New Phytologist

Research

Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity

Lachezar A. Nikolov[1], Philip Shushkov[2], Bruno Nevado[3], Xiangchao Gan[1], Ihsan A. Al-Shehbaz[4], Dmitry Filatov[3], C. Donovan Bailey[5] and Miltos Tsiantis[1]

[1]Department of Comparative Development and Genetics, Max Planck Institute for Plant Breeding Research, Cologne 50829, Germany; [2]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125, USA; [3]Department of Plant Sciences, University of Oxford, Oxford, OX1 3RB, UK; [4]Missouri Botanical Garden, 4344 Shaw Boulevard, St Louis, MO 63110, USA; [5]Department of Biology, New Mexico State University, Las Cruces, NM 88003, USA

Author for correspondence:
Miltos Tsiantis
Tel: +49 22 15062105
Email: tsiantis@mpipz.mpg.de

Received: *9 August 2018*
Accepted: *10 January 2019*

New Phytologist (2019)
doi: 10.1111/nph.15732

Key words: anchored phylogenomics, comparative transcriptomics, crucifers, leaf evolution, targeted sequence capture.

Summary

- The Brassicaceae family comprises *c.* 4000 species including economically important crops and the model plant *Arabidopsis thaliana*. Despite their importance, the relationships among major lineages in the family remain unresolved, hampering comparative research.
- Here, we inferred a Brassicaceae phylogeny using newly generated targeted enrichment sequence data of 1827 exons (> 940 000 bases) representing 63 species, as well as sequenced genome data of 16 species, together representing 50 of the 52 currently recognized Brassicaceae tribes. A third of the samples were derived from herbarium material, facilitating broad taxonomic coverage of the family.
- Six major clades formed successive sister groups to the rest of Brassicaceae. We also recovered strong support for novel relationships among tribes, and resolved the position of 16 taxa previously not assigned to a tribe. The broad utility of these phylogenetic results is illustrated through a comparative investigation of genome-wide expression signatures that distinguish simple from complex leaves in Brassicaceae.
- Our study provides an easily extendable dataset for further advances in Brassicaceae systematics and a timely higher-level phylogenetic framework for a wide range of comparative studies of multiple traits in an intensively investigated group of plants.

Introduction

Comparative biology relies on a firm phylogenetic framework to extend the mechanistic insights derived from a handful of model organisms to the broad diversity of life. In plants, research on the model *Arabidopsis thaliana* (L.) Heynh broadly informs our understanding of development, physiology, secondary metabolism and plant–microbe interactions, as well as natural variation of these processes (Krämer, 2015; Provart *et al.*, 2016). *Arabidopsis thaliana* belongs to the diverse and economically important family Brassicaceae, which includes *c.* 4000 species distributed across a wide range of habitats around the globe, important crop plants like cabbage, rapeseed and mustard domesticated for food and biofuel, ornamentals, and invasive weeds (Appel & Al-Shehbaz, 2003; Franzke *et al.*, 2010; Kiefer *et al.*, 2014). The family generally features small genomes, which enabled the sequencing of the first plant genome and the highest number of genome sequences for any plant lineage to date (Arabidopsis Genome Initiative, 2000; Koenig & Weigel, 2015). The wealth of genetic and genomic resources coupled with broad trait diversity and ecological adaptations make the Brassicaceae an attractive system for addressing important biological questions, such as genome and chromosome evolution, the evolution of form, adaptation to environmental change, crop domestication and adaptive physiology, plant–animal and plant–microbe interactions, and metabolic diversity (Appel & Al-Shehbaz, 2003; Franzke *et al.*, 2010; Koenig & Weigel, 2015; Krämer, 2015; Nikolov & Tsiantis, 2017).

Recent progress in the systematics of Brassicaceae has assigned most of the species to 52 monophyletic groupings (tribes) (Bailey *et al.*, 2006; Warwick *et al.*, 2010; Al-Shehbaz, 2012) in three major lineages (I, II and III) (Beilstein *et al.*, 2006, 2008), or six clades (A–F) (Huang *et al.*, 2015). The most comprehensive phylogenomic study of nuclear markers of Brassicaceae derives from 32 transcriptome samples representing 29 of the 52 tribes (Huang *et al.*, 2015). These analyses have resulted in the development of a draft tribal classification that provides a comprehensive catalog of the independent lineages in the family. Despite these efforts, relationships along the backbone of the phylogeny and among tribes remain largely unresolved. Moreover, 11 genera are not yet assigned to a tribe (Kiefer *et al.*, 2014). It has been argued that this lack of resolution reflects early rapid radiation, ancient and recent polyploidy, and hybridization across species, genera or even distant lineages that resulted in the lack of phylogenetically informative sequence variation (Franzke *et al.*, 2010; Huang *et al.*, 2015). Expanded genomic and taxonomic coverage are

necessary to test these scenarios, and to minimize phylogenetic error related to insufficient character and taxon sampling (Heath *et al.*, 2008) to resolve the relationships among mustard lineages.

The Brassicaceae exhibit considerable morphological diversity, especially in leaf, fruit and trichome characters (Appel & Al-Shehbaz, 2003; Koenig & Weigel, 2015; Nikolov & Tsiantis, 2017). Leaf shape is a model trait that has received significant attention in studies of the genetic basis of morphological change (Bar & Ori, 2015). The sister genus to the rest of Brassicaceae, *Aethionema*, has simple leaves with entire margins (Mohammadin *et al.*, 2017). The rest of the family features simple leaves with entire margins and minor serrations, or more complex lobed or dissected leaves. Lack of a robust Brassicaceae phylogeny has limited the understanding of the distribution of and the transitions between leaf character states, and it is not clear how leaf complexity has evolved across the family. Comparative genetics has revealed several key developmental regulators of leaf complexity in model Brassicaceae (Vlad *et al.*, 2014; Rast-Somssich *et al.*, 2015). Tracing the evolutionary history of the molecular players that underpin these developmental events is of key importance for understanding how leaf form develops and diversifies, but the necessary phylogenetic framework is currently lacking.

Targeted sequence capture has emerged as a powerful and cost-effective method to produce genome-scale data that facilitates orthologous gene comparisons for many species (Hedtke *et al.*, 2013; Lemmon & Lemmon, 2013; Buddenhagen *et al.*, 2016; Johnson *et al.*, 2018) that has not previously been employed in Brassicaceae phylogenetics. By contrast to other methods for genome reduction, such as transcriptome and restriction site-associated DNA sequencing, targeted sequence capture performs well with degraded archival samples that may be the only available source of material for rare or geographically isolated taxa. Targeting single-copy nuclear genes provides a promising way to generate abundant phylogenetically informative data while minimizing potential issues arising from complex gene lineage evolution (Lemmon & Lemmon, 2013). This approach has been used extensively to resolve difficult phylogenetic relationships in plants, including sunflowers (Mandel *et al.*, 2014), milkweeds (Weitemier *et al.*, 2014), sages (Fragoso-Martínez *et al.*, 2017), legumes (Vatanparast *et al.*, 2018), Dutchman's pipes (Wanke *et al.*, 2017) and breadfruit (Johnson *et al.*, 2016), and is widely used in metazoan phylogenomics (e.g. Faircloth *et al.*, 2015; Prum *et al.*, 2016; Alfaro *et al.*, 2018; Espeland *et al.*, 2018).

Here, we reconstruct the evolutionary history of Brassicaceae with high support using a phylogenomic dataset of 1827 target captured exons from 63 species, one-third of which were derived from otherwise difficult to sample herbarium material. When combined with data acquired from previously published genomes of 16 species, our sampling represents 50 of the currently recognized 52 tribes and 16 taxa unassigned to a tribe. We provide a phylogenetic hypothesis for Brassicaceae, test the robustness of our estimates using novel dataset partitioning and taxon selection schemes, and uncover previously unidentified clades. We use this robust phylogenetic framework to study the evolution of leaf shape in the family through comparative transcriptome analysis, which identified a core set of genes under selection at the gene expression and at the protein level associated with the evolution and development of leaf complexity in the Brassicaceae.

Materials and Methods

Information on Plant Material (Supporting Information Table S1), Library Preparation and Sequencing, and Obtaining corresponding exons from sequenced genomes (Tables S2, S3) is included in Methods S1.

Probe design

In order to generate targets for probe design with improved capture efficiency and enhanced phylogenetic utility throughout the Brassicaceae phylogeny, the genomes of *Arabidopis thaliana*, *Sisymbrium irio* and *Aethionema arabicum* were used to identify a set of putative single-copy orthologous genes by compiling the complete set of coding sequences from each genome and identifying single reciprocal hits using BLAT v.32x1 (Kent, 2002). Only coding sequences with length ≥ 960 bp, ≥ 85% sequence identity to the first best hit from *A. thaliana* and second best hit with sequence identity ≤ 40% were retained. The putative single-copy orthologous genes were split into their corresponding exons. The target set included exons > 180 bp to allow for sufficiently long sequences for probe design, 30% < CG content < 70% to improve in-solution hybridization, and no sequence similarity to annotated transposable elements or organellar sequences to avoid enrichment of nonrelevant targets. Approximately 40 000 unique enrichment probes (baits) of biotinylated RNA 120-mers were designed and synthesized at MYcroarray (Ann Arbor, MI, USA) with a 60-base overlap (2× tilling) between baits.

