Brassicales phylogeny inferred from 72 plastid genes: A reanalysis of the phylogenetic localization of two paleopolyploid events and origin of novel chemical defenses

Patrick P. Edger1,2,12, Jocelyn C. Hall3, Alex Harkess4,5, Michelle Tang6, Jill Coombs5, Setareh Mohammadin1, M. Eric Schranz1, Zhiyong Xiong8, James Leebens-Mack4, Blake C. Meyers5, Kenneth J. Sytsma9, Marcus A. Koch10, Ihsan A. Al-Shehbaz11, and J. Chris Pires6

PREMISE OF THE STUDY: Previous phylogenetic studies employing molecular markers have yielded various insights into the evolutionary history across Brassicales, but many relationships between families remain poorly supported or unresolved. A recent phylotranscriptomic approach utilizing 1155 nuclear markers obtained robust estimates for relationships among 14 of 17 families. Here we report a complete family-level phylogeny estimated using the plastid genome.

METHODS: We conducted phylogenetic analyses on a concatenated data set comprising 44,926 bp from 72 plastid genes for species distributed across all 17 families. Our analysis includes three additional families, Tovariaceae, Salvadoraceae, and Setchellanthaceae, that were omitted in the previous phylotranscriptomic study.

KEY RESULTS: Our phylogenetic analyses obtained fully resolved and strongly supported estimates for all nodes across Brassicales. Importantly, these findings are congruent with the topology reported in the phylotranscriptomic study. This consistency suggests that future studies could utilize plastid genomes as markers for resolving relationships within some notoriously difficult clades across Brassicales. We used this new phylogenetic framework to verify the placement of the At-α event near the origin of Brassicaceae, with median date estimates of 31.8 to 42.8 million years ago and restrict the At-β event to one of two nodes with median date estimates between 85 to 92.2 million years ago. These events ultimately gave rise to novel chemical defenses and are associated with subsequent shifts in net diversification rates.

CONCLUSIONS: We anticipate that these findings will aid future comparative evolutionary studies across Brassicales, including selecting candidates for whole-genome sequencing projects.

KEY WORDS: phylogenomics; glucosinolates; secondary metabolites; evolutionary novelties.
The order Brassicales, which contains ~4700 species or ~2.2% of eudicot diversity (Magallón et al., 1999; Kiefer et al., 2014; Cardinal-McTeague et al., 2016), is a monophyletic group consisting of 17 families including several model families for evolutionary biology (Rodman et al., 1996, 1998; Hall et al., 2004; Soltis et al., 2011; Chase et al., 2016). The primary families for such studies include Brassicaceae, Cleomaceae, and Caricaceae with dozens of sequenced reference genomes (Koenig and Weigel, 2015) and numerous genome projects that are currently underway. Brassicales are known for their rich diversity of morphological, physiological, developmental, and chemical traits, and a range of projects are investigating the genetic basis of this diversity, including transitioning from C₃ to C₄ photosynthesis [e.g., Gynandropsis gynandra (L.) Briq.; Külahgölü et al., 2014; van den Bergh et al., 2014], abiotic stress tolerances [e.g., Eutrema salsugineum (Pall.) Al-Shehbaz & S.I.Warwick; Yang et al., 2013], and evolution of sex chromosomes (e.g., Carica papaya L.; Ming et al., 2008).

