Component Model to Account for Sample Structure in Genome-Wide Association Studies.” *Nature Genetics* 42 (4): 348–54. https://doi.org/10.1038/ng.548.

Katoh, K., J. Rozewicki, and K. D. Yamada. 2017. “MAFFT Online Service: Multiple Sequence Alignment, Interactive Sequence Choice and Visualization. Briefings in Bioinformatics. Sep 6.”

Katz, Ella, Rammyani Bagchi, Verena Jeschke, Alycia R. M. Rasmussen, Aleshia Hopper, Meike Burow, Mark Estelle, and Daniel J. Kliebenstein. 2020. “Diverse Allyl Glucosinolate Catabolites Independently Influence Root Growth and Development.” *Plant Physiology* 183 (3): 1376–90. https://doi.org/10.1104/pp.20.00170.

Katz, Ella, Sophia Nisani, Mor Sela, Hila Behar, and Daniel A. Chamovitz. 2015. “The Effect of Indole-3-Carbinol on PIN1 and PIN2 in Arabidopsis Roots.” *Plant Signaling & Behavior* 10 (9): e1062200. https://doi.org/10.1080/15592324.2015.1062200.

Kerwin, Rachel E., Julie Feusier, Alise Muok, Catherine Lin, Brandon Larson, Daniel Copeland, Jason A. Corwin, et al. 2017. “Epistasis × Environment Interactions among Arabidopsis Thaliana Glucosinolate Genes Impact Complex Traits and Fitness in the Field.” *The New Phytologist* 215 (3): 1249–63. https://doi.org/10.1111/nph.14646.

Kerwin, Rachel, Julie Feusier, Jason Corwin, Matthew Rubin, Catherine Lin, Alise Muok, Brandon Larson, et al. 2015. “Natural Genetic Variation in Arabidopsis Thaliana Defense Metabolism Genes Modulates Field Fitness.” *eLife* 4 (April). https://doi.org/10.7554/eLife.05604.

Kim, Jeongwoon, Kiyoon Kang, Eliana Gonzales-Vigil, Feng Shi, A. Daniel Jones, Cornelius S. Barry, and Robert L. Last. 2012. “Striking Natural Diversity in Glandular Trichome Acylsugar Composition Is Shaped by Variation at the Acyltransferase2 Locus in the Wild Tomato Solanum Habrochaites.” *Plant Physiology* 160 (4): 1854–70. https://doi.org/10.1104/pp.112.204735.

Kliebenstein, D. J. 2004. “Secondary Metabolites and Plant/environment Interactions: A View through Arabidopsis Thaliana Tinged Glasses.” *Plant, Cell & Environment* 27 (6): 675–84. https://doi.org/10.1111/j.1365-3040.2004.01180.x.

Kliebenstein, D. J., and N. I. Cacho. 2016. “Chapter Three - Nonlinear Selection and a Blend of Convergent, Divergent and Parallel Evolution Shapes Natural Variation in Glucosinolates.” In *Advances in Botanical Research*, edited by Stanislav Kopriva, 80:31–55. Academic Press. https://doi.org/10.1016/bs.abr.2016.06.002.

Kliebenstein, D. J., J. Gershenzon, and T. Mitchell-Olds. 2001. “Comparative Quantitative Trait Loci Mapping of Aliphatic, Indolic and Benzylic Glucosinolate

Production in Arabidopsis Thaliana Leaves and Seeds." *Genetics* 159 (1): 359–70. https://www.ncbi.nlm.nih.gov/pubmed/11560911.

Kliebenstein, D. J., J. Kroymann, P. Brown, A. Figuth, D. Pedersen, J. Gershenzon, and T. Mitchell-Olds. 2001. "Genetic Control of Natural Variation in Arabidopsis Glucosinolate Accumulation." *Plant Physiology* 126 (2): 811–25. https://doi.org/10.1104/pp.126.2.811.

Kliebenstein, Daniel J. 2009. "A Quantitative Genetics and Ecological Model System: Understanding the Aliphatic Glucosinolate Biosynthetic Network via QTLs." *Phytochemistry Reviews: Proceedings of the Phytochemical Society of Europe* 8 (1): 243–54. https://doi.org/10.1007/s11101-008-9102-8.

Kliebenstein, Daniel J., Antje Figuth, and Thomas Mitchell-Olds. 2002. "Genetic Architecture of Plastic Methyl Jasmonate Responses in Arabidopsis Thaliana." *Genetics* 161 (4): 1685–96. https://www.ncbi.nlm.nih.gov/pubmed/12196411.

Kliebenstein, Daniel J., Virginia M. Lambrix, Michael Reichelt, Jonathan Gershenzon, and Thomas Mitchell-Olds. 2001. "Gene Duplication in the Diversification of Secondary Metabolism: Tandem 2-Oxoglutarate-Dependent Dioxygenases Control Glucosinolate Biosynthesis in Arabidopsis." *The Plant Cell* 13 (3): 681. https://doi.org/10.2307/3871415.

Kliebenstein, Daniel, Deana Pedersen, Bridget Barker, and Thomas Mitchell-Olds. 2002. "Comparative Analysis of Quantitative Trait Loci Controlling Glucosinolates, Myrosinase and Insect Resistance in Arabidopsis Thaliana." *Genetics* 161 (1): 325–32. https://www.ncbi.nlm.nih.gov/pubmed/12019246.

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy. 2017. “Canu: Scalable and Accurate Long-Read Assembly via Adaptive K-Mer Weighting and Repeat Separation. bioRxiv, 071282.”

Kroymann, J., S. Textor, J. G. Tokuhisa, K. L. Falk, S. Bartram, J. Gershenzon, and T. Mitchell-Olds. 2001. “A Gene Controlling Variation in Arabidopsis Glucosinolate Composition Is Part of the Methionine Chain Elongation Pathway.” *Plant Physiology* 127 (3): 1077–88. https://doi.org/10.1104/pp.127.3.1077.

Kroymann, Juergen, Susanne Donnerhacke, Domenica Schnabelrauch, and Thomas Mitchell-Olds. 2003. “Evolutionary Dynamics of an Arabidopsis Insect Resistance Quantitative Trait Locus.” *Proceedings of the National Academy of Sciences of the United States of America* 100 Suppl 2 (November): 14587–92. https://doi.org/10.1073/pnas.1734046100.

Kroymann, Juergen, and Thomas Mitchell-Olds. 2005. “Epistasis and Balanced Polymorphism Influencing Complex Trait Variation.” *Nature* 435 (7038): 95–98. https://doi.org/10.1038/nature03480.

Kuraku, Shigehiro, Christian M. Zmasek, Osamu Nishimura, and Kazutaka Katoh. 2013. “aLeaves Facilitates on-Demand Exploration of Metazoan Gene Family Trees on MAFFT Sequence Alignment Server with Enhanced Interactivity.” *Nucleic Acids Research* 41 (Web Server issue): W22–28. https://doi.org/10.1093/nar/gkt389.

Lankau, Richard A. 2007. “Specialist and Generalist Herbivores Exert Opposing Selection on a Chemical Defense.” *The New Phytologist* 175 (1): 176–84. https://doi.org/10.1111/j.1469-8137.2007.02090.x.

Lankau, Richard A., and Daniel J. Kliebenstein. 2009. “Competition, Herbivory and Genetics Interact to Determine the Accumulation and Fitness Consequences of a Defence Metabolite.” *The Journal of Ecology* 97 (1): 78–88. http://www.jstor.org/stable/20528833.

Lankau, Richard A., and Sharon Y. Strauss. 2007. “Mutual Feedbacks Maintain Both Genetic and Species Diversity in a Plant Community.” *Science* 317 (5844): 1561–63. https://doi.org/10.1126/science.1147455.

Lee, Cheng-Ruei, Hannes Svardal, Ashley Farlow, Moises Exposito-Alonso, Wei Ding, Polina Novikova, Carlos Alonso-Blanco, Detlef Weigel, and Magnus Nordborg. 2017. “On the Post-Glacial Spread of Human Commensal Arabidopsis Thaliana.” *Nature Communications* 8 (February): 14458. https://doi.org/10.1038/ncomms14458.

MacQueen, Alice, Xiaoqin Sun, and Joy Bergelson. 2016. “Genetic Architecture and Pleiotropy Shape Costs of Rps2 -Mediated Resistance in Arabidopsis Thaliana.” *Nature Plants* 2 (8): 1–8. https://doi.org/10.1038/nplants.2016.110.

Malcolm, Stephen B. 1994. “Milkweeds, Monarch Butterflies and the Ecological Significance of Cardenolides.” *Chemoecology* 5 (3): 101–17. https://doi.org/10.1007/BF01240595.

Malinovsky, Frederikke Gro, Marie-Louise F. Thomsen, Sebastian J. Nintemann, Lea Møller Jagd, Baptiste Bourgine, Meike Burow, and Daniel J. Kliebenstein. 2017. “An Evolutionarily Young Defense Metabolite Influences the Root Growth of Plants via the Ancient TOR Signaling Pathway.” *eLife* 6 (December). https://doi.org/10.7554/eLife.29353.

Mithen, R., J. Clarke, C. Lister, and C. Dean. 1995. “Genetics of Aliphatic Glucosinolates. III. Side Chain Structure of Aliphatic Glucosinolates in Arabidopsis Thaliana.” *Heredity* 74 (2): 210–15. https://doi.org/10.1038/hdy.1995.29.

Moghe, Gaurav D., and Robert L. Last. 2015. “Something Old, Something New: Conserved Enzymes and the Evolution of Novelty in Plant Specialized Metabolism.” *Plant Physiology* 169 (3): 1512–23. https://doi.org/10.1104/pp.15.00994.

Moore, Bethany M., Peipei Wang, Pengxiang Fan, Bryan Leong, Craig A. Schenck, John P. Lloyd, Melissa D. Lehti-Shiu, Robert L. Last, Eran Pichersky, and Shin-Han Shiu. 2019. “Robust Predictions of Specialized Metabolism Genes through Machine Learning.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (6): 2344–53. https://doi.org/10.1073/pnas.1817074116.

Opitz, Sebastian E. W., and Caroline Müller. 2009. “Plant Chemistry and Insect Sequestration.” *Chemoecology* 19 (3): 117–54. https://doi.org/10.1007/s00049-009-0018-6.

Paradis, Emmanuel, and Klaus Schliep. 2019. “Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R.” *Bioinformatics* 35 (3): 526–28. https://doi.org/10.1093/bioinformatics/bty633.

Petersen, Bent Larsen, Sixue Chen, Carsten Hørslev Hansen, Carl Erik Olsen, and Barbara Ann Halkier. 2002. “Composition and Content of Glucosinolates in

Developing Arabidopsis Thaliana." *Planta* 214 (4): 562–71. https://doi.org/10.1007/s004250100659.

Pfalz, Marina, Heiko Vogel, Thomas Mitchell-Olds, and Juergen Kroymann. 2007. "Mapping of QTL for Resistance against the Crucifer Specialist Herbivore Pieris Brassicae in a New Arabidopsis Inbred Line Population, Da(1)-12×Ei-2." *PloS One* 2 (6): e578. https://doi.org/10.1371/journal.pone.0000578.

Ramos-Onsins, Sebastián E., Barbara E. Stranger, Thomas Mitchell-Olds, and Montserrat Aguadé. 2004. "Multilocus Analysis of Variation and Speciation in the Closely Related Species Arabidopsis Halleri and A. Lyrata." *Genetics* 166 (1): 373–88. https://doi.org/10.1534/genetics.166.1.373.

Raybould, A. F., and C. L. Moyes. 2001. "The Ecological Genetics of Aliphatic Glucosinolates." *Heredity* 87 (Pt 4): 383–91. https://doi.org/10.1046/j.1365-2540.2001.00954.x.

Rodman, James E., Arthur R. Kruckeberg, and Ihsan A. Al-Shehbaz. 1981. "Chemotaxonomic Diversity and Complexity in Seed Glucosinolates of Caulanthus and Streptanthus (Cruciferae)." *Systematic Botany* 6 (3): 197–222. https://doi.org/10.2307/2418282.

Rodman, James Eric. 1980. "Population Variation and Hybridization in Sea-rockets (Cakile, Cruciferae): Seed Glucosinolate Characters." *American Journal of Botany* 67 (8): 1145–59. https://doi.org/10.1002/j.1537-2197.1980.tb07748.x.

Salehin, Mohammad, Baohua Li, Michelle Tang, Ella Katz, Liang Song, Joseph R. Ecker, Daniel J. Kliebenstein, and Mark Estelle. 2019. "Auxin-Sensitive Aux/IAA Proteins Mediate Drought Tolerance in Arabidopsis by Regulating Glucosinolate

Levels." *Nature Communications* 10 (1): 4021. https://doi.org/10.1038/s41467-019-12002-1.

Schilmiller, Anthony L., Eran Pichersky, and Robert L. Last. 2012. "Taming the Hydra of Specialized Metabolism: How Systems Biology and Comparative Approaches Are Revolutionizing Plant Biochemistry." *Current Opinion in Plant Biology* 15 (3): 338–44. https://doi.org/10.1016/j.pbi.2011.12.005.

Smith, S. A., and C. W. Dunn. 2008. "Phyutility: A Phyloinformatics Tool for Trees, Alignments and Molecular Data. Bioinformatics24: 715-716."

Sønderby, Ida E., Fernando Geu-Flores, and Barbara A. Halkier. 2010. "Biosynthesis of Glucosinolates--Gene Discovery and beyond." *Trends in Plant Science* 15 (5): 283–90. https://doi.org/10.1016/j.tplants.2010.02.005.

Stamatakis, A. 2014. "RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics30: 1312-1313." *Go to Original Source*.

Szakiel, Anna, Cezary Pączkowski, and Max Henry. 2011. "Influence of Environmental Abiotic Factors on the Content of Saponins in Plants." *Phytochemistry Reviews: Proceedings of the Phytochemical Society of Europe* 10 (4): 471–91. https://doi.org/10.1007/s11101-010-9177-x.

