BRASSICALES TO BRASSICA:
INTEGRATING PHYLOGENOMICS AND POPULATION 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
MAKENZIE ELIZABETH MABRY
Dr. J. Chris Pires, Dissertation Supervisor

JULY 2020

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

BRASSICALES TO BRASSICA:
INTEGRATING PHYLOGENOMICS AND POPULATION GENOMICS

Presented by Makenzie Elizabeth Mabry, 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. Elizabeth King

Dr. Gavin C. Conant

DEDICATION

“It's really about having been able to move somebody or inspire somebody for a brief moment to want to be a better version of themselves”

- *Kobe Bryant*

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, J. Chris Pires, who from the beginning provided me with numerous opportunities to gain experience in mentorship, research, and teaching. Thank you for sharing the “Dogs of the Plant World” with me. I have very much enjoyed working together to figure out this amazingly diverse group of plants. I also want to thank Chris for instilling in me the importance of creating a team and teaching me the skills needed to lead. I additionally thank my committee members, Libby King, Jim Birchler, and Gavin Conant for their support and availability in all things from research to surviving graduate school.

Further, the work presented here would not been possible without the help and support from the undergraduates that I have been fortunate enough to work with: Julia Brose, Sarah Gebken, Brandin Grindstaff, Evan Gallagher, and Liz Countee. You are all such amazing humans and I look forward to seeing all that you accomplish. Thank you for keeping me sane, keeping the plants alive, and motivating me to keep going.

I want to especially thank Marianne Slaten, Sarah Unruh, R. Shawn Abrahams, Katy Guthrie, Troy Rowan, Harly Durbin, Jo Morrow, Joe Gunn, Paul Petrowski, and Sarah Turner Hissong for their support and friendship. I would not have made it through living in Missouri without you all.

I would like to thank the University of Missouri faculty, staff, and facilities, especially Barb Sonderman for her support in just about everything from teaching to life and constantly being my cheerleader; Bob Schnabel, Bill Spollen, Christopher Bottoms, and Jake Gotburg for their help in computational analyses; Alan Marshall and Nick Valentine for their IT assistance, especially when teaching; Lauren Sullivan, Ruthie

Angelovici, Kathy Newton and Kate Gill for their overall support; Dean Bergstrom for keeping the LSC running; Loquita Ellis and Victoria Bence for all the many moments of laughter; Rodger Meissen and Melody Kroll for their work in advertising and promoting my research; David Mendoza for his support and encouragement in developing of the BIPS program; Nathan Bivens, Ming-Yi Zhou, and Karen Bromert for their assistance with sequencing; and Vicki Bryan and Rebecca Ballew for keeping the IPG and BioSci programs organized and running.

Finally, I would like to thank my amazing family and friends for their support both near and far. I especially want to thank my partner, Paul Blischak, for his support and patience in answering my constant questions. I truly could not have accomplished so much without you.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	II
LIST OF FIGURES	VI
LIST OF TABLES	VIII
ABSTRACT	IX
CHAPTER 1 : INTEGRATING NETWORKS, PHYLOGENOMICS, AND POPULATION GENOMICS FOR THE STUDY OF POLYPLOIDY	1
ABSTRACT	2
INTRODUCTION	3
PREVALENCE OF POLYPLOIDY	7
COMPLICATIONS FOR DETERMINING IF THERE ARE THERE RULES FOR POLYPLOIDY	11
MODELS FOR EVOLUTIONARY INFERENCES IN POLYPLOIDS	20
SUMMARY AND FUTURE DIRECTIONS	31
ACKNOWLEDGMENTS	34
LITERATURE CITED	35
CHAPTER 2 : PHYLOGENY AND MULTIPLE INDEPENDENT WHOLE-GENOME DUPLICATION EVENTS IN THE BRASSICALES.....	55
ABSTRACT	56
INTRODUCTION	57
MATERIALS AND METHODS	61
RESULTS	66
DISCUSSION.....	80
CONCLUSIONS	88
ACKNOWLEDGMENTS	89
AUTHOR CONTRIBUTIONS	89
DATA AVAILABILITY	90
LITERATURE CITED	91
CHAPTER 3 : THE EVOLUTIONARY HISTORY OF WILD AND DOMESTICATED, AND FERAL <i>BRASSICA OLERACEA</i> (BRASSICACEAE).....	123
ABSTRACT	124
INTRODUCTION	125
RESULTS	128
DISCUSSION.....	141
MATERIALS AND METHODS	145
AUTHOR CONTRIBUTIONS	151
ACKNOWLEDGMENTS	151
REFERENCES	152
CHAPTER 4 : IONOMIC VARIATION IN DIVERSE <i>BRASSICA OLERACEA</i> CULTIVAR AND WILD RELATIVE POPULATIONS FROM GERMPLASM COLLECTIONS	185
ABSTRACT	186
INTRODUCTION	187
MATERIALS AND METHODS	188

GERMPLASM AND DATA AVAILABILITY	193
RESULTS	193
DISCUSSION.....	204
CONCLUSIONS	208
ACKNOWLEDGMENTS.....	208
AUTHOR CONTRIBUTIONS	209
REFERENCES	210
CHAPTER 5 : A FERAL FUTURE	251
ABSTRACT	252
EMERGING INTEREST IN FERAL POPULATIONS.....	253
THE DIFFICULTIES IN DEFINING FERAL	253
FERALIZATION IS NOT MERELY THE REVERSE OF DOMESTICATION.....	256
FERAL POPULATIONS MAY MODEL NATURAL SELECTION BETTER THAN DOMESTICATION	258
FERAL POPULATIONS CAN HELP US UNDERSTAND LOCAL ADAPTATION AND IMPROVE AGRICULTURAL SPECIES.....	260
CONCLUDING REMARKS.....	262
ACKNOWLEDGMENTS.....	263
REFERENCES	264
VITA	271

LIST OF FIGURES

FIGURE 1.1 HYPOTHETICAL SCENARIOS DEMONSTRATING DIFFERENT APPROACHES FOR ANALYZING THE ALLOPOLYPLOID <i>BRASSICA NAPUS</i>	6
FIGURE 1.2 RESOLVING PATTERNS OF BIASED FRACTIONATION WITH POINT.....	26
FIGURE 2.1 ASTRAL-III SPECIES PHYLOGENY AND WHOLE-GENOME DUPLICATION EVENTS OF THE BRASSICALES.	70
FIGURE 2.2 COALESCENT-BASED SPECIES PHYLOGENY AND WHOLE-GENOME DUPLICATION EVENTS OF THE BRASSICACEAE.....	72
FIGURE 2.3 COALESCENT-BASED SPECIES PHYLOGENY AND WHOLE-GENOME DUPLICATION EVENTS OF THE CLEOMACEAE.	74
FIGURE 2.4 COMPARISON OF ORTHOLOG DIVERGENCES AND Ks PEAKS OF THE CLEOMACEAE TO TEST HYPOTHESES OF PLACEMENT FOR TH- α	77
FIGURE 2.5 COALESCENT-BASED SPECIES PHYLOGENIES AND WHOLE-GENOME DUPLICATION EVENTS OF THE (A) CAPPARACEAE AND (B) RESEDACEAE + OUTGROUPS.....	79
FIGURE 2.6 COMPARISON OF (A) MAXIMUM LIKELIHOOD WHOLE-CHLOROPLAST PHYLOGENY TO (B) COALESCENT-BASED SPECIES PHYLOGENY OF THE BRASSICALES.....	82
FIGURE 3.1 <i>BRASSICA OLERACEA</i> WILD RELATIVE AND CULTIVAR DIVERSITY.	126
FIGURE 3.2 SPECIES DISTRIBUTION MODELS.	130
FIGURE 3.3 SAMPLE DEMOGRAPHICS.	132
FIGURE 4.1 RAW ELEMENT ABUNDANCE DIVERSITY SEPARATED MY CULTIVARS AND WILD GERMPLASM...	195
FIGURE 4.2 CALCULATED HERITABILITY OF ELEMENTS.	196
FIGURE 4.3 PCA OF ELEMENTAL COMPOSITION FOR A) CULTIVARS VERSUS WILD GERMPLASM, B) AMONG CULTIVARS, AND C) BETWEEN EDIBLE LEAVES AND STEM (TUBER) CULTIVARS.....	197
FIGURE 4.4 GWAS RESULTS FOR EACH MODEL (SHOWN FOR AL) WITH SNPs MAPPING TO KNOWN IONOMIC GENES INDICATED BY GREEN TRIANGLE.	201
FIGURE 4.5 GWAS RESULTS OF ALL MODELS FOR PC1. LABELLED SNPs ARE TOP HITS FOR EACH CHROMOSOME.	202
SUPPLEMENTAL FIGURE 2.1 CURRENT UNDERSTANDING OF THE RELATIONSHIPS BETWEEN THE 17 FAMILIES OF THE BRASSICALES (APG IV).	104
SUPPLEMENTAL FIGURE 2.2 BUSCO ANALYSIS OF <i>DE NOVO</i> TRANSCRIPTOMES.	116
SUPPLEMENTAL FIGURE 2.3 MAXIMUM LIKELIHOOD PHYLOGENY OF THE BRASSICALES USING TWO CHLOROPLAST GENES, <i>MATK</i> AND <i>NDHF</i>	117
SUPPLEMENTAL FIGURE 2.4 MAXIMUM LIKELIHOOD WHOLE-CHLOROPLAST PHYLOGENY OF THE BRASSICALES.	118
SUPPLEMENTAL FIGURE 2.5 Ks PLOTS OF THE BRASSICACEAE USING BOTH FASTKS (McKAIN ET AL., 2016) AND DUPPIPE (BARKER ET AL., 2010).....	120
SUPPLEMENTAL FIGURE 2.6 ORTHOLOG DIVERGENCES AND Ks PEAKS OF THE CLEOMACEAE..	121
SUPPLEMENTAL FIGURE 2.7 ORTHOLOG DIVERGENCES AND Ks PEAKS OF THE (A) CAPPARACEAE AND (B) RESEDACEAE + OUTGROUPS.	122
SUPPLEMENTAL FIGURE 3.1 PCA OF A) PC2 vs PC3 AND B) PC3 vs PC4 OF WILD AND CULTIVAR SAMPLES.	166
SUPPLEMENTAL FIGURE 3.2 TREEMIX ANALYSIS OF WILD SAMPLES AND CULTIVARS	167
SUPPLEMENTAL FIGURE 3.3 INTROGRESSION TESTS PERFORMED.	168
SUPPLEMENTAL FIGURE 3.4 EXPRESSION PCA FOR A) PC1 vs PC2, B) PC2 vs PC3, C) PC3 vs PC4.....	169
SUPPLEMENTAL FIGURE 3.5 <i>BRASSICA OLERACEA</i> CULTIVARS AND WILD RELATIVE DENDROGRAM BASED ON EXPRESSION PROFILES.	170
SUPPLEMENTAL FIGURE 3.6 MODULE ASSIGNMENT AND DENDROGRAM FOR THE TOP 1000 VARIABLE GENES IN EXPRESSION.....	171

SUPPLEMENTAL FIGURE 4.1 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR AL.....	229
SUPPLEMENTAL FIGURE 4.2 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR B.....	230
SUPPLEMENTAL FIGURE 4.3 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR CA.....	231
SUPPLEMENTAL FIGURE 4.4 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR CD.....	232
SUPPLEMENTAL FIGURE 4.5 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR CO.....	233
SUPPLEMENTAL FIGURE 4.6 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR CU.....	234
SUPPLEMENTAL FIGURE 4.7 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR FE.....	235
SUPPLEMENTAL FIGURE 4.8 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR K.....	236
SUPPLEMENTAL FIGURE 4.9 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR MG.....	237
SUPPLEMENTAL FIGURE 4.10 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR MN.....	238
SUPPLEMENTAL FIGURE 4.11 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR MO.....	239
SUPPLEMENTAL FIGURE 4.12 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR NA.....	240
SUPPLEMENTAL FIGURE 4.13 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR NI.....	241
SUPPLEMENTAL FIGURE 4.14 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR P.....	242
SUPPLEMENTAL FIGURE 4.15 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR RB.....	243
SUPPLEMENTAL FIGURE 4.16 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR SR.....	244
SUPPLEMENTAL FIGURE 4.17 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR ZN.....	245
SUPPLEMENTAL FIGURE 4.18 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR PC1.....	246
SUPPLEMENTAL FIGURE 4.19 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR PC2.....	247
SUPPLEMENTAL FIGURE 4.20 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR PC3.....	248
SUPPLEMENTAL FIGURE 4.21 TUKEY’S TEST FOR SIGNIFICANT DIFFERENCES AMONG CULTIVARS OF <i>B.</i> <i>OLERACEA</i> FOR PC4.....	249
SUPPLEMENTAL FIGURE 4.22 OVERLAP OF SNPs AMONG MODELS FOR ELEMENTS WHICH HAD SIGNIFICANT SNPs IN MORE THAN ONE MODEL.....	250

LIST OF TABLES

TABLE 4.1 ANOVA P-VALUES WITH MORPHOTYPE AS PREDICTOR VARIABLE.	198
TABLE 4.2 ANOVA P-VALUES WITH WILD OR LEAFY AS PREDICTOR VARIABLES.	198
TABLE 4.3 OVERLAP OF IDENTIFIED SNPs AND KNOWN IONOMIC GENES.	200
TABLE 4.4 SIGNIFICANT SNPs IDENTIFIED IN GWAS OF PCs.....	203
SUPPLEMENTAL TABLE 2.1 TAXON SAMPLING WITH AUTHORSHIP (DETERMINED USING HTTP://WWW.IPNI.ORG/INDEX.HTML), SEED ACCESSIONS AND THE COLLECTIONS THEY ARE FROM, ADDITIONAL ANALYSIS INFORMATION, AND SRA ACCESSIONS. SSC = SMALL SINGLE COPY, LSC= LARGE SINGLE COPY, IR = INVERTED REPEAT.....	105
SUPPLEMENTAL TABLE 2.2 RNA AND DNA EXTRACTION METHOD, LIBRARY PREPARATION METHOD, SEQUENCING METHOD, READ SIZE, AND RAW READ NUMBERS.....	109
SUPPLEMENTAL TABLE 2.3 ORTHOGROUPS RETAINED FOR EACH ANALYSIS.	112
SUPPLEMENTAL TABLE 2.4 BIC SCORES FOR 1-4 COMPONENTS FOR BOTH FASTKS (MCKAIN ET AL., 2016) AND DUPPIPE (BARKER ET AL., 2010) KS PLOTS.	113
SUPPLEMENTAL TABLE 3.1 SAMPLE INFORMATION WITH SPECIES, VARIETY, CULTIVAR, ACCESSION, SAMPLEID, AND PHENOTYPES USED IN ANALYSES.....	172
SUPPLEMENTAL TABLE 3.2 F4 RESULTS, STANDARD ERROR, AND Z-SCORES FOR SAMPLES TESTED.	177
SUPPLEMENTAL TABLE 3.3 MODULES IDENTIFIES IN WGCNA, THE NUMBER OF GENE IN EACH MODULE, NUMBER OF SYntenic HITS AND BLAST HITS TO <i>ARABIDOPSIS</i> WITH GO BIOLOGICAL PROCESS INDICATE FOR 5 GENE WITH THE LARGEST FOLD ENRICHMENT.....	181
SUPPLEMENTAL TABLE 3.4 ARCHAEOLOGICAL <i>BRASSICA</i> REPORTS FROM EUROPE AND THE NEAR EAST. .	182
SUPPLEMENTAL TABLE 3.5 LITERARY AND ARTISTIC SOURCES COVERING THE CLASSICAL GREEK, ROMAN, AND MEDIEVAL AND POST-MEDIEVAL SOURCES.	183
SUPPLEMENTAL TABLE 4.1 KNOWN IONOMICS GENES IDENTIFIED IN <i>ARABIDOPSIS THALIANA</i> (ATH) WITH <i>BRASSICA OLERACEA</i> (BOL) DETERMINED ORTHOLOGS.....	225

ABSTRACT

The Brassicales are an economically important order of flowering plants. Many crop species such as kale, broccoli, cabbage, cauliflower, canola oil, capers, and papaya, as well as the model plant organism, *Arabidopsis thaliana*, all belong to this diverse order. The focus of my dissertation has been to use phylogenomics and population genomics to understand the evolution of the Brassicales, as well as the important crop species *Brassica oleracea*.

Within the Brassicales, the family Brassicaceae is notable for many whole-genome duplication (WGD) events, including one at the base of the family. The sister family to the Brassicaceae, the Cleomaceae, is also known to have a WGD event; however, it has never been placed within the family. Previous studies assessing the relationships between and within these families have either used a few genes or a few taxa. Phylo-transcriptomics, a quickly evolving field that uses RNA-seq data to make phylogenomic inferences from many more nuclear genes than traditional PCR, is a great method to no longer have to choose between too few genes or taxa. For my second chapter, I used phylo-transcriptomics to estimate relationships in the Brassicales and used the resulting phylogeny as a framework for inferring WGD events. These analyses recovered the known At- β and At- α events, as well as multiple independent WGD events in the Cleomaceae, Capparaceae, and Resedaceae.

The horticultural crop *Brassica oleracea* is a notable member of the Brassicales, playing an important role in global food systems. *Brassica oleracea* is unique in that it has been domesticated into several morphotypes (cultivars), including broccoli, Brussels sprouts, cabbage, cauliflower, kale, kohlrabi, and several lesser well known morphotypes,

such as walking stick kale and marrow cabbage. These crops are widely used as leaf and root vegetables, as well as for animal feed. There are several hypotheses on the origin of these crops. However, after resolving the relationships between wild relatives and domesticated types, I have determined that cultivation likely originated in the Eastern Mediterranean region, with additional domestications occurring around the world. I also identified feral populations that had previously been mistaken as progenitor populations. Using this system of diversity within domesticated *B. oleracea*, I examined patterns of relationships among morphotypes, including signals of hybridization and introgression.

Brassica oleracea vegetable crops are frequently regarded as superfoods (e.g., kale) and are an important diet source for key minerals and elements. Ionomics, the measurement of the total elemental composition of an organism, allows for high throughput characterization of the levels of these elements at the same time, including those important to diet such as iron and zinc. Using association mapping techniques, I have identified genes underlying quantitative phenotypic traits of economic importance such as iron, as well as total element composition. This knowledge can then allow for further decisions of the best accessions to target for human health, as well as provide an excellent resource for future studies wanting to enhance specific elements in different crops.

To conclude, I discuss the unique opportunity that feral populations provide for the study of natural selection and how this information can be leveraged for improvement of crops in breeding them to be more locally adapted.

CHAPTER 1 : INTEGRATING NETWORKS, PHYLOGENOMICS, AND POPULATION GENOMICS FOR THE STUDY OF POLYPLOIDY

Paul D. Blischak*¹, Makenzie E. Mabry*², Gavin C. Conant^{3,4}, and J. Chris Pires⁵

1 Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, Ohio 43210, USA

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

3 Division of Animal Sciences, University of Missouri, Columbia, Missouri 65211, USA

4 Current affiliation: Bioinformatics Research Center, Program in Genetics and Department of Biological Sciences, North Carolina State University, Raleigh, North Carolina 27695, USA;

5 Division of Biological Sciences and Bond Life Sciences Center, University of Missouri, Columbia, Missouri 65211-7310, USA

Please cite the published version in Annual Reviews of Ecology, Evolution, and Systematics:

Blischak PD*, ME Mabry*, GC Conant, & JC Pires 2018 Integrating networks, phylogenomics, and population genomics for the study of polyploidy. *Annu. Rev. Ecol. Evol. Syst.* 2018. 49:253–78. <https://doi.org/10.1146/annurev-ecolsys-121415-032302>

*These authors contributed equally as first authors.

ABSTRACT

Duplication events are regarded as sources of evolutionary novelty, but our understanding of general trends for the long-term trajectory of additional genomic material is still lacking. Organisms with a history of whole genome duplication (WGD) offer a unique opportunity to study potential trends in the context of gene retention/loss, gene and network dosage, and changes in gene expression. In this review, we discuss the prevalence of polyploidy across the tree of life, followed by an overview of studies investigating genome evolution and gene expression. We then provide an overview of methods in network biology, phylogenomics, and population genomics that will be critical for advancing our understanding of evolution post-WGD, highlighting the need for models that can accommodate polyploids. Finally, we close with a brief note on the importance of random processes in the evolution of polyploids with respect to neutral versus selective forces, ancestral polymorphisms, and the formation of autopolyploids versus allopolyploids

INTRODUCTION

Significance of Gene Duplications in the History of Life

Gene duplications are key drivers of evolutionary innovation, and the presence of duplicated material in the genome has important phenotypic consequences (Ohno 1970; Kondrashov & Kondrashov 2006; Hughes et al. 2014; Soltis & Soltis 2016; Van de Peer et al. 2017). For instance, the propensity of a gene to persist after duplication over evolutionary time is negatively associated with its tendency for being deleterious when present in increased copy number (Rice & McLysaght 2017a). Not surprisingly, extra gene copies have also been associated with diseases as diverse as Alzheimer's, Parkinson's, autism, and psoriasis (Almal & Padh 2012; Rice & McLysaght 2017b). Despite this importance of duplication and copy number variations (CNV) across biological timescales, important facets of their function and evolution are still unclear, with numerous gaps in our knowledge having been documented across disciplines (Barker et al. 2016; Soltis et al. 2016). This is especially the case for systems that have experienced a complete doubling of their chromosomes, a phenomenon known as polyploidy.

Polyploidy: Prevalence and Relevance of Whole Genome Duplications Across the Tree of Life

Beyond merely creating large sets of duplicated genes for study, polyploidy is itself of extraordinary interest. Polyploidization (also known as whole genome duplication: WGD) refers to the doubling (or more) of the entire genome through one of a variety of mechanisms. Its importance for evolutionary innovation was emphasized by Ohno (1970), and the advent of genome sequencing has unveiled a history of ancient WGD across the tree of life, spanning bacteria, archaea, and eukaryotes (Van de Peer et al. 2017).

Polyploidy has now been described in lineages including nematodes, arthropods, chordates, fungi, oomycetes, and land plants (Van de Peer et al. 2017; Li et al. 2018), and the pace of discovering polyploidy is arguably accelerating (One Thousand Plant Transcriptome Initiative 2018). Hence, it is probably safer to say that most lineages of life have simply not been sufficiently examined for polyploidy rather than claiming *a priori* that polyploidy is rare in certain lineages. More parochially, genome sequencing has also supported Ohno's (1970) hypothesis of two rounds of WGD at the base of the vertebrate phylogeny, referred to as the 2R events. The products of these ancient duplications continue to have implications for human health (Makino & McLysaght 2010; Almal & Padh 2012; Rice & McLysaght 2017a). Among the more obvious remnants of the vertebrate 2R events is the Hox gene family that regulates development, a duplication that may have been critical in allowing the vertebrate radiation (Holland 1999).

As we will describe in later sections, not all duplicates created by polyploidy are retained, and the over-retention of duplicated genes involved in the regulation of other genes after multiple polyploidization events allows for the rewiring of the cell regulatory network and may set the stage for developmental innovations (Freeling 2009). In some taxonomic groups, such as plants and yeast, polyploidy events have been involved in metabolic innovations (Conant & Wolfe 2007; van Hoek & Hogeweg 2009; Ruprecht et al. 2017a,b; Wisecaver et al. 2017; Lopez-Nieves et al. 2018), have escalated co-evolutionary arms races (Edger et al. 2015), and may have even enabled survival of mass extinction events (Soltis & Soltis 2016; Van de Peer et al. 2017).

Given that there appears to be a pattern in the fate of some gene duplicates, several researchers have investigated the question of whether or not there are rules governing the

evolution of polyploid genomes (Freeling 2009; Soltis et al. 2016). However, despite support for several hypotheses (e.g., dosage balance hypothesis, Birchler & Veitia 2012), there is not yet enough evidence to make strong conclusions about the universal outcomes of WGD. Among the open questions regarding gene and genome duplications is how interacting networks of the cell shape the evolution of gene expression in duplicates. More specifically, questions regarding the relative role of changes in genome structure vs. gene expression have yet to be answered in a comprehensive way for species experiencing gene and genome duplications (reviewed in Soltis et al. 2016). Complementary to this uncertainty is the idea that changes in these two factors may not occur in the same way, or at the same pace, across differing species (Soltis et al. 2016). In order to answer these questions, we advocate that methodological innovations, in concert with improved conceptual models, are necessary to advance our understanding of polyploid lineages. We echo Soltis et al. (2016), Wendel et al. (2018), and Edger et al. (2018) on the importance of including phylogenetic-level sampling to compare species, rather than conducting analyses in pairwise frameworks (illustrated in **Figure 1.1**). Likewise, studies that employ population level sampling to capture and reconstruct the variation inherent in the process of genome evolution will be especially useful for gaining a better understanding about general processes governing the fate of duplicated genes and genomes. Lastly, to truly understand if there are rules for polyploidy, analyses of duplicated genomes should be expanded to more lineages across the tree of life, rather than focusing solely on a single group (e.g., plants).

Figure 1.1 Hypothetical scenarios demonstrating different approaches for analyzing the allopolyploid *Brassica napus*. In each panel, we illustrate a hypothetical chromosome segment with several genes. Gene size and distances between genes are not included. **(a)** Pairwise: Comparisons between the different subgenomes of *B. napus* (A and C) are independently compared to their respective parental subgenomes from *B. oleracea* (C) and *B. rapa* (A). Mechanisms such as transposable element (TE) insertion (black bars), gene methylation (pink bars), and gene loss (empty gene blocks) can all be examined to understand their effects on gene expression (bars above gene blocks) and genome evolution. **(b)** Phylogenomics: Including more distant relatives of the focal species is used to help anchor gene order, show patterns of ancestral polyploidy (e.g., *Brassica* triplication) and homoeologous recombination (dark arrows), and reveal different patterns of gene retention, loss, and methylation in the context of the broader phylogeny. This is the framework used by POInT (Conant & Wolfe 2008a; Conant 2014). Note that the phylogeny and phasing of homoeologous segments allows us to distinguish "shared losses," where all taxa retain the same copy or copies of the homeologs, from "reciprocal losses," where different taxa retain different copies. Reciprocal losses have been implicated in reproductive isolation in polyploid organisms (Scannell et al. 2006). **(c)** Population genomics: Sampling multiple individuals per species reveals variation in patterns of gene loss (dark circles) and methylations (pink circles) owing to the genealogical process of coalescence. For simplicity, we illustrate only one possible genealogy. Events are marked on the gene tree with circles using the gene number and the corresponding genome number (G#) noted above (from left to right, time proceeds from past to present). For example, the first event when looking from left to right is a loss of gene 6 in Genome 3. This gene loss affects all sampled individuals except the bottom individual from *B. nigra*, which branched from the rest of the genealogy before the event occurred. In reality, each genome would have its own gene tree, and other processes such as recombination would need to be considered. Though not illustrated here, the frameworks presented in panels **(b)** and **(c)** are equally suitable for examining gene expression and TE insertion as well.

In this review, we survey recent developments on the prevalence of genome duplications across the phylogeny of eukaryotes, with particular emphasis on events in the teleost fishes, fungi, and land plants, where these events are especially prevalent and/or well studied. We then explore the role of polyploidy in genome evolution and ask if there are rules for polyploidy. Following this general discussion on genome evolution, we review recent studies addressing the effects of polyploidy on genome structure and gene expression. Next, we give several examples of models useful for making evolutionary inferences in polyploids at the gene network, phylogenomic, and population genomic levels, and highlight the importance of making models that are **ploidy-aware**. Finally, we provide guidelines for future work that can take advantage of these modeling frameworks to gain new insights into the general principles of polyploid evolution.

PREVALENCE OF POLYPLOIDY

Polyploidy and Fish – The 3R and 4R Events

Since Ohno's (1970) original hypothesis about the vertebrate 2R events, whole genome sequencing has revealed a complex history of polyploidy in fishes, including a WGD event near the origin of teleost fish (3R), as well as a more recent WGD in the salmonids (4R; Hermansen et al. 2016). These WGD events have been associated with phenotypic novelty, including the adaptation of hearts to aquatic life and the origins of electric organs in teleost fish (Moriyama & Koshiba-Takeuchi 2018). In their sequencing and analysis of the Atlantic salmon genome, Lien et al. (2016) sought to understand the process of diploidization following the salmonid WGD. They suggest that large chromosomal rearrangements were largely responsible for the breakdown of multivalent

pairings during meiosis, leading to disomic inheritance. Interestingly, this process of diploidization is incomplete: up to 10% of the genome is still experiencing tetrasomic inheritance. Additional studies on the evolution of gene retention and loss, as well as gene expression, have leveraged the genomic resources available for teleost fish to investigate lineage specific diploidization (Robertson et al. 2017) and patterns of sub- and neofunctionalization of duplicated genes (Pasquier et al. 2017). Others have reviewed the conceptual framework surrounding comparative genomic analyses of 3R and 4R gene duplicates and their fates across different lineages of teleosts (Hermansen et al. 2016). These serial WGD events have also led to novel methodological developments for inferring patterns of large-scale synteny in ancestral genome reconstructions. Using their new approach, Nakatani & McLysaght (2017) demonstrated that small-scale rearrangements were more prevalent than previously thought, adding an additional wrinkle to the evolution of post-WGD teleost fish genomes.

Polyploidy in Fungi

The systems biology of polyploidy in fungi, particularly the fermenting yeasts, has been studied for some time (Roman et al. 1955; Sémon & Wolfe 2007). However, questions remain: the ancient WGD in bakers' yeast and its relatives was only recently conclusively shown to be an allopolyploid (Marcet-Houben & Gabaldón 2015). It is also still a mystery why the fungi do not have more polyploid events. Campbell et al. (2016) explore this question and hypothesize that this conclusion may be an artifact of the difficulties in identifying ancient polyploidy in fungi. Most studies of ancient polyploidy in fungi are on fermenting yeasts and use comparative methods on the numerous species with complete genomes in the genus *Saccharomyces*. For example, Kuang et al. (2016) show that some

paralogs continue to affect regulatory networks long after they were originally retained, and that the loss of paralogs may eliminate redundancy, but also limits the long-term potential for adaptation. Additionally, genome-wide methods of measuring rates of adaptation in diploids and polyploids can now find a spectrum of mutations related to specific genomic positions, such as their proximity to recombination hotspots (Marad et al. 2018), allowing for direct comparisons of WGD and adaptation in polyploids and their progenitors.

Other studies of polyploidy in fungi focus on more recent WGD. For example, Gerstein et al. (2017) used a haploid, diploid, and polyploid strain of a human fungal pathogen to measure the rate at which the fungal species returned to their baseline polyploid level in different environments. They tested the “nutrient limitation hypothesis,” which suggests that because haploid cells have a higher surface to volume ratio, species that can exist as either diploids or haploids should be driven to the haploid state to increase nutrient uptake when they are limited (Mable 2001). Their results were surprising, in that a more complex picture emerged. Under phosphorus limitation and in minimal medium, lower ploidy levels were selected for. However, in complete medium or when nitrogen was limited, higher ploidy levels were enriched (Gerstein et al. 2017). These results suggest that the assumption that polyploids form most readily in nutrient poor or harsh environments may require additional qualification and that there seems to be much more to discover in how neo- and mesopolyploidy has shaped fungal evolutionary history

Polyploidy in Plants

Polyploidy has been most extensively studied in land plants (Soltis & Soltis 2016; Soltis et al. 2016; Zhan et al. 2016; Van de Peer et al. 2017; Landis et al. 2018). A recent

phylogenomic analysis integrated transcriptomes and genomes for 1,173 species across all major groups of green plants (Viridiplante) and relatives, including the red algae and glaucophytes (freshwater algae; One Thousand Plant Transcriptome Initiative 2018). The study mapped over 240 ancient polyploid events onto a robust phylogeny, including over 150 newly inferred events. Strikingly, at least one ancient WGD was inferred in the ancestry of every species of land plant, with the exception of *Selaginella* and some liverworts; in contrast, most chlorophytes (green algae), streptophyte algae, glaucophytes, and red algal lineages showed little or no evidence of WGD (One Thousand Plant Transcriptome Initiative 2018). A recent study of the liverwort *Marchantia* documented similar patterns of gene and genome duplication in the early evolution of land plants (Bowman et al. 2017). The One Thousand Plant Transcriptome Initiative (2018) not only phylogenetically localized WGD events, but also concluded that many ancient polyploid events are associated with key evolutionary innovations, including the origins of multicellularity, vascular systems, roots, seeds, flowers, and fruits. Independent genome duplication events in, for example, the bryophytes and angiosperms, suggest parallel evolution of pathways, such as land plants' cell walls (Ruprecht et al. 2017a,b). The association between polyploidy and novelty has also been found in studies of more recently diverged plant groups, such as in the Brassicales and the origin of novel metabolic pathways (Edger et al. 2015; Wisecaver et al. 2017) and red betalain pigments in the Caryophyllales (Smith et al. 2017; Lopez-Nieves et al. 2018).

As in the study of polyploidy in fish and fungi, the relationship between diversification and diploidization over time in plants has been investigated (Dodworth et al. 2015; Mandakova & Lysak 2018). These studies observe that novel traits and

diversification are usually not initiated by genome doubling *per se*; rather, differentiation of the primary polyploid populations through a range of processes, resulting in post-polyploid evolution, reproductive isolation among post-polyploid offspring, and subsequent speciation and cladogenetic events. One advantage of studying post-polyploid evolution in plants is that polyploidy is recurring across multiple lineages through time, providing the opportunity for studying any rules for polyploidy in the context of convergent evolution in recent polyploids (Mandakova & Lysak 2018, Novikova et al. 2018).

COMPLICATIONS FOR DETERMINING IF THERE ARE THERE RULES FOR POLYPLOIDY

With such prevalence of polyploidy across the tree of life, it seems reasonable to ask if there are rules associated with WGD. One major pattern often observed after WGD is the process of the organism returning to a diploid-like state (fractionation) and a general convergence in the genes lost during this process. That is to say that there is nonrandom gene loss in these genomes after WGD occurs (Birchler et al. 2005; Birchler & Veitia 2007). As discussed by Soltis et al. (2016), we are still in the early stages of collecting data to have a complete understanding of these processes. Not only are there missing data across different types of genomic and ecological studies (see Table 1 in Soltis et al. 2016), but there are also missing data on understanding how polyploid organisms respond over time to being polyploid (*see Figure 2* in Soltis et al. 2016). We do know that there are both short-term and long-term processes that act on these species (Edger et al. 2018; Wendel et al. 2018), but we do not yet have enough sampling, within enough species, to fully understand all of the patterns associated with WGD (Soltis et al. 2016). Looking for rules

gets further complicated by the different types of polyploidy, autopolyploidy and allopolyploidy, each of which has their own set of challenges in understanding the biological phenomena associated with them. Below, we review recent work, mostly addressing allopolyploid plants, that evaluate the changes associated with polyploid organisms returning to a diploid-like state, including patterns of genome structural and expression changes, as well as the eventual observation of sub- and neofunctionalization. Lastly, we discuss the dosage balance hypothesis as a mechanism for generating these observed patterns.

Genome Structural Changes - Homoeologous exchanges

One unique feature of polyploid genomes is that not only can they undergo homologous exchanges of chromosomes, but they can also undergo homoeologous exchanges between the duplicated chromosomes contributed by the diploid parents (Osborn et al. 2003). This often results in gene presence/absence variation (PAV). These homoeologous exchanges are especially interesting as they have been associated with many phenotypic changes seen in polyploids (Pires et al. 2004; Gaeta & Pires 2010; He et al. 2017; Hurgobin et al. 2017). Although homoeologous exchanges have been most extensively studied in the *Brassica* crops (He et al. 2017; Hurgobin et al. 2017; Stein et al. 2017), high quality genomes have enabled our ability to observe homoeologous exchanges in other plant genomes such as peanut (Nguepop et al. 2016), coffee (Lashermes et al. 2016), wheat, and other species (He et al. 2017). However, patterns of PAV can arise from multiple processes, therefore a challenging problem remains in terms of distinguishing and classifying homoeologous exchanges resulting in reciprocal crossover versus gene conversion events (Gaeta & Pires 2010; Scienski et al. 2015; He et al. 2017). We echo

Edger et al. (2018) in suggesting that a phylogenomic approach would improve our accuracy in assigning homoeologs to the correct parental subgenome. In particular, we argue that assessing homoeologs in both phylogenomic and, if possible, population genomic frameworks would help to explain patterns as seen in maize and soybean. These species differ in their levels of rearrangement, with maize being much more rearranged, possibly because its polyploidy event involved more diverged parents (Zhao et al. 2017). Phylogenomics combined with the rise of high-quality pan-genomes for some species like *Brachypodium* (Gordon et al. 2017), will play important roles in our ability to distinguish CNV and gene PAV that arises from homoeologous exchanges in the context of polyploid genome structural evolution, as well as their effects on phenotype (Golicz et al. 2016; Hurgobin et al. 2017).

Genome Structural Changes - Biased fractionation

One post-polyploidy pattern that can give rise to rearranged genomes is biased fractionation, where the losses of homoeologous genes after WGD are biased toward one of the parental genomes (reviewed in Freeling 2009). These losses break the synteny blocks shared between a polyploid and its diploid relatives, giving rise to new genome structures in the former. Recent studies in neopolyploid *Mimulus* (Edger et al. 2017) and paleopolyploid maize (Renny-Byfield et al. 2017) demonstrate the range of methods that can be employed to study the impact of polyploidy and fractionation on genome structure and gene expression. Many studies on fractionation use pairwise comparisons, resulting in debate on the true prevalence of biased fractionation. Therefore, to more accurately model biased fractionation, we propose moving from pairwise to phylogenomic comparisons (**Figure 1.1**). Recently, using a phylogenetic model of post-polyploidy evolution from the tool

POInT (see section *Phylogenomics and Phylogenetic Networks - Modeling polyploidy resolution with POInT*), Emery et al. (2018) presented robust evidence for the existence of biased fractionation after WGDs in *Arabidopsis*, the grasses, and yeasts, highlighting the importance of asking questions about polyploidy in a phylogenomic framework.

Expression Changes in Allopolyploids - Subgenome Dominance

Many studies have shown that, in allopolyploids, one genome becomes dominant to the other (reviewed in Edger et al. 2018). This pattern is observed both in terms of gene losses (the biased fractionation just mentioned) and changes in expression. What is not well understood is what makes one genome dominant over the other and how quickly this occurs in neopolyploids. To examine this, Edger et al. (2017) analyzed the established allopolyploid species *Mimulus peregrinus* (Monkeyflower), its progenitor species (*M. guttatus* and *M. luteus*), a resynthesized *M. peregrinus*, and a resynthesized triploid hybrid (*M. robertsii*). They found that subgenome dominance appeared immediately after polyploidization and increased over time. The primary drivers of the dominance were epigenetic modifications that led to homoeolog expression bias (Edger et al. 2017), both of which we discuss below in more detail.

Expression Changes in Allopolyploids -Epigenetic modifications

The importance of epigenetics in polyploid plant evolution is becoming increasingly clear with many recent studies and reviews (Edger et al. 2017; Gardiner et al. 2017; Horvath & Slotte 2017; Vicent & Casacuberta 2017; Baldrich et al. 2018; Ding & Chen 2018; Jiao et al. 2018; Rodriguez & Arkhipova 2018; Wendel et al. 2018). Epigenetic modifications have been suggested to play a large role in subgenome dominance and eventual diploidization after WGD (Edger et al. 2017; Vicent & Casacuberta 2017;

Wendel et al. 2018) by affecting the expression of nearby genes, often resulting in homoeolog expression bias. Wendel et al. (2018) discuss the complex role of epigenetics in polyploid plants after WGD in detail. Below, we focus on three main modifications: gene methylation, transposable element (TE) methylation, and small RNAs.

To evaluate the role of gene body methylation sites in polyploids, Gardiner et al. (2017) assessed gene body methylation across a diverse collection of hexaploid wheat from different environments. The authors found that not only were gene body methylation sites specific to the different environments, but that these patterns could eventually become hard-coded as methylation polymorphisms that are passed down from generation to generation (Gardiner et al. 2017). Over time, these gene body methylation sites influenced phenotypes by driving local adaption, affecting salt tolerance and heading date traits across these wheat collections (Gardiner et al. 2017).

Although gene body methylation has long been known to affect phenotype, TE insertion is also increasingly seen as a mechanism for disrupting expression of genes (Hollister & Gaut 2009; Horvath & Slotte 2017; Dubin et al. 2018), especially in polyploids (Vicent & Casacuberta 2017). Vicent and Casacuberta (2017) argue that TEs, particularly when combined with polyploidy, are major players in plant evolution, and it has been hypothesized that TEs in polyploids may play a role in subgenome dominance as the amount, efficiency, or proximity to genes of TEs in the parental genomes may affect which genome becomes dominant. Indeed, Edger et al. (2017), when testing this hypothesis in *Mimulus*, observed that the dominantly expressed subgenome had a lower TE abundance. This is true of other plant systems including *Brassica* (Cheng et al. 2016) and maize (Renny-Byfield et al. 2017). TEs also seem to have played a role in shaping polyploid

animal genomes such as salmonid fishes, amphibians, nematodes, and rotifers, however further work is needed to improve TE annotation in these systems before broader conclusions can be drawn (Rodriguez & Arkhipova 2018).

Most recently, the role of small RNAs in mediating expression to avoid the negative effects of TE expression is becoming appreciated. As stated above, there is a trade-off that develops between TE silencing and regulation of nearby genes (Horvath & Slotte 2017). To test how TE silencing by small RNAs and gene regulation affects selection on TE insertions, Horvath and Slotte (2017) analyzed 166 individuals of *Capsella grandiflora*. They determined that silenced TEs that were closest to genes were under stronger purifying selection than silenced TEs farther away, and TEs farthest away were not silenced at all (Horvath & Slotte 2017). The authors also note that these small RNAs are able to silence other, distant TEs (Horvath & Slotte 2017). If two such systems from different parents are merged in an allopolyploid, the effect might be drastic, as the small RNAs could interact with each genome, driving inappropriate silencing of critical genes. To further study these effects on the stability of an allopolyploid genome, Jiao et al. (2018) assessed small RNAs in two resynthesized wheat allotetraploids, using one subgenome from a common parent and the other from different parents (AADD and S^LS^LAA). They observed that small RNAs from the common parental subgenome affected each allotetraploid differently and that these differences resulted in unequal changes in expression of some genes due to activation of TEs (Jiao et al. 2018). Ultimately, the authors suggest that these differences in stability may in part explain the evolutionary trajectories of wheat allotetraploids and their progenitors (Jiao et al. 2018).

Expression Changes in Allopolyploids - Homoeolog expression bias

Arguably the final player in the story of subgenome dominance is homoeolog expression bias, the unequal contribution of expression from parental alleles in an allopolyploid. These expression patterns may occur for a number of reasons, including SNPs, indels, and the epigenetic and small RNA effects just discussed. While assessing segmental allopolyploid rice, Sun et al. (2017) found widespread homoeolog expression bias. They determined that homoeolog reorganization caused CNV, which in turn resulted in the homoeolog expression bias they were observing (Sun et al. 2017). This study is of particular interest as it addresses segmental allopolyploids, providing insight to how population level variation interacts when brought together in polyploidy. Although these patterns of biased expression have also often been observed after WGD, the underlying mechanisms that cause them remain unclear. To better understand if external stresses were triggering these patterns, Powell et al. (2017) assessed if homoeolog expression bias was driven by biotic or abiotic stresses, looking specifically at how homoeolog specific expression changed in wheat during a fungal infection. They concluded that homoeologs had differences in the degree of responsiveness to the fungal infection, resulting in homoeolog expression bias (Powell et al. 2017). These patterns are not unique to plants: Ren et al. (2017) observed similar patterns in teleost fish. They found that homoeolog expression bias appeared to result in the rapid growth of the allopolyploids compared to their parental species (e.g., heterosis; Ren et al. 2017; Cheng et al. 2018). Many studies also note the correlation between homoeolog expression bias and phenotypic changes (Yang et al. 2016; Powell et al. 201; Sun et al. 2017), suggesting that it plays an important role in polyploid genome evolution. Linked back to the question of biased fractionation, it

is argued that homeologs with lower expression may have lesser fitness defects when deleted, meaning that the expression bias might be the first step to the biases observed in homeolog retention (Schnable et al. 2011; Cheng et al. 2012).

Sub- and neofunctionalization

The role of neofunctionalization and subfunctionalization has attracted a great deal of interest in the field, with analyses arguably running beyond what the available data can support (Conant et al. 2014). Distinguishing between these two models (and indeed distinguishing them from many others less often cited, Conant & Wolfe 2008b), requires two types of inference that are difficult to make reliably: an inference as to the ancestral function of the unduplicated gene (including its expression pattern) and an inference as to the type of selection acting in the immediate aftermath of the duplication itself (Conant & Wolfe 2008b). In particular, both models require the rejection of a neutral model of functional divergence, a point we return to in a later section.

The dosage balance hypothesis

One possible mechanism for the convergent patterns seen in post-WGD fractionation (e.g., gene loss) seen in polyploids is the dosage balance hypothesis (DBH; Birchler et al. 2005; Freeling 2009; Birchler & Veitia 2012; Conant et al. 2014; Birchler et al. 2016), which states that genes that are highly connected (i.e., their products are involved in many interactions) are more likely to be retained post-WGD. The most basic rule about duplicate genes, and by extension, polyploidy, is that duplicates are unstable and are often lost through genetic drift (Nei & Roychoudhury 1973; Li 1980). And indeed, gene loss after polyploidy is extensive and rapid (in evolutionary terms; Scannell et al. 2007). With the availability of complete polyploid genomes for study, one of the unexpected patterns

to emerge was the strong patterns of convergent evolution in those losses (Freeling 2009; Mayfield-Jones et al. 2013; Barker et al. 2016; Soltis et al. 2016; Washburn et al. 2016). Most prominent among the genes not rapidly lost were genes involved in the regulation of other genes after multiple polyploidization events, which has been seen repeatedly (Maere et al. 2005; Freeling & Thomas 2006; Freeling 2009; Freeling et al. 2015). Freeling (2009) describes that the fates of genes duplicated by polyploidy (WGD) and small-scale duplications (SSD) have predictable yet different outcomes in being lost or kept, most likely due to the principles underlying the DBH (Freeling 2009; Birchler & Veitia 2012; Conant et al. 2014; Birchler et al. 2016).

At the most basic level, a gene's dosage is expected to be correlated with its expression (Birchler & Veitia 2007). The DBH extends on this principle by asking how dosage-induced changes in expression level alter a gene product's interactions with other gene products (Birchler & Veitia 2012). These interactions are complex and considerable biochemical knowledge may be required to predict the effects of a dosage change (Coate et al. 2016).

With this model, the proposed explanation of the outcome of SSD versus WGD is straightforward. Genes that, for reasons of expression buffering, sparsity of interactions, or other types of functional robustness, are tolerant to changes in dosage will be observed among the genes duplicated in SSD events. After polyploidy, however, these same genes will generally be unlikely to induce a selective cost if one of the two homoeologous copies is lost by drift, and will be less common among the surviving duplicates. On the other hand, those genes that are intolerant to dosage change will be, almost by definition, intolerant to duplication through small-scale duplication. After polyploidy, however, they will be part

of clusters of homoeologs that function in balance together, and the same forces that oppose their duplication in small-scale events will oppose the loss of homoeologous copies after polyploidy (Pires & Conant 2016). To further explain this concept, we use a baking analogy: adding an egg to a cake recipe (a small-scale duplication) will probably result in a runny mess; doubling the recipe (WGD) just makes a bigger cake (borrowed with permission from Jessica Neil). The DBH is now supported by observations from polyploids ranging across eukaryotes such as *Paramecium*, yeast, humans, and plants (Makino & McLysaght 2010; Birchler & Veitia 2012; Conant et al. 2014; Birchler et al. 2016; Li et al. 2016; Xie et al. 2016; Tasdighian et al. 2017). The DBH also has implications in other areas of biology: dosage effects may alter our ability to detect or model heritability and epistasis in quantitative genes (Birchler et al. 2016), yet it may help explain the duplication structure of plant genomes (Li et al. 2016).

MODELS FOR EVOLUTIONARY INFERENCES IN POLYPLOIDS

For species experiencing a WGD event, new tools have begun to help us understand post-polyploidy patterns of genetic polymorphism at both the temporal, spatial, and molecular levels, though there is ample room for future research. In general, our understanding of the evolutionary processes that shape genetic diversity within and among populations and species has flourished because of increasingly sophisticated and innovative statistical and computational models. However, modeling evolutionary forces in polyploids is complicated by a number of factors not experienced by diploids, including uncertainty in allelic dosage when inferring genotypes, variable patterns of allelic inheritance across generations (e.g., auto- vs. allopolyploidy), and differential patterns of

gene loss and fractionation over time (Sankoff et al. 2012; Dufresne et al. 2014). Approaches for genotyping single nucleotide polymorphisms from low-coverage sequence data that account for some of these factors are beginning to emerge (Blischak et al. 2016, 2017; Zohren et al. 2016), and future models that build on this work will continue to facilitate the study of evolutionary history in polyploid lineages

As new tools are developed for polyploids, it is important to consider building models that are ploidy-aware, meaning that they are capable of handling mixed-ploidy samples (e.g., tetraploids, hexaploids, and lower-ploidy progenitors), as well as mixtures of auto- and allopolyploids. In some cases, it may be possible to extend existing methods for diploids, but in others additional model layers may be required. Below, we review a selection of recent investigations in polyploid gene networks, phylogenomics, and population genomics, highlighting instances where building ploidy-aware models can lead to novel insights. We particularly focus on the integration of inferred population or species phylogenies (phylogenetic networks when there is hybridization) with comparative data to test specific models of polyploid genome evolution (e.g., see **Figure 1.1**).

Polyploidy and Gene Network Evolution

Once the role of gene and molecular network structure in the evolution of polyploid genomes is appreciated, it becomes apparent why polyploidy represents a unique opportunity for innovation: innovations can occur at the level of subnetworks, not merely single genes. We have formalized this proposal into a network partitioning procedure that can detect network subfunctionalization/partitioning for post-polyploid networks (Conant & Wolfe 2006). A recent example of this idea is the demonstration by DeSmet et al. (2017) that two groups of homoeologous genes in *Arabidopsis* diverged in their tissue-specific

expression (one group specializing to roots and pollen and the other to leaves) after the most recent WGD. The molecular functions of these genes and their tissue-specific expression patterns lead these authors to argue that a group of genes with a common ancestral function subspecialized to two groups, one involved in polarized growth in pollen and roots and the other in defense and stress response. This finding mirrors our own that stress responses specialized in yeast post-WGD (Conant & Wolfe 2006). Zhu et al. (2013) also used a network approach to study the fate of duplicated gene copies in yeast, finding that molecular interactions of genes evolved more slowly after WGD. These examples, which focus on network-level rather than single gene-level changes, are consistent with a neutral process of subfunctionalization. In the metabolic realm, however, there are a growing number of examples of pathway-level innovations with features more consistent with some form of neofunctionalization, including the evolution of glucose fermentation in yeast and of the diversification of the glucosinolate repertory in the Brassicales (Conant & Wolfe 2007; van Hoek & Hogeweg 2009; Edger et al. 2015).

Polyploidy and Gene Network Evolution - The DBH as a network hypothesis

The DBH found support from early analyses in plants and yeast (Papp et al. 2003; Birchler et al 2005). Papp et al. (2003) studied baker's yeast, which was one of the first organisms where significant efforts had been made to elucidate the structure of cellular networks. Hence, they illustrated that one way to frame the DBH is in network terms, proposing that highly connected genes will be more likely to show dosage-dependent phenotypes, and are hence more likely to be retained as homoeologous pairs post-polyploidy. Explicit analyses of the yeast and *Arabidopsis* polyploidy events in terms of their cellular networks have supported this contention (Papp et al. 2003; Bekaert et al.

2011; Conant 2014). In yeast, one can see that the single copy genes (e.g., genes whose homoeolog has been lost) show patterns consistent with the DBH: the earliest losses after the WGD were genes with few interactions and it was only later that genes with greater number of interactions started to be lost.

This network view of polyploidy is gaining currency (*Arabidopsis* Interactome Mapping Consortium 2011; Chae et al. 2012), with gene co-expression networks in particular being amenable to study in a wide range of taxa beyond just model organisms (Gallagher et al. 2016; Ruprecht et al. 2017a,b). One of the reasons for the increasing popularity of network approaches is that they can uncover unexpected relationships. For example, human homoeologs from the 2R duplication event are less likely to show copy number variations (Makino & McLysaght 2010). However, using chromatin interaction data that measured the three-dimensional co-localization of homoeolog pairs in the nucleus, Xie et al. (2016) were able to show that homoeologous gene pairs from 2R were more likely to be co-localized than other paralogs, and that the most co-localized homoeologous pairs were also those least likely to show copy number variation. Such a pattern would be explicable if gene spatial co-localization were a selectable trait, because co-localization helps maintain coexpression between homoeologous or functionally related genes (Homouz & Kudlicki 2013; Thévenin et al. 2014).

Phylogenomics and Phylogenetic Networks

Given the prevalence of WGD, incorporating polyploids in the inference of phylogeny is an important, though sometimes avoided, task. Among the most difficult problems in such phylogenetic analyses is that allopolyploids, formed by hybridization, do not have a tree-like pattern of evolutionary descent. Inferring the reticulation events that

form allopolyploids, and even homoploid hybrids, is a theoretically and computationally demanding task. However, methodological developments are available to explore the complexity of navigating the space of species networks (trees with reticulate edges), rather than just assuming a strictly bifurcating tree (Solís-Lemus & Ané 2016; Jones 2017; Kamneva & Rosenberg 2017; van Iersel et al. 2017; Wen & Nakhleh 2017). Furthermore, many of these new methods are being implemented in open source software packages such as PhyloNet (Wen et al. 2017) and PhyloNetworks (Solís-Lemus et al. 2017), making them accessible to the community.