Despite this incredible diversity in form, members of the order are united by chemical defenses called glucosinolates (i.e., mustard oils) (Rodman et al., 1998; Schranz et al., 2011). Glucosinolates are highly variable across the order (Fig. 1), including over 120 unique compounds that are grouped into distinct classes based on their amino acid precursor (Fahey et al., 2001; Kliebenstein et al., 2001; Mithen et al., 2010). Some glucosinolates have been described outside the Brassicales, including Euphorbiaceae (Rodman et al., 1998). However, methionine-derived and tryptophan-derived (indolic) glucosinolates are novel to only certain families within Brassicales (Fahey et al., 2001; Schranz et al., 2011; van den Bergh et al., 2016). These novel classes of glucosinolates evolved, at least in part, due to an ancient arms race with insect herbivores, including butterflies of the subfamily Pierinae (Pieridae, Lepidoptera) (Ehrlich and Raven, 1964; Edger et al., 2015). Gene and genome duplication events followed by neofunctionalization played a central role driving the escalation of novel glucosinolate diversity across Brassicales (Kliebenstein, 2008; Hofberger et al., 2013; Edger et al., 2015). For example, the regulatory and core biosynthetic pathways for these novel glucosinolate classes arose from retained duplicates from both single-gene (tandem) and an ancient whole-genome duplication (At-β) shared by most families of Brassicales (Edger et al., 2015). The duplicated pathways from the At-β event evolved to synthesize a novel class of chemical defense compounds from a new amino acid substrate. Similarly, new biosynthetic steps involved in novel structural elaborations unique to Brassicaceae arose due to retained duplicate genes from the At-a whole-genome and tandem duplications (Hofberger et al., 2013; Edger et al., 2015). A more recent study investigated the origin of certain specialized metabolites across Solanaceae, revealing that gene duplications and functional diversification also played a central role in generating novelty (Moghe et al., 2017). These studies, among others, support Susumu Ohno’s hypothesis (Ohno, 1970) that gene duplications have been instrumental in driving the origin of new features and increases in organismal complexity and species diversity over evolutionary time (Freeding and Thomas, 2006; Edger and Pires, 2009).

A robust and well-resolved phylogeny is needed to interpret the wealth of data available from Brassicales species in an evolutionary context, including glucosinolate diversity. As with many diverse plant lineages, resolving the Brassicales phylogeny has been a combination of robust relationships being evident in earliest analyses with other groups proving more challenging to place robustly. For example, early studies identified a species-rich core group of families, referred to as the core Brassicales (Rodman et al., 1994, 1996), based on synapomorphy of an extended rbcl sequence found in Brassicaceae, Capparaceae, Cleomaceae, Gyrostemonaceae, Resedaceae, and Tovariaceae. Additional studies expanded this well-supported clade to eight families that also have an aspartic acid rather than the usual stop codon at the end of the gene: Brassicaceae, Cleomaceae, Capparaceae, Emblingiaceae, Gyrostemonaceae, Pentadiplandraceae, Resedaceae, and Tovariaceae (Hall et al., 2004; Su et al., 2012; Cardinal-McTeague et al., 2016). Within the core Brassicales, many relationships are strongly supported based on multiple analyses. A Brassicaceae-Cleomaceae clade is strongly supported as sister to Capparaceae (Hall et al., 2002, 2004; Hall, 2008; Edger et al., 2015). Gyrostemonaceae and Resedaceae are a strongly supported clade (Hall et al., 2004; Ronse De Craene and Haston, 2006; Edger et al., 2015; Cardinal-McTeague et al., 2016) that are sister to Pentadiplandraceae (Edger et al., 2015; Cardinal-McTeague et al., 2016). The Brassicaceae-Cleomaceae-Capparaceae and Gyrostemonaceae-Resedaceae-Pentadiplandraceae clades are strongly supported as sister groups, with this clade being sister to Emblingiaceae (Edger et al., 2015). In contrast, the phylogenetic placement of Tovariaceae within the core Brassicales remains unresolved despite broad taxonomic sampling and five gene regions (e.g., in polytomy in Cardinal-McTeague et al., 2016). Notably, Edger et al. (2015) were able to robustly estimated relationships within core Brassicales using 1155 nuclear markers, but they did not include Tovariaceae. Thus, understanding relationships of Tovariaceae may be dependent on a substantial increase in data.