Textor, Susanne, Jan-Willem de Kraker, Bettina Hause, Jonathan Gershenzon, and James G. Tokuhisa. 2007. "MAM3 Catalyzes the Formation of All Aliphatic Glucosinolate Chain Lengths in Arabidopsis." *Plant Physiology* 144 (1): 60–71. https://doi.org/10.1104/pp.106.091579.

Thakur, Partap S., and Vinay K. Rai. 1982. “Dynamics of Amino Acid Accumulation of Two Differentially Drought Resistant Zea Mays Cultivars in Response to Osmotic Stress.” *Environmental and Experimental Botany* 22 (2): 221–26. https://doi.org/10.1016/0098-8472(82)90042-9.

Todesco, Marco, Sureshkumar Balasubramanian, Tina T. Hu, M. Brian Traw, Matthew Horton, Petra Epple, Christine Kuhns, et al. 2010. “Natural Allelic Variation Underlying a Major Fitness Trade-off in Arabidopsis Thaliana.” *Nature* 465 (7298): 632–36. https://doi.org/10.1038/nature09083.

Turner, Stephen D. 2014. “Qqman: An R Package for Visualizing GWAS Results Using Q-Q and Manhattan Plots.” *bioRxiv*. https://doi.org/10.1101/005165.

Uremis, I., Mehmet Arslan, M. K. Sangun, V. Uygur, and N. Isler. 2009. “Allelopathic Potential of Rapeseed Cultivars on Germination and Seedling Growth of Weeds.” *Asian Journal of Chemistry* 21 (3): 2170. https://www.researchgate.net/profile/Mehmet-Arslan-12/publication/283612250_Allelopathic_potential_of_rapeseed_cultivars_on_germination_and_seedling_growth_of_weeds/links/565dba6808ae4988a7bce567/Allelopathic-potential-of-rapeseed-cultivars-on-germination-and-seedling-growth-of-weeds.pdf.

Wentzell, Adam M., and Daniel J. Kliebenstein. 2008. “Genotype, Age, Tissue, and Environment Regulate the Structural Outcome of Glucosinolate Activation.” *Plant Physiology* 147 (1): 415–28. https://doi.org/10.1104/pp.107.115279.

Wentzell, Adam M., Heather C. Rowe, Bjarne Gram Hansen, Carla Ticconi, Barbara Ann Halkier, and Daniel J. Kliebenstein. 2007. “Linking Metabolic QTLs with

Network and Cis-eQTLs Controlling Biosynthetic Pathways." *PLoS Genetics* 3 (9): 1687–1701. https://doi.org/10.1371/journal.pgen.0030162.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Windsor, Aaron J., Michael Reichelt, Antje Figuth, Ales Svatos, Juergen Kroymann, Daniel J. Kliebenstein, Jonathan Gershenzon, and Thomas Mitchell-Olds. 2005. "Geographic and Evolutionary Diversification of Glucosinolates among near Relatives of Arabidopsis Thaliana (Brassicaceae)." *Phytochemistry* 66 (11): 1321–33. https://doi.org/10.1016/j.phytochem.2005.04.016.

Wolters, Hanno, and Gerd Jürgens. 2009. "Survival of the Flexible: Hormonal Growth Control and Adaptation in Plant Development." *Nature Reviews. Genetics* 10 (5): 305–17. https://doi.org/10.1038/nrg2558.

Wright, Stephen I., Beatrice Lauga, and Deborah Charlesworth. 2002. "Rates and Patterns of Molecular Evolution in Inbred and Outbred Arabidopsis." *Molecular Biology and Evolution* 19 (9): 1407–20. https://doi.org/10.1093/oxfordjournals.molbev.a004204.

Yamada, Kosumi, Tsuyoshi Hasegawa, Eiichi Minami, Naoto Shibuya, Seiji Kosemura, Shosuke Yamamura, and Koji Hasegawa. 2003. "Induction of Myrosinase Gene Expression and Myrosinase Activity in Radish Hypocotyls by Phototropic Stimulation." *Journal of Plant Physiology* 160 (3): 255–59. https://doi.org/10.1078/0176-1617-00950.

Yang, C. W., C. C. Lin, and C. H. Kao. 2000. "Proline, Ornithine, Arginine and Glutamic Acid Contents in Detached Rice Leaves." *Biologia Plantarum*. https://idp.springer.com/authorize/casa?redirect_uri=https://link.springer.com/article/10.1023/A:1002733117506&casa_token=j-jDEQjKD6kAAAAA:9hmztYuGZYZiohBLQ3w-C4S_V-7TZjoJ-g-IdyPUUpH5QtRsrfa6i4hJGcVSXAf6E3KjdscMlQWNUJj9.

Yu, Guangchuang. 2020. "Using Ggtree to Visualize Data on Tree-Like Structures." *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis... [et Al.]* 69 (1): e96. https://doi.org/10.1002/cpbi.96.

Zhu, Wangsheng, Maricris Zaidem, Anna-Lena Van de Weyer, Rafal M. Gutaker, Eunyoung Chae, Sang-Tae Kim, Felix Bemm, et al. 2018. "Modulation of ACD6 Dependent Hyperimmunity by Natural Alleles of an Arabidopsis Thaliana NLR Resistance Gene." *PLoS Genetics* 14 (9): e1007628. https://doi.org/10.1371/journal.pgen.1007628.

Züst, Tobias, and Anurag A. Agrawal. 2017. "Trade-Offs Between Plant Growth and Defense Against Insect Herbivory: An Emerging Mechanistic Synthesis." *Annual Review of Plant Biology* 68 (April): 513–34. https://doi.org/10.1146/annurev-arplant-042916-040856.

Züst, Tobias, Christian Heichinger, Ueli Grossniklaus, Richard Harrington, Daniel J. Kliebenstein, and Lindsay A. Turnbull. 2012. "Natural Enemies Drive Geographic Variation in Plant Defenses." *Science* 338 (6103): 116–19. https://doi.org/10.1126/science.1226397.

Appendix B: The contributions from the progenitor genomes of the mesopolyploid Brassiceae are evolutionarily distinct but functionally compatible

Yue Hao,[1] Makenzie E. Mabry,[2] Patrick P. Edger,[3,4] Michael Freeling,[5] Chunfang Zheng,[6] Lingling Jin,[7] Robert VanBuren,[3,8] Marivi Colle,[3] Hong An,[2] R. Shawn Abrahams,[2] Jacob D. Washburn,[9] Xinshuai Qi,[10] Kerrie Barry,[11] Christopher Daum,[11] Shengqiang Shu,[11] Jeremy Schmutz,[11,12] David Sankoff,[6] Michael S. Barker,[10] Eric Lyons,[13,14] J. Chris Pires,[2,15] and Gavin C. Conant[1,16,17,18]

1Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina 27695, USA; 2 Division of Biological Sciences, University of Missouri–Columbia, Columbia, Missouri 65211, USA; 3 Department of Horticulture, Michigan State University, East Lansing, Michigan 48824, USA; 4 Genetics and Genome Sciences, Michigan State University, East Lansing, Michigan 48824, USA; 5 Department of Plant and Microbial Biology, University of California, Berkeley, California 94720, USA; 6 Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada; 7 Department of Computer Science, University of Saskatchewan, Saskatoon, Saskatchewan S7N 5C9, Canada; 8 Plant Resilience Institute, Michigan State University, East Lansing, Michigan 48824, USA; 9 Plant Genetics Research Unit, USDA-ARS, Columbia, Missouri 65211, USA; 10Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA; 11Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA; 12HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA; 13School of Plant Sciences, University of Arizona, Tucson, Arizona 85721, USA; 14BIO5 Institute, University of Arizona, Tucson, Arizona 85721, USA; 15Informatics Institute, University of Missouri–Columbia, Columbia, Missouri 65211, USA; 16Program in Genetics, North Carolina State University, Raleigh, North Carolina 27695, USA; 17Department of Biological Sciences, North Carolina State University, Raleigh, North Carolina 27695, USA; 18Division of Animal Sciences, University of Missouri–Columbia, Columbia, Missouri 65211, USA

Please cite the published work here:

Hao, Yue, Makenzie E. Mabry, Patrick P. Edger, Michael Freeling, Chunfang Zheng, Lingling Jin, Robert VanBuren, Marivi Colle, Hong An, **R. Shawn Abrahams**, et al. "The contributions from the progenitor genomes of the mesopolyploid Brassiceae are evolutionarily distinct but functionally compatible." Genome research 31, no. 5 (2021): 799-810.

Authorial Contribution: Rearing of *Crambe* accessions and tissue collection. Taxonomic qualifications and partial methods section.

ABSTRACT

The members of the tribe Brassiceae share a whole-genome triplication (WGT), and one proposed model for its formation is a two-step pair of hybridizations producing hexaploid descendants. However, evidence for this model is incomplete, and the evolutionary and functional constraints that drove evolution after the hexaploidy are even less understood. Here, we report a new genome sequence of *Crambe hispanica*, a species sister to most sequenced Brassiceae. Using this new genome and three others that share the hexaploidy, we traced the history of gene loss after the WGT using the Polyploidy Orthology Inference Tool (POInT). We confirm the two-step formation model and infer that there was a significant temporal gap between those two allopolyploidizations, with about a third of the gene losses from the first two subgenomes occurring before the arrival of the third. We also, for the 90,000 individual genes in our study, make parental subgenome assignments, inferring, with measured uncertainty, from which of the progenitor genomes of the allohexaploidy each gene derives. We further show that each subgenome has a statistically distinguishable rate of homoeolog losses. There is little indication of functional distinction between the three subgenomes: the individual subgenomes show no patterns of functional enrichment, no excess of shared protein–protein or metabolic interactions between their members, and no biases in their likelihood of having experienced a recent selective sweep. We propose a "mix and match" model of allopolyploidy, in which subgenome origin drives homoeolog loss propensities but where genes from different subgenomes function together without difficulty.

INTRODUCTION

Fifty years ago, Ohno (1970) published a forceful opus on the role of gene duplication, and in particular of genome duplication (i.e., polyploidy), in evolutionary innovation. Since then, evidence both of polyploidy's ubiquity (Wolfe and Shields 1997; Van de Peer et al. 2009, 2017; Soltis and Soltis 2012) and of its role in evolutionary innovations such as yeast aerobic glucose fermentation, the organization of the retinas of teleost fishes, and in plant defensive compounds, has continued to accumulate (Conant and Wolfe 2007; Merico et al. 2007; van Hoek and Hogeweg 2009; Edger et al. 2015; Sukeena et al. 2016). Preeminent among the polyploid lineages are the flowering plants, in which more than 180 ancient polyploidies are known (One Thousand Plant Transcriptomes Initiative 2019).

When a new polyploid genome is created by the merging of similar but not identical progenitor species, it is referred to as an allopolyploid. Among allopolyploidies, the preferential retention of gene copies (homoeologs) from one of the parental subgenomes, known as biased fractionation, has been observed in yeast, maize, cotton, monkeyflower, *Arabidopsis*, *Brassica*, and nematodes (Thomas et al. 2006; Conant and Wolfe 2008a; Cheng et al. 2012; Parkin et al. 2014; Renny-Byfield et al. 2015; Edger et al. 2017; Emery et al. 2018; Schoonmaker et al. 2020). Allopolyploids also show a tendency for genes from one of the subgenomes to be more highly expressed, and silencing or loss of genes from the remaining subgenomes is correspondingly more likely (Thomas et al. 2006; Schnable et al. 2011; Yoo et al. 2014). A number of sources of these biases have been proposed, from variations in transposon silencing (Freeling et al. 2012; Woodhouse et al. 2014; Zhao et al. 2017; Alger and Edger 2020), to the disruption

of organelle-nucleus communication (Sharbrough et al. 2017; Costello et al. 2020) and epigenetic changes attributed to the genomic shock of polyploidy (McClintock 1984; Bird et al. 2018; Wendel et al. 2018). In this work, we sought to critically evaluate one such proposal: that allopolyploids might bring together coevolved and conflicting copies of multiprotein complexes (Codoñer and Fares 2008; Gong et al. 2012; Scienski et al. 2015; Emery et al. 2018). In this framework, early random gene losses from one subgenome that partly resolved these conflicts might then set the polyploidy down a path favoring losses from that subgenome. A related proposal was made by Makino and McLysaght (2012), who argued that selection to maintain dosage balance among interacting genomic neighbors could produce local, and eventually global, biases in fractionation.

It is also notable that not all homoeologs are equally likely to revert to single copy after a polyploidy, regardless of the level of biased fractionation. Duplicated genes coding for transcription factors, ribosomal proteins, and kinases are over-retained after independent polyploidies in flowering plants, yeasts, ciliates, and vertebrates (Seoighe and Wolfe 1998; Blanc and Wolfe 2004; Maere et al. 2005; Aury et al. 2006; Makino and McLysaght 2010). These patterns are best explained by a need to maintain dosage balance among highly interacting genes (Birchler et al. 2005; Hakes et al. 2007; Birchler and Veitia 2012, 2014; Conant et al. 2014). There are also genes that prefer not to be duplicated: genes for DNA repair and those targeted to organelles have returned to single copy rapidly after genome duplication (De Smet et al. 2013; Conant 2014).