Adapting these tools for phylogenetic inference with polyploids will provide a much-needed tool set for understanding patterns of hybridization and WGD. A recent study by Kamneva et al. (2017) used PhyloNet to test hypotheses about the formation of polyploid strawberries (*Fragaria*; Rosaceae) ranging from tetraploid to decaploid. Though not specifically built for polyploids, PhyloNet is able to model multiple haplotypes in each lineage. It can hence infer hybridization in these species when the assumptions of its model (the coalescent) are not violated, making it most appropriate for recent allopolyploids. The software package GRAMPA is useful for detecting older WGD events (Thomas et al. 2017). It uses a known species tree and a set of gene trees to resolve putative polyploidy events using a gene tree reconciliation algorithm. Both the species tree and the gene trees can be multi-labeled (containing multiple leaves with the same name), allowing for homoeologous gene copies and allopolyploid subgenomes to be modeled.

A potential shortcoming of species network inference methods is that they do not scale well to large numbers of species (Kamneva et al. 2017). The computational complexity of searching network space, along with inferring the number of hybridization

events, makes the algorithms extremely slow as more species are added (van Iersel et al. 2017). A potential solution for this problem would be to use faster parsimony-based methods to get a set of candidate networks that can then be used as input in PhyloNet or PhyloNetworks to calculate the likelihood on a small set of networks (Solís-Lemus et al. 2017; Wen et al. 2017; van Iersel et al. 2017). Species tree inference methods that use quartets have also proven to be fast and useful [e.g., SVDquartets (Chifman & Kubatko 2014) and ASTRAL-II (Mirarab & Warnow 2015)]. Extending these methods to model hybridization or the subgenomes of an allopolyploid could then allow for faster inference of species networks for bigger data sets.

Phylogenomics and Phylogenetic Networks - Modeling polyploidy resolution with POInT

Another example of a polyploid-aware modeling approach is the tool POInT (the Polyploidy Orthology Inference Tool; Conant & Wolfe 2008a; Conant 2014). It is a unified, phylogenetic and model-based framework for analyzing the evolution of polyploid genomes from across the eukaryotic tree of life. POInT scaffolds together chromosome segments from multiple genomes sharing a polyploidy event. Using an evolutionary model of gene loss, it then phases those segments into orthologs and homoeologs using the phylogenetic signal of shared gene losses. The result is shared blocks of double-conserved synteny between the genomes analyzed (see **Figure 1.2**). POInT's modeling framework allows one to make a variety of inferences, including: (1) The phylogeny of the post-polyploidy organisms; (2) The timing of the gene losses on the phylogeny; (3) The presence or absence of patterns such as biased fractionation or duplicate fixation, and (4) the occurrence of post-polyploidy gene conversion (Scienski et al. 2015). Our most recent POInT analysis compared and contrasted the yeast WGD with the last *Arabidopsis*

polyploidization event, as well as the grass ρ WGD event, finding strong evidence for biased fractionation in all three: the bias was short-lived in yeast but persistent in the two plant lineages (Emery et al. 2018). Moreover, the bias is linked to function: single-copy genes whose products physically interact are more likely to come from the same parental genome than expected.

Figure 1.2 Resolving patterns of biased fractionation with POInT. Double-conserved synteny blocks created by polyploidy and homoeolog loss after the recent *At- α* genome duplication in *Arabidopsis* and its relatives. POInT (section 4.2.1) was used to infer the orthologous regions of these six genomes: the probabilities above each column are its estimate of the probability of this orthology relationship relative to the 26-1 (63) other possible sets of orthology relationships. Genes in these blocks may either have surviving duplicates in at least some taxa (pink/tan) or may be single-copy across all species (green if derived from the first subgenome and blue if from the second). Adjacent genes are joined by lines if they are each other's genomic neighbors in the extant genome in question: the absence of such a line indicates a synteny break. The branch lengths (estimated at the species level) in the tree are proportional to the number of gene losses on that branch: hence larger values correspond to a relatively higher chance that a gene duplicated at the start of a branch will have been single-copy by the end. The effects of biased fractionation can be seen in the greater number of retained single-copy orthologs from subgenome #1 (Emery et al. 2018 gives a more complete description of this phenomenon).

Investigating patterns of polymorphism in populations of polyploids, or populations of different ploidy levels, presents a unique challenge when attempting to disentangle the effects of drift, migration, and selection on the distribution of genetic variation (Meirmans et al. 2018). Linking demographic factors with increases in ploidy and coupling them with methods for inferring population structure and connectivity through migration can provide important insights on the timing of WGD, the dispersal of polyploid lineages to new geographic regions, and the genomic divergence of polyploid taxa from their parental species (Jighly et al. 2018). In their work on the allotetraploid *Capsella bursa-pastoris* (shepherd's purse; Brassicaceae), Cornille et al. (2016) used genome-wide SNP data to model the demographic history of the species across Europe, the Middle East, and Asia. Using approximate Bayesian computation, they compared models with differing dispersal patterns and rates of gene flow and found support for a model of independent dispersal events from the Middle East into Europe and Asia with moderate levels of migration. Another study on British birch trees (*Betula*; Betulaceae) by Zohren et al. (2016) discovered unidirectional gene flow between an allotetraploid species (*B. pubescens*) and one of its putative diploid progenitors (*B. pendula*). Studies in the 4R salmonid fishes have sought to understand how polyploidy affects sequence divergence in sockeye salmon (*Oncorhynchus nerka*); Limborg et al. (2017) sorted genomic loci into duplicated (tetrasomic inheritance) and non-duplicated (disomic inheritance) SNPs using linkage maps and the R package PolyFreqs (Blischak et al. 2016) to estimate allele frequencies and population differentiation. When both duplicated and non-duplicated loci were included in

their scans for selection, duplicated loci revealed novel outlier regions that were not detected when using only non-duplicated loci.

In addition to patterns of DNA sequence variation, population-level variation in structural changes associated with polyploidy can have important consequences for gene expression, as well as the differential retention of duplicated gene copies. The concept of a pan-genome (Tettelin et al. 2005), a natural extension of a reference genome to include gene PAV and CNV across individuals, is helping to formalize the analysis of differential retention and loss of genes across populations. Gordon et al. (2017) recently used a pan-genome framework to analyze 54 lines of the grass *Brachypodium distachyon* (Poaceae) and found that variation in gene content correlated with population genetic structure.

Other structural features that influence gene retention and expression include TEs and DNA methylation. As discussed above (section 3) it is also likely that epigenetic changes have a strong influence on patterns in subgenome dominance, gene expression evolution, cellular network organization, and the maintenance of post-polyploidy redundancy. However, how this variation is distributed across populations remains largely unknown. Future work that builds on existing models for disentangling the effects of demography and selection (Sheehan & Song 2016; L. Wang et al. 2017), convergent adaptation (Lee & Coop 2017), and the spatial and temporal distribution of polyploidy will provide an important theoretical framework for testing hypotheses about genome evolution after WGD.

Integrating Inferences on Trees, Phylogenetic Networks, and Population Graphs

As more methods for studying the demographic and phylogenetic histories of polyploids emerge, a natural next step for understanding the broader patterns of evolution

following WGD is to use these types of inference as a framework for modeling comparative data. Tools such as phylogenetic comparative methods are already commonly applied to make inferences about trait evolution on trees (Pennell & Harmon 2013). Recent work has also extended these methods to phylogenetic networks, making it possible to include the effects of hybridization on species' traits (Bastide et al. 2017). Using a known species history can also be useful for many downstream analyses including comparative genomics, phylogenetic models of gene expression and molecular evolution, inferring patterns of nucleotide substitution that corrects for gene tree incongruence, and inferring complex patterns of gene duplication and loss (Rohlf & Nielsen 2015; Szöllősi et al. 2015; Dunn et al. 2018; O'Toole et al. 2018). At the population level, several spatially explicit tools have been developed to model population structure and admixture (Bradburd et al. 2016), as well as methods for inferring population graphs (also called admixture graphs) that model divergence and drift (Cheng et al. 2017). Demographic modeling using software such as *∂a∂i* (Gutenkunst et al. 2009) and *fastsimcoal2* (Excoffier et al. 2013) has been successfully applied to diploids (e.g., Beissinger et al. 2016; Qi et al. 2017), and extensions of these methods to polyploid lineages could allow for more robust inferences of historical changes in population size associated with WGD or gene flow with progenitor species. Methods for genome-wide association studies (GWAS) have also been developed for more complicated evolutionary scenarios (Hejase et al. 2017), and could be further modified to include polyploidy.

Using these methods as a base upon which additional complexity can be added is a potentially powerful approach to model ploidy-specific data patterns. For example, Racimo et al. (2017) used estimated admixture graphs to control for genetic relatedness and gene

flow when modeling polygenic adaptation in humans, an approach that could be used in polyploid species to identify candidate genes under selection. Kryvokhyzha et al. (2016) used previously estimated demographic information when analyzing gene expression data in *Capsella bursa-pastoris* and found that expression differences were mainly explained by population structure. We particularly note how these examples highlight the potential to integrate past knowledge to inform future inferences.

Ultimately, building a statistical framework for combining geographic, genomic, epigenetic, TE, gene expression, and other types of 'omics data, all while remaining ploidy-aware, is an enormous challenge. Depending on the question at hand, it may be possible to extend existing methods to accommodate the desired level of complexity. For highly multidimensional data with no obvious way to construct a model, approaches such as machine learning can offer computationally efficient approaches for extracting the features of interest. Though it is still a developing area in evolutionary genetics, machine learning will likely be an increasingly valuable tool for integrating large, multivariate data sets to make inferences about polyploid genomes (Sheehan & Song 2016; Schrider & Kern 2018). It is worth emphasizing here, however, that the complexity of polyploid genome evolution warrants special caution when applying parameter-rich models or black-box algorithms. Testing the absolute fit of a model through simulations, and understanding when and how a trained algorithm may be overfit, will be key components of ensuring that results are not misinterpreted. Moreover, these types of evaluations can provide additional insights by revealing features of the data that are not captured by an analysis (e.g., identifying outliers), pinpointing them as targets for future study.

SUMMARY AND FUTURE DIRECTIONS

Though we are still on our way to discovering universal rules for polyploidy, if there are any, the research discussed in this review highlights many of the recent approaches taken for analyzing, describing, and understanding how WGD affects genome evolution. We also note how prevalent polyploidy is across the tree of life, and how important it will be to include additional groups of taxa in studies of polyploidy for an overall picture of how organisms evolve after WGD events. We stress in particular that we are currently lacking methods suitable for extracting general patterns from polyploids. However, with new models for the analysis of gene networks, phylogenomic, and population genomic data, we expect that this deficiency will be ameliorated over the coming years. Nevertheless, from a conceptual point of view, we believe that one other crucial element has sometimes been neglected in discussions of the effects of polyploidy: stochasticity. Understanding how much of the variation that we observe in polyploid genomes is governed by random, neutral processes will be of fundamental importance for drawing conclusions about how they evolve over time. With this in mind, we touch on a few areas where considering random processes can guide future research.

Polyploidy as a Window into Evolution by Neutrality and Selection

The impact of selection (directional or purifying) on changes in gene expression after WGD is a central question with regard to the maintenance of dosage balance. While selection acting to alter expression levels is one explanation of observed expression changes in duplicate genes, we and others have noted (Birchler & Veitia 2012; Conant et al. 2014) that there is another at least equally plausible mechanism for such divergence: drift (Khaitovich et al. 2004, 2005; Yanai et al. 2006). While it might seem surprising that

gene expression would evolve neutrally, there are clear theoretical reasons to expect that the selection on expression level can be weak (Wagner 2007), especially in multicellular eukaryotes with small effective population sizes (Lynch & Marinov 2015). Indeed, explicit modeling of weak selection interacting with drift to influence the evolution of dosage balance in duplicate genes has been shown in teleost fishes (Thompson et al. 2016).

Polyploid plants are the natural system for testing whether homoeolog expression divergence is truly neutral: they offer many more duplicate genes, all of the same age, than are available in other groups with many sequenced genomes, such as mammals or fruit flies. While these analyses would require high-quality comparative gene expression data from multiple related polyploid genomes and their diploid relatives, the theoretical tools for making the evolutionary comparisons exist (Rohlf & Nielsen 2015; Dunn et al. 2018), and results will almost certainly prove illuminating. Moreover, gene expression studies are not the only area that would benefit from such approaches: almost any phenotype hypothesized to be influenced by polyploidy should probably also be tested for deviations from the expected evolutionary patterns under drift.

Ancestral Polymorphism and Hemiplasy

Patterns of gene retention/loss, TE insertion, and methylation often show variation within and between populations or species. Understanding the mechanisms that generate this variation can be important for determining how selection acts on polyploid genomes. A vital component of this problem is again the establishment of a null expectation for the amount neutral variation present in a population. A common framework for such a null hypothesis in evolutionary genetics is the coalescent model (Kingman 1982), a stochastic process describing the genealogical relationships among a group of sampled alleles. In the

context of multiple species, it also describes the relation between gene trees and species trees, and provides a baseline prediction for how often gene trees may differ from the underlying species tree (reviewed in Degnan & Rosenberg 2009). Failing to account for this gene tree variation can lead to mistaken inferences. An excellent illustration of this problem was given by Mendes & Hahn (2016), who demonstrated that tests for positive selection reported more false positives due to an increase in apparent substitutions resulting from incomplete lineage sorting (ILS), a pattern they refer to as SPILS: Substitutions Produced by ILS. More broadly speaking, this phenomenon, termed hemiplasy, results from the incomplete sorting of ancestral polymorphism in a character state (Avisé & Robinson 2008; Hahn & Nakhleh 2016). An illustration of this can be seen in **Figure 1.1C**, where gene losses and methylation are mapped onto a gene tree, and the individuals sampled from each species show variation in these patterns as a result of their genealogical relationships. Modeling this variation to elucidate patterns of post-WGD evolution will be critical, especially as multi-species comparisons continue to include increasing numbers of individuals.

Auto- vs. Allopolyploid Formation

The occurrence of subgenome dominance is generally regarded as a phenomenon affecting allopolyploids. However, the level of divergence between populations or individuals forming a polyploid might be such that subgenome dominance can happen in autopolyploids formed from divergent populations within the same species. For example, an ecologically isolated population could be subjected to more epigenetic modifications due to novel environmental stresses. And though it may be genetically similar to surrounding conspecific populations, the formation of a polyploid through interbreeding

could lead to different patterns of expression in the contributing genomes, leading one parental genome to dominate the other. The random inheritance of structural modifications could also lead to these types of patterns, making the study of polyploid formation more difficult. In general, when considering the level of divergence among interbreeding populations that form auto- and allopolyploids, it is likely that it falls on a speciation continuum (*sensu* de Quieroz 2007), complicating the rigid demarcation of polyploids into the categories of ‘auto’ and ‘allo.’ This ambiguity points to the importance of considering the role of segmental allopolyploids (Stebbins 1951; Xiong et al. 2011), as well how taxonomic units are delimited.

ACKNOWLEDGMENTS

We thank Jim Birchler for engaging conversation and Pat Edger, Barbara Mable, and Aoife McLysaght for thoughtful comments that greatly improved this paper. This work was supported by the National Science Foundation award #1339156.

LITERATURE CITED

- Almal SH, Padh H. 2012. Implications of gene copy-number variation in health and diseases. *J. Hum. Genet.* 57(1):6-13.
- Arabidopsis Interactome Mapping Consortium. 2011. Evidence for network evolution in an Arabidopsis interactome map. *Science* 333:601–607.
- Awise JC, Robinson TJ. 2008. Hemiplasy: a new term in the lexicon of phylogenetics. *Syst. Biol.* 57:503–507.
- Baldrich P, Beric A, Meyers B. 2018. Despacito: The slow evolutionary changes in plant microRNA. *Curr. Opin. Plant Biol.* 42:16-22.
- Barker MS, Husband BC, Pires JC. 2016. Spreading Winge and flying high: the evolutionary importance of polyploidy after a century of study. *Am. J. Bot.* 103(7):1139–45.
- Bastide P, Solís-Lemus C, Kriebel R, Sparks KW, Ané C. 2017. Phylogenetic comparative methods on phylogenetic networks with reticulations. *Syst. Biol.* (early view) <https://doi.org/10.1093/sysbio/syy033>.
- Beissinger TM, Wang L, Crosby K, Durvasula A, Hufford MB, Ross-Ibarra J. 2016. Recent demography drives changes in linked selection across the maize genome. *Nat. Plants* 2:16084.
- Bekaert M, Edger PP, Pires JC, Conant GC. 2011. Two-phase resolution of polyploidy in the *Arabidopsis* metabolic network gives rise to relative followed by absolute dosage constraints. *Plant Cell* 23(5):1719-28.

- Birchler JA, Riddle NC, Auger DL, Veitia RA. 2005. Dosage balance in gene regulation: biological implications. *Trends Genet.* 21:219–226.
- Birchler JA, Veitia RA. 2007. The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell* 19(2):395-402.
- Birchler JA, Veitia RA. 2012. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl. Acad. Sci USA* 109(37):14746-53.
- Birchler JA, Johnson AF, Veitia RA. 2016. Kinetics genetics: Incorporating the concept of genomic balance into an understanding of quantitative traits. *Plant Sci.* 245:128-34.
- Blischak PD, Kubatko LS, Wolfe AD. 2016. Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids. *Mol. Ecol. Resour.* 16(3):742–54.
- Blischak PD, Kubatko LS, Wolfe AD. 2017. SNP genotyping and parameter estimation in polyploids using low-coverage sequence data. *Bioinformatics* 34:407–415.
- Bowman JL, Kohchi T, Yamamoto KT, Jenkins J, Shu S, et al. 2017. Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell* 171(2):287-304.
- Bradburd GS, Ralph PL, Coop GM. 2016. A spatial framework for understanding population structure and admixture. *PLoS Genet.* 12(1):e1005703.
- Campbell MA, Ganley AR, Gabaldón T, Cox MP. 2016. The case of the missing ancient fungal polyploids. *Am. Nat.* 188(6):602–14.

- Chae L, Lee I, Shin J, Rhee SY. 2012. Towards understanding how molecular networks evolve in plants. *Curr. Opin. Plant Biol.* 15:177–184.
- Cheng F, Wu J, Fang L, Sun S, Liu B, et al. 2012. Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. *PloS one* 7(5), e36442.
- Cheng F, Sun C, Wu J, Schnable J, Woodhouse MR, et al. 2016. Epigenetic regulation of subgenome dominance following a whole genome triplication in *Brassica rapa*. *New Phytol.* 211(1):288-99.
- Cheng JY, Mailund T, Nielsen R. 2017. Fast admixture analysis and population tree estimation for SNP and NGS data. *Bioinformatics* 33(14):2148–55.
- Cheng F, Wu J, Cai X, Liang J, Freeling M, Wang X. 2018. Gene retention, fractionation and subgenome differences in polyploid plants. *Nat. Plants* 4:258–268.
- Chifman J, Kubatko LS. 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30(23):3317–3324.
- Coate JE, Song MJ, Bombarely A, Doyle JJ. 2016. Expression-level support for gene dosage sensitivity in three *Glycine* subgenus *Glycine* polyploids and their diploid progenitors. *New Phytol.* 212 (3):1083-1093.
- Conant GC, Wolfe KH. 2006. Functional partitioning of yeast co-expression networks after genome duplication. *PLoS Genet.* 4:e109.
- Conant GC, Wolfe KH. 2007. Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Mol. Syst. Biol.* 3:129.

- Conant GC, Wolfe KH. 2008a. Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast. *Genetics* 179(3):1681-92.
- Conant GC, Wolfe KH. 2008b. Turning a hobby into a job: how duplicated genes find new functions. *Nat. Rev. Genet.* 9(12):938.
- Conant GC. 2014. Comparative genomics as a time machine: How relative gene dosage and metabolic requirements shaped the time-dependent resolution of yeast polyploidy. *Mol. Biol. Evol.* 31(12):3184-93.
- Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: Clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr. Opin. Plant Biol.* 19:91-98.
- Cornille A, Salcedo A, Kryvokhyzha D, Glémin S, Holm K, et al. 2016. Genomic signature of successful colonization of Eurasia by the allopolyploid shepherd's purse (*Capsella bursa-pastoris*). *Mol. Ecol.* 25(2):616–29.
- de Quieroz K. 2007. Species concepts and species delimitation. *Syst. Biol.* 56:879–886.
- Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference, and the multispecies coalescent. *Trends Ecol. Evol.* 24:332–340.
- DeSmet R, Sabaghian E, Li Z, Saeys Y, Van De Peer. 2017. Coordinated functional divergence of genes after genome duplication in *Arabidopsis thaliana*. *Plant Cell* 29(11):2786-800.
- Ding M, Chen ZJ. 2018. Epigenetic perspective on the evolution and domestication of polyploid plants and crops. *Curr. Opin. Plant Biol.* 42:37-48.

- Dodsworth S, Chase MW, Leitch AR. 2015. Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Bot. J. Linn. Soc.* 180(1):1-5.
- Dubin MJ, Scheid OM, Becker C. 2018. Transposons: a blessing curse. *Curr. Opin. Plant Biol.* 42:23-29.
- Dufresne F, Stift M, Vergilino R, Mable BK. 2014. Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Mol. Ecol.* 23:40–69.
- Dunn CW, Zapata F, Munro C, Siebert S, Hejnal A. 2018. Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proc. Natl. Acad. Sci. USA* early online: January 4, 2018, doi: 10.1073/pnas.1707515115.
- Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, et al. 2015. The butterfly plant arms-race escalated by gene and genome duplications. *Proc. Natl. Acad. Sci. USA* 112(27):8362-66.
- Edger PP, Smith R, McKain MR, Cooley AM, Vallejo-Marin M, et al. 2017. Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. *Plant Cell* 29(9):2150-67.
- Edger PP, McKain MR, Bird KA, VanBuren R. 2018. Subgenome assignment in allopolyploids: challenges and future directions. *Curr. Opin. Plant Biol.* 42:76–80.
- Emery M, Willis MMS, Hao Y, Barry K, Oakgrove K, et al. 2018. Preferential retention of genes from one parental genome after polyploidy illustrates the nature and scope of the genomic conflicts induced by hybridization. *PLoS Genet.* 14:e1007267.

- Excoffier L, Dupanloup I, Huerta-Sánchez E, Susa VC, Foll M. 2013. Robust demographic inference from genomic and SNP data. *PLoS Genet.* 9(10):e1003905.
- Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* 16(7):805–814.
- Freeling M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.* 60:433-53.
- Freeling M, Scanlon MJ, Fowler JE. 2015. Fractionation and subfunctionalization following genome duplications: mechanisms that drive gene content and their consequences. *Curr. Opin. Genet. Dev.* 35:110-18.
- Gaeta RT, Pires JC. 2010. Homoeologous recombination in allopolyploids: the polyploid ratchet. *New Phytol.* 186(1), 18-28.
- Gallagher JP, Grover CE, Hu G, Wendel JF. 2016. Insights into the ecology and evolution of polyploid plants through network analysis. *Mol. Ecol.* 25(11): 2644-60.
- Gardiner L-J, Joynson R, Omony J, Rusholme-Pilcher R, Olohan L, et al. 2017. Hidden variation in polyploid wheat drives local adaptation. *bioRxiv* 217828. <https://doi.org/10.1101/217828>.
- Gerstein AC, Lim H, Berman J, Hickman MA. 2017. Ploidy tug-of-war: Evolutionary and genetic environments influence the rate of ploidy drive in a human fungal pathogen. *Evolution* 71(4):1025-38.

- Golicz AA, Bayer PE, Barker GC, Edger PP, Kim HR, et al. 2016. The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat. Commun.* 7:13390.
- Gordon SP, Contreras-Moreira B, Woods DP, Des Marais DL, Burgess D, et al. 2017. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun.* 8(1):2184.
- Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5(10):e1000695.
- Hahn MW, Nakhleh L. 2016. Irrational exuberance for resolved species trees. *Evolution* 70:7–17.
- He Z, Wang L, Harper AL, Havlickova L, Pradhan AK, Parkin IAP, Bancroft I. 2017. Extensive homoeologous genome exchanges in allopolyploid crops revealed by mRNAseq-based visualization. *Plant Biotechnol. J.* 15(5):594-604.
- Hejase HA, Pol NV, Bonito GM, Edger PP, Liu KJ. 2017. Coal-Miner: a statistical method for GWA studies of quantitative traits with complex evolutionary origins. *Proc. ACM Inter. Conf. Bioinform., Comp. Biol., Health Inform.* (ACM-BCB '17), 8th, Boston, Massachusetts, USA, August 20–23. ACM, New York, NY, USA, 107-114.
- Hermansen RA, Hvidsten TR, Sandve SR, Liberles DA. 2016. Extracting functional trends from whole genome duplication events using comparative genomics. *Biol. Proceed. Online* 18:11.

- Holland PWH. 1999. Gene duplication: Past, present and future. *Sem. Cell Dev. Biol.* 10:541-547.
- Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.* 19:1419-1428.
- Homouz D, Kudlicki AS. 2013. The 3D organization of the yeast genome correlates with co-expression and reflects functional relations between genes. *PLoS ONE* 8:e54699.
- Horvath R, Slotte T. 2017. The role of small RNA-based epigenetic silencing for purifying selection on transposable elements in *Capsella grandiflora*. *Genome Biol. Evol.* 9(10):2911-20.
- Hughes TE, Langdale JA, Kelly S. 2014. The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. *Genome Res.* 24(8):1348-55.
- Hurgobin B, Golicz AA, Bayer PE, Chan CK, Tirnaz S, et al. 2017. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnol. J.* (in press) doi: 10.1111/pbi.12867.
- Jiao W, Yuan J, Jiang S, Liu Y, Wang L, et al. 2018. Asymmetrical changes of gene expression, small RNAs and chromatin in two resynthesized wheat allotetraploids. *Plant J.* 93(5), 828-842.

- Jighly A, Lin Z, Forster JW, Spangenberg GC, Hayes BJ, Daetwyler HD. 2018. Insights into population genetics and evolution of polyploids and their ancestors. *Mol. Ecol. Resour.* (early view): <https://doi.org/10.1111/1755-0998.12896>.
- Jones GR. 2017. Bayesian phylogenetic analysis for diploid and allotetraploid species networks. *bioRxiv* 129361. <http://dx.doi.org/10.1101/129361>
- Kamneva OK, Rosenberg NA. 2017. Simulation-based evaluation of hybridization network reconstruction methods in the presence of incomplete lineage sorting. *Evol. Bioinform. Online* 13:1176934317691935.
- Kamneva OK, Syring J, Liston A, Rosenberg NA. 2017. Evaluating allopolyploid origins in strawberries (*Fragaria*) using haplotypes generated from target capture sequencing. *BMC Evol. Biol.* 17(1):180.
- Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, Wirkner U, Ansorge W, Paabo S. 2004. A neutral model of transcriptome evolution. *PLoS Biol.* 2(5):E132.
- Khaitovich P, Paabo S, Weiss G. 2005. Toward a neutral evolutionary model of gene expression. *Genetics* 170(2):929-39.
- Kingman JCF. 1982. On the genealogy of large populations. *J. Appl. Stat.* 19:27–43.
- Kondrashov FA, Kondrashov AS. 2006. Role of selection in fixation of gene duplications. *J. Theor. Biol.* 239(2):141-51.
- Kryvokhyzha D, Holm K, Chen J, Cornille A, Glémin S, et al. 2016. The influence of population structure on gene expression and flowering time variation in the

- ubiquitous weed *Capsella bursa-pastoris* (Brassicaceae). *Mol. Ecol.* 25(5):1106–21.
- Kuang MC, Hutchins PD, Russell JD, Coon JJ, Hittinger CT. 2016. Ongoing resolution of duplicate gene functions shapes the diversification of a metabolic network. *eLife* 5:e19027.
- Landis JB, Soltis DE, Li Z, Marx HE, Barker MS, et al. 2018. Impact of whole-genome duplication events on diversification rates in angiosperms. *Am. J. Bot.* (in press): <https://doi.org/10.1002/ajb2.1060>.
- Lashermes P, Hueber Y, Combes MC, Severac D, Dereeper A. 2016. Inter-genomic DNA exchanges and homeologous gene silencing shaped the nascent allopolyploid coffee genome (*Coffea arabica* L.). *G3* (Bethesda) 6(9):2937–48.
- Lee KM, Coop G. 2017. Distinguishing among modes of convergent adaptation using population genomic data. *Genetics* 207(4):1591–1619.
- Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, et al. 2016. The Atlantic salmon genome provides insights into rediploidization. *Nature* 533(7602):200–05.
- Limborg MT, Larson WA, Seeb LW, Seeb JE. 2017. Screening of duplicates loci reveals hidden divergence patterns in a complex salmonid genome. *Mol. Ecol.* 26(17):4509–22.
- Li, W-H. 1980. Rate of gene silencing at duplicate loci: A theoretical study and interpretation of data from tetraploid fish. *Genetics* 95:237–258.

- Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. 2016. Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* 28(2):326–44.
- Li Z, Tiley GP, Galuska SR, Reardon CR, Kidder TI, Rundell RJ, Barker MS. 2018. Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc. Natl. Acad. Sci.* (ahead of print) 115(18):4713-4718.
- Lopez-Nieves S, Yang Y, Timoneda T, Wang M, Feng T, et al. 2018. Relaxation of tyrosine pathway regulation underlies the evolution of betalain pigmentation in Caryophyllales. *New Phytol.* 217(2):896-908.
- Lynch M, Marinov GK. 2015. The bioenergetic costs of a gene. *Proc. Natl. Acad. Sci USA* 112(51):15690-95.
- Mable BK. 2001. Ploidy evolution in the yeast *Saccharomyces cerevisiae*: a test of the nutrient limitation hypothesis. *J. Evol. Biol.* 14:157–170.
- Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, et al. 2005. Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. USA* 102(15), 5454-5459.
- Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc. Natl. Acad. Sci. USA* 107:9270–9274.
- Mandakova T, Lysak M. 2018. Post-polyploid diploidization and diversification through dysploid changes. *Curr. Opin. Plant Biol.* 42:55-65.
- Marad DA, Burskirk SW, Lang GI. 2018. Altered access to beneficial mutations slows adaptation and biases fixed mutations in diploids. *Nat. Evol. Evol.* 2:882-889.

- Marcet-Houben M, Gabaldón T. 2015. Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biol.* 13: e1002220.
- Mayfield-Jones D, Washburn JD, Arias T, Edger PP, Pires JC, Conant GC. 2013. Watching the grin fade: Tracing the effects of polyploidy on different evolutionary time scales. *Sem. Cell Dev. Biol.* 24(4):320-31.
- Meirmans PG, Liu S, van Tienderen PH. 2018. The Analysis of Polyploid Genetic Data. *J. Hered.* 109(3):283-296.
- Mendes FK, Hahn MW. 2016. Gene tree discordance causes apparent substitution rate variation. *Syst. Biol.* 65:711–721.
- Mirarab S, Warnow T. 2015. ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31(12):i44–i52.
- Moriyama Y, Koshiba-Takeuchi K. 2018. Significance of whole-genome duplications on the emergence of evolutionary novelties. *Brief. Funct. Genomics* (epub ahead of print): doi 10.1093/bfpg/ely007.
- Nakatani Y, McLysaght A. 2017. Genomes as documents of evolutionary history: a probabilistic macrosynteny model for the reconstruction of ancestral genomes. *Bioinformatics* 33(14):i369-78.
- Nei M, Roychoudhury AK. 1973. Probability of fixation of nonfunctional genes at duplicate loci. *Am. Nat.* 107:362–372.

- Nguepjob JR, Tossim H-A, Bell JM, Rami J-F, Sharma S, et al. 2016. Evidence of genomic exchanges between homeologous chromosomes in a cross of peanut with newly synthesized allotetraploid hybrid. *Front. Plant Sci.* 7:1635.
- Novikova, Hohmann N, Van de Peer y. 2018. Polyploid *Arabidopsis* species originating around recent glaciation maxima. *Curr. Opin. Plant Biol.* 42:8-15.
- Ohno S. 1970. *Evolution by Gene Duplication*. Springer, New York.
- One Thousand Plant Transcriptome Initiative. 2018. A phylogenomic view of evolutionary complexity in green plants. In review.
- Osborn TC, Buttrill DV, Sharp AG, Pickering KJ, Parkin IAP, et al. 2003. Detection and effects of a homeologous reciprocal transposition in *Brassica napus*. *Genetics* 165:1569-1577
- O'Toole ÁN, Hurst LD, McLysaght A. 2018. Faster evolving primate genes are more likely to duplicate. *Mol. Biol. Evol.* 35(1):107–18.
- Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424(6945):194-7.
- Pasquier J, Braasch I, Batzel P, Cabau C, Montfor J, et al. 2017. Evolution of gene expression after whole genome duplication: new insights from the spotted gar genome. *J. Exp. Zool. B Mol. Dev. Evol.* 328(7):709-21.
- Pennell MW, Harmon LJ. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Ann. N. Y. Acad. Sci.* 1289:90-105.

- Pires JC, Zhao J, Schranz ME, Leon EJ, Quijada PA, et al. 2004. Flowering time divergence and genomic rearrangements in resynthesized *Brassica* polyploids (Brassicaceae). *Biol. J. Linnean Soc.* 82:675-688.
- Pires JC, Conant GC. 2016. Robust yet fragile: Expression noise, protein misfolding, and gene dosage in the evolution of genomes. *Annu. Rev. Genet.* 50:113-31.
- Powell JJ, Fitzgerald TL, Stiller J, Berkman PJ, Gardiner DM, Manners JM, et al. 2017. The defence-associated transcriptome of hexaploid wheat displays homoeolog expression and induction bias. *Plant Biotechnol. J.* 15(4):533-543.
- Qi X, An H, Ragsdale AP, Hall TE, Gutenkunst RN, et al. 2017. Genomic inferences of domestication events are corroborated by written records in *Brassica rapa*. *Mol. Ecol.* 26(13):3373–88.
- Racimo F, Berg JJ, Pickrell JK. 2017. Detecting polygenic adaptation in admixture graphs. *Genetics* (early online) <https://doi.org/10.1534/genetics.117.300489>.
- Ren L, Cui J, Wang J, Tan H, Li W, Tang C, et al. 2017. Analyzing homoeolog expression provides insights into the rediploidization event in gynogenetic hybrids of *Carassius auratus* red var. $\times$ *Cyprinus carpio*. *Sci. Rep.* 7(1):13679.
- Renny-Byfield S, Rodgers-Melnick E, Ross-Ibarra J. 2017. Gene fractionation and function in the ancient subgenomes of maize. *Mol. Biol. Evol.* 34(8):1825–32.
- Rice A, McLysaght A. 2017a. Dosage sensitivity is a major determinant of human copy number variant pathogenicity. *Nat. Commun.* 8:14366.
- Rice A, McLysaght A. 2017b. Dosage-sensitive genes in evolution and disease. *BMC Biol.* 15:78.

- Robertson FM, Gundappa MK, Grammes F, Hvidsten TR, Redmond AK, et al. 2017. Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biol.* 18:111.
- Rodriguez F, Arkhipova IR. 2018. Transposable elements and polyploid evolution in animals. *Curr. Opin. Genet. Dev.* 49:115-123.
- Rohlf RV, Nielsen R. 2015. Phylogenetic ANOVA: The expression variance and evolution model for quantitative trait evolution. *Syst. Biol.* 64(5):695–708.
- Roman H, Phillips MM, Sands SM. 1955. Studies of polyploid *Saccharomyces*. I. Tetraploid segregation. *Genetics* 40:546–561.
- Ruprecht C, Vaid N, Proost S, Persson S, Mutwil M. 2017a. Beyond genomics: studying evolution with gene coexpression networks. *Trends Plant Sci.* 22:298–307.
- Ruprecht C, Proost S, Hernandez-Coronado M, Ortiz-Ramirez C, Lang D, et al. 2017b. Phylogenomic analysis of gene co-expression networks reveals the evolution of functional modules. *Plant J.* 90(3):447-65.
- Sankoff D, Zheng C, Wong B. 2012. A model for biased fractionation after whole genome duplication. *BMC Genomics* 13(Suppl 1):S8.
- Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440:341–345.
- Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KP. 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc. Natl. Acad. Sci. USA* 104:8397-8402.

- Schnable JC, Springer NM, Freeling M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. USA* 108(10), 4069-4074.
- Schrider DR, Kern AD. 2018. Machine learning for population genetics: a new paradigm. *Trends Genet.* (early online) <https://doi.org/10.1016/j.tig.2017.12.005>.
- Scienski K, Fay JC, Conant GC. 2015. Patterns of gene conversion in duplicated yeast histones suggest strong selection on a coadapted macromolecular complex. *Genome Biol. Evol.* 7(12):3249-58.
- Sémon M, Wolfe KH. 2007. Rearrangement rate following the whole-genome duplication in teleosts. *Mol. Biol. Evol.* 24:860–867.
- Sheehan S, Song YS. 2016. Deep learning for population genetic inference. *PLoS Comput. Biol.* 12(3):e1004845.
- Smith SA, Brown JW, Yang Y, Bruenn R, Drummond CP, Brockington SF, Walker JF, Last N, Douglas NA, Moore JM. 2017. Disparity, diversity, and duplications in the Caryophyllales. *New Phytol.* 217(2):836-854.
- Solis-Lemus C, Ané C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet.* 12(3):e1005896.
- Solis-Lemus C, Bastide P, Ané C. 2017. PhyloNetworks: a package for phylogenetic networks. *Mol. Biol. Evol.* 34(12):3292–98.
- Soltis PS, Soltis DE. 2016. Ancient WGD events as drivers of key innovations in angiosperms. *Curr. Opin. Plant Biol.* 30:159–65.

- Soltis DE, Visger CJ, Marchant DB, Soltis PS. 2016. Polyploidy: Pitfalls and paths to a paradigm. *Am. J. Bot.* 103(7):1146-66.
- Stebbins GL. 1951. *Variation and Evolution in Plants*. Columbia University Press.
- Stein A, Coriton O, Rousseau-Geutin M, Samans B, Schiessl SV, et al. 2017. Mapping of homoeologous chromosome exchanges influencing quantitative trait variation in *Brassica napus*. *Plant Biotechnol. J.* 15(11):1478-89.
- Sun Y, Wu Y, Yang C, Sun S, Lin X, Liu L, et al. 2017. Segmental allotetraploidy generates extensive homoeologous expression rewiring and phenotypic diversity at the population level in rice. *Mol. Ecol.* 26(20):5451–66.
- Szöllősi GJ, Tannier E, Daubin V, Boussau B. 2015. The inference of gene trees with species trees. *Syst. Biol.* 64(1):42–62.
- Tasdighian S, Van Bel M, Li Z, Van de Peer Y, Carretero-Paulet L, Maere S. 2017. Reciprocally retained genes in the angiosperm lineage show the hallmarks of dosage balance sensitivity. *Plant Cell* 29(11):2766-2785.
- Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial “pan-genome”. *Proc. Natl. Acad. Sci. USA* 102:13950–13955.
- Thévenin A, Eib-Dor E, Ozery-Flato M, Shamir R. 2014. Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome. *Nucleic Acids Res.* 42:9854–9861.
- Thomas GWC, Ather SH, Hahn MW. 2017. Gene-tree reconciliation with MUL-trees to resolve polyploidy events. *Syst. Biol.* 66(6):1007–1018.

- Thompson A, Zakon HH, Kirkpatrick M. 2016. Compensatory drift and the evolutionary dynamics of dosage-sensitive duplicate genes. *Genetics* 202:765–774.
- Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. *Nat. Rev. Genet.* 18(7):411-424.
- van Hoek MJ, Hogeweg P. 2009. Metabolic adaptation after whole genome duplication. *Mol. Biol. Evol.* 26(11):2441-2453.
- van Iersel L, Jones M, Scornavacca C. 2017. Improved maximum parsimony methods for phylogenetic networks. *Syst. Biol.* 67:518–542.
- Vicent CM, Casacuberta JM. 2017. Impact of transposable elements on polyploid plant genomes. *Ann. Bot.* 120(2):195-207.
- Wagner A. 2007. Energy costs constrain the evolution of gene expression. *J. Exp. Zool. B Mol. Dev. Evol.* 308:322-324.
- Wang L, Beissinger TM, Lorant A, Ross-Ibarra C, Ross-Ibarra J, Hufford MB. 2017. The interplay of demography and selection during maize domestication and expansion. *Genome Biol.* 18(1):215.
- Washburn JD, Bird KA, Conant GC, Pires JC. 2016. Convergent evolution and the origin of complex phenotypes in the age of systems biology. *Int. J. Plant Sci.* 177:305-318.
- Wen D, Nakhleh L. 2017. Co-estimating reticulate phylogenies and gene trees from multi-locus sequence data. *Syst. Biol.* 67:439–457.
- Wen D, Yu Y, Zhu J, Nakhleh L. 2017. Inferring phylogenetic networks using PhyloNet. *Syst. Biol.* (early online) <https://doi.org/10.1093/sysbio/syy015>.

- Wendel JF, Lisch D, Hu G, Mason AS. 2018. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Curr. Opin. Genet. Dev.* 49:1–7.
- Wisecaver JH, Borowsky AT, Tzin V, Jander G, Kliebenstein DJ, Rokas A. 2017. A global co-expression network approach for connecting genes to specialized metabolic pathways in plants. *Plant Cell* 29(5):944-959.
- Xie T, Yang QY, Wang XT, McLysaght A, Zhang HY. 2016. Spatial colocalization of human ohnolog pairs acts to maintain dosage-balance. *Mol. Biol. Evol.* 33(9):2368-2375.
- Xiong Z, Gaeta RT, Pires JC. 2011. Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*. *Proc. Natl. Acad. Sci. USA* 108(19):7908-7913.
- Yanai I, Korbel JO, Boue S, McWeeney SK, Bork P, Lercher MJ. 2006. Similar gene expression profiles do not imply similar tissue functions. *Trends Genet.* 22(3):132-138.
- Yang J, Liu D, Wang X, Ji C, Cheng F, et al. 2016. The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. *Nature Genet.* 48(10):1225.
- Zhan SH, Drori M, Goldberg EE, Otto SP, Mayrose I. 2016. Phylogenetic evidence for cladogenetic polyploidization in land plants. *Am. J. Bot.* 103(7):1252-1258.

- Zhao M, Zhang B, Lisch D, Ma J. 2017. Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. *Plant Cell* (early online): <https://doi.org/10.1105/tpc.17.00595>.
- Zhu Y, Lin Z, Nakhleh L. 2013. Evolution after whole-genome duplication: a network perspective. *G3: Genes|Genom.|Genet.* 3:2049–2057.
- Zohren J, Wang N, Kardailsky I, Borrell JS, Joecker A, et al. 2016. Unidirectional diploid-tetraploid introgression among British birch trees with shifting ranges shown by restriction site-associated markers. *Mol. Ecol.* 25(11):2413–2426.

CHAPTER 2 : PHYLOGENY AND MULTIPLE INDEPENDENT WHOLE-GENOME DUPLICATION EVENTS IN THE BRASSICALES

Makenzie E. Mabry¹, Julia M. Brose¹, Paul D. Blischak², Brittany Sutherland², Wade T. Dismukes¹, Christopher A. Bottoms³, Patrick P. Edger⁴, Jacob D. Washburn⁵, Hong An¹, Jocelyn C. Hall⁶, Michael R. McKain⁷, Ihsan Al-Shehbaz⁸, Michael S. Barker², M. Eric Schranz⁹, Gavin C. Conant¹⁰, J. Chris Pires¹

1 Division of Biological Sciences and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, U.S.A.

2 Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, U.S.A.

3 Informatics Research Core Facility and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, U.S.A.

4 Department of Horticulture, Michigan State University, East Lansing, MI, U.S.A.

5 Plant Genetics Research Unit, USDA-ARS, Columbia, MO, U.S.A.

6 Department of Biological Sciences, University of Alberta, Edmonton, Canada

7 Department of Biological Sciences, The University of Alabama, Tuscaloosa, AL, U.S.A.

8 Missouri Botanical Garden, St. Louis, MO, U.S.A.

9 Wageningen University and Research, Wageningen, The Netherlands

10 Bioinformatics Research Center, Program in Genetics and Department of Biological Sciences, North Carolina State University, Raleigh, NC, U.S.A.

Please cite the published version in the American Journal of Botany:

Mabry ME, JM Brose, PD Blischak, B Sutherland, WT Dismukes, CA Bottoms, PP Edger, JD Washburn, H An, JC Hall, MR McKain, I Al-Shehbaz, MS Barker, ME Schranz, GC Conant, & JC Pires 2020. Phylogeny and Multiple Independent Whole-Genome Duplication Events in the Brassicales. American Journal of Botany

ABSTRACT

Whole-genome duplications (WGDs) are prevalent throughout the evolutionary history of plants. For example, dozens of WGDs have been phylogenetically localized across the order Brassicales, specifically, within the family Brassicaceae. A WGD event has also been identified in the Cleomaceae, the sister family to Brassicaceae, yet its placement, as well as that of WGDs in other families in the order, remains unclear.

Phylo-transcriptomic data were generated and used to infer a nuclear phylogeny for 74 Brassicales taxa. Genome survey sequencing was also performed on 66 of those taxa to infer a chloroplast phylogeny. These phylogenies were used to assess and confirm relationships among the major families of the Brassicales and within the family Brassicaceae. Multiple WGD inference methods were then used to assess the placement of WGDs on the nuclear phylogeny.

Well-supported chloroplast and nuclear phylogenies for the Brassicales and the putative placement of the Cleomaceae specific WGD event Th- α are presented. This work also provides evidence for previously hypothesized WGDs, including a well-supported event shared by at least two members of the Resedaceae family, and a possible event within the Capparaceae.

Phylogenetics and the placement of WGDs within highly polyploid lineages continues to be a major challenge. This study adds to the conversation on WGD inference difficulties by demonstrating that sampling is especially important for WGD identification and phylogenetic placement. Given its economic importance and genomic resources, the Brassicales continues to be an ideal group for assessing WGD inference methods.

INTRODUCTION

The Brassicales is an economically important order of flowering plants, home to crop species (kale, broccoli, cabbage, cauliflower, papaya, capers, and canola) as well as several model species including *Arabidopsis*. There are currently 17 accepted families within the order (APG IV, 2016), with the Brassicaceae being the most well studied family due to its many crop and model plant species. Together with the closely related Cleomaceae and Capparaceae families, the three families contain 94% of the species diversity in the order (Edger et al., 2015). Sister to these three families is an unresolved clade of four families: Tovariaceae, Gyrostemonaceae, Resedaceae, and Pentadiplandraceae. This clade is followed by Emblingiaceae, [[Salvadoraceae + Bataceae], Koeberliniaceae], Limnanthaceae, Setchellanthaceae, [Caricaceae + Moringaceae], and [Tropaeolaceae + Akaniaceae] (**Supplemental Figure 2.1**; APG IV, 2016). The order diverged from other Rosids around 103 mya and represents 2.2% of the total extant core eudicot diversity (Magallon et al., 1999; Cardinal-McTeague et al., 2016). Previous research has identified multiple whole-genome duplication (WGD) events across the order using a variety of comparative methods, including genomics, transcriptomics, and molecular cytogenetics (Vision et al., 2000; Schranz and Mitchell-Olds, 2006; Barker et al., 2009; Cheng et al., 2013; Kagale et al., 2014; Edger et al., 2015, 2018; Mandáková et al., 2017; Lysak, 2018; One Thousand Plant Transcriptomes Initiative, 2019; **Supplemental Figure 2.1**). Four of the most studied events include one near the base of the order (At- β ; Edger et al., 2015, 2018), an event at the base of the Brassicaceae (At- α ; Vision et al., 2000; Haudry et al., 2013; Edger et al., 2015), a triplication at the base of the tribe Brassiceae in the Brassicaceae

(Lysak et al., 2005; Tang et al., 2012), and an unplaced event within the Cleomaceae (Tha; Schranz and Mitchell-Olds, 2006; Barker et al., 2009).

Within the Brassicales, the Brassicaceae has the largest number of accepted species (>4,000; BrassiBase). It contains the model plant *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000) as well as important crops of the *Brassica* and *Raphanus* groups. Its clades have been placed into three major lineages (Lineage I, Lineage II, and Lineage III; Beilstein et al., 2006), with notable named clades acknowledged more recently (i.e., Clade C; Huang et al., 2016; Nikolov et al., 2019). The relationships among these lineages and clades are unclear. Besides elucidating the relationships within the Brassicaceae, another major area of research has focused on its considerable glucosinolate diversity within the family (Kliebenstein et al., 2001; Ratzka et al., 2002; Züst et al., 2018; Blažević et al., 2019), including the impact of WGD events on the glucosinolate chemical structures (Edger et al., 2015; Barco and Clay, 2019).

Sister to the Brassicaceae is the Cleomaceae. A mostly herbaceous family of ~270 species of pantropical plants, it diverged from the Brassicaceae around 40 mya (Edger et al., 2015). The Cleomaceae displays a much wider range of floral morphologies than its sister family, a characteristic that has been the focus of several recent studies (Bhide et al., 2014; Brock, 2014; Bayat et al., 2018). This family is unique among the Brassicales for containing species that utilize C₄ photosynthesis (*Gynandropsis gynandra* and *Coalisina angustifolia*, formerly *Cleome angustifolia*; Feodorova et al., 2010) as well as, though not unique to Cleomaceae (Schlüter et al., 2016), a C₃-C₄ intermediate (*Coalisina paradoxa*, formerly *C. paradoxa*; van den Bergh et al., 2014). Cleomaceae is known to have undergone at least one independent polyploidy event that occurred after the split from the

Brassicaceae, named Th- α (after *Tarenaya hassleriana*), and has been dated to around 13.7 mya (Schranz and Mitchell-Olds, 2006; Barker et al., 2009; Cheng et al., 2013). The analyses for this identification and placement used only partial genomic fragments, ESTs, or a single genome. It was subsequently determined that the Th- α event is shared with the species *G. gynandra*, a C₄ species (van den Bergh et al., 2014), but not with *Cleome violacea*, *Arivela viscosa*, or *Polanisia trachysperma* (Emery et al., 2018; One Thousand Plant Transcriptomes Initiative, 2019). The precise phylogenetic location of Th- α remains a mystery (van den Bergh et al., 2014; Bayat et al., 2018).

The Capparaceae--a mostly woody tropical family of 450 species--is less studied than either of its two sister families, Brassicaceae and Cleomaceae. Like the Cleomaceae, the Capparaceae is very diverse in its floral morphology (Endress, 1992), and, like other members of the order (with the exception of *Koeberlinia spinosa*; Tobe and Raven 2008), it produces glucosinolates but shares the production of unique methyl-glucosinolates with only the Cleomaceae (Hall et al., 2002; Mithen et al., 2010). In this group is the economically important species *Capparis spinosa*, or capers. Recent work using chromosome counts hypothesized that the Capparaceae and a more distant family, the Resedaceae, may possess unique WGD events (Lysak, 2018). Members of the Resedaceae, a relatively small clade of ~85 species, are mostly distributed across Europe, the Middle East, and Africa with one taxon occurring in North America (*Oligomeris linifolia*) due to a long-distance dispersal event (Martín-Bravo et al., 2007; 2009; Cardinal-McTeague et al., 2016).

To infer phylogenetic relationships within the Brassicales, we use phylo-transcriptomics, a quickly evolving subdiscipline of phylogenomics that uses RNA-seq

data as the basis of its inferences (Dunn et al., 2008; McKain et al., 2012; Yang et al., 2015; Washburn et al., 2017; Unruh et al., 2018; Godden et al., 2019). Transcriptomics gives access to many more nuclear genes than does traditional PCR-based approaches and is less expensive than sequencing an entire genome. RNA-seq data also allows for assessing gene and genome duplication events (Baker et al., 2009; McKain et al., 2012). One difficulty with using transcriptomes for phylogenetic inference is an inability to determine orthology (Dunn et al., 2013; Yang and Smith 2014; Washburn et al., 2017; Emms and Kelly, 2019). Several methods have been developed to address this problem, including those that aim to identify orthogroups, or sets of genes descended from a single gene in the last common ancestor of the group or species of interest (Duarte et al., 2010; Emms and Kelly, 2015). OrthoFinder2 (Emms and Kelly, 2019) offers improvements in both orthogroup inference accuracy and in computational speed, especially when using Diamond (an alternative to BLAST; Buchfink et al., 2015). Together, these methods have enabled phylo-transcriptomics to be extremely useful for inferring species relationships, understanding gene evolution, and elucidating WGD events.

The Brassicales are an intriguing group for the study of polyploidy. With well-established WGD events across the Brassicaceae, including At- α at the base (Vision et al., 2000; Edger et al., 2015), and the identification of a unique and more recent, albeit unplaced, event in the Cleomaceae (Th- α ; Schranz and Mitchell-Olds, 2006; Barker et al., 2009; Cheng et al., 2013) one wonders what makes this group unique? Here, we aim to answer the remaining questions on the placement of WGD events, including Th- α , using phylo-transcriptomics with a focus on sampling the Brassicaceae and Cleomaceae and additional sampling of the Capparaceae, Resedaceae, Bataceae, Caricaceae, and

Moringaceae families. We ask if Th-a is shared across the Cleomaceae or if the family, like the Brassicaceae, is characterized by multiple events. We also test the recent hypothesis that the Resedaceae and the Capparaceae possess independent WGD events (Lysak, 2018). We demonstrate that the Brassicales are a powerful resource for the study of WGD and an important group to test how WGD correlates with variation in floral morphology, photosynthesis types, metabolism (especially glucosinolates), and other traits of interest.