Sequencing data processing for phylogenomics

Please see Fig. S1 for phylogenetic reconstruction workflow. Raw reads were adaptor and quality trimmed in TRIMMOMATIC v.0.33 (Bolger *et al.*, 2014) with the following parameters: illuminaclip: TruSeq3-PE2.fa leading:20 trailing:20 slidingwindow:5:20 minlen:36, followed by deduplication with fastx_collapser in FASTX-Toolkit (Hannon Lab, CSHL, Cold Spring Harbor, New York, NY, USA). All reads were pooled together irrespective of direction. Reference-based mapping followed by *de novo* extension using the unmapped reads were performed with the hybrid assembler YASRA (Ratan, 2009) as incorporated in the Alignreads pipeline (Straub *et al.*, 2011), which accommodates high-sequence divergence between reads and reference, with percentage identity to reference set to 'medium' and reads aligned in a single step without iteration. To minimize sequencing errors and alleviate complications arising from the putative polyploid nature of some species, heterozygosity, and the inclusion of potentially mixed individuals from herbarium sheets, SNP calling was performed with SAMTOOLS (Li *et al.*, 2009) and VARSCAN v.2.4.1 (Koboldt *et al.*, 2012) with the following parameters: minimum coverage per position 5, minimum frequency of observed allele 0.6, *P*-value 0.1, minimum number of reads supporting position

5; if these criteria were not met, N was called. Contig identity was assigned with BLAT v.35 using translated DNA against the respective exon reference sets, selecting the highest scoring hit, and contigs with score > 20 and percentage identity > 75% were retained. The contigs corresponding to each target exon derived from each of the three references were aligned together in MAFFT v.6.851b (Katoh *et al.*, 2002) using [–maxiterate 1000 –genafpair] to call a consensus sequence for each exon; if coverage was absent, N was inserted, otherwise, majority rule applied (Table S4). Transcriptome-derived reads from *Turritis glabra* were processed similarly.

Exons from sequenced genomes were added to the assembled capture-derived exons and aligned in MAFFT v.6.851b according to the references, which were placed in frame. Alignments were trimmed at the border of the *A. thaliana* reference to remove overflowing noncoding sequence. Realignment by coding frame was performed in MACSE v.1.02 (Ranwez *et al.*, 2011), and trimmed to remove entire codon positions if internal stop codon indicative of misalignment was present in any of the species (a total of 1205 internal stop codons for the entire alignment of ~16 million codons), or if a codon position was too diverse (most prevalent amino acid identical for < 30% of the taxa). Positions with > 20% ambiguous amino acids resulting from unidentified nucleotides (Ns) were removed. Final sequences shorter than 35% of unambiguous nucleotide positions based on the reference exon length were removed. For plastome assembly, we utilized the FASTPLAST pipeline (McKain, 2017), which combines reference-guided and *de novo* assembly to generate plastid genomes using off-target organellar reads (genome skimming) (Table S5). No plastid assemblies with contigs long enough to meet our quality cut-offs were produced for *Cremolobus peruvianus* and *Lunaria rediviva*, or *Turritis glabra* (this likely resulted from polyA enrichment in transcriptome sequencing). Plastid genomes were annotated with GESEQ (https://chlorobox.mpimp-golm.mpg.de/geseq.html). Plastid protein-coding genes were processed similarly to the nuclear dataset to obtain multiple sequence alignments.

Exon-pruning and phylogenetic-signal analyses

Because some ingroup relationships can be sensitive to the choice of an outgroup, we constructed two matrices with different outgroup composition and shared exon sets (1101 exons) – one with *Carica papaya* and *Tarrenaya hassleriana*, and another including only *T. hassleriana*. We used PARTITIONFINDER v.2.0 (Lanfear *et al.*, 2014, 2016) and RAXML v.8.2.9 (Stamatakis, 2014) to reconstruct maximum likelihood trees after concatenation. Rooting with *C. papaya* for the more inclusive dataset, and *T. hassleriana* for the other dataset demonstrated that different outgroup schemes result in identical ingroup relationships for Brassicaceae. Therefore, to maximize the number of shared exons, we only included *T. hassleriana* as the outgroup in our in-depth analyses. We estimated the maximum-likelihood (ML) gene trees for each exon using the fast algorithm in RAXML with GTRGAMMA model of nucleotide substitution, where each exon has three partitions (for the 1st, 2nd and 3rd positions of a codon), and 100 rapid bootstrap replicates. To reduce sequence biases and assess potential sources of misleading signal, we calculated and excluded the most extreme outliers in each of seven metrics: upper quartile of long-branch score (L; 20 exons), standard deviation of the long branch score (36 exons), average patristic differences (33 exons), and R^2 of the saturation score (S; 103 exons) and saturation slope (24 exons) using TRESPEX v.1.1 (Struck, 2014). We further calculated Matching Splits (M; excluded 26 exons) and Robinson–Foulds (11 exons) tree distances for individual exons as implemented in TREECMP v.1.0 (Bogdanowicz *et al.*, 2012) and excluded outliers resulting from aberrant modes of molecular evolution or incorrect paralog assignment that could influence the combined analyses. To explore the effects of dataset partitioning and to test the stability of our results based on exon inclusion/exclusion, we employed two cut-offs for the S, L and M metrics to partition the resulting set of exons into three increasingly exclusive subsets were for each metric. These 27 unique exon sets were named according to the combination of metric intervals, such that the most inclusive dataset is *S1L1M1* and the most exclusive dataset is *S3L3M3* (Table S6). These unique exon sets were used to infer phylogenies via concatenation- and coalescence-based methods. The distances between trees were visualized with density plots of tree distances and multidimensional scaling plots in R.

Phylogenetic reconstruction

Maximum likelihood analyses of the concatenated 27 matrices were performed in RAXML v.8.2.9 after PARTITIONFINDER v.2.0 to find the computationally most efficient model of evolution that minimizes overall model complexity and accurately accounts for substitution processes (Table S6).

In order to evaluate the phylogenetic signal of each exon in a given phylogenetic matrix, we separately optimized the model of evolution with constrained tree topology to each of the eight unique topologies inferred previously in the maximum-likelihood (ML) analysis. The log-likelihood score of each exon at a given topology, $\log_e L(\text{Exon}\,|\text{Topology}\,i)$ was computed as the sum of the log-likelihood scores of all sites in the exon

$$\log_e L(\text{Exon}\,|\text{Topology}\,i) = \sum\nolimits_{\alpha\in\text{Exon}} \log_e L(\text{Site}\,\alpha|\text{Topology}\,i).$$

Using the log-likelihood scores of the exons for all eight unique topologies, the phylogenetic signal of each exon in the given phylogenetic matrix was estimated as the sum of the absolute pairwise differences of the exon log-likelihood scores of the unique tree topologies

$$\Delta\text{GLS} = \frac{1}{N_{\text{pairs}}}\sum\nolimits_{i<j}\left|\log_e L(\text{Exon}\,|\text{Topology}\,i) - \log_e L(\text{Exon}\,|\text{Topology}\,j)\right|.$$

(ΔGLS, phylogenetic signal of the exon; N_{pairs}, number of unique tree topology pairs; i and j run over the set of unique tree topologies). This formula extends the definition of phylogenetic

signal in Shen *et al.* (2017) to any number of tree topologies. The phylogenetic signals of the exons were computed separately for each of the eight most-inclusive phylogenetic matrices that generated the eight unique tree topologies (*S1L1M1*, *S1L2M2*, *S2L2M1*, *S2L2M2*, *S1L1M3*, *S2L2M3*, *S3L1M1*, *S3L3M1*). Exons that disproportionately contribute to the phylogenetic signal of a given phylogenetic matrix (exons with phylogenetic signal in the matrix > 10) were excluded from the matrix and ML tree estimations were performed in RAxML with the filtered datasets. The resulting topologies were renamed *S1L1M1R*, *S1L2M2R*, *S2L2M1R*, *S2L2M2R*, *S1L1M3R*, *S2L2M3R*, *S3L1M1R* and *S3L3M1R*.

Plastid-coding gene phylogenies were estimated in RAxML from the partial or complete plastid genomes obtained from off-target reads and plastid sequences obtained from whole plastomes from GenBank (Table S3).

Because the supermatrix approaches can fail to fully account for the influence of conflicting gene-tree signal due to processes such as incomplete lineage sorting, we estimated coalescent species trees with the gene tree summation method for each of the 27 exon sets in ASTRAL-II v.4.10.12 (currently the only method that can analyze a dataset of our scale under the multispecies coalescent model).

Quartet-based computations of internode certainty (LQ-IC) as a measure of phylogenetic incongruence were calculated according to Zhou *et al.*, 2017 with the *S1L1M1R* topology as the reference tree and sets of evaluation trees, corresponding to all bootstrap replicates from the RAxML analyses with ($27 \times 100 = 2700$ trees) and without ($8 \times 100 = 800$) disproportionate contributors from the phylogenetic informativeness analysis. To compute LQ-IC scores for the coalescent analyses, the 100 bootstrap replicates of each gene tree in a given dataset were used to compile 100 sets of tree replicates, where each exon is represented. These 100 sets were used to compute 100 trees under the coalescent model. This procedure was done separately for each of the 27 datasets to produce a total of 2700 evaluation trees.

Phylogenetic hypothesis testing (AU test) was performed in CONSEL v.1.20 (Shimodaira & Hasegawa, 2001) to test the statistical significance of topological differences between trees in the ML analyses and the ML analyses after excluding disproportionate contributors to the phylogenetic signal.