The Brassiceae are the most morphologically diverse tribe in the family Brassicaceae (Cheng et al. 2014) and contain important crops such as broccoli, cabbage,

kale, mustard, and canola. This tribe experienced a hexaploidy (i.e., whole-genome triplication [WGT]) between 5 and 9 million years ago after its divergence from *Arabidopsis thaliana* (Wang et al. 2011). This Brassiceae WGT is a valuable system for studying all the aforementioned phenomena because the triplication allows us to explore each in unusual detail. This polyploidy was originally inferred with comparative linkage mapping (Lagercrantz 1998; Lukens et al. 2004; Parkin et al. 2005; Schranz et al. 2006) and confirmed by chromosome painting (Lysak et al. 2005; Lysak 2009). The patterns of biased fractionation observed in the genome of *Brassica rapa* suggested that the triplication "event" was actually two separate allopolyploid hybridizations involving three distinct diploid progenitor species, with the merger of the two currently highly fractionated ancestral subgenomes occurring first, followed by the subsequent addition of a third subgenome, which currently possesses the most retained genes (Cheng et al. 2012; Tang et al. 2012). However, this proposal is worth revisiting as it rests on inferences from a single genome: a phylogenetically broader analysis of the genomes that descend from the hexaploidy would more firmly ground our descriptions of its early history. At the moment, we lack genomes from early diverging lineages with the hexaploidy, such as those in the genus *Crambe*, which is sister to the genus *Brassica* (Arias and Pires 2012). Biologically, *Crambe* species are not only important industrial oilseed sources because of their high erucic acid content (Lazzeri et al. 1997; Warwick and Gugel 2003; Carlsson et al. 2007) but also could serve as resources for *Brassica* crop development (Rudloff and Wang 2011).

Using a new genome sequence from *Crambe hispanica*, we analyzed the Brassiceae WGT with our tool for modeling post-polyploidy genome evolution: the

Polyploidy Orthology Inference Tool (POInT) (Conant and Wolfe 2008a). We sought to first confirm the two-step hexaploidy model and its relationship to the observed three subgenomes in the extant genomes. POInT, which we recently extended to allow the analysis of WGTs (Schoonmaker et al. 2020), is ideally suited to this task because it can model homoeolog losses phylogenetically and test for biases in fractionation without ad hoc assumptions. We then tested the proposal that functional differences between the allopolyploid progenitors contributed to the biases in homoeolog losses using functional hierarchies, gene coexpression information, protein interaction catalogs, and metabolic network data.

RESULTS

A well-assembled and annotated genome of *Crambe hispanica*

The genome of *Crambe hispanica* was assembled using Pacific Biosciences (PacBio) reads. This assembly had a contig N50 of 4.4 Mb across 1019 contigs with a total assembly length of 480 Mb. Eleven terminal telomeres were resolved by the Canu assembler (Koren et al. 2017). The assembly graph showed low heterozygosity and few assembly artifacts, with the exception of one megacluster consisting of a high copy number LTR across 500 contigs and spanning ~30 Mb. The draft assembly was then polished using Illumina paired-end data. We also used Hi-C proximity ligation sequencing data to scaffold the genome, which resulted in 18 scaffolds that include 99.5% of the original assembly with a scaffold N50 of 32.6 Mbp and scaffold N90 of 30.1 Mbp. The annotated genome is of high quality: we compared its gene set against the Benchmarking Universal Single-Copy Orthologs (BUSCO v.2) (Simão et al. 2015) plant

data set (embryophyta_odb9), finding that 95.8% of these expected genes were present in our annotation.

Inferring blocks of triple-conserved synteny in four triplicated Brassiceae genomes and estimating an ancestral gene order

Based on their phylogenetic placement and assembly quality, we selected and retrieved from CoGe (Lyons and Freeling 2008; Lyons et al. 2008a) three additional mesohexaploid genomes for our analyses: those of *Brassica rapa* (version 1.5, CoGe id 24668) (Wang et al. 2011), *Brassica oleracea* (TO1000 version 2.1, CoGe id 26018) (Liu et al. 2014; Parkin et al. 2014), and *Sinapis alba* (version 1.1, CoGe id 33284). For each of these four genomes, we inferred blocks of triple-conserved synteny (TCS), with the genome of *Arabidopsis thaliana* used as an unduplicated reference. We then merged these blocks across all of the four genomes: we refer to each such locus as a "pillar." Each pillar consists of between one and three surviving genes in each of the four genomes. As described in Methods, we used both a set of TCS blocks inferred with POInT containing 14,050 pillars (P_{pillars}) and a separate ancestral genome reconstruction that estimates the gene order that existed just before the WGT. The latter contains five reconstructed ancestral chromosomes involving 89 scaffolds with a total of 10,868 ancestral genes. When we match these genes to the TCS blocks computed with POInT, the result is 7993 ancestrally ordered pillars (A_{pillars}).

Inferring the evolutionary relationships of the four Brassiceae genomes from gene loss patterns

We fit models of WGT evolution (see below) to several different orderings of the 14,050 pillars in the P_{pillars} set and to the A_{pillars} (Supplemental Table S1). These orderings of the P_{pillars} differed in their number of synteny breaks: we used the ordering with the highest likelihood under the WGT 3rate G1Dom model for our remaining analyses (see below). Similarly, we compared the fit of three possible phylogenetic topologies to the pillars under this model: the remainder of our analyses use the topology shown in Figure 1, which has the highest likelihood. We note that one of the other two topologies, although having a lower likelihood under POInT's models (Supplemental Fig. S1), is the phylogeny estimated using plastid genomes (Arias and Pires 2012). Because the A_{pillars} give similar parameter estimates but comprise a smaller data set, we will discuss our results in terms of the P_{pillars}.

The three subgenomes differ in their propensity for homoeolog copy loss

POInT uses user-defined phylogenetic Markov models of gene loss after WGT. These models have seven states (Fig. 2): the triplicated state **T**, in which all three copies from the WGT are still present; the "duplicated" states $\mathbf{D_{1,2}}$, $\mathbf{D_{1,3}}$, $\mathbf{D_{2,3}}$, in which one out of the three gene copies has been lost, and three single-copy states, $\mathbf{S_1}$, $\mathbf{S_2}$, and $\mathbf{S_3}$. Previous work suggested that the three subgenomes that formed these hexaploids are distinct in their patterns of gene preservation (Cheng et al. 2012; Tang et al. 2012), consisting of a less fractionated (LF) genome, a subgenome with intermediate levels of gene loss (more fractionated 1 or MF1), and an even more fractionated subgenome (MF2). We hence defined state $\mathbf{S_1}$ to correspond to LF and $\mathbf{S_2}$ and $\mathbf{S_3}$ to MF1 and MF2, respectively

POInT statistically assigned genes from each of the four mesopolyploid genomes to the LF, MF1, and MF2 subgenomes with high confidence: 75% of the pillars have subgenome assignments with posterior probabilities >0.84 (Supplemental Fig. S3). We observe clear signals of biased fractionation: although we estimate that 2864 genes were lost from the LF subgenome along the shared root branch (e.g., before the split of *S. alba* from the other three species), the corresponding figures for MF1 and MF2 are 5373 and 6347, respectively (Fig. 1). These values are in qualitative agreement with previous findings (Cheng et al. 2012, 2014; Liu et al. 2014; Xie et al. 2019).

We assessed the statistical support for these estimated differences in the subgenomes' rates of homoeolog loss using a set of nested models of post-WGT gene loss. We started with a model (WGT Null) that did not differentiate between the subgenomes, meaning that the shared base transition rate from **T** to $\mathbf{D_{1,2}}$, $\mathbf{D_{1,3}}$, or $\mathbf{D_{2,3}}$ is defined to be α ($0 \leq \alpha < \infty$) (Fig. 2). The transition rate from $\mathbf{D_{1,2}}$, $\mathbf{D_{1,3}}$, or $\mathbf{D_{2,3}}$ to $\mathbf{S_1}$, $\mathbf{S_2}$, or $\mathbf{S_3}$ is scaled by σ; that is, it occurs at rate $\alpha \times \sigma$. We compared this model to a more complex one that allowed losses of both triplicated and duplicated genes to be less frequent from a posited LF subgenome (WGT 1Dom) (Fig. 2). This model introduces a fractionation parameter f_1 ($0 \leq f1 \leq 1$), which potentially makes the transitions between **T** and $\mathbf{D_{2,3}}$ rarer than the other T-to-D rates ($\alpha \times f1$) (Fig. 2). The WGT 1Dom model fits the pillar data significantly better than does WGT Null (Fig. 2) ($P < 10^{-10}$, likelihood ratio test with two degrees of freedom). We next compared the WGT 1Dom model to a WGT 1Dom_{G3} model that gives MF1 and MF2 separate loss rates. Again, this model gives a better fit to the pillar data than did WGT 1Dom ($P < 10^{-10}$, likelihood ratio test with two degrees of freedom) (Fig. 2). We hence confirm the presence of three

subgenomes, distinguishable by their patterns of homoeolog loss. Our approach does not require the identification of these three subgenomes a priori: the probabilistic assignment of genes to subgenomes is an integral part of the POInT orthology computation. As a result, the inherent uncertainty in these assignments is accounted for in estimating the various biased fractionation parameters. Our orthology inferences can be explored visually with $POInT_{browse}$ (http://wgd.statgen.ncsu.edu/).

Patterns of post-WGT gene loss support the two-step model of hexaploidy

To test the hypothesis that the WGT proceeded in two steps (Cheng et al. 2012; Tang et al. 2012), we used two approaches. First, we applied an extended version of the WGT $1Dom_{G3}$ model in which each model parameter was allowed to take on distinct values on the root branch and on the remaining branches (Root-spec. WGT $1Dom_{G3}$) (Fig. 2). This extended model fits the pillar data significantly better than does the original WGT $1Dom_{G3}$ model ($P < 10^{-10}$, likelihood ratio test with five degrees of freedom) (Fig. 2). The biased fractionation parameters for the root branch differ from those of the remaining branches: the value of *f1,3* on the root is smaller than on later branches (0.6445 vs. 0.7368), whereas *f2,3* is larger (0.6766 vs. 0.4078). These values are consistent with a two-step hypothesis: before the arrival of LF, there would have been a number of losses from MF1 and MF2, meaning that the relative preference for LF would be higher (smaller *f1,3*).

In our second approach, we developed a specific model of the two-step hexaploidy (WGT $1Dom_{G3}$+$Root_{LF}$) (Fig. 2). This model describes the transition from a genome duplication to a triplication. All pillars start in state $\mathbf{D_{2,3}}$: that is, the first

allopolyploidy has just occurred and the MF1 and MF2 genes are present but not the LF ones. We then model the addition of LF as transitions to either the **T**, $\mathbf{D_{1,2}}$, or the $\mathbf{D_{1,3}}$ states (with rates τ, $\beta_{1,2}$, or $\beta_{1,3}$, respectively). State **T** is seen when no losses occurred before the arrival of LF, the other states occur when either MF1 or MF2 experienced a loss before the arrival of LF. Any pillars that remain in $\mathbf{D_{2,3}}$ had no corresponding gene arrive from LF. Of course, at the level of the individual pillar, we have insufficient data to make such inferences; the utility of this model is to give global estimates of the degree of fractionation seen in MF1 and MF2 before the arrival of LF. This model offers a significantly improved fit over WGT $1Dom_{G3}$ ($P < 10^{-10}$, likelihood ratio test with three degrees of freedom) (Fig. 2). More important, we can propose other versions of this model in which either MF1 or MF2 is the last arriving subgenome; when we do so, the model fit is much worse than seen with WGT $1Dom_{G3}$+$Root_{LF}$ model (Supplemental Table S1). Hence, we can conclude that subgenomes MF1 and MF2 had already begun a process of (biased) fractionation before the addition of the LF subgenome. These conclusions derive only from genes that were inferred to be present in all three parental subgenomes, a requirement of the POInT models.

A gap between the two allopolyploidies

This root-specific model also allows us to estimate the state of MF1 and MF2 immediately before the arrival of LF. In particular, we can estimate the percentage of pillars that had already experienced losses before LF's arrival. About 28% of all the MF1 homoeologs inferred to have been lost on the root branch were lost before the arrival of

LF, with the equivalent number of MF2 losses being 38%. A negligible 0.3% of pillars do not appear to have received a copy of the LF homoeolog.

Mixed evidence for differences in selective constraint between subgenomes

In our data set there are 218 loci that have retained triplicates in all four genomes and have subgenome assignment confidence ≥ 95%. For each locus we calculated the selective constraints acting on the group of 12 genes using codeml (Yang 2007), allowing the genes from each subgenome to have a different d_N/d_S value. On average, among these retained triplets, genes from the LF subgenome show slightly smaller d_N/d_S values than do those from MF1 and MF2, but these differences are not statistically significant (Wilcoxon rank-sum tests LF to MF1: $P = 0.300$, LF to MF2: $P = 0.079$) (Supplemental Fig. S4).

Single-copy genes from multiple subgenomes are enriched in genes functioning in DNA repair

GO overrepresentation tests were performed with the *Arabidopsis* orthologs of genes returned to single copy by the end of the root branch from each subgenome. Similar to previous findings (De Smet et al. 2013), we found that single-copy genes are enriched in biological processes such as DNA repair and DNA metabolism (Supplemental Fig. S5). More specifically, single-copy genes from the LF subgenome are enriched in base-excision repair, whereas MF1 single-copy genes are enriched in nucleotide-excision repair, non-recombinational repair, and double-strand break repair (Supplemental Fig. S5A). Single-copy genes from both LF and MF1 show

overrepresented molecular functions in endo- and exodeoxyribonuclease activities (Supplemental Fig. S5B). LF single-copy genes are also enriched in RNA interference processes, suggesting that such interference, targeted to the MF1 and MF2 subgenomes, could be one mechanism by which biased fractionation was driven.

Genes from the same subgenome are not overly likely to physically or metabolically interact

For genes with high subgenome assignment confidence (≥95%), we mapped those assignments (LF, MF1, or MF2) and the duplication status at the end of the root branch onto the nodes (gene products) of the *A. thaliana* protein–protein interaction (PPI) network (Methods). For comparative purposes, we also produced a mapping of an extant network, based on the gene presence/absence data and subgenome assignments in *B. rapa*. In the "ancient" network inferred at the end of the common root branch, there are a relatively large number of nodes (1952) associated with surviving triplicated loci; these nodes were connected by a total of 2384 triplet-to-triplet edges. The *B. rapa*-specific network contains fewer nodes with retained triplets (662), and there were 263 edges connecting these nodes (Fig. 3A).

