nature plants

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41477-019-0486-9

In the format provided by the authors and unedited.

Interspecies association mapping links reduced CG to TG substitution rates to the loss of gene-body methylation

Christiane Kiefer[1,2,4], Eva-Maria Willing[1,3,4], Wen-Biao Jiao[1], Hequan Sun[1], Mathieu Piednoël[1], Ulrike Hümann[1], Benjamin Hartwig[1,3], Marcus A. Koch[2] and Korbinian Schneeberger[1]*

[1]Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, Cologne, Germany. [2]Department of Biodiversity and Plant Systematics, Centre for Organismal Studies, Heidelberg University, Heidelberg, Germany. [3]Present address: NEO New Oncology, Cologne, Germany. [4]These authors contributed equally: C. Kiefer, E.-M. Willing. *e-mail: schneeberger@mpipz.mpg.de

Supplementary Information

"**Quantitative inter-species association mapping links differences in CG⇒TG substitution rates in genes to the loss of gene-body methylation**"

Christiane Kiefer, Eva-Maria Willing et al. (2019)

Table of Contents

Supplementary Materials and Methods

Genome assembly and annotation

Selection and origin of species

In total, we selected 47 species based on their placement across the Brassicaceae phylogeny representing 23 of the 51 tribes and one genus (*Schrenkiella*) which had not been assigned to a tribe, yet. Often small tribes in the Brassicaceae comprise rare taxa which are only available as herbarium specimens. To be able to examine characters in live material availability of seed material was used as further criterion, implying that our selection mostly covers the largest tribes of the family. In addition, we used public genome size estimations as selection criterion to avoid neo-polyploids. Complexity to the dataset was added by choosing species from tribe Arabideae and Stevenieae, for which there was conflicting evidence of their evolutionary history[1,2], and morphological convergence might indicate deep reticulate processes. A second case study introducing complexity was intended with *Conringia planisiliqua*. This species has not been assigned yet with a particular tribe, but all other *Conringia* species are placed with tribe Conringieae sister to tribe Coluteocarpeae (*Noccaea* and *Raparia*). We selected two geographically diverse accessions of *Conringia planisiliqua* (from Turkey (labelled with (T)) and Greece (G)) to test for phylogenetic complexity and to use it as internal control for species-specific heterogeneity.

For 17 of the species reference sequences were published before[3–14]. In these cases, we ordered the matching genotypes (expect for *B. rapa*, where a different individual was chosen) and used the publicly available genome sequences. The remaining species were ordered from different botanic gardens or research institutes (Supplementary Table S1).

Plant growth and whole-genome sequencing

All plants were grown in the greenhouse at the Max Planck Institute for Plant Breeding Research, Cologne, Germany, under long day conditions (photoperiod 16h, photoperiod extended and light supplemented if necessary (lamps by Philips, Germany, MGR400 SONT 400W (yellow) and MGR400 HPIT 400 (white)). Plants were fertilized daily with Kristalon Scarlet 7.5 + 12 + 4.5 stock solution (12.8 kg/100L) and Calcinit 15% N, 26% CAO stock solution (9.3 kg/100L), mixed 50:50 and diluted to 1.2 µS. Leaves were harvested and shock frozen in liquid nitrogen. Genomic DNA was extracted using the Qiagen DNEasy Mini Kit (Qiagen, Germany) following the manufacturer's protocol.

Illumina short-read library preparation and whole-genome sequencing of 31 species has

been performed at the Max Planck Genome center using HiSeq2500/3000 sequencing machines. The genomes of *Turritis glabra* and *Pseudoturritis turrita* were in addition sequenced with PacBio technology using the RSII platform with P6C4 sequencing reagents. In total, 14.1 Gb reads from 25 SMRT cells and 22.6 Gb reads from 36 SMRT cell were generated for *T. glabra* and *P. turrita*. Amount of reads and estimated genome coverage are shown in Supplementary Table S1.

Illumina short-read assemblies

Illumina short read assemblies were generated using the assembly tool soap with default parameters (version 2.04)[15]. Best *k*-mer size was estimated using SPAdes (version 3.5.0)[16]. To remove redundant contigs that resulted from residual heterozygosity and individual haplotype assembly, we ran Redundans on each of the final soap assemblies using default parameters (version 0.12c)[17].