MATERIALS AND METHODS

Taxon sampling

Sampling of 74 taxa from 57 genera across the Brassicales spanned seven families (Brassicaceae, Cleomaceae, Capparaceae, Resedaceae, Bataceae, Moringaceae, Caricaceae), with a focus on the Brassicaceae (48 taxa) and Cleomaceae (17 taxa) (**Supplemental Table 2.1**). Seeds were grown at the University of Missouri or the University of Alberta in a sterile growth chamber environment. At maturity, but before flowering, leaf tissue was collected for both RNA and DNA extraction.

DNA and RNA isolation and sequencing

DNA was extracted from leaf tissue for 69 of the 74 taxa using a Qiagen DNeasy Plant Kit (Qiagen, Germantown, Maryland, USA). To increase yield, slight modifications to the manufacturer's protocol included increasing lysis buffer incubation time to one hour and using 25 µl of buffer to elute the final sample. TruSeq library preparation (Illumina, San Diego, California, USA) and genome survey sequencing (GSS, also known as skim

sequencing) on a NextSeq (Illumina, San Diego, California, USA) were carried out at the University of Missouri, resulting in 2 X 150 bp reads.

RNA from leaf tissue was collected and immediately flash frozen using liquid nitrogen. For 38 samples, RNA was isolated using a ThermoFisher Invitrogen PureLink RNA Mini Kit (Invitrogen, Carlsbad, California, USA) followed by TruSeq library preparation (Illumina, San Diego, California, USA) and sequencing on the NextSeq (Illumina, San Diego, California, USA), resulting in 2 X 75 bp reads (**Supplemental Table 2.2**). For 16 samples, RNA was isolated using a ThermoFisher Invitrogen PureLink RNA Mini Kit (Invitrogen, Carlsbad, California, USA) then sequenced on a HiSeq instrument (Illumina, San Diego, California, USA), resulting in 2 X 100 bp reads (**Supplemental Table 2.2**). For 17 samples, RNA was sequenced on a HiSeq instrument (Illumina, San Diego, California, USA) for 2 X 100 bp reads but using a Qiagen RNeasy Plant Kit (Qiagen, Germantown, Maryland, USA) for RNA isolation (**Supplemental Table 2.2**). Two samples were isolated using a ThermoFisher Invitrogen PureLink RNA Mini Kit (Invitrogen, Carlsbad, California, USA) and sequenced on a HiSeq (Illumina, San Diego, California, USA) for 2 X 250 bp reads (**Supplemental Table 2.2**). All sequencing and library preparation for the above samples was performed by the University of Missouri DNA Core Facility.

At the University of Alberta, the sample *Cleomella serrulata* had tissue pooled from leaves, apical meristematic tissue, and floral tissue of different developmental stages including small, medium, and large buds and open flowers from two plants. Total RNA was extracted using a RNeasy Plant Mini Kit (Qiagen, Germantown, Maryland, USA) following the manufacturer's protocol then treated with DNase I (New England Biolabs,

Ipswich, Massachusetts, USA) for 30 min at 37 °C to remove residual DNA from the total RNA. Sequencing was conducted by Plate-forme d'Analyses Génomique de l' Université Laval with the Illumina TruSeq RNASeq (Illumina, San Diego, California, USA) for library preparation and Illumina (Illumina, San Diego, California, USA) for sequencing for paired-end 2 X 100 bp reads.

Chloroplast assembly, alignment, and phylogenomics

To verify identification of taxa, an analysis was performed with two previously published chloroplast genes, *matK* and *ndhF*, for 91 taxa (Hall, 2008) plus 66 samples from this study. The two chloroplast genes were annotated and extracted from *de novo* whole chloroplast sequences (discussed below) using Geneious v8.1.9 (Kearse et al., 2012). We were unable to annotate and extract *ndhF* for the taxon *Batis maritima*. Alignment of resulting genes was performed in MAFFT v7 (Katoh et al., 2002) and cleaned with phyutility v2.7.1 (Smith and Dunn, 2008) using the parameter *-clean 0.5*. For maximum likelihood (ML) phylogenetic inference, RAxML v8 (Stamatakis, 2014) was run with a separate partition for each gene, GTRGAMMA as the model, and 1000 bootstrap replicates.

To assemble the *de novo* chloroplasts sequences from the GSS data, we used Fast-Plast v1.2.8 (McKain and Wilson, 2017). This method utilizes Trimmomatic v0.35 (Bolger et al., 2014) to clean the reads of adaptors using a Phred score of 33, Bowtie2 v2.3.4.3 (Langmead and Salzberg, 2012) to separate chloroplast reads by mapping them to a reference database of angiosperm chloroplasts, followed by both SPAdes v3.13.0 (Bankevich et al., 2012) and *afin* to assemble reads (<https://github.com/mrmckain/Fast-Plast/tree/master/afin>). For 13 samples that would not assemble with the default options,

the *--subsample* option yielded successful assemblies (**Supplemental Table 2.1**). Since we obtained only partial regions of the chloroplast genomes for *Polanisia dodecandra*, *Farsetia aegyptia*, and *Cardamine hirsuta*, these samples were excluded from the downstream analyses. Following assembly, MAFFT v7 (Kato et al., 2002) was used to align the large single copy (LSC), the small single copy (SSC), and one copy of the inverted repeat (IR). Alignments were cleaned with phyutility v2.7.1 (Smith and Dunn, 2008) using the parameter *-clean 0.5*. Maximum likelihood phylogenomic inference was performed in RAxML v8 (Stamatakis, 2014) with partitions for each region, GTRGAMMA as the model, and 1000 bootstrap replicates.

Transcriptome assembly, alignment, and phylogenomics

For transcriptome analyses, reads were trimmed with trimmomatic v0.35 (Bolger et al., 2014) using the parameters *SLIDINGWINDOW:4:5*, *LEADING:5*, *TRAILING:5*, and *MINLEN:25*, followed by assembly with Trinity v2.2 (Grabherr et al., 2011). The resulting *de novo* transcriptomes were checked for completeness in BUSCO v3 (Simão et al., 2015; Waterhouse et al., 2017) and compared to the Embryophyta database. Transcriptomes were translated to protein sequences by extracting the longest open reading frame, and coding regions were predicted using TransDecoder v3.0 (github.com/TransDecoder/TransDecoder). Finally, orthology was inferred in OrthoFinder v2.2.6 (Emms and Kelly, 2019), first using the parameter *-S diamond* (Buchfink et al., 2015) and then the parameter *-M msa -ot* for multiple sequence alignments and only trees. Using custom scripts, alignments were filtered for 80% taxon occupancy (github.com/MU-IRCF/filter_by_ortho_group) and alignment quality, allowing for only 40% gaps (github.com/MU-IRCF/filter_by_gap_fraction). To estimate gene trees using ML

inference, RAxML v8 was used (Stamatakis, 2014) under the PROTCATWAG model and 100 bootstrap replicates. Gene trees were analyzed in PhyloTreePruner v1.0 (Kocot et al., 2013) to remove any paralogous genes by using a cutoff of 10 for the minimum number of taxa required to keep a group. Resulting alignments were then used to estimate final gene trees with RAxML v8 (Stamatakis, 2014) under the PROTCATWAG model and 100 bootstrap replicates. Species tree estimation for the Brassicales was performed with ASTRAL-III v.5.6.1 (Zhang et al., 2018) and included the parameter *-t 2* to assess discordance among gene trees. Species tree analyses were also performed at the family level (Brassicaceae, Cleomaceae, Capparaceae, [Resedaceae + Bataceae + Moringaceae + Caricaceae]; **Supplemental Table 2.3**).

Whole-genome duplication

To estimate the phylogenetic placement of whole-genome duplications, we used PUG v2.1 (github.com/mrmckain/PUG) to query putative paralogs over multiple gene trees with the estimated ASTRAL-III tree as the input species tree. For each analysis, we used the original ML gene trees before running them through PhyloTreePruner (i.e., gene trees with all duplicates retained), the ASTRAL-III tree (rooted and with node labels removed), and parameters *--estimate_paralogs* and *--outgroups Carica_papaya, Moringa_oleifera* as input. Output duplicate gene counts were used for nodes with bootstrap values of 80% or better.

As another confirmation of duplication events, we constructed histograms giving the distribution of the synonymous divergence (*K_s*) between paralogs in each transcriptome. This method allows for the potential identification of peaks in the distribution that may be indicative of a WGD event. The position of the peak along the *K_s*

axis provides an estimate of when the event occurred. Typically, the peak closest to time zero (or $K_s \sim 0$) corresponds to recent tandem duplicates, not relevant to WGD events. Plots of K_s distributions were made for all taxa in FASTKs v1.1 (github.com/mrmckain/FASTKs), as described in McKain et al., (2016), and in DupPipe, following Barker et al. (2010). R v3.5.1 (R Core Team, 2018) was used to estimate normal mixture models for K_s values using mclust v.5.0.2 (Fraley and Raftery, 2002; Fraley et al., 2012). To assess for the best number of peaks to explain the data, we tested one to four components for each mixture model. We chose the component with the lowest Bayesian information criterion (BIC) score as the best fit (**Supplemental Table 2.4**).

Ortholog divergence was estimated using OrthoPipe, as described in Barker et al. (2010). Using the estimated ortholog divergence and the DupPipe K_s estimates, we bookended the positions of potential events by comparing when species diverged to the mean paralog divergence of an estimated WGD event: if the ortholog divergence between pairs of species is greater (larger K_s value) than the paralog divergence of a WGD event, then the species do not share the event; if the ortholog divergence between species is less than the WGD paralog divergence, then the species share the proposed event.

RESULTS

Sequence matrices

DNA read pools range in size from 6,637,717 to 13,335,392 reads per sample. After assembly of complete chloroplasts, the inferred genomes for the 66 taxa range in length from 137,110 to 160,272 bp. The LSC, SSC, and IR regions were isolated and aligned separately, with total alignment lengths of 84,350 bp, 17,931 bp, and 26,500 bp,

respectively. The analysis of two previously published copies of *matK* and *ndhF*, in combination with our own data, resulted in alignment lengths of 1,521 and 985 bp for each gene, respectively. Both chloroplast analyses had 100% occupancy for taxa included.

RNA read pools range in size from 5,555,024 to 59,723,745 reads per sample, with an average of 22,520,865 reads per sample. To check completeness of transcriptomes, the assemblies were run through BUSCO v3 (Simão et al., 2015; Waterhouse et al., 2017). All assemblies had greater than 66% complete genes with less than 12% of genes missing or fragmented (**Supplemental Figure 2.2**). OrthoFinder v.2.2.6 (Emms and Kelly, 2019) recovered 47,600 orthogroups across the Brassicales. Filtering for 80% taxon occupancy (59/74 taxa) yielded 10,968 orthogroups. After filtering for alignment quality by allowing for only 40% gaps, we recovered 2,663 orthogroups. Pruning trees for any remaining paralogs, by using a minimum of 10 taxa as a cutoff, resulted in 1,284 orthogroups, which were then used for species tree inference. Following the steps above for each family (Brassicaceae, Capparaceae, Cleomaceae, and Resedaceae + Bataceae + Moringaceae + Caricaceae), we recovered 2,100, 10,214, 3,626, and 8,476 orthogroups, respectively (**Supplemental Table 2.3**).

Chloroplast phylogenomics of the Brassicales

The analysis of the chloroplast genes *matK* and *ndhF*, the 91 taxa from the study by Hall (2008), and our 66 samples identified some inconsistencies of species placement but recovered the same overall relationships as published for other chloroplast phylogenies of the Brassicales (Hall, 2008; Cardinal-McTeague et al., 2016; Edger et al., 2018; **Supplemental Figure 2.3**). Species sampled in both Hall (2008) and in the present study that are not recovered as sister to one another include *Stanleya pinnata*, *Cleomella lutea*,

Andinocleome pilosa, and *Capparis tomentosa*. The lack of congruence for species placement may be due to species being mislabeled (e.g., *Cleomella lutea*) or species being more genetically diverse than previously thought. Due to this uncertainty in taxon identification, we refer to these samples as Brassicaceae sp., *Polanisia* sp., Cleomaceae sp., and Capparaceae sp., respectively.

For the whole chloroplast analyses, using just one copy of the IR, all nodes except four were recovered with 70% bootstrap support or better with a topology largely congruent with previous studies (Hall, 2008; Cardinal-McTeague et al., 2016; Edger et al., 2018). This agreement includes a clade of *Moringa oleifera* and *Carica papaya* sister to a clade of [Bataceae + Resedaceae + Capparaceae + Cleomaceae + Brassicaceae], followed by Bataceae sister to [Resedaceae + Capparaceae + Cleomaceae + Brassicaceae], Resedaceae sister to [Capparaceae + Cleomaceae + Brassicaceae], and Capparaceae sister to [Cleomaceae + Brassicaceae] (**Supplemental Figure 2.4**). The relationships among the major lineages within the Brassicaceae were also in agreement with previous studies (Guo et al., 2017). We recovered *Aethionema arabicum* as sister to the rest of the family, followed by Lineage I sister to [Lineage III + Clade C + Lineage II and Expanded Lineage II] and Lineage III sister to [Clade C + Lineage II and expanded Lineage II]. Relationships within the Cleomaceae were congruent with previous studies (Hall, 2008; Patchell et al., 2014), with *Cleome* sensu stricto (after Patchell et al., 2014) sister to *Polanisia* plus the rest of the family. Most likely due to sampling, our relationships among the Capparaceae were not congruent with previous studies (Hall, 2008; Tamboli et al., 2018). Previous studies with more sampling recovered *Boscia* sp. sister to *Cadaba*, while we recovered *Boscia* as sister to *Capparis*.

Analysis of nuclear data from the transcriptome with ASTRAL-III recovers a well-resolved tree with all nodes but four having a local posterior probability of 0.7 or higher (**Figure 2.1**). The overall relationships of the families and major lineages were congruent with previous studies using transcriptomics (Edger et al., 2015). As with the whole chloroplast phylogeny, we recover a clade of *Moringa oleifera* and *Carica papaya* sister to a clade of [Bataceae + Resedaceae + Capparaceae + Cleomaceae + Brassicaceae], Bataceae sister to [Resedaceae + Capparaceae + Cleomaceae + Brassicaceae], Resedaceae sister to [Capparaceae + Cleomaceae + Brassicaceae], and Capparaceae sister to [Cleomaceae + Brassicaceae]. Within Brassicaceae, the major lineages were recovered, as supported by previous literature (Huang et al., 2016; Nikolov et al., 2019), with *Aethionema arabicum* as sister to the rest of the family, followed by Lineage III sister to [Lineage I + Clade C + Lineage II and expanded Lineage II], and Lineage I sister to [Clade C + Lineage II and expanded Lineage II]. Within Cleomaceae, the relationships were mostly congruent with previous nuclear phylogenies (Patchell et al., 2014) with *Polanisia* sister to *Cleome* sensu stricto, but differing in the placement of *Gynandropsis* (unsupported in Patchell et al., 2014). Additionally, the sampling of only four Capparaceae limited our ability to say much about the relationships within the family; however, to date, there is no phylogeny based solely on nuclear data for this group of plants. Discordance analyses of the Brassicales showed agreement among gene trees for nodes along the backbone, except within and between Clade C, Lineage II, and expanded Lineage II in the Brassicaceae.

Brassicales

Figure 2.1 ASTRAL-III species phylogeny and whole-genome duplication events of the Brassicales. **(A)** Species tree with branch lengths proportional and discordance plotted as pie charts at nodes. Dark grey, proportion of gene trees agreeing with main topology; grey, proportion of gene trees agreeing with first alternative topology; light grey, proportion of gene trees agreeing with second alternative topology. Previously inferred events (At-a and At-β) as well as possible placement of Th-a are indicated. Branches colored by number of unique gene duplications as determined by PUG (github.com/mrmckain/PUG). Support values are indicated if below 0.7 local posterior probability. **(B)** Coalescent-based species tree with branch lengths.

WGD events across the Brassicales

Two popular (and most cost-effective) methods used to detect WGDs across a phylogeny include the analysis of (1) gene tree topologies and (2) Ks plots, which allowed for the identification of signatures left behind in paralogs after WGD. We used a combination of these approaches to test hypotheses of proposed WGD across the Brassicales. Using PUG (github.com/mrmckain/PUG), a gene tree topology WGD estimation method, we recovered some previously inferred events with high support (e.g. At- α and At- β) but did not find strong support for other WGD events, such as the more recent Brassiceae triplication (**Figure 2.1**). Notably, PUG identified only 65 unique gene duplications at the Brassiceae node when only gene trees with >80% bootstrap support were considered. This number is surprisingly low when compared to At- α and At- β with counts over 300 and 150, respectively. To increase the number of orthogroups used to infer species trees as well as to increase the number of gene trees to query putative paralogs against, we independently analyzed phylogenies of each family for evidence of WGD. By evaluating the Brassicaceae, Capparaceae, Cleomaceae, and [Resedaceae + Bataceae + Moriagaceae + Caricaceae] families separately, we were increased gene tree counts in the analyses and improved WGD detection of previously inferred events (**Figure 2.2**).

Figure 2.2 Coalescent-based species phylogeny and whole-genome duplication events of the Brassicaceae. Branches colored by number of unique gene duplications as determined by PUG (github.com/mrmckain/PUG). WGD events identified by PUG (black stars) FASTKs (black squares; McKain et al., 2016); and by DupPipe (black circles; Barker et al., 2010) are indicated. Support values are all above 0.7 local posterior probabilities.

WGD events across the Brassicales: The Brassicaceae

Analysis of the Brassicaceae identified At- α at the base of the family and also successfully identified the Brassicaceae whole genome triplication (Lysak et al., 2005; Tang et al., 2012; **Figure 2.2**). We also recovered neopolyploid (relatively young) events shared between (1) *Chorispora tenella* and *Diptychocarpus strictus*, (2) *Lepidium ruderae* and *L. sativum*, (3) *Descurainia sophioides* and *D. pinnata*, (4) *Turritis glabra* and *Erysimum cheiranthoides*, and (5) a clade of *Isatis lusitanica*, *I. tinctoria*, and *Myagrimum perfoliatum*. Using both FASTKs to estimate pairwise Ks values (github.com/mrmckain/FASTKs; McKain et al., 2016) and DupPipe to estimate Ks values using duplications in gene trees (Barker et al., 2010), the Ks plots mostly showed agreement with the WGD events inferred by the phylogenetic method, PUG. For example, Ks plots from both analyses recovered the Brassicaceae triplication ($K_s \sim 0.3$; **Supplemental Figure 2.5**). However, for the neopolyploid events and At- α , the Ks plots showed differing results between FASTKs and DupPipe, some with and others without evidence for WGD events (**Supplemental Figure 2.5**).

WGD events across the Brassicales: Independent events in Cleomaceae

When running PUG using the Cleomaceae family, we placed Th- α as potentially shared between *Tarenaya hassleriana* and Cleomaceae sp. We also identified additional events between (1) *Coalisina paradoxa* and *Coalisina angustifolia*, (2) four species of *Polanisia*, and (3) *Cleome amblyocarpa*, *Cleome africana*, and *Cleome arabica* (**Figure 2.3**).

Cleomaceae

Unique Gene Duplications

Figure 2.3 Coalescent-based species phylogeny and whole-genome duplication events of the Cleomaceae. Branches colored by number of unique gene duplications as determined by PUG (github.com/mrmckain/PUG). WGD events identified by PUG (black stars), by FASTKs (black squares; McKain et al., 2016), and by DupPipe (black circles; Barker et al., 2010) are indicated. Ks plots using both FASTKs and DupPipe are placed next to their corresponding branch. Y-axes of Ks plots are not congruent: FASTKs measures number of pairs, and DupPipe measures numbers of duplications. Support values are indicated if below 0.7 local posterior probability.

Both methods of Ks estimation provided support for the placement of Th- α with peaks at ~ 0.4 for *T. hassleriana* and Cleomaceae sp. and also for *Melidiscus giganteus*, *Gynandropsis gynandra*, and *Sieruela monophylla*, suggesting that Th- α is shared across more than just *T. hassleriana* and Cleomaceae sp. (**Figure 2.3**). We did not see evidence for this peak in *Arivela viscosa*, which is sister to the species above. As for the other three events, the story is more complicated. When compared to *C. violacea* (which lacks evidence for Th- α ; Emery et al., 2018; One Thousand Plant Transcriptomes Initiative, 2019), we concluded that there was no evidence for two of these novel events: the event shared by the four species of *Polanisia* and the event shared by *Cleome amblyocarpa*, *Cleome africana*, and *Cleome arabica*. However, the event between *Coalisina paradoxa* and *Coalisina angustifolia* does have a signal for a WGD in the Ks plots (**Figure 2.3**).

Due to incongruence of results for the placement of Th- α , we divided potential placements into four hypotheses, H1-H4, to test the age of ortholog divergence between taxa to the age of Th- α (Ks ~ 0.4). We found evidence that Th- α is shared with at least *T. hassleriana*, Cleomaceae sp., and *Melidiscus giganteus* and that Th- α occurred before the divergence between *Melidiscus giganteus* and *T. hassleriana* and around the same time as the divergence of *Gynandropsis gynandra* and *T. hassleriana* (Th- α H2; **Figure 2.4A**). When we compare the divergence between *A. viscosa* and *Gynandropsis gynandra* to the Ks values of these three species along with *Sieruela monophylla*, *A. viscosa*, and *Gynandropsis gynandra*, we found that *A. viscosa* and *Gynandropsis gynandra* diverged more recently than Th- α and that, as in earlier Ks plots, *A. viscosa* lacks evidence for Th- α (Th- α H3; **Figure 2.4B**). This perplexing result could indicate that the data from *A. viscosa* is of poor quality or the genome has lost such a large fraction of the duplicates that the

signal for this event is not detected. To further test for the placement of Th- α , we expanded our comparisons to include the ortholog divergence of *Coalisina angustifolia* and *T. hassleriana* as well as the divergence between *C. violacea* and *T. hassleriana* to test if the proposed independent WGD events between these two clades may be a single event (Th- α H4; **Figure 2.4C**). Surprisingly, the ortholog divergence for both pairs of taxa is about the same age as Th- α . Based on these results, Th- α is either shared across the whole clade (Th- α H4), or is two separate events that happened at approximately the same time. A comparison of ortholog divergence to Ks peaks for the two other identified WGD events using phylogenomics suggests that there is no other WGD event in the Cleomaceae (**Supplemental Figure 2.6**).

Figure 2.4 Comparison of ortholog divergences and Ks peaks of the Cleomaceae to test hypotheses of placement for Th- α . **(A)** Testing H2 by comparison of ortholog divergences of *Melidiscus giganteus* and *Gynandropsis gynandra* to *Tarenaya hassleriana* compared with Ks peaks of *Cleomaceae* sp., *Melidiscus giganteus*, and *Tarenaya hassleriana*. **(B)** Testing of H3 by comparison of ortholog divergence between *Arivela viscosa* and *Tarenaya hassleriana*, and *Gynandropsis gynandra* to *Tarenaya hassleriana* with Ks values of *Cleomaceae* sp., *Sieruela monophylla*, *Melidiscus giganteus*, *Arivela viscosa*, *Gynandropsis gynandra*, and *Tarenaya hassleriana*. **(C)** Testing the H4 hypothesis for placement of Th- α .

WGD events across the Brassicales: The Capparaceae

In agreement with Lysak (2018), our PUG recovered evidence for an independent WGD event in the Capparaceae that is shared between a species of *Capparis* and another species of Capparaceae included in our analyses (**Figure 2.5A**). This event was supported by Ks plots using FastKs, but not DupPipe, with a peak centered at $Ks \sim 0.3$ (**Figure 2.5A**). Ortholog divergences between members of the Capparaceae showed conflicting patterns. When comparing Ks values of *Boscia* sp, *Capparis fascicularis*, Capparaceae sp., and *Cadaba natalensis* from DupPipe to the ortholog divergence time between *Boscia* sp. and *Capparis fascicularis*, we found that the divergence between these two species occurs before the possible WGD event, agreeing with the PUG analysis. All four taxa shared a peak in their Ks distribution, although their Ks plots from both analyses were not in agreement, providing conflicting results for the identification of this event. The divergences tested between *Boscia* sp. and *Cadaba natalensis* and between *Capparis fascicularis* and *Cadaba natalensis* also occur before the proposed event. However, the divergence between *Capparis fascicularis* and a misidentified species of Capparaceae seems to occur at the same time as the peak in Ks values (**Supplemental Figure 2.7**).

Figure 2.5 Coalescent-based species phylogenies and whole-genome duplication events of the (A) Capparaceae and (B) Resedaceae + Outgroups. Branches colored by number of unique gene duplications as determined by PUG (github.com/mrmckain/PUG). WGD events identified by PUG (black stars), by FASTKs (black squares; McKain et al., 2016), and by DupPipe (black circles; Barker et al., 2010) are indicated. Ks plots using both FASTKs and DupPipe are placed next to their corresponding branch. Y-axes of Ks plots are not congruent: FASTKs measures number of pairs, and DupPipe measures numbers of duplications. At-β and At-γ events are noted above corresponding peaks in the Ks plots. Support values are all above 0.7 local posterior probabilities.

WGD events across the Brassicales: Resedaceae specific event

When combining the Resedaceae (*Ochradenus barcardis* and *Reseda odorata*), Bataceae, Moringaceae, and Caricaceae families together, we found strong evidence for a Resedaceae-specific WGD event in all three analyses, with Ks plots indicating a peak at ~ 0.4 (**Figure 2.5B**). The ortholog divergences seemed to support the proposal of this WGD as shared between the samples of Resedaceae. Both samples (*Reseda odorata* and *Ochradenus barcardis*) shared a Ks peak around ~ 0.4 , which occurred before the divergence between these two samples and after the divergence between Resedaceae from *Batis maritima* (**Supplemental Figure 2.7**). In addition, we recovered evidence for At- β using both PUG and DupPipe (Ks ~ 1.7 ; **Figure 2.5B**).

DISCUSSION

Studies of the relationships within the Brassicales have either included many taxa but few genes (Hall et al., 2004; Hall, 2008; Cardinal-McTeague et al., 2016), a few taxa and few genes (Rodman et al., 1998), or few taxa and many genes (Edger et al., 2015, 2018). In this study, we balance taxa and genes to present a well-supported chloroplast and nuclear phylogeny for the Brassicales, both being in overall agreement with previous studies at the interfamilial and intrafamilial levels (Edger et al., 2015, 2018; Cardinal-McTeague et al., 2016; Huang et al., 2016; Guo et al., 2017). Using the nuclear phylogeny, we highlight several potential placements of the Th- α WGD event and identify other possible novel events in the Cleomaceae, Capparaceae, and Resedaceae.

Incongruences between the chloroplast and nuclear trees across the Brassicaceae

Although relationships in our nuclear and chloroplast phylogenies are congruent with previous analyses, we highlight incongruence *between* the nuclear and chloroplast trees among the major lineages of the Brassicaceae, a well-documented pattern between these genomes (Beilstein et al., 2008; Huang et al., 2016; Nikolov et al., 2019; summarized in **Figure 2.6**). We find Lineage I sister to [Lineage III + Lineage II + Expanded Lineage II + Clade C] in the chloroplast tree and Lineage III sister to [Lineage I + Lineage II + Expanded Lineage II + Clade C] in the nuclear tree. Huang et al. (2016), using 113 low-copy nuclear genes from 55 Brassicaceae species, recovered a tree congruent with our nuclear phylogeny as did Nikolov et al. (2019) in their study using 79 species and 1,421 exons. Additionally, Guo et al. (2017), using 77 chloroplast genes from 53 samples, recovered a phylogeny in agreement with our chloroplast tree. With additional taxon sampling, an increase in data, and using the same samples across analyses, we too recover incongruent relationships between nuclear and chloroplast trees, leaving us to conclude that the trees from these genomes are in disagreement due to different evolutionary histories, which could include ancient hybridization or introgression events (Forsythe et al., 2018). For future users of these phylogenies, the differences between these two trees are important to consider when using the phylogeny to assess character evolution and divergence dating, as node ordering depends on which tree is used.

Figure 2.6 Comparison of (A) maximum likelihood whole-chloroplast phylogeny to (B) coalescent-based species phylogeny of the Brassicales. Major lineages and clades of the Brassicaceae are indicated. Support values are indicated if below 0.7 local posterior probabilities or 70% bootstrap support.

Putative placement of Th-a in the Cleomaceae

Previous studies have identified a WGD event unique to Cleomaceae (Th- α) using a variety of sources from syntenic regions to ESTs (Schranz and Mitchell-Olds, 2006; Barker et al., 2009; reviewed in Bayat et al., 2018; One Thousand Plant Transcriptomes Initiative, 2019). Placement of Th- α within the Cleomaceae had yet to be confirmed. By using Ks plots to assess for the signatures of WGD left behind in the paralogs they created, phylogenetics using individual gene tree topologies, gene counts, a known species tree, and ortholog divergences, we are able to putatively place Th- α as shared between *Tarenaya hassleriana*, Cleomaceae sp., *Melidiscus giganteus*, *Gynandropsis gynandra*, and *Sieruela*

monophyla, *Arivela viscosa*, *Coalisina angustifolia*, and *Coalisina paradoxa* (Th-α H4; **Figure 2.3**). We include these last three species due to both the evidence from ortholog divergences and the signatures in Ks plots that strongly suggest this event is shared with all species (**Figures 2.3 and 2.4**). It is possible that two separate events occurred independently and that *A. viscosa* does indeed lack a WGD. Although this hypothesis (Th-α H3) is in agreement with the most recent analysis of this group by the One Thousand Plant Transcriptomes Initiative (2019), their sampling of the Cleomaceae was limited and did not include *Coalisina angustifolia* and *Coalisina paradoxa*. Ks plots of all samples, other than *A. viscosa*, identify a peak at Ks ~ 0.4, agreeing with previous studies (Barker et al., 2009; van den Bergh et al., 2014) that first identified this peak in *Tarenaya hassleriana* followed by *Gynandropsis gynandra*. PUG, however, supports two separate events. The difficulty in placing this event may be due to the gene tree discordance found within this clade (**Figure 2.1**) or to the possibility that, like others in the order, it is a whole genome triplication, in which case it will be hard to tease apart due to having to resolve two serial duplication events that occurred in a short time period (as with the tribe Brassiceae; Tang et al. 2012).

Multiple WGD events in the Cleomaceae?

In addition to placing Th-α, we report two possible additional WGD events in the Cleomaceae. Both events are identified in the Brassicales and the Cleomaceae analyses, but with much greater support in the analysis of the Cleomaceae species. These WGD events are placed at common ancestors shared between (1) *Cleome amblyocarpa*, *C. africana*, and *C. arabica* and (2) four species of *Polanisia* (**Figure 2.3**). Ks plots provide contrasting support for these events. Ks plots from FASTKs of *C. africana* and *C. arabica*

show a small peak of duplicates at $K_s \sim 0.3$. Yet, there is no evidence of a WGD event when the same data are run through DupPipe. The K_s plots of *C. amblyocarpa* also give conflicting evidence for this event. The K_s plot from FASTKs looks much more similar to that of *C. violacea*, which we know based on genome sequencing does not show evidence of any recent WGD event (Emery et al., 2018). The second event shared between the four species of *Polanisia* is supported by a large number (2,200) of unique gene duplications using PUG but not by K_s plots from either FASTKs or DupPipe; the resulting plots look more similar to *C. violacea*. Analyses of ortholog divergence between *C. amblyocarpa*, *C. africana*, and *C. arabica* also lack support for a WGD (**Supplemental Figure 2.6**) as do analyses between the four species of *Polanisia* (**Supplemental Figure 2.6**). To further test how WGD and C_4 photosynthesis have evolved in this family, we suggest a study focusing primarily on Cleomaceae sampling. C_4 photosynthesis has evolved at least three times independently in Cleomaceae, specifically in (of the taxa sampled) *Gynandropsis gynandra* and *Coalisina angustifolia* with *Coalisina paradoxa* as a $C_3 - C_4$ intermediate in anatomy and physiology (Bhide et al., 2014). If our putative placement of Th- α is correct, then all of these samples share this event. It will be interesting to investigate the role of polyploidy and, more specifically, Th- α , in character evolution in this group.

WGD events in the Capparaceae and Resedaceae

Although our analysis included only two samples, we recover some support for an event between at least one species of *Capparis* and the misidentified species of Capparaceae (**Figure 2.5A**). Due to this ambiguous species identification, inconclusiveness from K_s plots, and no support in comparison between ortholog divergence and K_s peaks, this event, although supported by many unique duplicates in the

PUG analysis, should be interpreted carefully. That said, Lysak (2018), using chromosome counts, also proposed that Capparaceae had a unique event; although, chromosome counts alone may be misleading in concluding that a WGD event has occurred (Evans et al., 2017). There is also a lack of agreement between Ks plots derived using FASTKs and DupPipe, which estimate Ks values differently (i.e., pairwise Ks estimates in FASTKs versus estimates of Ks at nodes in gene trees in DupPipe), further confounding evidence for either presence or absence of a Capparaceae-specific event. Between information presented by Lysak (2018) and the evidence presented here, this possible event certainly warrants additional study.

We find good evidence to support the presence of a separate WGD event in the Resedaceae, also hypothesized by Lysak (2018). This polyploidy is one of the few events recovered with consensus between Ks plots (from both FASTKs and DupPipe), phylogenetics, and ortholog divergences (**Figure 2.4B** and **Supplemental Figure 2.7**). Therefore, we are confident in the identification of this event (named Rs-a here). This event was recently identified using a single species (*Reseda odorata*; One Thousand Plant Transcriptomes Initiative 2019); however, we determine that it is shared with at least one other species in the family, *Ochradenus barcardis*. The sister families, Caricaceae and Moringaceae, show no evidence of unique WGD events, which is in agreement with the recent whole-genome sequencing of *Moringa oleifera* (Chang et al., 2018). When Tian et al. (2015) compared the papaya genome, which shows no evidence of a (recent) WGD (Ming et al., 2008), to the newly sequenced genome of *Moringa oleifera*, they too concluded that Moringaceae did not experience a family-specific genome duplication. The

identified and well-supported Resedaceae event warrants additional sampling and investigation to test if it is shared across the whole family.

Methodological challenges with placing WGD events: sampling matters

Currently, three types of methods are used to detect WGD events: (1) Ks plots to assess for signatures left behind in paralogs after WGD, (2) identification of retained duplicate blocks in a genome, and (3) phylogenetics using individual gene tree topologies. Since all three methods have limitations in identifying WGD events, we use a combination of approaches to test hypotheses, reduce proposing events that may not exist, and simultaneously provide multiple lines of evidence for recovered events.

Recently, an abundance of papers have highlighted the difficulties and complexities of determining WGD events across the tree of life (Conover et al., 2018; Tiley et al., 2018; Li and Barker, 2019; Li et al., 2019; Nakatani and McLysaght, 2019; Zwaenepoel and van de Peer, 2019; Zwaenepoel et al., 2019). We add another dimension to this conversation by demonstrating that the different taxonomic levels from which we sample, such as the order or family, make a difference in support of previously identified WGDs (i.e., the Brassiceae triplication). Recent research has demonstrated that differences in taxonomic sampling and taxon occupancy in data matrices can influence the inference of WGDs, particularly if adding taxa decreases taxon occupancy in gene families (Yang et al., 2015; Li et al., 2018; Li and Barker, 2019; Zwaenepoel and van de Peer, 2019). Testing for WGD events across the Brassicales phylogeny led to less certain topologies; therefore, signals of WGD are missed when filtering for nodes with only high bootstrap support to count duplicates. To account for this and to increase taxon and gene family occupancy in our datasets, we reduce sampling to just the family level to infer WGD. However, at each level

of analysis, we choose an arbitrary cut-off for the number of duplicates that we feel is sufficient to infer a WGD event, a documented criticism of these types of methods (Zwaenepoel and van de Peer, 2019). Others note that it is also important to consider heterogeneity in substitution rates (Cui et al., 2006; Barker et al., 2009; Yang et al., 2015), but phylogenomic methods and denser sampling should obviate the need for rate corrections. Variation in the duplication and loss rate across the species tree may also impact tests for WGD events (Li et al., 2018; Zwaenepoel and van de Peer, 2019).

Although our Ks-based inferences of WGDs are largely consistent with phylogenomic inferences, there are some differences among the approaches. FASTKs and DupPipe use different estimates of Ks that likely produce the observable differences in the respective Ks plots. FASTKs uses a pairwise approach for estimating Ks values (github.com/mrmckain/FASTKs; McKain et al., 2016), whereas DupPipe estimates Ks values from nodes of gene trees (Barker et al., 2010). Tiley et al. (2018) explored the difference in Ks estimates from these types of approaches and found that the observed differences in peaks of duplications between the two different methods is consistent with simulations. The node-based estimates of Ks from DupPipe often yield sharper peaks in putative WGDs, with overall lower numbers of duplications because of the difference in number of nodes vs pairwise comparisons. However, the results of both approaches are largely consistent after close inspection. Perhaps more confounding for Ks analyses is the interpretation of mixture models to identify putative peaks associated with a WGD. Mixture models, which are typically fit to the distribution of duplicates, tend to overestimate the number of true peaks (Naik et al., 2007; Tiley et al., 2018; Zwaenepoel et al., 2019). Using two different methods across multiple species allowed us to evaluate and

compare putative peaks from different analyses to identify the expected signatures of WGDs. Furthermore, because paralogs from WGDs tend to be more highly expressed than those resulting from tandem duplications (Casneuf et al., 2006), transcriptomes may yield data that is more enriched for WGD duplicates than (fragmented) genomic data. Therefore, transcriptome data, as shown by Tiley et al. (2018), may actually improve success in detecting WGD events.

CONCLUSIONS

The Brassicales is an excellent group to compare methods of WGD identification because of the wealth of genomic data available and previously inferred WGDs. With many chromosome-level genomes available, analyses based on synteny, which seem to be regarded as most reliable in detecting these events (Nakatani and McLysaght, 2019), can be used as controls for comparing WGD methods. Sequenced genomes, which are placed throughout the Brassicales, provide strong evidence for taxa that we know do not have recent WGD events (i.e., *Cleome violacea* and *Carica papaya*) and for taxa that do show evidence for recent WGD events (i.e. *Arabidopsis thaliana* and many *Brassica* crops). These resources provide calibration points that can be used to verify results when testing for novel events. This group of plants, combined with recent insights on difficulties in placing WGD events, can help further the development of innovative methods in describing and identifying WGDs.

ACKNOWLEDGMENTS

We thank our funding sources, including NSF grants IOS-1339156 and EF-1550838 to MSB, NSF-DEB 1737898 to PPE, NSF IOS-1811784 to PDB, NIH PERT K12GM000708 to BS, and NSF DEB-1146603 to JCP. We thank the staff from the University of Missouri, specifically Nathan Bivens, Ming-Yi Zhou, and Karen Bromert with the DNA Core for their patience and assistance in getting successful libraries for all samples; Jake Gotburg with the Research Computing Support Services (RCSS) for assistance in running large jobs across the Lewis cluster; and Barb Sonderman with the Division of Biological Sciences for assistance in keeping many of these plants alive. Finally, we acknowledge Sarah Unruh and Melody Kroll and our anonymous reviewers for providing feedback that improved this manuscript.

AUTHOR CONTRIBUTIONS

MEM, JCP, GCC, JCH, PPE, and MES designed the project. MEM, JMB, HA, JDW, and WTD grew plants, collected tissue, and isolated RNA. MEM and JMB isolated DNA. JCH contributed *Cleomella serrulate*. MEM and JMB analyzed the data. PDB and JDW assisted with processing and analyzing the data. CAB wrote alignment filtering scripts. MSB and BS analyzed data using DupPipe and OrthoPipe. MRM helped with processing data using FastPlast. IA and JCH updated nomenclature and helped to correct plant identifications. MEM wrote original manuscript draft, which was read and approved by all authors.

DATA AVAILABILITY

Sequence data from this article can be found in the NCBI SRA data libraries under BioProject accession number PRJNA542714. Individual BioSample accession numbers can be found in **Supplemental Table 2.1**. Seeds corresponding to samples are available upon request.

LITERATURE CITED

- APG IV. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, 181(1): 1-20.
- Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814): 796.
- Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5): 455-477.
- Barco, B. and N. K. Clay. 2019. Evolution of glucosinolate diversity via whole-genome duplications, gene rearrangements, and substrate promiscuity. *Annual Review of Plant Biology*, 70: 585-604.
- Barker, M.S., H. Vogel, and M. E. Schranz. 2009. Paleopolyploidy in the Brassicales: analyses of the Cleome transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. *Genome Biology and Evolution*, 1: 391-399.
- Barker, M. S., K. M. Dlugosch, L. Dinh, R. S. Challa, N. C. Kane, M. G. King, and L. H. Rieseberg. 2010. EvoPipes.net: bioinformatic tools for ecological and evolutionary genomics. *Evolutionary Bioinformatics*, 6: EBO-S5861.
- Bayat, S., M. E. Schranz, E. H. Roalson, and J. C. Hall. 2018. Lessons from Cleomaceae, the sister of crucifers. *Trends in Plant Science* 23: 808–821
- Beilstein, M.A., I. A. Al-Shehbaz, and E. A. Kellogg. 2006. Brassicaceae phylogeny and trichome evolution. *American Journal of Botany*, 93(4): 607-619.

- Beilstein, M.A., I. A. Al-Shehbaz, S. Mathews, and E. A. Kellogg. 2008. Brassicaceae phylogeny inferred from phytochrome A and ndhF sequence data: tribes and trichomes revisited. *American Journal of Botany*, 95(10): 1307-1327.
- Bhide A., S. Schliesky, M. Reich, A. P. Weber, and A. Becker. 2014. Analysis of the floral transcriptome of *Tarenaya hassleriana* (Cleomaceae), a member of the sister group to the Brassicaceae: towards understanding the base of morphological diversity in Brassicales. *BMC Genomics* 15(1): 140.
- Blažević, I., S. Montaut, F. Burčul, C. E. Olsen, M. Burow, P. Rollin, and N. Agerbirk. 2019. Glucosinolate structural diversity, identification, chemical synthesis and metabolism in plants. *Phytochemistry*, 112100.
- Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15): 2114-2120.
- BrassiBase. Tools and biological resources for Brassicaceae character and trait studies. <https://brassibase.cos.uni-heidelberg.de>. Accessed 24 Mar 2019.
- Brock, K.C. 2014. Tracking the Evolutionary History of Development Genes: Implications for the Diversification of Fruits and Flowers in the Brassicaceae and Cleomaceae. *Master's thesis. University of Alberta, Edmonton, Alberta, Canada.*
- Buchfink, B., Xie, C., and Huson, D.H. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1): 59.
- Cardinal-McTeague W., K. J. Sytsma, and J. C. Hall. 2016. Biogeography and diversification of Brassicales: a 103 million year tale. *Molecular Phylogenetics and Evolution*, 99: 204-224

- Casneuf, T., S. De Bodt, J. Raes, S. Maere, and Y. Van de Peer. 2006. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. *Genome Biology*, 7(2): R13.
- Chang, Y., H. Liu, M. Liu, X. Liao, S. K. Sahu, Y. Fu, B. Song, et al. 2018. The draft genomes of five agriculturally important African orphan crops. *GigaScience* 8(3): giy152.
- Cheng, S., E. van den Bergh, P. Zeng, X. Zhong, J. Xu, X. Liu, J. Hofberger, et al. 2013. The *Tarenaya hassleriana* genome provides insight into reproductive trait and genome evolution of crucifers. *The Plant Cell*, 25(8): 2813-2830.
- Conover, J. L., N. Karimi, N. Stenz, C. Ané, C. E. Grover, C. Skema, J. A. Tate, et al. 2018. A Malvaceae mystery: a mallow maelstrom of genome multiplications and maybe misleading methods? *Journal of Integrative Plant Biology*, 61(1): 12-31.
- Cui, L., P. K. Wall, J. H. Leebens-Mack, B. G. Lindsay, D. E. Soltis, J. J. Doyle, P. S. Soltis, et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Research*, 16(6): 738-749.
- Duarte, J. M., P. K. Wall, P. P. Edger, L. L. Landherr, H. Ma, J. C. Pires, and J. Leebens-Mack. 2010. Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology*, 10(1): 61.
- Dunn, C. W., A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne, S. A. Smith, E. Seaver, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452(7188): 745.

- Dunn, C., Howison, M. & Zapata, F. 2013. Agalma: an automated phylogenomics workflow. *BMC Bioinformatics*, 14: 330.
- Edger P. P., H. M. Heidel-Fischer, M. Bekaert, J. Rota, G. Glöckner, A. E. Platts, D. G. Heckel, et al. 2015. The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences*, 112(27): 8362-8366.
- Edger, P.P., J. C. Hall, A. Harkess, M. Tang, J. Coombs, S. Mohammadin, M. E. Schranz, et al. 2018. Brassicales phylogeny inferred from 72 plastid genes: a reanalysis of the phylogenetic localization of two paleopolyploid events and origin of novel chemical defenses. *American Journal of Botany*, 105(3): 463-469.
- Emery, M., M. M. S. Willis, Y. Hao, K. Barry, K. Oakgrove, Y. Peng, J. 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): p.e1007267.
- Emms, D. M. and S. Kelly. 2015. OrthoFinder: solving fundamental biases in whole-genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1): 157.
- Emms, D. M. and S. Kelly. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1): 1-14.
- Endress, P. K. 1992. Evolution and floral diversity: the phylogenetic surroundings of *Arabidopsis* and *Antirrhinum*. *International Journal of Plant Sciences*, 153(3, Part 2): S106-S122.

- Evans, B. J., N. S. Upham, G. B. Golding, R. A. Ojeda, and A. A. Ojeda. 2017. Evolution of the largest mammalian genome. *Genome Biology and Evolution*, 9(6): 1711-1724.
- Feodorova, T. A., E. V. Voznesenskaya, G. E. Edwards, E. H. and Roalson. 2010. Biogeographic patterns of diversification and the origins of C4 in *Cleome* (Cleomaceae). *Systematic Botany*, 35(4), 811-826.
- Forsythe, E.S., A. D. Nelson, and M. A. Beilstein. 2018. Biased gene retention in the face of massive nuclear introgression obscures species relationships. *bioRxiv*, 197087.
- Fraley, C., and A. E. Raftery. 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458): 611-631.
- Fraley, C., A. E. Raftery, and L. Scrucca. 2012. Normal mixture modeling for model-based clustering, classification, and density estimation. *Department of Statistics University of Washington* 23, 2012.
- Godden, G. T., T. J. Kinser, P. S. Soltis, and D. E. Soltis. 2019. Phylotranscriptomic analyses reveal asymmetrical gene duplication dynamics and signatures of ancient polyploidy in mints. *Genome Biology and Evolution*, 11(12): 3393-3408.
- Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7): 644.
- Guo, X., J. Liu, G. Hao, L. Zhang, K. Mao, X. Wang, D. Zhang, et al. 2017. Plastome phylogeny and early diversification of Brassicaceae. *BMC Genomics*, 18(1):176.

- Hall, J. C., K. J. Sytsma, and H. H. Iltis. 2002. Phylogeny of Capparaceae and Brassicaceae based on chloroplast sequence data. *American Journal of Botany*, 89(11): 1826-1842.
- Hall J. C., H. H. Iltis, and K. J. Sytsma. 2004. Molecular phylogenetics of core Brassicales, placement of orphan genera Emblingia, Forchhammeria, Tirania, and character evolution. *Systematic Botany*, 29(3): 654-669.
- Hall, J.C. 2008. Systematics of Capparaceae and Cleomaceae: an evaluation of the generic delimitations of Capparis and Cleome using plastid DNA sequence data. *Botany*, 86(7): 682-696.
- Haudry, A., A. E. Platts, E. Vello, D. R. Hoen, M. Leclercq, R. J. Williamson, E. Forczek, et al. 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nature Genetics*, 45(8): 891.
- Huang C.H., R. Sun, Y. Hu, L. Zeng, N. Zhang, L. Cai, Q. Zhang, et al. 2016. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Molecular Biology and Evolution*, 33(2): 394-412.
- Kagale, S., S. J. Robinson, J. Nixon, R. Xiao, T. Huebert, J. Condie, D. Kessler, et al. 2014. Polyploid evolution of the Brassicaceae during the Cenozoic era. *The Plant Cell*, 26(7): 2777-2791.
- Katoh, K., K. Misawa, K. I. Kuma, and T. Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14): 3059-3066.

- Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12): 1647-1649.
- 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): 681-693.
- Kocot, K.M., M. R. Citarella, L. L. Moroz, and K. M. Halanych. 2013. PhyloTreePruner: a phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evolutionary Bioinformatics* 9: EBO-S12813.
- Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie2. *Nature Methods*, 9(4): 357.
- Li, Z., G. P. Tiley, R. J. Rundell, and M. S. Barker. 2019. Reply to Nakatani and McLysaght: Analyzing deep duplication events. *Proceedings of the National Academy of Sciences*, 116(6): 1819-1820.
- Li, Z, and M. S. Barker. 2019 Inferring putative ancient whole genome duplications in the 1000 Plants (1KP) initiative: access to gene family phylogenies and age distributions. *bioRxiv* 735076.
- Li, Z., G. P. Tiley, S. R. Galuska, C. R. Reardon, T. I. Kidder, R. J. Rundell, and M. S. Barker. 2018. Multiple large-scale gene and genome duplications during the

- evolution of hexapods. *Proceedings of the National Academy of Sciences*, 115(18): 4713-4718.
- Lysak, M.A. 2018. Brassicales: an update on chromosomal evolution and ancient polyploidy. *Plant Systematics and Evolution* 1-6.
- Lysak, M. A., M. A. Koch, A. Pecinka, and I. Schubert. 2005. Chromosome triplication found across the tribe Brassiceae. *Genome Research*, 15(4): 516-525.
- Magallon, S., P. R. Crane, and P. S. Herendeen. 1999. Phylogenetic pattern, diversity, and diversification of eudicots. *Annals of the Missouri Botanical Garden* 297-372.
- Mandáková, T., Z. Li, M. S. Barker, and M. A. Lysak. 2017. Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. *The Plant Journal*, 91(1): 3-21.
- Martín-Bravo, S., H. Meimberg, M. Luceño, W. Märkl, V. Valcárcel, C. Bräuchler, P. Vargas, and G. Heubl. 2007. Molecular systematics and biogeography of Resedaceae based on ITS and trnL-F sequences. *Molecular Phylogenetics and Evolution*, 44(3): 1105-1120.
- Martín-Bravo, S., P. Vargas, and M. Luceño. 2009. Is *Oligomeris* (Resedaceae) indigenous to North America? Molecular evidence for a natural colonization from the Old World. *American Journal of Botany*, 96(2): 507-518.
- McKain, M.R., N. Wickett, Y. Zhang, S. Ayyampalayam, W. R. McCombie, M. W. Chase, J. C. Pires, et al. 2012. Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in Agavoideae (Asparagaceae). *American Journal of Botany*, 99(2): 397-406.

- McKain M. R., and M. Wilson. 2017. Fast-Plast: rapid *de novo* assembly and finishing for whole chloroplast genomes. Version 1.2.6. <https://github.com/mrmckain/Fast-Plast>
- McKain, M. R., H. Tang, J. R. McNeal, S. Ayyampalayam, J. I. Davis, C. W. dePamphilis, T. J. Givnish, et al. 2016. A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. *Genome Biology and Evolution*, 8(4): 1150-1164.
- Ming, R., S. Hou, Y. Feng, Q. Yu, A. Dionne-Laporte, J. H. Saw, P. Senin, et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*, 452(7190): 991.
- Mithen, R., R. Bennett, and J. Marquez. 2010. Glucosinolate biochemical diversity and innovation in the Brassicales. *Phytochemistry*, 71(17-18): 2074-2086.
- Naik, P. A., P. Shi, and C. L. Tsai. 2007. Extending the Akaike information criterion to mixture regression models. *Journal of the American Statistical Association*, 102(477): 244-254.
- Nakatani, Y. and A. McLysaght. 2019. Macrosynteny analysis shows the absence of ancient whole-genome duplication in lepidopteran insects. *Proceedings of the National Academy of Sciences*, 116(6): 1816-1818.
- Nikolov, L. A., P. Shushkov, B. Nevado, X. Gan, I. A. Al-Shehbaz, D. Filatov, C. D. Bailey, and M. Tsiantis. 2019. Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. *New Phytologist*, 222(3): 1638-1651.
- One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, 574(7780): 679.

- Patchell, M. J., E. H. Roalson, and J. C. Hall. 2014. Resolved phylogeny of Cleomaceae based on all three genomes. *Taxon*, 63(2): 315-328.
- R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Ratzka, A., H. Vogel, D. J. Kliebenstein, T. Mitchell-Olds, and J. Kroymann. 2002. Disarming the mustard oil bomb. *Proceedings of the National Academy of Sciences*, 99(17): 11223-11228.
- Rodman, J. E., P. S. Soltis, D. E. Soltis, K. J. Sytsma, and K. G. Karol. 1998. Parallel evolution of glucosinolate biosynthesis inferred from congruent nuclear and plastid gene phylogenies. *American Journal of Botany*, 85(7): 997-1006.
- Schlüter, U., A. Bräutigam, U. Gowik, M. Melzer, P. A. Christin, S. Kurz, T. Mettler-Altmann, and A. P. Weber. 2016. Photosynthesis in C3–C4 intermediate *Moricandia* species. *Journal of Experimental Botany*, 68(2): 191-206.
- Schranz M. E. and T. Mitchell-Olds. 2006. Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. *The Plant Cell*, 18(5): 1152-1165
- Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19): 3210-3212.
- Smith S. A. and C. W. Dunn. 2008. Phyutility: a phyloinformatics tool for trees, alignments, and molecular data. *Bioinformatics*, 24: 715-716.

- Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9): 1312-1313.
- Tamboli, A. S., P. B. Yadav, A. A. Gothe, S. R. Yadav, and S. P. Govindwar. 2018. Molecular phylogeny and genetic diversity of genus *Capparis* (Capparaceae) based on plastid DNA sequences and ISSR markers. *Plant Systematics and Evolution*, 304(2): 205-217.
- Tang, H., M.R. Woodhouse, F. Cheng, J.C. Schnable, B.S. Pedersen, G. Conant, X. Wang, M. Freeling, and J.C. Pires. 2012. Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. *Genetics* 190: 1563–1574.
- Tian, Y., Y. Zeng, J. Zhang, C. Yang, L. Yan, X. Wang, C. Shi, et al. 2015. High quality reference genome of drumstick tree (*Moringa oleifera* Lam.), a potential perennial crop. *Science China Life Sciences*, 58(7): 627-638.
- Tiley, G. P., M. S. Barker, and J. G. Burleigh. 2018. Assessing the performance of Ks plots for detecting ancient whole-genome duplications. *Genome Biology and Evolution*, 10(11): 2882-2898.
- Tobe, H., and P. H. Raven. 2008. Embryology of *Koeberlinia* (Koeberliniaceae): Evidence for core-Brassicalean affinities. *American journal of botany*, 95(11), 1475-1486.
- Unruh, S. A., M. R. McKain, Y. I. Lee, T. Yukawa, M. K. McCormick, R. P. Shefferson, A. Smithson, et al. 2018. Phylotranscriptomic analysis and genome evolution of the Cypridioideae (Orchidaceae). *American Journal of Botany*, 105(4): 631-640.

- van den Bergh, E., C. Külahoglu, A. Bräutigam, J. M. Hibberd, A. P. M. Weber, X. G. Zhu, and M. E. Schranz. 2014. Gene and genome duplications and the origin of C4 photosynthesis: birth of a trait in the Cleomaceae. *Current Plant Biology*, 1: 2-9.
- Vision, T. J., D. G. Brown, and S. D. Tanksley. 2000. The origins of genomic duplications in *Arabidopsis*. *Science*, 290(5499): 2114-2117.
- Washburn, J. D., J. C. Schnable, G. C. Conant, T. P. Brutnell, Y. Shao, Y. Zhang, M. Ludwig, et al. 2017. Genome-guided phylo-transcriptomic methods and the nuclear phylogenetic tree of the Paniceae grasses. *Scientific Reports*, 7(1): 13528.
- Waterhouse, R. M., M. Seppey, F. A. Simão, M. Manni, P. Ioannidis, G. Klioutchnikov, E. V. Kriventseva, and E. M. Zdobnov. 2017. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, 35(3): 543-548.
- Yang, Y. and S. A. Smith. 2014. Orthology inference in nonmodel Organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution*, 31: 3081-3092.
- Yang, Y., M. J. Moore, S. F. Brockington, D. E. Soltis, G. K. S. Wong, E. J. Carpenter, Y. Zhang, et al. 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. *Molecular Biology and Evolution*, 32(8): 2001-2014.
- Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(6): 153.

- Züst, T., M. Mirzaei, and G. Jander. 2018. *Erysimum cheiranthoides*, an ecological research system with potential as a genetic and genomic model for studying cardiac glycoside biosynthesis. *Phytochemistry Reviews*, 17(6): 1239-1251.
- Zwaenepoel, A., Z. Li, R. Lohaus, and Y. van de Peer. 2019. Finding evidence for whole-genome duplications: a reappraisal. *Molecular Plant*, 12(2): 133-136.
- Zwaenepoel, A. and Y. van de Peer. 2019. Ancient whole-genome duplications and the evolution of the gene duplication and loss rate. *bioRxiv*, 556076.

Supplemental Figure 2.1 Current understanding of the relationships between the 17 families of the Brassicales (APG IV). An asterisk (*) indicates branch support between 50-80%. All other branches have greater than 80% support. Orange stars indicate whole-genome duplication events supported by the One Thousand Plant Transcriptomes Initiative (2019). It is uncertain if the At- β event is shared by Setchellanthaceae or not (latter scenario is indicated; but see Edger et al. 2018).

Supplemental Table 2.1 Taxon sampling with authorship (determined using <http://www.ipni.org/index.html>), seed accessions and the collections they are from, additional analysis information, and SRA accessions. SSC = small single copy, LSC= large single copy, IR = inverted repeat.

Family	Genus	Species and Authorship	Seed Accession	Collection	Transcriptome	Chloroplast SSC	Chloroplast LSC	Chloroplast IR	Fast-Plast Subsample	SRA Accession
Bataceae	<i>Batis</i>	<i>maritima</i> L.	NA	Pat Edger	X	X	X	X	3500000	SAMN12815478
Brassicaceae	<i>Aethionema</i>	<i>arabicum</i> (L.) Andr. ex DC.	1010	Eric Schranz	X	X	X	X		SAMN12815475
Brassicaceae	<i>Barbarea</i>	<i>bracteosa</i> Guss.	6203-84	Gomez Campo	X	X	X	X		SAMN12815477
Brassicaceae	Brassicaceae Burnett	sp.	Colorado diploid	Chris Pires	X	X	X	X		SAMN12815542
Brassicaceae	<i>Cakile</i>	<i>maritima</i> Scop.	B2005-875	Marcus Koch	X	X	X	X		SAMN12815481
Brassicaceae	<i>Calepina</i>	<i>irregularis</i> Thell.	2158-72	Gomez Campo	X	X	X	X		SAMN12815482
Brassicaceae	<i>Capsella</i>	<i>bursa-pastoris</i> (L.) Medik.	2006-956	Marcus Koch	X	X	X	X		SAMN12815485
Brassicaceae	<i>Cardamine</i>	<i>hirsuta</i> L.	2005-701	Marcus Koch	X	X				SAMN12815486
Brassicaceae	<i>Caulanthus</i>	<i>amplexicaulis</i> S. Watson	NA	Alan Pepper	X	X	X	X		SAMN12815488
Brassicaceae	<i>Chorisporea</i>	<i>tenella</i> DC.	2006-629	Marcus Koch	X	X	X	X	1000000	SAMN12815489
Brassicaceae	<i>Cochlearia</i>	<i>officinalis</i> L.	1910-84	Gomez Campo	X	X	X	X		SAMN12815502
Brassicaceae	<i>Crambe</i>	<i>hispanica</i> L.	PI388853	USDA	X	X	X	X		SAMN12815503
Brassicaceae	<i>Descurainia</i>	<i>pinnata</i> (Walter) Britton	B2004-26	Marcus Koch	X	X	X	X		SAMN12815505
Brassicaceae	<i>Descurainia</i>	<i>sophioides</i> (Fisch.) O.E. Schulz	BarbGH	Chris Pires	X	X	X	X		SAMN12815504
Brassicaceae	<i>Diptyocarpus</i>	<i>strictus</i> (Fisch. ex M.Bieb.) Trautv	05-0397-10-00	Klaus Mummenhoff	X					SAMN12815506
Brassicaceae	<i>Eruca</i>	<i>vesicaria</i> (L.) Cav.	PI633218	USDA	X	X	X	X	1000000	SAMN12815507
Brassicaceae	<i>Erysimum</i>	<i>cheiranthoides</i> L.	03-0030-10-00	Klaus Mummenhoff	X	X	X	X		SAMN12815508
Brassicaceae	<i>Euclidium</i>	<i>syriacum</i> (L.) W.T. Aiton	0587-68	Gomez Campo	X	X	X	X		SAMN12815509
Brassicaceae	<i>Farsetia</i>	<i>aegyptia</i> Turra	96-001-10-00	Klaus Mummenhoff	X		X			SAMN12815510
Brassicaceae	<i>Goldbachia</i>	<i>laevigata</i> DC.	NA	Martin Lysak	X	X	X	X		SAMN12815511
Brassicaceae	<i>Guillenia</i>	<i>lasiophylla</i> Green	Costal type	Chris Pires	X	X	X	X		SAMN12815512
Brassicaceae	<i>Hesperis</i>	<i>matronalis</i> L.	0039-68	Gomez Campo	X	X	X	X	1000000	SAMN12815514

Brassicaceae	<i>Hirschfeldia</i>	<i>incana</i> (L.) Lagr.-Foss	2024-71	Gomez Campo	X	X	X	X		SAMN12815515
Brassicaceae	<i>Iberis</i>	<i>amara</i> L.	2005-238	Marcus Koch	X	X	X	X		SAMN12815516
Brassicaceae	<i>Isatis</i>	<i>lusitanica</i> L.	01-0035-10-00	Klaus Mummenhoff	X	X	X	X		SAMN12815517
Brassicaceae	<i>Isatis</i>	<i>tinctorial</i> L.	B2009-0378	Marcus Koch	X	X	X	X		SAMN12815518
Brassicaceae	<i>Leavenworthia</i>	<i>uniflora</i> (Michx.) Britton	2004-557	Marcus Koch		X	X	X		SAMN12815549
Brassicaceae	<i>Lepidium</i>	<i>ruderales</i> L.	05-0190-10-00	Klaus Mummenhoff	X	X	X	X		SAMN12815519
Brassicaceae	<i>Lepidium</i>	<i>sativum</i> L.	2003-169	Marcus Koch	X	X	X	X		SAMN12815520
Brassicaceae	<i>Lobularia</i>	<i>maritima</i> (L.) Desv.	B2009-0390	Marcus Koch	X	X	X	X		SAMN12815521
Brassicaceae	<i>Lunaria</i>	<i>annua</i> L.	2006-705	Marcus Koch	X	X	X	X		SAMN12815522
Brassicaceae	<i>Malcolmia</i>	<i>maritima</i> (L.) W.T. Aiton	2008-777	Marcus Koch	X	X	X	X	2500000	SAMN12815523
Brassicaceae	<i>Matthiola</i>	<i>longipetala</i> DC.	1351-70	Gomez Campo	X	X	X	X		SAMN12815524
Brassicaceae	<i>Meniocus</i>	<i>linifolius</i> (Willd.) DC.	0893-68	Gomez Campo	X	X	X	X	1000000	SAMN12815476
Brassicaceae	<i>Myagrum</i>	<i>perfoliatum</i> L.	2006-432	Marcus Koch	X	X	X	X		SAMN12815526
Brassicaceae	<i>Nasturtium</i>	<i>officinale</i> W.T.Aiton	AMC7.20.15	Chris Pires	X	X	X	X		SAMN12815527
Brassicaceae	<i>Olimarabidopsis</i>	<i>pumila</i> (Stephan) Al-Shehbaz, O'Kane & R.A.Price	BarbGH	Chris Pires	X					SAMN12815529
Brassicaceae	<i>Physaria</i>	<i>acutifolia</i> Rydb.	01201.13 ALPLAINS 50	Chris Pires	X	X	X	X		SAMN12815530
Brassicaceae	<i>Psychine</i>	<i>stylosa</i> Desf.	B2004-78	Marcus Koch	X	X	X	X		SAMN12815535
Brassicaceae	<i>Rorippa</i>	<i>islandica</i> (Oeder) Borbás	2006-89	Gomez Campo	X	X	X	X		SAMN12815537
Brassicaceae	<i>Schizopetalum</i>	<i>walkeri</i> Sims	NA	Martin Lysak	X	X	X	X		SAMN12815538
Brassicaceae	<i>Sinapis</i>	<i>alba</i> L.	0560-79	Gomez Campo	X	X	X	X	1000000	SAMN12815539
Brassicaceae	<i>Sisymbrium</i>	<i>brassiciforme</i> C.A.Mey	04-0372-10-00	Klaus Mummenhoff	X	X	X	X		SAMN12815540
Brassicaceae	<i>Sisymbrium</i>	<i>leucocladum</i> (Boiss.) D.A.German & Al-Shehbaz	3760-75	Gomez Campo	X	X	X	X		SAMN12815534
Brassicaceae	<i>Sisymbrium</i> L.	sp.	2004-557	Marcus Koch	X					SAMN12815541
Brassicaceae	<i>Streptanthus</i>	<i>arizonicus</i> S.Watson	02-0095-00-00	Klaus Mummenhoff	X	X	X	X		SAMN12815543

Brassicaceae	<i>Streptanthus</i>	<i>heterophyllus</i> Nutt.	1730-71	Gomez Campo	X	X	X	X		SAMN12815544
Brassicaceae	<i>Teesdalia</i>	<i>nudicaulis</i> (L.) W.T.Aiton	2219-73	Gomez Campo	X	X	X	X		SAMN12815546
Brassicaceae	<i>Thlaspi</i>	<i>arvense</i> L.	1211-67	Gomez Campo	X	X	X	X		SAMN12815547
Brassicaceae	<i>Turritis</i>	<i>glabra</i> L.	04-0314-10-00	Klaus Mummenhoff	X	X	X	X	4500000	SAMN12815548
Capparaceae	<i>Boscia</i> Lam.	sp.	SHN 12-075	Silverhill seeds and books	X	X	X	X		SAMN12815479
Capparaceae	<i>Cadaba</i>	<i>natalensis</i> Sond.	SHN 12-076	Silverhill seeds and books	X	X	X	X		SAMN12815480
Capparaceae	Capparaceae Juss.	sp.	SHN 12-079	Silverhill seeds and books	X	X	X	X		SAMN12815484
Capparaceae	<i>Capparis</i>	<i>fascicularis</i> DC.	SHN 12-078	Silverhill seeds and books	X	X	X	X		SAMN12815483
Cariaceae	<i>Carica</i>	<i>papaya</i> L.	NA	Chris Pires	X	X	X	X		SAMN12815487
Cleomaceae	<i>Arivela</i>	<i>viscosa</i> (L.) Raf.	KEW 104126	Kew Gardens	X	X	X	X		SAMN12815500
Cleomaceae	Cleomaceae Horan.	sp.	KEW 36588	Kew Gardens	X	X	X	X		SAMN12815497
Cleomaceae	<i>Cleome</i>	<i>africana</i> Botsch.	NA	Chris Pires	X	X	X	X		SAMN12815490
Cleomaceae	<i>Cleome</i>	<i>amblyocarpa</i> Barratte & Murb.	151485	Jocelyn Hall	X					SAMN12815491
Cleomaceae	<i>Cleome</i>	<i>arabica</i> L.	NA	Jocelyn Hall	X	X	X	X		SAMN12815493
Cleomaceae	<i>Cleome</i>	<i>violacea</i> L.	NA	Jocelyn Hall	X	X	X	X		SAMN12815498
Cleomaceae	<i>Cleomella</i>	<i>serrulata</i> (Pursh) Roalson & J.C.Hall	NA	Jocelyn Hall	X					SAMN12815501
Cleomaceae	<i>Coalisina</i>	<i>angustifolia</i> Raf.	KEW 82633	Kew Gardens	X	X	X	X	2000000	SAMN12815492
Cleomaceae	<i>Coalisina</i>	<i>paradoxa</i> (R.Br. ex DC.) Roalson & J.C.Hall	KEW 109051	Kew Gardens	X	X	X	X	1000000	SAMN12815496
Cleomaceae	<i>Gynandropsis</i>	<i>gynandra</i> (L.) Briq.	NA	Jocelyn Hall	X	X	X	X		SAMN12815513
Cleomaceae	<i>Melidiscus</i>	<i>giganteus</i> Raf.	NA	Jocelyn Hall	X	X	X	X		SAMN12815499
Cleomaceae	<i>Polanisia</i>	<i>dodecandra</i> (L.) DC.	2006-372	Marcus Koch	X	X				SAMN12815531
Cleomaceae	<i>Polanisia</i>	<i>graveolens</i> Raf.	2004 -157	Marcus Koch	X	X	X	X		SAMN12815532
Cleomaceae	<i>Polanisia</i> Raf.	sp.	KEW 93745	Kew Gardens	X	X	X	X		SAMN12815494

Cleomaceae	<i>Polanisia</i>	<i>trachysperma</i> Torr. & A.Gray	2006-365	Marcus Koch	X	X	X	X	1600000	SAMN12815533
Cleomaceae	<i>Sieruela</i>	<i>monophylla</i> (L.) Roalson & J.C.Hall	KEW 413163	Kew Gardens	X					SAMN12815495
Cleomaceae	<i>Tarenaya</i>	<i>hassleriana</i> (Chodat) Iltis	KEW 48516	Kew Gardens	X	X	X	X		SAMN12815545
Moringaceae	<i>Moringa</i>	<i>oleifera</i> Lam.	NA	Chris Pires	X	X	X	X	1000000	SAMN12815525
Resadaceae	<i>Ochradenus</i>	<i>baccatus</i> Delile	DMJ0324	Chris Pires	X	X	X	X		SAMN12815528
Resadaceae	<i>Reseda</i>	<i>odorata</i> L.	NA	Larry Chandler	X	X	X	X	1000000	SAMN12815536

Supplemental Table 2.2 RNA and DNA extraction method, library preparation method, sequencing method, read size, and raw read numbers.

Family	Genus	Species	RNA Extraction Kit	RNA Library Prep	RNA Seq Machine	RNA Read Length	RNA Raw Reads	DNA Extraction Kit	DNA Library Prep	DNA Seq Machine	DNA Read Length	DNA Raw Reads
Bataceae	<i>Batis</i>	<i>maritima</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	16565629	Qiagen DNeasy	TruSeq	NextSeq	2X150	8309055
Brassicaceae	<i>Aethionema</i>	<i>arabicum</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	55433848	Qiagen DNeasy	TruSeq	NextSeq	2X150	8757856
Brassicaceae	<i>Barbarea</i>	<i>bracteosa</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	53269435	Qiagen DNeasy	TruSeq	NextSeq	2X150	8979330
Brassicaceae	Brassicaceae	sp.	ThermoFisher PureLink RNA Mini Kit	TruSeq	HiSeq	2X100	40618374	Qiagen DNeasy	TruSeq	NextSeq	2X150	7179797
Brassicaceae	<i>Cakile</i>	<i>maritima</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	HiSeq	2X250	5555024	Qiagen DNeasy	TruSeq	NextSeq	2X150	10054961
Brassicaceae	<i>Calepina</i>	<i>irregularis</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	16367748	Qiagen DNeasy	TruSeq	NextSeq	2X150	9472593
Brassicaceae	<i>Capsella</i>	<i>bursa-pastoris</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	23770081	Qiagen DNeasy	TruSeq	NextSeq	2X150	9962200
Brassicaceae	<i>Cardamine</i>	<i>hirsuta</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	23276338	Qiagen DNeasy	TruSeq	NextSeq	2X150	7163619
Brassicaceae	<i>Caulanthus</i>	<i>amplexicaulis</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	HiSeq	2X100	30561937	Qiagen DNeasy	TruSeq	NextSeq	2X150	8638165
Brassicaceae	<i>Chorispora</i>	<i>tenella</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	19065280	Qiagen DNeasy	TruSeq	NextSeq	2X150	10713329
Brassicaceae	<i>Cochlearia</i>	<i>officinalis</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	19278016	Qiagen DNeasy	TruSeq	NextSeq	2X150	9785784
Brassicaceae	<i>Crambe</i>	<i>hispanica</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	HiSeq	2X100	6677521	Qiagen DNeasy	TruSeq	NextSeq	2X150	8981402
Brassicaceae	<i>Descurainia</i>	<i>pinnata</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	HiSeq	2X100	8315473	Qiagen DNeasy	TruSeq	NextSeq	2X150	10532085
Brassicaceae	<i>Descurainia</i>	<i>sophioides</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	HiSeq	2X100	41235763	Qiagen DNeasy	TruSeq	NextSeq	2X150	7149869
Brassicaceae	<i>Diptyocarpus</i>	<i>strictus</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	HiSeq	2X100	7422199					7488205
Brassicaceae	<i>Eruca</i>	<i>vesicaria</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	HiSeq	2X100	7743240	Qiagen DNeasy	TruSeq	NextSeq	2X150	9696734
Brassicaceae	<i>Erysimum</i>	<i>cheiranthoides</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	19481894	Qiagen DNeasy	TruSeq	NextSeq	2X150	9627950
Brassicaceae	<i>Euclidium</i>	<i>syriacum</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	HiSeq	2X100	7871790	Qiagen DNeasy	TruSeq	NextSeq	2X150	8499185
Brassicaceae	<i>Farsetia</i>	<i>aegyptia</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	49764414	Qiagen DNeasy	TruSeq	NextSeq	2X150	8514984
Brassicaceae	<i>Goldbachia</i>	<i>laevigata</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	16809960	Qiagen DNeasy	TruSeq	NextSeq	2X150	9849790
Brassicaceae	<i>Guillenia</i>	<i>lasiophylla</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	23503171	Qiagen DNeasy	TruSeq	NextSeq	2X150	10824454
Brassicaceae	<i>Hesperis</i>	<i>matronalis</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	51877498	Qiagen DNeasy	TruSeq	NextSeq	2X150	7807320
Brassicaceae	<i>Hirschfeldia</i>	<i>incana</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	18013197	Qiagen DNeasy	TruSeq	NextSeq	2X150	11527537
Brassicaceae	<i>Iberis</i>	<i>amara</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	HiSeq	2X100	7587536	Qiagen DNeasy	TruSeq	NextSeq	2X150	6845991
Brassicaceae	<i>Isatis</i>	<i>lusitanica</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	22464165	Qiagen DNeasy	TruSeq	NextSeq	2X150	11695349

Brassicaceae	<i>Isatis</i>	<i>tinctoria</i>	Qiagen RNEasy	TruSeq	HiSeq	2X100	20649046	Qiagen DNeasy	TruSeq	NextSeq	2X150	9446020
Brassicaceae	<i>Leavenworthia</i>	<i>uniflora</i>	NA	NA	NA	NA	NA	Qiagen DNeasy	TruSeq	NextSeq	2X150	9120176
Brassicaceae	<i>Lepidium</i>	<i>ruderales</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	18623760	Qiagen DNeasy	TruSeq	NextSeq	2X150	9459705
Brassicaceae	<i>Lepidium</i>	<i>sativum</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	HiSeq	2X100	7083621	Qiagen DNeasy	TruSeq	NextSeq	2X150	9451177
Brassicaceae	<i>Lobularia</i>	<i>maritima</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	21143830	Qiagen DNeasy	TruSeq	NextSeq	2X150	9578197
Brassicaceae	<i>Lunaria</i>	<i>annua</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	HiSeq	2X100	6184646	Qiagen DNeasy	TruSeq	NextSeq	2X150	7316600
Brassicaceae	<i>Malcolmia</i>	<i>maritima</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	HiSeq	2X100	7408880	Qiagen DNeasy	TruSeq	NextSeq	2X150	7908019
Brassicaceae	<i>Matthiola</i>	<i>longipetala</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	18984562	Qiagen DNeasy	TruSeq	NextSeq	2X150	10044883
Brassicaceae	<i>Meniocus</i>	<i>linifolius</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	51075472	Qiagen DNeasy	TruSeq	NextSeq	2X150	9512838
Brassicaceae	<i>Myagrum</i>	<i>perfoliatum</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	HiSeq	2X100	6922905	Qiagen DNeasy	TruSeq	NextSeq	2X150	6637717
Brassicaceae	<i>Nasturtium</i>	<i>officinale</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	52146124	Qiagen DNeasy	TruSeq	NextSeq	2X150	8184647
Brassicaceae	<i>Olimarabidopsis</i>	<i>pumila</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	19934340	NA	NA	NA	NA	NA
Brassicaceae	<i>Physaria</i>	<i>acutifolia</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	21947053	Qiagen DNeasy	TruSeq	NextSeq	2X150	9285650
Brassicaceae	<i>Psychine</i>	<i>stylosa</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	17163612	Qiagen DNeasy	TruSeq	NextSeq	2X150	9336172
Brassicaceae	<i>Rorippa</i>	<i>islandica</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	HiSeq	2X100	7662263	Qiagen DNeasy	TruSeq	NextSeq	2X150	7161563
Brassicaceae	<i>Schizopetalum</i>	<i>walkeri</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	16524348	Qiagen DNeasy	TruSeq	NextSeq	2X150	9736577
Brassicaceae	<i>Sinapis</i>	<i>alba</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	HiSeq	2X100	7602469	Qiagen DNeasy	TruSeq	NextSeq	2X150	9902350
Brassicaceae	<i>Sisymbrium</i>	<i>brassiciforme</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	22582931	Qiagen DNeasy	TruSeq	NextSeq	2X150	9078512
Brassicaceae	<i>Sisymbrium</i>	<i>leucocladum</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	15845037	Qiagen DNeasy	TruSeq	NextSeq	2X150	10653471
Brassicaceae	<i>Sisymbrium</i>	sp.	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	16438394	NA	NA	NA	NA	NA
Brassicaceae	<i>Streptanthus</i>	<i>arizonicus</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	17359223	Qiagen DNeasy	TruSeq	NextSeq	2X150	10016265
Brassicaceae	<i>Streptanthus</i>	<i>heterophyllus</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	17818482	Qiagen DNeasy	TruSeq	NextSeq	2X150	10220458
Brassicaceae	<i>Teesdalia</i>	<i>nudicaulis</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	59723745	Qiagen DNeasy	TruSeq	NextSeq	2X150	8579441
Brassicaceae	<i>Thlaspi</i>	<i>arvense</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	HiSeq	2X100	7651483	Qiagen DNeasy	TruSeq	NextSeq	2X150	6871390
Brassicaceae	<i>Turritis</i>	<i>glabra</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	15406084	Qiagen DNeasy	TruSeq	NextSeq	2X150	10860792
Capparaceae	<i>Boscia</i>	sp.	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	17980307	Qiagen DNeasy	TruSeq	NextSeq	2X150	10870502
Capparaceae	<i>Cadaba</i>	<i>natalensis</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	18255177	Qiagen DNeasy	TruSeq	NextSeq	2X150	9366417

Capparaceae	Capparaceae	sp.	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	17848365	Qiagen DNeasy	TruSeq	NextSeq	2X150	9425035
Capparaceae	<i>Capparis</i>	<i>fascicularis</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	18084217	Qiagen DNeasy	TruSeq	NextSeq	2X150	10378771
Cariaceae	<i>Carica</i>	<i>papaya</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	18870089	Qiagen DNeasy	TruSeq	NextSeq	2X150	13335392
Cleomaceae	<i>Arivela</i>	<i>viscosa</i>	Qiagen RNEasy	TruSeq	HiSeq	2X100	23245927	Qiagen DNeasy	TruSeq	NextSeq	2X150	9098547
Cleomaceae	Cleomaceae	sp.	Qiagen RNEasy	TruSeq	HiSeq	2X100	25851821	Qiagen DNeasy	TruSeq	NextSeq	2X150	10240576
Cleomaceae	<i>Cleome</i>	<i>africana</i>	Qiagen RNEasy	TruSeq	HiSeq	2X100	21205456	Qiagen DNeasy	TruSeq	NextSeq	2X150	7267839
Cleomaceae	<i>Cleome</i>	<i>amblyocarpa</i>	Qiagen RNEasy	TruSeq	HiSeq	2X100	21519053	NA	NA	NA	NA	NA
Cleomaceae	<i>Cleome</i>	<i>arabica</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	14426573	Qiagen DNeasy	TruSeq	NextSeq	2X150	7960574
Cleomaceae	<i>Cleome</i>	<i>violacea</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	HiSeq	2X250	6720824	Qiagen DNeasy	TruSeq	NextSeq	2X150	7683480
Cleomaceae	<i>Cleomella</i>	<i>serrulata</i>	Qiagen RNEasy	TruSeq				NA	NA	NA	NA	NA
Cleomaceae	<i>Coalisina</i>	<i>angustifolia</i>	Qiagen RNEasy	TruSeq	HiSeq	2X100	19351183	Qiagen DNeasy	TruSeq	NextSeq	2X150	8842965
Cleomaceae	<i>Coalisina</i>	<i>paradoxa</i>	Qiagen RNEasy	TruSeq	HiSeq	2X100	21871851	Qiagen DNeasy	TruSeq	NextSeq	2X150	10046879
Cleomaceae	<i>Gynandropsis</i>	<i>gynandra</i>	Qiagen RNEasy	TruSeq	HiSeq	2X100	19102136	Qiagen DNeasy	TruSeq	NextSeq	2X150	9236435
Cleomaceae	<i>Melidiscus</i>	<i>giganteus</i>	Qiagen RNEasy	TruSeq	HiSeq	2X100	19954871	Qiagen DNeasy	TruSeq	NextSeq	2X150	9018300
Cleomaceae	<i>Polanisia</i>	<i>dodecandra</i>	Qiagen RNEasy	TruSeq	HiSeq	2X100	44674939	Qiagen DNeasy	TruSeq	NextSeq	2X150	8796223
Cleomaceae	<i>Polanisia</i>	<i>graveolens</i>	Qiagen RNEasy	TruSeq	HiSeq	2X100	25106398	Qiagen DNeasy	TruSeq	NextSeq	2X150	10218176
Cleomaceae	<i>Polanisia</i>	sp.	Qiagen RNEasy	TruSeq	HiSeq	2X100	25823291	Qiagen DNeasy	TruSeq	NextSeq	2X150	8242370
Cleomaceae	<i>Polanisia</i>	<i>trachysperma</i>	Qiagen RNEasy	TruSeq	HiSeq	2X100	38182441	Qiagen DNeasy	TruSeq	NextSeq	2X150	9082194
Cleomaceae	<i>Sieruela</i>	<i>monophylla</i>	Qiagen RNEasy	TruSeq	HiSeq	2X100	25080383	NA	NA	NA	NA	NA
Cleomaceae	<i>Tarenaya</i>	<i>hassleriana</i>	Qiagen RNEasy	TruSeq	HiSeq	2X100	19162929	Qiagen DNeasy	TruSeq	NextSeq	2X150	12209330
Moringaceae	<i>Moringa</i>	<i>oleifera</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	57179455	Qiagen DNeasy	TruSeq	NextSeq	2X150	10239447
Resadaceae	<i>Ochradenus</i>	<i>baccatus</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	18868053	Qiagen DNeasy	TruSeq	NextSeq	2X150	9340353
Resadaceae	<i>Reseda</i>	<i>ordorata</i>	ThermoFisher PureLink RNA Mini Kit	TruSeq	NextSeq	2X75	18754089	Qiagen DNeasy	TruSeq	NextSeq	2X150	10029969

Supplemental Table 2.3 Orthogroups retained for each analysis.

Group	# of genes assigned to groups	% assigned	OrthoGroups	G50	O50	# of groups with all species	Single-copy OrthoGroups	Taxon Occupancy (80%)	Alignment Quality (40% gaps)	# of Trees after first round of RAxML	Tree Pruning (cutoff #)	Final tree # for species tree inference
Brassicales	3516602	96.3	47600	252	4630	6444	0	59/74 = 10968	2663	2663	1284 (10)	1284
Brassicaceae	2224281	96.7	39809	135	5160	7822	0	39/48 = 12129	5587	5587	2110 (10)	2110
Capparaceae	143114	79.4	25179	5	9600	12350	3228	3/4 = 17571	15988	14338	13328 (3)	10214
Cleomaceae	883407	91.9	42218	37	7662	9706	55	13/17 = 14107	10732	10732	3626 (10)	3626
RBMC	165057	78.7	21372	8	7447	9098	881	4/5 = 11756	10038	10038	8476 (4)	8476

Supplemental Table 2.4 BIC scores for 1-4 components for both FASTKs (McKain et al., 2016) and DupPipe (Barker et al., 2010) Ks plots.

Species Information			DupPipe BIC Scores				FASTKs BIC Scores			
Family	Genus	Species	1 peak	2 peaks	3 peaks	4 peaks	1 peak	2 peaks	3 peaks	4 peaks
Bataceae	<i>Batis</i>	<i>maritima</i>	-6273.05	-4970.69	-4612.552	-4515.273	-29659.81	-29227.93	-28742.21	-28355.45
Brassicaceae	<i>Aethionema</i>	<i>arabicum</i>	-17888.73	-15930.41	-15712.96	-15489.97	-72664.46	-71361.21	-70739.96	-70405.8
Brassicaceae	<i>Barbarea</i>	<i>bracteosa</i>	-20416.34	-18389.58	-18047.26	-17929.21	-55743.7	-54258.87	-53867.64	-53868.36
Brassicaceae	<i>Cakile</i>	<i>maritima</i>	-49327.51	-48717.18	-48587.63	-48384.6	-140948	-139956.7	-139670.2	-139216.5
Brassicaceae	<i>Calepina</i>	<i>irregularis</i>	-34320.8	-30254.13	-29871.31	-29871.35	-67744.52	-64841.66	-63941.85	-63706.6
Brassicaceae	<i>Capsella</i>	<i>bursa-pastoris</i>	-9387.535	-8388.332	-8312.341	-8217.367	-51871.77	-50359.75	-49734.31	-49607.75
Brassicaceae	<i>Cardamine</i>	<i>hirsuta</i>	-18046.43	-15950.79	-15655.47	-15519.12	-27724.56	-27082.3	-26877.96	-26843.1
Brassicaceae	<i>Caulanthus</i>	<i>amplexicaulis</i>	-72240.47	-66365.55	-65466.22	-65312.87	-296250.7	-288865.5	-286690	-286481.9
Brassicaceae	<i>Chorispora</i>	<i>tenella</i>	-14681.18	-13104.49	-12870.93	-12797.43	-27157.47	-26754.78	-26675.19	-26592.95
Brassicaceae	<i>Cochlearia</i>	<i>officinalis</i>	-26835.57	-25203.31	-24977.02	-24724.24	-45997.75	-45297.29	-44895.66	-44896.44
Brassicaceae	<i>Crambe</i>	<i>hispanica</i>	-53282.57	-52870.1	-52509.33	-52201.88	-162684.3	-161909.2	-161518.4	-160916.1
Brassicaceae	<i>Descurainia</i>	<i>sophioides</i>	-20134.61	-17945.48	-17617.56	-17430.33	-78315.54	-76328.78	-74915.19	-74615.81
Brassicaceae	<i>Descurainia</i>	<i>pinnata</i>	-32177.51	-29224.79	-28697.38	-28298.75	-139705.5	-136187.3	-134855.7	-134853.8
Brassicaceae	<i>Diptycocarpus</i>	<i>strictus</i>	-20176.44	-18385.49	-18151.68	-18005.14	-51593.24	-49629.33	-49558.3	-49397.94
Brassicaceae	<i>Eruca</i>	<i>vesicaria</i>	-57302.99	-56279.26	-56171.14	-55940.44	-180363.8	-178936.7	-178670.7	-178002.3
Brassicaceae	<i>Erysimum</i>	<i>cheiranthoides</i>	-17035.26	-15470.34	-15201.41	-15048.28	-27870.51	-27251.97	-27169.5	-27167.07
Brassicaceae	<i>Euclidium</i>	<i>syriacum</i>	-7330.347	-6664.564	-6569.126	-6486.873	-75457	-74132.51	-74132.3	-73955.92
Brassicaceae	<i>Farsetia</i>	<i>aegyptia</i>	-87046.73	-78626.97	-77752.59	-77752.63	-27584.64	-26808.89	-26648.29	-26461.84
Brassicaceae	<i>Goldbachia</i>	<i>laevigata</i>	-46724.37	-40469.43	-39908.4	-39824.05	-126148.9	-120596.6	-119013.2	-118631.9
Brassicaceae	<i>Guillenia</i>	<i>lasiophylla</i>	-59735.11	-53526.87	-52969.1	-52857.97	-134706.6	-128952.9	-128027.7	-127946
Brassicaceae	<i>Hesperis</i>	<i>matronalis</i>	-8981.003	-8402.243	-8296.301	-8279.953	-116152.4	-114019.3	-113171.9	-113171.2
Brassicaceae	<i>Hirschfeldia</i>	<i>incana</i>	-57478.96	-56124.2	-56066.99	-55528.43	-129935.9	-129659.6	-129543	-128886.2
Brassicaceae	<i>Iberis</i>	<i>amara</i>	-37998.83	-36103.28	-35797.37	-35513.78	-104300.9	-102190.1	-101688.1	-101659.2

Brassicaceae	<i>Isatis</i>	<i>lusitanica</i>	-16637.79	-14632.94	-14324.95	-14221.36	-27652.87	-26943.13	-26871.77	-26692.73
Brassicaceae	<i>Isatis</i>	<i>tinctoria</i>	-32807.62	-28987.6	-28672.49	-28416.86	-64911.3	-62615.86	-61966.98	-61638.4
Brassicaceae	<i>Lepidium</i>	<i>ruderales</i>	-51478.63	-43198.8	-42721.1	-42559.56	-125387.6	-118218.6	-117136.5	-116627.7
Brassicaceae	<i>Lepidium</i>	<i>sativum</i>	-63920.94	-57354.73	-56523.84	-56506.45	-235772.5	-226791.3	-224613.5	-224338.9
Brassicaceae	<i>Lobularia</i>	<i>maritima</i>	-22482.68	-20460.54	-20234	-20233.39	-35962.92	-35163.19	-34845.46	-34845.36
Brassicaceae	<i>Lunaria</i>	<i>annua</i>	-49086.82	-48726.38	-48391.74	-48324.15	-156452.4	-156020.2	-155609.5	-155183.2
Brassicaceae	<i>Malcolmia</i>	<i>maritima</i>	-31503.91	-28447.88	-28044.13	-27721.6	-104663.4	-100716.8	-100089.1	-99712.58
Brassicaceae	<i>Matthiola</i>	<i>longipetala</i>	-18787.75	-16630.62	-16436.26	-16259.59	-28755.13	-28074.03	-27838.22	-27791.81
Brassicaceae	<i>Meniocus</i>	<i>linifolius</i>	-93350.47	-79319.88	-78028.4	-77924.67	-293255.5	-277515.4	-274443.6	-273907.6
Brassicaceae	<i>Myagrum</i>	<i>perfoliatum</i>	-18501.89	-16611.93	-16246.46	-16074.25	-46087.16	-44153.4	-43545.81	-43274.85
Brassicaceae	<i>Nasturtium</i>	<i>officinale</i>	-60471.95	-52138.19	-51393.83	-51165.92	-223823.9	-213019.7	-211035.3	-210681.9
Brassicaceae	<i>Olimarabidopsis</i>	<i>pumila</i>	-39183.53	-33871.53	-33459.93	-33419.88	-66715.53	-63165.63	-62249.17	-62076.87
Brassicaceae	<i>Physaria</i>	<i>acutifolia</i>	-41432.66	-40148.94	-40051.83	-39542.31	-100664.9	-100252.2	-100064.3	-99552.98
Brassicaceae	<i>Psychine</i>	<i>stylosa</i>	-42292.4	-41469.83	-41414.98	-41051.42	-109069.2	-108968.6	-108891.8	-108786.9
Brassicaceae	<i>Rorippa</i>	<i>islandica</i>	-22644.09	-20601.65	-20221.54	-20067.94	-53588.67	-51903.39	-51623.81	-51592.38
Brassicaceae	<i>Schizopetalum</i>	<i>walkeri</i>	-42672.93	-41045.12	-40688.12	-40594.02	-134672.9	-133861.7	-132864.5	-132864.6
Brassicaceae	<i>Sinapis</i>	<i>alba</i>	-58678.87	-58381.11	-57891.62	-57599.67	-214304.1	-213673.6	-213076.5	-212988.6
Brassicaceae	<i>Sisymbrium</i>	<i>brassiciforme</i>	-56912.82	-51040.69	-50499.11	-50441.32	-106873.4	-106872.7	-106631	-106195.7
Brassicaceae	<i>Sisymbrium</i>	<i>leucocladum</i>	-18171.84	-16346.14	-16073.08	-15879.77	-29525.35	-28937.77	-28532.62	-28429.9
Brassicaceae	<i>Sisymbrium</i>	sp.	-45927.74	-44982.75	-44902.77	-44505.49	-48002.97	-45832.08	-45213.05	-44905.1
Brassicaceae	Brassicaceae	sp.	-92159.93	-85720.01	-84367.26	-84131.4	-397328.4	-390193.2	-386693.6	-385884.6
Brassicaceae	<i>Streptanthus</i>	<i>arizonicus</i>	-44070.05	-41193.34	-40805.58	-40760.63	-92667.5	-89997.98	-89422.72	-89310.5
Brassicaceae	<i>Streptanthus</i>	<i>heterophyllus</i>	-47146.15	-43765.39	-43355.83	-43327.92	-98277.5	-95744.79	-94996.69	-94933.27
Brassicaceae	<i>Teesdalia</i>	<i>nudicaulis</i>	-70434.89	-62882.51	-41633.38	-61916.37	-210073	-202074	-199908.6	-199724.9
Brassicaceae	<i>Thlaspi</i>	<i>arvense</i>	-21362.18	-19322.8	-18939.83	-18772.49	-54264.15	-53094.77	-53083.27	-52492.75
Brassicaceae	<i>Turritis</i>	<i>glabra</i>	-49465.03	-42282.28	-41683.33	-41633.38	-92297.93	-86449.15	-85442.42	-85345.57

Capparaceae	<i>Boscia</i>	sp.	-15462.33	-13656.09	-12905.13	-12735.44	-57128.34	-55238.59	-54170.86	-54008.21
Capparaceae	<i>Cadaba</i>	<i>natalensis</i>	-10654.85	-8824.122	-8174.672	-8072.744	-47515.81	-45670.41	-45094.29	-44846.13
Capparaceae	<i>Capparis</i>	<i>fascicularis</i>	-13851.04	-12299.53	-11690.69	-11549.21	-43810.82	-42563.49	-41808.2	-41170.72
Capparaceae	Capparaceae	sp.	-14017.43	-12372.35	-11778.9	-11641.74	-55875.39	-54956.17	-54456.94	-54060.24
Cariaceae	<i>Carica</i>	<i>papaya</i>	-12964.8	-10810.21	-10361.01	-10318.83	-37739.28	-36667.15	-35950.16	-35652.17
Cleomaceae	Cleomaceae	sp.	-32790.84	-31907.16	-30767.01	-30457.21	-121041.3	-120137.4	-119236.3	-118223
Cleomaceae	<i>Cleome</i>	<i>africana</i>	-12033.86	-10203.68	-9700.341	-9570.846	-54351.44	-53335.11	-53115.26	-52575.36
Cleomaceae	<i>Cleome</i>	<i>amblyocarpa</i>	-10307.71	-8534.686	-8097.695	-8023.899	-39832.76	-38657.31	-37793.17	-37665.87
Cleomaceae	<i>Cleome</i>	<i>arabica</i>	-7773.53	-6276.361	-5974.505	-5878.58	-15333.58	-14779.72	-14518.46	-14386.6
Cleomaceae	<i>Sieruela</i>	<i>monophylla</i>	-34499.78	-33835.05	-32662.68	-32407.19	-125206.2	-124041.9	-124342.1	-123520.9
Cleomaceae	<i>Cleome</i>	<i>violacea</i>	-14478.17	-12665.1	-12143.09	-12031.81	-49680.57	-48555.81	-48077.79	-48054.83
Cleomaceae	<i>Melidiscus</i>	<i>giganteus</i>	-39082.77	-38144.81	-37158.06	-36845.59	-208644	-207264.7	-206914.5	-204958.6
Cleomaceae	<i>Arivela</i>	<i>viscosa</i>	-12125.13	-10601.82	-10094.69	-9995.74	-46372.67	-44884.62	-44811.79	-44461.78
Cleomaceae	<i>Polanisia</i>	sp.	-11321.78	-9636.747	-9186.818	-9072.467	-46405.07	-45258.65	-44304.96	-43969.71
Cleomaceae	<i>Cleomella</i>	<i>serrulata</i>	-35545.31	-29633.1	-28966.29	-28500.9	-104059.2	-98769.64	-97286.63	-96448.16
Cleomaceae	<i>Coalisina</i>	<i>angustifolia</i>	-32947.1	-32573.87	-31485.88	-31275.91	-70695.18	-70690.09	-70254.17	-69824.07
Cleomaceae	<i>Coalisina</i>	<i>paradoxa</i>	-45113.69	-43013.85	-42454.3	-42123.13	-232073.7	-230283.9	-228074.1	-226942.3
Cleomaceae	<i>Gynandropsis</i>	<i>gynandra</i>	-26445.88	-25136.17	-24368.74	-24136.14	-105952.6	-104635	-103678.3	-103482.8
Cleomaceae	<i>Polanisia</i>	<i>dodecandra</i>	-15250.73	-13157.53	-12519.87	-12362.36	-62510.34	-61483.52	-60632.61	-60304.93
Cleomaceae	<i>Polanisia</i>	<i>graveolens</i>	-11921.91	-9855.295	-9345.755	-9212.286	-33712.79	-32733.65	-32274.19	-31957.47
Cleomaceae	<i>Polanisia</i>	<i>trachysperma</i>	-13878.81	-11855.59	-11252.52	-11116.56	-56577.94	-55015.23	-54726.4	-54212.57
Cleomaceae	<i>Tarenaya</i>	<i>hassleriana</i>	-34991.25	-34328.86	-33198.46	-33171.58	-113691.1	-113078.4	-112644.8	-112107.3
Moringaceae	<i>Moringa</i>	<i>oleifera</i>	-8975.359	-6567.789	-6563.584	-6244.213	-19414.45	-18830.46	-18389.46	-18135.18
Resadaceae	<i>Ochradenus</i>	<i>baccatus</i>	-30000.1	-28976.74	-28050.13	-27748.79	-65666.52	-64702.93	-64643.43	-63693.03
Resadaceae	<i>Reseda</i>	<i>odorata</i>	-29667.79	-28450.24	-27615.78	-27382.23	-68356.46	-67711.71	-67638.78	-67120.38

Brassicales BUSCO Assessment Results

Supplemental Figure 2.2 BUSCO analysis of *de novo* transcriptomes. Legend indicates the percent of genes that are complete and single copy (light blue), complete and duplicate (dark blue), fragmented (yellow), and missing (red) in *de novo* transcriptomes

Supplemental Figure 2.4 Maximum likelihood whole-chloroplast phylogeny of the Brassicales. Support values are indicated if below 70% bootstrap support.

Supplemental Figure 2.5 Ks plots of the Brassicaceae using both FASTKs (McKain et al., 2016) and DupPipe (Barker et al., 2010). Whole-genome duplication events, At-a and the Brassicaceae triplication (T) event, are noted above corresponding peaks.

Supplemental Figure 2.6 Ortholog divergences and Ks peaks of the Cleomaceae. **(A)** Ortholog divergences between *C. amblyocarpa* and *C. africana*, *C. arabica*, and *C. violacea* and between *C. violacea* and *C. africana* to test placement of potential novel WGD event. **(B)** Ortholog divergences between *Polanisia* sp. and *C. violacea*, *Polanisia* sp., and *P. dodecandra*, *P. trachysperma*, and *Polanisia* sp. and between *P. trachysperma* and *P. dodecandra* to test for placement of the second potential novel WGD event.

Supplemental Figure 2.7 Ortholog divergences and Ks peaks of the (A) Capparaceae and (B) Resedaceae + Outgroups. Proposed Resedaceae whole-genome duplication event indicated.

**CHAPTER 3 : THE EVOLUTIONARY HISTORY OF WILD AND
DOMESTICATED, AND FERAL *BRASSICA OLERACEA*
(BRASSICACEAE)**

Makenzie E. Mabry¹, Sarah D. T. Hissong², Evan Y. Gallagher¹, Alex C. McAlvay³,
Patrick P. Edger⁴, Jonathan D. Moore⁵, Peter G. Walley⁶, Graham R. Teakle⁶,
Paul Hand⁷, David A. C. Pink⁷, Chris Stevens⁸, Guy Barker⁶, Hong An¹, Joanne Labate⁹,
Dorian Q. Fuller⁸, Robin G. Allaby⁶, Jared E. Decker¹⁰, Michael A. Gore¹¹,
J. Chris Pires¹

1 Division of Biological Sciences and Christopher S. Bond Life Sciences Center,
University of Missouri, Columbia, MO, U.S.A.

2 Department of Plant Biology, University of California Davis, Davis, CA, U.S.A.

3 Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY, USA

4 Department of Horticulture, Michigan State University, East Lansing, MI, U.S.A.

5 Systems Biology Centre, University of Warwick, UK

6 School of Life Science, University of Warwick, UK

7 Harper Adams University, UK

8 Institute of Archaeology, University College London, UK

9 USDA-ARS, Plant Genetic Resources Unit, Cornell University, Geneva, NY, U.S.A.

10 Division of Animal Sciences, University of Missouri, Columbia, USA.

11 Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell
University, Ithaca, NY, U.S.A.

ABSTRACT

Brassica oleracea is an important crop species that has intrigued researchers for centuries, including Darwin, due to its diversity in forms. The practice of artificial selection has resulted in several crop types including broccoli, Brussels sprouts, cabbage, kale, and kohlrabi, as well as several lesser well-known types, such as walking stick kale and marrow cabbage. However, over a century and a half after Darwin, we still know very little about the origins of this crop species. To date, the wild progenitor(s) and geographic origin have remained uncertain, with conflicting hypotheses supporting either a Sicilian or an English origin. Using a diversity panel of 224 accessions representing 14 different *B. oleracea* crop types and nine potential wild progenitor species, we integrate phylogenetic and population genetic techniques to examine patterns of relationships among cultivars and wild relatives. These analyses point to *B. cretica* as the closest living relative of *B. oleracea*, supporting an origin of cultivation in the Eastern Mediterranean region. Understanding the evolutionary history of crops, including identifying crop wild relatives may point to new approaches for future crop breeding efforts, especially in light of climate change.

INTRODUCTION

“Greek legend has it that the cabbage sprung from where Zeus' sweat hit the ground.”

-N. D. Mitchell (1976)

For centuries, the horticultural crop *Brassica oleracea* L. has played an important role in the global food systems, both as a source of leaf and root vegetables and of fodder and forage (Shyam et al. 2012). When first introduced to the species, Darwin drew many parallels between his theory on natural selection and the cultivation practices that led to the varied forms of this plant. Just as dogs have been domesticated into many popular breeds like the Chihuahua or the Golden Retriever, *B. oleracea* has been domesticated via artificial selection into several different popular vegetables, including broccoli, Brussels sprouts, cabbage, cauliflower, kale, and kohlrabi. One uniting characteristic of these vegetable crops is the presence of glucosinolates (GSLs), bitter tasting compounds that are useful to the plant for their herbivory defense and anti-carcinogenic properties (Traka et al. 2013). Although these six major crop types comprise the majority of the U.S. market (Agricultural Marketing Service, Market News Reports; <http://www.ams.usda.gov>), there are at least an additional 12 varieties (**Figure 3.1**) that are also cultivated. These varieties include Chinese white kale or Gai-lan, ornamental kale, which is used as sidewalk décor, romanesco with its beautiful fractal patterns, and the walking stick kale found on the Canary Islands. As a single species, *B. oleracea* encompasses a morphological diversity greater than that of some entire plant genera.

Figure 3.1 *Brassica oleracea* wild relative and cultivar diversity. **A)** *B. cretica*, **B)** *B. hilarionis*, **C)** *B. incana*, **D)** *B. insularis*, **E)** *B. macrocarpa*, **F)** *B. montana*, **G)** *B. oleracea*, **H)** *B. rupestris*, **I)** *B. villosa*, **J)** broccoli, **K)** Brussels sprouts, **L)** cauliflower, **M)** Chinese white kale, **N)** collards, **O)** curly kale, **P)** walking stick kale, **Q)** kohlrabi, **R)** lacinato kale, **S)** marrow-stem kale, **T)** ornamental kale, **U)** perpetual kale, **V)** red cabbage, **W)** romanesco, **X)** savoy cabbage, **Y)** tronchuda kale, **Z)** cabbage, **AA)** cone cabbage.

Surprisingly, little is known about the domestication center or progenitor species of this important crop, although several hypotheses have been proposed. Many authors refer to these cultivated crops along with their wild relatives as a single cytodeme, as they all have the same number of chromosomes ($2n = 18$) with a similar genomic organization. Those wild relatives include *B. cretica*, *B. hilarionis*, *B. incana*, *B. insularis*, *B. macrocarpa*, *B. montana*, *B. oleracea*, *B. rupestris*, and *B. villosa* (**Figure 3.1**). Along with uncertainty in domestication origins, there is also confusion in the taxonomy of this group of plants. L. H. Bailey, an American horticulturist, botanist, and cofounder of the American Society for Horticultural Science stated that “*Some of these plants appear to be more confused in literature than in nature*” (The Cultivated Brassicas Second Paper- Gentes

Herbarum; 1930). Although we note that several taxonomic schemes have been proposed for this group, here we elect to treat the wild relatives as individual species.

Hypotheses on *B. oleracea*'s origin include both a single and multiple domestication events (Neutrofal 1927; Schiemann 1932; Shultz 1936; Lizgunova 1959; Helm 1963; Snogerup 1980; Heaney et al. 1987; Song et al. 1988; Song et al. 1990; Swarup & Brahmi 2005; Allender et al. 2007). Different studies that have pointed to a single origin of domestication have nonetheless identified several different wild species suggested as either the progenitor species or closest living relative (Snogerup, Gustafsson, and Von Bothmer 1990; Song et al. 1990; Hodgkin 1995; Maggioni et al. 2018). Some authors suggest an origin in England, as there are weedy, cabbage-like plants found along the Atlantic coast there. These wild *B. oleracea* plants, which are actually found along the coasts of all of western Europe (England, France, Spain, and Portugal) have been described in the literature as either a progenitor species (Snogerup et al. 1990; Song et al. 1990; Hodgkin 1995) or as escaped populations of cultivated plants (Mitchell 1976; Michell & Richards 1979). Hodgkin (1995) hypothesized that the initial domestication took place on the Atlantic coast and early cultivated forms were brought from there to the Mediterranean, where selection for many of the early crop types occurred. Most recently, using linguistic and literary considerations, a Mediterranean domestication has been proposed, given that both early Greek and Latin literature is rich in expressions indicating a deep-rooted knowledge and use of the *B. oleracea* crops (Maggioni et al. 2018).

