Supplemental Materials:

Suppmentary Text S1.1-S1.6
Figures S1.1 to S1.3
Tables S1.1 to S1.7
Supplementary Text S2.1-S2.4
Figures S2.1 to S2.5
Tables S2.1 to S2.3
Supplementary Text S3.1-S3.2
Figures S3.1 to S3.3
Tables S3.1
Suppmentary Text S4.1-S4.3
Figures S4.1 to S4.7
Tables S4.1 to S4.4
Supplementary Text S5.1-S5.9
Figures S5.1 to S5.4
Tables S5.1 to S5.3
Supplementary Text S6.1
Figure S6.1
Tables S6.1

Other Supplemenary Materials for this manuscript includes the following:

Dataset S1 for Table S1.3, Dataset S2 for Tables S4.1, and Dataset S3 for Tables S6.1.

Supplementary Text S1

Estimating phylogenetic relationships across Brassicales families

The order Brassicales, which contains 4,765 species or ~2.2% of eudicot diversity (30), is a monophyletic group consisting of 17 families including the mustards (Brassicaceae) (31-35). Previous molecular phylogenetic studies were able to obtain robust estimates for some relationships. For example, early studies identified a core group of eight families, known as the core Brassicales, comprised of: Brassicaceae, Cleomaceae, Capparaceae, Emblingiaceae, Gyrostemonaceae, Pentadiplandraceae, Resedaceae, and Tovariaceae (31, 32, 35). Within this core group, a Brassicaceae-Cleomaceae clade is strongly supported as sister to the Capparaceae (32, 33, 36). Another stongly supported clade includes Borthwickiaceae, Gyrostemonaceae, Resedaceae, and two unplaced genera (*Forchhammeria* and *Stixis)* (32, 37). However, all other relationships within the core Brassicales are still either unresolved or lack statistical support (bootstrap less than 70%). In addition, a few unplaced genera are included in the core Brassicales, of which one was recently elevated to familial rank (Borthwickiaceae) (38).

There are also a number of supported familial relationships outside the core Brassicales. These include the Koeberliniaceae-Bataceae clade, the Tropaeolaceae-Akaniaceae clade, and the Caricaceae-Moringaceae clade (32, 39, 40). However, an alternate relationships for Bataceae has also been inferred, namely a Bataceae-Salvadoraceae clade (bootstrap 99%)(37). A clade that includes the core Brassicales, Koeberlinaceae, Bataceae, Salvadoraceae, Setchellanthaceae, and Limnanthaceae is strongly supported (32), with Limnanthaceae and Setchellanthaceae being early diverging lineages within it (31, 35, 37, 41, 42). The Caricaceae-Moringaceae and Akaniaceae-Tropaeolaceae clades are supported as the earliest diverging lineages (i.e. sister to all other families) (37), but the relationships among these remains unknown. In sum, the overwhelming majority of the nodes along the backbone of the phylogeny are still either unresolved or poorly supported. These previous studies estimated familial relationships using up to four phylogenetic markers, generally from the plastid, although Ronse De Craene and Haston (2006) also used a single nuclear maker (18S rRNA). Here, we report a Brassicales phylogeny estimated using nuclear markers obtained from transcriptomes and publicy available genomes of species distributed across 14 Brassicales families.

S1.1 Transcriptomes: RNA Isolation, Library Construction, and Assembly

Total RNA was extracted and pooled from all available young tissues (see Table S1.1) using the PureLink RNA Mini Kit (Invitrogen, Carlsbad, CA, USA). Next selected samples were normalized with the Evrogen TRIMMER and MINT kits (Evrogen, Moscow, Russia), converted into Illumina libraries using either the NEB prep E600L kit (New England Biolabs, Ipswich, MA, USA) or the TruSeq RNA kit (Illumina, San Diego, CA, USA), and sequenced paired-end on either the Illumina Genome Analyzer II or HiSeq-2000 instrument at the University of Missouri Sequencing Core. These data have been deposited into the NCBI Sequence Read Archive (SRA). Sequence data for *Tropaeolum majus* was obtained from the NCBI -SRA (SRX108504 – SRX108510). Illumina data was quality filtered and trimmed using NextGene v2.17 (SoftGenetics, State College, PA, USA), further processed using custom scripts to remove quality trimmed reads shorter than 40bp and any resulting unpaired “orphan” single reads, and assembled *de novo* with Trinity (43) (See Table S1.2 for assembly statistics).

Table S1.1: RNA extracted and Illumina reads generated

For each of the listed species, total RNA was extracted and pooled from all available young tissues (indicated in light grey) including seedlings, roots, leaves, stems, flowers (buds and mature floral organs), fruits, and other (e.g. tissues, developmental stages, and treatments). The total number of Illumina reads generated per species is shown in the last column.

Family	Species	Seedlings	Roots	Leaves	Stems	Flowers, Fruits	other	# reads
Brassicaceae	*Aethionema arabicum* ES1020							58,793,342
Brassicaceae	*Aethionema arabicum* 84-56-2							82,720,684
Akaniaceae	*Akania bidwillii*							114,205,326
Bataceae	*Batis maritima*							115,905,390
Capparaceae	*Capparis spinosa*							76,667,886
Cleomaceae	*Cleome violacea*							257,273,714
Emblingiaceae	*Emblingia calceoliflora*							65,578,620
Gyrostemonaceae	*Gyrostemon ramulosus*							94,656,836
Koeberliniaceae	*Koeberlina spinosa*							117,012,624
Limnanthaceae	*Limnanthes douglasii*							125,299,410
Moringaceae	*Moringa oleifera*							91,777,978
Pentadiplandraceae	*Pentadiplandra brazzeana*							72,317,888
Resedaceae	*Reseda odorata*							64,730,750

Table S1.2: Transcriptome and Genome Data and NCBI Accession Numbers

The top portion of the table summarizes the species with analyzed transcriptomes, Trinity *de novo* assembly statistics (number of contigs and average contig size), and NCBI Short Read Archive BioProject numbers (http://www.ncbi.nlm.nih.gov/sra). The bottom portion of the table summarizes the species that were analyzed with sequenced genomes, listing the number of annotated protein coding genes in those genomes and NCBI Genome Identification numbers (http://www.ncbi.nlm.nih.gov/genome).

Family	Species	No. Assembled Contigs	Average Contig Size	Data type	NCBI SRA BioProject
Brassicaceae	*Aethionema arabicum* ES1020	19,037	801	Transcriptome - RNAseq	PRJNA283303
Brassicaceae	*Aethionema arabicum* 84-56-2	86,597	479	Transcriptome - RNAseq	PRJNA283303
Cleomaceae	*Cleome violacea*	29,831	1,021	Transcriptome - RNAseq	PRJNA283303
Capparaceae	*Capparis spinosa*	72,912	1,138	Transcriptome - RNAseq	PRJNA283303
Pentadiplandraceae	*Pentadiplandra brazzeana*	39,052	1,008	Transcriptome - RNAseq	PRJNA283303
Gyrostemonaceae	*Gyrostemon ramulosus*	11,441	347	Transcriptome - RNAseq	PRJNA283303
Resedaceae	*Reseda odorata*	52,511	736	Transcriptome - RNAseq	PRJNA283303
Emblingiaceae	*Emblingia calceoliflora*	18,742	370	Transcriptome - RNAseq	PRJNA283303
Koeberliniaceae	*Koeberlina spinosa*	13,361	534	Transcriptome - RNAseq	PRJNA283303
Bataceae	*Batis maritima*	41,235	1,240	Transcriptome - RNAseq	PRJNA283303
Limnanthaceae	*Limnanthes douglasii*	49,068	744	Transcriptome - RNAseq	PRJNA283303
Moringaceae	*Moringa oleifera*	55,484	1,359	Transcriptome - RNAseq	PRJNA283303
Tropaeolaceae	*Tropaeolum majus*	9,577	665	Transcriptome - RNAseq	PRJNA80079
Akaniaceae	*Akania bidwillii*	98,343	677	Transcriptome - RNAseq	PRJNA283303
AVERAGE		42,657	794		

Family (Order)	Species	No. Protein Coding Genes	Citation	Data Type	NCBI Genome ID
Brassicaceae (Brassicales)	*Arabidopsis thaliana*	27,025	AGI et al., 2000 (Nature)	Genome	Genome ID 4
Brassicaceae (Brassicales)	*Arabidopsis lyrata*	32,534	Hu et al., 2011 (Nature Genetics)	Genome	Genome ID 493
Brassicaceae (Brassicales)	*Brassica rapa*	41,174	Wang et al., 2011 (Nature Genetics)	Genome	Genome ID 229
Brassicaceae (Brassicales)	*Thellungiella parvula*	28,901	Dassanayake et al., 2011 (Nature Genetics)	Genome	Genome ID 3585
Brassicaceae (Brassicales)	*Aethionema arabicum*	48,138	Haudry et al., 2013 (Nature Genetics)	Genome	Genome ID 17729
Caricaceae (Brassicales)	*Carica papaya*	21,784	Ming et al., 2008 (Nature)	Genome	Genome ID 513
Malvaceae (Malvales)	*Theobroma cacao*	28,993	Argout et al., 2011 (Nature Genetics)	Genome	Genome ID 572
Salicaceae (Malpighiales)	*Populus trichocarpa*	55,532	Tuskan et al., 2006 (Science)	Genome	Genome ID 98
Vitaceae (Vitales)	*Vitis vinifera*	28,268	Jaillon et al., 2007 (Nature)	Genome	Genome ID 401
AVERAGE		34,705			

S1.2 Identification Of Ortholog Groups Using OrthoMCL From Transcriptomes

An objective gene family classification built with gene models from 22 sequenced land plant plant genomes using OrthoMCL (44) was used to identify a large set of putatively nuclear single copy genes (nSCG) and populate a phylogenetic super matrix with gene sequences from 14 diverse Brassicales transcriptomes and two additional Brassicaceae genomes (*Brassica rapa (45)* and *Arabidopsis lyrata* (46)). The 22 sequenced land plant genomes include: *Solanum tuberosum* (47), *Solanum lycopersicum* (48), *Mimulus guttatus* (49), *Arabidopsis thaliana* (50), *Thellungiella (Eutrema) parvula* (51), *Carica papaya* (52), *Theobroma cacao* (53), *Populus trichocarpa* (54), *Fragaria vesca (55), Medicago trunculata* (56), *Glycine max* (57), *Vitis vinifera* (58), *Nelumbo nucifera* (59), *Aquilegia coerulea* (60), *Sorghum bicolor* (61), *Brachypodium distachyon* (62), *Oryza sativa* (63), *Musa acuminata* (64), *Phoenix dactylifera* (65), *Amborella trichopoda* (66), *Selaginella moellendorffii* (67), and *Physcomitrella patens* (68).

New sequences were sorted into the gene family classification using Hidden Markov Models (HMMs) built from orthogroup alignments from the 22 scaffold genomes. Using this classification, 1155 orthogroups containing a single gene in *Vitis vinifera*, *Populus trichocarpa*, *Theobroma cacao*, *Carica papaya*, *Arabidopsis lyrata*, *Arabidopsis thaliana*, and *Eutrema parvula* were selected for downstream analysis (Supplemental Table S1.3). When more than one unigene in a transcriptome data set was identified as belonging to a nSCG orthogroup, we attempted to generate a consensus scaffold sequence by aligning unigene sequences to a reference protein in *Carica papaya*. If unigene sequences were less than 5% divergent from each other in overlapping regions, a consensus sequence was generated using IUPAC ambiguity codes and filling undetermined sequence regions with Ns to produce a single scaffold sequence for each nSCG. If greater than 5% divergence among unigenes was observed for a taxon (3367 cases), this gene was identified as a scaffolding failure and omitted for that sample (Supplemental Table S1.3). A total of 3899 nSCG were not detected in the transcriptome data and are coded as missing data (Supplemental Table S1.3).

Individual nSCG orthogroups were aligned with MAFFT (69) and concatenated into a supermatrix containing all 1155 nSCG. The final alignment contained nearly 2.5 million nucleotide positions and 18 Brassicales taxa representing all families except Setchellanthaceae, Salvadoraceae, and Borthwickiaceae. Sites containing fewer than 30, 60, 80, 85, and 90 percent of the taxa were further trimmed from the alignment to assess the effect of missing data on phylogeny estimation (Supplemental Table S1.3).

Table S1.3: [Excel table] Summary of nuclear single copy gene coverage in the Brassicales transcriptome data

1155 orthogroups have a single gene member in *Vitis vinifera*, *Populus trichocarpa*, *Theobroma cacao*, *Carica papaya*, *Arabidopsis lyrata*, *Arabidopsis thaliana*, and *Eutrema parvula*. Coverage of the reference *Carica* gene in each taxon is given for each orthogroup. Genes excluded from the analysis because of conflict/divergence among unigenes is indicated by 'SCF' and genes not detected in the transcriptome are indicated as 'NA'. Summary information and functional annotations are also provided.

Table S1.4: The number of sites (base pairs) in the full and trimmed datasets filtered based on percent species completeness (30%, 60%, 80%, 85%, and 90%)

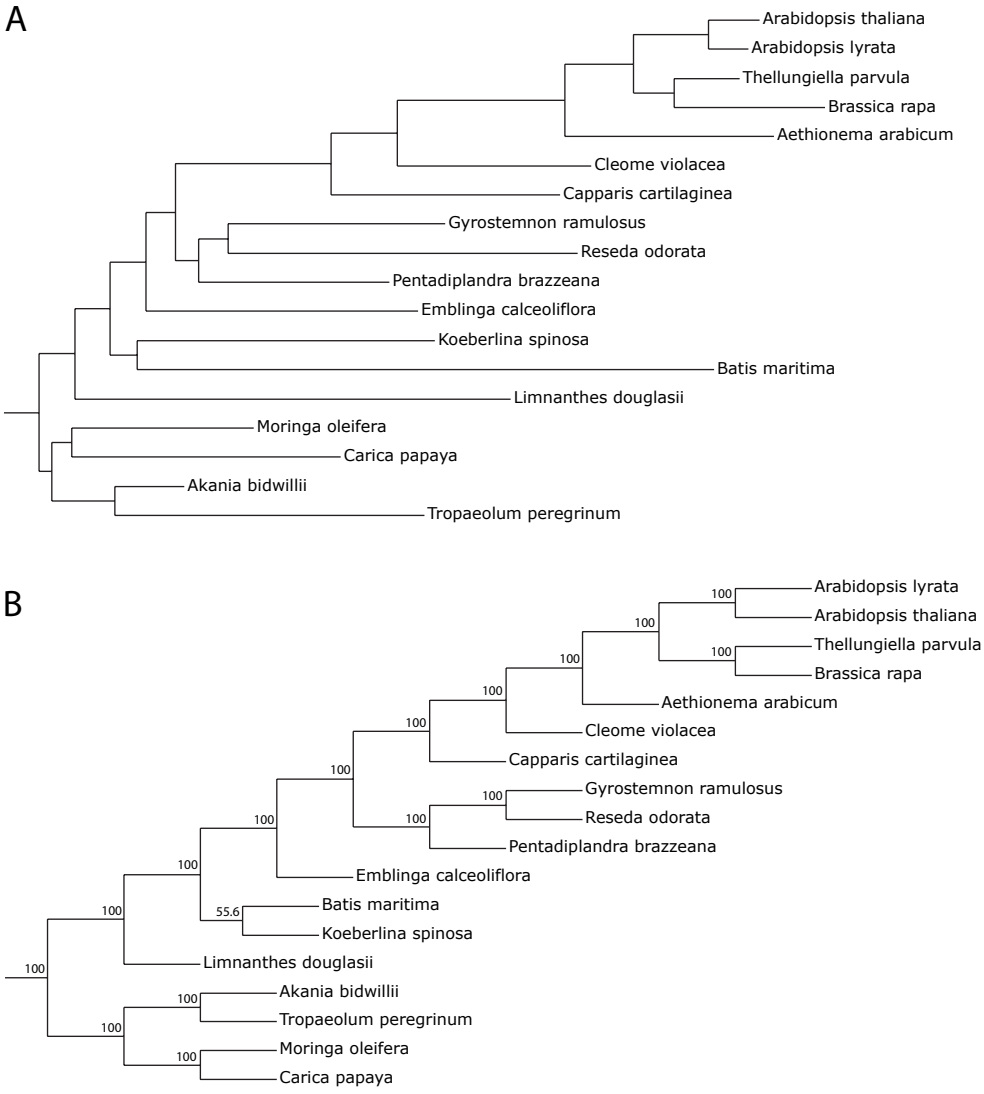
Alignment	Full Dataset	30% Species Completeness	60% Species Completeness	80% Species Completeness	85% Species Completeness	90% Species Completeness
Number of Sites	2,497,662	1,574,131	960,554	174,163	74,579	22,573

S1.3 Phylogenetic Analyses To Estimate Relationships Among Brassicales Families

The 85% species complete data matrix, consisting of 74,579 characters, was used to estimate the relationships among the 14 Brassicales families. All phylogenetic analyses were conducted through the CIPRES V3.1 portal (70). The program RAxML version 7.3.1 (71) was employed to search for the optimal maximum likelihood tree with the GTR+GAMMA substitution model. Node support was estimated with 1,000 bootstrap replicates. Maximum parsimony (MP) analyses were conducted using PAUP (72), and Bayesian analyses using BEAST (73). Consensus trees were summarized with Consense (70). All alignments and trees have been deposited in TreeBASE (http://www.treebase.org/).

The estimates for phylogenetic relationships across Brassicales are nearly identical from the ML, MP, and Bayesian analyses, with the exception for Bataceae and Koeberlinaceae. These two families are either sister families forming a strongly supported clade as observed in the MP tree (Figure 1C, 100% bootstrap support) or separate lineages as observed in the Bayesian tree (Figure 1D). These two trees were recovered in the ML analysis at nearly equal frequencies: 56% MP and 44% Bayesian. The phylogenetic estimates for all remaining relationships among Brassicales families are congruent, and with 100% bootstrap support for all other nodes. The slight difference in these two phylogenetic estimates does not impact our analysis on glucosinolate diversity, diversification rates, or any other subsequent analysis performed here since these two families consist only of three species total (Koeberlinaceae, 1 species; Bataceae, 2 species). The relative phylogenetic placement of these two families for both recovered trees is congruent with previously published studies. The Bayesian tree was used for all subsequent analyses.

Relationships were estimated using the 85% complete data-matrix derived from 1155 shared single copy nuclear genes (Supplemental Table S1.3), aligned using MUSCLE (74), and with maximum likelihood (RaxML (71)), maximum parsimony (PAUP (72)) and Bayesian (BEAST (73)) analyses. Panel A depicts the best-scoring maximum likelihood tree with branch length estimates, and Panel B depicts the majority-rule consensus tree of 1000 bootstrap replicates with branch lengths proportional to bootstrap support (support values shown above nodes). Bootstrap support for all nodes are 100%, except the *Batis maritima* (Bataceae) and *Koeberlina spinosa* (Koeberlinaceae) clade with only 55.6% of tree topologies supporting this relationship. Panel C depicts the most parsimonious tree with branch lengths equal to bootstrap support, based on consensus of 1000 bootstrap replicates. These analyses yielded a fully supported tree (i.e. 100% support for all nodes) with an identical topology to the best-scoring maximum likelihood tree. Panel D depicts the Bayesian tree with the best likelihood score, which reflects the only alternate supported topology in Panel B.


A
Arabidopsis thaliana
Arabidopsis lyrata
Thellungiella parvula
Brassica rapa
Aethionema arabicum
Cleome violacea
Capparis cartilaginea
Gyrostemnon ramulosus
Reseda odorata
Pentadiplandra brazzeana
Emblinga calceoliflora
Koeberlina spinosa
Batis maritima
Limnanthes douglasii
Moringa oleifera
Carica papaya
Akania bidwillii
Tropaeolum peregrinum
B
100
100
100
100
100
100
100
100
100
100
100
100
100
55.6
100
100
100
100
100
Arabidopsis lyrata
Arabidopsis thaliana
Thellungiella parvula
Brassica rapa
Aethionema arabicum
Cleome violacea
Capparis cartilaginea
Gyrostemnon ramulosus
Reseda odorata
Pentadiplandra brazzeana
Emblinga calceoliflora
Batis maritima
Koeberlina spinosa
Limnanthes douglasii
Akania bidwillii
Tropaeolum peregrinum
Moringa oleifera
Carica papaya

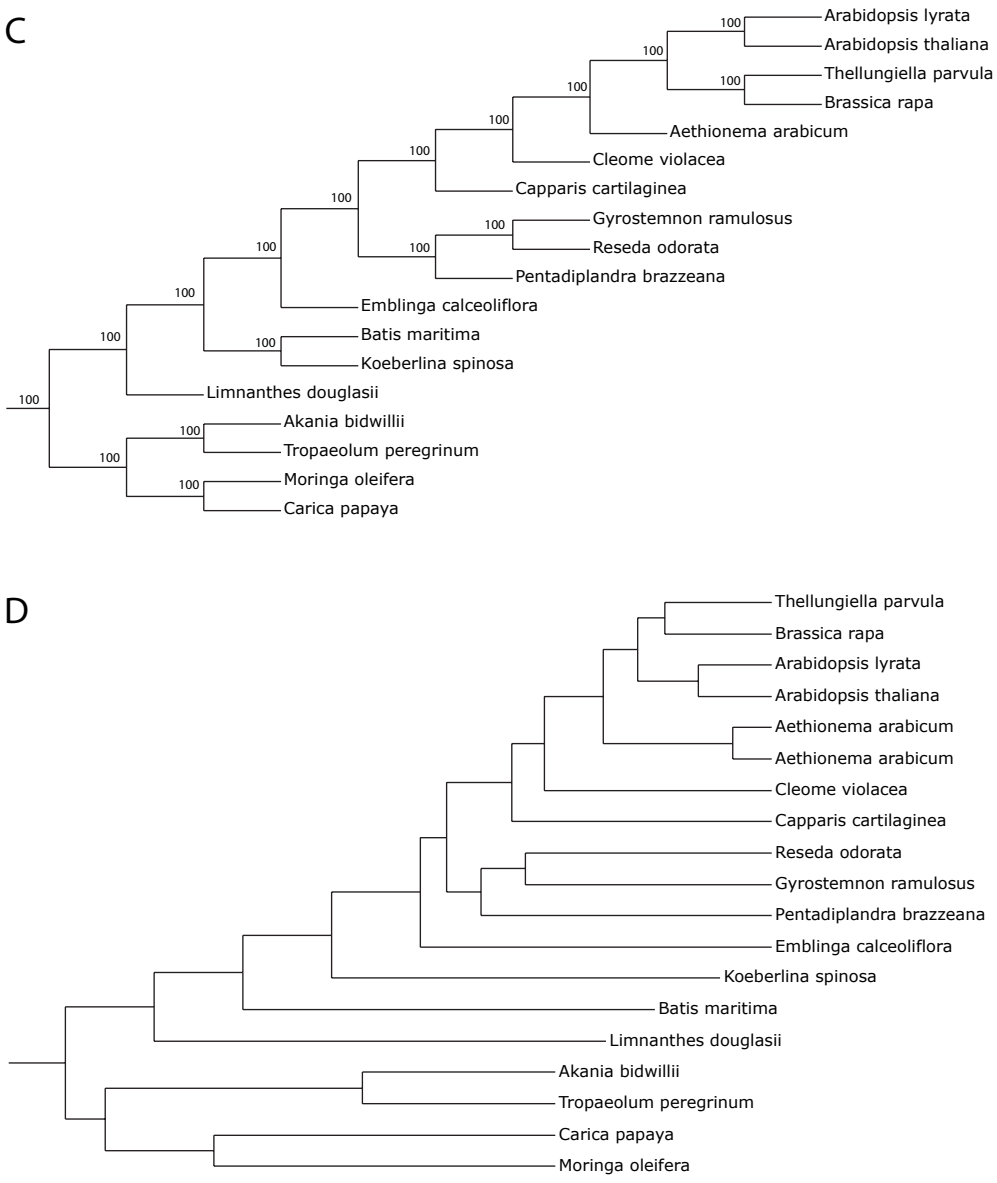


Figure S1.1: Relationships among the Brassicales families

Phylogenetic relationships of eighteen species distributed across fourteen Brassicales families.