The dosage constraints that affect surviving gene copies post-polyploidy will tend to result in the retention of genes involved in multiunit complexes or in the same signaling pathways (Birchler and Veitia 2007, 2012; Conant et al. 2014). Thus, we expected to see that the retained triplets showed higher network connectivity. And indeed, our permutation tests reveal that the retained triplets on the root branch are significantly overconnected to each other in the PPI network ($P = 0.018$) (Supplemental Fig. S6). We

also hypothesized that proteins coded for from the same subgenome would be more likely to be connected because of preferential retention of genes from a single complex from the same subgenome. To test this idea, we partitioned the gene products based on their subgenome of origin. The LF subgenome contains more genes and thus more exclusive connections (Fig. 3B). When considering only genes that had returned to single copy by the end of the root (Fig. 3C), we identified 188 LF-LF edges among 886 single-copy LF genes, with fewer edges exclusive to MF1 and MF2 genes (30 and 3, respectively). We used randomization (Methods) to test whether the numbers of such subgenome-specific edges differed from what would be expected by chance. When considering the network as a whole, we found that there were significantly fewer LF-LF edges than expected (P = 0.022) (Supplemental Fig. S6). However, when we considered only the single-copy genes in the network, the number of subgenome-specific edges did not differ from that seen in random networks for any of the three subgenomes (P = 0.286 for LF-LF edges) (Supplemental Fig. S6), suggesting that the original dearth of such edges was a statistical artifact resulting from the excess of triplet-to-triplet edges.

We also explored the association of between genes' role in metabolism and their pattern of post-hexaploidy evolution using the *A. thaliana* metabolic network (Methods). However, again considering the state of each pillar at the end of the root branch, we did not find an excess of shared metabolic interactions between triplicated or single-copy genes in this network (Supplemental Fig. S6).

Finally, we asked whether genes from the same subgenome are more likely to be coexpressed. We constructed a *B. rapa* coexpression network from the RNA-seq data described in Methods. In this network, edges connect pairs of genes that are highly

correlated in their expression (Methods). The inferred coexpression network contains 3933 nodes (e.g., genes) from the LF subgenome, 2310 nodes from MF1, and 1982 from MF2. We then counted the number of edges connecting pairs of nodes from the same subgenome. To assess whether there was an excess of such shared subgenome coexpression relationships, we randomly rewired the network 100 times and compared the edge count distributions from these randomized networks to those of the real network (Pérez-Bercoff et al. 2011). We found that the real network did not show a significant excess of shared edges between genes from the same subgenome when compared to the randomized networks (LF-LF, $P = 0.36$; MF1-MF1, $P = 0.82$; MF2-MF2 $P = 0.08$) (Fig. 4A–F).

Subgenome of origin does not affect the propensity to have experienced a selective sweep

We tested for associations between genes' subgenome of origin and their propensity to experience recent selective sweeps. Data on these sweeps was taken from a recent scan in *B. rapa* by Qi et al. (2021). No subgenome had either an excess or a deficit of observed sweeps relative to the other two (Supplemental Fig. S7). Genes from the MF1 subgenome showed slightly negative association with selective sweeps ($P = 0.0089$, χ^2 test).

DISCUSSION

The combination of the new genome sequence of *Crambe hispanica* and our modeling of the post-WGT evolution of the four Brassiceae genomes using POInT allowed us to draw a number of conclusions regarding the Brassiceae WGT. We confirmed previous work (Cheng et al. 2012; Tang et al. 2012) arguing that these genomes derive from a pair of ancient allopolyploidies: more subtly, we also show that, as had been proposed, the least fractionated subgenome (e.g., the one with the most retained genes) is very likely the genome that was added last. To these proposals, we add the new observation that these hybridization events were likely not particularly closely spaced in time: our model predicts that on the order of one-third of the gene losses from subgenomes MF1 and MF2 that occurred on the root branch occurred before the arrival of the LF subgenome. Of course, one should not take this result to necessarily imply a very large number of calendar years between the events; gene loss immediately after polyploidy can be quite rapid (Scannell et al. 2007; De Smet et al. 2013). In the future, it will be interesting to further refine the timing of these events; the problem, however, is a challenging one because the allopolyploid nature of the events means that molecular clock approaches will tend to estimate speciation times for the allopolyploid ancestors rather than hybridization times.

Many forces shape genome evolution after polyploidy. A tendency for genes that operate in multiunit complexes or that are involved in signaling cascades to remain duplicated post-polyploidy is best explained by the presence of dosage constraints driven by a need to maintain the stoichiometry and kinetics of assembly for such functional units

(Birchler et al. 2005, 2016; Birchler and Veitia 2007, 2012; Conant et al. 2014). On the other hand, genes involved in functions such as DNA repair very often return rapidly to singleton status after duplication (Freeling 2009; De Smet et al. 2013). Our results illustrate the importance of these dosage effects, with genes whose products interact with many other gene products in *A. thaliana* being overly likely to be retained in triplicate in these Brassicae genomes. This pattern is not observed for metabolic genes, a result we interpret as illustrating metabolism's dynamic robustness to gene dosage changes (Kacser and Burns 1981).

We had previously argued that one force driving the biased fractionation that distinguishes the LF, MF1, and MF2 subgenomes might be selection to maintain coadapted complexes from a single parental subgenome (Emery et al. 2018). That such coadapted complexes exist and respond to polyploidy is suggested by the gene conversions seen after the yeast polyploidy among the duplicated ribosomal and histone proteins (Evangelisti and Conant 2010; Scienski et al. 2015). However, these examples may be exceptions rather than the rule, meaning that pressure to maintain coadapted complexes is not a significant driver of biases in fractionation. We found that although there was some degree of functional distinction for single-copy genes from the LF subgenome (e.g., enrichment in biological processes such as DNA repair and RNA interference), more generally speaking, there was no significant evidence of functional incompatibilities between single-copy genes from different subgenomes. Thus, genes from the same subgenome were not more likely to interact with each other physically, nor were the genes returned to single copy on the common root branch functionally subdivided among the subgenomes. Even the DNA repair enzyme genes that rapidly

returned to single copy appear to derive from at least two of the three subgenomes. It hence appears that the original hypothesis of De Smet et al. (2013) that these genes may be prone to dominant negative interactions may best explain their preference for a single-copy state.

It remains to be seen if the "mix and match" pattern of subgenome retention observed here represents the dominant mode of evolution for allopolyploidies. Of course, whether or not subgenome conflicts exist may be partly a question of the preexisting differences between the progenitor species, and a more general survey of allopolyploidies that includes estimates of the progenitor genomes' divergence before the polyploidy events would be most enlightening. If the pattern holds, however, the implications would be significant, because hybridization represents an important means of adaption (Paterson 2005; Hollister 2015; Alix et al. 2017; Blanc-Mathieu et al. 2017; Smukowski Heil et al. 2017). Adding the effects of hybridization to polyploidy's known association with innovation (Edger et al. 2015) and to the tendency of dosage sensitive genes to remain duplicated for the longer periods needed for such innovations (Blanc and Wolfe 2004; Conant and Wolfe 2008b; Conant et al. 2014; Zhao et al. 2017; Liang and Schnable 2018; Qiu et al. 2020) makes a strong case for considering polyploidy a critical source of material for innovation at the genomic level.

METHODS

Crambe hispanica (PI 388853) sample preparation and genome sequencing

Leaf tissue was harvested from 36 dark treated inbred plants (selfed for nine generations; PI 388853). Dark treatment was performed to reduce chloroplast abundance

and involved leaving the plants in a dark room for 3–4 d. After treatment, 5 g of tissue was collected across 36 plants. This process was repeated three times, allowing us to obtain a total of 15 g of tissue. This tissue was then sent to the University of Delaware Sequencing and Genotyping Center at the Delaware Biotechnology Institute for high molecular weight DNA isolation and library preparation before PacBio and Illumina sequencing. Libraries were prepared using standard SMRTbell procedures, followed by sequencing of 11 PacBio SMRT cells on a PacBio sequel and one PacBio SMRT cell of RSII sequencing. Paired-end 150-bp reads were generated on an Illumina HiSeq 2500 system. For Hi-C scaffolding, 0.5 g tissue sample was sent to Phase Genomics.

Crambe hispanica v1.1 genome assembly and annotation

The assembly of the *Crambe hispanica* v1.1 genome was performed using Canu v1.6 (Koren et al. 2017). In total, 3.9 million raw PacBio reads spanning 48 Gb were used as input for Canu. The following parameters were modified for assembly: minReadLength = 1000, GenomeSize = 500 Mb, corOutCoverage = 200 “batOptions=-dg 3 -db 3 – dr 1 -ca 500 -cp 50”. All other parameters were left as default. The assembly graph was visualized using Bandage (Wick et al. 2015) to assess ambiguities in the graph related to repetitive elements and heterozygosity. The draft Canu assembly was polished reiteratively using high-coverage Illumina paired-end data (82 million reads) with Pilon v1.22 (Walker et al. 2014). Quality filtered Illumina reads were aligned to the genome using Bowtie 2 (v2.3.0) (Langmead and Salzberg 2012) under default parameters, and the resulting BAM file was used as input for Pilon with the following parameters: --flank 7, -

-K 49, and --mindepth 8. Pilon was run recursively three times using the updated reference each time to correct the maximum number of residual errors.

A Proximo Hi-C library was prepared as described (Phase Genomics) and sequenced on an Illumina HiSeq 2500 system with paired-end 150 bp reads. The de novo genome assembly of Hi-C library reads were used as input data for the Phase Genomics Proximo Hi-C genome scaffolding platform.

The genome was annotated using MAKER (Campbell et al. 2014), using evidence of protein sequences downloaded from the Araport 11 and Phytozome 12 plant databases (Goodstein et al. 2012; Cheng et al. 2017) and *C. hispanica* transcriptome data. The transcriptome data for genome annotation was extracted from bud, root, and leaf tissues under standard daylight conditions using the Thermo Fisher Scientific PureLink RNA Mini Kit. Library prep was done using Illumina TruSeq DNA PCR-free and sequenced for nonstranded mRNA-Seq 2 × 250 on Illumina HiSeq. *C. hispanica* transcriptomic data were assembled with StringTie (Pertea et al. 2015). Repetitive regions in the genome were masked using a custom repeat library and Repbase Update (Bao et al. 2015) through RepeatMasker Open-4.0 (Smit et al. 2015). Ab initio gene prediction was performed using SNAP (Korf 2004) and AUGUSTUS (Stanke and Waack 2003). The resulting MAKER gene set was filtered to select gene models with Pfam domain and annotation edit distance (AED) <1.0. Then, the amino acid sequences of predicted genes were searched against a transposase database using BLASTP and an *E*-value cutoff of 10^{-10} (Campbell et al. 2014). If >30% of a given gene aligned to transposases after the removal of low complexity regions, that gene was removed from the gene set.

Triple-conserved synteny reconstruction

We developed a three-step pipeline for inferring the conserved synteny blocks created by polyploidy (Emery et al. 2018). For the first step of this pipeline, we used *Arabidopsis thaliana* Col-0 version 10.29 (CoGe genome id 20342) as a nonhexaploid outgroup and identified homologous genes between it and each of the four hexaploid genomes using GenomeHistory (Conant and Wagner 2002). Genes were defined as homologous if their translated products shared 70% amino acid sequence identity and the shorter sequence was at least 80% of the length of the longer. In the second step, we sought to place genes from each of the hexaploid genomes into blocks of triple-conserved synteny (TCS) relative to their *A. thaliana* homologs. To do so, we inferred a set of "pillars," each of which contains a single gene (or group of tandem duplicates) from *A. thaliana* and between one and three genes from the hexaploidy genome. Using simulated annealing (Kirkpatrick et al. 1983; Conant and Wolfe 2006), we sought a combination of pillar gene assignments and relative pillar order that maximized the TCS. In the third and final step, we merged the pillars across the four hexaploid genomes, using their *A. thaliana* homologs as indices. We then sought a global pillar order that minimized the number of synteny breaks across all of the hexaploid genomes (Supplemental Fig. S2). These three steps resulted in a set of 14,050 ordered pillars, each with at least one surviving gene from each of the four genomes (Fig. 1) and a corresponding "ancestral" gene from *A. thaliana*. Supplemental Table S1 shows that POInT's model inferences are consistent across a number of such estimated ancestral orders.

An ancestral genome order reconstruction

As a verification of our POInT pipeline, we also sought an independent inference of the order of the genes in the parental subgenomes just before the first step of the *Brassica* triplication. First, we used CoGe's SynMap (Lyons et al. 2008b) to identify homologs between the *A. thaliana* and *Arabidopsis lyrata* genomes and those between *B. rapa* and *B. oleracea.* The SynMap algorithm was applied with a chaining distance of 50 genes and a minimum of five aligned gene pairs to identify likely orthologous genes in all pairwise comparisons of the four genomes. Paralogs were identified by self-comparisons of each of the two *Brassica* genomes with SynMap. Then these orthologs and paralogs were grouped into 24,011 homology sets with the OMG! program (Zheng et al. 2011). Every homology set consists of one to three *Brassica* paralogs from each of the three *Brassica* genomes and a single *Arabidopsis* gene from each of the two *Arabidopsis* genomes, representing one "candidate gene" in the reconstructed ancestral genome. Among these, 2178 homology sets contained the maximum of eight genes (one each from the two *Arabidopsis* genomes and three each from the two *Brassica* genomes).