Due to the vast amount of phenotypic diversity currently displayed in cultivars, multiple events have been proposed to best explain the domestication of *B. oleracea*. Neutrofal (1927) thought that *B. montana* was the progenitor to cabbages and *B. rupestris*

was the progenitor to kohlrabi, while Schulz (1936) identified *B. cretica* as the progenitor of cauliflower and broccoli. Helm (1963) proposed a triple origin in which cauliflower, broccoli, and sprouting broccoli were derived from a single progenitor species, thousand-headed kale and Brussels sprouts from another species, and all other crop forms from a third. Snogerup (1980) thought that while headed cabbages were derived from *B. oleracea* and stem kales were from both *B. rupestris* and *B. incana*, Chinese white kale was derived specifically from *B. cretica* subsp. *niva*. Others proposed that since Sicily had a large diversity of wild relatives, that Sicily was the center of domestication (Schiemann 1932; Lizgunova 1959), agreeing with Vavilov (1951), who argued that plants were not domesticated somewhere random in the world, but that the center of diversity should also be considered the center of origin.

Due to the lack of both genetic and archaeological evidence, the evolutionary history of this economically important species has been left untold. Using a diversity panel of 14 cultivar types and nine wild relatives, together the largest and most diverse collection of this species and its wild relatives, we integrate phylogenomics, population genomics, species distribution modeling, archaeological, and literary sources to identify the closest living wild relative, investigate the origin of domestication, and highlight feral populations of *B. oleracea*.

RESULTS

Species distribution models suitability of potential closest living wild relatives

To tease apart potential closest living wild relatives of domesticated *B. oleracea* types, species distributions were modeled for each potential wild relative during the mid-

Holocene (7,000-5,000 years ago), the estimated time of domestication (**Figure 3.2**). We compiled occurrence records from the GBIF (www.gbif.org) data portal and from Snogerup et al. (1990). In addition to occurrence records, nineteen bioclimate variables at 2.4 minute resolution in conjunction with the Climate Systems Model 4 (Meehl et al. 2012) from WorldClim 1.4 (Hijmans et al. 2005) were used to generate species distribution estimated in MaxEnt 3.4.1 (Phillips et al. 2017). The omission rate on training samples was found to be close to predicted omission indicating good model fit and the receiver operating characteristic curve showed an area under curve of 0.943 for *B. incana*, 0.949 for *B. cretica*, 0.888 for *B. macrocarpa*, 0.981 for *B. villosa*, 0.999 for *B. hilarionis*, 0.947 for *B. montana*, 0.999 for *B. rupestris*, 0.957 for *B. insularis*.

Plant samples, phylogeny, and population clustering suggest some wild relatives are not “wild”

Sampling of *B. oleracea* cultivars included seven types of kales (chinese white kale; 22, tronchuda kale; 13, lacinato kale; 8, curly kale; 12, perpetual kale; 13, walking stick kale; 2, and collard greens; 16), three types of cabbages (white cabbage; 15, savoy cabbage; 4, and marrow cabbage; 13), Brussels sprouts (17), broccoli (20), cauliflower (18), and kohlrabi (15; **Supplemental Table 3.1**). Together these cultivated types accounted for 189 of the 224 total samples. The remainder of the 224 samples included previously identified wild relatives: (putatively wild) *B. oleracea* (3), *B. cretica* (3), *B. incana* (5), *B. montana* (2), *B. hilarionis* (1), *B. insularis* (5), *B. macrocarpa* (4), *B. rupestris* (7), and *B. villosa* (5). RNA from these samples was sequenced across 10 lanes of a NextSeq flowcell using 2 X 75 paired-end Illumina sequencing. Reads were then mapped to the *B. oleracea*

TO1000 genome (Chinese kale; Parkin et al. 2014; release-41), resulting in a final dataset of 36,750 SNPs after filtering.

Figure 3.2 Species distribution models. **A)** Current species distribution of wild relatives. Mid-Holocene estimated species distribution for **B)** *B. cretica*, **C)** *B. hilarionis*, **D)** *B. incana*, **E)** *B. insularis*, **F)** *B. macrocarpa*, **G)** *B. montana*, **H)** *B. rupestris*, **I)** *B. villosa*.

Reconstruction of the phylogeny for all samples recovered *B. incana* as sister to a large clade of cultivated samples. However, within this clade we recover an additional group of wild samples consisting of three species, *B. incana* (205, 208, 209), one sample of *B. villosa* (233), and one sample of *B. cretica* (195). We refer to this as wild C - clade 2. We also find samples of *B. cretica* (196, 199), *B. montana* (222), and putatively wild *B. oleracea* (175, 176, 177) placed throughout the cultivated clade (samples in red; **Figure 3.3A**). Within the cultivar clade, Chinese white kale, broccoli, cauliflower, romanesco, kohlrabi, curly kale, and Brussels sprouts were recovered as monophyletic assemblages with 70% bootstrap support or better. Cabbages were also found as monophyletic, but with only 55% bootstrap support. In addition to the wild C - clade 2 clustering separately from other sampled wild relatives, these samples also showed differing proportions of shared ancestry (**Figure 3.3B**). At $K = 2$ in fastSTRUCTURE analyses, where the marginal likelihood is maximized, samples cluster as either wild or cultivated with wild C - clade 2 showing a significant amount of shared ancestry with other wild samples as does one sample of tronchuda kale (*costata*; sample 030). Other samples of wild relatives, including the two samples of *B. incana* (204 and 207) that are found sister to all cultivated samples share 100% of their ancestry with cultivated types, rather than any wild types, perhaps indicating that these are not truly wild samples. At $K = 3$, clades of broccoli (var. *italica*), cauliflower (var. *botrytis*), and Chinese white kale (var. *alboglabra*) separate from other cultivated types. Finally, at $K = 4$, white kale (var. *alboglabra*) separates from broccoli (var. *italica*) and cauliflower (var. *botrytis*).

Figure 3.3 Sample demographics. **A)** Individual sample phylogeny and **B)** population ancestry (Shown as a fastSTRUCTURE diagram for K = 2 to K = 4). Sample names in red = putatively wild samples. K=2 maximizes marginal likelihood (+). K=3 explains structure in data best (++). Asterisks (*) next to names indicate those samples that were recovered as 70% membership to one group or another in K=2 and used for further analyses. **C)** PCA of PC1 vs PC2 of wild and cultivar samples. Triangles = wild samples, circles = cultivars. Black triangles with colored outlines = wild C - clade 2 samples with species identification indicated by outline color. **D)** Species tree of wild and cultivar samples. Bootstrap support indicated above branches if below 70%.

PCA and Phylogenetic patterns highlight closest living wild relatives

Principal components analyses (PCA) of wild relatives and cultivars separates cultivars from most wild types (**Figure 3.3C**). PC1 separates wild types from cultivars and PC2 separates wild C - clade 2 from all other wild C species (indicated as black triangles with species identification as outlines). PC3 and PC4 show more clustering of cultivars (**Supplemental Figure 3.1**). One sample of *B. cretica* clusters closest to cultivated types while all samples of *B. incana* (those not in clade 2) cluster within the cultivars, along with one sample of *B. montana*. Other wild relatives cluster based on geography. Also notable are the putatively wild *B. oleracea* samples, which cluster with the cultivars (gray triangles). Current distributions of *B. cretica* are throughout the Eastern Mediterranean, primarily in Greece, highlighting a potential origin of domestication (**Figure 3.2A**). Species level phylogenomics using IQ-Tree (Nguyen et al. 2015) and the PoMo model (Schrempf et al. 2016) indicate that *B. incana* is either the closest living wild relative, or that it is also a domesticated species and that *B. cretica* is the closest living wild relative (**Figure 3.3D**). Within the cultivar clade, the backbone consists of much shorter branches, evidenced by weak support at four nodes, with the exception of the clade consisting of cauliflower (var. *botrytis*), broccoli (var. *italica*), Chinese white kale (var. *alboglabra*), and kohlrabi (var. *gongylodes*).

Admixture between wild and cultivated populations

To further assess why some wild samples grouped within the cultivars, we tested for signals of introgression by treating these samples as individuals in TreeMix (Pickrell & Pritchard 2012) and looking for evidence of gene flow. We recovered significant admixture between wild and cultivated populations with five migration events explaining 97.5% of

the variance within the data (**Supplemental Figure 3.2 E & I**). This analysis indicated gene flow from one sample of *B. cretica* (198) to three other lineages, both wild and cultivated (*B. hilarionis*, alboglabra, and an ancestor leading to gongylodes + *B. cretica*; 199). Other recovered gene flow events included one from *B. rupestris* and one from a later ancestor both to wild C - clade 2. The last recovered event is from Brussels sprouts to two samples of *B. incana* (207, 204). The phylogeny is largely congruent with the species phylogeny, with one difference being *B. hilarionis* sister to *B. montana* (sample 224) rather than as successive groups. We find two samples of *B. cretica* as successive with sample 198 followed by 196. Following these wild populations we recover a clade with, again, very short branches, but similar clades as the species tree (**Figure 3.3D**). These clades are [Brussels sprouts + collards + cabbage + savoy cabbage], [trouchuda kale + Chinese white kale + broccoli + cauliflower], [walking stick kale + lacinato kale], and a clade of [marrow-stem kale + perpetual kale]. Curly kale groups with two samples of wild *B. oleracea* (177 and 175) and one sample of *B. montana* (222).

To test explicitly if introgression was driving the patterns discussed above, we utilized F4 statistics as implemented in the ‘fourpop’ script in TreeMix (Pickrell & Pritchard 2012). Using *B. villosa* and *B. rupestris* as a sister group, we tested each wild sample that clustered within the cultivars as sister to cultivars (**Supplemental Figure 3.3**). We find that only two samples of *B. incana* (205, 208) show signals of introgression, with all other tested samples evolving treelike indicating that these samples are of a feral origin (**Supplemental Table 3.2**).

Expression variation and profiles place wild C - clade 2 sister to all cultivars

PCA of the top 1000 variable genes in expression levels recovered similar results as with the genetic variation (**Supplemental Figure 3.4**). PC1 versus PC2 separate most wild species from cultivar types with two distinct clouds of cultivars. We find the same samples of *B. incana*, *B. cretica*, and *B. montana* again clustering within the cultivars (**Supplemental Figure 3.4**). PC2 versus PC3 and PC3 versus PC4 show less differentiation. Hierarchical clustering of these top 1000 genes using expression profiles recovers wild C - clade 2 as sister to all cultivars (**Supplemental Figure 3.5**). This result agrees with phenotyping done by the authors indicating that these plants are phenotypically wild, rather than mis-identifications. Two samples, *B. hilarionis* and one sample of tronchuda kale (030), reveal unique expression profiles and are found sister to all other samples. To identify modules of genes that might be driving these differences between wild species and cultivars we used WGCNA, a co-expression network analysis (Langfelder & Horvath 2008). We found that a soft-thresholding power of six was best as it was the lowest power that satisfied the approximate scale-free topology criterion, resulting in five modules of genes (**Supplemental Figure 3.6**). Module gene number and top five GO terms enrichment in **Supplemental Table 3.3**. Generally, these genes are associated with response to stress, herbivory defense mechanisms, and response to light.

Archaeological evidence for Brassica oleracea domestication origin

Supplemental Table 3.4 lists finds from archaeological sites across Europe and the Near East. The earliest reported claim of *B. oleracea* comes from the Austrian Alps and comprises three seeds dated to the Middle Bronze Age (c. 1600-1400 cal. BC; Schmidl & Oeggel 2005). The lack of illustrations and discussion of separation criteria from other

Brassica spp. makes us question the reliability of this identification to species level (*see Discussion*). The only other find of similar antiquity comes from seeds of *B. oleracea* identified by scanning electron microscopy from Late Bronze Age/Early Iron Age Syria and radiocarbon dated directly between 1300-1020 ca. BC (Kaniewski et al. 2011). These finds are associated with destruction levels at Gibala, Tell Tweini in Western Syria on the Mediterranean coast. The next earliest finds are from the La Tène period in Switzerland, France, and Germany dating to around 400-100 BC, with Roman introductions to other parts of Europe and North Africa occurring between the 1st to the 4th centuries AD. These include desiccated remains from Mons Claudianus, a settlement associated with a quarry site in the Eastern desert, Egypt (Van der Veen 2011), possible seeds of *B. oleracea* from a Roman well in Liguria, north-west Italy (Arobba et al. 2013), and several finds from Roman Britain (Van der Veen, Livarda, & Hill 2008) and Roman Germany (Wiethold 2002).

The next group comprises seeds identified from sites dating between the 9th to the 13/14th Century AD. These are mainly from northern Europe, with one waterlogged record from Žatec in the north-west Czech Republic. Most of these are finds of waterlogged seeds; however, in one case the evidence relates not to seeds, but to pottery residues where lipids of *Brassica* leaf waxes were identified (Evershed et al. 1992; Evershed et al. 1994). The authors attribute this to the boiling of leaves of *B. oleracea*, and given the lack of evidence for other commonly eaten *Brassica* leaves in England at this time, this identification would appear likely.

The final archaeological finds date from the middle of the 14th century to the modern day, and predominantly come from sites in northern and central Europe. The vast

majority are, as with the earlier medieval finds, related to waterlogged and mineralized remains from waste pits, sewers, and wells (**Supplemental Table 3.4**).

Classical and medieval written sources

A number of authors have mentioned the absence of evidence for cabbages in the Near Eastern literature and pictorial record, while other *Brassica/Sinapis* crops, such as turnip and mustard are cited (Maggioni et al. 2010). **Supplemental Table 3.5** summarizes a selection of relevant literary and artistic sources.

The earliest references to cabbage date to Greek scholars of the 6th to 3rd centuries BC. Hipponax writing in the 6th century BC refers to a seven-leaf cabbage or κράμβην τὴν ἐπτάφυλλον (West 2011), while Hippocrates *On the Nature of Women* written around the mid 5th to mid 4th century BC refers to the use of cabbage, or krambe, κράμβη in a few recipes (Totelin 2009).

Two authors of the 4th to early 3rd Century BC have provided more detailed descriptions relating to four to five varieties of cabbage. The works of Eudermus are only known from Athenaeus of Naucratis, cited in the *The Deipnosophistae* written around the 3rd Century AD (f.85 quoted in *Anthenaeus* 9.9, Yonge 1854; pp. 582-3). Eudermus refers to three kinds of cabbage, κράμβης, a smooth-leaved type, λειόφυλλος, a parsley or curly-leaved type, σελινούσσης, and a salt variety, ἀλμυρίδος/ ἀλμυρίς, the latter he states has a delicate taste and appears to be distributed mainly in the eastern Aegean and on the modern Turkish coast. Theophrastus (Yonge 1854; VII, iv.4-6, *Anthenaeus The Deipnosophists* 9.9) also refers to three types, a curly-leaved type (οὐλόφυλλου/οὐλόφυλλος), a smooth-leaved type (λειόφυλλου/λειοφυλλος), and a wild type (ἀγρίας/ ἀγρία) with a bitter taste, many branches, and many small round leaves (Yonge 1854; p. 85). Nicander of Colophon

dating to the 2nd century BC also quoted by Athenaeus also refers to a smooth leaved type, sometimes found wild, and curly leaved type with many leaves, and one with a reddish color (Yonge 1854; p. 582).

Several Roman authors refer to cabbage: for instance Varro *De Rustica* (Hooper & Ash 1934), Columella in *On Agriculture* (Forster & Heffner 1968), and Virgil in *Georgics* (Fairclough 1916) all mention cabbage several times. However, the best descriptions are those given by Cato, writing also in the 2nd century BC, and Pliny the Elder writing in the 1st Century AD. Cato refers also to a smooth-leaved variety, large with broad leaves and a thick stem (Hooper & Ash 1934; 157.1-12), a curly leaved called “parsley cabbage”, a small staled, tender pungent variety and a wild cabbage. Pliny writing some 200 years later in his *Natural History* describes a great many more varieties than seen in the previous classical works (Rackham 1950; 19.41), referring to smooth large leafed, a crisp curly leaved, a thinned stalked, tender leaved variety, then others called Tritian, Cumanian (leaves close to the ground), the Aricinian (thinner of no height but more leaves), Pomeii (taller, tender, stalk thin at root), the Bruttium (or Calabrian) large pungent leaves and a thin stalk, the Sabine (crisp thick, sweet leaves), Lacuturres (large head, grown in Arcia), and the halmyridia (growing by the sea-shore- this has been identified as *Crambe maritima* - unlikely as not found in Italy, or *Convolvulus soldanella*). Pliny describes both the head of the Tritian as occasionally growing up to a foot in diameter.

Concerning the identification of these plants more recent scholars have been inclined to accept that most translations of Greek or Latin of cabbage, refer to *B. oleracea*, although both in the Greek-English Lexicon (Liddell & Scott 1940) and in Hort’s (1916)

translation of Theophrastus *Historia Plantarum*, both κράμβης and ῥάφανος, are translated as *B. cretica* not *B. oleracea*.

The modern-day distribution of *B. cretica* includes south-east Greece and the Aegean Islands; it is also recorded from the south of Turkey, the west Syrian Mediterranean coast, and north Lebanon (Kroll 2005; Hanelt, Buttner, & Mansfeld 2001; Davies, Cullen, & Coode 1965; Post 1932; Strid & Tan 2002). Given that neither *B. incana* nor even wild feral forms of *B. oleracea* are known from Greece, and in light of the genetic work presented here it would seem prudent to assign these earlier curly leaved and smooth leaved varieties known within the Greek World to *B. cretica*. Certainly, there are differences between *B. cretica* subsp. *cretica* and *B. cretica* subsp. *laconica* that might be reflective of such differences in the past. Regarding the wild types it should as Willis (2007; Maggioni et al. 2010) notes the term, ῥάφανος is used for a number of plants including radish ῥάφανος, (*Raphanus sativa*), potentially charlock or runch, ῥάφανος ἡ ἀγρία (*Raphanus raphanistrum*), and spurge ῥάφανος ἡ ὀρεία (*Euphorbia apios*). Although the description by Niciander (quoted by Athenaeus; Yonge 1854; p. 582) would seem to leave less doubt that wild or at least feral forms of *B. cretica* were known in at least Ionia in the 2nd Century BC.

Whether *B. cretica* was domesticated in mainland, island Greece, or on the coast of modern Turkey, Syria, or Lebanon is less clear. Given the situation for island diversity of feral populations of *B. incana* (Tutin et al. 2003; p. 407; see below) and that considerable diversity is also noted for *B. cretica* subsp. *aegaea*, it would seem probable that *B. cretica* was also introduced to many of the Aegean Islands by human populations.

Regarding the archaeological finds of *B. oleracea* from Gibala, Western Syria dated to 1300-1020 cal. BC, seeds of *B. cretica* are similar to *B. oleracea* having a finer reticulum than seen for example in *B. nigra* or *B. tournefortii*, whereas *B. elongata* has a much less distinctive surface reticulation. These latter two species' seeds are relatively small, while those of *B. deflexa*, the only other *Brassica* found in this region, are less than 1mm in size. As *B. oleracea* would not appear to be present in Western Syria at this date it seems probable that the seeds recovered from Late Bronze Age Gibala, Tell Tweini in Syria are more probably of *B. cretica* and therefore potentially represent the earliest evidence for the cultivation of this species. Whether it was domesticated in this area or to the north in south-west Turkey or to the south in Lebanon, or indeed in mainland modern Greece or on perhaps one of the larger Aegean Islands remains a topic for future research. However, linguistic evidence would favor a Greek domestication (Maggioni et al. 2010).

The distribution of *B. cretica* and/or *B. incana* appears uncertain. Mansfeld lists both *B. cretica* and *B. incana* from the south of Crimea (Hanelt, Buttner, & Mansfeld 2001), the Flora of Russia lists *B. cretica* and *B. oleracea* as growing wild in Crimea (Fedorov 2001), while Snogerup et al. (1990) suggest that identifications of *B. cretica* from the Crimea may be *B. incana*. Snogerup et al. (1990) list a number of varieties of *B. cretica* from the Aegean islands and while it is possible that it spread through natural means, it is quite probable that rather its distribution reflects feral populations from early cultivars spreading from the former Ionian coast.

DISCUSSION

Multiple lines of evidence support a single Eastern Mediterranean origin

Using genome-scale, multilocus data along with species distribution models, archeological, and literary sources, we find evidence to support a single Eastern Mediterranean origin for *B. oleracea*. Species distribution models show that *B. cretica*, *B. incana*, and *B. macrocarpa* all had suitable habitat in the areas suggested by literature sources at the time when domestication is thought to have occurred (Maggioni et al. 2018). Genomic evidence supports a single Eastern Mediterranean origin as we recover *B. cretica* and *B. hilarionis* together sister to all cultivars in our individual sample phylogeny. PCA further suggests that *B. cretica* is the most similar to cultivars as it clusters the most closely, agreeing with our species-level phylogeny. This evidence, together with the phylogeographic discontinuity of wild *B. oleracea* samples with their Eastern Mediterranean progenitors, the more distant placement of *B. insularis*, *B. macrocarpa*, and *B. villosa*, and strong patterns of shared ancestry among *B. incana* and cultivars, leads us to support the hypothesis of domestication in the Eastern Mediterranean from *B. cretica* populations.

However, our knowledge of the spread of *Brassica* crops past domestication is compounded by both the difficulties of identifying the seeds of individual species and weak backbone support in the sample level and species phylogeny. It should be noted that seeds of both *Raphanus* and *Sinapis* can be difficult to distinguish from *Brassica* seeds under low magnification when presented with a limited number of poorly preserved charred specimens. Thus, while the archeological evidence is generally poor, it is sufficient to make a number of comments. The first is that there are no firm archaeobotanical records for *B.*

oleracea preceding the Middle Bronze Age. The Middle Bronze Age record from Austria should be regarded with caution as a possible misidentification/misnaming. The seeds and identification are listed under “Herbaceous vegetation of frequently disturbed sites.” *Brassica oleracea* is not found wild in Austria and only *B. nigra* is likely to be found in this region growing among crops (Tutin et al. 2003) or potentially cultivated turnip (*B. rapa*). Leaving aside the record from Syria (Kaniewski et al. 2011), there is no good archaeological evidence to suggest the possible cultivation of cabbages in Europe or North Africa as a whole prior to the Late Iron Age and Roman periods. There are no records for *B. oleracea* from before this period within databases relating to the Near East (Reihl 2014), Europe (Kroll 2001; 2005), Britain (Tomlinson & Hall 1996), the Czech Republic (Kreuz & Schäfer 2002), or within pre-dynastic and Pharaonic Egypt (Murray 2000) despite, as stated above, having evidence for other *Brassica* spp. While the evidence is patchy it is believed that the increasing appearance of seeds of *B. oleracea* on archaeological sites across Europe and North Africa at this time (from the 2nd Century BC to the 4th century AD) can be attributed to the spread of crops both within and on the periphery of the Roman Empire (e.g. Van der Veen 2001). At the very least it would appear to represent the increasing ubiquity of the species during the Roman period. The second observation is that by comparison to the time periods and area covered, there is a clear increase in finds within medieval Europe that might also suggest that the crop had become more popular or widespread by this time.

The role of feral populations in the domestication of Brassica oleracea

There is an increasing quantity of literature that highlights the role of wild and feral populations as pools of diversity that contributed to crop diversification through

introgressions during domestication events (Allaby 2010; Fuller et al. 2014). We too find that this is the case of the domestication of *B. oleracea*: it appears that introgression from wild or feral populations contributed to the genetics of particular crops, and vice versa, which is revealed in an analysis of admixture (**Figure 3.3**). We identify several feral populations of not only wild *B. oleracea*, but also of *B. cretica* and *B. incana* as well as *B. montana* and *B. villosa*, indicating that these crops and their wild relatives revert to a feral state with relative ease. Wild *B. oleracea*, a suggested progenitor species, is recovered using several lines of evidence as a feral lineage, agreeing with the conclusions of Mitchel (1976) and Michell & Richards (1979). All three samples of wild *B. oleracea* included in our analyses fall out within the cultivar clade in the individual-level phylogeny and cluster with domesticates in all PCA analyses (using both SNPs and expression variation). *Brassica incana*, another suggested progenitor species, appears also to be feral as we find that all of our samples of *B. incana* share a large amount of shared ancestry with cultivars (**Figure 3.2B**). Tests for introgression show that only two of the five samples (205 and 208), show any signature of introgression, including events among other wild taxa such as *B. rupestris*, where their ranges cross (**Figure 3.3**). *Brassica incana* has been described as a ‘traditional wild vegetable’ (Tripodi et al. 2012), with Crimean *B. incana* described as having indehiscent mature siliques (non-shattering seed pods; Diederichsen et al. 2012), therefore it is likely represented by one of the cabbage types recorded in early Greek and Roman sources. Finally, while we do identify a single sample of *B. cretica* as a likely progenitor species, we also find three samples (195, 196, and 199) that cluster with cultivars, these likely also representing feral lineages as they show a large amount of shared ancestry with all other cultivars and show either very little or no signal of introgression.

For our identified wild C - clade 2, comprising three species and five samples, *B. incana* (205, 208, 209), one sample of *B. villosa* (233), and one sample of *B. cretica* (195) with shared ancestry with both cultivars and wild relatives, we believe that these samples may have a chloroplast similar to one another which is driving their grouping.

Dispersal by human translocation westward, ultimately to the Atlantic coast of Europe, appear to have established both regional feral populations and new crop morphotypes. Current archaeological evidence suggests that this process may have begun with Late Bronze Age seafaring (3000-3300 years ago), when the whole Mediterranean became linked in trade perhaps for the first time (Broodbank 2013) and to have continued to provide a corridor for introgression and varietal diversification through the Iron Age (up to 2000 years ago), including producing those varieties recorded in Classical written sources (Maggioni et al. 2010; 2018). By 2000 BP, trade links along the Atlantic seaboard from North Africa and Iberia through Britain and Ireland are clearly indicated in archaeology (Cunliffe 2001), and are associated with the first peopling of the Canary Islands from the north (~2200 BP), where feral *B. bourgaei* became established and Walking stick kale is endemic. Previous researchers have observed that wild *B. oleracea* and *B. cretica* populations are both typically found in coastal locations associated with ancient sea ports as well as their preferred ecological niche on chalk cliffs undisturbed by grazing (Snogerup et al. 1990; Mitchell 1976). Taking together the insights of phylogenomics (**Figure 3.2A**) and archaeology, we infer that the distribution of *B. cretica* is partly anthropogenic and associated with dispersal by seafaring human populations, prior to and during the domestication process. Additionally, while the present data has demonstrated this for European vegetable crops, we might expect similar processes

associated with the eastwards dispersal of Chinese white kale across overland Asian routes to China.

In the future, additional sampling of *B. cretica* populations and subspecies as well as more samples of *B. incana* may help to sort out who truly is the closest living relative. Due to the many confounding factors of potential backcrossings, feral populations, and partial domestications, wide geographic sampling combined with deep-level sequencing is necessary to uncover the evolutionary patterns that have led to the current *B. oleracea* vegetable crops we know today.

MATERIALS AND METHODS

Taxon sampling

Samples from domesticated morphotypes include seven types of kales (chinese white kale; 22, tronchuda kale; 13, lacinato kale; 8, curly kale; 12, perpetual kale; 13, walking stick kale; 2, and collard greens; 16), three types of cabbages (white cabbage; 15, savoy cabbage; 4, and marrow cabbage; 13), Brussels sprouts (17), broccoli (20), cauliflower (18), and kohlrabi (15). Together these accounted for 189 of the 224 total samples. The rest of the 224 samples included wild *B. oleracea* (3), *B. cretica* (3), *B. incana* (5), *B. montana* (2), *B. hilarionis* (1), *B. insularis* (5), *B. macrocarpa* (4), *B. rupestris* (7), and *B. villosa* (5; **Supplemental Table 3.1**). These include accessions from the United States Department of Agriculture, Agriculture Research Service (USDA-ARS) germplasm center (Geneva, NY, 114 accessions), The Leibniz Institute of Plant Genetics and Crop Plant Research (IPK; 71 accessions), Universidad Politécnica de Madrid (UPM; 4 accessions), The Nordic Genetic Resource Centre (NordGen; 2 accessions), Gomez Campo

Collection (2 accessions), John Innes Center (1 accession), Doubled Haploid lines (17 samples, some accessions sampled twice), or from our personal collection (13 accessions). Four replicates of each accession were grown from seed in a sterile growth chamber at the University of Missouri (MU; Columbia, MO) Bond Life Sciences Center in a randomized complete block design. At the seventh leaf stage, leaf four was collected from each plant and immediately flash frozen in liquid nitrogen for RNA extraction. Morphotype identity was validated in mature plants by growing all accessions twice over the span of two years (**Supplemental Table 3.1**).

Species distribution modeling

We compiled occurrence records of *B. incana*, *B. insularis*, *B. montana*, *B. macrocarpa*, *B. villosa*, *B. hilarionis*, and *B. cretica* from the GBIF (www.gbif.org) data portal and data from Snogerup (1990). From the GBIF data, we omitted duplicate records, records without location data, records without vouchers, and collections from the grounds of botanical gardens. Records clearly outside of the native range were also removed. From the Snogerup (1990) data, we omitted records that could not be georeferenced to <5km spatial uncertainty. In the case of *B. cretica*, populations in Lebanon and Israel and *B. incana* in Crimea are thought to be likely early human introductions (Snogerup 1990) so records from these areas were omitted. Occurrences above 1200m altitude were also omitted, as these species rarely occur above 1000m and observations above these altitudes may represent anthropogenic dispersals to disturbed areas (Ibid.) or misidentifications. To minimize spatial autocorrelation and sampling bias (Beck et al. 2014), we thinned the filtered occurrences to records greater than or equal to 10km apart using the ‘spThin’ package in R (Aiello-Lammens et al. 2015). After filtering and thinning, 172 records

remained for *B. cretica*, 65 for *B. incana*, 57 for *B. insularis*, 101 *B. montana*, 7 for *B. macrocarpa*, 15 for *B. villosa*, 6 for *B. hilarionis*. Next we obtained rasters for 19 bioclimatic variables at 2.5 minutes resolution based on contemporary climate data from WorldClim 2.0 (Fick and Hijmans, 2017) and rasters for 19 bioclimatic variables at 2.5 minutes resolution based on mid-Holocene climate projections using the Community Climate System Model 4 (Meehl et al., 2012) from WorldClim 1.4 (Hijmans et al., 2005). Rasters were clipped to fit the Mediterranean and surrounding areas. All variables were used in modeling as recent simulations have shown that removing highly correlated collinear variables has an insignificant impact on maximum entropy model performance (Feng et al., 2019). Projections for mid-Holocene habitat suitability were generated using MaxEnt 3.4.1 (Phillips et al., 2017). Linear, quadratic, product, and hinge features were used and jackknifes were used to measure variable importance. Model performance was evaluated with the adjusted area under receiver operating characteristic curve (DeLong et al., 1988).

RNA isolation and sequencing

RNA was isolated using the ThermoFisher Invitrogen PureLink RNA mini kit (Invitrogen, Carlsbad, CA, USA) followed by TruSeq library preparation (Illumina, San Diego, CA, USA) and sequencing on the NextSeq (Illumina, San Diego, CA, USA) for 2 X 75 bp reads. Library preparation and sequencing was performed through the MU DNA Core Facility. For eight flow cells, 24 samples were multiplexed and sequenced in a single flow-cell, followed by a ninth flow cell with 17 samples, and a tenth flow-cell with just 16 samples.

Mapping and SNP calling

Short reads were mapped to the *B. oleracea* TO1000 genome (Chinese kale; Parkin et al. 2014; release-41) using the STAR v. 2.5.2 (Dobin et al. 2013) two-pass alignment to first identify splice junctions, which were used in the second pass to improve mapping (Engström et al. 2013). Mapped reads (BAM format) were then processed following the GATK v. 3.8 best practices for RNA-seq reads (Mckenna et al. 2010; Van der Auwera et al. 2013; Poplin et al. 2017). To ensure that reads were mapping correctly, the GATK ‘Split’N’Trim’ function was used to split reads into exon segments and trim any overhanging reads in intron segments. In total, 7,564,168 variants were called before any filtering was performed. The resulting variants were filtered to exclude those with a Fisher strand (FS) value greater than 30 and quality depth (QD) less than 2.0. This recovered 942,357 variants in total, with 879,865 variants on chromosomes 1-9 and 62,492 variants on scaffolds. Only variants aligning to chromosomes were used in downstream analyses. The remaining variants were then filtered on the basis of the proportion of missing data (`-max-missing 0.4`), sites with mean depth values greater than 5 (`--min-meanDP 5`), and if they were indels (`--remove-indels`; vcfTools v. 0.1.17; Danecek et al. 2011) resulting in a total of 103,525 SNPs. Finally, SNPs were filtered for linkage disequilibrium (LD) using a window size of 80KB, or about two times the estimated length for 50% LD decay (Cheng et al. 2016) for a final dataset of 36,750 SNPs using PLINK v. 1.90 (`--indep 80kb 5 2`; Purcell et al. 2007).

Phylogenetic and Introgression Inference

To test how the different populations are related to one another and which wild relative is most closely related to the domesticated types, we used three different

phylogenetic programs; SNPhylo v. 20160204 (Lee et al. 2014) to assess individual sample relationships, IQ-Tree v. 1.6 (Nguyen et al. 2015) using the PoMo model (Schrempf et al. 2016) to test species level relationships, and TreeMix v. 1.13 (Pickrell and Pritchard 2012) to assess introgression. For SNPhylo (Lee et al. 2014) we ran analyses using 0.1 for LD threshold, 0.01 for minor allele threshold, 0.4 for missing rate, 1000 bootstrap replicates, and rooted with sample 238 (*B. villosa*). For IQ-Tree, we used the Polymorphism-aware phylogenetic Models (PoMo) (Schrempf et al. 2016; *-m GTR+P*) to perform phylogenetic comparisons using population genetic data, 1000 bootstrap replicates using their ultra fast bootstrap approximation method, and again *B. villosa* was used to root the tree. For our IQ-Tree analysis, we subsampled data to include only those sample which assigned as either a cultivar or wild relative with at least 70% ancestry in our $K = 2$ fastSTRUCTURE result. To test both topology of relationships and for gene flow among populations, we used TreeMix with the following parameters: *-noss -root villosa -bootstrap -k 500* and *-m* from 2-10. TreeMix (Pickrell and Pritchard 2012) was run with samples of *B. cretica*, *B. incana*, and *B. oleracea* as individuals and all cultivars and wild relatives as populations. To implicitly test for signatures of introgression, the ‘fourpop’ script in TreeMix (Pickrell & Pritchard 2012) was used with *B. villosa* and *B. rupestris* as a sister group and our tested samples sister to cultivars.

Population Structure Variation

Angsd v. 0.925 (Korneliussen et al. 2014) was run on two sets of bam files (all samples and just domesticates) using the parameters *-doGlf 2 -doMajorMinor 1 -doMaf 2 -minMapQ 30 -SNP_pval 1e-6* followed by PCAangsd v. 0.97 (Meisner and Albrechtsen 2018) to visualize population structure, estimate allele frequencies, and calculate

inbreeding coefficients among wild species and domesticates with the parameters *-admix -selection 1 -inbreed 2*.

To test ancestry proportions and identify likely genetic structure of described populations we used fastSTRUCTURE v 1.0 (Raj, Stephens, and Pritchard 2014). We tested K values 2 to 4 for our full dataset with wild and domesticated populations, and K values 2 to 10 for our dataset of just domesticated populations using default convergence criteria and priors. The best K value for runs using both datasets was estimated using *chooseK.py* in the fastSTRUCTURE package.

Clustering based on Expression Profiles

First Salmon v 1.2.1 (Patro, Duggal, and Kingsford 2015) was used to acquire abundances or transcript per million (TPM) measures for each sample and the estimated number of reads originating from transcripts. The input for expression profile analysis was prepared using tximport (Soneson, Love, and Robinson 2015), followed by correction for library size (*estimateSizeFactors*) and variance-stabilizing transformation (*vst*) in DESeq2 v 1.28.1 (Love, Huber, & Anders 2014). Correction for batch effect was performed using limma v 3.44.3 (*limma::removeBatchEffect*; Ritchie et al. 2015). Using only the top 1000 variable genes, we ran a PCA (*prcomp*) on the normalized expression values to test for clustering based on expression profiles. Finally, to assess networks of genes driving differences observed in the PCA, we used WGCNA v 1.68 (Langfelder & Horvath 2008) again using the top 1000 variable genes. Clustering of samples was based on Euclidean distance using *hclust* with *method= "average"*. Soft-thresholding powers were assessed from 10 to 20 to choose the best power for network construction and module detection.

AUTHOR CONTRIBUTIONS

MEM, PPE, JDM, PGW, GRT, PH, DACP, GB, JED, JH, MAG, and JCP designed the project. MEM, EYG, and SDTH grew plants and collected tissue. MEM and EYG extracted and isolated RNA. MEM analyzed the genetic data. ACM conducted the species distribution models. CS, DQF, and RGA researched the archeology and written data. SDTH, HA, and JED assisted with processing and analyzing the data. MEM wrote the original manuscript.

ACKNOWLEDGMENTS

We thank Bob Schnable, Troy Rowan, Harly Durbin, and Paul Blischak for their assistance with computational analyses, the Mizzou DNA core, Nathan Bivens, Ming-Yi Zhou, and Karen Bromert, for their assistance in getting quality data for sequencing, and our computing resources, specifically the Research Computing Support Services (RCSS) and Informatics Research Core Facility (IRCF) at the University of Missouri. Finally, we thank our funding USDA-ARS Project No. 8060-21000-024-00D.

REFERENCES

- Aiello-Lammens, M.E., Boria, R.A., Radosavljevic, A., Vilela, B. & Anderson, R.P. 2015. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, 38(5), pp.541-545.
- Allaby, R. 2010. Integrating the processes in the evolutionary system of domestication. *Journal of experimental Botany*, 61(4), 935-944.
- Allender, C. J., Allainguillaume, J., Lynn, J., & King, G. J. 2007. Simple sequence repeats reveal uneven distribution of genetic diversity in chloroplast genomes of *Brassica oleracea* L. and (n= 9) wild relatives. *Theoretical and Applied Genetics*, 114(4), 609-618.
- Ansorge, J., Igel, K., Schäfer, H., & Wiethold, J. 2003. Ein Holzschacht aus der Baderstr. 1 in Greifswald. Aus der materiellen Alltagskultur der sozialen Oberschicht einer Hansestadt in der 2. Hälfte des 14. Jahrhunderts. Bodendenkmalpflege in *Mecklenburg-Vorpommern, Jahrbuch* 50, 2002, 119-157.
- Arobba, D., Bulgarelli, F., Siniscalco, C., & Caramiello, R. 2013. Roman landscape and agriculture on the Ligurian coast through macro and microremains from a Vada Sabatia well (Vado Ligure, Italy). *Environmental Archaeology*, 18(2), 114-131.
- Beck, J., Böller, M., Erhardt, A. & Schwanghart, W. 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, 19, pp.10-15.
- Beranová, Zeli, 2002. cibule a česnek v kuchyni a v medicíně do konce 16. Století. Sborník Západočeského muz Plzni [Plizen] řada *History* 16, 185-194.

- Block, H. 1984. 1647: Horticultura danica. København. Ældste danske havebog, fotografisk optryk, Århus.
- Boretius, A. 1883. Capitularia regum francorum (Vol. 1). Monumenta Germaniae Historica. Societatis Aperiendis Fontibus Rerum Germanicarum Medii Aevi.
- Brinkkemper, O. 2002. Plantenresten uit beerputten aangetroffen op een bouwlocatie aan het Spaarne te Haarlem. Een “rijke” informatiebron!. Haarlems Bodemonderzoek, 36, 104-132.
- Brinkkemper, O. 1991. Wetland farming in the area to the south of the Meuse estuary during the Iron Age and Roman Period. An environmental and palaeo-economic reconstruction. *Analecta Praehistorica Leidensia*, 24.
- Broodbank, C. 2013. The Making of the Middle Sea: A History of the Mediterranean from the Beginning to the Emergence of the Classical World. *Thames & Hudson*.
- Cheng, F., Sun, R., Hou, X., Zheng, H., Zhang, F., Zhang, Y., Liu, B., et al. 2016. Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in *Brassica rapa* and *Brassica oleracea*. *Nature genetics* 48, 1218.
- Cooremans, B. 2000. Het macrobotanisch onderzoek: een speurtocht naar plantaardige (voedsel) resten. in Veeckman, J., van Hoof, W., Cooremans, B., Ervynck, A., van Neer, W. Eds., De inhoud van de afvalput van de Groote Schalien Looze: speuren naar de 17de eeuwse bewoners. BRABOM Berichten en rapporten over het Antwerps. Bodemonderzoek en Monumentenzorg 3. (Antwerpen, Stad Antwerpen,), pp. 131-139.

- Cunliffe, B. W. 2001. *Facing the ocean: the Atlantic and its peoples, 8000 BC-AD 1500*. Oxford University Press, USA.
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156-2158.
- Davies, P. H., Cullen, J., & Coode, M. J. E. 1965. *Flora of Turkey* (vols. 1–10) Edinburgh University Press.
- DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pp.837-845.
- Diederichsen, A., Rozhkov, R. V., Korzhenevsky, V. V., & Boguslavsky, R. L. 2012. Collecting genetic resources of crop wild relatives in Crimea, Ukraine, in 2009. *Crop wild relative*, 34.
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15-21.
- Dodoens, R. 1971. 1554. Cruydeboeck. Officina I. Loci, Antwerp. Facsimile at Cambridge Univ. Libr, 9, 1616.
- Engström, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., Alioto, T., et al. 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature methods*, 10(12), 1185.
- Evershed, R. P., Heron, C., Charters, S., & Goad, L. J. 1992. The survival of food residues: new methods of analysis, interpretation and application. *In Proceedings of the British Academy* (Vol. 77, No. 187, p. 2).

- Evershed, R. P., Arnot, K. I., Collister, J., Eglinton, G., & Charters, S. 1994. Application of isotope ratio monitoring gas chromatography–mass spectrometry to the analysis of organic residues of archaeological origin. *Analyst*, 119(5), 909-914.
- Fairclough, H.R. 1916. *Virgil. Eclogues, Georgics, Aeneid*. (Loeb Classical Library Volumes 63 & 64, Cambridge, Mass.).
- Fedorov, A. 2001. *Flora of Russia: the European part and bordering regions*. Volume 5: Magnoliophyta (= Angiospermae), Magnoliopsida (= Dicotyledones). Flora of Russia: the European part and bordering regions. Volume 5: Magnoliophyta (= Angiospermae), Magnoliopsida (= Dicotyledones).
- Feng, X., Park, D.S., Liang, Y., Pandey, R. and Papeş, M. 2019. Collinearity in ecological niche modeling: Confusions and challenges. *Ecology and Evolution*, 9(18), pp.10365-10376.
- Fick, S.E. and Hijmans, R.J. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology*, 37(12), pp.4302-4315.
- Forster S. & Heffner, E.H. 1968. Lucius Junius Moderatus Columella On Agriculture (Loeb Classical Library, Heinemann, London).
- Fuller, D. Q., Denham, T., Arroyo-Kalin, M., Lucas, L., Stevens, C. J., Qin, L., et al. 2014. Convergent evolution and parallelism in plant domestication revealed by an expanding archaeological record. *Proceedings of the National Academy of Sciences*, 111(17), 6147-6152.
- Godwin, H., & Bachem, K. 1961. Appendix III: plant material. KM Richardson, Excavations in Hungate, York, *Archaeol. J.*, II6, IO9.

- Greig, J., Girling, M., & Skidmore, P. 1982. *The plant and insect remains*. 60-71 in Barker, P., Higham, R. Hen Domen Montgomery. A timber castle on the English-Welsh Border. (I. Roy. Archaeol. Inst., London).
- Hall, A. R. 1983. Environment and living conditions at two Anglo-Scandinavian sites. Council for British Archaeology for the York Archaeological Trust.
- Hanelt, P., Buttner, R., & Mansfeld, R. 2001. Mansfeld's Encyclopedia of Agricultural and Horticultural Crops (except Ornamentals). Mansfeld's Encyclopedia of Agricultural and Horticultural Crops (except Ornamentals).
- Hansson, A.M. *Vikingabotanik*. in Pettersson, B. et al., Eds., Människan och naturen. Etnobiologi i Sverige 1. (Stockholm, 2001).
- Harlow, M., & Smith, W. 2001. Between fasting and feasting: the literary and archaeobotanical evidence for monastic diet in Late Antique Egypt. *Antiquity*, 75(290), 758-768.
- Harvey, J. H. 1992. Garden plants of Moorish Spain: a fresh look. *Garden History*, 20(1), 71-82.
- Heaney, R. K., Fenwick, G. R., Mithen, R. F., & Lewis, B. G. 1987. Glucosinolates of wild and cultivated *Brassica* species. *Phytochemistry*, 26(7), 1969-1973.
- Heimdahl, J. 2010. Barbariska trädgårdsmästare: nya perspektiv på hortikulturen i Sverige fram till 1200-talets slut.
- Helm, J. 1963. Morphologisch-taxonomische Gliederung der Kultursippen von *Brassica oleracea* L. *Die Kulturpflanze*, 11(1), 92-210.
- Hijmans, R.J., Cameron, S., Parra, J., Jones, P., Jarvis, A. and Richardson, K. 2005. WorldClim—global climate data. Version 1.4 (release 3).

- Hodgkin, T. 1995. *Cabbages, kales, etc.* 76-82. In. J. Smartt and N. Simmonds (ed.).
Evolution of crop plants.
- Hooper, A.D., & Ash, H.B. 1934 *Cato and Varro: On Agriculture*. Loeb Classical Library,
No. 283. (Harvard Univ. Press).
- Hort, A. 1916. *Theophrastus: enquiry into plants*. Cambridge (Mass.)
- Jacomet, S., & Vandorpe, P. 2011. Plantes anciennes et nouvelles: la région du Rhin
supérieur et l'Allemagne du Sud-Ouest.
- Jensen, H. A. 1985. Catalogue of late-and post-glacial macrofossils of Spermatophyta from
Denmark, Schleswig Scania, Halland, and Blekinge dates 13,000 BP to 1536 AD.
- Kaniewski, D., Van Campo, E., Van Lerberghe, K., Boiy, T., Vansteenhuyse, K., Jans, G.,
et al. 2011. The Sea Peoples, from cuneiform tablets to carbon dating. *PLoS One*,
6(6).
- Kočár, P., Čech, P., Kozáková, R., & Kočárová, R. 2010. Environment and economy of
the early medieval settlement in Žatec. *Interdisciplinaria Archaeologica*, 1, 45-60.
- Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. 2014. ANGSD: analysis of next
generation sequencing data. *BMC bioinformatics*, 15(1), 356.
- Kreuz, A., & Schäfer, E. 2002. A new archaeobotanical database program. *Vegetation
history and archaeobotany*, 11(1-2), 177-180.
- Kroll, H. 2001. Literature on archaeological remains of cultivated plants (1999/2000).
Vegetation history and archaeobotany, 10(1), 33-60.
- Kroll, H. 2005. Literature on archaeological remains of cultivated plants 1981–2004.
Available at archaeobotany.de/database.html. Accessed January, 2, 2016.

- Langfelder, P., & Horvath, S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9(1), 559.
- Lee, T. H., Guo, H., Wang, X., Kim, C., & Paterson, A. H. 2014. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC genomics*, 15(1), 162.
- Liddell, H. G., & Scott, R. 1940. A Greek-English Lexicon Perseus.
- Lizgunova, T. V. 1959. The history of botanical studies of the cabbage, *Brassica oleracea* L. *Bulletin of Applied Botany, Genetics and Plant Breeding*, 32, 37-70.
- Love, M.I., Huber, W. & Anders, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550.
- Maggioni, L., von Bothmer, R., Poulsen, G., & Branca, F. 2010. Origin and domestication of cole crops (*Brassica oleracea* L.): linguistic and literary considerations. *Economic botany*, 64(2), 109-123.
- Maggioni, L., von Bothmer, R., Poulsen, G., & Lipman, E. 2018. Domestication, diversity and use of *Brassica oleracea* L., based on ancient Greek and Latin texts. *Genetic resources and crop evolution*, 65(1), 137-159.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9), 1297-1303.
- Meehl, G.A., Washington, W.M., Arblaster, J.M., Hu, A., Teng, H., Tebaldi, C., et al. 2012. Climate system response to external forcings and climate change projections in CCSM4. *Journal of Climate*, 25(11), pp.3661-3683.
- Meisner, J., & Albrechtsen, A. 2018. Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, 210(2), 719-731.

- Mitchell, N. D. 1976. The status of *Brassica oleracea* L. subsp. *oleracea* (wild cabbage) in the British Isles. *Watsonia*, 11, 97-103.
- Mitchell, N. D., & Richards, A. J. 1979. *Brassica oleracea* L. ssp. *oleracea* (B. sylvestris (L.) Miller). *The Journal of Ecology*, 1087-1096.
- Moltsen, A.S.A. 2002. Arkæobotaniske undersøgelser fra middelalderbyen København - metodik og udvalgte eksempler in Viklund, K. Ed., Nordic archaeobotany - NAG 2000 in Umeå. *Archaeology and Environment* 15. (Univ. Umeå, Umeå). pp. 173-180.
- Murray, M. A. 2000. *Fruits, vegetables, pulses and condiments*. Cambridge University Press.
- Neutrofal, F. 1927. Zytologische Studien über die Kulturrassen von *Brassica oleracea*. *Oesterr. Bot. Z.* 76: 105–115.
- Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, 32(1), 268-274.
- Patro, R., Duggal, G., & Kingsford, C. 2015. Salmon: accurate, versatile and ultrafast quantification from RNA-seq data using lightweight-alignment. *Biorxiv*, 021592.
- Parkin, I. A., Koh, C., Tang, H., Robinson, S. J., Kagale, S., Clarke, W. E., et al. 2014. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. *Genome biology*, 15(6), R77.
- Phillips, S.J., Dudík, M. and Schapire, R.E. 2017. Maxent software for modeling species niches and distributions (Version 3.4. 1). *Biodiversity Informatics*.

- Pickrell, J., & Pritchard, J. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *Nature Precedings*, 1-1.
- Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., et al. 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 201178.
- Post, G. E. 1932. *Flora of Syria, Palestine, and Sinai* (Vol. 1). Рипол Классик.
- Prescott, J. 1989. [1315-1395] le Viandier de Taillevent: 14th Century Cookery, Based on the Vatican Library Manuscript Translated into English by James Prescott. (Hypatia Press, Eugene, ed. 2).
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3), 559-575.
- Rackham, H. & Pliny (the Elder). 1950. Natural History with an English Translation: Vol 5: Libri XVII-XIX. Harvard University Press.
- Raj, A., Stephens, M., & Pritchard, J. K. 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, 197(2), 573-589.
- Reihl, S. 2014. Archaeobotanical database of Eastern Mediterranean and Near Eastern sites. <http://www.ademnes.de/>.
- Ritchie, ME, Phipson, B, Wu, D, Hu, Y, Law, CW, Shi, W, & Smyth, GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43 (7), e47
- Schiemann, E. 1932. *Entstehung der Kulturpflanzen*. Gebrüder Borntraeger.

- Schmidl, A., & Oeggel, K. 2005. Subsistence strategies of two Bronze Age hill-top settlements in the eastern Alps—Friaga/Bartholomäberg (Vorarlberg, Austria) and Ganglegg/Schluderns (South Tyrol, Italy). *Vegetation History and Archaeobotany*, 14(4), 303-312.
- Schrempf, D., Minh, B. Q., De Maio, N., von Haeseler, A., & Kosiol, C. 2016. Reversible polymorphism-aware phylogenetic models and their application to tree inference. *Journal of theoretical biology*, 407, 362-370.
- Shultz, O. E. 1936 in *Die Natürlichen Pflanzfamilien*, No. 2, Vol. 17b, pp. 227-658, edited by A. Engler and H. Harm.
- Shyam, P., Wu, X., & Bhat, S. R. 2012. History, evolution, and domestication of *Brassica* crops. *Plant breeding reviews*, 35, 19-84.
- Sillasoo, I. 2003. Plant depictions in late medieval religious art. People and nature in historical perspective. *Central European University & Archaeolingua*, Budapest, 377-393.
- Sillasoo, I. 2002. *Gardens and garden products in Medieval Turku, Estonia*, in Viklund, K. Ed., *Nordic archaeobotany - Archaeology and Environment* 15. (Univ. of Umeå, Umeå), pp. 181-192.
- Snogerup, S. 1980. The wild forms of the *Brassica oleracea* group ($2n=18$) and their possible relations to the cultivated ones. *The wild forms of the Brassica oleracea group ($2n=18$) and their possible relations to the cultivated ones.*, 121-132.
- Snogerup, S., Gustafsson, M. & Von Bothmer, R. 1990. *Brassica* sect. *Brassica* (Brassicaceae) I. Taxonomy and variation. *Willdenowia*, pp.271-365.