S1.4 Divergence time estimates using BEAST, nuclear data, and fossil calibrations

Given our sampling, only two fossils were used for calibration in the Brassicales, with their age priors modeled as normal distributions as this takes into account their bi-directional uncertainty (75). First, following previous literature (30, 76), the Brassicales root was modeled with mean 89.5 my, and we used a standard deviation that was 25% of this value to reflect the uncertainly in the estimate (std. dev. 22.3 my). Second, the *Tropeaolum+Akania* node was modeled with mean 61 my (standard deviation 15.25 my) based upon the age of the oldest known *Akania* sp. fossil (77), as used previously in Brassicales analyses (78).

For the divergence estimates, only the 1st and 2nd positions of each codon were used, resulting in 116,109 bp of data of the 80% species-complete data matrix (Table S1.4). The HKY + G model of substitution was used with four discrete categories, and the uncorrelated lognormal distribution was used to model branch lengths with a Yule model speciation process. The remaining prior distributions were left to their defaults. Analyses were run for 100 million generations sampling every 1000 generations, with the number of independent runs depending on the analysis (minimally two independent runs). The convergence of the likelihood traces of the independent runs was assessed with Tracer v1.5 and the ESS values were verified to be above 100 for all parameters.

The newly resolved, robust phylogenetic framework for the Brassicales (Figure S1.2) allows us for the first time to accurately estimate variation in diversification rates at the family level across Brassicales. In addition, we assessed the phylogenetic diversity of glucosinolate compounds across the order to identify the origin of glucosinolate classes that are synthesized from novel substrates. Such analyses were previously impractical due to the lack of a resolved Brassicales phylogeny.

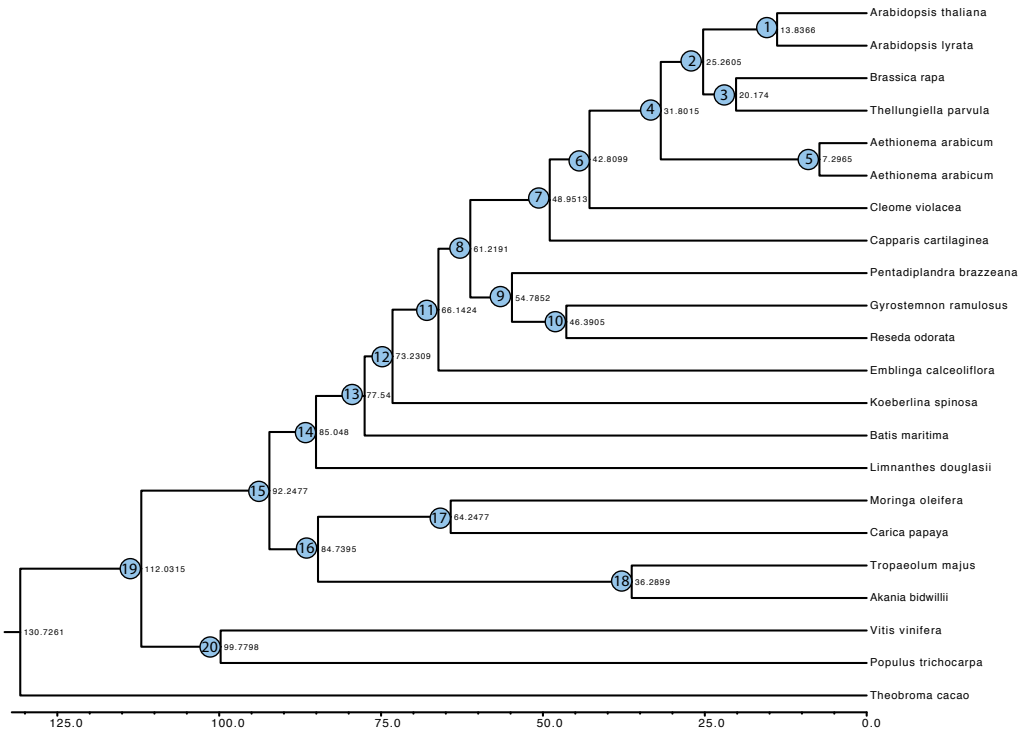


Figure S1.2: Divergence-date estimates based on 1155 single copy nuclear genes in BEAST analyses

Median date estimates (in million years) are shown adjacent to each node and along bottom bar, and 95% highest posterior density intervals for each estimate date are provided in Supplemental Table S1.5.

Table S1.5: Divergence times, both median dates and 95% confidence intervals (in millions of years), for all nodes across the Brassicales

Dates were estimated using BEAST analyses (73). Refer to Supplemental Figure S1.2 for identifying nodes (numbers within blue circles).

Node	Divergence	Median Estimate	95% Confidence Interval
1	*Arabidopsis thaliana - Arabidopsis lyrata*	13.8366	6.58 - 20.85
2	*Arabidopsis thaliana - Brassica rapa*	25.2605	13.5 - 36.76
3	*Brassica rapa - Thellungiella parvula*	20.174	10.93 - 30.03
4	*Arabidopsis thaliana - Aethionema arabicum*	31.8015	16.81 - 45.89
5	*Aethionema arabicum - Aethionema arabicum*	7.2965	2.27 - 13.17
6	*Arabidopsis thaliana - Cleome violacea*	42.8099	23.26 - 62.13
7	*Arabidopsis thaliana - Capparis cartilaginea*	48.9513	26.64 - 71.5
8	*Arabidopsis thaliana - Pentadiplandra brazzeana*	61.2191	33.37 - 89.22
9	*Pentadiplandra brazzeana - Reseada odorata*	54.7852	29.46 - 80.69
10	*Reseda odorata - Gyrostemnon ramulosus*	46.3905	22.56 - 67.16
11	*Arabidopsis thaliana - Emblinga calceoliflora*	66.1424	34.27 - 95.31
12	*Arabidopsis thaliana - Koeberlina spinosa*	73.2309	40.05 - 106.92
13	*Arabidopsis thaliana - Batis maritima*	77.5401	41.95 - 111.88
14	*Arabidopsis thaliana - Limnanthes douglasii*	85.048	45.99 - 123.25
15	*Arabidopsis thaliana - Carica papaya*	92.2477	49.57 - 133.08
16	*Carica papaya - Akania bidwillii*	84.7395	45.56 - 122.6
17	*Carica papaya - Moringa oleifera*	64.2477	33.94 - 92.51
18	*Akania bidwillii - Tropaeolum majus*	36.2899	19.1 - 54.37
19	*Arabidopsis thaliana - Populus trichocarpa*	112.0315	59.66 - 162.99
20	*Populus trichocarpa - Vitis vinifera*	99.7798	54.43 - 152.18

S1.5 Detecting diversification rate variation across Brassicales

A. Relative Cladogenesis Test

To detect significant shifts in diversification rates across Brassicales phylogeny, we used species richness data from the Angiosperm Phylogeny Website V12 (www.mobot.org/MOBOT/research/APweb/) (Figure 1) and analyzed using the R package GEIGER- Relative Cladogenesis Test (79). Two significant shifts in diversification rates were detected across Brassicales. The oldest radiation, dated near the Cretaceous-Paleogene (K-Pg) boundary (formerly referred to as the Cretaceous-Tertiary or K-T extinction event), occurred at node 8 in Figure S1.2 (p-value = 0.0019349845; Bonferroni corrected p-value = 0.038699690). The other signficant radiation occurred within the Brassicaceae at node 2 in Figure S1.2 (p-value = 0.0003430522; Bonferroni corrected p-value = 0.006861044).

B. Modeling evolutionary diversification using stepwise AIC (MEDUSA) analysis

Changes in diversification rates can also be investigated using ultrametric tree data without any *a priori* selection of specific nodes. To do this, we used a method called MEDUSA (modeling evolutionary diversification using stepwise AIC) (80). In this method, a constant parameter model of diversification is fit to the data, and then birth and death rates are allowed to shift at each node. Each node is then tested, first singly then in larger groupings, with models having significant increases in fit, evaluated by an increase in the AIC value, selected and further compared. Terminal tips of the tree represent genera, and for each the number of species in that genera was used.

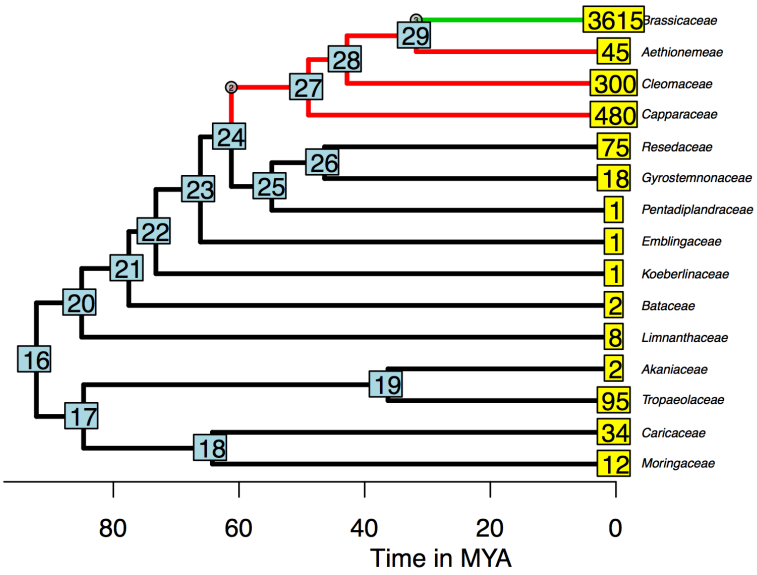


Figure S1.3 Phylogenetic tree with color shading for the three rate partitions identified using MEDUSA

Optimal MEDUSA birth-death model for tree with 15 tips representing 4689 taxa. Numbers in yellow are the number of species in each family, numbers in blue indicate those nodes on the tree. Node 27 and its descendants are colored red in the tree above, indicating one rate shift, while the second rate shift is green shaded (leading to node 13).

Resulting parameter estimates are listed below, with low and high values indicating the bounds of the 95% confidence intervals. The 95% confidence intervals on parameter values shown below is calculated from profile likelihoods. The appropriate AICC-threshold for a tree of 28 tips is 1.091845.

Table S1.6 Optimal MEDUSA birth-death model with parameter values show

Model	Shift.Node	Ln.Lik.part	R	epsilon	r.low	r.high	eps.low	eps.high
1	16	−88.647	0.029	0.881	0.021	0.040	0.792	1
2	27	−28.633	0.113	0.452	0.092	0.145	0.000	0.864
3	13	−9.1927	0.207	0.802	0.160	0.297	0.127	1

Thus, the diversification rate of the Brassicales had two significant increases. The first was at the origin of the clade Capparaceae+Cleomaceae+Brassicaceae, and the second within the Brassicaceae.

S1.6 Glucosinolates: Diversity and Novelty across Brassicales

We used the known glucosinolate chemical diversity across the Brassicales summarized by previous reviews (81, 82) to assess the phylogenetic distribution and origin of glucosinolate classes. The phylogenetic distribution of these compounds shows that Indolic glucosinolates (class I), which are the only class synthesized from the amino acid tryptophan, are unique to the most recent common ancestor of Bataceae-Brassicaceae (Table S1.5, Figure 1). Additionally, our phylogenetic analysis indicates that methionine-derived glucosinolates are unique to the Capparaceae-Cleomaceae-Brassicaceae clade. Finally, over half of all described glucosinolate compounds have only been identified in the core Brassicaceae and are not shared with the earliest diverging tribe Aethionemeae (81) (Figure 1). The origin of these new compounds can be explained by a near-doubling of the glucosinolate pathway in the Brassicaceae, with a significant over-retention of duplicates following the At-α event and a nearly 10 fold higher rate of tandem duplications than that observed genome-wide (83). These duplicate genes are involved in a number of processes during glucosinolate biosynthesis including core-structure formation and side-chain modifications (84, 85).

Table S1.7: Glucosinolate diversity classified based on chemical structure across Brassicales families

The presence of indolic and methionine (met-) derived glucosinolates across Brassicales families is summarized here, and the phylogenetic localization of these novel groups is shown in Figure 1. *Koeberlina spinosa*, the sole species in Koeberliniaceae, does not synthesize any known glucosinolate compounds.

Family	Indolic Glucosinolates	Met-derived Glucosinolates
Brassicaceae	Present	Present
Tribe Aethionemeae (Brassicaceae)	Present	Present
Cleomaceae	Present	Present
Capparidaceae	Present	Present
Pentadiplandraceae	Present	
Gyrostemonaceae	Present	
Resedaceae	Present	
Emblingiaceae		
Koeberliniaceae		
Bataceae	Present	
Limnanthaceae		
Moringaceae		
Caricaceae		
Tropaeolaceae		
Akaniaceae		

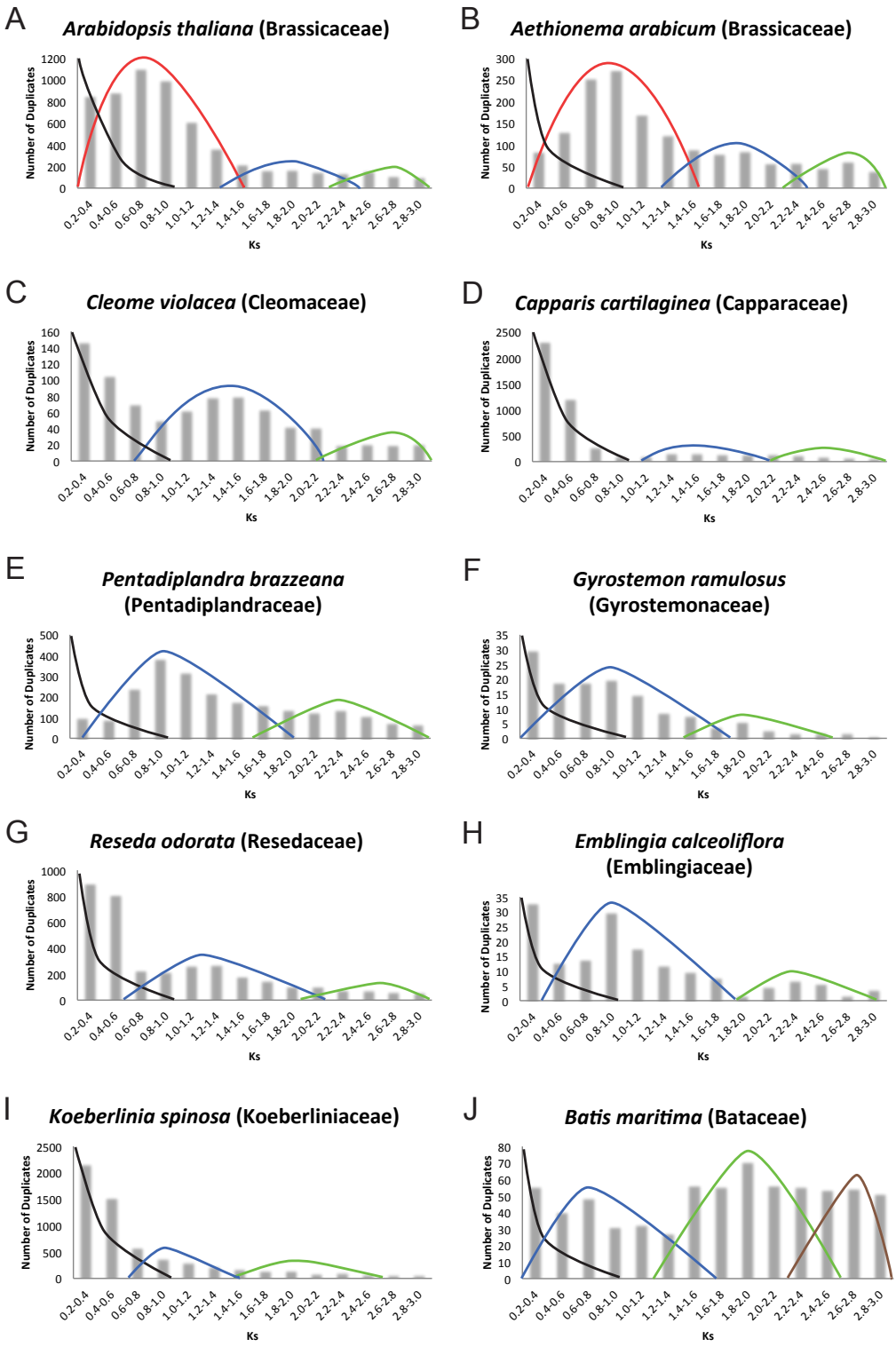
Supplementary Text S2

Phylogenetic localization of whole genome duplications (Brassicales)

Plant evolutionary history is rich with whole genome duplications (WGDs), including at least two ancient events shared by all angiosperms (86). In addition, genomic analyses of *Arabidopsis thaliana* (Brassicales) revealed the remnants of three later WGDs (87, 88). The most ancient of these events, termed At-γ, occurred at the origin of the eudicots (89), and the more recent At-β and At-α WGDs are phylogenetically restricted to the order Brassicales (52, 90). However, inferring the exact phylogenetic placement for these events has been difficult for two major reasons; a dearth of genomic resources from the families of the Brassicales, making the presence of the two events in particular lineages uncertain, and the lack of a robust phylogenetic framework with which to localize the events. Supplementary Text S1 summarizes our analyses to resolve the phylogenetic relationships of Brassicales families. Here, we report the phylogenetic localization of the At-β and At-α WGDs within the order Brassicales estimated using both transcriptome data (i.e. age distributions and phylogenetic analyses of gene duplicates) and comparing publicly available genomes via syntenic analyses of species distributed across 14 Brassicales families. Our results refute previous inferences for the phylogenetic location of the At-α WGD event based on the distribution of species-richness (91), and further restrict the localization of the event compared to more recent estimates (92). See Figure 1 for phylogenetic placement of both WGDs.

S2.1 Age Distributions to Detect Ancient Whole Genome Duplications

Ancient whole genome duplications were detected across the order Brassicales by calculating substitutions per synonymous sites (Ks) between all gene duplicates filtered from transcriptome and genome data (*Arabidopsis thaliana* and *Carica papaya*) and constructing Ks age distributions. The DupPipe pipeline (http://EvoPipes.net) was used to calculate Ks values between duplicate pairs (93); duplicate pairs were identified with BLAST analyses (94), aligned with GeneWise 2.2.2 and MUSCLE 3.6 (74, 95), Ks values calculated with the PAML package (96), and distributions statistically evaluated with a mixture model analysis using EMMIX (97). The Ks distributions for each of the species are shown in Figure S2.1 and select species comparisons to identify shared duplications in Figure S2.2. EMMIX results are provided in Tables S2.1 and S2.2.