The homology sets were used to retrieve the ancestral gene order from an adjacency graph using an efficient algorithm called Maximum Weight Matching (MWM) (Zheng et al. 2013). We identified all the gene adjacencies in the four genomes, considering only the genes in the homology sets. Each adjacency was then weighted according to how many of the eight possible adjacencies were actually observed. The MWM produced an optimal set of 10,944 linear contigs containing all 24,001 putative ancestral genes from the homology sets that included 13,057 of 45,982 total adjacencies

in the data set, with the remaining adjacencies being inconsistent with this optimal set. We used the contigs in the output of the MWM to reconstruct each of the five ancestral chromosomes. There were 34 contigs containing large proportions of genes originating in two or more of the ancient chromosomes that were discarded, as were any contigs containing four or fewer genes from a *Brassica* genome. Although the 9712 contigs so omitted represent 89% of all contigs, they represent only 55% of the genes, leaving a small group of large contigs with strong synteny relations in our ancestral reconstruction. We next identified adjacencies among the contigs themselves and applied a second iteration of MWM on them, giving the optimal ordering of those contigs. Combining these orders with the existing gene order information within each contig yields the position of all the genes on each ancestral chromosome. This order was mapped to our set of pillars of TCS, giving a subset of those pillars ordered by this ancestral order estimate.

The phylogenetic relationships of the triplicated members of the Brassicaceae

POInT fits the models shown in Figure 2 to the pillar data under an assumed phylogenetic topology using maximum likelihood, allowing us to use that likelihood statistic to compare different phylogenetic relationships among these four hexaploid taxa. POInT's computational demands were too great to allow testing all 15 rooted topologies of four species (POInT's models are not time reversible). However, by making the reasonable assumption that *B. rapa* and *B. oleracea* are sister to each other, we were able to test the three potential relationships of *C. hispanica* and *S. alba* to the two *Brassicas.* Figure 1 gives the maximum likelihood topology: the two alternative topologies and their likelihoods are given in Supplemental Figure S1.

Selective constraints of the retained triplets

We identified 218 pillars that retained triplicated genes across all four genomes and for which the confidence in their subgenome assignments was ≥ 95%. For each such pillar, the 12 sequences were aligned using T-coffee (Notredame et al. 2000). The cladogram for each such set of 12 genes consists of three subtrees grouping four sequences that belong to the same subgenome in the same sister group (Supplemental Fig. S4). Using codeml in PAML (Yang 2007) with CodonFreq set to F3 × 4, we inferred three distinct d_N/d_S ratios, one for each of the three subtrees deriving from the three parental subgenomes.

Functional analysis of single-copy genes from different subgenomes

We performed functional analysis for genes where we have high (≥95%) confidence that they returned to single copy along the common root branch. Using the corresponding "ancestral" locus from *A. thaliana,* we performed individual Gene Ontology analyses with PANTHER (Mi et al. 2019) overrepresentation tests (release date 20190711) for genes from each subgenome. The background list used in all cases was the loci that remained duplicated or triplicated at the end of the root branch.

Protein–protein interaction and metabolic network analysis

The *A. thaliana* protein–protein interaction (PPI) network was downloaded from BioGRID (Arabidopsis Interactome Mapping Consortium 2011; Stark et al. 2011). The root branch post-WGT subgenome assignments for each "ancestral" locus represented by

an *Arabidopsis* gene were mapped onto the nodes (gene products) of the PPI network, so long as our confidence in those subgenome assignments was ≥95%. Similarly, for the extant *B. rapa*, we took loci with high subgenome assignment confidence ≥95% and mapped their *A. thaliana* orthologs onto network nodes. The resulting PPI network (Fig. 3) was visualized using Gephi 0.9.2 (Bastian et al. 2009) with the Fruchterman Reingold and Yifan Hu layout algorithms (Fruchterman and Reingold 1991; Hu 2006). To test whether gene products from the same subgenome are overconnected in this network, we permutated the subgenome assignments 1000 times, holding the network topology unchanged. We then compared the actual number of edges connecting single-copy genes from the same subgenome with the distribution of this value seen in the randomized networks (Supplemental Fig. S6). We also asked whether the ancestral genes corresponding to retained triplets showed an excess of connections among themselves. Because the number of edges between retained triplets and between single-copy genes are not independent, we performed an additional set of permutations, in which we held all the triplet rows constant and only shuffled the subgenome assignments of the remaining nodes.

We performed similar analyses using the AraGEM v1.2 metabolic network from *A. thaliana* (de Oliveira Dal'Molin et al. 2010; Bekaert et al. 2012). In this network, each node represents a biochemical reaction, and pairs of nodes are connected by edges if their respective reactions share a metabolite. For each *A. thaliana* gene encoding an enzyme catalyzing one such reaction, we mapped the root branch subgenome assignments (again requiring ≥95% confidence), assigning to that gene three presence/absence variables (one per subgenome). Then, for each subgenome, we counted

the number of edges between pairs of nodes with at least one pair of single-copy genes from a common subgenome. We assessed significance by holding the network topology and *Arabidopsis* gene assignments constant and randomizing the subgenome assignments 1000 times. We then compared the distributions of the single-subgenome edge counts from the simulations with the actual values (Supplemental Fig. S6).

Brassica rapa coexpression network analysis

We generated a gene expression data set for *Brassica rapa* spanning diverse experimental conditions, including the following: a cold treatment in leaves (4 h and 28 h post), methyl jasmonate treatment in leaves (4 h and 28 h post), anaerobic treatment in leaves (4 and 8 h post), salt treatment in roots (4 h and 28 h post), and a diurnal time course in leaves (every 4 h, six time points) in standard light-dark conditions but also in complete dark and complete light conditions. Total RNA was extracted from above organs using the Invitrogen Purelink RNA Mini Kit (Thermo Fisher Scientific), converted into a library using the Illumina TruSeq RNA kit, and paired-end 100-bp reads were sequenced on the HiSeq 2000 instrument at the VJC Genomics Sequencing Laboratory at the University of California, Berkeley. The NextGENe V2.17 (SoftGenetics) software package was used to remove low-quality Illumina data, map reads to the *B. rapa* FPsc (v1.0, CoGe id 20101) reference genome, and calculate normalized reads per kilobase of transcript per million (RPKM) values for all genes.

We filtered the data set to only include genes that were missing a measured expression value for at most one of the 32 RNA-seq libraries, leaving 24,907 *B. rapa* genes in it. The gene identifiers used for the expression data set were from the *B.*

rapa FPsc (v1.0, CoGe id 20101) reference genome, so we translated these identifiers to those from *B. rapa* Chiifu (v1.5, id 24668) using CoGe SynMap (Lyons et al. 2008b). In so doing, we only used *B. rapa* genes with one-to-one matches between the two references. For any pair of genes in the expression data set, we calculated the Spearman's correlation coefficient of their RPKM values. A coexpression network was then constructed using highly correlated gene pairs, for example, pairs having Spearman's correlation coefficients ≥0.9 (positive correlations), or ≤−0.9 (negative correlations). Thus, the nodes of this coexpression network are *B. rapa* genes, and the edges represent correlation in expression. The coexpression network was randomized 100 times by rewiring the edges while holding the nodes and their subgenome assignments unchanged. In other words, all edges were broken and randomly reconnecting, preserving the degree of every node (Pérez-Bercoff et al. 2011). The distributions of inter-subgenome and intra-subgenome edge counts are shown in Figure 4.

Association between recent selective sweeps in *B. rapa* and subgenomes origin

B. rapa genes were divided into those in the regions of selective sweeps detected by SweeD (Pavlidis et al. 2013) in either turnip, toria, Indian sarson, pak choi, or Chinese cabbage (vegetable types of *B. rapa*) and those showing no such signatures (Qi et al. 2017, 2021). We tested whether particular subgenomes (posterior probability ≥0.95) were unusually likely or unlikely to have experienced a selective sweep using χ^2 test. The association plot as shown in Supplemental Figure S7 was visualized using the vcd package version 1.4-4 in R 3.6.0 (Meyer et al. 2006; Zeileis et al. 2007; R Core Team 2019).

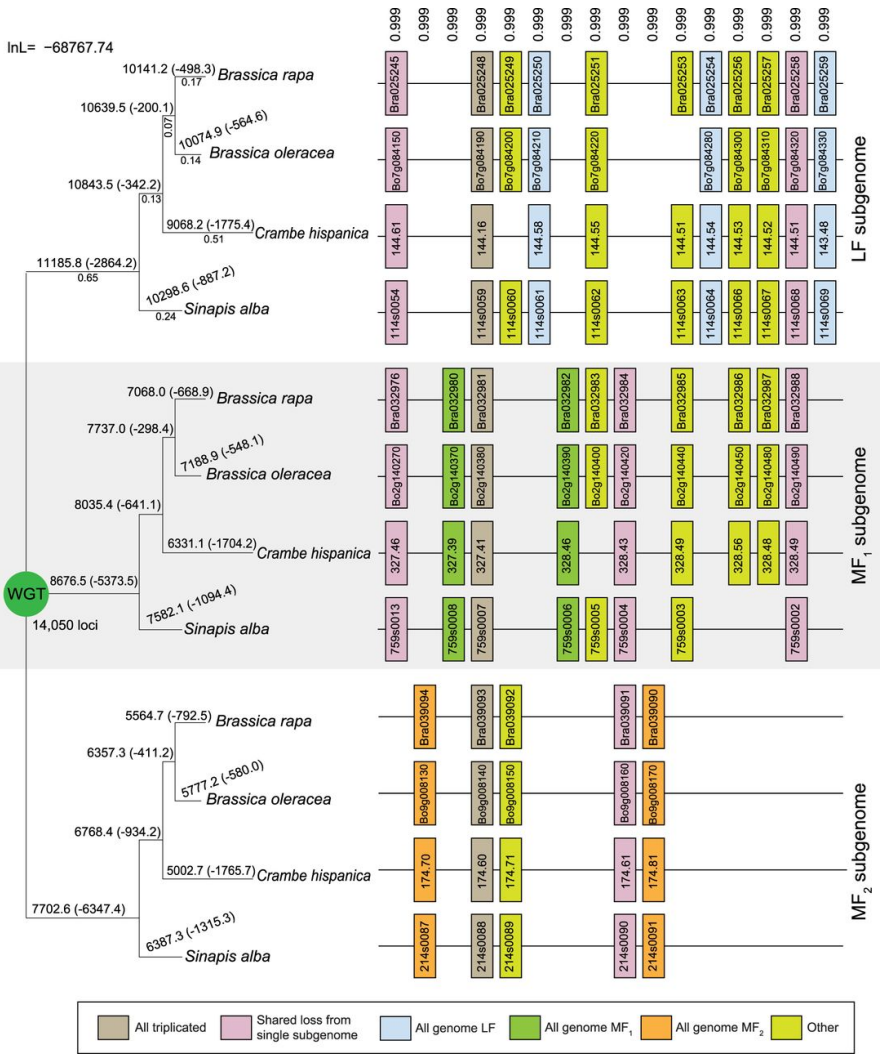


Figure B-1: Subgenome assignment and inference of gene loss after the shared WGT in four species. After the WGT, each ancestral locus could potentially expand to three gene copies, but owing to biases in the loss events, the number of surviving genes from the subgenomes are unequal. Our analyses (Results) indicate the presence of a less fractionated (LF) subgenome and two more fractionated ones (MF1 and MF2). These inferences are based on the gene losses observed across four genomes and along the phylogeny depicted. Shown here is a window of 16 post-WGT loci (of the total 14,050 such loci) in four species that share the WGT: *Brassica rapa*, *Brassica oleracea*, *Crambe hispanica*, and *Sinapis alba*. Each pillar corresponds to an ancestral locus, and the boxes represent extant genes. Pairs of genes are connected by lines if they are genomic neighbors (e.g., in synteny). The numbers *above* each pillar are the posterior probabilities assigned to this combination of orthology relationships relative to the other $(3!)^4-1 = 1295$ possible orthology states. The numbers *above* each branch of the tree give the number of genes in each subgenome surviving to that point, with the number of gene losses in parentheses. The gene loss inferences made by POInT are probabilistic: because some gene losses cannot be definitively assigned to a single branch, the resulting loss estimates are not integers. The numbers *below* the branches in the first subtree are POInT's branch length estimates (αt).