Comparative transcriptomics and identifying shifts in gene expression

De novo transcriptome assembly was performed with TRINITY v.2.4.0 (Grabherr *et al.*, 2011) using default parameters after combining data from all three biological replicates for each species. Coding sequences within contigs were identified with Transdecoder (http://transdecoder.github.io/), and all open reading frames with homology to the *A. thaliana* proteome (GenBank build UP000006548_3702) longer than 100 amino acids were used in subsequent analyses. Considering only the longest isoform of each gene, OrthoFinder (Emms & Kelly, 2015) was used to identify orthogroups consisting of a single gene for all eight species. Expression values were calculated for each triplicate with RSEM (Li & Dewey, 2011) in Trinity, using the species-specific transcriptome assemblies. Gene expression values for each species independently were expressed as transcripts per kilobase million (TPM; the most readily comparable measurement between species), and $\log_2$-transformed after adding 0.0001 to each value to avoid –inf errors. Between-sample normalization of expression values was performed with the package R/POISSONSEQ, which implements the method of Li *et al.* (2012). We built a NJ tree with the R/APE package (Paradis *et al.*, 2004) using as input a correlation matrix of the normalized expression values of all samples for the 3188 orthologous genes identified, to demonstrate that all replicates of each species clustered together, as expected.

Genes that deviate from background gene expression in the core Brassicaceae or in the complex-leaved species were identified in the context of the Ornstein–Uhlenbeck process modeling framework (Butler & King, 2004; Rohlfs *et al.*, 2014) after excluding genes exhibiting high-expression variability in at least one species (SD > 0.5 × mean; 846 genes). The framework takes into account phylogenetic information (tree topology including only the species with the newly generated transcriptomes and branch lengths calculated from the in-frame alignment of 1421 genes) to fit alternative models to the expression of each of the resulting 2342 genes using the average of the normalized, $\log_2$-transformed expression values for each species. Directional expression denotes a deviation from an optimal expression level, which accommodates phylogenetic relationships and drift, and does not specify the actual direction of the change. Two sets of tests were performed: one based on phylogeny alone comparing a null model of a single expression value for all species with models that specify distinct expression levels separately for the core Brassicaceae, and a second set of tests based on leaf morphology (simple/complex). Alternative models (H0a, uniform expression along all branches of the tree; H1a, shift in expression in the core Brassicaceae; H0b, uniform expression along all branches of the tree; and H1b, shift in expression along the branches leading to the complex-leaved species, where the character state of internal nodes is not specified) were compared using likelihood score and likelihood ratio chi-square test with one degree of freedom and FDR-corrected *P*-values to identify genes where a model with two optima fits the data better than a model with a single optimum for all lineages.

Detecting positive selection on the protein sequence

In order to estimate the frequency of positive selection acting on coding sequences, codon-based multiple species alignments for all orthogroups were used to fit alternative phylogenetic models of evolution of nonsynonymous to synonymous substitution rate ($\omega = dN/dS$). Codons with > 60% missing data and sequences with > 60% missing codons were excluded (nine genes). A further 32 genes were excluded because they showed very large synonymous divergence indicative of alignment errors or aberrant molecular evolution ($dS > 2.5$). Individual ML gene trees estimated in phyml (Guindon & Gascuel, 2003) were compared to the species tree using the Shimodaira–Hasegawa test (Shimodaira & Hasegawa, 1999) implemented in CONSEL v.1.20 (Shimodaira

& Hasegawa, 2001) to exclude genes exhibiting significant phylogenetic conflict, that is, genes that preferred the individual gene tree over the species tree (18.6% of the initial 3188 genes). The remaining 2554 genes were used to fit alternative models with 'Branch-site test of positive selection', which allows ω to vary among sites of the gene and branches of the phylogeny in codeml/PAML v.4.8 (Yang, 2007). The test identifies sites that are evolving neutrally or under negative selection in part of the tree (background) but exhibit a shift towards positive selection along branches of interest (foreground). We performed two sets of tests, one based on phylogeny and another based on leaf morphology, using a tree including only the taxa with newly generated transcriptomes. Significance was tested with likelihood ratio test with two degrees of freedom.

Data availability

Raw sequence read data are available from NCBI Short Read Archive (PRJNA518905).

Results and Discussion

Data generation

We designed the first exome targeted enrichment probe set for Brassicaceae based on single-copy nuclear markers derived from three reference genomes, *Arabidopsis thaliana*, *Sisymbrium irio*, and *Aethionema arabicum*, and targeted 1827 exons of average size 516 bp (range 180–6059 bp) from 764 genes, representing all Brassicaceae linkage groups (Fig. 1a,b). The focus on single-copy genes aimed to reduce issues of paralogy. The exons were selected for size and nucleotide composition to maximize sequence capture and phylogenetic utility, and their collective length measured on the longest alignment of the three references was 942 066 bp, cumulatively representing c. 1.5% of the exome of *A. thaliana*. We collected novel exome-targeted enrichment data for 63 species (Methods S1; Fig. 1c). To break up long branches, we included 16 taxa with controversial or ambiguous phylogenetic position that may represent independent evolutionary lineages. One third of the samples were derived from herbarium material, which produced high-quality data despite limited input DNA quality and quantity. Our complete dataset includes 79 Brassicaceae species representing all currently recognized lineages of Brassicaceae except two recently described tribes (Hillieleae (Chen *et al.*, 2016) and Shehbazieae (German & Friesen, 2014)).

Nearly all targeted regions were captured for all species with average coverage *c.*100×, resulting in a dataset with few missing data per terminal (Fig. 2a,b; Table S4). Mapping to each of the three references resulted in comparable exon recovery from the target species, suggesting that phylogenetic distance was not a significant factor for capture success in Brassicaceae. At the level of mapped reads, we observed polymorphisms that could reflect allelic variation (heterozygosity), copy number variation (reads from paralogous sequences) and the sampling of multiple individuals (from herbarium material) in all samples. Because the lack of synteny data precludes differentiating among these scenarios, we masked these positions using conservative criteria for base calling in the initial contig assembly to minimize the effect of this variation on tree estimation.

In order to assess whether our targeted loci can be recovered from transcriptome data without target enrichment to enable direct comparisons with previous phylotranscriptomic studies, we sequenced the transcriptome of *Turritis glabra* seedlings. We analyzed this transcriptome similarly to the targeted sequence data, which resulted in near complete (1781 of 1827) exon recovery, demonstrating the potential for future merging of our dataset with similarly generated targeted enrichment datasets and phylotranscriptomic datasets. Although we did not specifically target plastid sequences, we were able to assemble partial or complete (*Arabis ottonis-schulzii*, *Alyssopsis mollis*, *Bunias erucago*, *Dipoma iberideum*, *Kernera saxatilis*, *Murbeckiella pinnatifida*, *Pseudofortuynia esfandiarii*, *Subularia aquatica* and *Stanleya elata*) plastid genomes from off-target capture reads (genome skimming) for all but three species (Fig. S2; Table S5), enabling phylogenetic comparison of both biparentally-inherited nuclear and maternally-inherited plastid genomes. The extensive sequencing information we obtained also allowed for dataset partitioning to evaluate the contributions of different loci to the phylogenetic signal.

Phylogenomic analyses

Because phylogenomic datasets capture the unique evolutionary history of many genomic loci, we investigated relationships using concatenation and coalescence-based (ASTRAL-II) species tree approaches, and characterized potential systematic bias via three metrics often used to detect misleading signal in phylogenetic reconstructions: evolutionary rate heterogeneity (L) that may result in long-branch attraction; sequence saturation (S) that obscures phylogenetic signal; and distance between gene trees (M) generated from individual exons that may indicate hidden paralogy resulting from complex gene-lineage evolution or polyploidization (Fig. S3). Excluding the most extreme outliers for the three metrics (20 exons of 18 genes for L metric, 103 exons of 96 genes for S metric, and 26 exons of 25 genes for M metric; some exons are excluded by more than one metric) and loci not recovered from the genome of the outgroup *Tarenaya hassleriana* (79 exons of 25 genes) resulted in 1540 exons from 673 genes. To assess the robustness of phylogenetic estimates based on locus selection under different partitioning schemes, we concatenated the resulting exons into 27 matrices based on combinations of metric cut-off values (Dataset S1; Fig. S3; Table S6) and conducted maximum-likelihood (ML) analyses under the best-fitting models of evolution. These estimates resulted in species trees well supported along the backbone that produced eight unique topologies, which differed by the relationship between *Idahoa* and *Subularia*, the branching order of *Cochlearia*, *Conringia orientalis* and relatives (*Conringia* clade), and *Kernera* + *Petrocalis* in respect to the rest of the clade, and the placements of *Megacarpaea* and Biscutelleae + (*Lobularia* + *Iberis*) (Dataset S2; Fig. S4). Different phylogenetic matrices could not reject all alternative topologies based on the AU test, indicating that some