A
Arabidopsis thaliana (Brassicaceae)
B
Aethionema arabicum (Brassicaceae)
C
Cleome violacea (Cleomaceae)
D
Capparis cartilaginea (Capparaceae)
E
Pentadiplandra brazzeana (Pentadiplandraceae)
F
Gyrostemon ramulosus (Gyrostemonaceae)
G
Reseda odorata (Resedaceae)
H
Emblingia calceoliflora (Emblingiaceae)
I
Koeberlinia spinosa (Koeberliniaceae)
J
Batis maritima (Bataceae)
Number of Duplicates
Ks

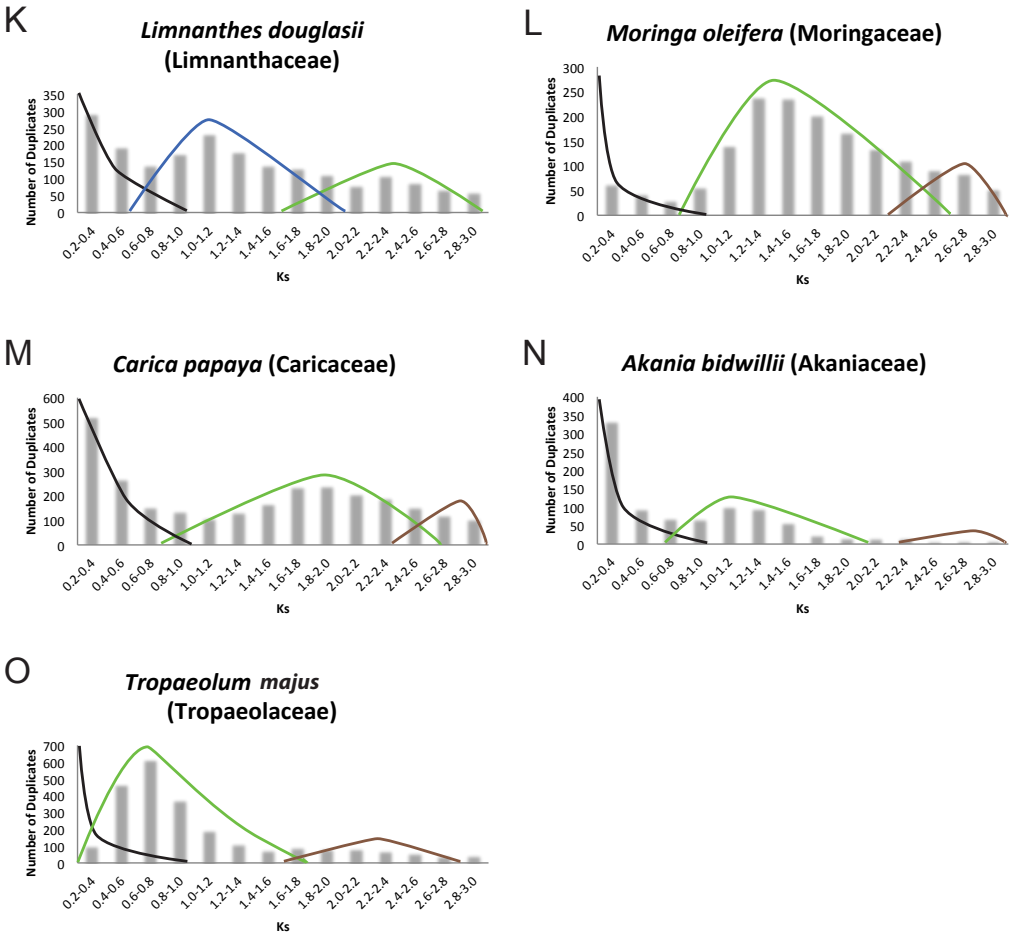


Figure S2.1: Whole Genome Duplications identified in Brassicales Transcriptomes

These are indicated beneath colored curves: At-α (red), At-β (blue), At-γ (green), and more ancient events (brown). Smaller-scale duplications (e.g. tandem duplicates), follows a power-law distribution, are beneath the black curves. These results show that the At-α event occurred at the origin of the family Brassicaceae, shared by both *Arabidopsis thaliana* (Panel A) and *Aethionema arabicum* (Panel B) but not detected in *Cleome violacea* (Panel C). All Brassicales families, except Moringaceae (Panel L), Caricaceae (Panel M), Akaniaceae (Panel N), and Tropaeolaceae (Panel O), share the At-β duplication. The At-γ event, shared by all eudicots, was detected in every Ks distribution. The more ancient events, detected in *Batis maritima* (Panel J) and all families not sharing the At-β event, may be the angiosperm-specific event (86). The placement of both At-α and At-β is shown in Figure 1.

Table S2.1: EMMIX mixture model results from the analyses of Ks distributions

These are shown in Figure S2.1 calculated from transcriptome and genome datasets using the DupPipe pipeline (93, 97). The total number of components and the estimated mean for each component is listed for smaller-scale duplicates (e.g. tandem duplicates) and three ancient whole genome duplications shared by *Arabidopsis thaliana*.

Family	Species	No. Components	Tandem duplicates	At-alpha	At-beta	At-gamma	At-delta
Brassicaceae	*Arabidopsis thaliana*	5	0.24268	0.63464, 1.0874	1.9876	2.7788	
Brassicaceae	*Aethionema arabicum*	4	0.1855	0.828	1.7589	2.7347	
Cleomaceae	*Cleome violacea*	8	0.000024375, 0.0055378, 0.018361, 0.061113, 0.1887, 0.4759		1.4646	2.7228	
Capparaceae	*Capparis cartilaginea*	10	0.000035353, 0.0056959, 0.013704, 0.030639, 0.062232, 0.1242, 0.32373, 0.52533		1.5314	2.4911	
Pentadiplandra	*Pentadiplandra brazzeana*	9	0.00001508, 0.0050629, 0.016682, 0.061524, 0.21366		0.918, 1.66	2.45, 2.8964	
Gyrostemonaceae	*Gyrostemnon ramulosus*	9	0.000014882, 0.011284, 0.085188, 0.24244, 0.37410, 0.53412		0.82579, 1.0607	1.8093	
Resedaceae	*Reseda odorata*	6	0.04289, 0.3833		1.1461, 2.0656	2.6333, 2.9296	
Emblingiaceae	*Emblinga calceoliflora*	4	0.0046699, 0.18427		0.96628	2.3681	
Koeberliniaceae	*Koeberlinia spinosa*	3	0.43929		0.88782	1.8076	
Bataceae	*Batis maritima*	4	0.18498		0.63107	1.9075	2.7689
Limnanthaceae	*Limnanthes douglasii*	5	0.1948, 0.44432		1.04, 1.6153	2.5427	
Moringaceae	*Moringa oleifera*	4	0.31549			1.467, 2.2365	2.729
Caricaceae	*Carica papaya*	9	0.00001, 0.02373, 0.08525, 0.19217, 0.39895, 0.83318			1.867, 2.6017	2.9362
Tropaeolaceae	*Tropaeolum peregrinum*	5				0.59732, 0.73756	1.6649, 2.4504, 2.9207
Akaniaceae	*Akania bidwillii*	10	0.000058371, 0.0080483, 0.016821, 0.031955, 0.0562, 0.088493, 0.15322, 0.30333			1.0826	2.3495

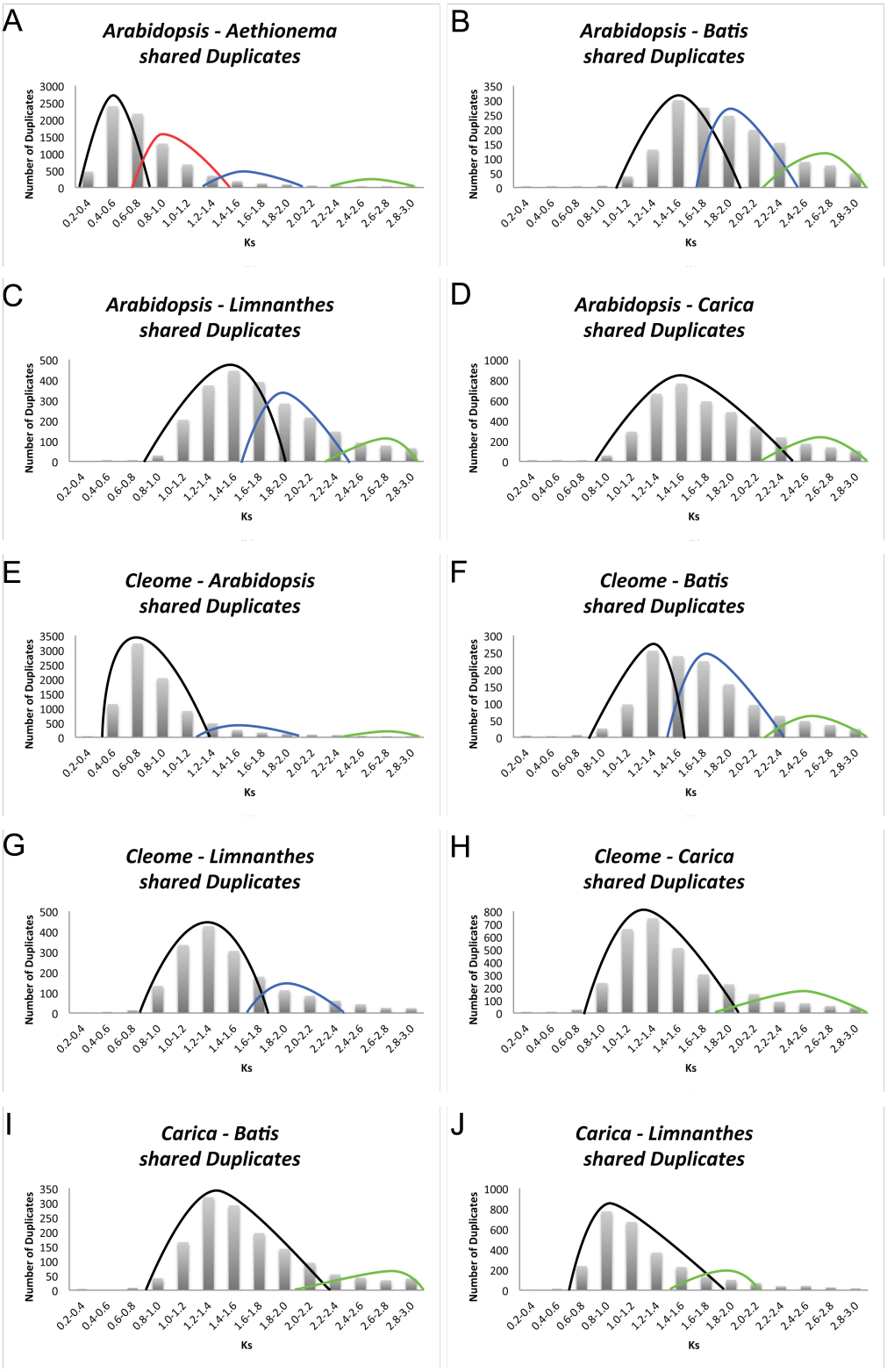

A
Arabidopsis - Aethionema
shared Duplicates
B
Arabidopsis - Batis
shared Duplicates
C
Arabidopsis - Limnanthes
shared Duplicates
D
Arabidopsis - Carica
shared Duplicates
E
Cleome - Arabidopsis
shared Duplicates
F
Cleome - Batis
shared Duplicates
G
Cleome - Limnanthes
shared Duplicates
H
Cleome - Carica
shared Duplicates
I
Carica - Batis
shared Duplicates
J
Carica - Limnanthes
shared Duplicates
Number of Duplicates
Ks
0.2-0.4
0.4-0.6
0.6-0.8
0.8-1.0
1.0-1.2
1.2-1.4
1.4-1.6
1.6-1.8
1.8-2.0
2.0-2.2
2.2-2.4
2.4-2.6
2.6-2.8
2.8-3.0

Figure S2.2: Synonymous substitutions (Ks) age distributions of shared duplicates between paired Brassicales species

Analysis was constructed using the DupPipe pipeline (http://EvoPipes.net)(93) and evaluated statistically with a mixture model analysis using EMMIX (97). EMMIX mixture model results are provided in Table S2.2. Species divergence (i.e. divergence of orthologs) is shown beneath the black curves, and whole genome duplications are beneath colored curves: At-α (red), At-β (blue), and At-γ (green). The At-α event is shared by both *Arabidopsis thaliana* and *Aethionema arabicum* (Panel A) but not present in *Cleome violacea*, *Batis maritima*, *Limnanthes douglasii* or *Carica papaya*. The At-β and At-γ duplication is shared by all of these species, except that *C. papaya* lacks At-β (Panels D, H, I, & J).

Table S2.2: EMMIX mixture model results from the analyses of Ks distribution

Data calculated from transcriptome and genome datasets for paired species to identify shared whole genome duplications using the DupPipe pipeline (93, 97). The total number of components and the estimated mean for each component are listed for the species divergence (i.e. divergence of orthologs) and three ancient whole genome duplications shared by *Arabidopsis thaliana*.

Species 1	Species 2	No. Components	Species Divergence	At-alpha	At-beta	At-gamma
Arabidopsis thaliana	*Aethionema arabicum*	7	0.040944, 0.44598, 0.58866	0.77488, 0.90259	1.5479	2.6389
Arabidopsis thaliana	*Batis maritima*	4	0.058412, 1.5332		1.9181	2.7841
Arabidopsis thaliana	*Limnanthes douglasii*	3	1.3977		1.8408	2.7935
Arabidopsis thaliana	*Carica papaya*	5	0.15385, 1.3749, 1.8150			2.5920, 2.8722
Cleome violacea	*Arabidopsis thaliana*	6	0.022428, 0.63677, 0.81077, 1.0786		1.5333	2.8487
Cleome violacea	*Batis maritima*	5	0.049949, 0.37187, 1.3298		1.6731	2.6082
Cleome violacea	*Limnanthes douglasii*	3	0.0015999, 1.2737		1.8974	
Cleome violacea	*Carica papaya*	4	0.16018, 1.2207			1.7435, 2.6609
Carica papaya	*Batis maritima*	4	0.09581, 1.368			1.9020, 2.8908
Carica papaya	*Limnanthes douglasii*	3	0.94914, 1.2664			1.8741

S2.2 Shared Ancient Gene Duplicates between *Limnanthes* and *Arabidopsis*

The retained At-β duplicates identified in *Arabidopsis thaliana* and *Limnanthes douglasii* were mapped onto the *Arabidopsis* karyotype to validate a shared At-β whole genome duplication. A list of retained At-β duplicates in the *Arabidopsis* genome (88) was screened against the PAML analyses of the combined *Limnanthes* and *Arabidopsis* datasets (Figure S2.2 Panel C). The identified At-β duplicates would include only those expressed in the Limnanthes transcriptome data and using the DupPipe Pipeline.

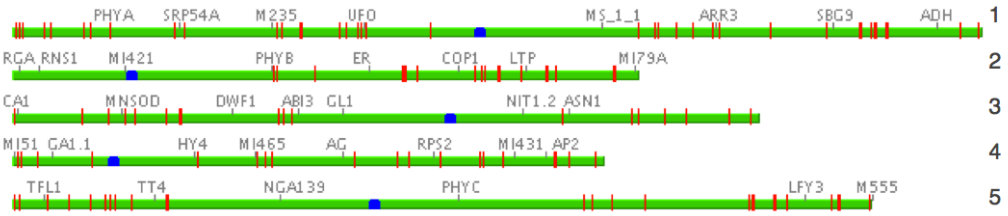


Figure S2.3: The retained At-β duplicates identified in the Limnanthes douglasii and Arabidopsis thaliana comparison (Figure S2.2 Panel C)

The At-β duplicates found in both *L. douglasii* and *A. thaliana* were mapped on each of the *Arabidopsis* chromosome (1-5). Blue markers indicate centromeres, standard gene markers are indicated with black tick marks and abbreviated names, and red markers indicate shared *Arabidopsis-Limnanthes* At-β duplicates. See Table S2.3 for the list of identified shared At-β duplicates from the *Limnanthes* transcriptome using the DupPipe pipeline. The image was generated using TAIR Map Viewer (https://www.arabidopsis.org/servlets/mapper).

Table S2.3: Retained At-β duplicates shared by both Arabidopsis thaliana and Limnanthes douglasii

These are listed as gene names per *Arabidopsis* chromosome. See Figure S2.3 for distribution across each Arabidopsis chromsome. The list of retained At-β duplicates in the *Arabidopsis* genome was obtained from Bowers et al. (2003).

Chromosome 1	Chromosome 2	Chromosome 3	Chromosome 4	Chromosome 5
AT1G01340	AT2G18960	AT3G01300	AT4G00360	AT5G01240
AT1G01600	AT2G19160	AT3G07010	AT4G00660	AT5G01620
AT1G01960	AT2G22290	AT3G09840	AT4G01830	AT5G04140
AT1G03930	AT2G28540	AT3G11250	AT4G04940	AT5G05980
AT1G04440	AT2G28620	AT3G12110	AT4G09160	AT5G07720
AT1G07380	AT2G28760	AT3G14400	AT4G13210	AT5G09410
AT1G07940	AT2G29650	AT3G15240	AT4G13710	AT5G09950
AT1G09540	AT2G34410	AT3G15500	AT4G14160	AT5G10260
AT1G14840	AT2G34850	AT3G15610	AT4G19710	AT5G11700
AT1G15750	AT2G34890	AT3G23340	AT4G23010	AT5G15020
AT1G23340	AT2G35155	AT3G23660	AT4G23920	AT5G15080
AT1G23870	AT2G36350	AT3G24230	AT4G26100	AT5G44480
AT1G23900	AT2G36390	AT3G46830	AT4G26600	AT5G45030
AT1G26150	AT2G38120	AT3G52370	AT4G29950	AT5G46340
AT1G26270	AT2G40010	AT3G52890	AT4G30190	AT5G48900
AT1G29400	AT2G40150	AT3G55140	AT4G31860	AT5G57110
AT1G29890	AT2G40810	AT3G57140	AT4G35880	AT5G57270
AT1G30620	AT2G45740	AT3G60860	AT4G36070	AT5G57350
AT1G30820	AT2G45810	AT3G62570	AT4G36860	AT5G57410
AT1G31120	AT2G45970		AT4G37100	AT5G59150
AT1G35580			AT4G39350	AT5G59290
AT1G52730				AT5G60390
AT1G53900				AT5G64220
AT1G54280				AT5G64740
AT1G55690				AT5G64990
AT1G57560				AT5G67380
AT1G59740				
AT1G60070				
AT1G64060				
AT1G68060				
AT1G70450				
AT1G70550				
AT1G71530				
AT1G71810				
AT1G71830				
AT1G71940				
AT1G72000				
AT1G72700				
AT1G72960				
AT1G72990				
AT1G79000				
AT1G80490				

S2.3. Constructing Gene Trees to Validate Phylogenetic Placement of At-β event

We used gene family tree analyses to validate phylogenetic localization of the At-β event. A total of 1263 gene family trees were constructed using our transcriptome data with SATé package 2.2 and RAxML 7.2.6 (71, 98), following clustering gene families using reciprocal best BLAST hits and MCL clustering (99). Tree topologies were summarized for gene duplication events using custom scripts. The most common topology supports Ks distribution data (Figure S2.4, Panel A), and consistent with placement of the At-β event in Figure 1.

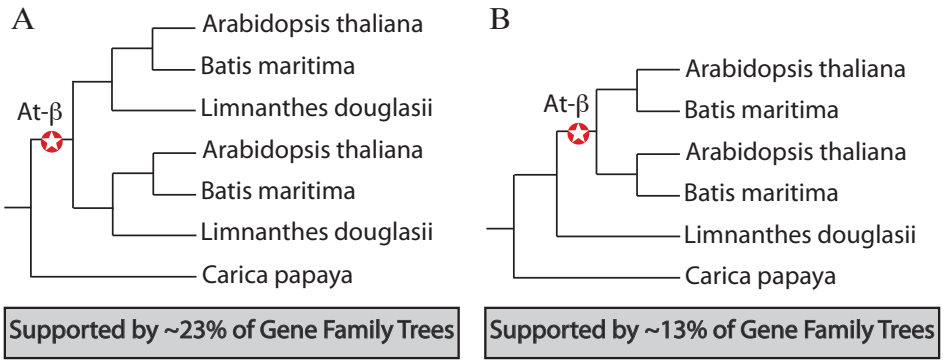
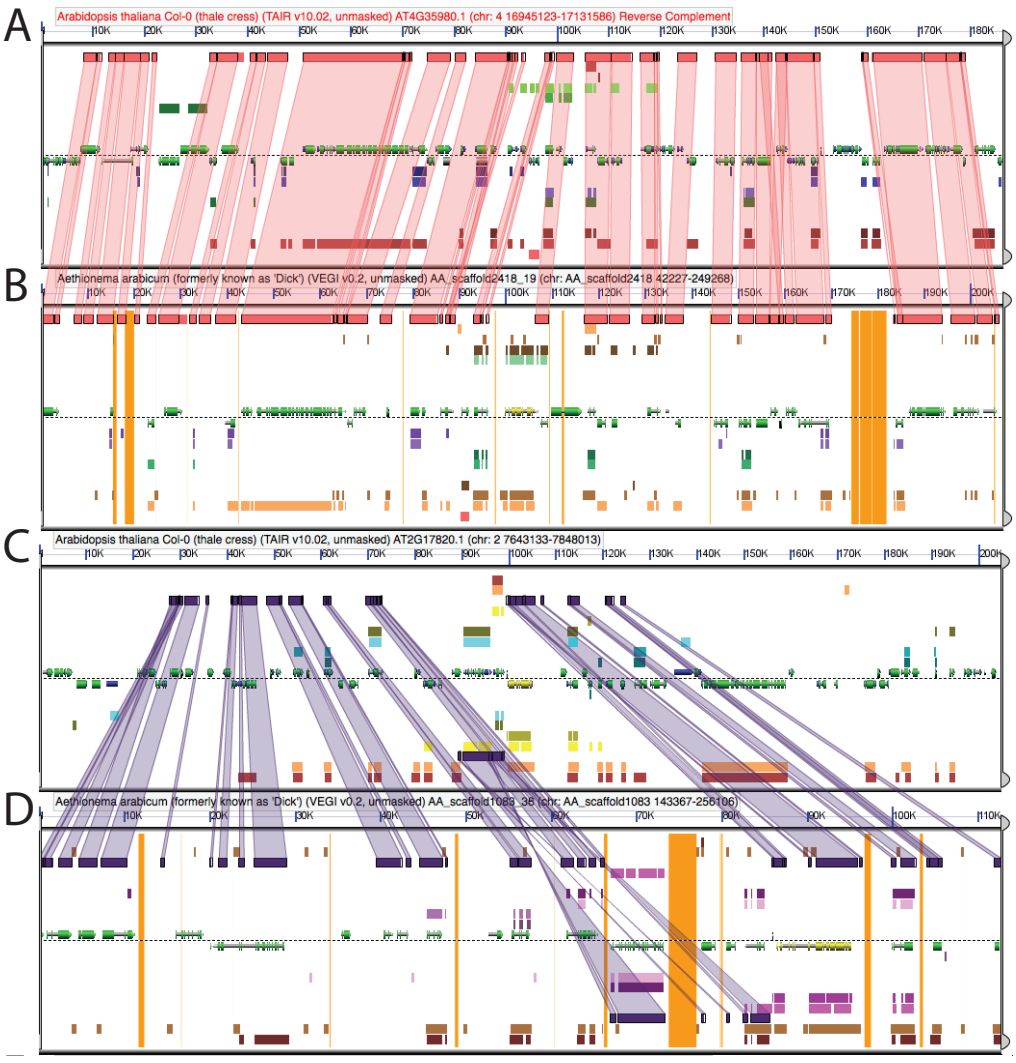


Figure S2.4: Gene family tree analyses to support At-β event

A. The most common tree topology supports that placement of the At-β event at the most common recent ancestor of *Arabidopsis thaliana*, *Batis maritima*, and *Limnanthes douglasii*. **B**. The second most common tree topology, supported by nearly half as many trees, places the At-β event at the most common recent ancestor of *Arabidopsis* and *Batis*, not shared by *Limnanthes*. The remaining topologies support a variety of non-duplicated species trees (i.e. gene loss following At-β in each species). These results support the family Limnanthaceae sharing the At-β whole genome duplication (Figure 1).