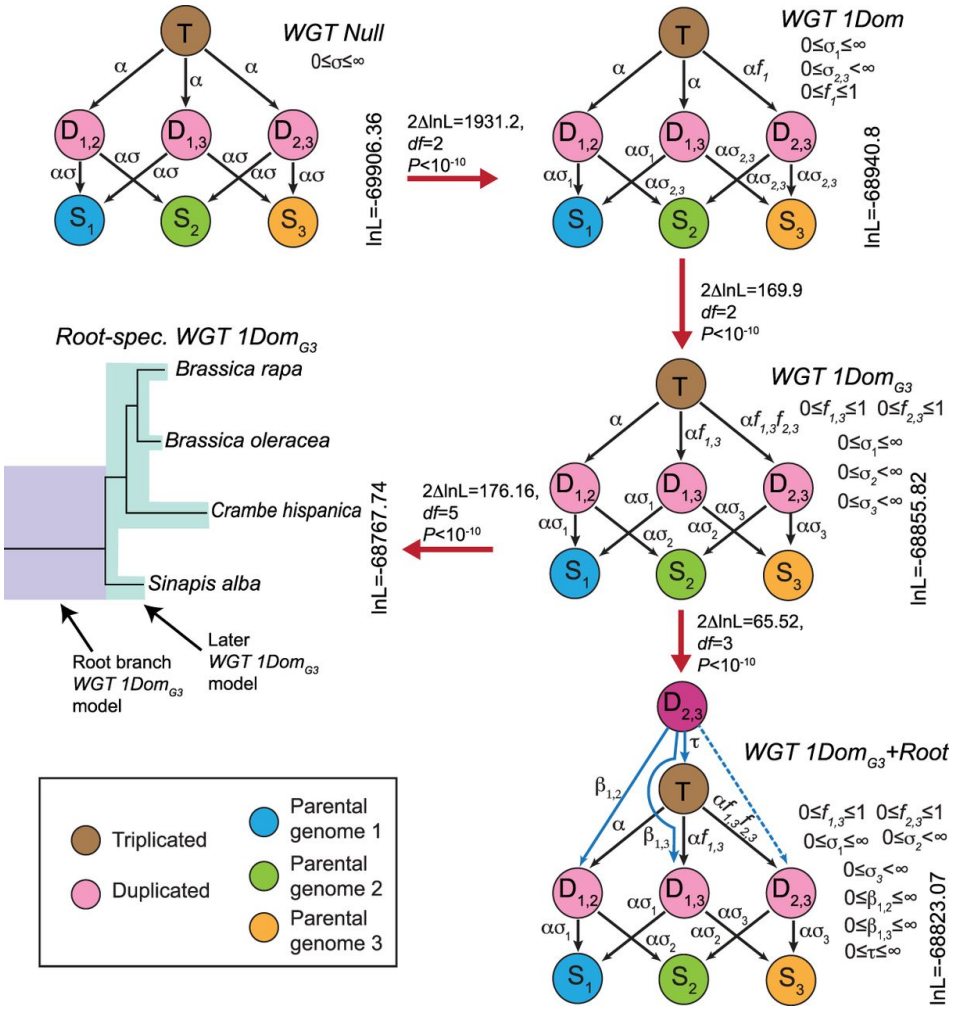


Figure B-2: POInT's models for inferring WGT. Five different models of post-WGT evolution and their ln-likelihoods are shown. In each model, the colored circles represent different states. The brown circle represents the triplicated state (**T**); the pink circles are duplicated states (**$D_{1,2}$**, **$D_{1,3}$**, and **$D_{2,3}$**); the blue, green, and yellow circles are three single-copy states (**S_1** for the LF subgenome, **S_2** for the MF1 subgenome, and **S_3** for the MF2 subgenome). The transition rates between states are shown *above* the arrows: (α) transition rate from triplicated state to duplicated states; (ασ) transition rates from duplicated states to single-copy states; (f) fractionation parameters; (β and τ) root model parameters. Red arrows connect pairs of models compared using likelihood ratio tests (Methods). In the WGT Null model, transition rates are the same across three subgenomes, modeling the scenario of no biased fractionation. In the WGT 1Dom model with the biased fractionation parameter *f1* ($0 \leq f1 \leq 1$), the MF1 and MF2 subgenomes are more fractionated than LF subgenome. In the WGT $1Dom_{G3}$ model, two fractionation parameters *f1,3* and *f2,3* were introduced, distinguishing the three subgenomes: MF2 is more fractionated than MF1, and MF1 is more fractionated than LF. The Root-spec. WGT $1Dom_{G3}$ model is similar to the previous model, but with two sets of parameters, one set for the root branch and the other for the remainder of the branches. The WGT $1Dom_{G3}$ + Root model is a two-step hexaploidy model created by starting each pillar in an intermediate state **$D_{2,3}$**. This state represents the merging of the MF1 and MF2 subgenomes as the first step of the hexaploid formation. The **T, $D_{1,2}$**, and **$D_{1,3}$** states represent the second allopolyploidy, with either no prior homoeolog losses (**T**) or a loss from one of the two MF subgenomes before that event (**$D_{1,2}$,** or **$D_{1,3}$**).

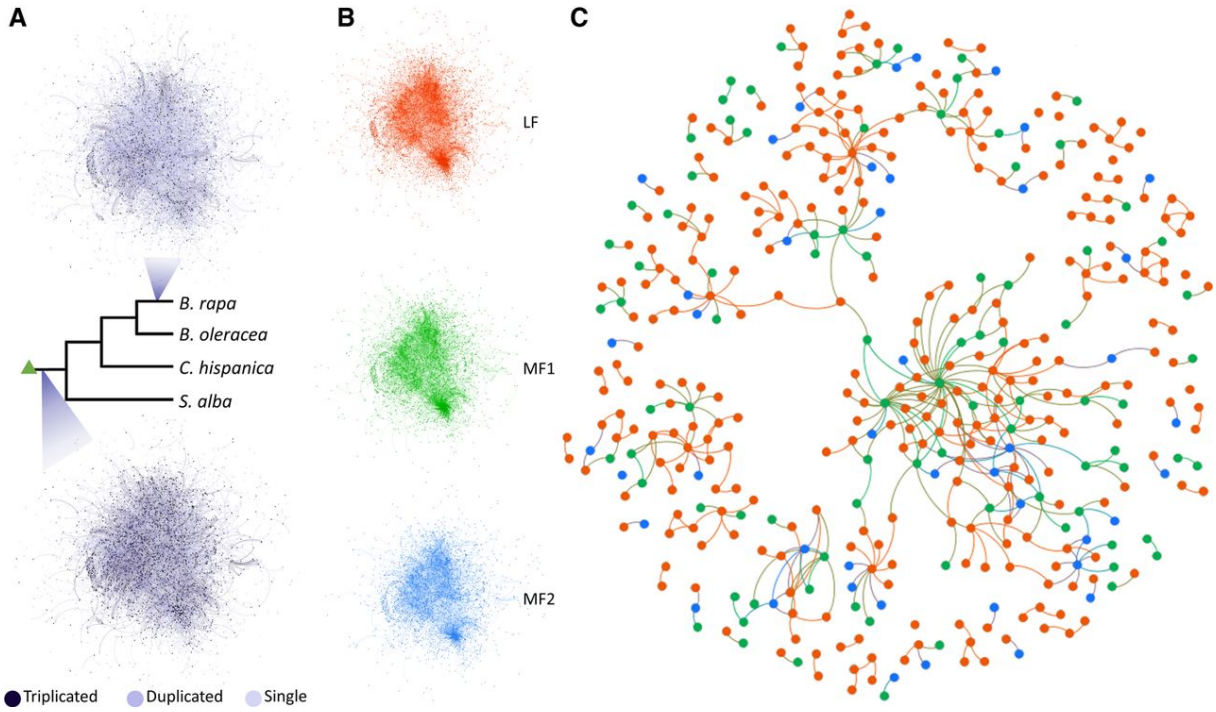


Figure B-3: Protein–protein interaction networks after the WGT. (*A*) The *Arabidopsis* PPI network at the root branch (*bottom*), and the same PPI network colored by the *Brassica rapa* gene retention status (*top*). The dark purple nodes represent retained triplets. (*B*) The PPI network partitioned by subgenome assignment at the root branch: (LF) red, 4249 nodes and 8454 edges; (MF1) green, 3379 nodes and 6442 edges; (MF2) blue, 3073 nodes and 4961 edges. (*C*) A subset of the PPI network where only nodes encoded by single copies genes and connected to other single-copy nodes are shown. Red nodes are from the LF subgenome, green nodes are from the MF1 subgenome, and blue nodes are from the MF2 subgenome.

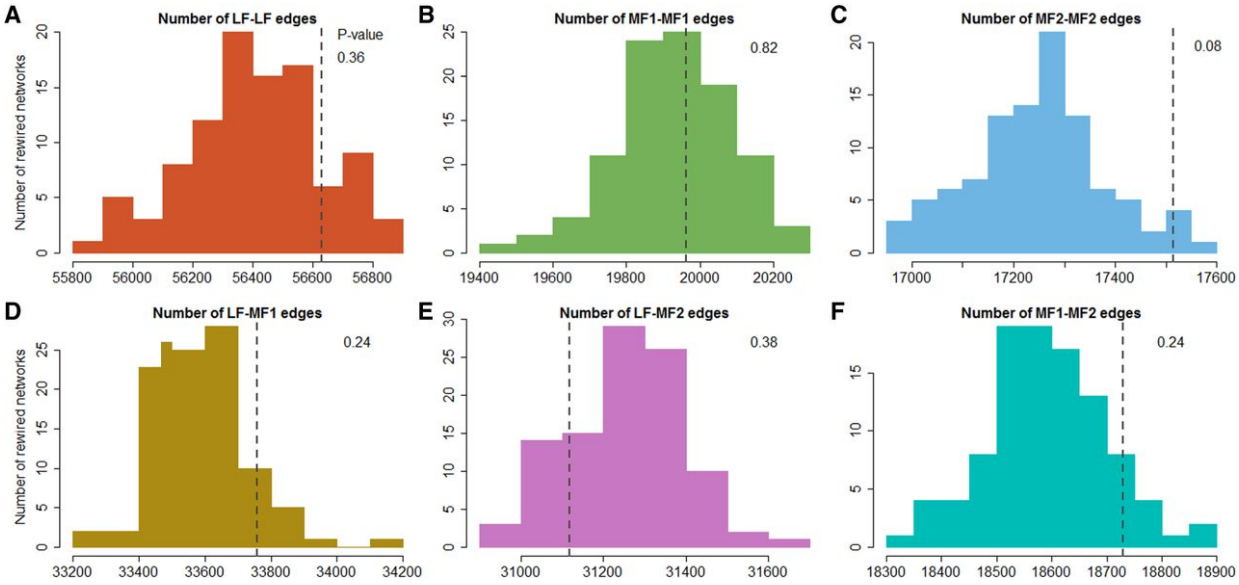


Figure B-4: Subgenome-specific edge counts for 100 rewired *Brassica rapa* coexpression networks compared to those from the actual network. (*A*) Distribution of the number of edges connecting pairs of *B. rapa* genes from the LF subgenome in 100 rewired networks. (*B*) Distribution of the number of edges connecting pairs of genes from the MF1 subgenome. (*C*) Distribution of the number of edges connecting pairs of genes from the MF2 subgenome. (*D*) Distribution of the number of edges connecting LF genes to MF1 genes. (*E*) Distribution of the number of edges connecting LF genes to MF2 genes. (*F*) Distribution of the number of edges connecting MF1 and MF2 genes. In each panel, the dark gray dashed line shows the number of edges with that set of subgenome assignments for the true network.

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. https://doi.org/10.1016/j.pbi.2020.03.004.

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. https://doi.org/10.1093/aob/mcx079.

Arabidopsis Interactome Mapping Consortium. 2011. "Evidence for Network Evolution in an Arabidopsis Interactome Map." *Science* 333 (6042): 601–7. https://doi.org/10.1126/science.1203877.

Arias, Tatiana, and J. Chris Pires. 2012. "A Fully Resolved Chloroplast Phylogeny of the Brassica Crops and Wild Relatives (Brassicaceae: Brassiceae): Novel Clades and Potential Taxonomic Implications." *Taxon* 61 (5): 980–88. https://doi.org/10.1002/tax.615005.

Aury, Jean-Marc, Olivier Jaillon, Laurent Duret, Benjamin Noel, Claire Jubin, Betina M. Porcel, Béatrice Ségurens, et al. 2006. "Global Trends of Whole-Genome Duplications Revealed by the Ciliate Paramecium Tetraurelia." *Nature* 444 (7116): 171–78. https://doi.org/10.1038/nature05230.

Bao, Weidong, Kenji K. Kojima, and Oleksiy Kohany. 2015. "Repbase Update, a Database of Repetitive Elements in Eukaryotic Genomes." *Mobile DNA* 6 (June): 11. https://doi.org/10.1186/s13100-015-0041-9.

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. 2009. “Gephi: An Open Source Software for Exploring and Manipulating Networks.” In *Third International AAAI Conference on Weblogs and Social Media*. aaai.org. https://www.aaai.org/ocs/index.php/ICWSM/09/paper/viewPaper/154.

Bekaert, Michaël, Patrick P. Edger, Corey M. Hudson, J. Chris Pires, and Gavin C. Conant. 2012. “Metabolic and Evolutionary Costs of Herbivory Defense: Systems Biology of Glucosinolate Synthesis.” *The New Phytologist* 196 (2): 596–605. https://doi.org/10.1111/j.1469-8137.2012.04302.x.

Birchler, James A., Adam F. Johnson, and Reiner A. Veitia. 2016. “Kinetics Genetics: Incorporating the Concept of Genomic Balance into an Understanding of Quantitative Traits.” *Plant Science: An International Journal of Experimental Plant Biology* 245 (April): 128–34. https://doi.org/10.1016/j.plantsci.2016.02.002.

Birchler, James A., Nicole C. Riddle, Donald L. Auger, and Reiner A. Veitia. 2005. “Dosage Balance in Gene Regulation: Biological Implications.” *Trends in Genetics: TIG* 21 (4): 219–26. https://doi.org/10.1016/j.tig.2005.02.010.

Birchler, James A., and Reiner A. Veitia. 2007. “The Gene Balance Hypothesis: From Classical Genetics to Modern Genomics.” *The Plant Cell* 19 (2): 395–402. https://doi.org/10.1105/tpc.106.049338.

———. 2012. “Gene Balance Hypothesis: Connecting Issues of Dosage Sensitivity across Biological Disciplines.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (37): 14746–53. https://doi.org/10.1073/pnas.1207726109.

———. 2014. “The Gene Balance Hypothesis: Dosage Effects in Plants.” *Methods in Molecular Biology* 1112: 25–32. https://doi.org/10.1007/978-1-62703-773-0_2.

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. https://doi.org/10.1111/nph.15256.

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. https://doi.org/10.1105/tpc.021410.

Blanc-Mathieu, Romain, Laetitia Perfus-Barbeoch, Jean-Marc Aury, Martine Da Rocha, Jérôme Gouzy, Erika Sallet, Cristina Martin-Jimenez, et al. 2017. “Hybridization and Polyploidy Enable Genomic Plasticity without Sex in the Most Devastating Plant-Parasitic Nematodes.” *PLoS Genetics* 13 (6): e1006777. https://doi.org/10.1371/journal.pgen.1006777.