- Soneson, C., Love, M. I., & Robinson, M. D. 2015. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *FL1000Research*, 4.
- Song, K. M., Osborn, T. C., & Williams, P. H. 1988. *Brassica* taxonomy based on nuclear restriction fragment length polymorphisms (RFLPs). *Theoretical and Applied Genetics*, 75(5), 784-794.
- Song, K., Osborn, T. C., & Williams, P. H. 1990. *Brassica* taxonomy based on nuclear restriction fragment length polymorphisms (RFLPs). *Theoretical and Applied Genetics*, 79(4), 497-506.
- Strid, A., & Tan, K. 2002. Flora Hellenica Vol 2: Nymphaeaceae to Platanaceae.
- Swarup, V., & Brahmi, P. 2005 *Kole crops in plant genetic resources; horticultural crops*. Dhillon, B.S., Tyagi, R.K., Saxena, S., Randhawa, G.J. Eds. (Narosa Publishing House pvt Ltd New Delhi) pp. 75-88.
- Tomlinson, P., & Hall, A. R. 1996. A review of the archaeological evidence for food plants from the British Isles: an example of the use of the Archaeobotanical Computer Database (ABCD). *Internet Archaeology*, 1.
- Totelin, L. M. (Ed.). 2009. Hippocratic recipes: oral and written transmission of pharmacological knowledge in fifth-and fourth-century Greece (Vol. 34). Brill.
- Tóth, A. J., Daróczi-Szabó, L., Kovács, Z. E., Gál, E., & Bartosiewicz, L. 2010. In the light of the crescent moon: reconstructing environment and diet from an Ottoman-period deposit in sixteenth to seventeenth century Hungary. *In Integrating Zooarchaeology and Paleoethnobotany* (pp. 245-280). Springer, New York, NY.

- Traka, M. H., Saha, S., Huseby, S., Kopriva, S., Walley, P. G., Barker, G. C., et al. 2013. Genetic regulation of glucoraphanin accumulation in Beneforté® broccoli. *New Phytologist*, 198(4), 1085-1095.
- Tripodi, G., Verzera, A., Dima, G., Condurso, C., & Ragusa, S. 2012. *Brassica fruticulosa* Cyr. and *Brassica incana* Ten.(Brassicaceae) as Mediterranean traditional wild vegetables: a valuable source of bioactive compounds. *Journal of essential oil research*, 24(6), 539-545.
- Tusser, T. 1878. Five hundred pointes of good husbandrie (Vol. 8). English dialect society.
- Tutin, T.G. et al., *Flora Europaea*: Volume 1: Psiotaceae to Platanaceae. (Cambridge Univ. Press, Cambridge, 2003).
- Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, 43(1), 11-10.
- Van der Veen, M. 2001. *The botanical evidence*. in Maxfield, V.A., Peacock, D.P.S.. Mons Claudianus. Survey and excavation 1987-1993. Vol 2. Excavations: Part 1. Fouilles (IFAO [Inst Franç Archéol Orientale] 43. Paris), pp. 174-222; appendix 1-2; Figure 8.1-8.13.
- Van der Veen, M. 2011. *Consumption, trade and innovation* (Vol. 6). Africa Magna Verlag.
- Van der Veen, M., Livarda, A., & Hill, A. 2008. New plant foods in Roman Britain—dispersal and social access. *Environmental Archaeology*, 13(1), 11-36.

- Vavilov, N. I. 1951. *The origin, variation, immunity and breeding of cultivated plants* (Vol. 72, No. 6, p. 482). LWW.
- Vehling, J. D. 1936. trans. *Apicius Cookery and Dining in Imperial Rome*. Walter M. Hill, Chicago.
- West, M. L. 2011. *Studies in Greek elegy and iambus* (Vol. 14). Walter de Gruyter.
- Willerding, U. 2003. Obstarten und Nüsse; Garen-, Obst- und Weinbau; Gartenbau, Obstbau, Weinbau in Benecke, N., Donat, P., Gringmuth-Dallmer, E., Willerding, U. Eds., *Frühgeschichte der Landwirtschaft in Deutschland Beitr Ur- u Frühgesch Mitteleuropa* 14. Langenweißbach, pp.30-31, pp. 162-172, pp. 258-262.
- Wieserowa, A. 1979. Plant remains from the Early and Late Middle Ages found in the settlement layers of the Main Market Square in *Cracow*. *Acta Palaeobotanica*, 20(2), 137-212.
- Wiethold, J. 2012. Hirse, Hanf und Hohldotter-Pflanzenfunde aus einem römischen Brunnen in Otterbach, Kr. Kaiserslautern [Restes végétaux provenant d'un puits gallo-romain à Ottersbach, Kr. Kaiserslautern].
- Wiethold, J. 2003a. Archäobotanische Ergebnisse. in Fries, H., Wiethold, J. Bemerkenswertes aus Stralsunds Altstadt die Grabung Apollonienmarkt 6 und ihre Ergebnisse. (Archäol Ber Mecklenburg-Vorpommern 10), pp. 220-247.
- Wiethold, J. 2003b. Archäobotanische Untersuchungen zur Ernährungs- und Wirtschaftsgeschichte des Mittelalters und der frühen Neuzeit. in Noël, R., Paquay, I., Sosson, J.-P. Eds., *Au-delà de l'écrit. Les hommes et leurs vécus matériels au Moyen Âge à la lumière des sciences et des techniques. Nouvelles perspectives*. (Louvain-la-Neuve), pp. 461-499.

- Wiethold, J. 2002. Giff in de schottele. Strowe dar peper up ...Botanische Funde als Quellen zur mittel-alterlichen Ernährungs- und Umweltgeschichte in Einbeck. in Heege A. (Hrsg.), Einbeck im Mittelalter : eine archäologisch-historische Spurensuche. Studien zur Einbecker Geschichte 17 (Oldenburg), pp. 240-246.
- Wiethold, J. 2000. Ernährung und Umwelt im spätmittelalterlichen Rostock. Archäobotanische Ergebnisse der Analyse zweier Kloaken in der Kröpeliner Str. 55–56/Kuhstraße. Bodendenkmalpflege Mecklenburg-Vorpommern, Jahrbuch, 47, 351-378.
- Willis, R. J. 2007. *The history of allelopathy*. Springer Science & Business Media.
- Yonge, C. D. 1854. *Athenaeus of Naucratis. The deipnosophists, or, Banquet of the learned of Athenæus*. London: Henry G. Bohn.
- Zemanek, A., & Wasylikowa, K. 1996. Historia botaniki i archeobotanika w poszukiwaniu danych o użytkowaniu roślin w średniowiecznym Krakowie. *Analecta: studia i materiały z dziejów nauki*, 5(1 (9)), 123-138.
- Zeven, A. C., & Brandenburg, W. A. 1986. Use of paintings from the 16th to 19th centuries to study the history of domesticated plants. *Economic Botany*, 40(4), 397-408.
- Zoltán, A., & Sárghadinnye, A. 2006. (*Cucumis melo*) Archeogenetikája, ITS- és SSR – Heterogenitása Egy 15. századi Lelettől a Mai Fajtákig, *Unpublished Doctoral Thesis*. Gödöllő..

Supplemental Figure 3.1 PCA of **A)** PC2 vs PC3 and **B)** PC3 vs PC4 of wild and cultivar samples. Triangles = wild samples, circles = cultivars. Black triangles with colored outlines = wild C - clade 2 samples with species identification indicated by outline color.

Supplemental Figure 3.3 Introgression tests performed. Samples in red indicate signals for introgression, green are evolving tree like, and in black are mixed in results.

Supplemental Figure 3.4 Expression PCA for **A)** PC1 vs PC2, **B)** PC2 vs PC3, **C)** PC3 vs PC4. Triangles = wild samples, circles = cultivars. Black triangles with colored outlines = wild C - clade 2 samples with species identification indicated by outline color.

Supplemental Figure 3.5 *Brassica oleracea* cultivars and wild relative dendrogram based on expression profiles. Wild types indicated by unique color, cultivars in grey. Wild C - clade 2 indicated in red in bottom bar plot.

Supplemental Figure 3.6 Module assignment and dendrogram for the top 1000 variable genes in expression.

Supplemental Table 3.1 Sample information with Species, variety, cultivar, accession, sampleID, and phenotypes used in analyses.

Genus	Species	Variety	Cultivar	Accession	Sample	Phenotype	Group	Type
<i>Brassica</i>	<i>oleracea</i>	<i>sabauda</i>	Savoy Cabbage	PI246117	100 S19	sabauda	domesticata	heading
<i>Brassica</i>	<i>oleracea</i>	<i>sabauda</i>	Savoy Cabbage	PI507856	101 S20	sabauda	domesticata	heading
<i>Brassica</i>	<i>oleracea</i>	<i>sabauda</i>	Savoy Cabbage	GroenkopVanHemert+Co	102 S21	sabauda	domesticata	heading
<i>Brassica</i>	<i>oleracea</i>	<i>ramosa</i>	Perpetual Kale	IPK:BRA1213	103 S22	ramosa	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>ramosa</i>	Perpetual Kale	IPK:BRA2952	104 S23	ramosa	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>ramosa</i>	Perpetual Kale	IPK:BRA1891	105 S3	ramosa	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>ramosa</i>	Perpetual Kale	IPK:BRA3138	106 S24	ramosa	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>ramosa</i>	Perpetual Kale	IPK:K9829	107 S1	ramosa	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>ramosa</i>	Perpetual Kale	IPK:BRA2782	108 S2	ramosa	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>ramosa</i>	Perpetual Kale	IPK:BRA1259	109 S3	ramosa	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>gemmifera</i>	Brussels Sprouts	G30872	10 S10	gemmifera	domesticata	heading
<i>Brassica</i>	<i>oleracea</i>	<i>ramosa</i>	Perpetual Kale	IPK:BRA67	110 S4	ramosa	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>ramosa</i>	Perpetual Kale	IPK:BRA2905	111 S5	ramosa	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>ramosa</i>	Perpetual Kale	IPK:BRA2382	112 S2	ramosa	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>ramosa</i>	Perpetual Kale	IPK:BRA2784	113 S6	ramosa	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>ramosa</i>	Perpetual Kale	IPK:BRA2956	114 S7	ramosa	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>ramosa</i>	Perpetual Kale	G30718	115 S8	ramosa	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>ramosa</i>	Perpetual Kale	G30724	116 S9	ramosa	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>botrytis</i>	Cauliflower	PI269312	117 S10	botrytis	domesticata	heading
<i>Brassica</i>	<i>oleracea</i>	<i>botrytis</i>	Cauliflower	PI291996	118 S11	botrytis	domesticata	heading
<i>Brassica</i>	<i>oleracea</i>	<i>botrytis</i>	Cauliflower	PI231208	119 S7	botrytis	domesticata	heading
<i>Brassica</i>	<i>oleracea</i>	<i>gemmifera</i>	Brussels Sprouts	G30870	11 S11	gemmifera	domesticata	heading
<i>Brassica</i>	<i>oleracea</i>	<i>botrytis</i>	Cauliflower	PI271445	120 S12	botrytis	domesticata	heading
<i>Brassica</i>	<i>oleracea</i>	<i>botrytis</i>	Cauliflower	G30857	121 S13	botrytis	domesticata	heading
<i>Brassica</i>	<i>oleracea</i>	<i>botrytis</i>	Cauliflower	G28888	122 S14	botrytis	domesticata	heading
<i>Brassica</i>	<i>oleracea</i>	<i>botrytis</i>	Cauliflower	G30769	123 S15	botrytis	domesticata	heading
<i>Brassica</i>	<i>oleracea</i>	<i>botrytis</i>	Cauliflower	G30438	124 S16	botrytis	domesticata	heading
<i>Brassica</i>	<i>oleracea</i>	<i>botrytis</i>	Cauliflower	G30436	125 S17	botrytis	domesticata	heading
<i>Brassica</i>	<i>oleracea</i>	<i>botrytis</i>	Cauliflower	G30435	126 S18	botrytis	domesticata	heading
<i>Brassica</i>	<i>oleracea</i>	<i>botrytis</i>	Cauliflower	PI267724	127 S8	botrytis	domesticata	heading
<i>Brassica</i>	<i>oleracea</i>	<i>botrytis</i>	Cauliflower	HIRGRU007826	128 S19	botrytis	domesticata	heading
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	TO1000	129 S20	alboglabra	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>gemmifera</i>	Brussels Sprouts	G30867	12 S12	gemmifera	domesticata	heading
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	HIRGRU007543	130 S21	alboglabra	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	A12	131 S22	alboglabra	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	HIRGRU013023	132 S23	alboglabra	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	IPK:BRA1210	133 S24	alboglabra	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	IPK:BRA1006	134 S1	alboglabra	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	IPK:BRA990	135 S2	alboglabra	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	IPK:BRA1909	136 S3	alboglabra	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	IPK:BRA169	137 S4	alboglabra	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	IPK:BRA1658	138 S9	alboglabra	domesticata	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	IPK:BRA1143	139 S5	alboglabra	domesticata	leafy

<i>Brassica</i>	<i>oleracea</i>	<i>gemmifera</i>	Brussels Sprouts	PI365170	13 S13	<i>gemmifera</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	IPK:BRA1747	140 S6	<i>alboglabra</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	IPK:BRA1211	141 S7	<i>alboglabra</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	IPK:BRA1244	142 S8	<i>alboglabra</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	Bot.Interests-Kailann,whiteflowered	143 S9	<i>alboglabra</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>capitata</i>	Ornamental cabbage	PI662701	144 S10	<i>capitata</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	TO1000	145 S11	<i>alboglabra</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	HIRGRU007543	146 S4	<i>alboglabra</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	HIRGRU013023	147 S12	<i>alboglabra</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	A12	148 S13	<i>alboglabra</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	PI249556	149 S5	<i>alboglabra</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>gemmifera</i>	Brussels Sprouts	PI209942	14 S14	<i>gemmifera</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	PI435900	150 S14	<i>alboglabra</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>alboglabra</i>	Chinese White Kale	PI662520	151 S15	<i>alboglabra</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>botrytis</i>	Cauliflower	HIRGRU007826	152 S16	<i>botrytis</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>botrytis</i>	Romanesco	Bot.Interests-Romanesco	153 S10	<i>botrytis</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>botrytis</i>	Romanesco	UrbanFarmer-Romanesco	154 S11	<i>botrytis</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>botrytis</i>	Romanesco	EdenBros-Romanesco	155 S17	<i>botrytis</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>botrytis</i>	Romanesco	SeedSavers-Romanesco	156 S6	<i>botrytis</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>botrytis</i>	Cauliflower	PI267724	157 S18	<i>botrytis</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>capitata</i>	Red Cabbage	PI187232	159 S19	<i>capitata</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>gemmifera</i>	Brussels Sprouts	PI243050	15 S15	<i>gemmifera</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>capitata</i>	Red Cabbage	PI329197	160 S20	<i>capitata</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>capitata</i>	Red Cabbage	PI246060	161 S21	<i>capitata</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>capitata</i>	Red Cabbage	PI246059	162 S22	<i>capitata</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>capitata</i>	Red Cabbage	PI246046	163 S23	<i>capitata</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>capitata</i>	Red Cabbage	PI245012	164 S7	<i>capitata</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>costata</i>	Tronchuda Kale	PI662668	165 S3	<i>costata</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>gongylodes</i>	Kohlrabi	HIRGRU005389	167 S24	<i>gongylodes</i>	domesticate	tuber
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	Early-Big	168 S12	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	Mar34-DH	169 S1	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>gemmifera</i>	Brussels Sprouts	PI244839	16 S1	<i>gemmifera</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	GD33-DH	170 S2	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>longata</i>	walking stick kale	TenerifeWSKale	171 S3	<i>longata</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>longata</i>	walking stick kale	alsagarden	172 S4	<i>longata</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>longata</i>	walking stick kale	mm1	173 S5	<i>longata</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>palmifolia</i>	Giant Jersey Kale	G30723	174 S6	<i>palmifolia</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>oleracea</i>	Wild oleracea	G30186	175 S7	<i>oleracea</i>	wild	wild
<i>Brassica</i>	<i>oleracea</i>	<i>oleracea</i>	Wild oleracea	NGB21675	176 S8	<i>oleracea</i>	wild	wild
<i>Brassica</i>	<i>oleracea</i>	<i>oleracea</i>	Wild oleracea	NGB16241	177 S9	<i>oleracea</i>	wild	wild
<i>Brassica</i>	<i>oleracea</i>	<i>viridis</i>	Collard Greens	PI66281599OI	178 S10	<i>viridis</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>viridis</i>	Collard Greens	PI66281706GI	179 S11	<i>viridis</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>gemmifera</i>	Brussels Sprouts	PI312902	17 S16	<i>gemmifera</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>viridis</i>	Collard Greens	PI66279606GI	180 S12	<i>viridis</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>viridis</i>	Collard Greens	PI66280606GI	181 S13	<i>viridis</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>viridis</i>	Collard Greens	PI66281406GI	182 S14	<i>viridis</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>viridis</i>	Collard Greens	PI66280906GI	183 S15	<i>viridis</i>	domesticate	leafy

<i>Brassica</i>	<i>oleracea</i>	<i>viridis</i>	Collard Greens	PI66280406GI	184 S16	<i>viridis</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>viridis</i>	Collard Greens	PI66283106GI	185 S17	<i>viridis</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>viridis</i>	Collard Greens	PI66280706GI	186 S18	<i>viridis</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>viridis</i>	Collard Greens	PI66283307GI	187 S19	<i>viridis</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>viridis</i>	Collard Greens	PI66282706GI	188 S20	<i>viridis</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>viridis</i>	Collard Greens	PI66280506GI	189 S21	<i>viridis</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>gemmifera</i>	Brussels Sprouts	PI343670	18 S17	<i>gemmifera</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>viridis</i>	Collard Greens	PI66279806GI	190 S22	<i>viridis</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>viridis</i>	Collard Greens	PI66280806GI	191 S23	<i>viridis</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>viridis</i>	Collard Greens	PI66282806GI	192 S24	<i>viridis</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>viridis</i>	Collard Greens	PI66281606GI	193 S13	<i>viridis</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>viridis</i>	Collard Greens	PI66284104OI	194 S14	<i>viridis</i>	domesticate	leafy
<i>Brassica</i>	<i>cretica</i>		Wild C	PI662588	195 S4	<i>cretica</i>	wild	wild
<i>Brassica</i>	<i>cretica</i>		Wild C	UPM6346	196 S5	<i>cretica</i>	wild	wild
<i>Brassica</i>	<i>cretica</i>		Wild C	BRA3053	198 S6	<i>cretica</i>	wild	wild
<i>Brassica</i>	<i>cretica</i>		Wild C	BRA3092	199 S15	<i>cretica</i>	wild	wild
<i>Brassica</i>	<i>oleracea</i>	<i>gemmifera</i>	Brussels Sprouts	PI249534	19 S18	<i>gemmifera</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>palmifolia</i>	Lacinato kale	AriasCollection	1 S1	<i>palmifolia</i>	domesticate	leafy
<i>Brassica</i>	<i>hilarionis</i>		Wild C	HIRGRU011463	203 S16	<i>hilarionis</i>	wild	wild
<i>Brassica</i>	<i>incana</i>		Wild C	BRA1262	204 S17	<i>incana</i>	wild	wild
<i>Brassica</i>	<i>incana</i>		Wild C	BRA2918	205 S7	<i>incana</i>	wild	wild
<i>Brassica</i>	<i>incana</i>		Wild C	UPM5974	207 S18	<i>incana</i>	wild	wild
<i>Brassica</i>	<i>incana</i>		Wild C	PI662584	208 S19	<i>incana</i>	wild	wild
<i>Brassica</i>	<i>incana</i>		Wild C	PI662591	209 S8	<i>incana</i>	wild	wild
<i>Brassica</i>	<i>oleracea</i>	<i>gemmifera</i>	Brussels Sprouts	PI343671	20 S19	<i>gemmifera</i>	domesticate	heading
<i>Brassica</i>	<i>insularis</i>		Wild C	PI662587	212 S20	<i>insularis</i>	wild	wild
<i>Brassica</i>	<i>insularis</i>		Wild C	BRA2996	213 S21	<i>insularis</i>	wild	wild
<i>Brassica</i>	<i>insularis</i>		Wild C	BRA3050	214 S22	<i>insularis</i>	wild	wild
<i>Brassica</i>	<i>insularis</i>		Wild C	BRA3051	215 S9	<i>insularis</i>	wild	wild
<i>Brassica</i>	<i>insularis</i>		Wild C	BRA3052	216 S8	<i>insularis</i>	wild	wild
<i>Brassica</i>	<i>macrocarpa</i>		Wild C	BRA2854	218 S10	<i>macrocarpa</i>	wild	wild
<i>Brassica</i>	<i>macrocarpa</i>		Wild C	BRA2944	219 S11	<i>macrocarpa</i>	wild	wild
<i>Brassica</i>	<i>oleracea</i>	<i>gemmifera</i>	Brussels Sprouts	PI261775	21 S20	<i>gemmifera</i>	domesticate	heading
<i>Brassica</i>	<i>macrocarpa</i>		Wild C	CO7031	220 S23	<i>macrocarpa</i>	wild	wild
<i>Brassica</i>	<i>macrocarpa</i>		Wild C	PI662585	221 S9	<i>macrocarpa</i>	wild	wild
<i>Brassica</i>	<i>montana</i>		Wild C	BRA1644	222 S24	<i>montana</i>	wild	wild
<i>Brassica</i>	<i>montana</i>		Wild C	GC3607-75	224 S12	<i>montana</i>	wild	wild
<i>Brassica</i>	<i>rupestris</i>		Wild C	BRA2851	226 S10	<i>rupestris</i>	wild	wild
<i>Brassica</i>	<i>rupestris</i>		Wild C	BRA2945	227 S11	<i>rupestris</i>	wild	wild
<i>Brassica</i>	<i>rupestris</i>		Wild C	BRA2992	228 S13	<i>rupestris</i>	wild	wild
<i>Brassica</i>	<i>rupestris</i>		Wild C	GC3822-75	229 S14	<i>rupestris</i>	wild	wild
<i>Brassica</i>	<i>oleracea</i>	<i>gemmifera</i>	Brussels Sprouts	PI343669	22 S21	<i>gemmifera</i>	domesticate	heading
<i>Brassica</i>	<i>rupestris</i>		Wild C	K10259	230 S12	<i>rupestris</i>	wild	wild
<i>Brassica</i>	<i>rupestris</i>		Wild C	K10260	231 S13	<i>rupestris</i>	wild	wild
<i>Brassica</i>	<i>rupestris</i>		Wild C	UPM6575	232 S14	<i>rupestris</i>	wild	wild
<i>Brassica</i>	<i>villosa</i>		Wild C	BRA1896	233 S15	<i>villosa</i>	wild	wild
<i>Brassica</i>	<i>villosa</i>		Wild C	BRA2853	234 S16	<i>villosa</i>	wild	wild

<i>Brassica</i>	<i>villosa</i>		Wild C	K10263	236 S15	<i>villosa</i>	wild	wild
<i>Brassica</i>	<i>villosa</i>		Wild C	K10264	237 S16	<i>villosa</i>	wild	wild
<i>Brassica</i>	<i>villosa</i>		Wild C	UPM6581	238 S17	<i>villosa</i>	wild	wild
<i>Brassica</i>	<i>oleracea</i>	<i>gemmifera</i>	Brussels Sprouts	PI343673	23 S22	<i>gemmifera</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>gemmifera</i>	Brussels Sprouts	G30869	24 S23	<i>gemmifera</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>costata</i>	Tronchuda Kale	PI662702	25 S24	<i>costata</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>costata</i>	Tronchuda Kale	TronchudaBeiraRenee'sGard	26 S1	<i>costata</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>costata</i>	Tronchuda Kale	PI662565	27 S2	<i>costata</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>costata</i>	Tronchuda Kale	PI662564	28 S3	<i>costata</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>costata</i>	Tronchuda Kale	PI662668	29 S1	<i>costata</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>palmifolia</i>	Lacinato kale	BotanicalInterests	2 S2	<i>palmifolia</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>costata</i>	Tronchuda Kale	PI662659	30 S1	<i>costata</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>costata</i>	Tronchuda Kale	PI662566	31 S4	<i>costata</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>costata</i>	Tronchuda Kale	PI662567	32 S2	<i>costata</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>costata</i>	Tronchuda Kale	PI662568	33 S2	<i>costata</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>costata</i>	Tronchuda Kale	PI662569	34 S5	<i>costata</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>medullosa</i>	Marrow Cabbage	G30717	35 S6	<i>medullosa</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>medullosa</i>	Marrow Cabbage	G30859	36 S3	<i>medullosa</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>medullosa</i>	Marrow Cabbage	G30862	37 S7	<i>medullosa</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>medullosa</i>	Marrow Cabbage	PI662703	38 S8	<i>medullosa</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>medullosa</i>	Marrow Cabbage	IPK:CR2545	39 S4	<i>medullosa</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>palmifolia</i>	Lacinato kale	Renee'sGarden	3 S3	<i>palmifolia</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>medullosa</i>	Marrow Cabbage	IPK:CR1431	40 S9	<i>medullosa</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>medullosa</i>	Marrow Cabbage	IPK:CR1454	41 S10	<i>medullosa</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>medullosa</i>	Marrow Cabbage	IPK:CR1317	42 S11	<i>medullosa</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>medullosa</i>	Marrow Cabbage	IPK:CR1256	43 S12	<i>medullosa</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>medullosa</i>	Marrow Cabbage	IPK:CR1235	44 S13	<i>medullosa</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>medullosa</i>	Marrow Cabbage	IPK:CR1410	45 S14	<i>medullosa</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>medullosa</i>	Marrow Cabbage	IPK:CR1295	46 S15	<i>medullosa</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>medullosa</i>	Marrow Cabbage	IPK:CR1238	47 S16	<i>medullosa</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>sabellica</i>	Curly Kale	IPK:BRA2456	48 S17	<i>sabellica</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>sabellica</i>	Curly Kale	IPK:BRA2449	49 S18	<i>sabellica</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>palmifolia</i>	Lacinato kale	IPK:BRA1906	4 S4	<i>palmifolia</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>sabellica</i>	Curly Kale	IPK:BRA2187	50 S19	<i>sabellica</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>sabellica</i>	Curly Kale	IPK:BRA2239	51 S20	<i>sabellica</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>sabellica</i>	Curly Kale	IPK:BRA2286	52 S21	<i>sabellica</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>sabellica</i>	Curly Kale	IPK:BRA1841	53 S22	<i>sabellica</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>sabellica</i>	Curly Kale	IPK:BRA1491	54 S23	<i>sabellica</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>sabellica</i>	Curly Kale	IPK:BRA1899	55 S24	<i>sabellica</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>sabellica</i>	Curly Kale	IPK:BRA960	56 S1	<i>sabellica</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>sabellica</i>	Curly Kale	IPK:BRA963	57 S2	<i>sabellica</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>sabellica</i>	Curly Kale	IPK:BRA964	58 S3	<i>sabellica</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>sabellica</i>	Curly Kale	IPK:BRA2116	59 S4	<i>sabellica</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>palmifolia</i>	Lacinato kale	IPK:BRA1905	5 S5	<i>palmifolia</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	PI231210	60 S5	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	PI662597	61 S5	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	PI662596	62 S6	<i>italica</i>	domesticate	heading

<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	PI662524	63 S7	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	PI662674	64 S8	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	PI662529	65 S9	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	PI662712	66 S10	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	PI662786	67 S11	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	PI662525	68 S12	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	G30778	69 S13	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>palmifolia</i>	Lacinato kale	IPK:BRA2846	6 S6	<i>palmifolia</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	G30416	70 S14	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	G28865	71 S15	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	G21111	72 S16	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	G30779	73 S17	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	Mar34-DH	74 S18	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	GD33-DH	75 S19	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>italica</i>	Broccoli	Early-Big	76 S20	<i>italica</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>gongylodes</i>	Green Kohlrabi	PI662563	77 S21	<i>gongylodes</i>	domesticate	tuber
<i>Brassica</i>	<i>oleracea</i>	<i>gongylodes</i>	Green Kohlrabi	PI662704	78 S22	<i>gongylodes</i>	domesticate	tuber
<i>Brassica</i>	<i>oleracea</i>	<i>gongylodes</i>	Green Kohlrabi	PI662623	79 S23	<i>gongylodes</i>	domesticate	tuber
<i>Brassica</i>	<i>oleracea</i>	<i>palmifolia</i>	Lacinato kale	IPK:BRA3073	7 S7	<i>palmifolia</i>	domesticate	leafy
<i>Brassica</i>	<i>oleracea</i>	<i>gongylodes</i>	Green Kohlrabi	PI662560	80 S24	<i>gongylodes</i>	domesticate	tuber
<i>Brassica</i>	<i>oleracea</i>	<i>gongylodes</i>	Green Kohlrabi	PI662671	81 S1	<i>gongylodes</i>	domesticate	tuber
<i>Brassica</i>	<i>oleracea</i>	<i>gongylodes</i>	Green Kohlrabi	PI662624	82 S2	<i>gongylodes</i>	domesticate	tuber
<i>Brassica</i>	<i>oleracea</i>	<i>gongylodes</i>	Green Kohlrabi	PI662625	83 S6	<i>gongylodes</i>	domesticate	tuber
<i>Brassica</i>	<i>oleracea</i>	<i>gongylodes</i>	Purple Kohlrabi	G30935	84 S3	<i>gongylodes</i>	domesticate	tuber
<i>Brassica</i>	<i>oleracea</i>	<i>gongylodes</i>	Purple Kohlrabi	IPK:BRA1889	85 S4	<i>gongylodes</i>	domesticate	tuber
<i>Brassica</i>	<i>oleracea</i>	<i>gongylodes</i>	Purple Kohlrabi	PI096969	86 S5	<i>gongylodes</i>	domesticate	tuber
<i>Brassica</i>	<i>oleracea</i>	<i>gongylodes</i>	Purple Kohlrabi	PI188610	87 S6	<i>gongylodes</i>	domesticate	tuber
<i>Brassica</i>	<i>oleracea</i>	<i>gongylodes</i>	Purple Kohlrabi	IPK:BRA42	88 S7	<i>gongylodes</i>	domesticate	tuber
<i>Brassica</i>	<i>oleracea</i>	<i>gongylodes</i>	Purple Kohlrabi	IPK:BRA2349	89 S8	<i>gongylodes</i>	domesticate	tuber
<i>Brassica</i>	<i>oleracea</i>	<i>gemmifera</i>	Brussels Sprouts	PI385957	8 S8	<i>gemmifera</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>gongylodes</i>	Purple Kohlrabi	G30762	90 S9	<i>gongylodes</i>	domesticate	tuber
<i>Brassica</i>	<i>oleracea</i>	<i>capitata</i>	Cabbage	PI320913	91 S10	<i>capitata</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>capitata</i>	Cabbage	PI324236	92 S11	<i>capitata</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>capitata</i>	Cabbage	PI343514	93 S12	<i>capitata</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>capitata</i>	Cabbage	PI263061	94 S13	<i>capitata</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>capitata</i>	Cabbage	PI518837	95 S14	<i>capitata</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>capitata</i>	Cabbage	PI182149	96 S15	<i>capitata</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>capitata</i>	Cabbage	PI171529	97 S16	<i>capitata</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>capitata</i>	Cabbage	PI225859	98 S17	<i>capitata</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>sabauda</i>	Savoy Cabbage	PI343603	99 S18	<i>sabauda</i>	domesticate	heading
<i>Brassica</i>	<i>oleracea</i>	<i>gemmifera</i>	Brussels Sprouts	G30601	9 S9	<i>gemmifera</i>	domesticate	heading

Supplemental Table 3.2 F4 results, standard error, and Z-scores for samples tested. Bold rows indicated significant difference from 0, meaning introgression has occurred.

A	B	C	D	F4	Standard error	Z-score
villosa	rupestris	incana 207	alboglabra	0.00023707	0.000251059	0.944297
villosa	rupestris	incana 207	botrytis	-0.0002034	0.000348798	-0.58324
villosa	rupestris	incana 207	capitata	0.00024372	0.000239324	1.01835
villosa	rupestris	incana 207	costata	-0.0002089	0.000249182	-0.838172
villosa	rupestris	incana 207	gemmaefera	-0.0002678	0.000298926	-0.895879
villosa	rupestris	incana 207	gongylodes	0.00040561	0.000247749	1.63718
villosa	rupestris	incana 207	italica	-0.0005715	0.000304867	-1.87464
villosa	rupestris	incana 207	longata	0.00024014	0.000331068	0.725357
villosa	rupestris	incana 207	medullosa	0.00028387	0.000243447	1.16602
villosa	rupestris	incana 207	palmifolia	0.00029966	0.000269983	1.1099
villosa	rupestris	incana 207	ramosa	0.00020889	0.000253683	0.823412
villosa	rupestris	incana 207	sabauda	0.00022713	0.000258791	0.877668
villosa	rupestris	incana 207	sabellica	2.28E-05	0.000270922	0.0842503
villosa	rupestris	incana 207	viridis	0.00034021	0.000238328	1.42748
villosa	rupestris	incana 204	alboglabra	1.93E-05	0.0002182	0.0882659
villosa	rupestris	incana 204	botrytis	1.44E-05	0.000309978	0.0463931
villosa	rupestris	incana 204	capitata	2.59E-05	0.000199395	0.129899
villosa	rupestris	incana 204	costata	8.96E-06	0.000189282	0.0473207
villosa	rupestris	incana 204	gemmaefera	-5.00E-05	0.000190726	-0.262092
villosa	rupestris	incana 204	gongylodes	0.0001878	0.000211333	0.88863
villosa	rupestris	incana 204	italica	-0.0003537	0.000273732	-1.29215
villosa	rupestris	incana 204	longata	2.23E-05	0.00028798	0.0775342
villosa	rupestris	incana 204	medullosa	6.61E-05	0.000174152	0.379268
villosa	rupestris	incana 204	palmifolia	8.18E-05	0.000201892	0.405372
villosa	rupestris	incana 204	ramosa	-8.93E-06	0.000166644	-0.05358
villosa	rupestris	incana 204	sabauda	9.32E-06	0.000199471	0.0467159
villosa	rupestris	incana 204	sabellica	-0.000195	0.000179777	-1.08462
villosa	rupestris	incana 204	viridis	0.00012239	0.000175707	0.696583
villosa	rupestris	cretica 195	alboglabra	-6.25E-05	0.000298549	-0.209482
villosa	rupestris	cretica 195	botrytis	9.62E-05	0.000354986	0.270943
villosa	rupestris	cretica 195	capitata	-5.59E-05	0.000282004	-0.198221
villosa	rupestris	cretica 195	costata	9.08E-05	0.000280609	0.32343
villosa	rupestris	cretica 195	gemmaefera	3.18E-05	0.000279203	0.11394
villosa	rupestris	cretica 195	gongylodes	0.000106	0.000309974	0.341954
villosa	rupestris	cretica 195	italica	-0.0002719	0.000350678	-0.775363
villosa	rupestris	cretica 195	longata	-5.95E-05	0.00033252	-0.178852
villosa	rupestris	cretica 195	medullosa	1.57E-05	0.000248729	0.0633215
villosa	rupestris	cretica 195	palmifolia	-4.11E-08	0.000243512	-0.0001688
villosa	rupestris	cretica 195	ramosa	-9.07E-05	0.000247572	-0.366476
villosa	rupestris	cretica 195	sabauda	-7.25E-05	0.000269218	-0.269231
villosa	rupestris	cretica 195	sabellica	-0.0002768	0.000240809	-1.14941
villosa	rupestris	cretica 195	viridis	4.06E-05	0.000259094	0.156678
villosa	rupestris	incana 205	alboglabra	-0.0012924	0.000351526	-3.67658
villosa	rupestris	incana 205	botrytis	0.00132605	0.00041692	3.18059
villosa	rupestris	incana 205	capitata	-0.0012858	0.000359874	-3.57284
villosa	rupestris	incana 205	costata	0.00132063	0.000341155	3.87106
villosa	rupestris	incana 205	gemmaefera	0.00126169	0.000353239	3.57176
villosa	rupestris	incana 205	gongylodes	-0.0011239	0.000313321	-3.58697
villosa	rupestris	incana 205	italica	0.00095797	0.00039318	2.43647
villosa	rupestris	incana 205	longata	-0.0012894	0.000332826	-3.87393
villosa	rupestris	incana 205	medullosa	-0.0012456	0.000327121	-3.80784
villosa	rupestris	incana 205	palmifolia	-0.0012298	0.000300674	-4.09026
villosa	rupestris	incana 205	ramosa	-0.0013206	0.000319468	-4.13376
villosa	rupestris	incana 205	sabauda	-0.0013024	0.000333669	-3.90313
villosa	rupestris	incana 205	sabellica	-0.0015067	0.000329965	-4.56612
villosa	rupestris	incana 205	viridis	-0.0011893	0.000338778	-3.5105
villosa	rupestris	villosa 233	alboglabra	-0.0005665	0.000346423	-1.63527
villosa	rupestris	villosa 233	botrytis	0.00060014	0.000421632	1.42337
villosa	rupestris	villosa 233	capitata	-0.0005599	0.000364317	-1.53673
villosa	rupestris	villosa 233	costata	0.00059471	0.000323765	1.83687
villosa	rupestris	villosa 233	gemmaefera	0.00053577	0.000344822	1.55375
villosa	rupestris	villosa 233	gongylodes	-0.000398	0.000324247	-1.22733

villosa	rupestris	villosa_233	italica	0.00023205	0.000410267	0.565614
villosa	rupestris	villosa_233	longata	-0.0005634	0.000360855	-1.56137
villosa	rupestris	villosa_233	medullosa	0.00051971	0.000312704	1.66198
villosa	rupestris	villosa_233	palmifolia	0.00050392	0.000305032	1.65201
villosa	rupestris	villosa_233	ramosa	-0.0005947	0.000306828	-1.93817
villosa	rupestris	villosa_233	sabauda	-0.0005764	0.000325124	-1.77298
villosa	rupestris	villosa_233	sabellica	-0.0007807	0.000326091	-2.39426
villosa	rupestris	villosa_233	viridis	-0.0004634	0.000328966	-1.40854
villosa	rupestris	incana_209	alboglabra	-0.000559	0.000337467	-1.65636
villosa	rupestris	incana_209	botrytis	0.00059261	0.000396252	1.49553
villosa	rupestris	incana_209	capitata	-0.0005523	0.000317595	-1.73909
villosa	rupestris	incana_209	costata	0.00058718	0.000314141	1.86917
villosa	rupestris	incana_209	gemmaifera	0.00052824	0.000334798	1.57778
villosa	rupestris	incana_209	gongylodes	-0.0003904	0.000307106	-1.27132
villosa	rupestris	incana_209	italica	0.00022452	0.000348452	0.644344
villosa	rupestris	incana_209	longata	-0.0005559	0.000328183	-1.69386
villosa	rupestris	incana_209	medullosa	0.00051218	0.000289566	1.76877
villosa	rupestris	incana_209	palmifolia	0.00049639	0.000266851	1.86015
villosa	rupestris	incana_209	ramosa	-0.0005872	0.000291722	-2.01272
villosa	rupestris	incana_209	sabauda	-0.0005689	0.000303237	-1.87612
villosa	rupestris	incana_209	sabellica	-0.0007732	0.00027267	-2.83572
villosa	rupestris	incana_209	viridis	-0.0004558	0.000296656	-1.53657
villosa	rupestris	incana_208	alboglabra	-0.0007998	0.000375155	-2.13181
villosa	rupestris	incana_208	botrytis	0.0008334	0.000459908	1.8121
villosa	rupestris	incana_208	capitata	-0.0007931	0.000383202	-2.06971
villosa	rupestris	incana_208	costata	0.00082797	0.000360152	2.29896
villosa	rupestris	incana_208	gemmaifera	0.00076903	0.00036702	2.09534
villosa	rupestris	incana_208	gongylodes	-0.0006312	0.000363712	-1.7355
villosa	rupestris	incana_208	italica	0.00046531	0.0003994	1.16503
villosa	rupestris	incana_208	longata	-0.0007967	0.000396436	-2.00963
villosa	rupestris	incana_208	medullosa	0.00075297	0.000338837	2.22221
villosa	rupestris	incana_208	palmifolia	0.00073718	0.000333056	2.21337
villosa	rupestris	incana_208	ramosa	-0.0008279	0.000329658	-2.51153
villosa	rupestris	incana_208	sabauda	-0.0008097	0.000352642	-2.29609
villosa	rupestris	incana_208	sabellica	-0.001014	0.000330781	-3.06549
villosa	rupestris	incana_208	viridis	-0.0006966	0.000353975	-1.968
villosa	rupestris	cretica_199	alboglabra	-0.0003031	0.000199754	-1.51755
villosa	rupestris	cretica_199	botrytis	-0.0002695	0.000278437	-0.967886
villosa	rupestris	cretica_199	capitata	0.00030978	0.000204074	1.51797
villosa	rupestris	cretica_199	costata	-0.0002749	0.000175553	-1.56602
villosa	rupestris	cretica_199	gemmaifera	-0.0003339	0.00022496	-1.48411
villosa	rupestris	cretica_199	gongylodes	0.00047167	0.000176874	2.66672
villosa	rupestris	cretica_199	italica	-0.0006376	0.000274997	-2.3185
villosa	rupestris	cretica_199	longata	0.00030621	0.000259798	1.17863
villosa	rupestris	cretica_199	medullosa	-0.0003499	0.000184632	-1.89527
villosa	rupestris	cretica_199	palmifolia	-0.0003657	0.000204609	-1.7874
villosa	rupestris	cretica_199	ramosa	0.00027495	0.000179491	1.53182
villosa	rupestris	cretica_199	sabauda	0.0002932	0.000204041	1.43694
villosa	rupestris	cretica_199	sabellica	8.89E-05	0.000223647	0.397446
villosa	rupestris	cretica_199	viridis	0.00040627	0.000195827	2.07465
villosa	rupestris	oleracea_177	alboglabra	-0.0002966	0.000232874	-1.27367
villosa	rupestris	oleracea_177	botrytis	0.00033025	0.000290797	1.13566
villosa	rupestris	oleracea_177	capitata	-0.00029	0.000204552	-1.41755
villosa	rupestris	oleracea_177	costata	0.00032482	0.000189055	1.71813
villosa	rupestris	oleracea_177	gemmaifera	0.00026588	0.000230635	1.1528
villosa	rupestris	oleracea_177	gongylodes	-0.0001281	0.000186761	-0.68573
villosa	rupestris	oleracea_177	italica	-3.78E-05	0.000265632	-0.142447
villosa	rupestris	oleracea_177	longata	-0.0002935	0.000293626	-0.999695
villosa	rupestris	oleracea_177	medullosa	-0.0002498	0.000183167	-1.36386
villosa	rupestris	oleracea_177	palmifolia	-0.000234	0.000206753	-1.13189
villosa	rupestris	oleracea_177	ramosa	-0.0003248	0.00017538	-1.85194
villosa	rupestris	oleracea_177	sabauda	-0.0003065	0.000195878	-1.56498
villosa	rupestris	oleracea_177	sabellica	-0.0005109	0.00019722	-2.59027
villosa	rupestris	oleracea_177	viridis	-0.0001935	0.000175587	-1.10185
villosa	rupestris	oleracea_175	alboglabra	0.00017855	0.000247866	0.720364
villosa	rupestris	oleracea_175	botrytis	0.00021219	0.000305876	0.693726

villosa	rupestris	oleracea 175	capitata	-0.0001719	0.000193181	-0.889902
villosa	rupestris	oleracea 175	costata	0.00020677	0.000227848	0.907493
villosa	rupestris	oleracea 175	gemmafera	0.00014783	0.000276306	0.535007
villosa	rupestris	oleracea 175	gongyloides	1.00E-05	0.000232961	0.0429952
villosa	rupestris	oleracea 175	italica	-0.0001559	0.000291434	-0.534906
villosa	rupestris	oleracea 175	longata	0.00017549	0.00029665	0.591557
villosa	rupestris	oleracea 175	medullosa	0.00013176	0.000217881	0.604749
villosa	rupestris	oleracea 175	palmifolia	0.00011597	0.00023531	0.492848
villosa	rupestris	oleracea 175	ramosa	0.00020674	0.000230983	0.895052
villosa	rupestris	oleracea 175	sabauda	0.0001885	0.000214308	0.87955
villosa	rupestris	oleracea 175	sabellica	0.0003928	0.000241164	1.62878
villosa	rupestris	oleracea 175	viridis	-7.54E-05	0.000218207	-0.34563
villosa	rupestris	montana 222	alboglabra	0.00020657	0.000277042	0.745609
villosa	rupestris	montana 222	botrytis	0.00024021	0.000328126	0.732055
villosa	rupestris	montana 222	capitata	0.00019992	0.000246643	0.810579
villosa	rupestris	montana 222	costata	0.00023478	0.000280598	0.836718
villosa	rupestris	montana 222	gemmafera	0.00017584	0.000319945	0.549585
villosa	rupestris	montana 222	gongyloides	3.80E-05	0.00027931	0.136149
villosa	rupestris	montana 222	italica	-0.0001279	0.000346116	-0.369466
villosa	rupestris	montana 222	longata	0.0002035	0.000349653	0.581997
villosa	rupestris	montana 222	medullosa	0.00015978	0.000273292	0.584629
villosa	rupestris	montana 222	palmifolia	0.00014398	0.000282088	0.510421
villosa	rupestris	montana 222	ramosa	0.00023475	0.000280298	0.837514
villosa	rupestris	montana 222	sabauda	0.00021651	0.000267481	0.809427
villosa	rupestris	montana 222	sabellica	0.00042081	0.000296757	1.41804
villosa	rupestris	montana 222	viridis	-0.0001034	0.000269977	-0.383108
villosa	rupestris	cretica 196	alboglabra	1.94E-06	0.000255525	0.00759724
villosa	rupestris	cretica 196	botrytis	3.17E-05	0.000317371	0.0998804
villosa	rupestris	cretica 196	capitata	8.58E-06	0.000227046	0.0378019
villosa	rupestris	cretica 196	costata	2.63E-05	0.000240932	0.109057
villosa	rupestris	cretica 196	gemmafera	-3.27E-05	0.000235394	-0.138786
villosa	rupestris	cretica 196	gongyloides	0.00017048	0.000252445	0.675311
villosa	rupestris	cretica 196	italica	-0.0003364	0.000277628	-1.21164
villosa	rupestris	cretica 196	longata	5.01E-06	0.000308879	0.0162197
villosa	rupestris	cretica 196	medullosa	-4.87E-05	0.000220419	-0.221088
villosa	rupestris	cretica 196	palmifolia	-6.45E-05	0.000235383	-0.274119
villosa	rupestris	cretica 196	ramosa	-2.62E-05	0.000228233	-0.115001
villosa	rupestris	cretica 196	sabauda	-8.00E-06	0.000208251	-0.0384145
villosa	rupestris	cretica 196	sabellica	-0.0002123	0.000232377	-0.913633
villosa	rupestris	cretica 196	viridis	0.00010508	0.000229088	0.458672
villosa	rupestris	oleracea 176	alboglabra	0.00028099	0.000249023	1.12837
villosa	rupestris	oleracea 176	botrytis	-0.0002474	0.000291297	-0.849135
villosa	rupestris	oleracea 176	capitata	0.00028763	0.000213178	1.34926
villosa	rupestris	oleracea 176	costata	-0.0002528	0.000220768	-1.14498
villosa	rupestris	oleracea 176	gemmafera	-0.0003117	0.000225008	-1.38537
villosa	rupestris	oleracea 176	gongyloides	0.00044953	0.000261567	1.7186
villosa	rupestris	oleracea 176	italica	-0.0006154	0.000293627	-2.09598
villosa	rupestris	oleracea 176	longata	0.00028406	0.000307005	0.925262
villosa	rupestris	oleracea 176	medullosa	0.00032778	0.000198491	1.65137
villosa	rupestris	oleracea 176	palmifolia	0.00034357	0.0002322	1.47964
villosa	rupestris	oleracea 176	ramosa	0.0002528	0.000212908	1.18738
villosa	rupestris	oleracea 176	sabauda	0.00027105	0.00022309	1.21498
villosa	rupestris	oleracea 176	sabellica	6.67E-05	0.000211395	0.315725
villosa	rupestris	oleracea 176	viridis	0.00038413	0.000205849	1.86606
villosa	rupestris	cretica 198	alboglabra	-0.0005059	0.000322732	-1.56761
villosa	rupestris	cretica 198	botrytis	-0.0004723	0.000358214	-1.31842
villosa	rupestris	cretica 198	capitata	0.00051256	0.000317022	1.61679
villosa	rupestris	cretica 198	costata	-0.0004777	0.000303031	-1.57641
villosa	rupestris	cretica 198	gemmafera	-0.0005366	0.000320762	-1.67304
villosa	rupestris	cretica 198	gongyloides	0.00067446	0.000360377	1.87153
villosa	rupestris	cretica 198	italica	-0.0008404	0.000369548	-2.27403
villosa	rupestris	cretica 198	longata	0.00050899	0.000438253	1.1614
villosa	rupestris	cretica 198	medullosa	-0.0005527	0.0003133	-1.76415
villosa	rupestris	cretica 198	palmifolia	-0.0005685	0.000358699	-1.5849
villosa	rupestris	cretica 198	ramosa	0.00047773	0.000315116	1.51605
villosa	rupestris	cretica 198	sabauda	0.00049598	0.000313279	1.58318

villosa	rupestris	cretica_198	sabellica	0.00029167	0.000344178	0.847439
villosa	rupestris	cretica_198	viridis	0.00060905	0.000325176	1.873
villosa	rupestris	hilarionis	alboglabra	0.0003352	0.000399915	0.838167
villosa	rupestris	hilarionis	botrytis	-0.0003016	0.000467056	-0.645651
villosa	rupestris	hilarionis	capitata	0.00034184	0.000394559	0.866378
villosa	rupestris	hilarionis	costata	-0.000307	0.000404653	-0.758623
villosa	rupestris	hilarionis	gemmifera	-0.0003659	0.000402743	-0.908578
villosa	rupestris	hilarionis	gongylodes	0.00050373	0.000436322	1.1545
villosa	rupestris	hilarionis	italica	-0.0006696	0.000471594	-1.41995
villosa	rupestris	hilarionis	longata	0.00033826	0.000497975	0.679279
villosa	rupestris	hilarionis	medullosa	0.00038199	0.000411776	0.927656
villosa	rupestris	hilarionis	palmifolia	0.00039778	0.000444093	0.895707
villosa	rupestris	hilarionis	ramosa	0.00030701	0.000427581	0.71801
villosa	rupestris	hilarionis	sabauda	0.00032525	0.000398645	0.8159
villosa	rupestris	hilarionis	sabellica	0.00012095	0.000446419	0.270927
villosa	rupestris	hilarionis	viridis	0.00043833	0.000406541	1.07819

Supplemental Table 3.3 Modules identifies in WGCNA, the number of gene in each module, number of syntenic hits and blast hits to *Arabidopsis* with GO biological process indicate for 5 gene with the largest fold enrichment.

Module color	Gene number	# of <i>Arabidopsis</i> hits (Syteny)	# of <i>Arabidopsis</i> hits (Blast)	Fold enrichment	P-value	GO biological process
Grey	406	138	233	44.15	1.16E-02	anthocyanin-containing compound biosynthetic process (GO:0009718)
				41.39	8.91E-04	anthocyanin-containing compound metabolic process (GO:0046283)
				6.55	4.31E-02	response to jasmonic acid (GO:0009753)
				4.69	7.23E-04	response to oxidative stress (GO:0006979)
				4.61	2.30E-02	response to water deprivation (GO:0009414)
Brown	111	82	82	83.59	9.68E-04	response to fructose (GO:0009750)
				83.59	9.68E-04	syncytium formation (GO:0006949)
				39.33	5.29E-05	response to far red light (GO:0010218)
				32.78	1.99E-03	response to far red light (GO:0010218)
				23.55	9.20E-03	response to red light (GO:0010114)
Blue	179	105	111	> 100	9.94E-03	icosanoid metabolic process (GO:0006690)
				77.18	5.31E-05	indole glucosinolate metabolic process (GO:0042343)
				69.63	2.09E-03	response to herbivore (GO:0080027)
				65.86	2.61E-03	response to herbivore (GO:0080027)
				57	8.81E-06	jasmonic acid biosynthetic process (GO:0009695)
Yellow	77	52	53	33.92	2.16E-02	systemic acquired resistance (GO:0009627)
				22.49	9.89E-04	response to chitin (GO:0010200)
				19.1	2.15E-02	response to chitin (GO:0010200)
				18.61	2.44E-02	aging (GO:0007568)
				17.93	2.89E-02	response to salicylic acid (GO:0009751)
Turquoise	227	90	127	62.49	7.64E-09	S-glycoside biosynthetic process (GO:0016144)
				62.49	7.64E-09	glucosinolate biosynthetic process (GO:0019761)
				62.49	7.64E-09	glycosinolate biosynthetic process (GO:0019758)
				50.77	5.83E-03	leucine metabolic process (GO:0006551)
				44.28	1.25E-07	glucosinolate biosynthetic process (GO:0019761)

Supplemental Table 3.4 Archaeological *Brassica* reports from Europe and the Near East.