The theme of variable resolution of phylogenetic relationships extends beyond the core Brassicales, which is compounded by the omission of key taxa in some analyses. There are a number of family-level phylogenetic relationships that have been established outside the core Brassicales. Studies support the following sister-family pairs: Koebeliniaceae-Bataceae, Tropaeolaceae-Akaniaeceae, and Caricaceae-Moringaceae (Olson, 2002a, b; Hall et al., 2004; Cardinal-McTeague et al., 2016). However, an alternate relationship for Bataceae has been estimated, a Bataceae-Salvadoraceae clade (Ronse De Craene and Haston, 2006), which is congruent with findings of separate Bataceae and Koebeliniaceae lineages (Edger et al., 2015). Salvadoraceae was also not included in our previous phylotranscriptomic study (Edger et al., 2015). A clade that includes the core Brassicales, Koebeliniaceae, Bataceae, Salvadoraceae, Limnanthaceae, and Setchellanthaceae is strongly supported (Hall et al., 2004), with Setchellanthaceae as the earliest-diverging lineage of this group (Rodman et al., 1996, 1998; Karol et al., 1999; Chandler and Bayer, 2000; Ronse De Craene and Haston, 2006). Notably, our previous phylotranscriptomic study (Edger et al., 2015) did not sample either Salvadoraceae or Setchellanthaceae, of which the latter family was previously hypothesized as an important family to phylogenetically localize the At-β event (Schranz et al., 2011). The Caricaceae-Moringaceae and Akaniaeceae-Tropaeolaceae clades are strongly supported clades and are the earliest-diverging families of the order Brassicales (Ronse De Craene and Haston, 2006; Edger et al., 2015; Cardinal-McTeague et al., 2016). Last, the nodes along the backbone of the phylogeny are strongly supported in recent studies (Edger et al., 2015; Cardinal-McTeague et al., 2016), but neither study included every Brassicales family. Phylogenomic analyses utilizing coding sequences from the plastid genome have obtained strongly supported phylogenetic estimates for some notoriously difficult plant clades at various taxonomic levels, including angiosperms (Jansen et al., 2007),
FIGURE 1. Phylogenetic relationships across Brassicales: ancient whole-genome duplications, glucosinolate diversity, and shifts in net diversification rates. The Brassicales topology of the 17 families, estimated using 72 genes from the plastid genome, was utilized to phylogenetically localize the At-α (red star) and At-β (purple star) whole-genome duplications using data from Edger et al. (2015). Bootstrap values for nodes less than 100% are shown below adjacent branches. Arabidopsis, Brassica, and Aethionema are Brassicaceae (orange box). The core Brassicales are highlighted within the green box. Species richness for every family is provided in the column to the right of the phylogeny, with Brassicaceae split into the Aethionema lineage (48 total) (Mohammadin et al., 2017) and all other species (3925 total) (Al-Shehbaz, 2012; Koch et al., 2012). The two possible phylogenetic positions for the At-β whole-genome duplication are depicted with purple stars. The origins of indolic glucosinolates, Met-derived glucosinolates, and novel structural elaborations are shown with blue arrows; see Edger et al. (2015) for additional details and timing of these events. Date estimates for nodes are shown above adjacent branches. Three significant shifts in net diversification rates (black circles) were previously reported, including at the origin of Met-glucosinolates, base of the Gyrostemonaceae-Resedaceae clade, and origin of novel structural elaborations (Edger et al., 2015; Cardinal-McTeague et al., 2016). The whole-genome duplication unique to Cleomaceae is shown with a green star (Barker et al., 2009).
basal angiosperms (Leebens-Mack et al., 2005), eudicots (Moore et al., 2010), grasses (Cotton et al., 2015; Saarela et al., 2015), and Brassicaceae (Hohmann et al., 2015; Mandáková et al., 2017). Here, we report a phylogenetic estimate of the relationships among all 17 Brassicales families based on 72 plastid coding regions assembled from Illumina next-generation sequencing (NGS) data. We had three major goals for this study: (1) obtain a strongly supported phylogenetic estimate for all family-level relationships across Brassicales, (2) phylogenetically localize the At-α and At-β events, and (3) re-evaluate the phylogenetic distribution of novel glucosinolates that are associated with significant shifts in net diversification rates. Last, we hope these results will help guide the selection of species for future genome sequencing projects for evolutionary studies across Brassicales.