Campbell, Michael S., Meiyee Law, Carson Holt, Joshua C. Stein, Gaurav D. Moghe, David E. Hufnagel, Jikai Lei, et al. 2014. “MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations.” *Plant Physiology* 164 (2): 513–24. https://doi.org/10.1104/pp.113.230144.

Carlsson, A. S., D. Clayton, M. Toonen, and E. M. J. Salentijn. 2007. “Oil Crop Platforms for Industrial Uses: Outputs from the EPOBIO Project.” https://library.wur.nl/WebQuery/wurpubs/525126.

Cheng, Chia-Yi, Vivek Krishnakumar, Agnes P. Chan, Françoise Thibaud-Nissen, Seth Schobel, and Christopher D. Town. 2017. “Araport11: A Complete Reannotation

of the Arabidopsis Thaliana Reference Genome." *The Plant Journal: For Cell and Molecular Biology* 89 (4): 789–804. https://doi.org/10.1111/tpj.13415.

Cheng, Feng, Jian Wu, Lu Fang, Silong Sun, Bo Liu, Ke Lin, Guusje Bonnema, and Xiaowu Wang. 2012. "Biased Gene Fractionation and Dominant Gene Expression among the Subgenomes of Brassica Rapa." *PloS One* 7 (5): e36442. https://doi.org/10.1371/journal.pone.0036442.

Cheng, Feng, Jian Wu, and Xiaowu Wang. 2014. "Genome Triplication Drove the Diversification of Brassica Plants." *Horticulture Research* 1 (May): 14024. https://doi.org/10.1038/hortres.2014.24.

Codoñer, Francisco M., and Mario A. Fares. 2008. "Why Should We Care about Molecular Coevolution?" *Evolutionary Bioinformatics Online* 4 (February): 29–38. https://www.ncbi.nlm.nih.gov/pubmed/19204805.

Conant, Gavin C. 2014. "Comparative Genomics as a Time Machine: How Relative Gene Dosage and Metabolic Requirements Shaped the Time-Dependent Resolution of Yeast Polyploidy." *Molecular Biology and Evolution* 31 (12): 3184–93. https://doi.org/10.1093/molbev/msu250.

Conant, Gavin C., James A. Birchler, and J. Chris Pires. 2014. "Dosage, Duplication, and Diploidization: Clarifying the Interplay of Multiple Models for Duplicate Gene Evolution over Time." *Current Opinion in Plant Biology* 19 (June): 91–98. https://doi.org/10.1016/j.pbi.2014.05.008.

Conant, Gavin C., and Andreas Wagner. 2002. "GenomeHistory: A Software Tool and Its Application to Fully Sequenced Genomes." *Nucleic Acids Research* 30 (15): 3378–86. https://doi.org/10.1093/nar/gkf449.

Conant, Gavin C., and Kenneth H. Wolfe. 2006. “Functional Partitioning of Yeast Co-Expression Networks after Genome Duplication.” *PLoS Biology* 4 (4): e109. https://doi.org/10.1371/journal.pbio.0040109.

———. 2007. “Increased Glycolytic Flux as an Outcome of Whole-Genome Duplication in Yeast.” *Molecular Systems Biology* 3 (July): 129. https://doi.org/10.1038/msb4100170.

———. 2008a. “Probabilistic Cross-Species Inference of Orthologous Genomic Regions Created by Whole-Genome Duplication in Yeast.” *Genetics* 179 (3): 1681–92. https://doi.org/10.1534/genetics.107.074450.

———. 2008b. “Turning a Hobby into a Job: How Duplicated Genes Find New Functions.” *Nature Reviews. Genetics* 9 (12): 938–50. https://doi.org/10.1038/nrg2482.

Costello, Rona, David M. Emms, and Steven Kelly. 2020. “Gene Duplication Accelerates the Pace of Protein Gain and Loss from Plant Organelles.” *Molecular Biology and Evolution* 37 (4): 969–81. https://doi.org/10.1093/molbev/msz275.

De Smet, Riet, Keith L. Adams, Klaas Vandepoele, Marc C. E. Van Montagu, Steven Maere, and Yves Van de Peer. 2013. “Convergent Gene Loss Following Gene and Genome Duplications Creates Single-Copy Families in Flowering Plants.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (8): 2898–2903. https://doi.org/10.1073/pnas.1300127110.

Edger, Patrick P., Hanna M. Heidel-Fischer, Michaël Bekaert, Jadranka Rota, Gernot Glöckner, Adrian E. Platts, David G. Heckel, et al. 2015. “The Butterfly Plant Arms-Race Escalated by Gene and Genome Duplications.” *Proceedings of the*

National Academy of Sciences of the United States of America 112 (27): 8362–66. https://doi.org/10.1073/pnas.1503926112.

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. https://doi.org/10.1105/tpc.17.00010.

Emery, Marianne, M. Madeline S. Willis, Yue Hao, Kerrie Barry, Khouanchy Oakgrove, Yi Peng, Jeremy Schmutz, et al. 2018. "Preferential Retention of Genes from One Parental Genome after Polyploidy Illustrates the Nature and Scope of the Genomic Conflicts Induced by Hybridization." *PLoS Genetics* 14 (3): e1007267. https://doi.org/10.1371/journal.pgen.1007267.

Evangelisti, Annette M., and Gavin C. Conant. 2010. "Nonrandom Survival of Gene Conversions among Yeast Ribosomal Proteins Duplicated through Genome Doubling." *Genome Biology and Evolution* 2 (October): 826–34. https://doi.org/10.1093/gbe/evq067.

Freeling, Michael. 2009. "Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem, Whole-Genome, Segmental, or by Transposition." *Annual Review of Plant Biology* 60: 433–53. https://doi.org/10.1146/annurev.arplant.043008.092122.

Freeling, Michael, Margaret R. Woodhouse, Shabarinath Subramaniam, Gina Turco, Damon Lisch, and James C. Schnable. 2012. "Fractionation Mutagenesis and Similar Consequences of Mechanisms Removing Dispensable or Less-Expressed

DNA in Plants." *Current Opinion in Plant Biology* 15 (2): 131–39. https://doi.org/10.1016/j.pbi.2012.01.015.

Fruchterman, Thomas M. J., and Edward M. Reingold. 1991. "Graph Drawing by Force-Directed Placement." *Software: Practice & Experience* 21 (11): 1129–64. https://doi.org/10.1002/spe.4380211102.

Gong, Lei, Armel Salmon, Mi-Jeong Yoo, Kara K. Grupp, Zining Wang, Andrew H. Paterson, and Jonathan F. Wendel. 2012. "The Cytonuclear Dimension of Allopolyploid Evolution: An Example from Cotton Using Rubisco." *Molecular Biology and Evolution* 29 (10): 3023–36. https://doi.org/10.1093/molbev/mss110.

Goodstein, David M., Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D. Hayes, Joni Fazo, Therese Mitros, et al. 2012. "Phytozome: A Comparative Platform for Green Plant Genomics." *Nucleic Acids Research* 40 (Database issue): D1178–86. https://doi.org/10.1093/nar/gkr944.

Hakes, Luke, John W. Pinney, Simon C. Lovell, Stephen G. Oliver, and David L. Robertson. 2007. "All Duplicates Are Not Equal: The Difference between Small-Scale and Genome Duplication." *Genome Biology* 8 (10): R209. https://doi.org/10.1186/gb-2007-8-10-r209.

Hoek, Milan J. A. van, and Paulien Hogeweg. 2009. "Metabolic Adaptation after Whole Genome Duplication." *Molecular Biology and Evolution* 26 (11): 2441–53. https://doi.org/10.1093/molbev/msp160.

Hollister, Jesse D. 2015. "Polyploidy: Adaptation to the Genomic Environment." *The New Phytologist* 205 (3): 1034–39. https://doi.org/10.1111/nph.12939.

Hu, Yifan. 2005. “Efficient, High-Quality Force-Directed Graph Drawing.” *Mathematica Journal* 10 (1): 37–71. http://asus.myds.me:6543/paper/ktall/37%20-%201984%20-%20Efficient,%20High-Quality%20Force-Directed%20Graph%20Drawing.pdf.

Kacser, H., and J. A. Burns. 1981. “The Molecular Basis of Dominance.” *Genetics* 97 (3-4): 639–66. https://www.ncbi.nlm.nih.gov/pubmed/7297851.

Kirkpatrick, S., C. D. Gelatt Jr, and M. P. Vecchi. 1983. “Optimization by Simulated Annealing.” *Science* 220 (4598): 671–80. https://doi.org/10.1126/science.220.4598.671.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. “Canu: Scalable and Accurate Long-Read Assembly via Adaptive K-Mer Weighting and Repeat Separation.” *Genome Research* 27 (5): 722–36. https://doi.org/10.1101/gr.215087.116.

Korf, Ian. 2004. “Gene Finding in Novel Genomes.” *BMC Bioinformatics* 5 (May): 59. https://doi.org/10.1186/1471-2105-5-59.

Lagercrantz, U. 1998. “Comparative Mapping between Arabidopsis Thaliana and Brassica Nigra Indicates That Brassica Genomes Have Evolved through Extensive Genome Replication ….” *Genetics*. https://academic.oup.com/genetics/article-abstract/150/3/1217/6034672.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Lazzeri, Luca, Franca De Mattei, Fiorenzo Bucelli, and Sandro Palmieri. 1997. "Crambe Oil-a Potential New Hydraulic Oil and Quenchant." *Industrial Lubrication and Tribology*. https://www.emerald.com/insight/content/doi/10.1108/00368799710163893/full/html.

Liang, Zhikai, and James C. Schnable. 2018. "Functional Divergence between Subgenomes and Gene Pairs after Whole Genome Duplications." *Molecular Plant* 11 (3): 388–97. https://doi.org/10.1016/j.molp.2017.12.010.

Liu, Shengyi, Yumei Liu, Xinhua Yang, Chaobo Tong, David Edwards, Isobel A. P. Parkin, Meixia Zhao, et al. 2014. "The Brassica Oleracea Genome Reveals the Asymmetrical Evolution of Polyploid Genomes." *Nature Communications* 5 (May): 3930. https://doi.org/10.1038/ncomms4930.

Lukens, Lewis N., Pablo A. Quijada, Joshua Udall, J. Chris Pires, M. Eric Schranz, and Thomas C. Osborn. 2004. "Genome Redundancy and Plasticity within Ancient and Recent Brassica Crop Species." *Biological Journal of the Linnean Society. Linnean Society of London* 82 (4): 665–74. https://doi.org/10.1111/j.1095-8312.2004.00352.x.

Lyons, Eric, and Michael Freeling. 2008. "How to Usefully Compare Homologous Plant Genes and Chromosomes as DNA Sequences." *The Plant Journal: For Cell and Molecular Biology* 53 (4): 661–73. https://doi.org/10.1111/j.1365-313X.2007.03326.x.

Lyons, Eric, Brent Pedersen, Josh Kane, Maqsudul 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. https://doi.org/10.1104/pp.108.124867.

Lysak, Martin A., Marcus A. Koch, Ales Pecinka, and Ingo Schubert. 2005. "Chromosome Triplication Found across the Tribe Brassiceae." *Genome Research* 15 (4): 516–25. https://doi.org/10.1101/gr.3531105.

Lysák, Martin, and Others. 2009. "Comparative Cytogenetics of Wild Crucifers." https://www.med.muni.cz/en/science-and-research/publikacni-cinnost/839232.

Maere, Steven, Stefanie De Bodt, Jeroen Raes, Tineke Casneuf, Marc Van Montagu, Martin Kuiper, and Yves Van de Peer. 2005. "Modeling Gene and Genome Duplications in Eukaryotes." *Proceedings of the National Academy of Sciences of the United States of America* 102 (15): 5454–59. https://doi.org/10.1073/pnas.0501102102.

Makino, Takashi, and Aoife McLysaght. 2010. "Ohnologs in the Human Genome Are Dosage Balanced and Frequently Associated with Disease." *Proceedings of the National Academy of Sciences of the United States of America* 107 (20): 9270–74. https://doi.org/10.1073/pnas.0914697107.

———. 2012. "Positionally Biased Gene Loss after Whole Genome Duplication: Evidence from Human, Yeast, and Plant." *Genome Research* 22 (12): 2427–35. https://doi.org/10.1101/gr.131953.111.

McClintock, B. 1984. "The Significance of Responses of the Genome to Challenge." *Science* 226 (4676): 792–801. https://doi.org/10.1126/science.15739260.

Merico, Annamaria, Pavol Sulo, Jure Piskur, and Concetta Compagno. 2007. "Fermentative Lifestyle in Yeasts Belonging to the Saccharomyces Complex." *The FEBS Journal* 274 (4): 976–89. https://doi.org/10.1111/j.1742-4658.2007.05645.x.

Meyer, David, Achim Zeileis, and Kurt Hornik. 2006. "The Strucplot Framework: Visualizing Multi-Way Contingency Tables with Vcd." *Journal of Statistical Software, Articles* 17 (3): 1–48. https://doi.org/10.18637/jss.v017.i03.

Mi, Huaiyu, Anushya Muruganujan, Dustin Ebert, Xiaosong Huang, and Paul D. Thomas. 2019. "PANTHER Version 14: More Genomes, a New PANTHER GO-Slim and Improvements in Enrichment Analysis Tools." *Nucleic Acids Research* 47 (D1): D419–26. https://doi.org/10.1093/nar/gky1038.

Notredame, C., D. G. Higgins, and J. Heringa. 2000. "T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment." *Journal of Molecular Biology* 302 (1): 205–17. https://doi.org/10.1006/jmbi.2000.4042.

Ohno, Susumu. 2013. *Evolution by Gene Duplication*. Springer Science & Business Media. https://play.google.com/store/books/details?id=5SjqCAAAQBAJ.