Date	Countries	Notes	References
c. 1600–1400 BC Middle Bronze Age	Austria, Friaga	3 charred seeds, hilltop settlement in the southern Alps. Unusually early date, but no reason to suspect contamination from later occupation etc.	Schmidl and Oegg 2005
c. 1300-1020 BC	Gibala, NW Syria	C14 dated seeds from pottery vessels associated with a destruction layer. Seeds have been SEM; Seeds could be <i>B. cretica</i> .	Kaniewski et al. 2011
La Tène (450-100 BC)	Switzerland/France	Preservation not specified but is absent in Roman samples. Refers to identification problems.	Jacomet and, Vandrope 2011
Late Iron Age (400-100 BC)	Germany	Review of various finds including <i>B. oleracea</i> from the German Late Iron Age	Willerding, et al. 2003
Roman-Iron Age (1st-3rd C AD)	Egypt, Mons Claudianus	charred/desiccated seeds resembling <i>B. oleracea</i>	van der Veen 2001 p.200
1st-4th Centuries AD	Italy, Vado Ligure, Liguria	possible seeds found in Roman well	Arobba et al. 2013
43-410 AD	England	recorded from 3 military and 2 town sites	van der Veen et al. 2008
Roman 100-300 AD	Germany, Otterbach, Kaiserslautern	11 waterlogged seeds of probable <i>Brassica</i> cf. <i>oleracea</i> from a well in Otterbach	Weithold 2012
400-1100 AD	Sweden, Denmark, Netherlands	seeds of <i>Brassica oleracea</i>	Hansson 2001, Heimdahl 2010
800-1250 AD	Czech Republic, Žatec	Finds of waterlogged <i>Brassica</i> cf. <i>oleracea</i>	Kočár et al. 2010
900-1100 AD	Great Britain, York	Finds of waterlogged <i>Brassica</i> cf. <i>oleracea</i>	Godwin and Bachem 1961
900-1100 AD	Great Britain, York	Finds of waterlogged <i>Brassica</i> cf. <i>oleracea</i>	Hall et al. 1983
1100 AD	Denmark	One to four chance finds from cultural layers	Jensen 1985
Early Medieval (12th century)	England, Raunds, Northants	Residues of <i>Brassica</i> leaves identified within pottery sherds.	Evershed et al. 1992, Evershed et al. 1994
1223-1300 AD	Montgomery, Powys, Wales	Single waterlogged seed comparable to <i>Brassica oleracea</i> from pit fill associated with castle	Greig et al. 1982
13-14th Century/Late 14th Century	Germany, Einbeck / Greifswald	Archaeobotanical material including mid. 14th century sewers. Ansorge et al. 2003 p.150	Wiethold 2002, Ansorge et al. 2003
Late Middle Ages (1300 to 1500)	Denmark, København	Archaeobotanical finds	Moltsen 2002
Medieval (15th Century)	Hungary, Budapest	Waterlogged seeds in well	Zoltán and Sárghadinye 2006 (Table 6.1)
14th/15th-16th Century AD	Germany, Stralsund, Kiel; Lüneburg; Greifswald	Mainly waterlogged seeds of <i>Brassica</i> cf. <i>oleracea</i>	Wiethold 2003a, Wiethold 2003b
16th Century	Netherlands, Haarlem	Uncertain identification of mineralised seed of <i>Brassica</i> cf. <i>oleracea</i>	Brinkkemper 2002
16th Century AD	Germany, Rostock	archaeobotanical finds	Wiethold 2000
17th Century	Belgium, Antwerpen	Seeds found in contents of 17th Century waste pit at Antwerpen	Cooremans 2000
17th Century	Hungary, Buda	Probable charred seed of cabbage (<i>Brassica</i> cf. <i>oleracea</i>) found within waste pit	Tóth et al. 2010

Supplemental Table 3.5 Literary and artistic sources covering the Classical Greek, Roman, and medieval and post-medieval sources.

Date	Countries	Notes	References
550-30 BC	Ionian Greece	Earliest textual reference. Refers to a "seven - leaved" cabbage in iambic verse. Perhaps as an oath or offering.	Hipponax, Fr.104, (West 2011) p. 145
460-370 BC	Classical Greece	Refers to cabbage in a few medicinal recipes	Hippocrates, <i>Nat Mul.</i> 32, <i>Mul</i> 1.78 (Totelin 2009) p. 75
370-300 cal. BC	Classical Greece	3 varieties smooth, parsley-leaved and salty. Last has delicate taste and grows in Eretria, Cyme and Rhodes (and in Cnidos and Ephesus).	Eudemus f. 85, Quoted in Athenaeus <i>The Deipnosophists</i> 9.9 (Yonge 1854) pp. 582-3
370-288/85 cal. BC	Classical Greece	3 varieties: smooth ?seedless (capita), larger leafed, sweeter parsley/curly-leaved, and a wild type bitter taste with many branches and smaller leaves may be <i>B. cretica</i> (see Willis 2007 p. 20)	Theophrastus <i>Historia Plantarum</i> VII. iv .4-6 (Hort 1916) pp. 84-85 Athenaeus <i>The Deipnosophists</i> 9.9 (Yonge 1854) pp. 582-3
early 3rd Century BC	Cyclades, Greece	States the finest cabbages grow in Cyme, but bitter in Alexandria. Seed brought from Rhodes produces a sweet cabbage than after a year taste degenerates.	Diphilus of Siphnus (the Siphnian). Athenaeus <i>The Deipnosophists</i> 9.9 (Yonge 1854) pp. 582-3
c.200-100 BC	Greece +	Smooth leaved, sometimes found wild. Also a curly leaved and one of reddish color.	Nicander <i>Georgics</i> . f.85. Athenaeus <i>The Deipnosophists</i> 9.9 (Yonge 1854) pp. 582-3
c. 160 BC	Classical Roman Italy	Lists smooth-leaved, curly leaved (<i>apiaca</i>), a small stalked, pungent tender variety and wild type	Cato, Marcus (Hooper and Ash 1934) <i>De Agricultura</i> . Capitula CLVI CLVII:156-157, 1 pp. 144-151
AD 77-79	Classical Roman Italy	List a large number of varieties.	(Rackham and Pliny the Elder 1950). 41.19.
250-450 AD	Egypt	List of Monastic sources from Egyptian papyri	Oxyrhynchus Papyri XIV 1656 (BL), (Harlow and Smith 2001)
Late 4th/early 5th C. AD	Roman	Cited in nine recipes attributed to Apicius: dealing with <i>Cimæ</i> (<i>cymae</i>) & <i>Coliculi</i> (refers to sprouts, young cabbage and soft cabbage)	(Vehling 1936) <i>Apicus</i> III, ix.87-92, x.94, xii.99 & xv.103
770-800 AD	France/ Germany	Capitularies of Charlemagne <i>Capitula</i> & <i>Ravacaulos</i> (Kohlrabi)	<i>Capitulare de Villis</i> (Boretius 1883) p. 90
9th - 13th century/Medieval	Poland, Krakow	archaeobotanical and literary sources. See Wieserowa 1979 for identification criteria (also Brinkkemper 1991)	Wieserowa 1979; Zemanek and Wasylikowa 1996
c.960-1350 AD (med. Spain)	Spain	Translations of <i>qannabit</i> & <i>kurunb</i> would indicate cauliflower & cabbage from c.960 AD	Harvey 1992.
13th-14th C.	France/ Switzerland	Le Viandier de Taillevent: Collection of medieval recipes. Mentioned in one recipe.	Prescott 1989

1300-1500 AD.	Czech Republic	Reference to cabbage, garlic and onion in cooking and medicine	Beranová 2002
Late Middle Ages	Estonia, Tartu	Written and artistic depictions of <i>Brassica oleracea</i>	Sillasoo 2002; 2003
16th - 17th C. AD	Flanders and Netherlands	Depictions of cauliflower and cabbages in art.	Zeven and Brandenburg 1986
16th C. AD	Russia		Sil'vestr's Domostroi
1554 AD	Flemish	lists 5 types of cabbage - white, red, savoy, Roosken (pale red) and curly kale	Dodoens 1971
1580 AD	England	Refers to time of sowing and Ionians holding them as sacred	Tusser 1878
1647	Denmark	<i>Brassica oleracea capita</i> referred to as garden plant in <i>Horticultura Danica</i>	Block 1984

**CHAPTER 4 : IONOMIC VARIATION IN DIVERSE *BRASSICA*
OLERACEA CULTIVAR AND WILD RELATIVE POPULATIONS
FROM GERMPLASM COLLECTIONS**

Makenzie E. Mabry¹, Elizabeth Countee¹, Greg Ziegler², Sarah D. T. Hisson³,
Evan Y. Gallagher¹, Joanne Labate⁴, Michael A. Gore⁵, Ivan Baxter⁵, J. Chris Pires¹

1 Division of Biological Sciences and Christopher S. Bond Life Sciences Center,
University of Missouri, Columbia, MO, U.S.A.

2 Donald Danforth Plant Sciences Center, St. Louis, MO, U.S.A

3 Department of Plant Biology, University of California Davis, Davis, CA, U.S.A.

4 Plant Genetics Research Unit, USDA-ARS, Columbia, MO, U.S.A.

5 USDA-ARS Plant Genetics Research Unit, Donald Danforth Plant Sciences Center, St.
Louis, MO, U.S.A

6 Bioinformatics Research Center, Program in Genetics and Department of Biological
Sciences, North Carolina State University, Raleigh, NC, U.S.A.

ABSTRACT

Brassica oleracea vegetable crops, frequently regarded as superfoods (e.g., kale), are an important diet source for key minerals and elements. Ionomics, the measurement of the total elemental composition of an organism, allows for high throughput characterization of the levels of many of these elements at the same time, including those important in human diets, such as iron and zinc. Using association mapping techniques, we have identified genes underlying quantitative phenotypic traits of importance to both human and plant health such as iron, as well as total elemental composition. This knowledge can then allow for further decisions of the best accessions to target for human health, as well as provide a resource for future studies seeking to enhance specific elements in different crops.

INTRODUCTION

Many *Brassica* crops are regarded as a type of ‘superfood’ due their perceived nutritional properties. For example, by eating one cup of raw kale, one can satisfy 26% of their daily value requirement of manganese, 10% of copper, 9% of calcium, 9% of potassium, 6% of iron, 6% of magnesium, and 4% of phosphorus (<https://nutritiondata.self.com/>). These elements are not only important for human health, but also for the health of the plants themselves. In addition to oxygen, carbon dioxide, and water, plants require additional mineral elements for adequate nutrition (Marschner, 1995; Mengel et al. 2001). The minerals (or elements) that plants extract from the soil are important for almost every plant function. For example, sulfur is required to produce the universal amino acids methionine and cysteine (Taliencio et al. 2019). Additionally, other elements play vital roles as enzyme cofactors in metabolic processes (Lavres et al. 2016; Weisany et al. 2013), including photosynthesis (Semin et al. 2018; Zhao et al. 2018).

One way to study the abundance of elements in an organism is to use ionomics. Ionomics allows for the simultaneous measurement of the elemental composition of an organism for any given tissue (Salt et al. 2008). Previous studies have leveraged ionomics combined with genome-wide association studies (GWAS) to identify underlying genetic factors responsible for differences in elemental accumulation in seeds (Ziegler et al. 2018; Taliencio et al. 2019) and shoots (Dhanapal et al. 2018) of soybean using diversity panels. However, loci identified in GWAS tend to be difficult to map to known genes, especially in non-model crops and organisms. To address this gap, Whitt et al. (2018) curated a known ionomic gene list, focusing on genes which have been shown experimentally to change uptake, accumulation, and distribution of elements. This curated list can then be used as a

guide in identifying variants that may play a role in the non-model organism elemental accumulation.

Of the all the *Brassica* crops, the species *Brassica oleracea* is comprised of the most vegetables which are typically thought of as having high nutritional value (i.e., broccoli, Brussels sprouts, cauliflower, cabbage, kale, and kohlrabi). This, combined with a developed diversity panel consisting of both cultivar and wild germplasm, and its close relation to *Arabidopsis thaliana*, a species in which many ionomic genes have been identified, makes *B. oleracea* an excellent system for surveying its ionomic abundance and composition. Previous work on the elemental composition and abundance of *B. oleracea* have typically focused on only a single or just a few of elements at a time (Broadley et al. 2008; Broadley et al. 2009; Hammond et al. 2009; Broadley et al. 2010; White et al. 2010). The focus of these studies being how elemental accumulation was correlated with different environments. Here we utilize association mapping techniques, which have yet to be used in *B. oleracea*, to identify both known ionomic genes in *B. oleracea* as well as novel genes that require further study to determine their role in ionomic abundance and composition.

MATERIALS AND METHODS

Plant samples

Samples include seven types of kales (chinese white kale; 22, tronchuda kale; 13, lacinato kale; 8, curly kale; 12, perpetual kale; 13, walking stick kale; 2, and collard greens; 16), three types of cabbages (white cabbage; 15, savoy cabbage; 4, and marrow cabbage; 13), Brussels sprouts (17), broccoli (20), cauliflower (18), and kohlrabi (15). Together these accounted for 189 of the 224 total samples. The rest of the 224 samples included wild

germplasm of *B. oleracea* (3), *B. cretica* (3), *B. incana* (5), *B. montana* (2), *B. hilarionis* (1), *B. insularis* (5), *B. macrocarpa* (4), *B. rupestris* (7), and *B. villosa* (5). Samples were grown from seed from the United States Department of Agriculture, Agriculture Research Service (USDA-ARS) germplasm center (Geneva, NY, 114 accessions), The Leibniz Institute of Plant Genetics and Crop Plant Research (IPK; 71 accessions), Universidad Politécnica de Madrid (UPM; 4 accessions), The Nordic Genetic Resource Centre (NordGen; 2 accessions), Gomez Campo Collection (2 accessions), John Innes Center (1 accession), Doubled Haploid lines (17 samples, some accessions sampled twice), or from our personal collection (13 accessions). Four replicates of each accession were grown in Promax soil in a sterile growth chamber at the University of Missouri (MU; Columbia, MO) Bond Life Sciences Center in a randomized complete block design with a 16 hour light/8 hour dark cycle. Average temperature and humidity ranged from 26.8C and 56.5% during the light cycle to 20.3C and 87.5% during the dark cycle. Light was kept at 137 uMol (measured with a LI-COR model LI-250 light meter). At the seventh leaf stage, leaf three was collected from all four replicates, immediately flash frozen using liquid nitrogen, and pooled for RNA extraction. Leaf four was collected per replicate and lyophilized for ionomic phenotyping.

Ionomic phenotyping by ICP-MS

Lyophilized samples were phenotyped at the Donald Danforth Plant Science Center (St. Louis, MO, USA) for the elemental concentrations of boron (B11), sodium (Na23), magnesium (Mg26), aluminum (Al27), phosphorus (P31), potassium (K39), calcium (Ca44), iron (Fe54), manganese (Mn55), cobalt (Co59), nickel (Ni60), copper (Cu63), Zinc (Zn66), rubidium (Rb85), strontium (Sr88), molybdenum (Mo98), and cadmium (Cd111)

following normalization to weights and analytical methods described in Ziegler et al. (2018). To visualize differences of elemental abundance across samples, boxplots for each element, separated by cultivar, were made in R v 3.5.1 (R Core Team 2018) using the package ggplot2 (Wickham 2016).

Heritability estimation and data transformation

After outlier removal (Ziegler et al. 2018), we performed a Shapiro test for normality using the R statistical package v 3.5.1. If sample did not meet the assumptions of normality, Box Cox transformation were performed using the *powerTransform* function with “bcPower” from the car package v 2.1-4 (Fox & Weisberg 2019). Variance components were estimated using mixed linear models (lme4 package; Bates et al. 2014; model below) with sample, replicate, and plantout as random effects for each element.

```
Lmer (element ~ (1|sample) + (1|rep) + (1|plantout))
```

Variance components were then extracted and used to calculate heritability on a line mean basis using the equation below.

$$\text{var}(\text{sample}) / (\text{var}(\text{sample}) + (\text{var}(\text{rep}) / 4) + (\text{var}(\text{plantout}) / 2) + (\text{var}(\text{residual}) / 8))$$

RNA sequencing and Genotyping

RNA was isolated using the ThermoFisher Invitrogen PureLink RNA mini kit (Invitrogen, Carlsbad, CA, USA) followed by TruSeq library preparation (Illumina, San Diego, CA, USA) and sequencing on the NextSeq (Illumina, San Diego, CA, USA) for 2

X 75 bp reads. Library preparation and sequencing was performed through the MU DNA Core Facility. For eight flow cells, 24 samples were multiplexed and sequenced in a single flow-cell, followed by a ninth flow cell with 17 samples, and a tenth flow-cell with just 16 samples.

Short reads were mapped to the *B. oleracea* TO1000 genome (Chinese kale; Parkin et al. 2014; release-41) using the STAR v. 2.5.2 (Dobin et al. 2013) two-pass alignment to first identify splice junctions, which were used in the second pass to improve mapping (Engström et al. 2013). Mapped reads (BAM format) were then processed following the GATK v. 3.8 best practices for RNA-seq reads (Mckenna et al. 2010; Van der Auwera et al. 2013; Poplin et al. 2017). To ensure that reads were mapping correctly, the GATK ‘Split’N’Trim’ function was used to split reads into exon segments and trim any overhanging reads in intron segments. In total, 7,564,168 variants were called before any filtering was performed. The resulting variants were filtered to exclude those with a Fisher strand (FS) value greater than 30 and quality depth (QD) less than 2.0. This recovered 942,357 variants in total, with 879,865 variants on chromosomes 1-9 and 62,492 variants on scaffolds. Only variants aligning to chromosomes were used in downstream analyses. The remaining variants were then filtered on the basis of the proportion of missing data (*-max-missing 0.4*), sites with mean depth values greater than 5 (*--min-meanDP 5*), and if they were indels (*--remove-indels*; vcfTools v. 0.1.17; Danecek et al. 2011) resulting in a total of 103,525 SNPs. Finally, the resulting vcf file was sorted and converted to HapMap format using Tassel v 5.2.50 (Bradbury et al. 2007).

BLUP calculations, PCA, and ANOVAs

To capture the phenotypic variation among replicates for each sample, BLUPs (best linear unbiased predictions) were calculated for each element from ionomic phenotyping by log transformation and fitting to a linear model with random effects for variance components (ranef function in lme4 package; Bates et al. 2014). PCAs of BLUPs were performed in R v 3.5.1 (R Core Team 2018) using prcomp (Stats package v 3.5.1). PCAs were run on all data (wild and cultivated samples), only cultivated samples, and then further separated as cultivars with edible leaves (chinese white kale, tronchuda kale, lacinato kale, curly kale, perpetual kale, walking stick kale, collard greens, white cabbage, savoy cabbage, marrow cabbage, and Brussels sprouts) and cultivars with edible stems (kohlrabi). To test for significant differences among cultivars, ANOVAs (Stats package v 3.5.1) were performed on BLUP transformed data in R v 3.5.1 (R Core Team 2018) for both elemental traits and PCs. Post-hoc analyses, specifically Tukey's test, was run using the *glht* function (multcomp v 1.4-13; Hothorn et al. 2016) to determine which cultivars exhibited significant differences.

Genome wide association studies

GWAS was performed using variants with a minimum minor allele frequency of 0.05, a calculated kinship matrix, and the first three principle components. Four models, the general linear model (GLM), the mixed linear model (MLM: Zhang et al. 2010), the multiple loci mixed model (MLMM; Segura et al. 2012), and the fixed and random model circulating probability unification model (FarmCPU; Liu et al. 2016), were run as implemented in the GAPIT R package (Lipka et al., 2012). GWAS results of the BLUPs for each element and PC from each model were sorted to extract significant SNPs with FDR-Adjusted p-values less than 0.05 and 0.1. Using an existing list of primary ionome

genes in *A. thaliana* (Whitt et al. 2018), orthology was determined for *B. oleracea* genes using EnsemblPlants (*accessed May 3, 2020*) between the Tair10 *A. thaliana* genome and BOL *B. oleracea* genome (**Supplemental Table 4.1**). In cases where there were multiple *B. oleracea* genes matched to a single *A. thaliana* gene, both were kept. To search for overlapping genes against the master list, SNPs from the *B. oleracea* gene list and GWAS results were formatted into bed file format. The *B. oleracea* gene start and end positions were calculated with a 40kb LD window. Files were compared with bedtools intersect (Quinlan & Hall 2010) and overlap of identified SNPs was determined using Venny (Oliveros 2007).

GERMPLASM AND DATA AVAILABILITY

Lines used can be found at the USDA Germplasm Resources Information Network (GRIN), Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Universidad Politécnica de Madrid (UPM), The Nordic Genetic Resource Centre (NordGen), Gomez Campo Collection, John Innes Center or upon reasonable request to the JCP. Raw RNA reads are available under BioProject PRJNA544934. All scripts and data used can be found at https://github.com/mmabry/Brassica_oleracea_Ionomics_GWAS_TWAS.

RESULTS

Variation and heritability of elemental composition across cultivars

Across the 17 elements tested, we found considerable amounts of diversity in abundance among cultivars (**Figure 4.1**). In many cases ranges for each element differed by several orders of magnitude. For example, Fe and Zn ranged in abundance from

0.0000178-138.4343 and 0.0000264-457.4144, respectively. For Fe, the smallest recorded abundance is from a kohlrabi sample while the largest is from a tronchuda kale. For Zn, the smallest is again from a kohlrabi sample while the largest is from the wild C group. Overall, however, the wild samples are less abundant in Na, Ca, Fe, Ni, Sr, and Cd, while most abundant in Mo compared to cultivar types. Heritability of traits ranged from 0.36 to 0.86, with a median value of 0.72 (**Figure 4.2**).

Figure 4.1 Raw element abundance diversity separated my cultivars and wild germplasm.

Figure 4.2 Calculated heritability of elements.

PCA and ANOVAs of elemental traits

PCA showed no distinction between wild vs cultivated germplasm with PC1 and PC2 together explaining 46.5% of the variation in the data (**Figure 4.3A**). PCA of just cultivars also recovered mostly overlapping clusters (**Figure 4.3B**). Analyses separated by the part of the plant eaten (leafy versus stem), again recovered no separation of types (**Figure 4.3C**). For all PCAs, PC1 and PC2, we recover two large groups of elements as positively correlated with one another. The first group includes Zn, Co, P, Fe, Mn, Ni, Cd, Cu, and Al, while the second group is composed of Rb and Na, Sr, Ca, and Mg. The first group was negatively correlated with Mo and the second with Rb. ANOVAs for elemental abundance with cultivar as the predictor recovered all elements except Cu as significantly different between morphotypes (**Table 4.1**). Specifically, we find that for Co,

Brussels sprouts are mostly strongly differentiated from most other groups, as is walking stick kale for Mg, and tronchuda kale for Mn (*additional significant differences highlighted in red*; **Supplemental Figures 4.1 – 4.17**). We also find significant differences between wild and cultivated samples when using PC scores for PC1, PC3, and PC4, and between leafy cultivars versus those which we eat the stems for PC1 and PC4 (**Table 4.2**). Ad-hoc Tukey tests for PCs are reported in **Supplemental Figure 4.18 – 4.21**.

Figure 4.3 PCA of elemental composition for **A)** cultivars versus wild germplasm, **B)** among cultivars, and **C)** between edible leaves and stem (tuber) cultivars.

Table 4.1 ANOVA p-values with Morphotype as predictor variable.

DEPENDENT VARIABLE	P-VALUE
AL27	0.03772*
B11	7.02E-07*
CA44	0.03213*
CD111	3.63E-05*
CO59	3.64E-11*
CU63	0.09802
FE54	2.70E-09*
K39	2.52E-09*
MG26	9.11E-08*
MN55	1.80E-07*
MO98	4.46E-08*
NA23	2.64E-10*
NI60	6.85E-08*
P31	7.86E-05*
RB85	9.37E-09*
SR88	0.006748*
ZN66	4.19E-06*

Table 4.2 ANOVA p-values with wild or leafy as predictor variables.

DEPENDENT VARIABLE	WILD P-VALUE	LEAFY P-VALUE
PC1	0.001251*	1.59E-05*
PC2	0.8326	0.07098
PC3	1.99E-06*	0.06089
PC4	1.01E-05*	0.000559*

Genome wide association studies

GWAS for all four models (GLM, MLM, MLMM, and FarmCPU) of the 17 different elements, recovered a total of 214 significant SNPs using a FDR corrected p-value of 0.05. Only three of these SNPs overlapped with known ionomic genes (**Table 4.3**). In the GLM model of Al one SNP on chromosome 5 (5_28993074; **Figure 4.4**) overlapped with the *B. oleracea* ortholog (Bo5g087130) of AtNRT1.5/AtNPF7.3 (At1g32450). This gene is only one of two copies in *B. oleracea*, the other copy being Bo5g049080. Two

additional significant SNPs were identified using the FarmCPU model. Both of these SNPs mapped to Bo2g095650, an ortholog of AtNRAMP1(At1g80830) and were identified in the analysis of Cd (2_26991422) and of K (2_26998120). *B. oleracea* has three copies of this gene compared to *A. thaliana* with the other identified copies being Bo6g080720 and Bo6g124200.

Using a less restrictive FDR corrected P-value of 0.1, we find an additional 736 significant SNPs with an additional 11 overlaps with known ionomics genes (**Table 4.3**). All of these SNPs were identified in the GLM model of Al (**Figure 4.4**) and mapped to *B. oleracea* orthologs of ABH1, AtHAC1, AtMYB36, AtNRT1.5/AtNPF7.3, BTSL1, CCC, CNGC2/DND1, FPN1, and ZIFL2 (**Table 4.3**). Many of the identified *B. oleracea* orthologs are either kept in duplicate or triplicate compared to *A. thaliana*. Three genes, however, are retained as single copy - AtHAC (Bo7g014500), CNGC2/DND1 (Bo9g164130), and ZIFL2 (Bo8g048300). Within each element some significant SNPs are identified in more than one model (**Supplemental Figure 4.22**). For Al, SNP 8_25307136 is identified using the FarmCPU and MLMM model, and SNP 5_39443707 is identified in the GLM and MLM model. For Ca, SNP 1_16452246 is found in the FarmCPU and MLMM models. In the analysis of Cu, SNP 4_10188098 is found in GLM and MLM while SNP 7_26370280 is found in FarmCPU and MLMM. For Mg, four SNPs (2_5771825, 3_2483334, 4_49448054, and 8_30937885) are all found in the FarmCPU and MLMM models. Finally, for Sr, SNP 1_16452246 is found in both FarmCPU and MLMM. Significant SNPs for Cd, Fe, K, Mn, Na, and Ni were only found using the FarmCPU model.

For multivariate GWAS of PC1, PC2, PC3, and PC4 we identified 38 significant SNPs using a FDR corrected P-value of 0.05, and only seven additional SNPs using a FDR corrected P-value of 0.1. Unfortunately, we do not identify any SNPs that overlap with currently known ionomics genes (Whitt et al. 2018) for either p-value. We do however identify novel significant SNPs that may play a role in ionic composition for *B. oleracea* specifically (**Figure 4.5; Table 4.4**).

Table 4.3 Overlap of identified SNPs and known ionic genes.

FDR corrected p-Value	Model	Element	SNP	BOL gene	ATH gene	Other copies of ATH gene in BOL	Known Ionomics gene
0.05	FarmCPU	Cd	2_26991422	Bo2g095650	At1g80830	Bo6g080720, Bo6g124200	AtNRAMP1
0.05	FarmCPU	K	2_26998120	Bo2g095650	At1g80830	Bo6g080720, Bo6g124200	AtNRAMP1
0.05	GLM	Al	5_28993074	Bo5g087130	At1g32450	Bo5g049080	AtNRT1.5/ AtNPF7.3
0.1	GLM	Al	3_52966306	Bo3g149030	At1g30450	Bo7g043580	CCC
0.1	GLM	Al	5_28993063	Bo5g087130	At1g32450	Bo5g049080	AtNRT1.5/ AtNPF7.3
0.1	GLM	Al	6_37178451	Bo6g118960	At1g74770	Bo6g106530	BTSL1
0.1	GLM	Al	7_16051276	Bo7g043580	At1g30450	Bo3g149030	CCC
0.1	GLM	Al	7_35777629	Bo7g093140	At2g38460	Bo4g027290, Bo4g027330	FPN1
0.1	GLM	Al	7_5261009	Bo7g014500	At2g21045	none	AtHAC1
0.1	GLM	Al	8_15710632	Bo8g048300	At3g43790	none	ZIFL2
0.1	GLM	Al	8_15710704	Bo8g048300	At3g43790	none	ZIFL2
0.1	GLM	Al	9_40804717	Bo9g133340	At5g57620	Bo2g030240, Bo3g019230	AtMYB36
0.1	GLM	Al	9_48544533	Bo9g164130	At5g15410	none	CNGC2/DND1
0.1	GLM	Al	9_7014317	Bo9g022460	At2g13540	Bo3g087110	ABH1

Figure 4.4 GWAS results for each model (shown for A1) with SNPs mapping to known ionomic genes indicated by green triangle.

Figure 4.5 GWAS results of all models for PC1. Labelled SNPs are top hits for each chromosome.

Table 4.4 Significant SNPs identified in GWAS of PCs. BOL = *Brassica oleracea* gene which the SNP overlapped with. ATH = *Arabidopsis thaliana* ortholog and gene name. Bold SNPs are indicated in Fig. 4.5.

FDR corrected p-Value	Model	PC	SNP	BOL gene	ATH gene	ATH gene name
0.05	FarmCPU	1	1_16452246	-	-	-
0.05	FarmCPU	1	1_41859446	Bo1g147770	At3g06670	-
0.05	FarmCPU	1	2_26133950	Bo2g094610	-	-
0.05	FarmCPU	1	3_19955998	Bo3g048610	At4g04350	EMB2369
0.05	FarmCPU	1	3_55126764	Bo3g153860	-	-
0.05	FarmCPU	1	3_5754934	-	-	-
0.05	FarmCPU	1	4_39225921	Bo4g144050	At5g38530	TSBtype2
0.05	FarmCPU	1	5_35731304	-	-	-
0.05	FarmCPU	1	5_907657	Bo5g003710	At3g10990, At3g23960	-
0.05	FarmCPU	1	7_48189706	-	-	-
0.05	FarmCPU	1	8_31524392	-	-	-
0.05	FarmCPU	1	9_14523959	Bo9g044070	-	-
0.05	FarmCPU	1	9_54021704	Bo9g183110	At5g03210	-
0.05	FarmCPU	2	2_43025905	Bo2g136130	At5g38280	PR5K
0.05	FarmCPU	2	3_53148616	Bo3g149320	At4g38350	-
0.05	FarmCPU	2	3_63261241	Bo3g182660	At1g53040	-
0.05	FarmCPU	2	5_1728761	Bo5g005340	At1g05210	-
0.05	FarmCPU	2	6_32164346	Bo6g101270	-	-
0.05	FarmCPU	2	7_30440765	Bo7g080100	At2g01950	BRL2
0.05	FarmCPU	2	8_33938417	Bo8g098880	At2g26060	CIA1
0.05	FarmCPU	3	1_11809200	Bo1g039170	At4g24610	-
0.05	FarmCPU	3	2_36820702	Bo2g118200	-	-
0.05	FarmCPU	3	6_39441974	Bo6g124030	At1g79640	-
0.05	FarmCPU	3	7_40721750	-	-	-
0.05	FarmCPU	3	9_47909448	Bo9g163080	-	-
0.05	FarmCPU	4	2_36905896	-	-	-
0.05	FarmCPU	4	3_36931697	-	-	-
0.05	FarmCPU	4	4_51196825	Bo4g190370	At2g40730	-
0.05	FarmCPU	4	7_22420144	Bo7g060210	At1g22180	-
0.05	FarmCPU	4	7_31864084	Bo7g081850	At3g25620	ABCG21
0.05	FarmCPU	4	9_34607118	Bo9g108350	At5g52640	HSP90-1
0.05	FarmCPU	4	9_46321441	Bo9g154690	At5g18630	-
0.05	MLMM	1	1_16454763	Bo1g056560	-	-
0.05	MLMM	1	3_5754934	Bo3g017550	At5g60170	-
0.05	MLMM	1	5_35731304	-	-	-
0.05	MLMM	4	3_10484808	Bo3g027730	At2g34410	RWA3
0.05	MLMM	4	3_36931697	-	-	-
0.05	MLMM	4	9_46321441	Bo9g154690	-	-

DISCUSSION

Cultivar variation exists

Among the different cultivars we tested, several differences were identified, but for many elements, as only small differences in abundance. These differences may be beneficial to further explore in crop improvement for this species, both for plant and human health. We also find high heritability in elemental traits, agreeing with a previous study of elements in *B. oleracea* (Broadley et al. 2008).

*Known ionomic genes are identified as important contributors to elemental abundance in *B. oleracea**

One interesting pattern observed is that almost all of the significant SNPs which overlapped with known ionomic genes were found in the analysis of Al using the GLM model. Since Al is positively correlated with Zn, Co, P, Fe, Mn, Ni, Cd, and Cu, it is surprising that we do not identify any of these known genes in analyses of those additional elements, however it may be due to Al having more spread in abundance therefore giving it more power in the model. One important caveat in using this particular model is that it does not take into consideration population structure, which we know exists between *B. oleracea* cultivars and their wild relatives. However, here we argue that GWAS results of individual elements are still notable as we identify genes already known to be important in ionomic abundance, but here identified for the first time in *B. oleracea*, rather than identifying novel variants.

As *B. oleracea* is hexaploid in relation to *A. thaliana* (Brassicaceae Triplication; Lysak et al. 2005; Tang et al. 2012; Hao et al. 2020), we further examined if any of the identified *B. oleracea* orthologs were retained as duplicates or if they returned to single

copy. We would expect that after the tribe Brassiceae became hexaploidy relative to its sister groups, this shared common ancestor would have three copies for every single-copy gene in *A. thaliana*. However, present day species including *B. oleracea* have lost the majority of these duplicate genes (Hao et al. 2020). We further know that genes are not lost randomly during diploidization, so identifying these patterns of genes that remain duplicated versus those which return back to a single copy may provide insight on how they function in ionomic abundance, for example their sensitivity to dosage (Birchler et al. 2005; Freeling 2009; Birchler & Veitia 2012; Conant et al. 2014; Birchler et al. 2016). Of the 45 known *A. thaliana* ionomic genes (Whitt et al. 2018), we find that in *B. oleracea* 23 are duplicate, six are triplicate, 21 are single copy, and one gene is quadruplicate (**Supplemental Table 4.1**). For the genes that had significant SNPs, three were duplicates (AtNRAMP1, FPN1, AtMYB36), four were triplicates (AtNRT1.5/AtNPF7.3, CCC, BTSL1, ABH1), and three were identified as single copy (AtHAC1, ZIFL2, CNGC2/DND1). The identified functions of the duplicate genes varied from Mn and Fe uptake to Casparian strip formation, which is crucial for efficient uptake of nutrients (Morrissey et al. 2009; Cailliatte et al. 2010; Kamiya et al. 2015), while triplicate genes are overall important in regulation, distribution, and homeostasis (McDowell et al. 2013; Hindt et al. 2017; Li et al. 2017), and single copy genes are more generally categorized as regulating elemental composition (McDowell et al. 2013; Chao et al. 2014; Remy et al. 2015).

In addition to identifying known ionomic genes, we also identify potentially new genes that play a role in ionomic abundance in *B. oleracea*. For Al, two SNPs, 8_25307136 and 5_39443707, were identified in more than one model. SNP 8_25307136 mapped to

Bo8g077270 which has no ortholog in *A. thaliana*, but does have one with *B. rapa* (Bra018037) potentially indicating that this is a *Brassica* specific gene. However, the function of the gene is unknown. SNP 5_39443707 mapped to Bo5g127260 which does have an *A. thaliana* ortholog (At5g36780; PPI3A). PPI3A is protein pump-interactor and may help with the regulation of protein transport (UniProt Consortium). For Ca we identify one SNP, 1_16452246 in both the FarmCPU and MLMM models, however we do not find any overlap of annotated *B. oleracea* genes. Interestingly, this same SNP is found as significant in FarmCPU analyses of Cd and Mg, and in both the FarmCPU and MLMM analyses of Sr. Two SNPs, 4_10188098 (identified in GLM and MLM) and 7_26370280 (identified in FarmCPU and MLMM), were found in analyses of Cu. SNP 4_10188098 mapped to Bo4g045250 and is orthologous to At1g43670 or CYFBP in *A. thaliana*, which functions to regulate starch levels (Rojas-González et al. 2015). The other identified SNP, 7_26370280, did not map to any annotated gene in *B. oleracea*. In analyses of Mg, four SNPs (2_5771825, 3_2483334, 4_49448054, and 8_30937885) were all found in both FarmCPU and MLMM models. SNP 2_5771825 mapped to Bo2g022060, but had no identified orthologs in any other species. SNP 3_2483334 mapped to Bo3g007010 which did have an identified ortholog (At5g11410), but with unknown function. The last two SNPs (4_49448054 and 8_30937885) both mapped to genes (Bo4g186410 and Bo8g091410, respectively), both with orthologs in *A. thaliana* (At2g36810; SGR6 and At3g57040; ARR9, respectively). Gene SGR6 has been identified as being involved in gravitropism (Yamauchi et al. 1997) and ARR9 functions as response regulator involved in His-to-Asp phosphorelay signal transduction (To et al. 2004). Finally, in analyses of Sr,

SNP 1_16452246 was found using both the FarmCPU and MLM models, however it did not map to any annotated *B. oleracea* gene.

Role of newly identified genes in ionomic composition

For multivariate analyses of ionomic composition using PC scores, we do not identify any significant SNPs that map to known ionomic genes; however, we do identify 38 significant SNPs using a FDR corrected P-value of 0.05. Of these SNPs, 28 map to annotated *B. oleracea* genes with 20 of these genes having at least one *A. thaliana* ortholog. Of the *A. thaliana* orthologs eight are named genes (EMB2369, TSBtype2, PR5K, BRL2, CIA1, ABCG21, HSP90-1, and RWA3). In *A. thaliana* EMB2369 plays a role in the chloroplast for embryo development (Berg et al. 2005), however, its connection to ionomic composition is unclear. TSBtype2 plays a role in tryptophan activity (Yin et al. 2010). The kinase domain of PR5K has been shown to be related to a family of protein-serine/threonine kinases that are involved in the expression of self-incompatibility and disease resistance (Wang et al. 1996). The BRL2 gene seems to be involved in vascular differentiation (Caño-Delgado et al. 2004) which could be important in transporting elements from the soil to the rest of the plant. The CIA1 gene is an essential component of the cytosolic iron-sulfur (Fe-S) protein assembly (CIA) machinery (Luo et al. 2012). This gene has not been identified as important in *A. thaliana*, but has clear importance for ionomic composition, especially as Fe is positively correlated with many other elements. The ABCG21 is part of a large group of genes which are involved with ATP-binding (Sánchez-Fernández et al. 2001). Another identified gene, HSP90-1 is a molecular chaperone which plays a role in stabilizing unfolding protein intermediates and then in the release of them once stress is relieved (Takahashi et al. 2003; Cha et al. 2013). Finally, the

last identified gene, RWA3, has a probable function in the acetylation of xylan during secondary wall biosynthesis (Lee et al. 2011; Manabe et al. 2013). Together these genes provide opportunities for further research on ionomic composition in *B. oleracea*.

CONCLUSIONS

Using a diversity panel of 14 cultivar types and nine wild relatives, we successfully identify known ionomics genes from *A. thaliana* in *B. oleracea* as well as novel SNPs that may be important contributing factors in elemental accumulation for this vital crop species. Further work to verify the role of these genes, for example using a larger sample size (ideally in the thousands) or focusing on a single crop type, could be leveraged for breeding not only improved lines of *B. oleracea*, but also close relatives such as *B. napus* which shares half of its genome with *B. oleracea*. Additionally, while here we attempt to survey the overall ionomic profile of *B. oleracea*, we suggest that future work focuses on a single crop type in multiple environments, which would ultimately be more powerful in determination of casual ionomic genes. When plants are grown in nutrient-poor soils, many times farmers apply elemental rich fertilizers, costing them money. Breeding improved lines could hopefully both decrease cost for farmers, while at the same time provide essential nutrients for humans as food.

ACKNOWLEDGMENTS

We thank Marianne Slaten for her assistance in computing BLUPs and troubleshooting GAPIT, Paul Blischak for his assistance with statistical analyses, Libby King for her help with furthering our understanding of GWAS as well as her DataPhiles group which provided excellent feedback on analyses. We additionally thank our

computing resources, the Research Computing Support Services (RCSS) and Informatics Research Core Facility (IRCF), at the University of Missouri. Finally, we thank our funding USDA-ARS Project No. 8060-21000-024-00D.

AUTHOR CONTRIBUTIONS

MEM, JL, MAG, IB, and JCP designed the project. MEM, EYG, and SDTH grew plants and collected tissue. MEM and EYG measured and crushed leaves for elemental analysis. GZ conducted the elemental analysis using ICP-MS. MEM and LC ran the statistical tests of ANOVA, PCA, and GWAS. MEM wrote the original manuscript.

REFERENCES

- Arrivault, S., Senger, T., & Krämer, U. 2006. The *Arabidopsis* metal tolerance protein AtMTP3 maintains metal homeostasis by mediating Zn exclusion from the shoot under Fe deficiency and Zn oversupply. *The Plant Journal*, 46(5), 861-879.
- Assunção, A. G., Herrero, E., Lin, Y. F., Huettel, B., Talukdar, S., Smaczniak, C., et al. 2010. *Arabidopsis thaliana* transcription factors bZIP19 and bZIP23 regulate the adaptation to zinc deficiency. *Proceedings of the National Academy of Sciences*, 107(22), 10296-10301.
- Barberon, M., Dubeaux, G., Kolb, C., Isono, E., Zelazny, E., & Vert, G. 2014. Polarization of IRON-REGULATED TRANSPORTER 1 (IRT1) to the plant-soil interface plays crucial role in metal homeostasis. *Proceedings of the National Academy of Sciences*, 111(22), 8293-8298.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.
- Baxter, I., Brazelton, J. N., Yu, D., Huang, Y. S., Lahner, B., Yakubova, E., et al. 2010. A coastal cline in sodium accumulation in *Arabidopsis thaliana* is driven by natural variation of the sodium transporter AtHKT1; 1. *PLoS genetics*, 6(11).
- Baxter, I. 2009. Ionomics: studying the social network of mineral nutrients. *Current opinion in plant biology*, 12(3), 381-386.
- Baxter, I., Hosmani, P. S., Rus, A., Lahner, B., Borevitz, J. O., Muthukumar, B., et al. 2009. Root suberin forms an extracellular barrier that affects water relations and mineral nutrition in *Arabidopsis*. *PLoS genetics*, 5(5).

- Baxter, I., Muthukumar, B., Park, H. C., Buchner, P., Lahner, B., Danku, J., et al. 2008. Variation in molybdenum content across broadly distributed populations of *Arabidopsis thaliana* is controlled by a mitochondrial molybdenum transporter (MOT1). *PLoS Genetics*, 4(2).
- Berg, M., Rogers, R., Muralla, R., & Meinke, D. 2005. Requirement of aminoacyl-tRNA synthetases for gametogenesis and embryo development in *Arabidopsis*. *The Plant Journal*, 44(5), 866-878.
- Bernal, M., Casero, D., Singh, V., Wilson, G. T., Grande, A., Yang, H., et al. 2012. Transcriptome sequencing identifies SPL7-regulated copper acquisition genes FRO4/FRO5 and the copper dependence of iron homeostasis in *Arabidopsis*. *The Plant Cell*, 24(2), 738-761.
- Birchler, J. A., Johnson, A. F., & Veitia, R. A. 2016. Kinetics genetics: Incorporating the concept of genomic balance into an understanding of quantitative traits. *Plant Science*, 245, 128-134.
- Birchler, J. A., & Veitia, R. A. 2012. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences*, 109(37), 14746-14753.
- Birchler, J. A., Riddle, N. C., Auger, D. L., & Veitia, R. A. 2005. Dosage balance in gene regulation: biological implications. *Trends in Genetics*, 21(4), 219-226.
- Borghi, M., Rus, A., & Salt, D. E. 2011. Loss-of-function of Constitutive Expresser of Pathogenesis Related Genes5 affects potassium homeostasis in *Arabidopsis thaliana*. *PLoS One*, 6(10).

- Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19), 2633-2635.
- Broadley, M., Lochlainn, S., Hammond, J., Bowen, H., Cakmak, I., Eker, S., et al. 2010. Shoot zinc (Zn) concentration varies widely within *Brassica oleracea* L. and is affected by soil Zn and phosphorus (P) levels. *The Journal of Horticultural Science and Biotechnology*, 85(5), 375-380.
- Broadley, M. R., Hammond, J. P., King, G. J., Bowen, H. C., Hayden, R., Spracklen, W. P., et al. 2009. Biofortifying *Brassica* with calcium (Ca) and magnesium (Mg).
- Broadley, M. R., Hammond, J. P., King, G. J., Astley, D., Bowen, H. C., Meacham, M. C., et al. 2008. Shoot calcium and magnesium concentrations differ between subtaxa, are highly heritable, and associate with potentially pleiotropic loci in *Brassica oleracea*. *Plant Physiology*, 146(4), 1707-1720.
- Cailliatte, R., Schikora, A., Briat, J. F., Mari, S., & Curie, C. 2010. High-affinity manganese uptake by the metal transporter NRAMP1 is essential for *Arabidopsis* growth in low manganese conditions. *The Plant Cell*, 22(3), 904-917.
- Caño-Delgado, A., Yin, Y., Yu, C., Vafeados, D., Mora-García, S., Cheng, J. C., et al. 2004. BRL1 and BRL3 are novel brassinosteroid receptors that function in vascular differentiation in *Arabidopsis*. *Development*, 131(21), 5341-5351.
- Cha, J. Y., Ahn, G., Kim, J. Y., Kang, S. B., Kim, M. R., Su'udi, M., et al. 2013. Structural and functional differences of cytosolic 90-kDa heat-shock proteins (Hsp90s) in *Arabidopsis thaliana*. *Plant physiology and biochemistry*, 70, 368-373.

- Chao, D. Y., Baraniecka, P., Danku, J., Koprivova, A., Lahner, B., Luo, H., et al. 2014. Variation in sulfur and selenium accumulation is controlled by naturally occurring isoforms of the key sulfur assimilation enzyme ADENOSINE 5'-PHOSPHOSULFATE REDUCTASE2 across the *Arabidopsis* species range. *Plant physiology*, 166(3), 1593-1608.
- Chao, D. Y., Chen, Y., Chen, J., Shi, S., Chen, Z., Wang, C., et al. 2014. Genome-wide association mapping identifies a new arsenate reductase enzyme critical for limiting arsenic accumulation in plants. *PLoS biology*, 12(12).
- Chao, D. Y., Silva, A., Baxter, I., Huang, Y. S., Nordborg, M., Danku, J., et al. 2012. Genome-wide association studies identify heavy metal ATPase3 as the primary determinant of natural variation in leaf cadmium in *Arabidopsis thaliana*. *PLoS genetics*, 8(9).
- Chao, D. Y., Gable, K., Chen, M., Baxter, I., Dietrich, C. R., Cahoon, E. B., et al. 2011. Sphingolipids in the root play an important role in regulating the leaf ionome in *Arabidopsis thaliana*. *The Plant Cell*, 23(3), 1061-1081.
- Conant, G. 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), 3184-3193.
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156-2158.
- Desbrosses-Fonrouge, A. G., Voigt, K., Schröder, A., Arrivault, S., Thomine, S., & Krämer, U. 2005. *Arabidopsis thaliana* MTP1 is a Zn transporter in the vacuolar

- membrane which mediates Zn detoxification and drives leaf Zn accumulation. *FEBS letters*, 579(19), 4165-4174.
- Dhanapal, A. P., Ray, J. D., Smith, J. R., Purcell, L. C., & Fritschi, F. B. 2018. Identification of novel genomic loci associated with soybean shoot tissue macro- and micronutrient concentrations. *The plant genome*, 11(2).
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15-21.
- Eide, D., Broderius, M., Fett, J., & Guerinot, M. L. 1996. A novel iron-regulated metal transporter from plants identified by functional expression in yeast. *Proceedings of the National Academy of Sciences*, 93(11), 5624-5628.
- Engström, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., Alioto, T., et al. 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature methods*, 10(12), 1185-1191.
- Eroglu, S., Giehl, R. F., Meier, B., Takahashi, M., Terada, Y., Ignatyev, K., et al. 2017. Metal tolerance protein 8 mediates manganese homeostasis and iron reallocation during seed development and germination. *Plant physiology*, 174(3), 1633-1647.
- Eroglu, S., Meier, B., von Wirén, N., & Peiter, E. 2016. The vacuolar manganese transporter MTP8 determines tolerance to iron deficiency-induced chlorosis in *Arabidopsis*. *Plant physiology*, 170(2), 1030-1045.
- Fox, J., & Weisberg, S. 2019. An R Companion to Applied Regression (Third).
- Freeling, M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual review of plant biology*, 60, 433-453.

- Gao, Y. Q., Chen, J. G., Chen, Z. R., An, D., Lv, Q. Y., Han, M. L., et al. 2017. A new vesicle trafficking regulator CTL1 plays a crucial role in ion homeostasis. *PLoS biology*, 15(12), e2002978.
- Gobert, A., Park, G., Amtmann, A., Sanders, D., & Maathuis, F. J. 2006. Arabidopsis thaliana cyclic nucleotide gated channel 3 forms a non-selective ion transporter involved in germination and cation transport. *Journal of experimental botany*, 57(4), 791-800.
- Hammond, J. P., Broadley, M. R., White, P. J., King, G. J., Bowen, H. C., Hayden, R., et al. 2009. Shoot yield drives phosphorus use efficiency in *Brassica oleracea* and correlates with root architecture traits. *Journal of Experimental Botany*, 60(7), 1953-1968.
- Hao, Y., Mabry, M.E., Edger, P.E., Zheng, C., Jin, L., VanBuren, R., et al. 2020 The Contributions of Different Parents are Evolutionary but Not Functionally Distinct in the Meso-Allopolyploid Brassiceae. *Submitted*.
- Hindt, M. N., Akmajian, G. Z., Pivarski, K. L., Punshon, T., Baxter, I., Salt, D. E., & Guerinot, M. L. 2017. BRUTUS and its paralogs, BTS LIKE1 and BTS LIKE2, encode important negative regulators of the iron deficiency response in *Arabidopsis thaliana*. *Metallomics*, 9(7), 876-890.
- Hothorn, T., Bretz, F., Westfall, P., Heiberger, R. M., Schuetzenmeister, A., Scheibe, S., & Hothorn, M. T. 2016. Package ‘multcomp’. Simultaneous inference in general parametric models. *Project for Statistical Computing*, Vienna, Austria.

- Huang, X. Y., Deng, F., Yamaji, N., Pinson, S. R., Fujii-Kashino, M., Danku, J., et al. 2016. A heavy metal P-type ATPase OsHMA4 prevents copper accumulation in rice grain. *Nature communications*, 7(1), 1-13.
- Hussain, D., Haydon, M. J., Wang, Y., Wong, E., Sherson, S. M., Young, J., et al. 2004. P-type ATPase heavy metal transporters with roles in essential zinc homeostasis in *Arabidopsis*. *The plant cell*, 16(5), 1327-1339.
- Isayenkov, S. V., & Maathuis, F. J. 2008. The *Arabidopsis thaliana* aquaglyceroporin AtNIP7; 1 is a pathway for arsenite uptake. *Febs Letters*, 582(11), 1625-1628.
- Kamiya, T., Borghi, M., Wang, P., Danku, J. M., Kalmbach, L., Hosmani, P. S., et al. 2015. The MYB36 transcription factor orchestrates Casparian strip formation. *Proceedings of the National Academy of Sciences*, 112(33), 10533-10538.
- Kim, S. A., Punshon, T., Lanzirotti, A., Li, L., Alonso, J. M., Ecker, J. R., et al. 2006. Localization of iron in *Arabidopsis* seed requires the vacuolar membrane transporter VIT1. *Science*, 314(5803), 1295-1298.
- Koprivova, A., Giovannetti, M., Baraniecka, P., Lee, B. R., Grondin, C., Loudet, O., & Kopriva, S. 2013. Natural variation in the ATPS1 isoform of ATP sulfurylase contributes to the control of sulfate levels in *Arabidopsis*. *Plant physiology*, 163(3), 1133-1141.
- Kühnlenz, T., Hofmann, C., Uraguchi, S., Schmidt, H., Schempp, S., Weber, M., et al. 2016. Phytochelatin synthesis promotes leaf Zn accumulation of *Arabidopsis thaliana* plants grown in soil with adequate Zn supply and is essential for survival on Zn-contaminated soil. *Plant and Cell Physiology*, 57(11), 2342-2352.

- Lanquar, V., Ramos, M. S., Lelièvre, F., Barbier-Brygoo, H., Krieger-Liszkay, A., Krämer, U., & Thomine, S. 2010. Export of vacuolar manganese by AtNRAMP3 and AtNRAMP4 is required for optimal photosynthesis and growth under manganese deficiency. *Plant physiology*, 152(4), 1986-1999.
- Lavres, J., Castro Franco, G., & de Sousa Câmara, G. M. 2016. Soybean seed treatment with nickel improves biological nitrogen fixation and urease activity. *Front Environ Sci* 4: 37.
- Lee, C., Teng, Q., Zhong, R., & Ye, Z. H. 2011. The four *Arabidopsis* reduced wall acetylation genes are expressed in secondary wall-containing cells and required for the acetylation of xylan. *Plant and Cell Physiology*, 52(8), 1289-1301.
- Li, H., Yu, M., Du, X. Q., Wang, Z. F., Wu, W. H., Quintero, F. J., et al. 2017. NRT1.5/NPF7.3 functions as a proton-coupled H⁺/K⁺ antiporter for K⁺ loading into the xylem in *Arabidopsis*. *The Plant Cell*, 29(8), 2016-2026.
- Li, X., Zhang, H., Ai, Q., Liang, G., & Yu, D. 2016. Two bHLH transcription factors, bHLH34 and bHLH104, regulate iron homeostasis in *Arabidopsis thaliana*. *Plant Physiology*, 170(4), 2478-2493.
- Lindsay, E. R., & Maathuis, F. J. 2016. *Arabidopsis thaliana* NIP 7; 1 is involved in tissue arsenic distribution and tolerance in response to arsenate. *FEBS letters*, 590(6), 779-786.
- Liu, X., Huang, M., Fan, B., Buckler, E. S., & Zhang, Z. 2016. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS genetics*, 12(2).