PacBio long-read assemblies

Raw sequencing reads were filtered to remove short (<500 bp) or low-quality reads (QV < 80) using the SMRT Analysis software (v2.3). Filtered subreads were used for *de novo* assembly using FALCON (v3.0)[18]. For *T. glabra*, reads longer than 2,000 bp were used for self-correction and corrected reads longer than 6,000 bp were used for assembly. For *P. turrita* reads longer 3,000 bp and corrected reads longer than 5,000 bp were used for assembly. Other parameters were set as "--max_diff 100 --max_cov 150 --min_cov 5" for both genome assemblies. The assembled sequence was further polished using the filtered subreads using Quiver (https://github.com/PacificBiosciences/GenomicConsensus) and finally corrected using Illumina short-reads. Around 6.0 Gb (*T. glabra*) and 9.1 Gb (*P. turrita*) of Illumina reads were mapped to the polished assemblies using BWA v0.7.12[19]. Fixed differences between short read alignments and new assembly were called using SAMtools[20], and further used for generating the final corrected sequences.

Gene annotations

Gene annotations were performed for all newly assembled species, while for species with existing assembly we used the publicly available annotations. First we aligned all predicted protein sequences of *Arabidopsis thaliana*, *Arabidopsis lyrata, Cardamine hirsuta, Capsella rubella, Eutrema salsugnineum, Schrenkiella parvula, Arabis alpina* and *Arabis montbretiana* (defined as the core species) to all new assemblies using Scipio[21] (min_identity=80 min_coverage=75). For gene annotations we used Augustus[22] (--UTR=off,

version v3.0), glimmer[23] (default parameter, version v3.0.3) and snap[24] (default parameter, version 2013-11-29) on each of the assemblies separately. After this we combined the four output files (three from the de novo predictors and one from the homology based scipio predictors) using EVM[25] (min_intron_length 5, version v2012-06-25). Afterwards we determined the assembly completeness using BUSCO[26] for each of the assemblies (including the public assemblies) (default parameters).

TE identification and filtering of genes

To identify DNA repeats from our assemblies, we performed homology-based searches using RepeatMasker (http://www.repeatmasker.org; parameters: '-pa 10 -nolow -qq'). For that purpose, we designed our own custom Brassicaceae repeat database. It compiles every known Brassicaceae DNA repeats from the P_MITE[27], PlantSat[28], TIGR[29], Repbase[30] and PGSB[31] databases. To also consider possibly more distant and unknown repeats, we *de novo* identified and annotated repeats from each of the Illumina read datasets using dnaPipeTE[32] that we modified to adapt it to our data. The corresponding repeats were then integrated to our Brassicaceae repeat database, which we used to overlap the gene annotations with the RepeatMasker repeat annotation file and for each gene the exonic overlap with the annotated repeats has been calculated.
Final (and filtered) gene annotations included only those genes which were predicted by three gene annotation tools, which had a repeat overlap of less than 30% of their exonic sequence and which did overlap with at least one protein alignment of any of the core species.

Orthologous groups and gene family calculation

Ortholog calculations using all 48 samples led to instable orthogroups (OGs), where the clustering of genes relied on the presence or absence of individual species. To overcome this problem and to generate stable orthogroups, we first calculated orthologous groups for the core species only (core OGs) using OrthoFinder[33] (default parameters, version 1.1.4). Afterwards, we added each of the remaining species separately and included the genes from the additional species if they did not affect the integrity of the orthogroup as compared to the core OGs. If a core OGs was not present in the new calculation (including an additional genome), it was not clear whether the respective species features an ortholog of this group or not and hence was marked as "na". Only if a core OG was conserved but did not show any gene member of the newly added gene, we assumed that the gene was absent in the newly added species. This is a very conservative approach and ensures that genes are not

falsely annotated as missing due to complications in the assignment of OG.