New Phytologist

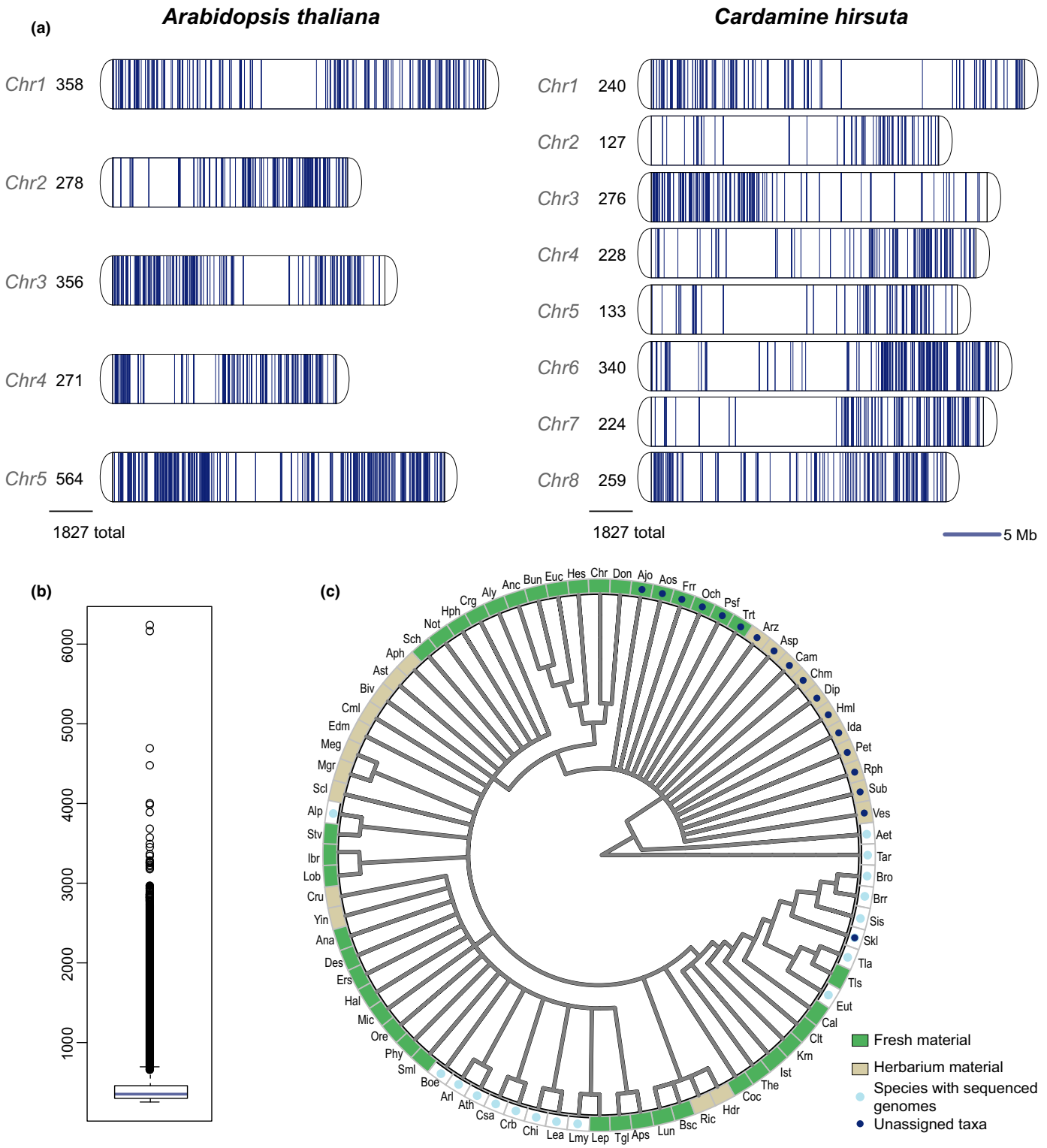

Fig. 1 Sources of plant material and features of the targeted loci. (a) The targeted exons (blue lines) span all linkage groups in the genomes of *Arabidopsis thaliana* ($n = 5$, a more derived karyotype) and *Cardamine hirsuta* ($n = 8$), and are largely absent from the centromeric and pericentromeric gene-poor regions. (b) Size distribution of targeted exons (bp) based on the longest alignment of the three reference species. (c) Taxon sampling summarized on a Brassicaceae phylogeny before the current study (phylogeny based on multigene analyses at the intertribal level, e.g. Huang *et al.*, 2015). One third of the samples were obtained from herbarium material (brown). Species with sequenced genomes are marked with a light blue dot; taxa previously unassigned to a tribe are marked with a dark blue dot. See Supporting Information Tables S1, S2 for species abbreviations and tribal assignments.

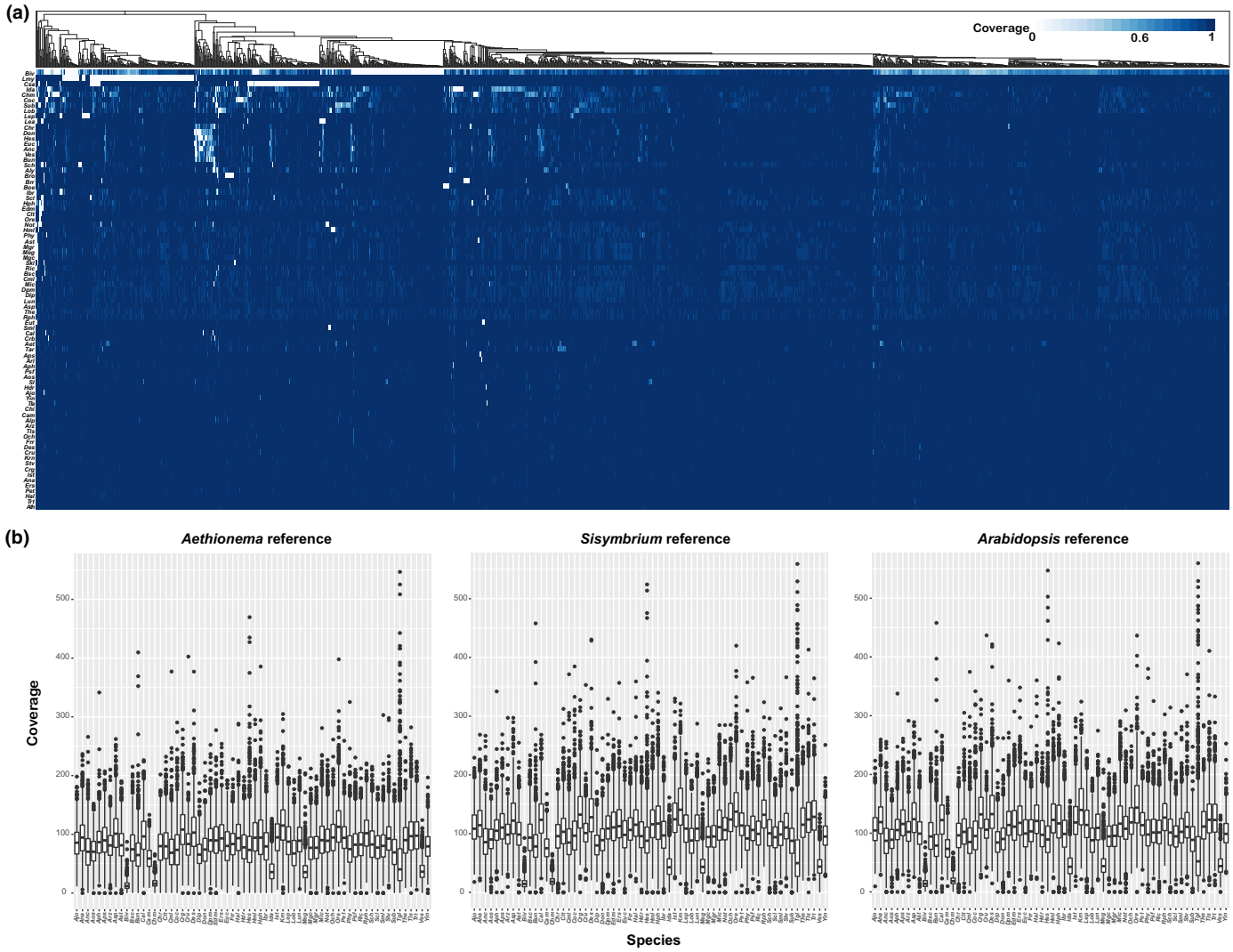

Fig. 2 Exon coverage. (a) Across taxa displaying near universal capture and even distribution of the sampled orthologous gene space. The *x*-axis represents all sampled 1827 exons per species (*y*-axis), with dendrogram at the top displaying the similarity in species-wide coverage between different exons. (b) Per-species coverage based on the three references showing comparable capture success irrespective of phylogenetic distance.

of these unique topologies are indistinguishable based on the data (Table S7).

Concatenation averages out phylogenetic signal from numerous loci to estimate the species phylogeny but it has been demonstrated that genes with strong phylogenetic signal may bias these estimates (Shen *et al.*, 2017). To study the effect of such strong contributors, we evaluated the phylogenetic informativeness of the exons measured by their relative contribution to the total likelihood of each of the eight unique topologies based on the most-inclusive concatenation dataset. As expected, some loci contributed disproportionally to the likelihood of a given topology (Fig. S5). Because misleading strong signal from a minority of loci can potentially overpower the phylogenetic signal of other, more reliable loci, we excluded these strong contributors from the respective most-inclusive phylogenetic matrices (*S1L1M1*, *S1L1M3*, *S1L2M2*, *S1L3M1*, *S1L3M3*, *S2L2M1*, *S3L1M1*, *S3L3M1*). The ML estimations of these eight matrices produced eight unique topologies, some indistinguishable based on the data, which by contrast with previous analyses consistently resolve a strongly supported clade of *Megacarpaea*, *Cochlearia*, *Lobularia* and *Iberis* (Dataset S2; Fig. S6). The topologies differed by the relationship between *Idahoa* and *Subularia*, the relationship among *Cochlearia*, *Lobularia* and *Iberis*, and the positions of Biscutelleae and Lineage V (see 'Backbone of the phylogeny'; Fig. S6; Table S7).

In order to determine how taxa with variable positions in the concatenation analyses influence tree topology, we conducted a series of taxon-pruning experiments, systematically removing *Cochlearia officinalis*, *Lobularia maritima*, *Iberis linifolia*, *Megacarpaea* spp., *Subularia aquatica* and *Idahoa scapigera* from the largest matrix excluding strong contributors (*S1L1M1R*) (Fig. S7). These removals led to variable placements of the other unstable taxa, demonstrating the high sensitivity of the analysis to the exclusion of even a single 'keystone' taxon, thus underscoring the importance of our broad sampling of recognized Brassicaceae tribes for accurate phylogenetic inference.