S2.4 Comparative Genomic Analyses to Validate At-α and At-β Placement

The genomes of *Arabidopsis thaliana*, *Aethionema arabicum*, *Carica papaya,* and *Vitis vinifera* (outgroup) were compared to identify shared whole genome duplications across Brassicales families. This analysis revealed a 4At:4Aa:1Cp:1Vv syntenic relationship, which support the absence of both At-α and At-β in *Carica* and *Vitis* (52), and the presence of both events in *Aethionema* (92). These results are consistent with previous analyses for the relative placement of both events in Figure 1.


A
Arabidopsis thaliana Col-0 (thale cress) (TAIR v10.02, unmasked) AT4G35980.1 (chr: 4 16945123-17131586) Reverse Complement
B
Aethionema arabicum (formerly known as 'Dick') (VEGI v0.2, unmasked) AA_scaffold2418_19 (chr: AA_scaffold2418 42227-249268)
C
Arabidopsis thaliana Col-0 (thale cress) (TAIR v10.02, unmasked) AT2G17820.1 (chr: 2 7643133-7848013)
D
Aethionema arabicum (formerly known as 'Dick') (VEGI v0.2, unmasked) AA_scaffold1083_36 (chr: AA_scaffold1083 143367-256106)

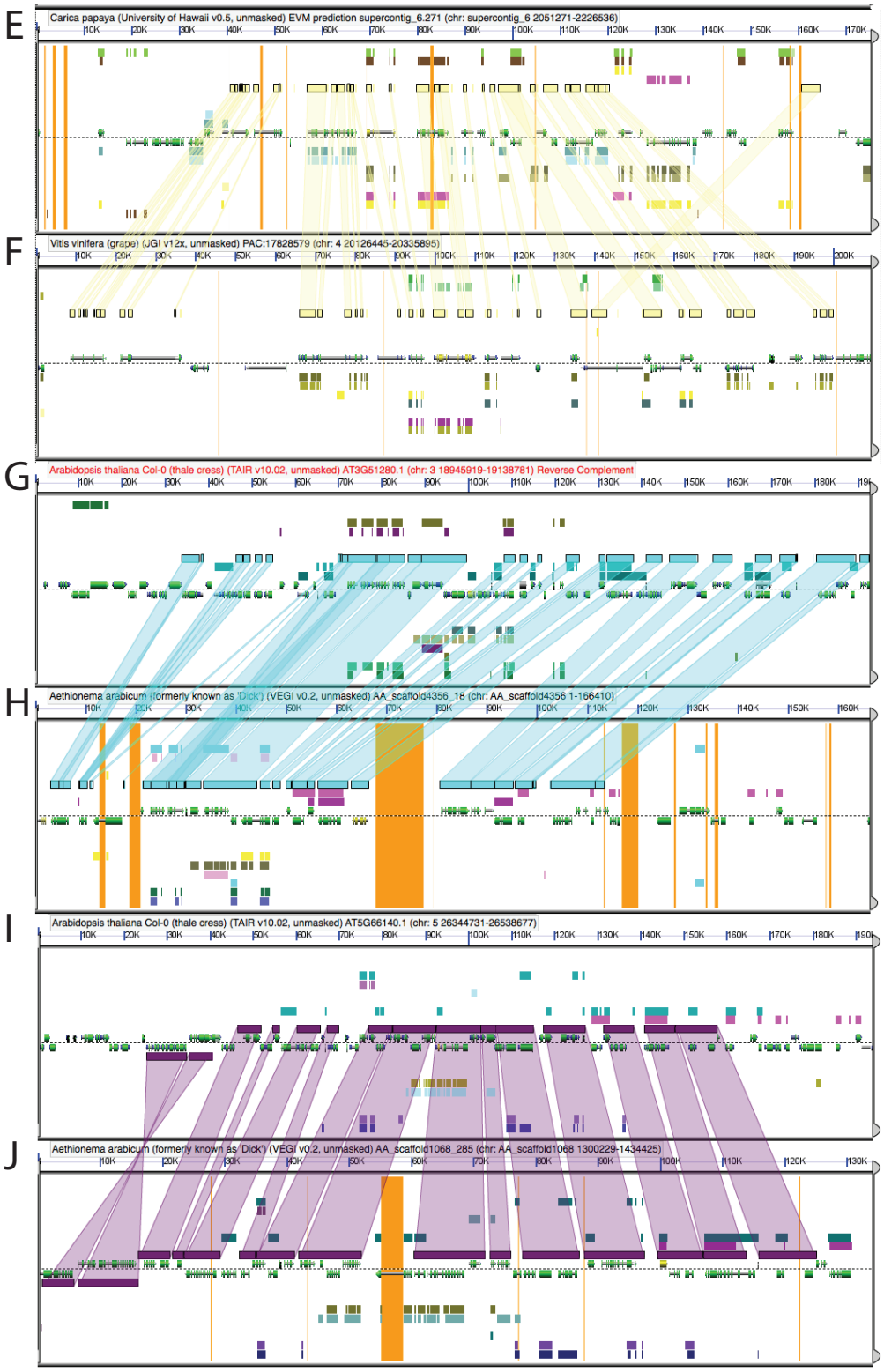

E
Carica papaya (University of Hawaii v0.5, unmasked) EVM prediction supercontig_6.271 (chr: supercontig_6 2051271-2226536)
F
Vitis vinifera (grape) (JGI v12x, unmasked) PAC:17828579 (chr: 4 20126445-20335895)
G
Arabidopsis thaliana Col-0 (thale cress) (TAIR v10.02, unmasked) AT3G51280.1 (chr: 3 18945919-19138781) Reverse Complement
H
Aethionema arabicum (formerly known as 'Dick') (VEGI v0.2, unmasked) AA_scaffold4356_18 (chr: AA_scaffold4356 1-166410)
I
Arabidopsis thaliana Col-0 (thale cress) (TAIR v10.02, unmasked) AT5G66140.1 (chr: 5 26344731-26538677)
J
Aethionema arabicum (formerly known as 'Dick') (VEGI v0.2, unmasked) AA_scaffold1068_285 (chr: AA_scaffold1068 1300229-1434425)

Figure S2.5: Syntenic comparisons of selected Brassicales species

Arabidopsis thaliana (Panels A, C, G, & I) *Aethionema arabicum* (Panels *B, D, H, & J)*, *Carica papaya (Panel E)* and *Vitis vinifera (Panel F)* genomic regions, revealed a 4At:4Aa:1Cp:1Vv syntenic relationship. These results support the absence of both At-α and At-β in *Carica* and *Vitis* (52), and the presence of both events in *Aethionema* (92). Results can be regenerated at: (http://genomevolution.org/r/93xd). Genes are depicted in the middle of each panel along a dashed line as green and yellow models, and shared syntenic regions as color-coded blocks to each of the other panels. For example, all *Arabidopsis* and *Aethionema* regions share syntenic blocks to the single copy regions in both *Carica* and *Vitis* (e.g. light green blocks in the *Carica* panel matches *Arabidopsis* Panel A). The light yellow shaded connectors between light yellow blocks highlight conserved syntenic blocks between the *Carica* and *Vitis* regions. The other four shaded connectors between blocks (e.g. red blocks in Panel A & B) highlight conserved regions between *Arabidopsis* and *Aethionema* regions.

Supplementary Text S3

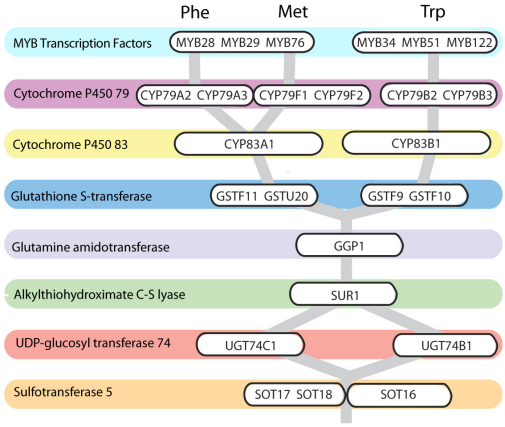
Origin of pathways that synthesize novel glucoinolate classes across Brassicales

The Brassicales are united by their ability to synthesize glucosinolates (i.e. mustard oils), which function as chemical defense compounds against oviposition and herbivory (100). More than 120 glucosinolates compounds have been characterized, comprising distinctly different classes (81), the presence of which variable across the Brassicales families (82). Several of the classes are entirely novel to more derived Brassicales families (Figure 1; Table S1.5). Gene duplications are known to be a major evolutionary force causing variation across glucosinolate biosynthetic pathways within the Brassicaceae family (101, 102). Here we present for the first time the gene duplication history for the core glucosinolate pathways across the Brassicales, and the origin of pathways that synthesize novel glucosinolate classes (Figure 2).

The core pathway for families that lack the At-β event (Figure 1; e.g. Caricaceae) uses only phenylalanine and branched chained amino acids as a substrate (103). Our results from the comparison of these metabolic pathways across the Brassicales show numerous retained duplicates across the core pathway for families sharing the At-β event. The retained duplicates shared by these Brassicales families now function in the biosynthesis of indolic glucosinolates from the novel amino acid substrate tryptophan. The key enzymatic steps and upstream transcriptional regulators encoded by these duplicate genes are specific to the biosynthesis of indolic glucosinolates (Table S3.1), thus they are novel gene functions and distinct from the core-pathway found in families lacking the At-β event. The ancestral aliphatic pathway from phenylalanine is still encoded by the other sets of duplicates. The glucosinolate pathways further expanded near the common recent ancestor of Capparaceae-Cleomeaceae-Brassicaceae, involving again the origin of novel gene functions among retained duplicates that use methionine as a novel amino acid substrate within the broader aliphatic glucosinolate pathway (Figure S3.1). The novel gene functions among these gene duplicates have been experimentally validated (Table S3.1). Collectively, these results show that retained gene duplicates (Figure S3.2) underwent functional diversification over evolutionary time to synthesize novel classes of compounds from new amino acid substrates.

S3.1 Comparison of Glucosinolate Pathways between *A. thaliana* and *C. papaya*

Gene duplication histories and phylogenetic relationships for transcriptional regulators and metabolic steps that synthesize indolic glucosinolates and aliphatic glucosinolates in *Arabidopsis thaliana* and *Carica papaya* were estimated to identify how novel pathways evolved. Pathway information, including genes and directionality, were obtained from Sonderby et al. (2010) and the Plant Metabolic Network (www.plantcyc.org/). Syntenic comparative genomic analyses were performed using the iPlant Comparative Genomics tool (http://genomevolution.org/CoGe/) to elucidate the duplication history for all glucosinolate genes in both *Arabidopsis* and *Carica*. Results are reported in Figure S3.1 (e.g. MYB transcription factors: six copies in *Arabidopsis* and one *Carica* ortholog), as well as the phylogenic relationships of duplicates are shown in Figure S3.2. Our phylogenetic estimates are congruent with previous estimates (103, 104), which show that the indolic glucosinolate pathway is largely distinct from the aliphatic glucosinolate pathway. Here, we show that the indolic glucosinolate pathway arose from gene duplications following the divergence from *Carica papaya*.



Arabidopsis Genome No. Genes	*Carica* Genome No. Genes	Genome Syntenic Analyses of Genes	Gene Phylogeny
6	1	http://genomevolution.org/r/960g	Figure S3.2-A
6	2	http://genomevolution.org/r/9624	Figure S3.2-B
2	1	http://genomevolution.org/r/9626	Figure S3.2-C
4	2	http://genomevolution.org/r/9629	Figure S3.2-D
1	1	http://genomevolution.org/r/9639	Figure S3.2-E
1	1	http://genomevolution.org/r/962h	Figure S3.2-F
2	1	http://genomevolution.org/r/962c	Figure S3.2-G
3	1	http://genomevolution.org/r/962f	Figure S3.2-H

Figure S3.1: The regulatory and biosynthetic pathways for glucosinolate biosynthesis in *Arabidopsis*

The genes and reaction directionality (top to bottom) for Aliphatic (Phenylalanine and Methionine as substrates) and Indolic (Tryptophan as a substrate) compounds is depicted. See Plant Metabolic Network (www.plantcyc.org/) for more detailed description of these biosynthetic pathways. Syntenic comparative genomic analyses were performed using the iPlant Comparative Genomics tool (http://genomevolution.org/CoGe/) to elucidate the duplication history for each biosynthetic step in both *Arabidopsis thaliana* and *Carica papaya*. Results are reported in the table (e.g. MYB transcription factors: six copies in *Arabidopsis* and one *Carica* ortholog) and phylogenic relationships of duplicates shown in Supplemental Figure S3.2. These results support the presence of only the Aliphatic (Phenylalanine) pathway in *Carica papaya*.

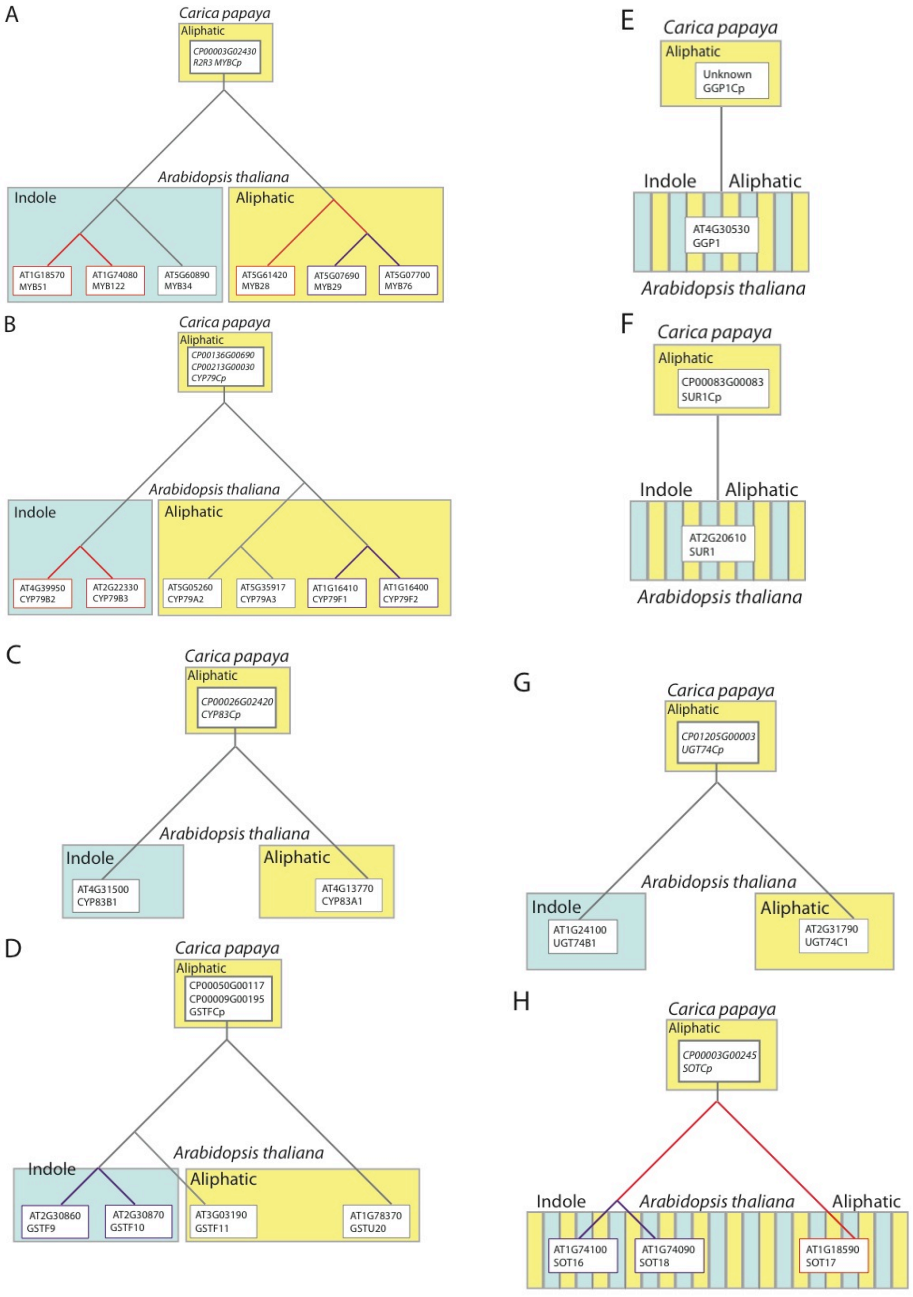

A
Carica papaya
Aliphatic
CP00003G02430
R2R3 MYBCp
Arabidopsis thaliana
Indole
Aliphatic
AT1G18570
MYB51
AT1G74080
MYB122
AT5G60890
MYB34
AT5G61420
MYB28
AT5G07690
MYB29
AT5G07700
MYB76
B
Carica papaya
Aliphatic
CP00136G00690
CP00213G00030
CYP79Cp
Arabidopsis thaliana
Indole
Aliphatic
AT4G39950
CYP79B2
AT2G22330
CYP79B3
AT5G05260
CYP79A2
AT5G35917
CYP79A3
AT1G16410
CYP79F1
AT1G16400
CYP79F2
C
Carica papaya
Aliphatic
CP00026G02420
CYP83Cp
Arabidopsis thaliana
Indole
Aliphatic
AT4G31500
CYP83B1
AT4G13770
CYP83A1
D
Carica papaya
Aliphatic
CP00050G00117
CP00009G00195
GSTFCp
Arabidopsis thaliana
Indole
Aliphatic
AT2G30860
GSTF9
AT2G30870
GSTF10
AT3G03190
GSTF11
AT1G78370
GSTU20
E
Carica papaya
Aliphatic
Unknown
GGP1Cp
Indole
Aliphatic
AT4G30530
GGP1
Arabidopsis thaliana
F
Carica papaya
Aliphatic
CP00083G00083
SUR1Cp
Indole
Aliphatic
AT2G20610
SUR1
Arabidopsis thaliana
G
Carica papaya
Aliphatic
CP01205G00003
UGT74Cp
Arabidopsis thaliana
Indole
Aliphatic
AT1G24100
UGT74B1
AT2G31790
UGT74C1
H
Carica papaya
Aliphatic
CP00003G00245
SOTCp
Indole
Arabidopsis thaliana
Aliphatic
AT1G74100
SOT16
AT1G74090
SOT18
AT1G18590
SOT17

Figure S3.2: Phylogeny and gene duplication history for transcriptional regulators and genes that synthesize indolic and aliphatic glucosinolates in *Arabidopsis thaliana* and *Carica papaya*

See Figure S3.1 for description of biosynthetic pathways (Panels A-H follow pathway directionality beginning with MYB (myeloblastosis) transcription factors). Our estimates of phylogenetic relationships are congruent with previous estimates (103, 105), and genes are color-coded as encoding products synthesizing only Indolic (blue), only Aliphatic (yellow), or both glucosinolate groups (stripped blue and yellow). These inferences are based on Sonderby et al. (2010) and on the Plant Metabolic Network (www.plantcyc.org/). Red branches indicate retained duplicates from the At-α event, purple branches are tandem duplicates: grey branches are cases where the available comparative genomic data are insufficient to reconstruct the duplication history (See Figure S3.1 for hyperlinks to regenerate analyses). The genes that synthesize indolic glucosinolates arose from duplications unique to the *Arabidopsis thaliana* lineage (Figure 1) following its divergence from *Carica papaya*. Analyses of our transcriptome data, shown in Figure 2, suggests that these duplicates arose due to either the At-β event or a series of other duplications occurring within the time frame of this whole genome duplication.

Table S3.1: Substrate specificity of specific duplicate genes

This table summarizes the studies (104-119) that have established the substrate specificity of duplicate glucosinolate biosynthetic genes (Figure S3.2), functioning in either the aliphatic or indolic pathways (Figure S3.1). Our results support this body of published work and put that work into the broader context of the origin of novel glucosinolate biosynthetic pathways.

Biosynthetic Steps	Aliphatic Pathway	Citation
MYB Transcription Factors	MYB28, MYB29, & MYB76	Sonderby et al., 2007; Sonderby et al., 2010
Cytochrome P450 79	CYP79A2 & CYP79A3 (Phe); CYP79F1 & CYP79F2 (Met)	Wittstock and Halkier, 2000 ; Chen et al., 2003; Sonderby et al., 2007
Cytrochrome P450 83	CYP83A1	Naur et al., 2003
Glutathione S-transferase	GSTF11 & GSTU20	Sonderby et al., 2010
UDP-glucosyl Transferase 74	UGT74C1	Grubb et al., 2014
Biosynthetic Steps	**Indolic Pathway**	**Citation**
MYB Transcription Factors	MYB34, MYB51, & MYB122	Celenza et al., 2005; Gigolashvili et al., 2007; Frerigmann and Gigolashvili, 2014
Cytochrome P450 79	CYP79B2 & CYP79B3	Hull et al., 2000; Mikkelsen et al., 2000; Zhao et al., 2002
Cytrochrome P450 83	CYP83B1	Bak et al., 2001
Glutathione S-transferase	GSTF9 & GSTF10	Sonderby et al., 2010
UDP-glucosyl Transferase 74	UGT74B1	Grubb et al., 2004

S3.2 Constructing glucosinolate pathways across Brassicales families

In order to further evaluate how novel glucosinolate pathways evolved over time, we investigated the origin of novel transcriptional regulators and metabolic steps that synthesize indolic and aliphatic glucosinolates derived from phenylalanine, tryptophan, and methionine (Figure S3.1) using the transcriptomes and genomes spanning 14 Brassicales families (Supplemenary Text S1). Pathway information in *Arabidopsis thaliana*, includes genes and directionality, were obtained from Sonderby et al. (2010) and the Plant Metabolic Network (www.plantcyc.org/). Transcriptome sequencing and assembly is described in Supplementary Note S1, genes encoding glucosinolate pathway were identified using protein BLAST analyses (Altschul et al., 1997) and results analyzed using the known *Arabidopsis* pathways and our phylogenetic framework (Figure 1). These results shown in Figure S4.3 are congruent with Figure S4.1, which shows that the Indolic glucosinolate pathway arose following the At-β event, and prior to the origin of the Methionine (Met) derived aliphatic glucosinolate pathway. The average copy number across all GLS biosynthetic steps increased during the evolutionary history of *Arabidopsis* with an average of 1.71 copies for post- At-β and 2.89 copies for post- At-α families (Figure S4.3). The Citric Acid (TCA) Cycle, which was selected as a control metabolic pathway of similar size, remained nearly constant in copy number across these polyploid intervals, with only a single tandem duplication unique to the Brassicaceae. Similar results were observed for the Glycolysis I pathway (Figure S4.3).