Orthogroups are calculated based on sequence similarity. This includes clustering of ancient gene duplicates and does not separate ancient paralogs from orthologs. Therefore, for each OG we calculated gene trees based on pairwise alignments of all member of each group (based on hamming distance) and split the trees (respectively the OG) by removing the root node of the tree as long as the resulting sub-trees included overlapping sets of species. This efficiently separated gene duplications that happened before or at the onset of the evolution of the Brassicaceae. However, this also separated outlier genes (placed at the basis of a gene tree). We reintroduced these outlier genes if they otherwise led to species without gene members in this OG.

Calculation of phylogeny

For the calculation of a phylogeny we collected a set of orthologous genes based on the gene sets defined in Zhang et al. and Duarte et al.[34,35]. From these lists, we selected a total of 100 gene families which were reported as being low or single copy in either of the studies and which were consistent with our gene families. We further added gene families from our set (with at least one member of each species) which showed the lowest copy number across the Brassicaceae used here.

We then generated a genotype matrix for each of these gene families based on pairwise alignments between each gene and the *A. thaliana* gene sequence of this gene family. If multiple *A. thaliana* genes were included in this family, we calculated a consensus sequence of these sequences used those as alignment partner for each species. The positions of the *A. thaliana* sequence of each alignment were used to match all pairwise alignments of a gene family and thereby converted them into a table (genotype matrix).

For the phylogenetic tree calculation, all sub-alignments were concatenated into one matrix but genes remained annotated. PartitionFinder[36–38] was used to determine the best partitions of the dataset as well as models to be used for each partition in the phylogenetic tree reconstruction. Phylogenetic reconstruction was done using the program RAxML-ng[39] using the concatenated alignment as well as partitions and corresponding models as input and setting number of bootstrap replicates to 1000. Bootstrap values were plotted on the best ML-tree.

In a second approach, we performed phylogenetic reconstructions by RAxML-ng[39] for each gene independently using the evolutionary models determined above. The resulting best trees were used as input for the program ASTRAL5.6.3[40] in order to calculate a combined phylogenetic tree giving a likelihood for each quartet on each node.

Speciation dating

We dated the split of the basal group including *A. arabicum* as well as the split of *Lineage III* (represented by *E. syriacum*) from the core of the Brassicaceae phylogeny. To date the split of *A. arabicum*, we first determined the best partition scheme of the alignment using PartitionFinder[36–38]. The concatenated alignment, the information on partitions along with the most suitable molecular evolutionary models were used to generate an input file for analysis by BEAST2.5.1[41] using the program BEAUTI2[41]. The splits between *Aethionema* and the remaining taxa as well as *A. thaliana* and Cochlearia were used as calibration points using the dates obtained by Hohmann et al[42]. Partitions and models were kept as unlinked while clocks and trees were treated as linked. The analysis was run for approximately 55 million generations in three independent runs and every 100.000th tree was sampled. Using the program Tracer1.7.1[43] we determined whether the analysis had reached a stationary phase. Trees and logs of the three runs were combined by using the program logcombiner included in the BEAST2.5.1 package. Burn-in was set to 10%. The information obtained from the analysis was summarized and plotted on the best tree using TreeAnnotator, which is also included in the BEAST2.5.1 package.

Analysis of polyploidization events

Dating of the At-α duplication event

Already the initial reference assembly of *A. thaliana* from the year 2000 had revealed an ancestral polyploidization event, which later was found to be shared by all Brassicaceae species and is commonly referred to as the At-α event[6,12,44–47]. We followed a recently published approach to approximate the time of this paleo-polyploidization event using differences within paralogous genes[48,49]. This method is based on individual genomes. We selected 14 diploid species to estimate the dating of the paleo-polyploidization event separately (Supplementary Fig. S4).

First, the gene sets of each of these species were blasted against themselves (ignoring self-hits with e-value cutoff <1e-10). The genes were then clustered based on their existing blast hit connections using MCL with default parameters (version 14-137)[50,51]. Clusters larger than 100 were ignored, for all other clusters we performed a codon-informed pairwise alignments using all possible pairs of genes within each cluster and thereby estimated *Ks* for each such gene pair. We then hierarchically clustered the genes of each cluster using *Ks* as distance measure between any given genes (gene pairs which failed to generate proper *Ks* estimates were assigned an artificial *Ks* value of 6). The cluster-based trees were split into subtrees in

a way that each subtree did not show any *Ks* value larger than 5 in order to split false or ancient duplicates. The resulting subtrees were then taken to refine the set of gene clusters. Finally, we removed clusters, which were likely to relate to transposable elements, by removing those clusters, which were larger than 12 members or did not show orthologs in at least four of the species in the core set (see orthologous group calculation). The *Ks* values calculated between all genes pairs of all remaining cluster were then used as final *Ks* distribution for the respective genome.