In order to assess gene conflict arising from deep coalescence, we estimated species trees for each of the 27 sets of gene trees with ASTRAL-II, which calculates the support of a topology based on the prevalence of quartet trees derived from the individual gene trees (Dataset S2). These analyses resulted in seven unique species tree topologies containing similar groupings as in the concatenated ML analyses that differed in the position of *Idahoa*, the placements of *Iberis* and *Lobularia*, and the relationships among *Cremolobus*, *Brayopsis* and *Schizopetalon*, and among *Arabidopsis*, *Physaria* and the rest of Lineage I (Fig. S8). By contrast to their conflicting placements in the ML trees, *Cochlearia,* (*Kernera* + *Petrocalis*) and the *Conringia* clade were resolved as successive sisters to the rest of Lineage II in all ASTRAL-II trees with high support. In all concatenation analyses, (*Camelina* + *Capsella*) is sister to (*Boechera* + *Halimolobos* + *Crucihimalaya*) + (*Dipoma* + *Hemilophia* + *Geococcus*), whereas in all ASTRAL-II it is sister to (*Boechera* + *Halimolobos* + *Crucihimalaya*). The difference in topologies suggests complex speciation in this part of the tree.

Backbone of the phylogeny

Analyses of nuclear loci produced well-resolved and largely congruent topologies consisting of six clades that form successive sister groups to the rest of Brassicaceae: (1) *Aethionema*; (2) a clade approximately representing the previously circumscribed Lineage III (Beilstein *et al.*, 2006); (3) a clade comprising the tribes Arabideae, Stevenieae and Alysseae, which we call Lineage IV; (4) a comprehensive clade consisting of representatives of Lineage I (Beilstein *et al.*, 2006); (5) a clade including relatives of the previously circumscribed Lineage II (Beilstein *et al.*, 2006); and (6) a novel clade of taxa distributed primarily in the Southern Hemisphere, designated hereafter as Lineage V (Fig. 3; Notes S1). At lower phylogenetic levels, our results support previously recognized relationships among tribes, reveal novel clades and assign previously unassigned taxa to a lineage, some of which may justify the erection of new tribes (Fig. S9; Tables S8, S9). The topology of the reconstructed plastid tree (Fig. S10) revealed conflicting signal between the nuclear and plastid genomes at deeper nodes, notably in the branching order of lineages I and III, and among the more terminally branching taxa in Lineage I (Fig. S11). Such cytonuclear discordance is consistent with prior findings that have invoked both chloroplast capture (Beilstein *et al.*, 2008) and substantial nuclear introgression (Forsythe *et al.*, 2017) in Lineage I, suggesting that such phenomena may influence phylogenetic relationships more broadly in the family.

Points of incongruence among analyses

Our results reveal areas of topological stability and highlight recalcitrant nodes that exhibit discordance at the individual gene tree level and after the interrogation of combinations of exons and gene trees. Much of the topological conflicts concern only few taxa with unstable positions in different analyses and datasets. For example, the placements of *Subularia* and *Idahoa*, although firmly resolved within the novel Lineage V, vary with respect to each other and to the Cremolobeae-Eudemeae-Schizopetaleae (CES) and the *Asta* + *Scoliaxon* clades in all analyses. Both *Subularia* and *Idahoa* exhibit higher rates of molecular evolution indicated by long branches, which can obscure phylogenetic relationships making them intractable to resolve even with datasets of our scale (e.g. King & Rokas, 2017).

Incongruent phylogenetic histories also may result from duplication and subsequent gene loss, incomplete lineage sorting and introgression (e.g. Wendel & Doyle, 1998). Another area of conflict in the Brassicaceae phylogeny is the relationship among *Cochlearia*, *Iberis*, *Lobularia* and *Megacarpaea*. This clade received little support in the coalescence and the initial concatenation analyses. High support was obtained only after exclusion of loci that contribute disproportionally to the phylogenetic signal in the concatenation analyses, suggesting difficulties in assigning gene orthology among taxa (Walker *et al.*, 2017) as the reason for the observed results. Such difficulties may arise from whole-genome duplication (polyploidization) events and subsequent asymmetric gene loss of paralogs, resulting in long branches (Fares *et al.*, 2005), as observed here. The larger average genome sizes of Anastaticeae, Cochlearieae and Iberideae (Megacarpaeeae genome size is not known), as well as the variable base chromosome number in all four tribes (Hohmann *et al.*, 2015) lend support to a possible shared mesopolyploid origin of this clade. A more targeted approach, such as distribution of synonymous substitutions of paralogs over time (e.g. Mandáková *et al.*, 2017) and syntenic information from whole-genome sequences will clarify this issue. However, note that whole-genome duplication events along terminal branches are unlikely to result in phylogenetic uncertainty. Relationships in Lineage III, where we sampled several polyploid species, are consistently resolved in all analyses, suggesting that the polyploidization events they feature may not be shared among tribes in this clade.

The basal-most nodes of Lineage I were universally resolved, but conflicting resolution was evident for the branching order associated with *Arabidopsis* spp. and *Physaria*, and the relationship among *Capsella rubella* + *Camelina sativa*, *Geococcus pusillus* + *Hemilophia rockii* + *Dipoma iribideum* and *Boechera stricta* + *Halimolobos pubens* + *Crucihimalaya himalaica*. This uncertainty is reflected in cytonuclear discordance, where the plastid phylogeny unites *Arabidopsis* spp. and *Capsella rubella* + *Camelina sativa* (a monophyletic Camelineae), and *Geococcus pusillus* + *Dipoma iberideum* + *Crucihimalaya himalaica*, strongly suggesting a complex speciation. It has been suggested recently that this observation reflects massive nuclear introgression (Forsythe *et al.*, 2017). Our broad sampling suggests that these processes are more pervasive and explain the inconsistent placement of Microlepidieae in Huang *et al.* (2015) better than a putative hybrid origin for the tribe. Incomplete lineage sorting likely contributed to some of the observed patterns although its extent in this clade is unclear. These three examples demonstrate the complexity of evolutionary processes shaping the current Brassicaceae diversity and further emphasize the utility of this model clade in studying complex evolutionary histories.

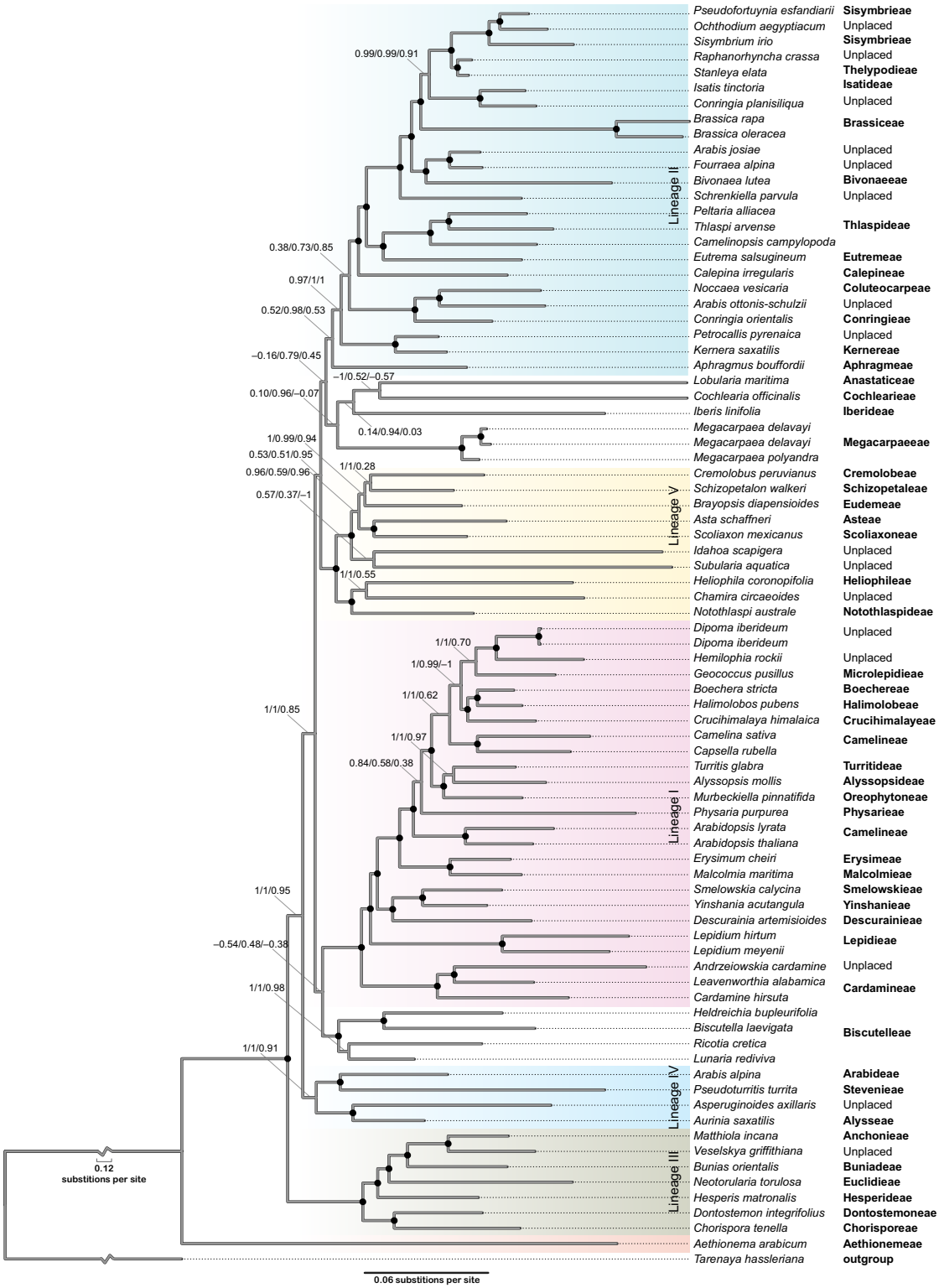

Fig. 3 Maximum-likelihood topology of the Brassicaceae relationships inferred from 79 species and 1421 exons. Nodes marked with a black circle are universally supported by all concatenation and coalescence analyses. Numbers associated with certain internodes are a quartet-based measure, lowest quartet internode certainty (LQ-IC) (Zhou *et al.*, 2017), for quantifying the similarity between this reference tree and three sets of evaluation trees: RAxML bootstrap trees (2700 trees)/the RAxML bootstrap trees after removal of strong contributors (800 trees)/trees used to calculate local support in the coalescence analyses (2700 trees). LQ-IC varies between −1 and 1: it approaches '1' when the reference internode is prevalent among the evaluation trees, it is close to '0' when alternative topologies are common, suggesting incongruence between the reference tree and the evaluation trees with respect to this internode, and approaches '−1' when the evaluation trees strongly contradict the internode present in the reference topology. For example, the branch leading to *Cremolobus peruvianus* + *Schizopetalon walkeri* (1/1/0.28; Lineage V) is universally supported in both maximum-likelihood analyses (LQ-IC score = 1) but not present in all coalescent evaluation trees, reflected by a lower LQ-IC score (0.28).