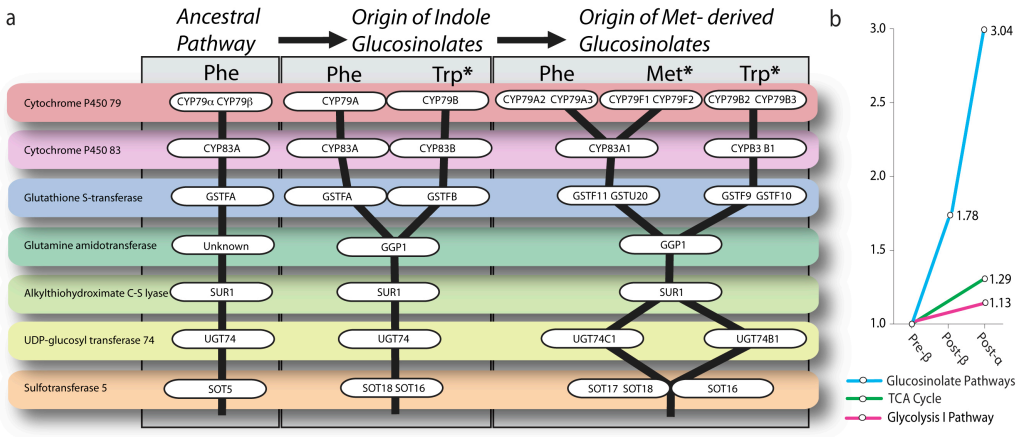


Figure S3.3. An illustration of the expansion of glucosinolate biosynthetic pathways

Substrates tryptophan (Trp), phenylalanine (Phe), and methionine (Met) are indicated, with enzymes depicted as white circles, and each pathway as black lines. These pathways were constructed using available transcriptome and genome data (Figure 1; Supplementary Text S1 & S2), revealing that the average copy number for many of the upstream biosynthetic step increases following both the At-β and At-α. For example, there is only a single copy of cytochrome P450 79 (CYP79) among pre- At-β families, two copies among post- At-β families, and six copies of this enzyme for post- At-α Brassicaceae. Each of these CYP79s have unique substrate specificities (Table S3.1). These results suggest that the indolic glucosinolate pathway arose following the At-β event, and later the origin of the Methionine (Met-) derived glucosinolate pathway evolved at the most common recent ancestor of Capparaceae and Cleomaceae (Figure 1).

Supplementary Text S4

Pieridae: estimation of the pylogeny & timing of divergences, analysis of divergences, hostplant usage, and diversification dynamics

The butterfly family Pieridae is comprised of a total of about 1000 species that are spread across 85 genera and 15 subgenera (120). Previously, the most thorough study of the Pieridae examined 90 taxa, which were representatives of 74 genera (121). While one nuclear gene was sequenced from all of these specimens (elongation factor-1alpha: EF-1α), 30 of these species had sequence data from an additional two nuclear gene regions and one mitochondrial gene region. Analyses found strong support for each of the 4 subfamilies, with Coliadinae being sister to Pierinae. However, support for the phylogenetic relationships at lower evolutionary levels was poor, and this poor resolution prevents a robust analysis of the evolutionary relationships and times of divergence among tribes and genera (121).

Recently we have significantly extended this previous genus level study to include eight gene regions covering a total of ~6700 bp (one mitochondrial and seven nuclear protein coding genes) in an analysis solely focused upon the systematic relationships among taxa (120). In our manuscript here, we use other methods upon the same data to simultaneously reconstruct both the phylogenetic relationships and times of divergences among taxa using fossil calibrations (described below). Briefly, the 96 taxa of Pieridae used were for the most part derived from the same specimens used in the previous genus level study (121), with a few novel specimens collected specifically for this study, and a few taxa for which sequence data was downloaded from NCBI (Supplementary Text S1). In addition, the outgroup taxa used in the study (n=14) were from the families Nymphalidae, Lycaenidae and Riodinidae; outgroup taxon sequences were taken from (122).

Once the time calibrated phylogeny was generated, this was then used in a series of analyses to determine diversification dynamics across a range of assumptions.

S4.1 Phylogeny & timing of divergences

The eight gene regions used were: the mitochondrial gene region *cytochrome oxidase subunit I* (COI) and the nuclear gene regions *elongation factor-1α* (EF-1α), *ribosomal protein S5* (RpS5), *carbamoylphosphate synthase domain protein* (CAD), *cytosolic malate dehydrogenase* (MDH), *glyceraldehyde-3-phosphate dehydrogenase* (GAPDH), *isocitrate dehydrogenase* (IDH) and *wingless*. PCR and sequencing protocols follow Wahlberg and Wheat (123). Alignment of gene regions was trivial, as all are protein-coding genes with a conserved codon structure. These gene regions have been used successfully for studies on butterfly relationships (122, 124). Again, this data has solely been used to infer the systematic relationships among taxa (120). Here we report on independent analyses that use the same data, but are primarily focused upon simultaneously reconstructing the timing of divergences among the studied genera.

Bayesian analyses using BEAST 1.7.4 (125) were run for 10 million generations sampling every 1000 generations, with the number of independent runs depending on the analysis (minimally two independent runs). The convergence of the likelihood traces of the independent runs was assessed with Tracer v1.5 and the ESS values were verified to be above 200 for all parameters.

Partitioning of large datasets is necessary, especially when different gene regions with different mutational dynamics are being used. Traditionally datasets are partitioned by gene region, sometimes divided into codon positions (126). Recently a new method partitioning data by relative rates of evolution has been advocated (127, 128) and here we compare the two strategies using Bayes Factors. The data were first partitioned by GENE with eight subdivisions. The second strategy followed that of Rota and Wahlberg (128). The data were sorted according to relative rates of evolution as calculated by the program TIGER (127). The relative rates were divided into 30 equal bins and all nucleotide sites were then placed into the 30 bins based on their relative rates. For phylogenetic analyses bins 1 to 23 were combined, as were bins 24 and 25, in order to achieve partitions with more than 100 characters in each (see (120) for additional details). As a result, there were seven partitions in the TIGER partitioned data. Analyses in BEAST were done with both partitioning strategies on datasets that did not include outgroups. We find, as in Walhberg et al. (120), that the TIGER partition is decisively better than the GENE partition based upon Bayes Factors. We also find that the TIGER partition runs significantly faster in our BEAST analysis. Thus we report the TIGER results.

BEAST runs indicated that a few long-branch taxa were causing problems with the analyses and thus monophyly constraints were used. Based upon the previous study (120), two clades were constrained to be monophyletic: (Pseudopontinae+Coliadinae+Pierinae) and Pierinae. In addition, three clades were defined for fossil calibrations: *Talbotia+Pieris* (clade *Stolopsyche*), two species of *Pontia* (clade *Miopieris*) and the clade including *Aporia, Delias, Leuciacria, Melete, Leodonta, Pereute, Eucheira, Neophasia, Catasticta, Charonias* and *Archonias* (clade *Coliates*). The fossils used for these calibrations are discussed in Braby et al. (121), but see de Jong (129) for a critical overview. Here they are used as age priors for the crown nodes described above and they were modeled as normal distributions (130): *Miopieris* 10 million years (standard deviation 2 my), *Coliates* 33 my (s.d. 2 my) and *Stolopsyche* 34 my (s.d. 2 my). The GTR + G model of substitutions was used for each partition, the clock models were unlinked across partitions and the uncorrelated lognormal distribution was used to model branch lengths, the Birth-Death process was used as a tree prior. Most prior distributions were left to their defaults, but the ucld.mean prior was changed to an exponential distribution with a mean of 0.1. Marginal likelihoods were estimated in BEAST based on the stepping-stone sampling method (131) as described by Baele et al. (132, 133).

Table S4.1 [Excel table] Genes and their coverage per species used in the phylogenetic analysis of Pieridae

The excel table presents in detail the genes used in the Wahlberg et al. (120), indicating the exact number of base pairs recovered and used for each specimen, their % coverage in relation to the total number of base pairs potentially able to be recovered, and their source when other than Wahlberg et al. (120). See Wahlberg et al. (120) for GenBank accession numbers.

S4.2 Hostplant usage and species counts

Hostplant usage data was derived from the Funet web resource, which is a large compendium of literature focused upon the natural history of the Lepidoptera.
http://www.nic.funet.fi/pub/sci/bio/life/insecta/lepidoptera/ditrysia/papilionoidea/pieridae/index.html
This web resource is a compendium of the leading text based resources (e.g. (134)).

There are only two credible records of Pierinae feeding upon Brassicales plants not containing indolic glucosinolates and these involve feeding on the family Tropaeolaceae, which consists of one or three genera depending on the literature source. All occur in Central or South America and several are garden plants commonly known as nasturtium. The first instance is from the genus *Leptophobia*, living in Costa Rica. This is likely a valid observation. The second comes from reports of *Pieris* butterflies feeding on nasturtium garden ornamentals. Given the historical distribution of the genus *Pieris*, these species likely never interacted with Tropaeolaceae until the Holocene, when anthropogenic impacts extended the range of both these plants and butterflies (e.g. agriculture, horticulture, crop transportation, pest invasions). Regardless, both of these butterfly genera are located in the only two clades that have invaded Brassicaceae and they are therefore expected to be among the most capable of detoxifying the widest possible range of glucosinolates produced by Brassicales as a whole. We therefore conclude that the use of Tropaeolaceae postdates the colonization of Brassicaceae.

A. Species numbers of Pierinae
Species numbers were extracted from the Global Butterfly Information System database (135) .

Haeuser, C. L., Holstein, J. & Steiner, A. (2012): The Global Butterfly Information System. http://www.globis.insects-online.de/ Last updated 08.04.2012.

B. Datafile, comma separated values.

tip.label,species_count,subfam,tribe,binary_brassicales,binary_capparaceae,binary_brassicaceae
Anteos_clorinde,3,Coliadinae,Coliadini,0,0,0
Anthocharis_cardamine,16,Pierinae,Anthocharidini,1,0,1
Aoa_affinis,1,Pierinae,Pierini,0,0,0
Aphrissa_statira,8,Coliadinae,Coliadini,0,0,0
Aporia_crataegi,32,Pierinae,Pierini,0,0,0
Appias_drusilla,39,Pierinae,Pierini,1,1,0
Archonias_brassolis,1,Pierinae,Pierini,0,0,0
Ascia_monuste,5,Pierinae,Pierini,1,1,1
Baltia_butleri,2,Pierinae,Pierini,0,0,0
Belenois_java,30,Pierinae,Pierini,1,1,0
Catasticta_cerberus,91,Pierinae,Pierini,0,0,0
Catopsilia_pomona,9,Coliadinae,Coliadini,0,0,0
Cepora_perimale,23,Pierinae,Pierini,1,1,0
Charonias_eurytele,2,Pierinae,Pierini,0,0,0
Colias_eurytheme,83,Coliadinae,Coliadini,0,0,0
Colotis_danae,47,Pierinae,Teracolini,1,1,0
Cunizza_hirlanda,1,Pierinae,Anthocharidini,0,0,0
Delias_belladonna,235,Pierinae,Pierini,0,0,0
Dercas_gobrias,4,Coliadinae,Coliadini,0,0,0
Dismorphia_zathoe,29,Dismorphiinae,Dismorphini,0,0,0
Dixeia_charina,10,Pierinae,Pierini,1,1,0
Elodina_angulipennis,26,Pierinae,Elodinini,1,1,0
Elphinstonia_charlonia,4,Pierinae,Anthocharidini,1,0,1
Enantia_lina,9,Dismorphiinae,Dismorphini,0,0,0
Eroessa_chiliensis,1,Pierinae,Anthocharidini,0,0,0
Eronia_cleodora,2,Pierinae,Teracolini,1,1,0
Eucheira_socialis,1,Pierinae,Pierini,0,0,0
Euchloe_ausonides,16,Pierinae,Anthocharidini,1,0,1
Eurema_hecabe,25,Coliadinae,Coliadini,0,0,0
Gandaca_harina,2,Coliadinae,Coliadini,0,0,0
Ganyra_josephina,3,Pierinae,Pierini,1,1,0
Gideona_lucasi,1,Pierinae,Teracolini,0,0,0
Gonepteryx_cleopatra,11,Coliadinae,Coliadini,0,0,0
Hebomoia_glaucippe,2,Pierinae,Anthocharidini,1,1,0
Hesperocharis_crocea,11,Pierinae,Anthocharidini,0,0,0
Hypsochila_wagenknechti,6,Pierinae,Pierini,0,0,0
Infraphulia_ilyodes,3,Pierinae,Pierini,0,0,0

Itaballia_demophile,3,Pierinae,Pierini,1,1,0
Ixias_pyrene,16,Pierinae,Teracolini,1,1,0
Kricogonia_lyside,2,Coliadinae,Coliadini,0,0,0
Leodonta_tellane,5,Pierinae,Pierini,0,0,0
Leptidea_sinapis,8,Dismorphiinae,Dismorphini,0,0,0
Leptophobia_aripa,17,Pierinae,Pierini,1,0,1
Leptosia_nina,9,Pierinae,Leptosiaini,1,1,0
Leuciacria_olivei,2,Pierinae,Pierini,0,0,0
Leucidia_brephos,2,Coliadinae,Coliadini,0,0,0
Lieinix_nemesis,6,Dismorphiinae,Dismorphini,0,0,0
Mathania_leucothea,4,Pierinae,Anthocharidini,0,0,0
Melete_lycimnia,6,Pierinae,Pierini,0,0,0
Moschoneura_pinthous,1,Dismorphiinae,Dismorphini,0,0,0
Mylothris_agathina,55,Pierinae,Pierini,0,0,0
Nathalis_iole,2,Coliadinae,Coliadini,0,0,0
Neophasia_menapia,2,Pierinae,Pierini,0,0,0
Nepheronia_thalassina,4,Pierinae,Nepheroniini,1,0,0
Pareronia_valeria,10,Pierinae,Nepheroniini,1,1,0
Patia_orize,3,Dismorphiinae,Dismorphini,0,0,0
Pereute_charops,9,Pierinae,Pierini,0,0,0
Perrhybris_pamela,3,Pierinae,Pierini,1,1,0
Phoebis_sennae,8,Coliadinae,Coliadini,0,0,0
Phulia_nymphula,4,Pierinae,Pierini,0,0,0
Pieriballia_viardi,1,Pierinae,Pierini,1,1,0
Pieris_napi,22,Pierinae,Pierini,1,1,1
Pierphulia_rosea,3,Pierinae,Pierini,0,0,0
Pinacopteryx_eriphia,1,Pierinae,Teracolini,1,1,0
Pontia_callidice,14,Pierinae,Pierini,1,1,1
Prioneris_philonome,7,Pierinae,Pierini,1,1,0
Pseudopieris_nehemia,2,Dismorphiinae,Dismorphini,0,0,0
Pseudopontia_paradoxa,1,Pseudopontiinae,Pseudopontini,0,0,0
Pyrisitia_proterpia,11,Coliadinae,Coliadini,0,0,0
Saletara_liberia,3,Pierinae,Pierini,0,0,0
Talbotia_naganum,1,Pierinae,Pierini,1,0,0
Tatochila_autodice,11,Pierinae,Pierini,0,0,0
Teracolus_eris,1,Pierinae,Teracolini,1,1,0
Teriocolias_zelia,1,Coliadinae,Coliadini,0,0,0
Theochila_maenacte,1,Pierinae,Pierini,0,0,0
Zegris_eupheme,3,Pierinae,Anthocharidini,1,0,1
Zerene_cesonia,2,Coliadinae,Coliadini,0,0,0

S4.3 Diversification dynamics

A. G-tests

Diversification rate comparisons can be made for specific nodes on a tree by making the reasonable assumption that sister nodes should have similar levels of species diversity, since they are of the same age and starting evolutionary material. With this conservative assumption, we tested whether there was a significantly higher level of extant species diversity in the two butterfly clades that independently colonized the Brassicaceae around 30 million years ago. The G-test of goodness-of-fit was performed in R using the Williams correct for a better approximation of the chi-square distribution, resulting in a more conservative test (136).

Anthocharidini tribe

The main split within Anthocharidini occurred around 40 million years ago, resulting in two clades that will be referred to as the *Eroessa* and *Anthocharis* clades. These respectively contain the following recognized species numbers and genera (17: *Eroessa, Cunizza, Hesperocharis, Mathania*; 39: *Anthocharis*, *Elphinstonia, Zegris, Euchloe*). Assuming equal species in both clades, we use the G-test of goodness-of-fit to quantify if our observed values are significantly different from this expectation. There are significantly more species in the *Anthocharis* clade as compared to the *Eroessa* clade ($G = 8.802$, $df = 1$, P-value = 0.003).

Pierina subtribe

This subtribe is composed of two clades, which we refer to as the *Ascia* and *Pieris* clades. These respectively contain the following recognized species numbers and genera (36: *Ascia, Ganyra, Tatochila, Theochila, Hypsochila, Pierphulia, Phulia, Infraphylia*; 63: *Pieris, Pontia, Baltia, Talbotia, Leptophobia, Pieriballia, Itaballia, Perrhybris*). Assuming equal species in both clades, we use the G-test of goodness-of-fit to quantify if our observed values are significantly different from this expectation. There are significantly more species in the *Pieris* clade as compared to the *Ascia* clade ($G = 7.973$, $df = 1$, P-value = 0.004). Within the *Ascia* clade, the *Ascia* genus is the only one that feeds on Brassicaceae plants, and thus the test above is a conservative test. Since *Ascia* feeding on Brassicaceae appears to be an independently derived trait as it is nested within a clade where it is the only Brassicaceae feeder, we can conduct the comparison above without the species in this genus (n=5). The results remained unchanged (31 species vs. 63 in the *Pieris* clade; $G = 11.056$, $df = 1$, P-value < 0.001).

B. Binary-state speciation and extinction (BiSSE) analysis

Diversification rate estimates can also be conducted in an explicit phylogenetic context using the trait state of species on an ultrametric tree. Here we used a binary-state speciation and extinction model (BiSSE) (137), with an extension for incompletely resolved phylogenies (138), as implemented in the diversitree R package (139). The incompletely resolved phylogenies option was used, not because of a weakness in support for the analyzed tree, but so we could include the number of species within each genera, as this is a genus level tree. In this model, speciation and extinction follow a birth-death process, while the rate of these two processes is allowed to vary with the two trait states in question. We partitioned the two traits as those that are feeding upon non-Brassicaceae Brassicales (group0), or on Brassicaceae (group1).

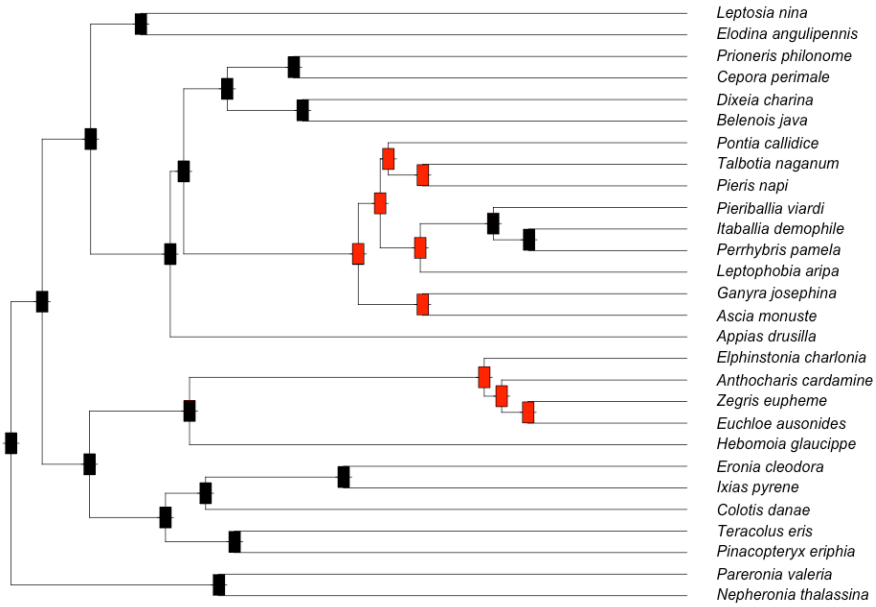


Figure S4.1 Pieridae phylogeny to infer ancestral states

Only taxa feeding on Brassicales were included, with inferred ancestral states shown below (group0 black, group1 red).

Table S4.2 Five models used and their resulting maximum likelihood estimates of parameter values

	lambda0	lambda1	mu0	mu1	q01	q10
Full	0.064	0.196	0	0.109	0.001	0.003
equal.null	0.123	0.123	0.064	0.064	0.001	0.001
equal.l	0.107	0.107	0.056	0	0	0.005
equal.mu	0.063	0.117	0	0	0.001	0.004
equal.q	0.064	0.278	0	0.21	0.002	0.002
equal.muq	0.064	0.116	0	0	0.001	0.001

The full model allowed all three parameters to be estimated in both trait groups, while the equal.null model constrained all values to be the same within each of the three parameters (i.e. set the values equal between trait groups) although the value is estimated. The qual.l model only constrained lambda (diversification rate) to be equal, equal.mu only constrained mu (extinction rate) to be equal, and equal.q only constrained the transition rate between the two states to be equal. The final model, equal.muq, constrains the extinction rate and the transition rate to be the same in both groups.

The full model is only significantly different from models that constrain lambda to be equal between the two groups (equal.null and equal.l), while constraints on extinction and transition, and both, are not significantly different from the full model. Note however that in these latter three models, the estimates of lambda show lambda0 > lambda1, as in the full model.

Table S4.3 Statistical comparisons of each constrained model to the full model

		Df	lnLik	AIC	ChiSq	Pr(>|Chi|)
Full		6	196.96	405.91		
Equal.null	1	3	204.89	415.78	15.8639	0.001209
Equal.l	2	5	199.29	408.59	4.6721	0.030657
Equal.q	3	5	197.42	404.84	0.929	0.335123
Equal.mu	4	5	197.29	404.58	0.6634	0.41535

Given these results, it is clear that group1 has a much higher diversification rate than group0. If we constrain these two groups to have the same extinction rate, and we constraint the transition between these groups to be equal, then we find that the diversification rate of group1 is nearly twice that of group0 (equal.muq model). While a lower estimate of the difference than the full model (which estimates group1 as being more than three times as high as group0), we consider this close to the lower bound estimate of the diversification rate difference.