- Long, T. A., Tsukagoshi, H., Busch, W., Lahner, B., Salt, D. E., & Benfey, P. N. 2010. The bHLH transcription factor POPEYE regulates response to iron deficiency in *Arabidopsis* roots. *The Plant Cell*, 22(7), 2219-2236.
- Loudet, O., Saliba-Colombani, V., Camilleri, C., Calenge, F., Gaudon, V., Koprivova, A., et al. 2007. Natural variation for sulfate content in *Arabidopsis thaliana* is highly controlled by APR2. *Nature genetics*, 39(7), 896-900.
- Luo, D., Bernard, D. G., Balk, J., Hai, H., & Cui, X. 2012. The DUF59 family gene AE7 acts in the cytosolic iron-sulfur cluster assembly pathway to maintain nuclear genome integrity in *Arabidopsis*. *The Plant Cell*, 24(10), 4135-4148.
- Lysak, M. A., Koch, M. A., Pecinka, A., & Schubert, I. 2005. Chromosome triplication found across the tribe Brassiceae. *Genome research*, 15(4), 516-525.
- Manabe, Y., Verhertbruggen, Y., Gille, S., Harholt, J., Chong, S. L., Pawar, P. M. A., et al. 2013. Reduced wall acetylation proteins play vital and distinct roles in cell wall O-acetylation in *Arabidopsis*. *Plant physiology*, 163(3), 1107-1117.
- Mengel, K., Kirkby, E. A., Kosegarten, H., & Appel, T. 2001. Principles of plant nutrition (pp. 397-434). Springer, Dordrecht.
- Marschner, H. 1995. Mineral nutrition of higher plants 2nd edition. Academic, Great Britain.
- McDowell, S. C., Akmakjian, G., Sladek, C., Mendoza-Cozatl, D., Morrissey, J. B., Saini, N., et al. 2013. Elemental concentrations in the seed of mutants and natural variants of *Arabidopsis thaliana* grown under varying soil conditions. *PloS one*, 8(5).

- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9), 1297-1303.
- Morrissey, J., & Guerinot, M. L. 2009. Iron uptake and transport in plants: the good, the bad, and the ionome. *Chemical reviews*, 109(10), 4553-4567.
- Oliveros, J.C. 2007-2015 Venny. An interactive tool for comparing lists with Venn's diagrams. <https://bioinfogp.cnb.csic.es/tools/venny/index.html>
- Olsen, Lene Irene, Thomas H. Hansen, Camille Larue, Jeppe Thulin Østerberg, Robert D. Hoffmann, Johannes Liesche, Ute Krämer et al. "Mother-plant-mediated pumping of zinc into the developing seed." *Nature plants* 2, no. 5 2016: 1-6.
- Palmer, C. M., Hindt, M. N., Schmidt, H., Clemens, S., & Guerinot, M. L. 2013. MYB10 and MYB72 are required for growth under iron-limiting conditions. *PLoS genetics*, 9(11).
- Parkin, I. A., Koh, C., Tang, H., Robinson, S. J., Kagale, S., Clarke, W. E., et al. 2014. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. *Genome biology*, 15(6), R77.
- Peiter, E., Montanini, B., Gobert, A., Pedas, P., Husted, S., Maathuis, F. J., et al. 2007. A secretory pathway-localized cation diffusion facilitator confers plant manganese tolerance. *Proceedings of the National Academy of Sciences*, 104(20), 8532-8537.
- Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., et al. 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 201178.

- Quinlan, A. R., & Hall, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841-842.
- R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Remy, E., Cabrito, T. R., Batista, R. A., Teixeira, M. C., Sá-Correia, I., & Duque, P. 2015. The major facilitator superfamily transporter ZIFL2 modulates cesium and potassium homeostasis in *Arabidopsis*. *Plant and Cell Physiology*, 56(1), 148-162.
- Robinson, N. J., Procter, C. M., Connolly, E. L., & Guerinot, M. L. 1999. A ferric-chelate reductase for iron uptake from soils. *Nature*, 397(6721), 694-697.
- Rojas-González, J. A., Soto-Suárez, M., García-Díaz, Á., Romero-Puertas, M. C., Sandalio, L. M., Mérida, Á., et al. 2015. Disruption of both chloroplastic and cytosolic FBPase genes results in a dwarf phenotype and important starch and metabolite changes in *Arabidopsis thaliana*. *Journal of experimental botany*, 66(9), 2673-2689.
- Salt, D. E., Baxter, I., & Lahner, B. 2008. Ionomics and the study of the plant ionome. *Annu. Rev. Plant Biol.*, 59, 709-733.
- Sánchez-Fernández, R., Davies, T. E., Coleman, J. O., & Rea, P. A. 2001. The *Arabidopsis thaliana* ABC protein superfamily, a complete inventory. *Journal of Biological Chemistry*, 276(32), 30231-30244.
- Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., & Nordborg, M. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature genetics*, 44(7), 825.

- Semin, B. K., Davletshina, L. N., Seibert, M., & Rubin, A. B. 2018. Creation of a 3Mn/1Fe cluster in the oxygen-evolving complex of photosystem II and investigation of its functional activity. *Journal of Photochemistry and Photobiology B: Biology*, 178, 192-200.
- Sunkar, R., Kaplan, B., Bouché, N., Arazi, T., Dolev, D., Talke, I. N., et al. 2000. Expression of a truncated tobacco NtCBP4 channel in transgenic plants and disruption of the homologous *Arabidopsis* CNGC1 gene confer Pb²⁺ tolerance. *The Plant Journal*, 24(4), 533-542.
- Takahashi, A., Casais, C., Ichimura, K., & Shirasu, K. 2003. HSP90 interacts with RAR1 and SGT1 and is essential for RPS2-mediated disease resistance in *Arabidopsis*. *Proceedings of the National Academy of Sciences*, 100(20), 11777-11782.
- Taliercio, E., Scaboo, A., Baxter, I., & Locke, A. M. 2019. The Ionome of a Genetically Diverse Set of Wild Soybean Accessions. *Crop Science*, 59(5), 1983-1991.
- Tang, H., Woodhouse, M. R., Cheng, F., Schnable, J. C., Pedersen, B. S., Conant, G., et al. 2012. Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy. *Genetics*, 190(4), 1563-1574.
- Tian, H., Baxter, I. R., Lahner, B., Reinders, A., Salt, D. E., & Ward, J. M. 2010. *Arabidopsis* NPCC6/NaKR1 is a phloem mobile metal binding protein necessary for phloem function and root meristem maintenance. *The Plant Cell*, 22(12), 3963-3979.
- To, J. P., Haberer, G., Ferreira, F. J., Deruere, J., Mason, M. G., Schaller, G. E., et al. 2004. Type-A *Arabidopsis* response regulators are partially redundant negative regulators of cytokinin signaling. *The Plant Cell*, 16(3), 658-671.

- UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1), D506-D515.
- Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, 43(1), 11-10.
- Vitart, V., Baxter, I., Doerner, P., & Harper, J. F. 2001. Evidence for a role in growth and salt resistance of a plasma membrane H⁺-ATPase in the root endodermis. *The Plant Journal*, 27(3), 191-201.
- Wang, X., Zafian, P., Choudhary, M., & Lawton, M. 1996. The PR5K receptor protein kinase from *Arabidopsis thaliana* is structurally related to a family of plant defense proteins. *Proceedings of the National Academy of Sciences*, 93(6), 2598-2602.
- Waters, B. M., Chu, H. H., DiDonato, R. J., Roberts, L. A., Eisley, R. B., Lahner, B., et al. 2006. Mutations in *Arabidopsis* yellow stripe-like1 and yellow stripe-like3 reveal their roles in metal ion homeostasis and loading of metal ions in seeds. *Plant Physiology*, 141(4), 1446-1458.
- Weisany, W., Raei, Y., & Allahverdipour, K. H. 2013. Role of some of mineral nutrients in biological nitrogen fixation. *Bulletin of Environment, Pharmacology and Life Sciences*, 2(4), 77-84.
- Wickham, H. 2016. ggplot2: elegant graphics for data analysis. Springer.
- Wild, M., Davière, J. M., Regnault, T., Sakvarelidze-Achard, L., Carrera, E., Diaz, I. L., et al. 2016. Tissue-specific regulation of gibberellin signaling fine-tunes *Arabidopsis* iron-deficiency responses. *Developmental cell*, 37(2), 190-200.

- White, P. J., Hammond, J. P., King, G. J., Bowen, H. C., Hayden, R. M., Meacham, M. C., et al. 2010. Genetic analysis of potassium use efficiency in *Brassica oleracea*. *Annals of Botany*, 105(7), 1199-1210.
- Whitt, L., Ricachenevsky, F., Ziegler, G., Clemens, S., Walker, E., Maathuis, F. J. M., et al. 2018. A curated list of genes that control elemental accumulation in plants. *BioRxiv*, 456384.
- Yamauchi, Y., Fukaki, H., Fujisawa, H., & Tasaka, M. 1997. Mutations in the SGR4, SGR5 and SGR6 loci of *Arabidopsis thaliana* alter the shoot gravitropism. *Plant and cell physiology*, 38(5), 530-535.
- Yan, J., Chia, J. C., Sheng, H., Jung, H. I., Zavodna, T. O., Zhang, L., et al. 2017. *Arabidopsis* pollen fertility requires the transcription factors CITF1 and SPL7 that regulate copper delivery to anthers and jasmonic acid synthesis. *The Plant Cell*, 29(12), 3012-3029.
- Yin, R., Frey, M., Gierl, A., & Glawischnig, E. 2010. Plants contain two distinct classes of functional tryptophan synthase beta proteins. *Phytochemistry*, 71(14-15), 1667-1672.
- Zhai, Z., Gayomba, S. R., Jung, H. I., Vimalakumari, N. K., Piñeros, M., Craft, E., et al. 2014. OPT3 is a phloem-specific iron transporter that is essential for systemic iron signaling and redistribution of iron and cadmium in *Arabidopsis*. *The Plant Cell*, 26(5), 2249-2264.
- Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. 2010. Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, 42(4), 355.

- Zhao, J., Zhang, W., Qiu, Q., Meng, F., Zhang, M., Rao, D., et al. 2018. Physiological regulation associated with differential tolerance to iron deficiency in soybean. *Crop Science*, 58(3), 1349-1359.
- Ziegler, G., Nelson, R., Granada, S., Krishnan, H. B., Gillman, J. D., & Baxter, I. 2018. Genomewide association study of ionomic traits on diverse soybean populations from germplasm collections. *Plant Direct*, 2(1), e00033.

Supplemental Table 4.1 Known ionomics genes identified in *Arabidopsis thaliana* (ATH) with *Brassica oleracea* (BOL) determined orthologs. Orthology was determined for *B. oleracea* genes using EnsemblPlants (accessed May 3, 2020) between the Tair10 *A. thaliana* genome and BOL *B. oleracea* genome. In many cases *B. oleracea* has multiple copies of the corresponding *A. thaliana*, all copies are indicated below. Plus signs (+) indicate ATH genes that are found duplicate in BOL. Single asterisks (*) indicate genes that are found triplicated in BOL. Double asterisks (**) indicate a ATH gene found in quadruplicate in BOL.

ATH Gene	BOL Gene	Chromosome	ATH Gene Name	Elements	Tissue	Citation
At1g01580+	Bo5g002390	chr5	FRO2	Fe	Root	Robinson et al. 1999
At1g01580+	Bo8g117920	chr8	FRO2	Fe	Root	Robinson et al. 1999
At1g18910	Bo8g103000	chr8	BTSL2	Fe,Mn,Zn	Leaf	Hindt et al. 2017
At1g20110*	Bo5g029170	chr5	FYVE1	Fe,Zn,Co,Mn	Root	Barberon et al. 2014
At1g20110*	Bo7g061820	chr7	FYVE1	Fe,Zn,Co,Mn	Root	Barberon et al. 2014
At1g20110*	Bo8g070000	chr8	FYVE1	Fe,Zn,Co,Mn	Root	Barberon et al. 2014
At1g30450+	Bo3g149030	chr3	CCC	Ca,K,Na,S	Seeds	McDowell et al. 2013
At1g30450+	Bo7g043580	chr7	CCC	Ca,K,Na,S	Seeds	McDowell et al. 2013
At1g32450+	Bo5g049080	chr5	AtNRT1.5/AtNPF7.3	K, NO3-	Shoots, Roots	Li et al. 2017
At1g32450+	Bo5g087130	chr5	AtNRT1.5/AtNPF7.3	K, NO3-	Shoots, Roots	Li et al. 2017
At1g36370	Bo8g025610	chr8	AtMSA1	S, Se	Shoots	Huang et al. 2016
At1g56160	Bo3g185830	chr3	myb72	Fe,Cd,Zn,Co,Mo	Leaf	Palmer et al. 2013
At1g56430	Bo1g087340	chr1	NAS4	Fe,Cd,Co,Mo	Leaf	-
At1g62180	Bo4g091470	chr4	AtAPR2	S, Se	Shoots	Loudet et al. 2007, Chao et al. 2014
At1g71200*	Bo6g112770	chr6	AtCITF1	Cu	Shoots, Anthers	Yan et al. 2017
At1g71200*	Bo7g080650	chr7	AtCITF1	Cu	Shoots, Anthers	Yan et al. 2017
At1g71200*	Bo01532s010	Scaffold01532	AtCITF1	Cu	Shoots, Anthers	Yan et al. 2017
At1g74770+	Bo6g118960	chr6	BTSL1	Fe,Mn,Zn	Leaf	Hindt et al. 2017
At1g74770+	Bo6g106530	chr6	-	-	-	-
At1g80830*	Bo2g095650	chr2	AtNRAMP1	Mn	Shoots, Roots	Cailliatte et al. 2010
At1g80830*	Bo6g080720	chr6	AtNRAMP1	Mn	Shoots, Roots	Cailliatte et al. 2010
At1g80830*	Bo6g124200	chr6	AtNRAMP1	Mn	Shoots, Roots	Cailliatte et al. 2010
At2g01770	Bo2g134570	chr2	VIT1	Fe	Seeds	Kim et al. 2006

At2g13540+	Bo3g087110	chr3	ABH1	S	Seeds	McDowell et al. 2013
At2g13540+	Bo9g022460	chr9	ABH1	S	Seeds	McDowell et al. 2013
At2g16770+	Bo3g091670	chr3	AtbZIP23	Zn	Shoots, Roots	Assunção et al. 2010
At2g16770+	Bo9g024230	chr9	AtbZIP23	Zn	Shoots, Roots	Assunção et al. 2010
At2g19110	Bo7g114220	chr7	AtHMA4	Zn	Shoots, Seeds	Hussain et al. 2004, Olsen et al. 2016
At2g21045	Bo7g014500	chr7	AtHAC1	As	Shoots	Chao et al. 2014
At2g23150+	Bo3g042290	chr3	AtNRAMP4	Fe,Mn,Zn	Shoots	Languar et al. 2010
At2g23150+	Bo8g100600	chr8	AtNRAMP4	Fe,Mn,Zn	Shoots	Languar et al. 2010
At2g25680+	Bo1g014650	chr1	MOT1	Mo	Leaf	Baxter et al. 2008
At2g25680+	Bo8g099010	chr8	MOT1	Mo	Leaf	Baxter et al. 2008
At2g28670+	Bo4g069300	chr4	ESB1	Ca,Mn,Zn,Na,S,K,As,Se,Mo	Leaf	Baxter et al. 2009
At2g28670+	Bo4g166640	chr4	ESB1	Ca,Mn,Zn,Na,S,K,As,Se,Mo	Leaf	Baxter et al. 2009
At2g38460*	Bo4g027330	chr4	FPN1	Co	Leaf	Morrissey et al. 2009
At2g38460*	Bo4g027290	chr4	FPN1	Co	Leaf	Morrissey et al. 2009
At2g38460*	Bo7g093140	chr7	FPN1	Co	Leaf	Morrissey et al. 2009
At2g39450+	Bo4g026360	chr4	AtMTP11	Mn	Shoots, Roots	Peiter et al. 2007
At2g39450+	Bo4g188200	chr4	AtMTP11	Mn	Shoots, Roots	Peiter et al. 2007
At2g46430	Bo4g009240	chr4	CNGC3	K	Leaf	Gobert et al. 2006
At2g46800	Bo4g006960	chr4	AtMTP1	Zn	Shoots	Desbrosses-Fonrouge et al. 2005
At3g06060	Bo5g144810	chr5	TSC10a	Na,K,Rb,Mg,Ca,Fe,Mo	Leaf	Chao et al. 2011
At3g06100	Bo5g144770	chr5	NIP7	As		Lindsay and Maathius 2016, Isayenkov and Maathius 2008
At3g12820+	Bo1g138940	chr1	myb10	Fe,Cd,Zn,Co,Mo	Leaf	Palmer et al. 2013
At3g12820+	Bo3g065150	chr3	myb10	Fe,Cd,Zn,Co,Mo	Leaf	Palmer et al. 2013
At3g14280+	Bo1g134480	chr1	-	S	Seeds	McDowell et al. 2013
At3g14280+	Bo5g130440	chr5	-	S	Seeds	McDowell et al. 2013
At3g15380	Bo3g066820	chr3	AtCTL1	Na, Fe, Zn, Mn, Mo	Shoots, Roots	Gao et al. 2017
At3g18290+	Bo1g117680	chr1	BTS	Fe,Zn,Mn	Leaf	Hindt et al. 2017
At3g18290+	Bo5g118870	chr5	BTS	Fe,Zn,Mn	Leaf	Hindt et al. 2017
At3g22890+	Bo3g081470	chr3	AtATPS1	S	Shoots	Koprivova et al. 2013

At3g22890+	Bo5g093400	chr5	AtATPS1	S	Shoots	Koprivova et al. 2013
At3g23210+	Bo1g102270	chr1	bHLH34	Fe	Root, shoot	Li et al. 2016
At3g23210+	Bo7g040830	chr7	bHLH34	Fe	Root, shoot	Li et al. 2016
At3g43790	Bo8g048300	chr8	ZIFL2	Cs	Leaf	Remy et al. 2015
At3g47640**	Bo1g072220	chr1	PYE	Fe,Zn,Mn,Co	Root	Long et al. 2010
At3g47640**	Bo3g114950	chr3	PYE	Fe,Zn,Mn,Co	Root	Long et al. 2010
At3g47640**	Bo3g107710	chr3	PYE	Fe,Zn,Mn,Co	Root	Long et al. 2010
At3g47640**	Bo3g168750	chr3	PYE	Fe,Zn,Mn,Co	Root	Long et al. 2010
At3g47950	NA	NA	AHA4	Na	Root	Vitart et al. 2001
At3g58060*	Bo6g074430	chr6	AtMTP8	Mn	Shoots, seeds	Eroglu et al. 2016, Eroglu et al. 2017
At3g58060*	Bo8g092400	chr8	AtMTP8	Mn	Shoots, seeds	Eroglu et al. 2016, Eroglu et al. 2017
At3g58060*	Bo00895s030	Scaffold00895	AtMTP8	Mn	Shoots, seeds	Eroglu et al. 2016, Eroglu et al. 2017
At3g58810	NA	-	AtMTP3	Zn	Shoots	Arrivault et al. 2006
At4g02780	Bo9g006080	chr9	GA1	Fe	Root	Wild et al. 2016
At4g10310	Bo2g110470	chr2	AtHKT1;1	Na	Leaf	Baxter et al. 2010
At4g14410	Bo4g126940	chr4	bHLH104	Fe	Root, shoot	Li et al. 2016
At4g16370	Bo8g043560	chr8	OPT3	Fe,Cd	Leaf	Zhai et al. 2014
At4g19690+	Bo1g021930	chr1	IRT1	Fe,Mn,Co,Cd,Zn	Root	Eide et al. 1996
At4g19690+	Bo7g106800	chr7	IRT1	Fe,Mn,Co,Cd,Zn	Root	Eide et al. 1996
At4g24120*	Bo1g037550	chr1	YSL1	Fe,Zn,Cu	NA	Waters et al. 2006
At4g24120*	Bo3g154250	chr3	YSL1	Fe,Zn,Cu	NA	Waters et al. 2006
At4g24120*	Bo7g109330	chr7	YSL1	Fe,Zn,Cu	NA	Waters et al. 2006
At4g30110	Bo1g014570	chr1	AtHMA2	Zn	Shoots, seeds	Hussain et al. 2004, Olsen et al. 2016
At4g30120+	Bo1g014550	chr1	HMA3	Cd	Leaf	Chao et al. 2012
At4g30120+	Bo7g114260	chr7	HMA3	Cd	Leaf	Chao et al. 2012
At4g35040+	Bo1g006780	chr1	AtbZIP19	Zn	Shoots, roots	Assunção et al. 2010
At4g35040+	Bo3g167300	chr3	AtbZIP19	Zn	Shoots, roots	Assunção et al. 2010
At5g02600+	Bo2g002660	chr2	NaKR1	Na,K,Rb	Leaf	Tian et al. 2010
At5g02600+	Bo9g183490	chr9	NaKR1	Na,K,Rb	Leaf	Tian et al. 2010

At5g03570	Bo3g002430	chr3	NaKR1	Na,K,Rb	Leaf	Tian et al. 2010
At5g15410	Bo9g164130	chr9	CNGC2/DND1	Ca,Mg	Seeds	McDowell et al. 2013
At5g18830+	Bo9g154530	chr9	AtSPL7	Cu	Shoots, roots	Bernal et al. 2012
At5g18830+	Bo00629s160	Scaffold00629	AtSPL7	Cu	Shoots, roots	Bernal et al. 2012
At5g35410	Bo4g139520	chr4	SOS2	Na	seeds	McDowell et al. 2013
At5g44070+	Bo2g120840	chr2	PCS1	Zn,Cd	Leaf	Kühnlenz et al. 2016
At5g44070+	Bo7g067020	chr7	PCS1	Zn,Cd	Leaf	Kühnlenz et al. 2016
At5g53130	NA	NA	CNGC1	Pb	Leaf	Sunkar et al. 2000
At5g53550+	Bo9g111420	chr9	YSL3	Fe,Zn,Cu	NA	Waters et al. 2006
At5g53550+	Bo04588s010	Scaffold04588	YSL3	Fe,Zn,Cu	NA	Waters et al. 2006
At5g57620*	Bo2g030240	chr2	AtMYB36	Li, B, Na, Mg, K, Ca, Mn, Fe, Co, Ni, Cu,Zn, Rb, Sr, Mo, Cd	Shoots	Kamiya et al. 2015
At5g57620*	Bo3g019230	chr3	AtMYB36	Li, B, Na, Mg, K, Ca, Mn, Fe, Co, Ni, Cu,Zn, Rb, Sr, Mo, Cd	Shoots	Kamiya et al. 2015
At5g57620*	Bo9g133340	chr9	AtMYB36	Li, B, Na, Mg, K, Ca, Mn, Fe, Co, Ni, Cu,Zn, Rb, Sr, Mo, Cd	Shoots	Kamiya et al. 2015
At5g64930+	Bo3g100730	chr3	CPR5	K	Leaf	Borghi et al. 2011
At5g64930+	Bo6g072900	chr6	CPR5	K	Leaf	Borghi et al. 2011
At5g67330	Bo7g063710	chr7	AtNRAMP3	Fe,Mn,Zn	Shoots	Languar et al. 2010

AI27 95% family-wise confidence level

Supplemental Figure 4.1 Tukey's test for significant differences among cultivars of *B. oleracea* for AI.

B11 95% family-wise confidence level

Supplemental Figure 4.2 Tukey's test for significant differences among cultivars of *B. oleracea* for B.

Ca44 95% family-wise confidence level

Supplemental Figure 4.3 Tukey's test for significant differences among cultivars of *B. oleracea* for Ca.

Cd111 95% family-wise confidence level

Supplemental Figure 4.4 Tukey's test for significant differences among cultivars of *B. oleracea* for Cd.

Supplemental Figure 4.5 Tukey's test for significant differences among cultivars of *B. oleracea* for Co.

Supplemental Figure 4.6 Tukey's test for significant differences among cultivars of *B. oleracea* for Cu.

Supplemental Figure 4.7 Tukey's test for significant differences among cultivars of *B. oleracea* for Fe.

Supplemental Figure 4.8 Tukey's test for significant differences among cultivars of *B. oleracea* for K.

Supplemental Figure 4.9 Tukey's test for significant differences among cultivars of *B. oleracea* for Mg.

Supplemental Figure 4.10 Tukey's test for significant differences among cultivars of *B. oleracea* for Mn.

Supplemental Figure 4.11 Tukey's test for significant differences among cultivars of *B. oleracea* for Mo.

Na23 95% family-wise confidence level

Supplemental Figure 4.12 Tukey's test for significant differences among cultivars of *B. oleracea* for Na.

Supplemental Figure 4.13 Tukey's test for significant differences among cultivars of *B. oleracea* for Ni.

P31 95% family-wise confidence level

Supplemental Figure 4.14 Tukey's test for significant differences among cultivars of *B. oleracea* for P.

Supplemental Figure 4.15 Tukey's test for significant differences among cultivars of *B. oleracea* for Rb.

Sr88 95% family-wise confidence level

Supplemental Figure 4.16 Tukey's test for significant differences among cultivars of *B. oleracea* for Sr.

Zn66 95% family-wise confidence level

Supplemental Figure 4.17 Tukey's test for significant differences among cultivars of *B. oleracea* for Zn.

Supplemental Figure 4.18 Tukey's test for significant differences among cultivars of *B. oleracea* for PC1.

Supplemental Figure 4.19 Tukey's test for significant differences among cultivars of *B. oleracea* for PC2.

Supplemental Figure 4.20 Tukey's test for significant differences among cultivars of *B. oleracea* for PC3.

Supplemental Figure 4.21 Tukey's test for significant differences among cultivars of *B. oleracea* for PC4.

A. Al27

B. B11

C. Ca44

D. Co59

E. Cu63

F. Mg26

G. P31

H. Sr88

Supplemental Figure 4.22 Overlap of SNPs among models for elements which had significant SNPs in more than one model.

CHAPTER 5 : A FERAL FUTURE

Makenzie E. Mabry*¹, Troy Rowan*², Jared E. Decker², J. Chris Pires¹

1 Division of Biological Sciences and Christopher S. Bond Life Sciences Center,
University of Missouri, Columbia, MO, U.S.A.

2 Division of Animal Sciences, University of Missouri, Columbia, USA.

*These authors contributed equally as first authors.

ABSTRACT

Feralization, the process of becoming de-domesticated, has intrigued researchers for centuries; however, the rigorous study of feral plants and animals is limited. Feral populations have been investigated frequently as invasive species and for biological control; however, we argue that ferality can be useful for understanding complex biological process such as natural selection. Specifically, we propose that feral plants and animals provide unique opportunities to study complex traits and are better populations to target for identification of traits for crop and livestock improvement. With genomic resources from the domesticated progenitors of these feral populations readily available, the time is ripe for research on feralization.

EMERGING INTEREST IN FERAL POPULATIONS

Recently, there has been a surge of interest in the investigation of feral animals and plants. Studies include chicken (Johnsson et al. 2016; Gering et al. 2019), pigeon (Giunchi et al. 2020; Carlen & Munshi-South 2020), pig (Fang et al. 2009), dingoes (Zhang et al. 2020), eggplant (Page et al. 2019), *Brassica* crops (Chen et al. 2020), rice (Qiu et al. 2017), the American bullfrog (Cruz-Cordovez et al. 2020), and even koi (Kohlmann et al. 2005). Collectively, these studies highlight that feralization, the process by which animals and plants revert back to a wild-like state, occurs across the tree of life. Although these feral populations have fascinated researchers since Darwin (1868), many questions remain about ferality, including how a feral population is defined and the genomic signatures of ferality. While there has been long standing interest in understanding feral populations to control population growth (i.e. cats and goats, invasive plants), there has recently been an interest in using these populations to further our understanding of evolution, especially on shorter timescales, as well as how these populations can be used for improving agricultural species (Gering et al. 2019). Thus, feral populations should not only be an interest for those wishing to control populations, but also more broadly to evolutionary biologists, systematists, breeders, and conservation biologists. However, we argue that research into feralization must be designed to powerfully confront the complex processes involved in the feralization process.

THE DIFFICULTIES IN DEFINING FERAL

When people hear the word “feral” they likely associate it with an animal which has escaped domestication and become more similar to their wild ancestors; however,

plants can become feral as well. These escaped populations have fascinated researchers, including Darwin, who have noted that although feral populations tend to revert to a wild type, this tendency is influenced by the environment (Darwin 1868). Examples and definitions for what a feral plant or animal is has been recently discussed by Gering et al. (2019). Broadly, feral populations can be described as endoferal or exoferal. Endoferal populations are those which result from domestication populations crossing with themselves while escaping cultivation, while exoferal populations outcross with either an already escaped population, wild population of the same species, other domesticated species, or other wild species (Gering et al. 2019). However, these descriptions of feral populations have many confounding factors such as weediness, timescale in which you are investigating, mating systems, introgression levels, and backcrossing frequency. We discuss how each of these make defining a feral species difficult below.

Weediness/Invasive

In many cases, feral plants and animals are growing or living in an area where they are not wanted. In some cases they may cause disruption to the environment which for some populations, has resulted in direct action in reducing these populations. Care needs to be taken when using these terms, for example weedy plants are those that are growing in areas in which they did not evolved or are not desired (Baker 1965), invasive is specific to causing harm to an area (Thompson 1991), while feral are populations no longer under cultivation (Gering et al. 2019). The challenge here is that populations may be described as more than one of these or the terms are used differently for different groups of organisms. For example, feral hogs cause a lot of economic harm, but are not called

invasive hogs. The important distinction being that feral populations may not cause harm, so this terminology creates confusion.

Timescales

Populations will range in the amount of time since the release of cultivation. We expect, especially with exoferal populations that there will be immediate effects vs long lasting effects, however, we have yet to fully understand these confounding factors. If feral populations are sampled soon after release of cultivation, they will look very similar to cultivar populations. But if the same population is sampled generations later, we may expect it to look like a wild relative (if still growing in the same area). So we ask how long does a population have to be uncultivated to be considered feral? Is there a genomic requirement to ferality? We hope that these questions can be addressed with additional studies across the tree of life addressing patterns of ferality.

Mating systems

Mating systems can further confuse terminology around feralization. Plant mating systems are diverse with respect to mechanisms of selfing or outcrossing, yet these differences have profound effects on what the genomics of the population looks like. In selfing populations, especially in the initial small populations that feral plants start as, drift will act quickly and may look different from cultivated populations soon after escape from cultivation. In outcrossing populations, feral plants may still breed with cultivated types, as well as wild relatives if they are in the area (Warwick & Stewart 2005). This can occur for a short time or longtime and will affect the ability to identify feral populations as unique entities.

Introgression from wild relatives

For those populations which are cultivated in the same area where initial domestication occurred, feral populations may breed with wild relatives (Chen et al. 2004; Arnaud et al. 2013). The genomic signatures of these populations would therefore have significant signatures of wild relatives, presenting difficulty in properly identifying the feral history of these plants. An additional difficulty is that to be able to identify these patterns, genomes are necessary.

Backcrossing to cultivation populations

Typically, feral populations occur near or adjacent to their cultivated counterparts, in some cases (potentially many cases for plants), backcrossing to the still cultivated population can occur (Warwick & Stewart 2005). This, like introgression from wild relatives, would leave signatures in the genome that could only be seen using full genomes

FERALIZATION IS NOT MERELY THE REVERSE OF DOMESTICATION

It is easy to think *a priori* that the process of fertilization would just reverse domestication, or that genes that were once under selection during cultivation would merely switch back to a wild state. However, studies show this is not the case (Johnsson et al. 2016; Qiu et al. 2017, Page et al. 2019; Zhang et al. 2020), rather feral populations take unique paths after they are no longer under cultivation. For traits controlled by single genes like branch number in maize (Doebley et al. 1997; Doebley 2004), shattering and compactness of wheat (Simons et al. 2006), fruit size in tomato (Frary et al. 2000), coat color for many domesticated animals (horses, sheep, goats, pig), and horns in ruminants (goats, sheep, cows), it may be more likely the gene could switch back to a wild state. Jain

(1977) found evidence for this in rye, when they found that the gene controlling shattering was under positive selection in feral populations in California. However for many traits, especially those which are quantitative in nature and influence the whole genome (i.e., metabolomic and yield), simply reverting back to a wild state would be a complicated process and would affect an entire network of genes (Washburn et al. 2016). Several studies find that this is indeed true (Johnsson et al. 2016; Qiu et al. 2017, Page et al. 2019; Zhang et al. 2020). Once these populations are released from cultivation, new pressures from the environment, such as drought, herbivory, and predation as well as interactions with both wild and cultivated populations lead to unique populations, which is further complicated if it is in a new location, or one different than originally occurring.

Feralized populations, are not simply escaped, isolated domesticates existing in the wild (with some exceptions such as Dingoes; Zhang et al. 2020). Rather, ferals are constantly admixing with natural, domestic, and other feral populations (Gering et al. 2015; Wang et al. 2017). This mirrors the initial domestication event, which is also accompanied by partial domestications, re-domestications, admixture, and wild introgression (Frantz et al. 2020). This introgression leads to mosaic genomes where beneficial alleles from each source population are changing in frequency. Both wild and domestic alleles can increase fitness in the feral population (Barbato et al. 2017; Bianco et al. 2015). We believe that feral populations are ideal for studying the role of admixture and adaptive introgression in the wild. Genomic resources and cataloged diversity from agricultural populations make the identification of preferentially introgressed regions possible.

FERAL POPULATIONS MAY MODEL NATURAL SELECTION BETTER THAN DOMESTICATION

The transition of a population from domestic to feral is accompanied by a range of genetic, phenotypic, and demographic changes. Studying domestication has uncovered the downstream impacts of bottlenecks and directional selection (Wright et al. 2005; Erye-Walker et al. 1998). Darwin was especially interested in selection on domesticated species, and discussed their connection to natural selection in his *On the Origin of Species*, and more in depth in his *The Variation of Animals and Plants Under Domestication* (Darwin 1859; Darwin 1868). He noticed a clear connection between the ability of farmers to alter characteristics of domesticated plants and animals and the selection pressures of natural selection. Feral populations allow us to observe the earliest effects of natural selection when a population is exposed to new environmental stressors, and a large shift in fitness optimum.

The natural selection pressures placed on newly feralized populations are twofold; a change in type, and increased specificity for adaptations to a local environment. We expect a shift in the type of selective pressure placed on the population. Directional selection tends to be the major pressure on domesticated populations (Andersson and Georges 2004). We would expect both early selection on domestication genes or pathways and ongoing selection for increased tameness, yields, growth rates, or other production-based characteristics to be directional in nature (García-Ruiz et al. 2016; Xue et al. 2016; Trut et al. 2009). The transition to ferality would initially be most likely accompanied by a transition to the stabilizing selection experienced by most natural populations (Charlesworth et al. 1982; Qiu et al. 2017). Phenotypes involved in local adaptation,

immune function, and disease resistance are under selection at varying levels in domestic populations, due mostly to their positive correlation with production traits (Rowan et al. 2020; Fariello et al. 2014). While instances of stabilizing selection occur in domesticated populations, they are substantially more infrequent in natural populations.

In addition to the population-wide shift to stabilizing selection, we would expect subpopulations to experience increased selection on phenotypes that modulate adaptation in their local environment. Environmental stressors such as drought, local pests or parasites, and novel predators can be managed in an agricultural context, but will exert different selective pressures on feral populations across landscapes. For example, local adaptation acted as the major selective pressure following the de-domestication of weedy rice. While seed shattering genes were identified as being under selection, the vast majority of differences were found in genes that modulate disease or pest resistance and salt or drought tolerance (Qiu et al. 2017). Broad geographic sampling of feralized populations will allow us to begin disentangling the changes feral populations undergo due to natural selection versus local adaptation. This has implications for understanding the biology underlying local adaptation in both domestic and wild populations as well as insights into the management of regional feral populations.

The genomic resources available for many feral agricultural species will allow us to better understand how natural selection acts on the genome in real time. Like in domestication, feralized populations can exhibit radical phenotypic changes over relatively short timescales. The genomic changes that underlie these rapid phenotypic changes are not well-understood. Some of these are driven by selection on mutations that control Mendelian phenotypes that otherwise would have been purged from the population, like

selection on coat color in animals (Linderholm and Larson 2013). In domestic pig populations, multiple mutations exist in the *melanocortin receptor 1* (*MC1R*) gene lead to substantial coat color diversity amongst domestic populations that can quickly fluctuate when artificial selection pressures are removed, leading to substantial coat color variation in feral populations (Fang et al. 2009). Beyond the appearance-based Mendelian phenotypes, little work has explored the ability of polygenic selection to rapidly change phenotypes in feralized populations. Subtle changes at thousands of loci can drive rapid phenotypic change across a population (Hayward and Sella 2019; Höllinger et al. 2019; Simons et al. 2018).

FERAL POPULATIONS CAN HELP US UNDERSTAND LOCAL ADAPTATION AND IMPROVE AGRICULTURAL SPECIES

As the climate changes, there has been increased interest in integrating locally-adaptive variants from wild or related populations into domesticated plant and animal species (Turner-Hissong et al. 2020; Box 2). This process of crossing and subsequent selection on an introgressed variant that increases agricultural production or fitness is referred to as adaptive introgression (Hendrick 2013). Adaptive introgression is well-documented in natural populations, and is being further exploited as a method for increasing the adaptability of agricultural species (Dempewolf et al. 2017; Burgarella et al. 2019). Crop species such as maize (Hufford et al. 2013), sunflower (Whitney et al. 2010), and potato (Jansky et al. 2019) have utilized wild variation to introgress adaptive variants into domesticated varieties. However, feral populations can perhaps provide a better source for adaptive introgression. For example, some species, instead of moving in space, escape

climate change ‘in time’ by shifting traits such as flowering time to more favorable conditions; this response has resulted in earlier flowering dates for many temperate plants (Primack et al. 2004; Franks et al. 2007). While cultivated crops are bred for specific timing in flowering and fruiting to ensure maximal yield and ease of harvest, crop species that have escaped cultivation are free to adapt to what is best in their local environment. Similarly, feral carp, which have resistance to the cyprinid herpes virus, a devastating disease to cultivated carp, have been studied as sources for introgression to breed resistant fish. Therefore we argue that feral populations can advance the practice of adaptive introgression and improvement of agricultural species in three ways: 1) Through traditional introgression with feral-domesticated hybridization and backcrossing, 2) By identifying locally adaptive variants segregating in feral individuals as candidates for precision introgression via gene-editing, or 3) Through the re-domestication of locally-adapted feral species.

Using feral or weedy relatives of agricultural species for hybridization and adaptive introgression may work better than wild relatives. Gene flow from wild relatives can benefit domesticated species through returning genetic variation lost during selective breeding (Aitken and Whitlock 2013). While crossing with feral individuals may not increase the genetic variation at the same rate as with a wild species, we expect that having a genetic background more similar to the domesticated would increase the efficiency of adaptive introgression (Dwivedi et al. 2007; Prohens et al. 2017). Feral populations also have adapted to their environment, potentially making them more fit in a given area. Therefore, adaptive alleles of large or small effect, could be introgressed without large

initial drops in productivity, reducing the necessary number of backcrosses, and make species more locally fit.

Feral populations offer intriguing opportunities to identify candidate loci for genome editing techniques such as CRISPR that can enhance adaptive phenotypes in crops or livestock. The transition from domestic to feral is accompanied by a shift in selection pressure from production traits to fitness in an unmanaged environment. In this sense, we suspect that soft selective sweeps on standing variation are likely to occur in the feral population. We expect that the strongest selection pressures placed on recently-feralized populations will be related to immune function and response to abiotic stressors, not necessarily on reversal of domesticated traits. We would expect that most feral populations experience extensive gene flow with related wild populations (Gering et al. 2019), producing areas of the genome enriched for introgressed wild alleles.

In addition to creating locally adapted lines, the process of “re-domesticating” feral populations could also generate novel genetic variation for breeding programs, or even entirely new crops species. Re-domestication has been explored in the context of performing modern domestications of domestic crops (Zsögön et al. 2018; Fernie & Yan 2019; <https://teosinte.wisc.edu/redomestication>), but not in the context of feral individuals. However, at least one example of a feral plant (Cherry tomatoes; Bauchet & Causse 2012) becoming re-domesticated exists.

CONCLUDING REMARKS

Above we argue how useful and exciting the study of feralization can be, including the potential for feral populations to be studied as independent natural selection

experiments that occur across the tree of life. These populations, although simply, consequences of domestication offer exciting new avenues for research as well as potential sources to make cultivated plants and animals more locally adapted. The wealth of associated genomic data due to the proximity to cultivated populations, combined with thorough sampling should prove feral populations to be powerful datasets to understand complex biological systems.

ACKNOWLEDGMENTS

We thank Sarah Turner-Hissong for thought provoking conversations on the usefulness and excitement of feral populations.

REFERENCES

- Aitken, S. N., & Whitlock, M. C. 2013. Assisted gene flow to facilitate local adaptation to climate change. *Annual Review of Ecology, Evolution, and Systematics*, 44, 367-388.
- Andersson, L., & Georges, M. 2004. Domestic-animal genomics: deciphering the genetics of complex traits. *Nature Reviews Genetics*, 5(3), 202-212.
- Arnaud, J. F., Viard, F., Delescluse, M., & Cuguen, J. 2003. Evidence for gene flow via seed dispersal from crop to wild relatives in *Beta vulgaris* (Chenopodiaceae): consequences for the release of genetically modified crop species with weedy lineages. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1524), 1565-1571.
- Baker, H. G. 1965. In HG Baker & GL Stebbins (Eds.), *Characteristics and modes of origin of weeds. The genetics of colonizing species* (pp. 147-172).
- Barbato, M., Hailer, F., Orozco-ter, Wengel, P., Kijas, J., Mereu, P., Cabras, P., et al. 2017. Genomic signatures of adaptive introgression from European mouflon into domestic sheep. *Scientific Reports*, 7(1), 1-13.
- Bauchet, G., & Causse, M. 2012. Genetic diversity in tomato (*Solanum lycopersicum*) and its wild relatives. *Genetic diversity in plants*, 133, 162.
- Bianco, E., Soto, H. W., Vargas, L., & Pérez-Enciso, M. 2015. The chimerical genome of Isla del Coco feral pigs (Costa Rica), an isolated population since 1793 but with remarkable levels of diversity. *Molecular ecology*, 24(10), 2364-2378.

- Burgarella, C., Barnaud, A., Kane, N. A., Jankowski, F., Scarcelli, N., Billot, C., et al. 2019. Adaptive introgression: An untapped evolutionary mechanism for crop adaptation. *Frontiers in plant science*, 10, 4.
- Carlen, E., & Munshi-South, J. 2020. Widespread genetic connectivity of feral pigeons across the Northeastern megacity. *Evolutionary Applications*.
- Charlesworth, B., Lande, R., & Slatkin, M. 1982. A neo-Darwinian commentary on macroevolution. *Evolution*, 36(3), 474-498.
- Chen, L. J., Lee, D. S., Song, Z. P., Suh, H. S., & LU, B. R. 2004. Gene flow from cultivated rice (*Oryza sativa*) to its weedy and wild relatives. *Annals of botany*, 93(1), 67-73.
- Chen, R., Shimono, A., Aono, M., Nakajima, N., Ohsawa, R., & Yoshioka, Y. 2020. Genetic diversity and population structure of feral rapeseed (*Brassica napus* L.) in Japan. *Plos one*, 15(1), e0227990.
- Cruz-Cordovez, C., Herrera, I., Espinoza, F., Rizzo, K., Sarmiento, M. B., Rodas, N., et al. 2020. New record of a feral population of *Lithobates catesbeianus* Shaw, 1802 in a protected area (Santay Island) in the Ecuadorian coast. *BioInvasions Records*, 9(2), 421-433.
- Darwin, C. 1859. *On the origin of species*. London: John Murray.
- Darwin, C. R. 1868) *Variation of plants and animals under domestication*. London: John Murray.
- Dempewolf, H., Baute, G., Anderson, J., Kilian, B., Smith, C., & Guarino, L. 2017. Past and future use of wild relatives in crop breeding. *Crop Science*, 57(3), 1070-1082.
- Doebley, J. 2004. The genetics of maize evolution. *Annu. Rev. Genet.*, 38, 37-59.

- Doebley, J., Stec, A., & Hubbard, L. 1997. The evolution of apical dominance in maize. *Nature*, 386(6624), 485-488.
- Dwivedi, S. L., Crouch, J. H., Mackill, D. J., Xu, Y., Blair, M. W., Ragot, M., et al. 2007. The molecularization of public sector crop breeding: progress, problems, and prospects. *Advances in agronomy*, 95, 163-318.
- Eyre-Walker, A., Gaut, R. L., Hilton, H., Feldman, D. L., & Gaut, B. S. 1998. Investigation of the bottleneck leading to the domestication of maize. *Proceedings of the National Academy of Sciences*, 95(8), 4441-4446.
- Fang, M., Larson, G., Ribeiro, H. S., Li, N., & Andersson, L. 2009. Contrasting mode of evolution at a coat color locus in wild and domestic pigs. *PLoS genetics*, 5(1).
- Fariello, M. I., Servin, B., Tosser-Klopp, G., Rupp, R., Moreno, C., San Cristobal, M., et al. 2014. Selection signatures in worldwide sheep populations. *PLoS One*, 9(8).
- Fernie, A. R., & Yan, J. 2019. De novo domestication: an alternative route toward new crops for the future. *Molecular plant*.
- Franks, S. J., Sim, S., & Weis, A. E. 2007. Rapid evolution of flowering time by an annual plant in response to a climate fluctuation. *Proceedings of the National Academy of Sciences*, 104(4), 1278-1282.
- Frantz, L. A., Bradley, D. G., Larson, G., & Orlando, L. 2020. Animal domestication in the era of ancient genomics. *Nature Reviews Genetics*, 1-12.
- Frary, A., Nesbitt, T. C., Frary, A., Grandillo, S., Van Der Knaap, E., Cong, B., et al. 2000. fw2. 2: a quantitative trait locus key to the evolution of tomato fruit size. *Science*, 289(5476), 85-88.

- García-Ruiz, A., Cole, J. B., VanRaden, P. M., Wiggans, G. R., Ruiz-López, F. J., & Van Tassell, C. P. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proceedings of the National Academy of Sciences*, 113(28), E3995-E4004.
- Gering, E., Incorvaia, D., Henriksen, R., Wright, D., & Getty, T. 2019. Maladaptation in feral and domesticated animals. *Evolutionary applications*, 12(7), 1274-1286.
- Gering, E., Johnsson, M., Willis, P., Getty, T., & Wright, D. 2015. Mixed ancestry and admixture in Kauai's feral chickens: invasion of domestic genes into ancient Red Junglefowl reservoirs. *Molecular ecology*, 24(9), 2112-2124.
- Giunchi, D., Mucci, N., Bigi, D., Mengoni, C., & Baldaccini, N. E. 2020. Feral pigeon populations: their gene pool and links with local domestic breeds. *bioRxiv*.
- Hayward, L. K., & Sella, G. 2019. Polygenic adaptation after a sudden change in environment. *bioRxiv*, 792952.
- Hedrick, P. W. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular ecology*, 22(18), 4606-4618.
- Höllinger, I., Pennings, P. S., & Hermisson, J. 2019. Polygenic adaptation: From sweeps to subtle frequency shifts. *PLoS genetics*, 15(3), e1008035.
- Hufford, M. B., Lubinsky, P., Pyhäjärvi, T., Devengenzo, M. T., Ellstrand, N. C., & Ross-Ibarra, J. 2013. The genomic signature of crop-wild introgression in maize. *PLoS genetics*, 9(5).
- Jain, S. K. 1977. Genetic Diversity of Weedy Rye Populations in California 1. *Crop Science*, 17(3), 480-482.

- Jansky, S., Haynes, K., & Douches, D. 2019. Comparison of Two Strategies to Introgress Genes for Resistance to Common Scab from Diploid *Solanum chacoense* into Tetraploid Cultivated Potato. *American Journal of Potato Research*, 96(3), 255-261.
- Johnsson, M., Gering, E., Willis, P., Lopez, S., Van Dorp, L., Hellenthal, G., et al. 2016. Feralisation targets different genomic loci to domestication in the chicken. *Nature communications*, 7, 12950.
- Kohlmann, K., Kersten, P., & Flajšhans, M. 2005. Microsatellite-based genetic variability and differentiation of domesticated, wild and feral common carp (*Cyprinus carpio* L.) populations. *Aquaculture*, 247(1-4), 253-266.
- Linderholm, A., & Larson, G. 2013. The role of humans in facilitating and sustaining coat colour variation in domestic animals. In *Seminars in cell & developmental biology* (Vol. 24, No. 6-7, pp. 587-593). Academic Press.
- Page, A., Gibson, J., Meyer, R. S., & Chapman, M. A. 2019. Eggplant Domestication: Pervasive Gene Flow, Feralization, and Transcriptomic Divergence. *Molecular biology and evolution*, 36(7), 1359-1372.
- Primack, D., Imbres, C., Primack, R. B., Miller-Rushing, A. J., & Del Tredici, P. 2004. Herbarium specimens demonstrate earlier flowering times in response to warming in Boston. *American Journal of Botany*, 91(8), 1260-1264.
- Prohens, J., Gramazio, P., Plazas, M., Dempewolf, H., Kilian, B., Díez, M. J., et al. 2017. Introgressiomics: a new approach for using crop wild relatives in breeding for adaptation to climate change. *Euphytica*, 213(7), 158.

- Qiu, J., Zhou, Y., Mao, L., Ye, C., Wang, W., Zhang, J., et al. 2017. Genomic variation associated with local adaptation of weedy rice during de-domestication. *Nature communications*, 8(1), 1-12.
- Rowan, T. N., Durbin, H. J., Seabury, C. M., Schnabel, R. D., & Decker, J. E. 2020. Powerful detection of polygenic selection and environmental adaptation in US beef cattle populations. *bioRxiv*.
- Simons, Y. B., Bullaughey, K., Hudson, R. R., & Sella, G. 2018. A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS biology*, 16(3), e2002985.
- Simons, K.J., Fellers, J.P., Trick, H.N., Zhang, Z., Tai, Y.S., Gill, B.S., & Faris, J.D. 2006. Molecular characterization of the major wheat domestication gene q. *Genetics* 172, 547–555.
- Thompson, J. D. 1991. The biology of an invasive plant. *BioScience*, 41(6), 393-401.
- Trut, L. N., Herbeck, Y. E., Kharlamova, A. V., Gulevich, R. G., & Kukekova, A. V. 2013. Fox domestication: molecular mechanisms involved in selection for behavior. *Russian Journal of Genetics: Applied Research*, 3(6), 419-425.
- Turner-Hissong, S. D., Mabry, M. E., Beissinger, T. M., Ross-Ibarra, J., & Pires, J. C. 2020. Evolutionary insights into plant breeding. *Current Opinion in Plant Biology*, 54, 93-100.
- Wang, H., Vieira, F. G., Crawford, J. E., Chu, C., & Nielsen, R. 2017. Asian wild rice is a hybrid swarm with extensive gene flow and feralization from domesticated rice. *Genome research*, 27(6), 1029-1038.

- Warwick, S. I., & Stewart, C. N. 2005. Crops come from wild plants: how domestication, transgenes, and linkage together shape ferality. *Crop ferality and volunteerism*, 36(1), 9-30.
- Washburn, J.D., Birk, K.A., Conant, G.C., & Pires, J.C. 2016. Convergent Evolution and the Origin of Complex Phenotypes in the age of Systems Biology. *International Journal of Plant Sciences* 177: 305-318.
- Whitney, K. D., Randell, R. A., & Rieseberg, L. H. 2010. Adaptive introgression of abiotic tolerance traits in the sunflower *Helianthus annuus*. *New Phytologist*, 187(1), 230-239.
- Wright, S. I., Bi, I. V., Schroeder, S. G., Yamasaki, M., Doebley, J. F., McMullen, M. D., & Gaut, B. S. 2005. The effects of artificial selection on the maize genome. *Science*, 308(5726), 1310-1314.
- Xue, S., Bradbury, P. J., Casstevens, T., & Holland, J. B. 2016. Genetic architecture of domestication-related traits in maize. *Genetics*, 204(1), 99-113.
- Zhang, S. J., Wang, G. D., Ma, P., Zhang, L. L., Yin, T. T., Liu, Y. H., et al. 2020. Genomic regions under selection in the feralization of the dingoes. *Nature communications*, 11(1), 1-10.
- Zsögön, A., Čermák, T., Naves, E. R., Notini, M. M., Edel, K. H., Weinl, S., et al. 2018. De novo domestication of wild tomato using genome editing. *Nature biotechnology*, 36(12), 1211-1216.

VITA

Makenzie E. Mabry grew up in Los Angeles county, California. Throughout high school and into college Makenzie was a competitive runner. In college Makenzie participated on the San Diego State Track and Cross-Country teams where she competed in steeplechase and the 5K. During college, Makenzie met Dr. Michael Simpson who introduced her to the world of plants. After graduating college Makenzie worked for two years as a veterinarian assistance and assistant supervisor for Sea World San Diego. During this time, she continued to volunteer with Dr. Simpson in studying the morphology of *Cryptantha pterocarya*, which eventually turned into a Master's project on the monophyly and biogeography of the genus. During her Master's she met Dr. J. Chris Pires at a BOTANY conference in New Orleans. Her first memory of him of course includes twitter. Somehow, Dr. Pires convinced the California native to pack up her things and two cats to move to Missouri to work on a PhD. After the completion of her degree, Makenzie, her fiancé, and their two cats (Pokee and Lani), will move to Gainesville, Florida where Makenzie will start as a Postdoctoral researcher with Drs. Pam and Doug Soltis at the University of Florida.