MATERIALS AND METHODS

Data sets

Total DNA was extracted from either fresh or silica-dried leaf material using the Qiagen DNeasy Plant Mini Kit (Qiagen, Germantown, MD, USA) (Invitrogen, Carlsbad, CA, USA). A complete list of taxa, voucher information, and access to sequence data is given in Appendix S1 (see the Supplemental Data with this article). The barcoded DNA Illumina libraries were constructed using the NEB prep E600L kit (New England Biolabs, Ipswich, MA, USA) and sequenced single end with 100-bp reads on both the Illumina (San Diego, CA, USA) GAII and HiSeq-2000 instrument at the University of Missouri DNA core. For Gyrostemon racemiger H.Walter, total RNA was extracted, instead of DNA, from fresh young leaf tissue using the Invitrogen PureLink RNA Mini Kit, converted into an Illumina library using the TruSeq RNA kit (Illumina), and paired-end 100-bp reads were sequenced on the HiSeq-2000 instrument at the University of Missouri DNA core. Sequence data for Arabidopsis thaliana (L.) Heynh. (NC_001284.2 and NC_000932.1), Carica papaya (NC_012116.1 and NC_010323.1), and Brassica oleracea L. (SRP010908) were obtained from NCBI-GenBank and NCBI-SRA. Raw illumina data generated in this study have been deposited in either the Dryad Digital Repository (https://doi.org/10.5061/dryad.kc5b8) or the NCBI database under BioProject PRJNA283303 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA283303).

Assembly and phylogenetic analyses

The NextGENE V2.17 (SoftGenetics, State College, PA, USA) software package was used to remove low-quality Illumina data. The quality-filtered data was assembled de novo using both the Maximum Overlap assembler in NextGENe and CAP3 (Huang and Madan, 1999), and assembly errors were manually screened utilizing NextGENe-Viewer. We utilized BLASTn (Altschul et al., 1990) combined with published reference sequences from Arabidopsis thaliana (NC_001284.2 and NC_000932.1) to parse plastid contigs larger than 300 bp. The Gyrostemon transcriptome library was assembled using Trinity (Grabherr et al., 2011). Target coding sequences (CDS) were identified using Arabidopsis thaliana organellar CDS (GenBank Y08501.2 and AP000423.1). We only utilized loci shared by all species. All alignments were generated using MUSCLE (Edgar, 2004), and columns were removed (i.e., across all taxa) if a base-pair position was missing in one species. We obtained a complete 44,926-bp data matrix derived from 72 different coding regions of the plastid genome. All phylogenetic analyses were conducted using the CIPRES V3.1 portal (Miller et al., 2010). The program RAxML version 7.3.1 (Stamatakis, 2006) was utilized to search for the optimal maximum likelihood (ML) tree using the GTR+GAMMA substitution model, combined with 1000 bootstrap replicates. Consensus trees were summarized with the program Consense (Felsenstein, 1989). The final alignment of assembled plastid sequences used in phylogenetic analyses available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.kc5b8. Glucosinolate profiles and Ks age distribution data for species was extracted from the data of Edger et al. (2015).

RESULTS

Previous phylogenetic analyses have provided important insights into the Brassicales phylogeny, including delineating all ~4700 species into 17 monophyletic families. However, these studies have been unable to resolve many familial relationships, specifically along the entire backbone of the tree. Here, we have employed a 72-marker approach to resolve the major relationships among all families. We obtained a strongly supported topology with 100% bootstrap support for the majority of nodes and two nodes with over 87% bootstrap support. Our phylogenetic estimates are a significant improvement in the overall phylogeny (e.g., core Brassicales, back-bone, earlier-diverging lineages), permitting us to more accurately pinpoint two ancient whole-genome duplications and the origin of two novel classes of glucosinolates.