In order to visualize these results, we also used an MCMC approach to obtain a posterior distribution of lambda values for group0 and group1, parameterized using the full model. To do this parameterized the BiSSE software (according to the manual):

```
p=starting.point.bisse(Pierinae_brassicales_feeding_tree)
prior <- make.prior.exponential(1 / (2 * p[1]))
mcmc(lik, coef(fit.full), 10000, prior=prior, print.every=0) # 10,000 samples
```

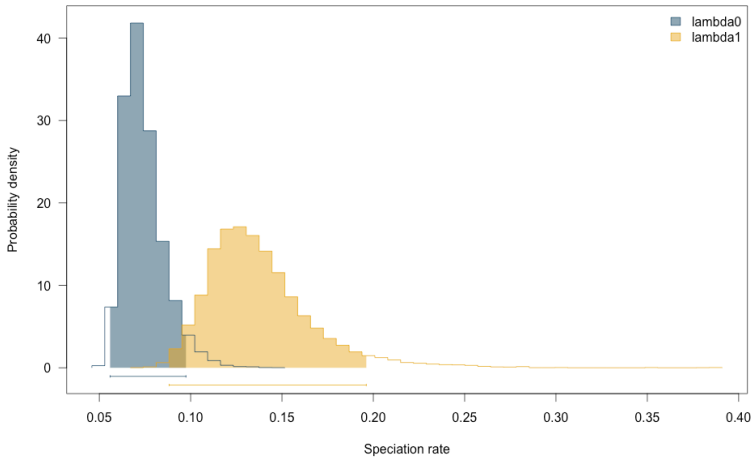


Figure S4.2 Plot of the probability density for the lambda estimates for two groups
The line below the histograms shows the 95% most probable region of the distribution.

To estimate the significance of the posterior estimates, we took the difference between each of these values (lambda group1 – lambda group0; histogram shown below). Note that the 95% most probable region of the distribution does not overlap with 0 and the values are positive, indicating that there is a significantly higher lambda in group1. Thus, of the Brassicales feeding Pierinae, those feeding on Brassicaceae have had a higher diversification rate compared to those not feeding upon Brassicaceae.

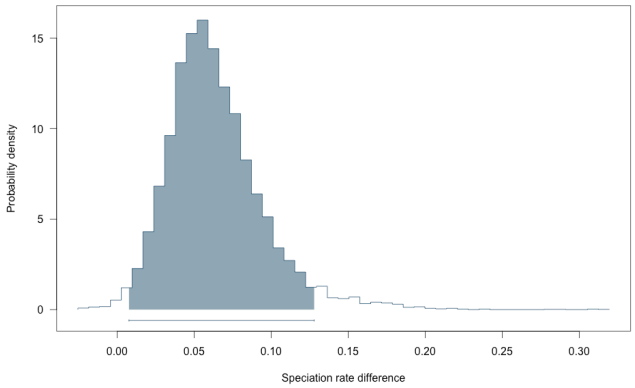


Figure S4.3 Plot of the difference in lambda estimates for the two groups

Results show lambda group1 – lambda group0. The line below the histograms shows the 95% most probable region of the distribution.

Finally, the application of BiSSE should be viewed as a complement to the previous section, as recent simulation studies have shown that the BiSSE approach, when using phylogenies having less than 300 taxa, has low power and potentially reduced accuracy and precision in its estimates. However, power does not appear to be a problem in our analysis and our estimated differences are not marginally significant. Thus, our findings using BiSSE are informative and consistent with the independent analyses reported in the previous section.

C. Modeling evolutionary diversification using stepwise AIC (MEDUSA) analysis

Changes in diversification rates can also be investigated using ultrametric tree data without any *a priori* selection of specific nodes. To do this, here we implemented a method called MEDUSA (modeling evolutionary diversification using stepwise AIC)(140). In this method, a constant parameter model of diversification is fit to the data, and then birth and death rates are allowed to shift at each node. Each node is then tested, first singly then in high groupings, with models having significant increases in fit, evaluated by an increase in the AIC value, selected and further compared. Terminal tips of the tree represent genera, and for each the number of species in that genus was used.

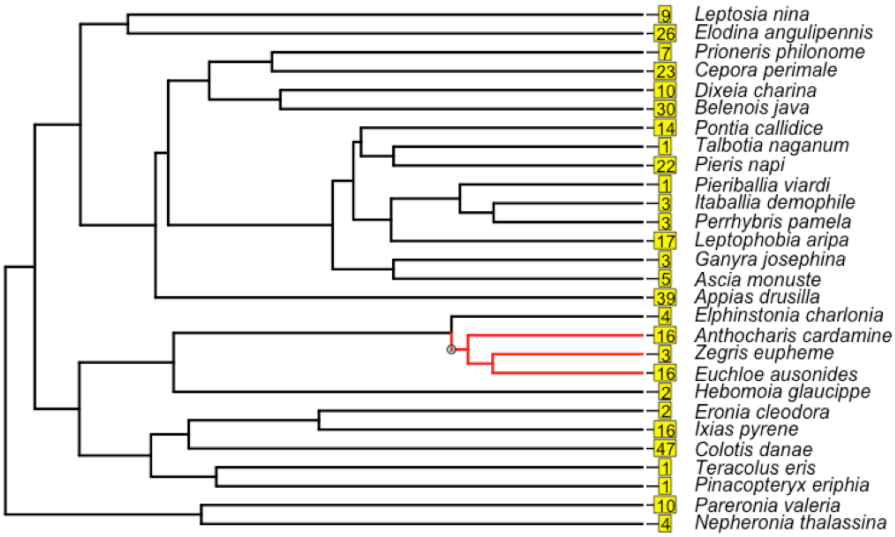


Figure S4.4 Phylogenetic tree with color shading for the two rate partitions identified using MEDUSA

Optimal MEDUSA birth-death model for tree with 28 tips representing 335 taxa. Numbers in yellow are the number of species in each genus. Node 39 and its descendants are colored red in the tree above. Lineage exemplars of genera used in sequencing are shown.

Output from model, showing the likelihood scores of the two models, and the improvement in AICC score resulting from allowing a shift at node 39:

Step 1: lnLik=-181.0129; aicc=366.2566; model=bd
Step 2: lnLik=-175.9411; aicc=363.1067; shift at node 39; model=bd; cut=stem

Resulting parameter estimates are listed below, with low and high values indicating the bounds of the 95% confidence intervals. The 95% confidence intervals on parameter values shown below is calculated from profile likelihoods. Appropriate AICC-threshold for a tree of 28 tips is: 1.091845.

Table S4.4 Optimal MEDUSA birth-death model with parameter values show

Model	Shift.Node	Ln.Lik.part	r	epsilon	r.low	r.high	eps.low	eps.high
1	29	-161.325	0.069	0.000	0.060	0.081	0.000	0.329
2	39	-14.616	0.153	0.000	0.099	0.237	0.000	0.725

Thus, the diversification rate of the *Anthocharis* clade is significantly higher than the background rate of Pierinae, when assuming a constant rate of diversification and extinction across the entire tree. This clade is estimated to approximately twice the rate of diversification as the rest of the Pierinae, which is similar to our previous findings using BiSSE.

D. Bayesian analysis of macroevolutionary mixtues (BAMM) analysis

We also investigated the ultrametric tree data for a more complex scenario of rate shifts. While the MEDUSA approach assumed a constant-rate diversification process, BAMM allows for an investigation of rate shifts that vary through time or in a diversity-dependent manner. Like MEDUSA, it allows for automatic detection of rate shifts that are maximally supported by the data without any *a priori* specifications as to their location within the tree. Like in the MEDUSA example, we also used our tree as a genus level tree by assigning the terminal tips to have the number of species in that representative's genus.

A poissonRatePrior priors of one failed to reach convergence, so a value of 0.3 was used. Other priors were selected using BAMMtools::setBAMMpriors, with the rest remaining at default settings. Samples were run for 10,000,000 generations. Convergence and stabilization of the run was visualized and confirmed by looking at the MCMC likelihood output after a 10% burn-in was discarded. The *effective sample sizes* of the log-likelihood estimate of the number of shifts and the log-likelihood were > 300 in both cases.

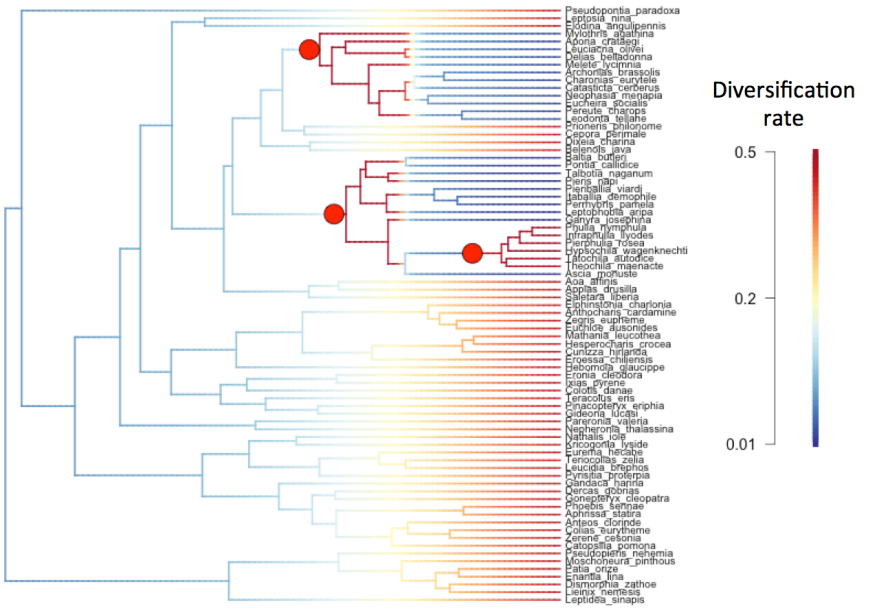


Figure S4.5: Diversification Rate Shifts

While Bayes Factor analysis found strong support for a model with 5 rate shifts, with a Bayes Factor = 90, in comparison to the null model that assumed 2 rate shifts (which is a result of using the poissonRatePrior = 0.3), identification of the best shift configuration set using default settings (getBestShiftConfiguration) identified 3 credible shifts.

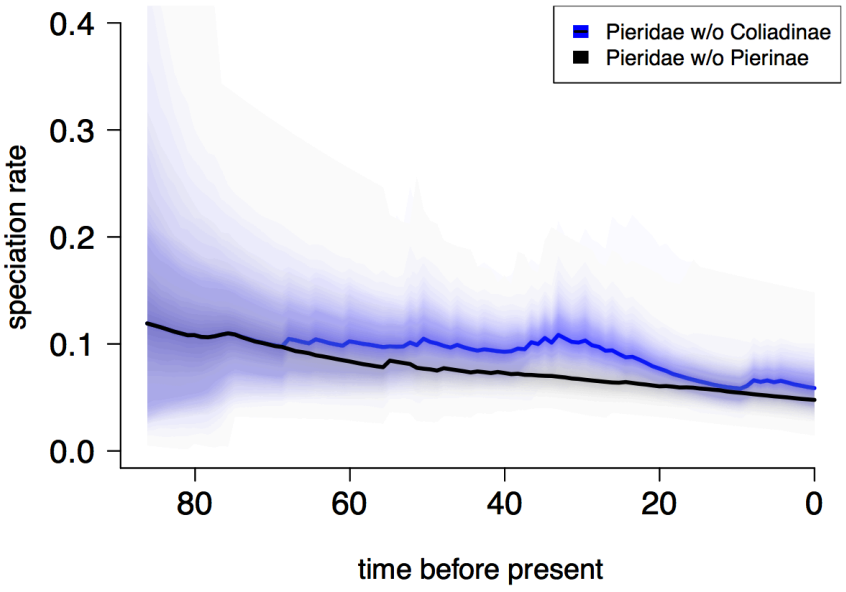


Figure S4.6: Rate Through Time Matrix

In order to estimate the changes in diversification through time in more detail, the getRateThroughTimeMatrix command was invoked. This creates a matrix of diversification rates through time, allowing for the inclusion or exclusion of specific nodes. The difference between matrices estimates for Pieridae without Coliadinae and Pieridae without Pierinae was then plotted through time and is shown in Figure 2A.

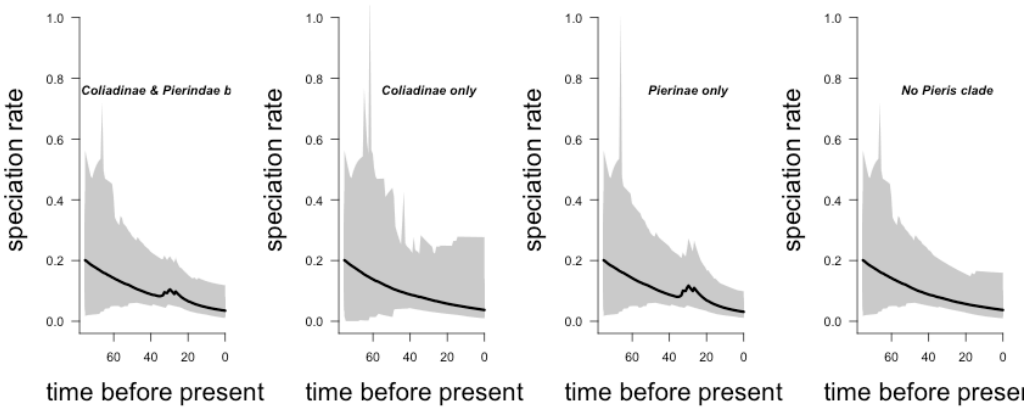


Figure S4.7 Coliadinae and Pierinae speciation rate comparison

In order to assess the potential impact of having non-GS feeding Pierinae in the analysis, we also investigate these dynamics in another analysis at the level of directly comparing only Coliadinae vs. Pierinae, with only Brassicales feeders in the latter. This plot is shown above. Only datasets including Pierinae have the peak in diversification rate.

Supplementary Text S5

S5.1 Nitrile Specificer Protein (NSP) genomics and evolutionary dynamics

The nitrile specifier gene (NSP) has been shown to be the primary detoxification mechanism used by Pierinae butterflies to break down the glucosinolate chemical defense system of their Brassicales hostplants (141). Analysis of the protein sequence of NSP revealed a motif that was repeated three times across the length of the gene (142) (Figure S5.1).

Further analysis of transcriptome data, exon-intron boundaries in the genome, and database searching revealed the following. NSP evolved from an ancestral gene having only one domain. Within Lepidoptera the one domain gene is found in all species studied to date, but only the Pierinae (the Brassicales feeding Pieridae) have a multiple domain version. Within the Pierinae, we know there are two copies of the NSP gene in *Pieris rapae*, at least one of which is active against the glucosinolates used in the functional assay.

With this understanding, there were a series of open questions that the manuscript here worked to address. First, we wanted to know how the NSP gene family was evolving since it appeared at the base of the Pierinae roughly 70 Mya. Since two copies were found in *P. rapae*, we specially wanted to know how often other species had paralogous copies of NSP, as well as the evolutionary history among these copies. Second, we wished to investigate the functional performance of these different copies of NSP. While previous work had demonstrated the connection between NSP function and a single protein sequence of *P. rapae* (143), and other work had demonstrated NSP enzymatic activity in the gut of various Pierinae butterflies (141), there was no understanding of the connection between NSP gene sequence and enzymatic activity. Stated another way, we needed to know if most species had only one copy of NSP or several, and whether or not those copies differed in their detoxification performance. Finally, we wanted to know whether the evolutionary history of these genes showed any molecular signatures of positive selection, as this would be consistent with an important role in coevolutionary dynamics.

These findings are summarized in the figure below, which is complementary to their reporting in the main text of the paper (Figure 1). Within *P. rapae*, the two paralogous copies of NSP differ in the performance in the glucosinolate-myrosinase assay, with the NSP version most closely related to *Anthocharis cardamines* showing divergent activity. This is reflected in similar performance of gene copies from these lineages in *P. brassicae* and *P. napi*. However, the two paralogous copies in *P. rapae* have identical expression pattern across tissues and in response to starvation, suggesting that this divergent function is likely still NSP activity, but that our assays are not able to currently detect this complexity (Figure S5.3). While both of the two divergent alleles of *A. cardamines* have NSP activity, they differ in their performance on two different glucosinolate compounds, further highlighting this dynamic of divergent activity. Supporting the importance of NSP genes in detoxifying glucosinolates, our investigation of the genome of a Pierinae species that stopped feeding on Brassicales nearly 35 Mya indicates that all NSP gene sequences are absent; only the single domain ancestral gene remains present in the genome. Thus, while NSP gene paralogs differ in their glucosinolate detoxification, maintenance of the NSP gene family within Pierinae genomes requires feeding upon glucosinolates.

Biological Materials:

C. eurytheme, *P. daplidicae, A. cardamines*, *G. rhamni, Pieris rapae*, *P. brassicae,* and *P. napi* butterflies were collected in Jena (Germany). *P. protodice* larvae were field collected in Missoula (Montana, USA), larvae of *B. creona*, *B. gidica* and *D. pigea* were field collected in the Cape region of South Africa and *E. socialis* larvae were collected in Mexico and were dissected and used for extraction of RNA and DNA.

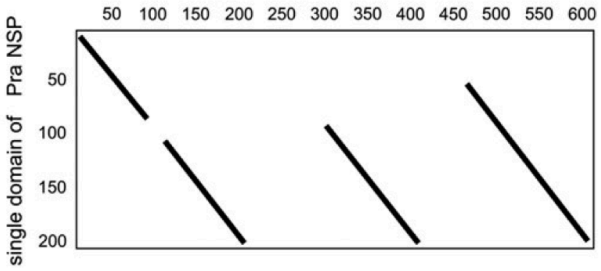


Figure S5.1 Repeated domain motif in the nitrile specifier protein

On the X axis is the full protein sequence of *P. rapae* NSP, while on the Y axis is a single domain of a second paralogous copy of NSP. Lines are dot plots indicating amino acid identity with numbers indicating length of amino acid sequence along axis. Figure taken from (142)).

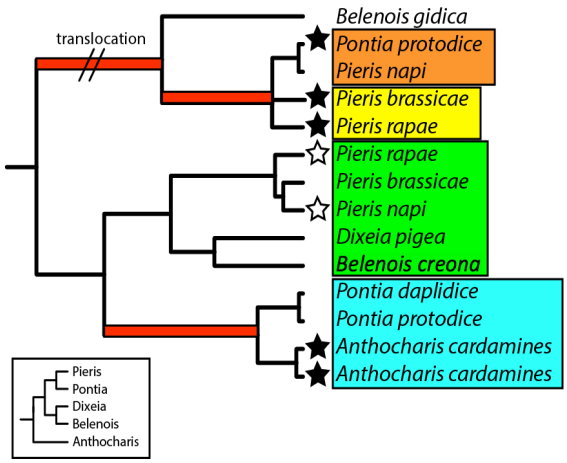


Figure S5.2. Phylogeny of nitrile specifier genes and pierid species

Independent genetic loci that were inferred (Figure S5.1) have their species name shaded in different colors and a translocation event indicated. Stars by species names indicate genes for which enzyme constructs were made and their function assayed. Those with detected NSP activity are black filled. Branches colored red indicate branches of the tree where specific codons have been detected to have undergone diversifying selection (Table S5.3). All nodes have > .9 posterior probability support (Where is this described). Boxed insert indicates the species level relationship among these genera resolved by the phylogenomic analysis (Figure 1).

S5.2 RNA and DNA extractions, dscDNA generation and preparation of cDNA libraries

RNA was extracted with TRIzol Reagent (Invitrogen) according to the manufacturer's protocol and pooled from 3rd and 4th instar larvae. An additional DNAse (Turbo DNAse, Ambion) treatment was included prior to the second purification step to eliminate any contaminating DNA. The DNAse enzyme was removed and the RNA was further purified by using the RNeasy MinElute Clean up Kit (Qiagen) following the manufacturer's protocol and eluted in 20 µl of RNA Storage Solution (Ambion). This second purification step was performed to eliminate contaminating polysaccharides, proteins and the DNAse enzyme. RNA integrity and quantity was verified on an Agilent 2100 Bioanalyzer using the RNA Nano chips (Agilent Technologies, Palo Alto, CA). RNA quantity was determined on a Nanodrop ND-1000 spectrophotometer. For the isolation of genomic DNA the abdomens of adult butterflies were ground to a fine powder in liquid nitrogen and DNA was isolated using the genomic tip 20/G and genomic DNA buffer kit following the manufacturer's protocol (Qiagen). For the generation of the dscDNA for454 GS-FLX sequencing, RNA from larval tissue material of *A. cardamines*, *D. pigea*, *B. gidica* and *G. rhamni* was used to generate full-length enriched, non-normalized cDNA libraries using a combination of the MINT cDNA synthesis kit (Evrogen) and the Trimmer Direct cDNA kit (Evrogen) using a polyT primed method generally following the manufacturer's protocol. Each step of the cDNA library generation procedure was carefully monitored to avoid the generation of artifacts and overcycling.

Additional cDNA libraries were generated for *P. daplidicae, P. protodice*, *P. rapae, P. brassicae*, *A. cardamines*, *G. rhamni, L. sinapis* and *E. socialis*. Double-stranded, full-length enriched cDNA from dissected larvae were generated by primer extension with the SMART cDNA library construction kit (Clontech) according to the manufacturers protocol but with several modifications. In brief, 2 µg of poly(A)+ mRNA was used for each cDNA library generated. cDNA size fractionation was performed with SizeSep 400 spun columns (GE Healthcare) that resulted in a cutoff at ~200 bp. The full-length-enriched cDNAs were ligated to the pDNR-Lib plasmid vector (Clontech). Ligations were transformed into *E. coli* DH5α-E electro-competent cells (Invitrogen). Furthermore, for *C. eurytheme, A. cardamines, P. rapae, P. brassicae* normalized full length-enriched cDNA libraries were generated using a combination of the SMART cDNA library construction kit and the Trimmer Direct cDNA normalization kit (Evrogen) following the manufacturer's protocol but with several modifications. In brief, reverse transcription was performed with a mixture of several reverse transcription enzymes (ArrayScript, Ambion; BioScript, Bioline; PrimeScript, TaKaRa; SuperScript II, Invitrogen) for 1h at 42 °C and 90 minutes at 50 °C. The normalization procedure was carefully monitored to avoid any overcycling of the resulting cDNAs. The full-length-enriched, normalized cDNAs were cut with SfiI and ligated to the SfiI-digested pDNR-Lib plasmid vector as described above.