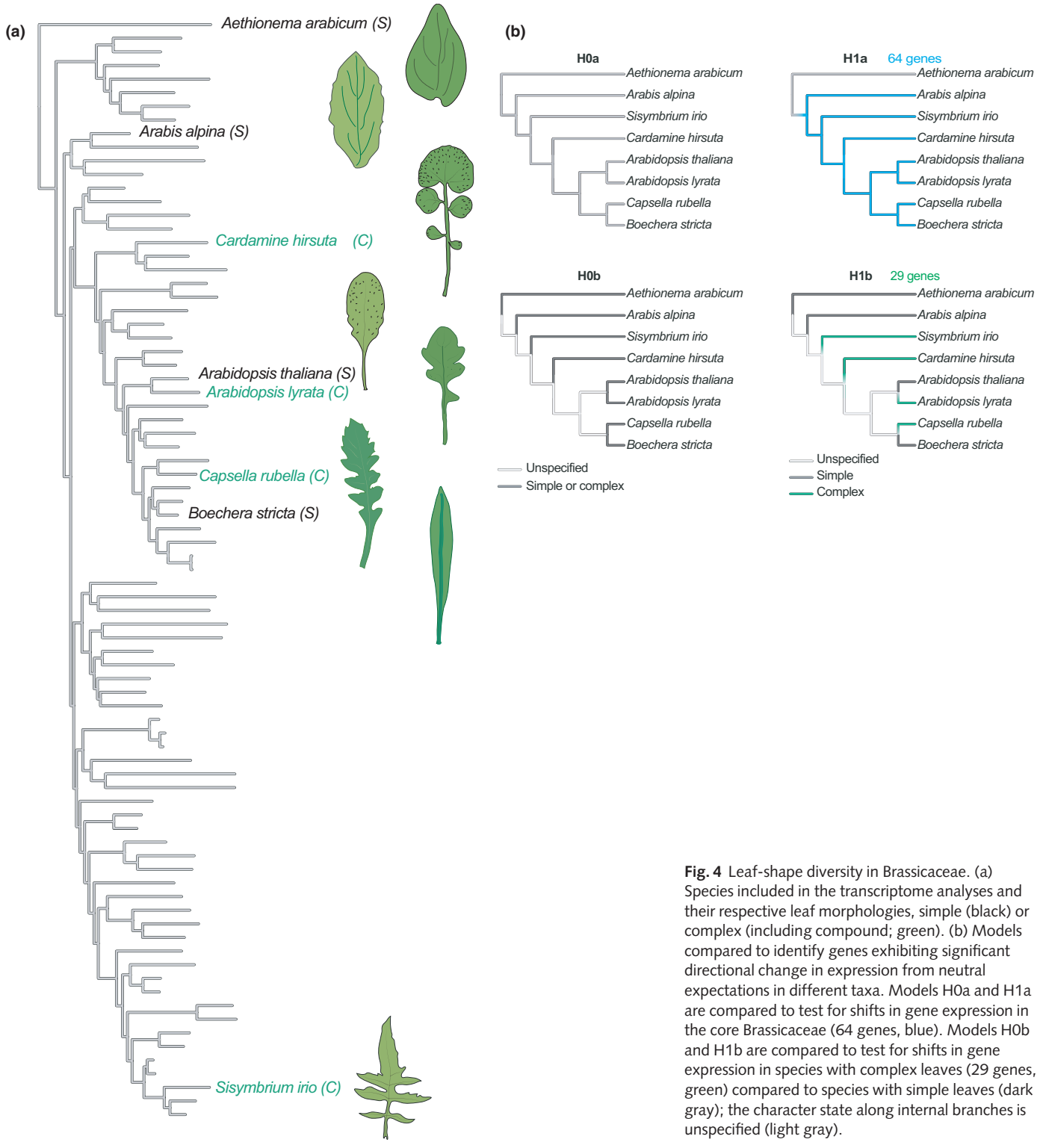

Fig. 4 Leaf-shape diversity in Brassicaceae. (a) Species included in the transcriptome analyses and their respective leaf morphologies, simple (black) or complex (including compound; green). (b) Models compared to identify genes exhibiting significant directional change in expression from neutral expectations in different taxa. Models H0a and H1a are compared to test for shifts in gene expression in the core Brassicaceae (64 genes, blue). Models H0b and H1b are compared to test for shifts in gene expression in species with complex leaves (29 genes, green) compared to species with simple leaves (dark gray); the character state along internal branches is unspecified (light gray).

Leaf-form evolution in Brassicaceae

The newly developed understanding of relationships along the backbone of the Brassicaceae phylogeny offers considerable opportunities for comparative studies focused on the evolution of critical traits. Leaf shape is a model trait for understanding the genetic basis of phenotypic diversity (Nikolov & Tsiantis, 2017) and our phylogeny provides a macroevolutionary framework to study the distribution of leaf complexity states in Brassicaceae. For example, simple leaves with entire, dentate or serrate margins are present in all six major clades in the family. All species of *Aethionema* have undivided, entire leaves, and have an articulation (joint) between the petiole-like base and the rest of the leaf blade, a trait which is exceptionally rare in the rest of the

Brassicaceae. Deeply lobed leaves are present in several independent lineages, including Euclidieae, Descurainieae, Lepidieae, Sisymbrieae, Brassiceae, Isatideae and Thelypodieae, among others. The tribes Descurainieae, Smelowskieae and Yinshanieae, which formed a distinct clade in the phylogeny, have predominantly dissected leaves, and in some species of *Descurainia* (e.g. the South American *D. nuttallii*) the leaves are finely 3-pinnatisect. The presence of simple leaves in *Smelowskia porsildii* is a derived case. The tribe most diverse in leaf morphology is the Cardamineae, the fourth largest in the family, where the vast majority of species have variously divided leaves. Pinnately compound leaves are present in three genera, *Cardamine*, *Nasturtium* and *Andrzeiowskia*. *Cardamine* includes species with all types of compound leaves: trifoliolate, pinnate, palmately compound and bipinnately compound, as well as variously dissected simple leaves and simple entire leaves. The only other genus in the family with truly compound leaves is *Hilliella* (Hillielleae) of China, but its placement is currently unknown (Chen *et al.*, 2016). More targeted phylogenies are needed to resolve transitions among character states.

Identifying gene expression differences that may underlie phenotypic differences is crucial for understanding the genetic basis of morphological evolution (Peter & Davidson, 2011; Rowan *et al.*, 2011). Previously we demonstrated the significant overrepresentation of transcription factors among the differentially expressed genes between *Cardamine hirsuta* and *Arabidopsis thaliana* during early leaf development (Gan *et al.*, 2016). To extend these analyses to the family level and identify gene expression and coding sequence differences associated with leaf diversity in Brassicaceae, we generated the transcriptional profiles of young developing leaves from eight diploid species with sequenced genomes and contrasting leaf shapes – simple or complex (including compound leaves) – throughout the Brassicaceae phylogeny (Fig. 4a, Table S10). Comparing gene expression across species requires prior knowledge of phylogenetic relationships (Dunn *et al.*, 2017) to account for neutral fluctuations in gene expression levels in different lineages (Rohlfs *et al.*, 2014). Such analyses could not previously be performed with sufficient confidence because (1) branch lengths, as required for downstream analyses, could not be estimated accurately, and (2) the relationship among taxa with newly generated transcriptomes, in particular the placement of the emerging model species *Arabis alpina*, has not been resolved consistently before our phylogenomic study (Willing *et al.*, 2015).