Core Brassicales

The core Brassicales is the most speciose monophyletic clade in the order, comprising of the Brassicaceae (3973 species), Cleomaceae (300 species), Capparaceae (480 species), Emblingiaceae (1 species), Gyrostemonaceae (18 species), Pentadiplandraceae (1 species), Resedaceae (75 species), and Tovariaceae (2 species) (Hall et al., 2004). For the most up-to-date species numbers, see BrassiBase (Koch et al., 2012; Kiefer et al., 2014). Our analyses using markers from the plastid genome yielded a fully resolved phylogenetic estimate for the entire core Brassicales (Fig. 1) and is identical to the tree estimated from 1155 nuclear genes (Edger et al., 2015) with the sole exception of the inclusion of Tovariaceae. The previously reported Brassicaceae-Cleomaceae clade is still strongly supported (bootstrap 100%), which is sister to the Capparaceae (bootstrap 100%). Our phylogenetic estimates suggest that Tovariaceae is sister to the Brassicaceae-Cleomaceae-Capparaceae clade (bootstrap 89.5%). In addition, our analyses also recovered a Gyrostemonaceae-Resedaceae clade (bootstrap 100%), with Pentadiplandraceae as the immediate sister lineage (bootstrap 100%). Last, Emblingia calceoliflora E.Muell., the sole species in the family Emblingiaceae, is strongly supported as being sister to the remainder of the core Brassicales (bootstrap 100%). Thus, we present the first robust and well-resolved phylogeny that includes all families in the core Brassicales.

Earlier-diverging lineages

Our estimates for the relationships among the nine families distributed across the earliest-diverging lineages are also congruent with
our previous phylotranscriptomic study (Edger et al., 2015) with the exception of the inclusion of two new families (Salvadoraceae and Setchellanthaceae). These nine families, with 166 species (~3.5% Brassicales) that are distributed across six major lineages, includes Koeberliniaceae (1 species), Bataceae (2 species), Salvadoraceae (11 species), Setchellanthaceae (1 species), Limnanthaceae (8 species), Caricaceae (35 species), Moringaceae (12 species), Tropaeolaceae (95 species), and Akaniaaceae (2 species). The monotypic Koeberliniaceae, which consists of Koeberlinia spinosa Zucc., is the lineage sister to the core Brassicales (bootstrap 100%) with the Bataceae-Salvadoraceae clade sister to that clade (bootstrap 100%). Limnanthaceae (bootstrap 87.4%) is sister to the aforementioned clades with Setchellanthaceae (bootstrap 100%) sister to this clade in turn. Caricaceae-Moringaceae (bootstrap 100%) are sister to all Brassicales except for the Tropaeolaceae-Akaniaaceae clade (100%), which represents the earliest-diverging lineage in the order (100%). These relationships are consistent with previously published phylogenies, including the Bataceae-Salvadoraceae clade (Rone De Craene and Haston, 2006) and placement of Setchellanthaceae (Cardinal-McTeague et al., 2016).

**DISCUSSION**

Our phylogenetic analysis of coding regions from the plastid genome obtained fully resolved and strongly supported estimates for all family-level relationships across the Brassicales. However, this analysis reflects only the maternal history of the order due to the inheritance of organellar genomes. Importantly, these estimates are congruent with the topology reported in our previous phylotranscriptomic study (Edger et al., 2015) with the exception of the addition of the three new families. This consistency suggests that future studies could utilize plastid genomes as markers for resolving relationships within some notoriously difficult clades across Brassicales, as was previously demonstrated within Brassicaceae (Hohmann et al., 2015; Mandáková et al., 2017). The sequencing and assembly of plastid genomes has remained an easier and more cost-effective approach compared to phylotranscriptomics (Wysocki et al., 2015; Mandáková et al., 2015). This approach compared to phylotranscriptomics (Wysocki et al., 2015; Mandáková et al., 2015) with an independent polyploid event (Barker et al., 2009). We were also able to phylogenetically restrict the At-β event to one of two nodes. The At-β event occurred either after the divergence of the Caricaceae-Moringaceae clade or after the divergence of Setchellanthaceae (Fig. 1). We previously determined that the At-β event was not shared by Akaniaaceae, Caricaceae, Moringaceae, and Tropaeolaceae, but present in the evolutionary history of all other surveyed families (Edger et al., 2015). Based on the phylogenetic placement of Salvadoraceae and Tovariaceae in this study, we are able to determine that these families must also share the At-β event. However, we are unable to determine whether Setchellanthaceae shares the At-β event. Thus, we are able to restrict the phylogenetic placement of this ancient whole-genome duplication to the entire order with the exception of a single species (Setchellanthus caeruleus; Setchellanthaceae). The exact phylogenetic placement of the At-β could be determined from future analyses of either transcriptome or a reference genome of Setchellanthus caeruleus.