S5.3 Sanger sequencing and generation of EST databases

Plasmid minipreparation from bacterial colonies grown in 96 deep-well plates was performed using the 96 robot plasmid isolation kit (Nextec) on a Tecan Evo Freedom 150 robotic platform (Tecan). Single-pass sequencing of the 5' termini of cDNA libraries was carried out on an ABI 3730 xl automatic DNA sequencer (PE Applied Biosystems). Vector clipping, quality trimming and sequence assembly was done with the Lasergene software package (DNAStar Inc.). We set up individual searchable databases for each of the species and used them to identify the genes we

describe in more detail in the text. Blast searches were conducted on a local server using the National Center for Biotechnology Information (NCBI) blastall program. Homology and gene ontology (GO; www.geneontology.com), enzyme classification codes (EC) and metabolic pathway analysis of the assembled sequences were determined using the BLAST2GO software (www.blast2go.de). Sequences were searched against the NCBI non-redundant (nr) protein database using an E-value cut-off of 10-3, with predicted polypeptides of a minimum length of 18 amino acids.

S5.4 2nd Generation Sequencing, assembly and candidate gene identification

We selected four key species (*A. cardamines*, *Dixeia pigea*, *Belenois gidica*, *Gonepteryx rhamni*) for the generation of non-normalized cDNAs and ultra-deep sequencing by 454 GS-FLX (Table S5.1). For transcriptome analysis of these four Pieridae species we used double-stranded cDNA, generated from total RNA extracted from 10 individuals each, generated by primer extension with the SMART cDNA library construction kit (Clontech) according to the manufacturers protocol but with several modifications mentioned above. The resulting dscDNA was sheared and 500-800 base pair long fragments were recovered after size selection. Using these fragments we constructed sequencing libraries for the Roche 454 machine according to the manufacturers protocols. For each library we used a quarter of a Titanium chemistry sequencing run yielding more than 50 Mb sequence information for each library. The reads were assembled using the Newbler assembler with standard settings and option "–large". Results of these runs is reported (Table S5.1)

In addition to transcriptome sequencing we performed paired-end Illumina (Solexa) sequencing on sheared genomic DNA fragments of *Belenois creona* and *Delias nigrina*. In order to construct genomic paired end libraries for sequencing using the Illumina machine, total genomic HMW DNA was sheared to yield a mean fragment size of ~200 bases. The sequencing libraries were generated according to standard Illumina protocols. Two lanes each were loaded with a library from one species and sequenced in both directions. This yielded 2.1 Gb for *Dixeia pigea* and 2.3 Gb for *Belenois creona* genomic DNA (Table S5.1). The corresponding read pair sequences were connected to one fragment with a spacer of 10 Ns between the individual sequences in order to allow the generation of combined paired-end reads into one contiguous sequence. This allowed for increased efficiency when generating BLASTable databases and our searches for both candidate and housekeeping (e.g. ribosomal protein) genes.

We assessed our 454 and Illumina coverage through estimating the number of hits against the complete ribosomal protein dataset of *B. mori* and by estimating the number of hits against the Unigene dataset (14,623) from the genome assembly of *B. mori* V2 (Table S2). Both the 454 and Illumina sequencing provided deep coverage of the transcriptome and genome, respectively, with the Illumina reads providing an estimated 5-fold genome coverage. We set up individual searchable databases for each of the species and used them to identify the genes we describe in the text. For the NSP, NSP-like and 1D-NSP-like gene searches, the following sequences were used: *P. rapae*, *P. brassicae*, *P. napi*, *P. daplidice* NSP, NSP-like, 1d-NSP-like; *A. cardamines* NSP-like, 1d-NSP-like; *G. rhamni*, *C. eurytheme* 1d-NSP-like (142).

Table S5.1 NSP genomic and transcriptomic datasets

Transcriptome and genome sequence data for selected Pieridae used in this study. The estimated sequence coverage is based on a combination of the number of hits (RibProt Species) against the complete ribosomal protein dataset of *B. mori* (RibProt *B. mori*) and by estimating the number of hits against the Unigene dataset from the genome assembly of *B. mori* V2.

Pieridae Species	Host plant family	Source material (Sequencing method)	Total no. of HQ reads	Average read length	Total no. of contigs	RibProt Species / RibProt *B. mori*	Estimated sequence coverage
Belenois gidica	*Capparidaceae*	cDNA (454 FLX)	222.531	315 bp	2.890	79/79	> 90%
Belenois creona	*Capparidaceae*	Genomic DNA (Illumina)	13.989.071	76 bp	n.a.	79/79	5fold
Dixeia pigea	*Capparidaceae*	cDNA (454 FLX)	184.049	320 bp	4.783	78/79	> 90%
Anthocharis cardamines	*Brassicaceae*	cDNA (454 FLX/Sanger)	163,462/5,572	378 bp	6,364/2,818	78/79	> 90%
Pieris rapae	*Brassicaceae*	cDNA (Sanger)	25.850	815 bp	10.943	77/79	> 90%
Pieris brassicae	*Brassicaceae*	cDNA (Sanger)	6.594	790 bp	4.359	72/79	> 85%
Delias nigrina	*Loranthaceae*	Genomic DNA (Illumina)	12.823.030	76 bp	n.a.	79/79	5fold
Delias nigrina	*Loranthaceae*	cDNA (Sanger)	5.412	820 bp	3.126	62/79	> 70%
Goneptery rhamni	*Fabaceae*	cDNA (454 FLX/Sanger)	221,451/2165	375 bp	7,329/718	79/79	> 90%
Colias eurytheme	*Fabaceae*	cDNA (Sanger)	9.140	835 bp	4.577	74/79	> 90%
Pontia spp.	*Brassicaceae*	cDNA	1.799	785 bp	655	59/79	>55%

S5.5 Fosmid library generation

Genomic DNA was isolated from several pupae of *Pieris rapae*, using the genomic tip 500/G isolation kit (Qiagen) as described above. Genomic DNA quantity was measured photospectrometrically on a Nanodrop ND1000 and DNA quality and size was checked by pulsed-field gel electrophoresis on a CHEF Mapper XA (Bio-Rad). The genomic DNAs were sheared to the desired 40 kb range with a Hydroshear device (Molecular Devices). For the generation of the fosmid libraries ~ 3 µg of sheared genomic DNA was used as starting material in a CopyControl fosmid library production kit protocol (Epicentre), resulted in a library of *E. coli* EPI300 clones, each carrying a ~ 40 kb DNA fragment in the pCC1FOS vector. Appr. 28000 colonies for each species were picked into 384well microtiter plates with a QPix II robotic colony picker (Genetix) and subsequently spotted onto large Performa II nylon membranes (Genetix). Colony picking, replicating, membrane spotting and quality testing was performed by the RZPD (German Resource Center for Genome Research). The library was stored as −80°C glycerol stocks in 384well microtiter plates. The randomly picked (n = 28000) clones represent a 2-3 fold genome coverage, assuming a genome size G of 450 Mbp and an insert size i of 40 kb.

A first quality check of the library for DNA insert size and clone diversity was done by restriction analysis of fosmid DNA isolated from twelve randomly selected library clones for each library. This revealed a total of 24 different restriction patterns and an average insert size of 34 kb. Overnight cultures of *E. coli* EPI300 clones were diluted 10× in LB containing 12.5 µg ml^{-1} chloramphenicol and 1× induction solution (Epicentre) and incubated for 5 h at 37°C, 300 rpm. Fosmids were isolated with the Nucleobond Xtra Midi Kits according to the manufacturers´ instructions (Macherey-Nagel). Fosmid library nylon filters were washed, blocked and hybridized with horseradish peroxidase (HRP)-labeled DNA fragments containing the Pieris 1d-NSP-like, NSP-like and NSP genes. Labeling, hybridization and probe detection were done according to specifications in the ECL DNA labeling and detection kit (GE Healthcare).

S5.6 Genomic localization and orientation of NSP, NSP-like and 1d-NSP-like genes

Sequencing of fosmid library clones and blasting the chromosomal area against the NCBI databases (http://www.ncbi.nlm.nih.gov/) allowed us to identify the neighboring genes of NSP and NSP-like in the genome of *P. rapae*. By blasting those ORF (open reading frame) regions against the *B. mori* genome assembly (http://sgp.dna.affrc.go.jp/KAIKO/) we gained the most likely paralogous genes of the NSP and NSP-like neighboring genes in the *B. mori* genome. The existent strong microsynteny between lepidopteran genomes allowed us to draw some conclusion about the genomic localization of the NSP gene family members in e.g. the *P. rapae* genome (Figure S5.1).

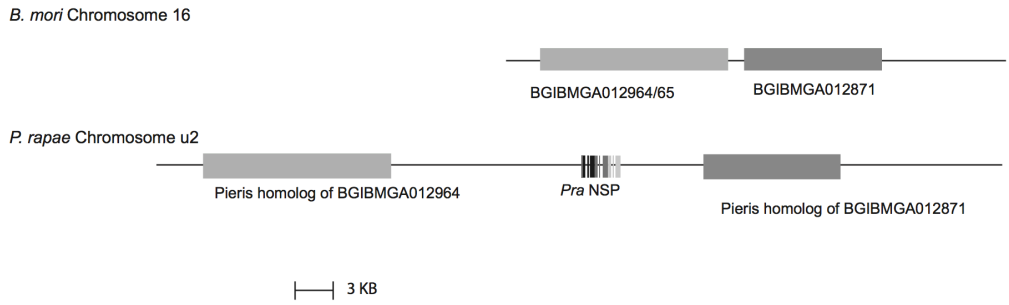


Figure S5.3 Genomic organization analysis of NSP genes

Genomic orientation and structure of the *P. rapae* 1d-NSP-like, NSP-like, NSP and flanking genes relative to the *B. mori* chromosomes. Gene abbreviations and numbers refer to the *B. mori* genome assembly (http://sgp.dna.affrc.go.jp/KAIKO/). Genomic orientation of the NSP-like gene to the *P. rapae* homolog of *B. mori* BGIBMGA005023 and the assumed localization of 1d-NSP-like relative to the *B. mori* ORF is depicted. Shown is the relative orientation of NSP to its flanking genes based on *P. rapae* fosmid sequences (A). Genomic orientation of BGIBMGA012964 and ORF BGIBMGA012871 in the *B. mori* genome and below the relative orientation of the *P. rapae* homologs of those genes to the NSP locus (B).

S5.7 qRT-PCR

P. rapae larvae were either starved for 30 hours or provided with plant material and adults were collected from lab-reared cultures. Gut and rest of body of the larvae were dissected and stored and complete *P. rapae* adults were directly shock-frozen in in liquid nitrogen. For both the larvae and the adults, two biological replicates were conducted with three larvae or three adults pooled for this experiment, respectively. Out of each pool 500 ng of DNA-free total RNA was converted into single-stranded DNA using a mix of random and oligo-dT20 primers according to the ABgene protocol (ABgene). Real-time PCR oligonucleotide primers were designed using the online Primer3 internet based interface (http://frodo.wi.mit.edu). Gene-specific primers were designed on the basis of sequences obtained from *P. rapae* and two additional genes as potential house-keeping genes (ribosomal protein subunit 18S and elongation initiation factor 4 a) to serve as the endogenous control (normalizer). Both house-keeping gene primers were thoroughly tested for linearity and uniformity. RPS18 was the most consistent gene, and subsequently used for further analyses. QRT-PCR was done in optical 96-well plates on a MX3000P Real-Time PCR Detection System (Stratagene) using the Absolute QPCR SYBR green Mix (ABgene) to monitor double-stranded DNA synthesis in combination with ROX as a passive reference dye included in the PCR master mix. For analysis the qBase software package for the automated analysis for real time quantities PCR data was used (http://www.genequantification.de). Expression of each gene in each tissue was calculated relative to the control gene RPS18. For each gene the lowest expression was then set to one and expression in all other tissue was set relative to that.

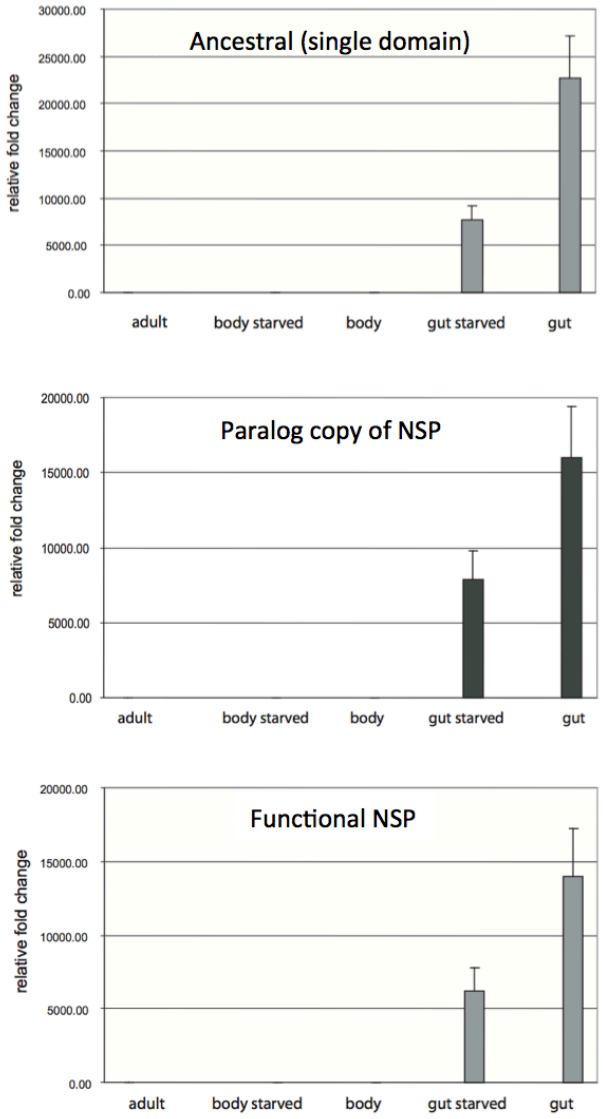


Figure S5.4 qRT-PCR of NSP and other genes in P. rapae larvae and adults

Pieris rapae larvae were either fed on plant material or starved for 30 hours and expression was measured separately for gut and rest body tissue. Two biological replicates were conducted on 3 larvae or adults, respectively, pooled for RNA extraction, subsequent cDNA synthesis and qRT-PCR.

S5.8 Heterologous expression and enzyme assays

Members of the NSP gene family, from *A. cardamines* and *P. rapae*, were amplified from cDNA clones by PCR using gene-specific primers and were inserted in pIB/V5-His TOPO (Invitrogen) in frame with the C-terminal His-tag by TA cloning according to the supplier's instructions. Positive clones were selected and correct insertion was confirmed by sequencing on an ABI 3730 xl automatic DNA sequencer (PE Applied Biosystems). Cells were transformed with Insect GeneJuice (Novagen) and the respective plasmids and after 48 h both cells and growth media were harvested and extracted in Insect Pop Culture Protein Extraction Reagent (Novagen) with 1 µL of benzonase (Novagen) and 10 µL of Proteinase Inhibitor Cocktail (Thermo Scientific) per ml of extraction volume. Assays were essentially done as described in (1) in 50 mM Tris-HCl, pH 7.5, containing 2 mM benzylglucosinolate or 4MSOB, respectively and an appropriate amount of protein extract or fraction thereof in a total volume of 500 µl at room temperature. The reaction was started by addition of myrosinase (Sigma-Aldrich) to a final concentration of 50 µg/ml. At the end of the incubation time the assay mixtures with benzylglucosinolate were stopped and extracted with 1 ml dichloromethane and analyzed by GC-MS and GC-FID. Assays with 4MSOB were heat inactivated at 70°C for 30 min and loaded onto an LC-MS.

Table S5.2 Heterologous expression results of NSP gene copies

All constructs except the *P. rapae* NSP from the *A. cardamines* lineage have detectable NSP activity, with the two alleles of *A. cardamines* differing in their performance on the two glucosinolates indicated below.

	4MSOB			Benzylglucosinolates		
	CN	ITC		CN	ITC	
	Peak area	Peak area	ratio CN/ITC	Peak area	Peak area	ratio CN/ITC
pIB vector control cells	19350797	633969649	0.03	4.64	356.06	0.01
pIB vector control media	15652705	637434244	0.02	3.26	333.21	0.01
A. cardamines NSP-allele1 cells	25188524	821876399	0.03	84.83	426.27	0.20
A. cardamines NSP-allele1 media	50861344	626424003	0.08	431.35	120.58	3.58
A. cardamines NSP-allele2 cells	46323568	767058745	0.06	34.38	476.52	0.07
A. cardamines NSP-allele2 media	94367727	569858472	0.17	103.55	302.29	0.34
P. rapae NSP(A.card_lineage) cells	18386155	788753567	0.02	4.47	783.98	0.01
P. rapae NSP(A.card_lineage) media	16377117	631376330	0.03	2.48	296.54	0.01
myrosinase only	0	0		0	0	
positive control (*A. cardamines* gut homogenate)	391282449	461251972	0.85	891.37	4.57	195.04

Peaks areas for 4MSOB were extracts of ion traces in LCMS. Peak area of Benzylglucosinolates were measures by GC-FID.

S5.9 Evolutionary analysis of NSP genes: detection of positive selection

In order to test the hypothesis that diversifying selection was acting upon the branches leading to the functional copies of NSP in the *Pieris* and *Anthocharis* lineages, a maximum likelihood analysis of codon evolution was conducted: a branch-site model was implemented using the CODEML package of the PAML software suite (96). This model allows for dN/dS, or ω, to vary both among sites in the protein and along specific branches of a given phylogenetic tree. Two separate analyses were conducted, each focusing upon the specific branch leading to either the *Pieris* or *Anthocharis* functional NSP. The branch-site test of positive selection conducts a likelihood ratio test of Model A against a null model. Model A estimates the frequency of two ω values over 4 site classes, for all the branches other than the specified branch (background), and then the specified branch (foreground). Importantly, the two values of ω are $0< \omega < 1$, or $\omega = 1$ in the background while in the foreground, an extra ω is allowed which can have values ≥ 1. A likelihood ratio test of the Model A vs. null model is then conducted, with comparisons of one degree of freedom in a X^2 test (PAML manual). Note that only full length genes were included in the analysis, and thus *Belenois gidica* and *B. creona* as seen in Figure S4.5 were not included in this PAML analysis.

Table S5.3. Codons identified as having experienced positive selection

Two butterfly clades are listed showing the P-value results of a model allowing $\omega > 1$ on specific branches leading to either the *Pieris* or *Anthocharis* lineage of functional NSP, denoted with the *Pieris* lineage or *Anthocharis* lineage, respectively. Codons identified as having $\omega > 1$, with an individual P-value < 0.05 in either lineage, are indicated.

Foreground branch	2 * lnL diff.	Pvalue	Sites under selection
Pieris lineage	17.15	9.1E-06	33, 86, 300, 490
Anthocharis lineage	21.42	9.6E-07	20, 68, 96, 157

Supplementary Text S6

Comparative genomic analyses to investigate gene loss patterns following At-a event

Ancient whole genome duplications (WGDs) are ubiquitous across eukaryotic kingdoms, having occurred at major evolutionary transitions including near the origins of the angiosperms (86), vertebrates (144), and teleost fishes (145). A common feature of polyploid lineages is that they are more speciose (species-rich) compared to their sister lineages, though the mechanisms responsible for greater species diversity remains a central question in evolutionary biology (146). The primary mechanism(s) that may drive diversification following WGD varies across kingdoms, and may involve prezygotic isolation following the origin of novel traits in teleost fishes (147) and hybrid incompatibility leading to postzygotic isolation in yeast (148). In the polyploid yeast lineage, speciation was driven by hybrid incompatibility caused by a version of the Bateson-Dobzhansky-Muller model, involving the reciprocal loss of duplicated genes across homoeologous regions among different populations (148). In contrast, the amazing biodiversity of teleost fishes, which represent nearly half of all vertebrates (~29,000 species), was not due to a hybrid-incompatibility mechanism (147). Rather, the teleost radiation was likely driven by the origin of novel pigmentation and cognitive traits that are encoded by novel genes that arose due to the WGD (149). The reciprocal gene loss mechanism has been proposed as a possible driver for speciose plant clades sharing whole genome duplications (89, 150, 151). However, the actual mechanisms driving diversification following WGDs in plants remains poorly understood.

Here, we investigated whether the most recent and largest species radiations in the Brassicales was caused by a yeast-like reciprocal gene loss (hybrid incompatibility) mechanism following At-α (Figure 1). Our analyses of the *Arabidopsis thaliana* (50) and *Aethionema arabicum* (92) genomes revealed that the family shares the duplication status for the majority of At-α duplicates, which includes ~53% of duplicates having returned to single copy prior to the divergence of the two earliest diverging lineages and ~21% of all genes still retained in duplicate in both genomes (Table S6.1; Figure S6.1). These results suggest that the majority of duplicate genes were lost relatively rapidly after At-α. The species-poor *Aethionema* clade, which includes only 45 species, lost an additional 2156 lineage specific duplicates (~18% At-α duplicates). The *Arabidopsis* genome belonging to the species-rich group, which includes 3615 species, has only 939 lineage specific duplicate losses (~8% At-α duplicates). If the primary mechanism driving speciation following At-α was reciprocal gene loss, the species-rich group would be expected to have lost far more lineage specific duplicates.

The most recent mass diversification occurred at the base of the *Arabidopsis* lineage and is not shared by the species-poor tribe Aethionemeae lineage with far more lineage-specific duplicate losses (Figure 1). Additionally, we did not find a single ancestral locus that exhibited a reciprocal gene loss pattern between *Aethionema* and *Arabidopsis*. Thus, these results suggest that a genetic hybrid incompatibility mechanism following the At-α was likely not a major contributor to this species radiation. Instead, the radiations following At-α were very likely driven due to other reasons including novel chemical defenses, following the evolution of novel biosynthetic steps (Supplementary Text S4), which triggered an adaptive radiation of the core Brassicaceae. This conclusion as to the possible mechanism driving the radiation is reinforced by our genomic analyses rejecting reciprocal gene loss as a major contributor to the most recent and largest radiation.

The origin of novel traits following gene and whole genome duplications, with the events giving rise to the novel gene functions encoding these traits, has been observed for teleost fishes (149) and yeasts (152). Major diversification rate shifts were also observed at the base of both yeast and teleost fishes (147, 148). Similarly across land plants, two ancient whole genome duplications likely gave rise to the origin of the seed and flower via novel developmental pathways (86). Recent studies show that the origins of many novel pathways, particularly the upstream dosage-sensitive regulators, involved in complex traits, are highly dependent on WGDs and likely would not arise following a series of smaller-scale duplications (e.g. tandem or segmental) due to the reduced fitness of the required evolutionary intermediates resulting from stoichiometric imbalances produced by single-gene duplications (153, 154). Thus it has been argued that these events have played a key role in driving macroevolutionary transitions by providing the building blocks for increases in morphological complexity (154).