Oliveira Dal'Molin, Cristiana Gomes de, Lake-Ee Quek, Robin William Palfreyman, Stevens Michael Brumbley, and Lars Keld Nielsen. 2010. "AraGEM, a Genome-Scale Reconstruction of the Primary Metabolic Network in Arabidopsis." *Plant Physiology* 152 (2): 579–89. https://doi.org/10.1104/pp.109.148817.

One Thousand Plant Transcriptomes Initiative. 2019. "One Thousand Plant Transcriptomes and the Phylogenomics of Green Plants." *Nature* 574 (7780): 679–85. https://doi.org/10.1038/s41586-019-1693-2.

Parkin, Isobel A. P., Sigrun M. Gulden, Andrew G. Sharpe, Lewis Lukens, Martin Trick, Thomas C. Osborn, and Derek J. Lydiate. 2005. "Segmental Structure of the Brassica Napus Genome Based on Comparative Analysis with Arabidopsis Thaliana." *Genetics* 171 (2): 765–81. https://doi.org/10.1534/genetics.105.042093.

Parkin, Isobel A. P., Chushin Koh, Haibao Tang, Stephen J. Robinson, Sateesh Kagale, Wayne E. Clarke, Chris D. Town, et al. 2014. "Transcriptome and Methylome Profiling Reveals Relics of Genome Dominance in the Mesopolyploid Brassica Oleracea." *Genome Biology* 15 (6): R77. https://doi.org/10.1186/gb-2014-15-6-r77.

Paterson, Andrew H. 2005. "Polyploidy, Evolutionary Opportunity, and Crop Adaptation." *Genetica* 123 (1-2): 191–96. https://doi.org/10.1007/s10709-003-2742-0.

Pavlidis, Pavlos, Daniel Živkovic, Alexandros Stamatakis, and Nikolaos Alachiotis. 2013. "SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes." *Molecular Biology and Evolution* 30 (9): 2224–34. https://doi.org/10.1093/molbev/mst112.

Peer, Yves Van de, Yves Van de Peer, Eshchar Mizrachi, and Kathleen Marchal. 2017. "The Evolutionary Significance of Polyploidy." *Nature Reviews Genetics*. https://doi.org/10.1038/nrg.2017.26.

Pérez-Bercoff, Åsa, Aoife McLysaght, and Gavin C. Conant. 2011. "Patterns of Indirect Protein Interactions Suggest a Spatial Organization to Metabolism." *Molecular bioSystems* 7 (11): 3056–64. https://doi.org/10.1039/c1mb05168g.

Pertea, Mihaela, Geo M. Pertea, Corina M. Antonescu, Tsung-Cheng Chang, Joshua T. Mendell, and Steven L. Salzberg. 2015. "StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads." *Nature Biotechnology* 33 (3): 290–95. https://doi.org/10.1038/nbt.3122.

Qi, Xinshuai, Hong An, Tara E. Hall, Chenlu Di, Paul D. Blischak, Michael T. W. McKibben, Yue Hao, Gavin C. Conant, J. Chris Pires, and Michael S. Barker. 2021. "Genes Derived from Ancient Polyploidy Have Higher Genetic Diversity and Are Associated with Domestication in Brassica Rapa." *The New Phytologist* 230 (1): 372–86. https://doi.org/10.1111/nph.17194.

Qi, Xinshuai, Hong An, Aaron P. Ragsdale, Tara E. Hall, Ryan N. Gutenkunst, J. Chris Pires, and Michael S. Barker. 2017. "Genomic Inferences of Domestication Events Are Corroborated by Written Records in Brassica Rapa." *Molecular Ecology* 26 (13): 3373–88. https://doi.org/10.1111/mec.14131.

Qiu, Yichun, Yii Van Tay, Yuan Ruan, and Keith L. Adams. 2020. "Divergence of Duplicated Genes by Repeated Partitioning of Splice Forms and Subcellular Localization." *The New Phytologist* 225 (2): 1011–22. https://doi.org/10.1111/nph.16148.

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. https://doi.org/10.1093/molbev/msv001.

Scannell, Devin R., A. Carolin Frank, Gavin C. Conant, Kevin P. Byrne, Megan Woolfit, and Kenneth H. Wolfe. 2007. "Independent Sorting-out of Thousands of

Duplicated Gene Pairs in Two Yeast Species Descended from a Whole-Genome Duplication." *Proceedings of the National Academy of Sciences of the United States of America* 104 (20): 8397–8402. https://doi.org/10.1073/pnas.0608218104.

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. https://doi.org/10.1073/pnas.1101368108.

Schoonmaker, Ashley, Yue Hao, David Mck Bird, and Gavin C. Conant. 2020. "A Single, Shared Triploidy in Three Species of Parasitic Nematodes." *G3* 10 (1): 225–33. https://doi.org/10.1534/g3.119.400650.

Schranz, M. Eric, Martin A. Lysak, and Thomas Mitchell-Olds. 2006. "The ABC's of Comparative Genomics in the Brassicaceae: Building Blocks of Crucifer Genomes." *Trends in Plant Science* 11 (11): 535–42. https://doi.org/10.1016/j.tplants.2006.09.002.

Scienski, Kathy, Justin C. Fay, and Gavin C. Conant. 2015. "Patterns of Gene Conversion in Duplicated Yeast Histones Suggest Strong Selection on a Coadapted Macromolecular Complex." *Genome Biology and Evolution* 7 (12): 3249–58. https://doi.org/10.1093/gbe/evv216.

Seoighe, C., and K. H. Wolfe. 1998. "Extent of Genomic Rearrangement after Genome Duplication in Yeast." *Proceedings of the National Academy of Sciences of the United States of America* 95 (8): 4447–52. https://doi.org/10.1073/pnas.95.8.4447.

Shanmugasundaram, S. 2012. “Wild Crop Relatives: Genomic and Breeding Resources. Oilseeds. Edited by C. Kole. Heidelberg, Dordrecht, London: Springer (2011), Pp. 295,£ 135.00. ISBN 978-3-642-14870-5.” *Experimental Agriculture* 48 (1): 156–156. https://www.cambridge.org/core/journals/experimental-agriculture/article/wild-crop-relatives-genomic-and-breeding-resources-oilseeds-edited-by-kole-c-heidelberg-dordrecht-london-springer-2011-pp-295-13500-isbn-9783642148705/B7F496224CAFEBD4C4C84242CB4B3A2B.

Sharbrough, Joel, Justin L. Conover, Jennifer A. Tate, Jonathan F. Wendel, and Daniel B. Sloan. 2017. “Cytonuclear Responses to Genome Doubling.” *American Journal of Botany* 104 (9): 1277–80. https://doi.org/10.3732/ajb.1700293.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. “BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.” *Bioinformatics* 31 (19): 3210–12. https://doi.org/10.1093/bioinformatics/btv351.

Smit, A. F. A., R. Hubley, and P. Green. 2015. “RepeatMasker Open-4.0. 2013--2015.”

Smukowski Heil, Caiti S., Christopher G. DeSevo, Dave A. Pai, Cheryl M. Tucker, Margaret L. Hoang, and Maitreya J. Dunham. 2017. “Loss of Heterozygosity Drives Adaptation in Hybrid Yeast.” *Molecular Biology and Evolution* 34 (7): 1596–1612. https://doi.org/10.1093/molbev/msx098.

Soltis, P. S., and D. E. Soltis. 2012. *Polyploidy and Genome Evolution*. Edited by Pamela S. Soltis and Douglas E. Soltis. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-31442-1.

Stanke, Mario, and Stephan Waack. 2003. “Gene Prediction with a Hidden Markov Model and a New Intron Submodel.” *Bioinformatics* 19 Suppl 2 (October): ii215–25. https://doi.org/10.1093/bioinformatics/btg1080.

Stark, Chris, Bobby-Joe Breitkreutz, Andrew Chatr-Aryamontri, Lorrie Boucher, Rose Oughtred, Michael S. Livstone, Julie Nixon, et al. 2011. “The BioGRID Interaction Database: 2011 Update.” *Nucleic Acids Research* 39 (Database issue): D698–704. https://doi.org/10.1093/nar/gkq1116.

Sukeena, Joshua M., Carlos A. Galicia, Jacob D. Wilson, Tim McGinn, Janette W. Boughman, Barrie D. Robison, John H. Postlethwait, Ingo Braasch, Deborah L. Stenkamp, and Peter G. Fuerst. 2016. “Characterization and Evolution of the Spotted Gar Retina.” *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution* 326 (7): 403–21. https://doi.org/10.1002/jez.b.22710.

Tang, Haibao, Margaret R. Woodhouse, Feng Cheng, James C. Schnable, Brent S. Pedersen, Gavin Conant, Xiaowu Wang, Michael Freeling, and J. Chris Pires. 2012. “Altered Patterns of Fractionation and Exon Deletions in Brassica Rapa Support a Two-Step Model of Paleohexaploidy.” *Genetics* 190 (4): 1563–74. https://doi.org/10.1534/genetics.111.137349.

Team, R. Core, and Others. 2013. “R: A Language and Environment for Statistical Computing.” http://r.meteo.uni.wroc.pl/web/packages/dplR/vignettes/intro-dplR.pdf.

Van de Peer, Yves, Steven Maere, and Axel Meyer. 2009. “The Evolutionary Significance of Ancient Genome Duplications.” *Nature Reviews. Genetics* 10 (10): 725–32. https://doi.org/10.1038/nrg2600.

Walker, Bruce J., Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, et al. 2014. "Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement." *PloS One* 9 (11): e112963. https://doi.org/10.1371/journal.pone.0112963.

Wang, Xiaowu, Hanzhong Wang, Jun Wang, Rifei Sun, Jian Wu, Shengyi Liu, Yinqi Bai, et al. 2011. "The Genome of the Mesopolyploid Crop Species Brassica Rapa." *Nature Genetics* 43 (10): 1035–39. https://doi.org/10.1038/ng.919.

Warwick, Suzanne I., and Richard K. Gugel. 2003. "Genetic Variation in the Crambe Abyssinica - C. Hispanica - C. Glabrata Complex." *Genetic Resources and Crop Evolution* 50 (3): 291–305. https://doi.org/10.1007/s10722-004-2910-9.

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. https://doi.org/10.1016/j.gde.2018.01.004.

Wick, Ryan R., Mark B. Schultz, Justin Zobel, and Kathryn E. Holt. 2015. "Bandage: Interactive Visualization of de Novo Genome Assemblies." *Bioinformatics* 31 (20): 3350–52. https://doi.org/10.1093/bioinformatics/btv383.

Wolfe, K. H., and D. C. Shields. 1997. "Molecular Evidence for an Ancient Duplication of the Entire Yeast Genome." *Nature* 387 (6634): 708–13. https://doi.org/10.1038/42711.

Woodhouse, Margaret R., Feng Cheng, J. Chris Pires, Damon Lisch, Michael Freeling, and Xiaowu Wang. 2014. "Origin, Inheritance, and Gene Regulatory

Consequences of Genome Dominance in Polyploids." *Proceedings of the National Academy of Sciences of the United States of America* 111 (14): 5283–88. https://doi.org/10.1073/pnas.1402475111.

Xie, Ting, Fu-Gui Zhang, Hong-Yu Zhang, Xiao-Tao Wang, Ji-Hong Hu, and Xiao-Ming Wu. 2019. "Biased Gene Retention during Diploidization in Brassica Linked to Three-Dimensional Genome Organization." *Nature Plants* 5 (8): 822–32. https://doi.org/10.1038/s41477-019-0479-8.

Yang, Ziheng. 2007. "PAML 4: Phylogenetic Analysis by Maximum Likelihood." *Molecular Biology and Evolution* 24 (8): 1586–91. https://doi.org/10.1093/molbev/msm088.

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. https://doi.org/10.1146/annurev-genet-120213-092159.

Zeileis, Achim, David Meyer, and Kurt Hornik. 2007. "Residual-Based Shadings for Visualizing (Conditional) Independence." *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* 16 (3): 507–25. https://doi.org/10.1198/106186007X237856.

Zhao, Tao, Rens Holmer, Suzanne de Bruijn, Gerco C. Angenent, Harrold A. van den Burg, and M. Eric Schranz. 2017. "Phylogenomic Synteny Network Analysis of MADS-Box Transcription Factor Genes Reveals Lineage-Specific Transpositions, Ancient Tandem Duplications, and Deep Positional Conservation." *The Plant Cell* 29 (6): 1278–92. https://doi.org/10.1105/tpc.17.00312.

Zheng, Chunfang, Eric Chen, Victor A. Albert, Eric Lyons, and David Sankoff. 2013. "Ancient Eudicot Hexaploidy Meets Ancestral Eurosid Gene Order." *BMC Genomics* 14 Suppl 7 (November): S3. https://doi.org/10.1186/1471-2164-14-S7-S3.

Zheng, Chunfang, Krister Swenson, Eric Lyons, and David Sankoff. 2011. "OMG! Orthologs in Multiple Genomes – Competing Graph-Theoretical Formulations." In *Lecture Notes in Computer Science*, 364–75. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-23038-7_30.

VITA

R. Shawn Abrahams was born in Plantation, Florida. They attended South Plantation High to participate in the magnet program focused on Environmental Science and Everglades Restoration for high school. There, Shawn nurtured their interest and aptitude for scientific research and the field of Botany. Shawn would attend the University of Florida to pursue a B.S. in Botany and participate in undergraduate research, student leadership, and LGBTQ activism. After graduating, Shawn completed a one-year internship with the Everglades foundation studying the impacts of sea-level rise on the vulnerable marsh ecosystems of their South Florida home. At a conference hosted by the Botanical Society of America in New Orleans, Shawn met their eventual Ph.D. advisor J. Chris Pires. Almost two years later, they left their coastal ways behind to study polyploidy and evolution as a member of the Pires lab. Following the completion of their degree, Shawn intended to be transient to New Haven, CT, and start as a postdoctoral fellow in the labs of Erika Edwards and Jen Wisecaver as part of their NSF Postdoctoral Fellowship.