To translate the *Ks* distributions into a time estimate, we used a kernel density estimation (KDE) of the *Ks* distribution and used resampling for confidence interval calculation. It has been described before that local maxima of the KDE (of a *Ks* distribution) within the interval 0.5 and 1.3 refers to the At-α event within Brassicaceae genomes. For each of the 14 genomes we found a single peak in the KDE within this region. This value was used as species-specific estimate for the dating of the At-α event. Based on 1000 re-samplings (i.e. randomly selecting half of the *Ks* values) we repeated estimation of this value to get a 90% confidence interval. Both dating value and confidence interval were then translated into a time estimate using the mutation rate estimated by Ossowski et al[52].

Finally, the final estimate for the date of the At-α event was an average value of the 14 individual estimates (~56 mya) while the final confidence interval was selected to include all individually estimated confidence intervals (47.9-62.8 mya) (Supplementary Fig. S4).

Identification of meso/neo-polyploids from assembly data using machine learning

Though many of the extant Brassicaceae species re-diploidized after this whole-genome duplication, neo-polyploidization is very common among the Brassicaceae[53] and 13 meso-polyploidization events that happened after the At-α have been identified by comparative chromosome painting and transcriptome analyses[54,55], even though the distinction between neo-polyploidy and meso-polyploidy remains difficult for some genomes[42].

We used a support vector machine (svm, R library “e1071”; https://cran.r-project.org/web/packages/e1071/index.html) to identify meso/neo-polyploids in the set of species, for which no clear record on ploidy level could be found. We used 18 species for which ploidy was reported before to train the svm using four different features from six meso-polyploids and 12 diploids (Supplementary Table S2) (parameter used for training: method="C-classification", kernel="linear"). The features used included number of peaks in the *Ks* distribution, percentage of duplicated genes in the BUSCO analysis (see section on gene annotations), number of single-copy genes across 2,036 gene families with genes from all species, and average gene family size.

The *Ks* distributions for feature selection were calculated for each species separately and were established similarly to the *Ks* distributions used for the dating of the At-α duplication event, with the difference that we omit cluster splitting based on high *Ks* values (Supplementary Fig. S5). Peaks, which were indicating duplication events younger than the At-α event were counted. Occasionally we could observe multiple of such peaks, however, we only counted the second peaks if they were larger than the first one to avoid noise. Moreover, for some of the genomes with two peaks no clear peak relating to the At-alpha duplication could be observed. In these cases, we also annotated only one peak. Finally, we found 21 species with at least one additional peak in their *Ks* distribution (including all six known meso-polyploids). Only *Physaria fenderli* showed two clear peaks, which were not shared with the other Physaria species in our set.

We tested the performance of the svm on the species with known ploidy within a leave-one-out experiment. For this we trained the svm 18 different times. For each of these runs, we left out one of the species, and then used the respective model to predict the poylploidization of the one that was left out. In 17 out of 18 cases the svm predicted the correct ploidy level, expect for *Cochlearia officinalis*. The feature values for this genome tend to be between the polyploids and the diploids. As this genome is known to have two additional rounds of polyploidizations after the At-α event[54] this might result from imperfect assembly of the highly polyploid plant, where the short-read based assembly was not sufficient to assemble the highly similar copies of the genome. Finally, we trained the svm on all 18 species and applied it to the remaining 30 species. This identified 16 additional meso-/neo-polyploid genomes.

Shared polyploidization event identification

We implemented a modified version of MAPS (Multi-tAxon Paleopolyploidy Search) to find shared polyploidization events between species[56]. For this we tested each closely-related pair of polyploid species if there polyploidization could be shared. In case it was shared, other closely-related species were checked as well if they also share the same event. This iterative method tested all branches of the phylogeny, which potentially could reveal shared events (see Fig. 1, all branches with numbers have been tested).