Our results provide evidence that the origin of novel traits, not Bateson-Dobzhansky-Muller – type hybrid incompatibilities due to duplicate gene loss patterns, spurred adaptive co-radiation events for Brassicales species and their specialist insect herbivores over the last 80 million years. In part, these radiations likely occurred as a result of open environmental niches, including those vacated by mass extinction events (155) (e.g. Creataceous-Tertiary event) and prolonged unstable climatic conditions (156), and new niches permitted by the evolution of novel traits. We also identified a substantial time-lag between both WGDs and the subsequent radiations (146), which was similarly observed for the teleost radiation (147). These time-lag events likely reflect the time required to evolve a novel trait from initially fully redundant duplicated ancestral pathways. Collectively, these studies suggest that the origin of novelty derived from whole genome duplications likely spurred successful lineages across Brassicales.

S6.1 Comparative genomic analyses of Arabidopsis and Aethionema

The two genomes were aligned against themselves using LASTZ (157). The primary alignments of regions against themselves were removed, as were alignments to regions within 100KBase of each other representing local duplications. The secondary alignments that were left were chained together and the highest scoring chains for each region of the genome were selected. These chains were netted to generate candidate extended regions with high similarity within genomes. For each gene model in *Aethionema* and *Arabidopsis*, the chains were used to liftover internally each gene to the coordinates where a duplicate gene should be located if that region had been retained. The location of the duplicate gene was checked for the presence of a gene model and if one was found, the protein sequence of the two models were compared. Where a blastp generated an E value <.01, the two genes were associated as candidate paralogs. The same approach was undertaken between *Arabidopsis* and *Aethionema* to generate candidate orthologs. The two lists were then merged to create Supplemental Table S6.1.

To further validate these results, a random set of ten syntenic regions that have remained duplicated in both species with at least 30 gene loss events in either species were manually screened for Reciprocal Gene Loss (RGL) (i.e. the loss of alternate duplicate copies across homoeolous regions). These genomic regions are distributed across all five *Arabidopsis* chromosomes. No RGL was discovered in any of these ten regions (0 out of 300 lost duplicates). Thus, RGL occurred at less than 0.34% of ancestral At-α duplicated loci. Based on the estimated

ancestral gene content size of ~14,800 genes for *Arabidopsis* prior to the At-α event (158), a maximum of 50 loci may have undergone RGL. These few loci would not account for the rich species diversity observed for the Brassicaceae (Figure 1).

Table S6.1: [Excel table] Summary of *Arabidopsis* and *Aethionema* orthologs and retained paralogs from the At-α event

Aethionema and *Arabidopsis* gene names are shown in Columns A & B and D and E, respectively. Those gene pairs supported by syntenic analyses are indicated in Column F.

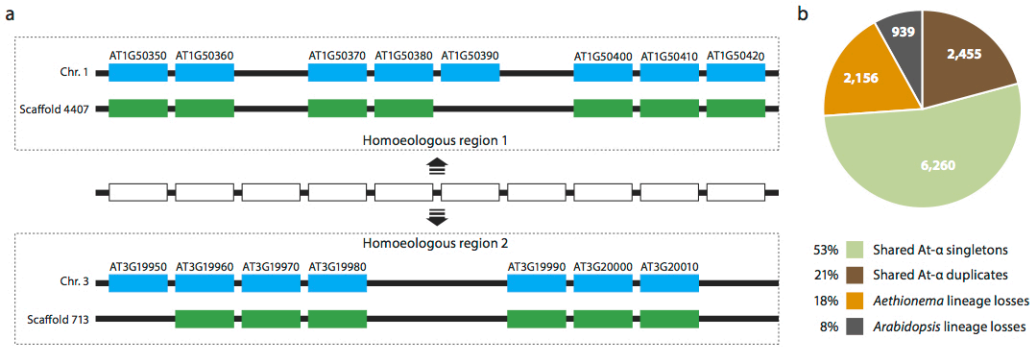


Figure S6.1 – Genome wide analyses of gene loss patterns following the At-α whole genome duplication (WGD) in *Arabidopsis* and *Aethionema*

A. Comparison of *Arabidopsis* and *Aethionema* homoeologous genomic regions. The *Arabidopsis* and *Aethionema* genomes represent the two earliest diverging lineages in the Brassicaceae following the At-α WGD. Here, homoeologous regions, encoded on *Arabidopsis* chromosomes 1 & 3 and *Aethionema* scaffolds 4407 & 713, are illustrated to show genome wide patterns. The ancestral pre-duplicated genomic region is centered showing ten gene models (white boxes). The homoeologus regions duplicated by the At-α event are shown above and below each ancestral gene model for *Arabidopsis thaliana* (blue models) and *Aethionema arabicum* (green models). Gene loss is indicated with missing gene models. Shared duplicate gene losses and retention between these two genomes is the most commonly observed state. B. Genome Wide Summary of At-α Duplicates. A total of 11,810 ancestral loci were identified between the *Arabidopsis* and *Aethionema* genomes that are supported by synteny (Supplemental Table S6.1). Most of these loci (~74%) are either shared as singletons due to gene loss or retained in duplicate in both genomes, 18% duplicates were lost in the *Aethionema* lineage only, and 8% duplicates were lost unique to the *Arabidopsis* lineage.

References:

30. Magallón S, Crane PR, & Herendeen PS (1999) Phylogenetic pattern, diversity, and diversification of Eudicots. *Ann Missouri Bot Gard* 86(2):297-372.
31. Rodman JE, Karol KG, Price RA, & Sytsma KJ (1996) Molecules, morphology, and Dahlgren's expanded order capparales. *Syst Bot* 21(3):289-307.
32. Hall JC, Iltis HH, & Sytsma KJ (2004) Molecular phylogenetics of core Brassicales, placement of orphan genera Emblingia, Forchhammeria, Tirania, and character evolution. *Syst Bot* 29(3):654-669.
33. Hall JC (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.
34. Soltis DE*, et al.* (2011) Angiosperm phylogeny: 17 genes, 640 taxa. *Am J Bot* 98(4):704-730.
35. Rodman JE, Soltis PS, Soltis DE, Sytsma KJ, & Karol KG (1998) Parallel evolution of glucosinolate biosynthesis inferred from congruent nuclear and plastid gene phylogenies. *Am J Bot* 85(7):997-1006.
36. Hall JC, Sytsma KJ, & Iltis HH (2002) Phylogeny of Capparaceae and Brassicaceae based on chloroplast sequence data. *Am J Bot* 89(11):1826-1842.
37. Ronse De Craene LP & Haston E (2006) The systematic relationships of glucosinolate-producing plants and related families: A cladistic investigation based on morphological and molecular characters. *Bot J Linn Soc* 151(4):453-494.
38. Su JX, Wang W, Zhang LB, & Chen ZD (2012) Phylogenetic placement of two enigmatic genera, Borthwickia and Stixis, based on molecular and pollen data, and the description of a new family of Brassicales, Borthwickiaceae. *Taxon* 61(3):601-611.
39. Olson ME (2002) Combining data from DNA sequences and morphology for a phylogeny of Moringaceae (Brassicales). *Syst Bot* 27(1):55-73.
40. Olson ME (2002) Intergeneric relationships within the Caricaceae-Moringaceae clade (Brassicales) and potential morphological synapomorphies of the clade and its families. *Int J Plant Sci* 163(1):51-65.
41. Karol KG, Rodman JE, Conti E, & Sytsma KJ (1999) Nucleotide sequence of rbcL and phylogenetic relationships of Setchellanthus caeruleus (Setchellanthaceae). *Taxon* 48(2):303-315.
42. Chandler GT & Bayer RJ (2000) Phylogenetic placement of the enigmatic western Australian genus emblingia based on rbcL sequences. *Plant Spec Biol* 15(1):67-72.
43. Grabherr MG*, et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29(7):644-652.
44. Li L, Stoeckert Jr CJ, & Roos DS (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res* 13(9):2178-2189.
45. Wang X*, et al.* (2011) The genome of the mesopolyploid crop species Brassica rapa. *Nat Genet* 43(10):1035-1039.
46. Hu TT*, et al.* (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat Genet* 43(5):476-483.
47. Xu X*, et al.* (2011) Genome sequence and analysis of the tuber crop potato. *Nature* 475(7355):189-195.

48. Sato S, *et al.* (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485(7400):635-641.
49. Hellsten U, *et al.* (2013) Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. *Proc Natl Acad Sci U S A* 110(48):19478-19482.
50. Kaul S, *et al.* (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408(6814):796-815.
51. Dassanayake M, *et al.* (2011) The genome of the extremophile crucifer Thellungiella parvula. *Nat Genet* 43(9):913-918.
52. Ming R, *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature* 452(7190):991-996.
53. Argout X, *et al.* (2011) The genome of Theobroma cacao. *Nat Genet* 43(2):101-108.
54. Tuskan GA, *et al.* (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* 313(5793):1596-1604.
55. Shulaev V, *et al.* (2011) The genome of woodland strawberry (Fragaria vesca). *Nat Genet* 43(2):109-116.
56. Young ND, *et al.* (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* 480(7378):520-524.
57. Schmutz J, *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature* 463(7278):178-183.
58. Jaillon O, *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449(7161):463-467.
59. Ming R, *et al.* (2013) Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.). *Genome Biol* 14:R41.
60. Fang GC, *et al.* (2010) Genomic tools development for Aquilegia: construction of a BAC-based physical map. *BMC genomics* 11:621.
61. Paterson AH, *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature* 457(7229):551-556.
62. Vogel JP, *et al.* (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* 463(7282):763-768.
63. Goff SA, *et al.* (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* 296(5565):92-100.
64. D'Hont A, *et al.* (2012) The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* 488(7410):213-217.
65. Al-Mssallem IS, *et al.* (2013) Genome sequence of the date palm Phoenix dactylifera L. *Nat Commun* 4.
66. DePamphilis CW, *et al.* (2013) The Amborella genome and the evolution of flowering plants. *Science* 342(6165).
67. Banks JA, *et al.* (2011) The selaginella genome identifies genetic changes associated with the evolution of vascular plants. *Science* 332(6032):960-963.
68. Rensing SA, *et al.* (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science* 319(5859):64-69.
69. Katoh K, Misawa K, Kuma KI, & Miyata T (2002) MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30(14):3059-3066.

70. Miller MA, Pfeiffer W, & Schwartz T (2011) The CIPRES science gateway: a community resource for phylogenetic analyses. in *Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery* (ACM, Salt Lake City, Utah), pp 1-8.
71. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(btl446).
72. Wilgenbusch JC & Swofford D (2003) Inferring evolutionary trees with PAUP*. *Curr Protoc Bioinformatics* 6.4.
73. Drummond AJ & Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7(1).
74. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792-1797.
75. Wheat CW & Wahlberg N (2013) Phylogenomic insights into the cambrian explosion, the colonization of land and the evolution of flight in Arthropoda. *Syst Biol* 62(1):93-109.
76. Wikström N, Savolainen V, & Chase MW (2001) Evolution of the angiosperms: calibrating the family tree. *Proc R Soc B* 268(1482):2211-2220.
77. Iglesias A*, et al.* (2007) A Paleocene lowland macroflora from Patagonia reveals significantly greater richness than North American analogs. *Geology* 35(10):947-950.
78. Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, & Mathews S (2010) Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. *Proc Natl Acad Sci U S A* 107:18724-18728.
79. Harmon LJ, Weir JT, Brock CD, Glor RE, & Challenger W (2008) GEIGER: Investigating evolutionary radiations. *Bioinformatics* 24(1):129-131.
80. Alfaro ME*, et al.* (2009) Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proc Natl Acad Sci U S A* 106(32):13410-13414.
81. Fahey JW, Zalcmann AT, & Talalay P (2001) The chemical diversity and distribution of glucosinolates and isothiocyanates among plants. *Phytochemistry* 56(1):5-51.
82. Mithen R, Bennett R, & Marquez J (2010) Glucosinolate biochemical diversity and innovation in the Brassicales. *Phytochemistry* 71(17-18):2074-2086.
83. Hofberger JA, Lyons E, Edger PP, Chris Pires J, & Eric Schranz M (2013) Whole genome and tandem duplicate retention facilitated glucosinolate pathway diversification in the mustard family. *Genome Biol Evol* 5(11):2155-2173.
84. Kliebenstein DJ (2008) A role for gene duplication and natural variation of gene expression in the evolution of metabolism. *PloS ONE* 3(3).
85. Kliebenstein DJ, Lambrix VM, Reichelt M, Gershenzon J, & Mitchell-Olds T (2001) Gene duplication in the diversification of secondary metabolism: Tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in arabidopsis. *Plant Cell* 13(3):681-693.
86. Jiao Y*, et al.* (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature* 473(7345):97-100.
87. Vision TJ, Brown DG, & Tanksley SD (2000) The origins of genomic duplications in Arabidopsis. *Science* 290(5499):2114-2117.
88. Bowers JE, Chapman BA, Rong J, & Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422(6930):433-438.
89. Jiao Y*, et al.* (2012) A genome triplication associated with early diversification of the core eudicots. *Genome Biol* 13(1):R3.

90. Soltis DE*, et al.* (2009) Polyploidy and angiosperm diversification. *Am J Bot* 96(1):336-348.
91. Franzke A, Lysak MA, Al-Shehbaz IA, Koch MA, & Mummenhoff K (2011) Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends in plant science* 16(2):108-116.
92. Haudry A*, et al.* (2013) An Atlas of over 90,000 Conserved Non-Coding Sequences Yields Detailed Insights Into Crucifer Regulatory Regions. *Nature genetics* 45:891-898.
93. Barker MS*, et al.* (2010) EvoPipes.net: Bioinformatic Tools for Ecological and Evolutionary Genomics. *Evol Bioinform Online* 6:143-149.
94. Altschul SF*, et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389-3402.
95. Birney E, Clamp M, & Durbin R (2004) GeneWise and Genomewise. *Genome Research* 14:988-995.
96. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 5:555-556.
97. McLachlan GJ & Krishnan T (1997) The EM Algorithm and Extensions. *New Yowk: Wiley.*
98. Liu K, Raghavan S, Nelesen S, Linder CR, & Warnow T (2009) Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees. *Science* 324(5934):1561-1564.
99. Enright AJ, Van Dongen S, & Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30(7):1575-1584.
100. Halkier BA & Gershenzon J (2006) Biology and biochemistry of glucosinolates. pp 303-333.
101. Kliebenstein DJ, Kroymann J, & Mitchell-Olds T (2005) The glucosinolate-myrosinase system in an ecological and evolutionary context. *Current opinion in plant biology* 8(3 SPEC. ISS.):264-271.
102. Hofberger JA, Lyons E, Edger PP, Pires JC, & Schranz EM (2013) Glucosinolate metabolic versatility in Brassicaceae is leverage by a complex interplay of increased ohnolog and tandem duplicate retention.
103. Bekaert M, Edger PP, Hudson CM, Pires J, & Conant GC (2012) Metabolic and evolutionary costs of herbivory defense: Systems biology of glucosinolate synthesis. *New Phytologist* 196(2):596-605.
104. Sønderby IE, Burow M, Rowe HC, Kliebenstein DJ, & Halkier BA (2010) A complex interplay of three R2R3 MYB transcription factors determines the profile of aliphatic glucosinolates in Arabidopsis. *Plant physiology* 153(1):348-363.
105. Sønderby IE*, et al.* (2007) A systems biology approach identifies a R2R3 MYB gene subfamily with distinct and overlapping functions in regulation of aliphatic glucosinolates. *PloS one* 2(12).
106. Sønderby IE, Geu-Flores F, & Halkier BA (2010) Biosynthesis of glucosinolates - gene discovery and beyond. *Trends in plant science* 15(5):283-290.
107. Wittstock U & Halkier BA (2000) Cytochrome P450 CYP79A2 from Arabidopsis thaliana L. catalyzes the conversion of L-phenylalanine to phenylacetaldoxime in the biosynthesis of benzylglucosinolate. *Journal of Biological Chemistry* 275(19):14659-14666.

108. Chen S*, et al.* (2003) CYP79F1 and CYP79F2 have distinct functions in the biosynthesis of aliphatic glucosinolates in Arabidopsis. *Plant Journal* 33(5):923-937.
109. Naur P*, et al.* (2003) CYP83A1 and CYP83B1, two nonredundant cytochrome P450 enzymes metabolizing oximes in the biosynthesis of glucosinolates in Arabidopsis. *Plant physiology* 133(1):63-72.
110. Celenza JL*, et al.* (2005) The arabidopsis ATR1 Myb transcription factor controls indolic glucosinolate homeostasis. *Plant physiology* 137(1):253-262.
111. Gigolashvili T*, et al.* (2007) The transcription factor HIG1/MYB51 regulates indolic glucosinolate biosynthesis in Arabidopsis thaliana. *Plant Journal* 50(5):886-901.
112. Gigolashvili T, Yatusevich R, Berger B, Müller C, & Flügge UI (2007) The R2R3-MYB transcription factor HAG1/MYB28 is a regulator of methionine-derived glucosinolate biosynthesis in Arabidopsis thaliana. *Plant Journal* 51(2):247-261.
113. Hull AK, Vij R, & Celenza JL (2000) Arabidopsis cytochrome P450s that catalyze the first step of tryptophan-dependent indole-3-acetic acid biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America* 97(5):2379-2384.
114. Mikkelsen MD, Hansen CH, Wittstock U, & Halkier BA (2000) Cytochrome P450 CYP79B2 from Arabidopsis catalyzes the conversion of tryptophan to indole-3-acetaldoxime, a precursor of indole glucosinolates and indole-3-acetic acid. *Journal of Biological Chemistry* 275(43):33712-33717.
115. Zhao Y*, et al.* (2002) Trp-dependent auxin biosynthesis in Arabidopsis: Involvement of cytochrome P450s CYP79B2 and CYP79B3. *Genes and Development* 16(23):3100-3112.
116. Bak S, Tax FE, Feldmann KA, Galbraith DW, & Feyereisen R (2001) CYP83B1, a cytochrome P450 at the metabolic branch point in auxin and indole glucosinolate biosynthesis in Arabidopsis. *The Plant cell* 13(1):101-111.
117. Grubb D*, et al.* (2014) Comparative analysis of Arabidopsis UGT74 glucosyltransferases reveals a special role of UGT74C1 in glucosinolate biosynthesis. *Plant Journal* (Epub 12541).
118. Frerigmann H & Gigolashvili T (2014) MYB34, MYB51, and MYB122 Distinctly Regulate Indolic Glucosinolate Biosynthesis in Arabidopsis thaliana. *Mol Plant.* (Epub: SSU004).
119. Grubb CD*, et al.* (2004) Arabidopsis glucosyltransferase UGT74B1 functions in glucosinolate biosynthesis and auxin homeostasis. *Plant Journal* 40(6):893-908.
120. Wahlberg N, Rota J, Braby MF, Pierce NE, & Wheat CW (2014) Revised systematics and higher classification of pierid butterflies (Lepidoptera: Pieridae) based on molecular data. *Zoologica Scripta* 43(6):641-650.
121. Braby MF, Vila R, & Pierce NE (2006) Molecular phylogeny and systematics of the Pieridae (Lepidoptera: Papilionoidea): Higher classification and biogeography. *Zoological Journal of the Linnean Society* 147(2):239-275.
122. Heikkilä M, Kaila L, Mutanen M, Peña C, & Wahlberg N (2012) Cretaceous origin and repeated tertiary diversification of the redefined butterflies. *Proceedings of the Royal Society B: Biological Sciences* 279(1731):1093-1099.
123. Wahlberg N & Wheat CW (2008) Genomic outposts serve the phylogenomic pioneers: Designing novel nuclear markers for genomic DNA extractions of lepidoptera. *Systematic Biology* 57(2):231-242.

124. Wahlberg N, *et al.* (2009) Nymphalid butterflies diversify following near demise at the Cretaceous/Tertiary boundary. *Proceedings of the Royal Society B: Biological Sciences* 276(1677):4295-4302.
125. Drummond AJ, Suchard MA, Xie D, & Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 29(8):1969-1973.
126. Rota J (2011) Data partitioning in Bayesian analysis: Molecular phylogenetics of metalmark moths (Lepidoptera: Choreutidae). *Systematic Entomology* 36(2):317-329.
127. Cummins CA & McInerney JO (2011) A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. *Systematic Biology* 60(6):833-844.
128. Rota J & Wahlberg N (2012) Exploration of data partitioning in an eight-gene data set: Phylogeny of metalmark moths (Lepidoptera, Choreutidae). *Zoologica Scripta* 41(5):536-546.
129. de Jong R (2007) Estimating time and space in the evolution of the Lepidoptera. *Tijdsch Entomol*.
130. Wheat CW & Wahlberg N (2013) Phylogenomic insights into the cambrian explosion, the colonization of land and the evolution of flight in Arthropoda. *Systematic Biology* 62(1):93-109.
131. Xie W, Lewis PO, Fan Y, Kuo L, & Chen MH (2011) Improving marginal likelihood estimation for bayesian phylogenetic model selection. *Systematic Biology* 60(2):150-160.
132. Baele G, Li WLS, Drummond AJ, Suchard MA, & Lemey P (2013) Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Molecular Biology and Evolution* 30(2):239-243.
133. Baele G, *et al.* (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution* 29(9):2157-2167.
134. DeVries PJ (1987) *The Butterflies of Costa Rica* (Princeton University Press).
135. Haeuser CL, Holstein J, & Steiner A (2012) The Global Butterfly Information System. (http://www.globis.insects-online.de/).
136. Sokal R & Rohlf FJ (1995) *Biometry: the principals and practice of statistics in biological research* (WH Freeman and Company).
137. Maddison WP, Midford PE, & Otto SP (2007) Estimating a binary character's effect on speciation and extinction. *Systematic Biology* 56(5):701-710.
138. Fitzjohn RG, Maddison WP, & Otto SP (2009) Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Systematic Biology* 58(6):595-611.
139. Fitzjohn RG (2012) Diversitree: Comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution* 3(6):1084-1092.
140. Alfaro ME, *et al.* (2009) Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proceedings of the National Academy of Sciences of the United States of America* 106(32):13410-13414.
141. Wheat CW, *et al.* (2007) The genetic basis of a plant-insect coevolutionary key innovation. *Proc Natl Acad Sci U S A* 104:20427-20431.
142. Fischer HM, Wheat CW, Heckel DG, & Vogel H (2008) Evolutionary origins of a novel host plant detoxification gene in butterflies. *Mol Biol Evol* 25(5):809-820.