**Phylogenetic distribution of novel glucosinolate compounds**

Methionine-derived and tryptophan-derived (indolic) glucosinolates, which are unique to the order Brassicales, are a set of novel chemical defenses (Rodman et al., 1998; Kliebenstein et al., 2002; Windsor et al., 2005; Bekaert et al., 2012). It is these compounds that give mustard and horseradish that distinctive taste and unique aroma. Our previous studies showed that these novel classes of glucosinolates are synthesized from new amino acid substrates via divergently evolved, duplicated pathways that arose due to the At-a and At-β whole-genome duplications (Hofberger et al., 2013; Edger et al., 2015). For example, in Arabidopsis thaliana, two paralogous groups of Myb transcription factors regulate different glucosinolate pathways, with Myb28/29/76 regulating the biosynthesis of Met-derived glucosinolates and Myb34/51/122 regulating the biosynthesis of indolyl glucosinolates. In Carica papaya (Caricaceae), which does not share the At-β event in its evolutionary history, the ancestral glucosinolate pathway is regulated by a single Myb transcription factor gene (Schranz et al., 2011; Bekaert et al., 2012).

A reanalysis of the phylogenetic distribution of these novel glucosinolate classes, using the new family-level phylogeny, placed the origin of indolic glucosinolates, methionine-derived glucosinolates, and novel structural variants at nodes consistent with our previous analysis (Edger et al., 2015) (Fig. 1). In our previous work, we also constructed and compared glucosinolate pathways for the majority of families, revealing that Limnanthaceae did not have the genes required for indolic glucosinolate biosynthesis. This analysis further supports the placement for indolic glucosinolates shown in Fig. 1. A reference genome for any Limnanthaceae (e.g., Limnanthes douglasii R.Br.), combined with additional targeted glucosinolate profiling, would be tremendously helpful to verify these findings. Furthermore, we are confident in the placement for the origin of methionine-derived glucosinolates since these are unique to the Brassicaceae-Cleomeae-Capparaceae clade. Similarly, novel-structural elaborations unique to the core Brassicaceae are synthesized from neo-functionalized duplicated genes retained from the At-a event (Kliebenstein, 2008; Hofberger et al., 2013).

The origin of these novel glucosinolates are also associated with significant shifts in net diversification rates (Fig. 1) (Edger et al., 2015; Cardinal-McTeague et al., 2016), supporting predicted evolutionary arms-race dynamics with certain butterflies (Ehrlich and Raven, 1964). However, a shift in net diversification rates is not observed for the origin of indolic glucosinolates but this is likely due to the age of this event (~85–92.2 million years ago; Edger et al.,
2015) pre-dating the Cretaceous-Tertiary (KT) mass extinction event (Fawcett et al., 2009). Regarding the diversification rate shift detected at the base of the Reseaceae-Gyrostemonmon clade (Cardinal-McTeague et al., 2016), the causes of this radiation remain unknown. Last, our findings are consistent with the predictions of the whole-genome duplication (WGD) radiation-lag time model (Schranz et al., 2012) in which the whole-genome duplication and the evolution of a novel adaptive trait (e.g., glucosinolates) derived from neofunctionalized duplicate genes should be separated by long time periods (e.g., millions of years). Neofunctionalization, developing novel gene functions from background mutations, requires long time periods especially without perturbing the established network stoichiometry (Birchler and Veitia, 2007; Edger and Pires, 2009). In conclusion, we anticipate that this new Brassicales phylogeny will aid future evolutionary studies across the order and guide the selection of future candidates for genome projects.

ACKNOWLEDGEMENTS

We thank J. Birchler, G. Conant, K. Newton, and J. Taylor for valuable comments on the manuscript; R. Schnabel and K. Wells for computational assistance and the National Science Foundation (DEB 1146603 and NPGI 1202793) and USDA-HATCH (1009804) for support. We also thank the editors and reviewers for their helpful comments during the review of this manuscript.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

LITERATURE CITED