To test if a given pair of polyploids shared an event we used gene families with two copies in both species. For each gene family, gene trees were calculated based on codon-informed multiple sequence alignments between the genes of each family including the respective ortholog of a diploid outgroup species (using the respective *A. thaliana* ortholog as outgroup unless otherwise indicated (Supplementary Table S3). We only considered the subtrees of the entire gene trees, which had the *A. thaliana* gene as outgroup. If this was not possible

we did not analyze the gene family further (outgroup test failed). Moreover, also those gene subtrees, which did include only a single copy of one of the species, were not further processed (single copy genes). For all others, we counted the number of gene trees that supported a duplication event in the common ancestry of the two species and compared this number of the number of gene trees that did not support a common duplication (Supplementary Fig. S7).

Functional annotation of convergent gene loss events in meso/neo-polyploids

Diploidization-based gene loss in angiosperms has been described as a non-random, convergent process where lost paralogs are enriched for basic house-holding functions[54,57] and retained paralogs are enriched for genes with function in secondary metabolite biosynthesis[54]. Our set of species, including comparatively young polyploids, is particularly well-suited for the analysis of genes which are consistently lost soon after polyploidization. When requiring single-copy genes in all mesopolyploids (or all mesopolyploids except for one) we found 494 gene families. Of those, 490 gene families featured an *A. thaliana* member as well, which were used to for the analysis of functional enrichment of these gene families. GO enrichment was assessed with agriGO v2.0[58]. 'Biological Process' were enriched for reproduction including embryo development, meiosis and recombination, while 'Molecular Functions' were enriched for ATP binding and helicase activity. Unexpectedly, however, this set of genes was strikingly enriched for genes connected to cell organelles within the 'Cellular component' analysis. Overall, 136 (28%; p-value < 10^{-12}) and 115 (23%; p-value < 10^{-8}) out of 490 genes were related to "chloroplasts" or "mitochondria", respectively (vs. 15% or 13% among all genes; Supplementary Fig. S16). This suggests that organelle-related nuclear genes in allo-polyploids are among the first genes to be removed after genome hybridization, where only one parental species contributes the organellar genomes.

Phylogenetic association mapping

Measuring leaf shape complexity between species

Plants were grown under long day conditions (16 hours photoperiod) in the greenhouses at the Max Planck Institute for Plant Breeding Research/Cologne/Germany (details see above). We took all rosette and cauline leaves at the stage of flowering from one individual plant of 39 species. We used ImageJ to analyze scans of each of these leaves. Drawing a

convex hull around each leaf allowed to measure the percentage of (white) pixels not related to each leaf within each convex hull. The average percentage across all leaves of a species was used as measurement of leave shape complexity (Supplementary Table S4).

Assessing genic substitution rates for each species

Genic substitution rates were estimated based on the genotype matrices, which we had generated for the phylogenetic tree calculation (see above). For each species, we calculated the percentage of observed base substitutions as compared to all positions with the respective ancestral base. Ancestral bases were defined using the two species at the base of the phylogeny (*E. syriacum* and *A. arabicum*). In order to account for faster and slower evolving species, these values were normalized with the respective branch length in the phylogenetic tree in order receive a substitution rate value.

Between-species association mapping statistics

Closely-related species are assumed to have more similar traits and genotypes because of their shared phylogenetic relationship[59]. This presents a problem for phylogenetic association mapping as it implies that the individual data points are not independent and introduce covariance, which is only based on the shared evolutionary history. Phylogenetic comparative methods can control for this unknown covariance. Their basic idea is to use the phylogenetic relationship (i.e. phylogenetic tree) of the species to produce an estimate of the expected covariance. These estimates can then be added to the equations in the form of a variance-covariance matrix and thereby control for the evolutionary relationship of the samples.