Employing Ornstein–Uhlenbeck process modeling (Rohlfs *et al.*, 2014) using our explicit phylogenetic framework, we identified 64 genes exhibiting significant directional change in expression from neutral expectations in the core Brassicaceae (the clade sister to *Aethionema*), where diversity in leaf margin geometry is pronounced (q-value < 0.05) (Dataset S3). These genes included the developmental regulators *AUXIN RESPONSE TRANSCRIPTION FACTOR 3*/*ETTIN* (*AT2G33860*), *CYCLIN D6;1* (*AT4G03270*), *CHROMATIN REMODELING 9*/*SWITCH 2* (*AT1G03750*), *Enhancer of polycomb-like transcription factor* (*AT4G32620*) and the ethylene receptor *ETHYLENE RESPONSIVE 2* (*AT3G23150*) (Table S11). Next, we tested for directional change in gene expression not compatible with neutral evolution along the branches leading to the species with complex leaves relative to species with simple leaves, and identified 29 such genes (Table S12, Dataset S3). These genes included the putative signaling components *MAP KINASE 17* (*AT2G01450*), *MAP KINASE 20* (*AT2G42880*), the microtubule-associated *GROWING PLUS-END TRACKING PROTEIN 2* (*AT3G53320*) and several post-translational modifiers (metallopeptidase M24 family, trypsin family, peptidase family M48), which emerge as candidates for a shared core set of regulators associated with the evolution and development of leaf complexity in Brassicaceae. Because microtubules play important roles in growth polarity, proteins that spatially organize this portion of the plant cytoskeleton may have been recruited to contribute to marginal growth polarization associated with complex leaf morphogenesis (Barkoulas *et al.*, 2008). One of the genes, *AT1G19485*, a WD40-repeat protein hypothesized to be a part of a CUL4-RING E3 ubiquitin ligase complex (Jackson & Xiong, 2010) exhibits both shift in expression and evidence for positive selection (Table S13) on residues in its N-terminus in complex-leaved species. This finding highlights a potential role of protein ubiquitination in the diversification of leaf form. Taken together, the gene expression results identify candidates for further functional studies into the genetic basis of leaf diversity in Brassicaceae.

Conclusion

We have reconstructed the backbone of the Brassicaceae phylogeny, identified six major clades that harbor the diversity of this economically important family, and provided putative placement of 16 taxa that were not resolved previously within the family. We also identified genes exhibiting significant directional change in expression or evidence for adaptive evolution that may be associated with leaf complexity. Our phylogenomic study provides a phylogenetic framework for future genomic, developmental and evolutionary studies among mustards.

Acknowledgements

We thank: Angela Hay and Laura Lagomarsino for critically reading the manuscript; Aaron Liston for advice on target capture sequencing; Bonn Botanical Garden (BONN), Harvard University Herbaria (GH and A), Berlin Botanical Garden and Botanical Museum (B), Missouri Botanical Garden (MO), the Royal Botanical Garden Kew (K), the Herbarium at the Bulgarian Academy of Sciences (SOM), BGV-UPM 'César Gómez Campo', Laura Lagomarsino, Daniel Santamaría-Aguilar, Mike Kintgen and Carlos Alonso-Blanco for plant material; Asis Hallab and Baoxing Song for providing alignments of the orthologous genes used in probe design; and Alison Devault and Jake Enk at Arbor Biosciences for probe design, library preparation and capture. We are grateful to BrassiBase (http://brassibase.cos.uni-heidelberg.de) for sharing seeds and expertise. Work on crucifer development and diversity in the Tsiantis Laboratory is supported by Deutsche Forschungsgemeinschaft 'Adaptomics' grants

TS 229/1-1 and SFB (Sonderforschungbereit) 680, and core grant by the Max Planck Society. MT also acknowledges support from CEPLAS Cluster of Excellence. LAN and PS were supported by Alexander von Humboldt Research Fellowships. CDB was supported by NSF Plant Genome Research Grant 1238731.

Author contributions

LAN and MT conceived and planned the study; LAN conducted all experimental work, and collected and curated sequence data; LAN and PS performed the phylogenomic analyses with input from CDB; LAN, BN and DF performed the transcriptome analyses, XG contributed sequenced genome data and bioinformatics expertise; IAAl-S provided taxonomic expertise and advice on the biology of Brassicaceae; LAN and MT wrote the paper with input from CDB whose discussions with MT over many years were instrumental for conceiving the study; and all authors approved the manuscript. MT provided funding and directed the study.

ORCID

Ihsan A. Al-Shehbaz https://orcid.org/0000-0003-1822-4005
Dmitry Filatov https://orcid.org/0000-0001-8077-5452
Bruno Nevado https://orcid.org/0000-0002-9765-2907
Lachezar A. Nikolov https://orcid.org/0000-0003-1594-6416

References

Alfaro ME, Faircloth BC, Harrington RC, Sorenson L, Friedman M, Thacker CE, Oliveros CH, Cerny D, Near TJ. 2018. Explosive diversification of marine fishes at the Cretaceous-Paleogene boundary. *Nature Ecology and Evolution* **2**: 688–696.

Al-Shehbaz IA. 2012. A generic and tribal synopsis of the Brassicaceae (Cruciferae). *Taxon* **61**: 931–954.

Appel O, Al-Shehbaz IA. 2003. Cruciferae. In: Kubitzki K, ed. *The families and genera of vascular plants*, vol 5. Berlin, Germany: Springer, 75–174.

Bailey CD, Koch MA, Mayer M, Mummenhoff K, O'Kane SL Jr, Warwick SI, Windham MD, Al-Shehbaz IA. 2006. Toward a global phylogeny of the Brassicaceae. *Molecular Biology and Evolution* **23**: 2142–2160.

Bar M, Ori N. 2015. Compound leaf development in model plant species. *Current Opinion in Plant Biology* **23**: 61–69.

Barkoulas M, Hay A, Kougioumoutzi E, Tsiantis M. 2008. A developmental framework for dissected leaf formation in the *Arabidopsis* relative *Cardamine hirsuta*. *Nature Genetics* **40**: 1136–1141.

Beilstein MA, Al-Shehbaz IA, Kellogg EA. 2006. Brassicaceae phylogeny and trichome evolution. *American Journal of Botany* **93**: 607–619.

Beilstein MA, Al-Shehbaz IA, Mathews S, Kellogg EA. 2008. Brassicaceae phylogeny inferred from phytochrome A and *ndhF* data: tribes and trichomes revisited. *American Journal of Botany* **95**: 1307–1327.

Bogdanowicz D, Giaro K, Wrobel B. 2012. TreeCmp: comparison of trees in polynomial time. *Evolutionary Bioinformatics* **8**: 475–487.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequencing data. *Bioinformatics* **30**: 2114–2120.

Buddenhagen C, Lemmon AR, Lemmon EM, Bruhl J, Cappa J, Clement WL, Donoghue M, Edwards EJ, Hipp AL, Kortyna M *et al.* 2016. Anchored Phylogenomics of angiosperms I: assessing the robustness of phylogenetic estimates. *bioRxiv* 086298; doi: 10.1101/086298.

Butler MA, King AA. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *American Naturalist* **164**: 683–695.

Chen H, Deng T, Yue J, Al-Shehbaz IA, Sun H. 2016. Molecular phylogeny reveals the non-monophyly of tribe Yinshanieae (Brassicaceae) and description of a new tribe, Hilliellieae. *Plant Diversity* **38**: 171–182.

Dunn CW, Zapata F, Munro C, Siebert S, Hejnol A. 2017. Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proceedings of the National Academy of Sciences, USA* **115**: E409–E417.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**: 157.

Espeland M, Breinholt J, Willmott KR, Warren AD, Vila R, Toussaint EFA, Maunsell SC, Aduse-Poku K, Talavera G, Eastwood R *et al.* 2018. A comprehensive and dated phylogenomic analysis of butterflies. *Current Biology* **28**: P770–P778.

Faircloth BC, Branstetter MG, White ND, Brady SG. 2015. Target enrichment of ultraconservative elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Molecular Ecology Resources* **15**: 489–501.

Fares MA, Byrne KP, Wolfe KH. 2005. Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of *Saccharomyces* species. *Molecular Biology and Evolution* **23**: 245–253.

Forsythe ES, Nelson ADL, Beilstein MA. 2017. Epistatic interactions drive biased gene retention in the face of massive nuclear introgression. *bioRxiv* doi: 10.1101/197087.

Fragoso-Martínez I, Salazar GA, Martínez-Gordillo M, Magallón S, Sánchez-Reyes L, Lemmon ME, Lemmon AR, Sazatornil F, Mendoza CG. 2017. A pilot study applying the plant Anchored Hybrid Enrichment method to New World sages (*Salvia* subgenus *Calosphace*; Lamiaceae). *Molecular Phylogenetics and Evolution* **117**: 124–134.

Franzke A, Lysak MA, Al-Shehbaz IA, Koch MA, Mummenhoff K. 2010. Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends in Plant Science* **16**: 108–116.

Gan X, Hay A, Kwantes M, Haberer G, Hallab A, Dello Ioio R, Hofhuis H, Pieper B, Cartolano M, Neumann U *et al.* 2016. The *Cardamine hirsuta* genome offers insight into the evolution of morphological diversity. *Nature Plants* **2**: 16167.

German DA, Friesen NW. 2014. *Shehbazia* (Shehbazieae, Cruciferae), a new monotypic genus and tribe of hybrid origin from Tibet. *Turczaninowia* **17**: 17–23.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al.* 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**: 644–665.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**: 696–704.

Heath TA, Hedtke SM, Hillis DM. 2008. Taxon sampling and the accuracy of phylogenetic analyses. *Journal of Systematics and Evolution* **46**: 239–257.

Hedtke SM, Morgan MJ, Cannatella DC, Hillis DM. 2013. Targeted enrichment: maximizing orthologous gene comparisons across deep evolutionary time. *PLoS ONE* **8**: e67908.

Hohmann N, Wolf EM, Lysak MA, Koch MA. 2015. A time-calibrated road map of Brassicaceae species radiation and evolutionary history. *Plant Cell* **27**: 2770–2784.

Huang C-H, Sun R, Hu Y, Zeng L, Zhang N, Cai L, Zhang Q, Koch MA, Al-Shehbaz I, Edger PP *et al.* 2015. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Molecular Biology and Evolution* **33**: 394–412.

Initiative The Arabidopsis Genome. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.

Jackson S, Xiong Y. 2010. CRL4s: the CUL4-RING E3 ubiquitin ligases. *Trends in Biochemical Sciences* **34**: 562–570.

Johnson MG, Gardner EM, Liu Y, Medina R, Goffinet B, Shaw AJ, Zerega NJC, Wickett NJ. 2016. HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* **4**: apps.1600016.

Johnson MG, Pokorny L, Dodsworth S, Botigue LR, Cowan RS, Devault A, Eiserhardt WL, Epitawalage N, Forest F, Kim JT *et al.* 2018. A universal

probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *bioRxiv* doi: 10.1101/361618.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research* **30**: 3059–3066.

Kent WJ. 2002. BLAT – the BLAST-like alignment tool. *Genome Research* **12**: 656–664.

Kiefer M, Schmickl R, German DA, Lysak M, Al-Shehbaz IA, Franzke A, Mummenhoff K, Stamatakis A, Koch MA. 2014. BrassiBase: introduction to a novel database on Brassicaceae evolution. *Plant and Cell Physiology* **55**: e3.

King N, Rokas A. 2017. Embracing uncertainty in reconstructing early animal evolution. *Current Biology* **27**: R1081–R1088.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* **22**: 568–576.

Koenig D, Weigel D. 2015. Beyond the thale: comparative genomics and genetics of *Arabidopsis* relatives. *Nature Reviews Genetics* **16**: 285–298.

Krämer U. 2015. Planting molecular functions in an ecological context with *Arabidopsis thaliana*. *eLife* **4**: e06100.

Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evolutionary Biology* **14**: 82.

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2016. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution* **34**: 772–773.

Lemmon EM, Lemmon AR. 2013. High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* **44**: 19.1–19.23.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li J, Witten DM, Johnstone IM, Tibshirani R. 2012. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* **13**: 523–538.

Mandáková T, Li Z, Barker MS, Lysak MA. 2017. Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. *The Plant Journal* **91**: 3–21.

Mandel JR, Dikow RB, Funk VA, Masalia RR, Staton SE, Kozik A, Michelmore RW, Rieseberg LH, Burke JM. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. *Applications in Plant Sciences* **2**: apps.1300085.

McKain M. 2017. *Fast-Plast automated de novo assembly of whole chloroplast genomes.* [WWW document] URL https://github.com/mrmckain/Fast-Plast. [accessed 10 October 2017].

Mohammadin S, Peterse K, van de Kerke SJ, Chatrou LW, Dönmez AA, Mummenhoff K, Pires JC, Edger PP, Al-Shehbaz IA, Schranz ME. 2017. Anatolian origins and diversification of *Aethionema*, the sister lineage to the core Brassicaceae. *American Journal of Botany* **104**: 1042–1054.

Nikolov LA, Tsiantis M. 2017. Using mustard genomes to explore the genetic basis of evolutionary change. *Current Opinion in Plant Biology* **36**: 119–128.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**: 289–290.

Peter IS, Davidson EH. 2011. Evolution of gene regulatory networks controlling body plan development. *Cell* **144**: 970–985.

Provart NJ, Alonso J, Assmann SM, Bergmann D, Brady SM, Brkljacic J, Browse J, Chapple C, Colot V, Cutler S *et al.* 2016. 50 years of Arabidopsis research: highlights and future directions. *New Phytologist* **209**: 921–944.

Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. 2016. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* **534**: S7–S8.

Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: multiple alignment of coding sequences accounting for frameshifts and stop codons. *PLoS ONE* **6**: e22594.

Rast-Somssich MI, Broholm S, Jenkins H, Canales C, Vlad D, Kwantes M, Bilsborough G, Dello Ioio R, Ewing RM, Laufs P *et al.* 2015. Alternate wiring of a KNOXI genetic network underlies differences in leaf development of *A. thaliana* and *C. hirsuta*. *Genes & Development* **29**: 2391–2404.

Ratan A. 2009. *Assembly algorithms for next-generation sequencing data.* Ph.D. Dissertation, Pennsylvania State University. [WWW document] URL etda.libraries.psu.edu.

Rohlfs RV, Harrigan P, Nielsen R. 2014. Modeling gene expression evolution with an extended Ornstein-Uhlenbeck process accounting for within-species variation. *Molecular Biology and Evolution* **31**: 201–211.

Rowan B, Weigel D, Koenig D. 2011. Developmental genetics and new sequencing technologies: the rise of nonmodel organisms. *Developmental Cell* **21**: 65–76.

Shen XX, Hittinger CT, Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology and Evolution* **1**: 0126.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution* **16**: 1114–1116.

Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**: 1246–1247.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Straub SCK, Fishbein M, Livshultz T, Foster Z, Parks M, Weitemier K, Cronn RC, Liston A. 2011. Developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics* **12**: 211.

Struck TH. 2014. TreSpEx – detection of misleading signal in phylogenetic reconstructions based on tree information. *Evolutionary Bioinformatics* **10**: 51–67.

Vatanparast M, Powell A, Doyle JJ, Egan AN. 2018. Targeting legume loci: a comparison of three methods for target enrichment bait design in Leguminosae phylogenomics. *Applications in Plant Sciences* **6**: e1036.

Vlad D, Kierzkowski D, Rast MI, Vuolo F, Dello Ioio R, Galinha C, Gan X, Hajheidari M, Hay A, Smith RS *et al.* 2014. Leaf shape evolution through duplication, regulatory diversification, and loss of a homeobox gene. *Science* **343**: 780–783.

Walker JF, Yang Y, Moore MJ, Mikenas J, Timoneda A, Brockington SF, Smith SA. 2017. Widespread paleopolyploidy, gene tree conflict, and recalcitrant relationships among the carnivorous Caryophyllales. *American Journal of Botany* **104**: 858–867.

Wanke S, Granados Mendoza C, Müller S, Paizanni Guillén A, Neinhuis C, Lemmon AR, Lemmon EM, Samain M-S. 2017. Recalcitrant deep and shallow nodes in *Aristolochia* (Aristolochiaceae) illuminated using anchored hybrid enrichment. *Molecular Phylogenetics and Evolution* **117**: 111–123.

Warwick SI, Mummenhof K, Sauder CA, Koch MA, Al-Shehbaz IA. 2010. Closing the gaps: phylogenetic relationships in the Brassicaceae based on DNA sequence data of nuclear ribosomal ITS region. *Plant Systematics and Evolution* **285**: 209–232.

Weitemier K, Straub SC, Cronn RC, Fishbein M, Schmickl R, McDonnell A, Liston A. 2014. Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* **2**: apps.1400042.

Wendel JF, Doyle JJ. 1998. Phylogenetic incongruence: window into genome history and molecular evolution. In: Soltis DE, Soltis PS, Doyle JJ, eds. *Molecular systematics of plants II*. Boston, MA, USA: Springer, 265–296.

Willing E-M, Rawat V, Mandáková T, Maumus F, James GV, Nordström KJV, Becker C, Warthmann N, Chica C, Szarzynska B *et al.* 2015. Genome expansion of *Arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation. *Nature Plants* **1**: 14023.

Yang Z. 2007. PAML 4: phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**: 1586–1591.

Zhou Z, Lutteropp S, Czech L, Stamatakis A, von Looz M, Rokas A. 2017. Quartet-based computation of internode certainty provide accurate and robust measures of phylogenetic incongruence. *bioArxiv* doi: 10.1101/168526.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Dataset S1 Exons included in each dataset.

Dataset S2 Phylogenetic trees derived from each dataset.

Dataset S3 Expression levels of leaf-expressed genes in 8 species of Brassicaceae that deviate from neutral expectations.

Fig. S1 Phylogenetic reconstruction workflow.

Fig. S2 Plastid coding sequences and sequenced plastid genomes obtained from genome skimming.

Fig. S3 Density plots of metrics used to partition the exon set into 27 datasets after excluding the extreme outliers for each metric

Fig. S4 Summary of the concatenation results represented by the tree space of the eight unique topologies from the maximum-likelihood analysis, and the topological differences among them.

Fig. S5 Phylogenetic informativeness of the loci from the most inclusive datasets that produced unique topologies.

Fig. S6 Summary of the concatenation results after excluding exons with phylogenetic informativeness score > 10.

Fig. S7 Pruning of taxa with unstable placement in different datasets and analyses.

Fig. S8 Summary of the ASTRAL-II results represented by the tree space of the seven unique topologies, and the topological differences among them.

Fig. S9 Placing unassigned to a tribe species with expanded taxon sampling of associated sister tribes.

Fig. S10 Plastid coding gene phylogeny.

Fig. S11 Comparison between tree topologies derived from nuclear and plastid loci.

Methods S1 Supplementary Materials and Methods.

Notes S1 Tree description and placing taxa unassigned to a tribe.

Table S1 Taxon and voucher information for species with newly generated sequence data.

Table S2 Accession information for species with sequenced nuclear genomes included in this study.

Table S3 Accession information for species with sequenced plastid genomes.

Table S4 Sequencing reads and contig statistics for nuclear loci.

Table S5 Skimmed reads and whole plastome statistics.

Table S6 Concatenated dataset statistics.

Table S7 AU tests to evaluate the support of a given matrix for a given topology.

Table S8 Accession information of marker genes used for the infratribal placement of unassigned to a tribe taxa.

Table S9 Summary of the phylogenetic placement of unassigned to a tribe taxa and taxonomic notes.

Table S10 Transcriptome assembly statistics.

Table S11 Genes that deviate from the background optimal expression value in the family in the core Brassicaceae.

Table S12 Genes that deviate from the background optimal expression value in the family along the branches leading to species with complex leaves.

Table S13 Genes and protein residues under positive selection in species with complex leaves.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.