To calculate these regressions, we used the R library “caper”. The phylogenetic tree used for the phylogenetic correction was identical to the phylogenetic tree as described above. The genotype was either the size distribution of the gene families (i.e. the number of genes found in each of the species) or the absence/presence patterns of genes in each of the speces. The phenotypes used here were leaf shape complexity or CG=>TG substitution rates (see above). For both phenotypes we individually compared a gene family against the phenotypic values if at least three species featured genes in this family, at least three species did not show any genes and there were not more than 20% of the species without reliable information on the number of genes in this family. In order to include binary discrete data in the genotype (absence/presence patterns), we used the function *brunch*, which is geared for categorial variables (differences can only occur if polytomies are present in the phylogeny, which was not the case here)[60]. For the regression of gene-copy number

differences we used caper's analogous function *crunch* geared for continuous data. In each of the regressions the phenotype was used as response variable, while a fixed effect was used for the genotype. The p-value of the slope of the regression was calculated with ANOVA and was taken as the individual p-value for the significance of the association of the genotype and the respective phenotype.

Power simulations

We simulated genotypes (i.e. gene families with different gene copy numbers) following an evolutionary model. Based on the actual changes of gene copy number in the phylogeny of the Brassicaceae we simulated the evolution of gene families along the individual branches of simulated phylogenies with 16, 32, and 64 species. The probability for a change at each branch was approximated after the deletion rates in the actual Brassicaceae species analyzed and defined by the depth in the phylogeny and linearly-scaled up to a maximum of 0.2. If a change was observed, there was a 60% chance for a duplication compared to a 40% chance for a deletion. If a gene copy number reached 0, it was no subject for change anymore.

For each of the three different phylogenies we simulated 250 different gene sets. For each of the 750 gene sets we simulated different phenotypes. The phenotypes were calculated based on 1, 5 or 10 different effects (i.e. gene families). The values were calculated either based on gene absence/presence patterns or on the copy number (to simulate dosage effects). For phenotypes, which were dependent on more effects, we lowered the effect size for the additional genes (to simulate major and minor effects). For each phenotype, we further calculated five different levels of environmental effects ranging from 0 to 100% of the fraction of the effect size that is determined by the genotype alone. Effects were only calculated for gene families with sufficient genetic diversity (i.e. at least five values for two different gene count categories including absence of genes as one of the categories).

Methylome sequencing

We have generated methylome data for seven species (including two replicates for each species) and combined them with the data published in Seymour et al[61]. Our data from *C. planisiliqua* (G) were already published before by Bewick et al[62]. The data for *E. salsugineum*, a species that was also analyzed in Bewick et al, was sequenced independently.

Plants cultivation

Seeds were sown on standard soil (Balster, Sinntal-Altengronau/Germany, type Mini-Tray)

and stratified at 4°C for 7 days. Afterwards, plants were grown under long day conditions (16 hours light photoperiod) in the greenhouse at the Max Planck Institute for Plant Breeding Research (details see above). Seeds which did not germinate on soil were sterilized by washing in 70% ethanol for 7 minutes followed by an additional wash step with 100% ethanol for 5 minutes. The sterilized seeds were sown on standard growth medium under sterile conditions and stratified at 4°C for one week. For germination, the seeds were placed in a growth chamber with long day conditions (16 hours photoperiod). Seedlings were then transferred to standard soil in the greenhouse and grow under long day conditions (16 hours photoperiod, details see above). Leaf samples were taken three weeks after sowing and shock frozen in liquid nitrogen.

Plant methylome sequencing

DNA was extracted using the Qiagen DNEasy Plant Mini Kit (Qiagen, Germany) according to the manufacturer's protocol. For methylome sequencing the NEXTflex Bisulfite Library Prep Kit for Illumina Sequencing (Bioo Scientific) was used including a gDNA conversion step with the EZ DNA Methylation-Gold kit (Zymo Research). The resultant DNA was used for conventional high-throughput sequencing on an Illumina HiSeq machine.

Data analysis

Short reads of each sample (i.e. each replicate) were aligned against the respective reference sequences using BS-Seeker2 (v2.0.8; bs_seeker2-build.py: --aligner=bowtie2; bs_seeker2-align.py: --aligner=bowtie2 -m 0.1 --bt2-p 30 --bt2--end-to-end)[63]. Alignments were sorted and duplicated read pair alignments were removed to minimize the effect of PCR duplicates. After alignments, BS-Seeker2 was used again to call methylated sites (bs_seeker2-call_methylation.py: default parameters). Bisulfite conversion rates were calculated with lambda phage DNA, which was spiked in the DNA of each sample. All fourteen samples showed conversion rates of over 99%.

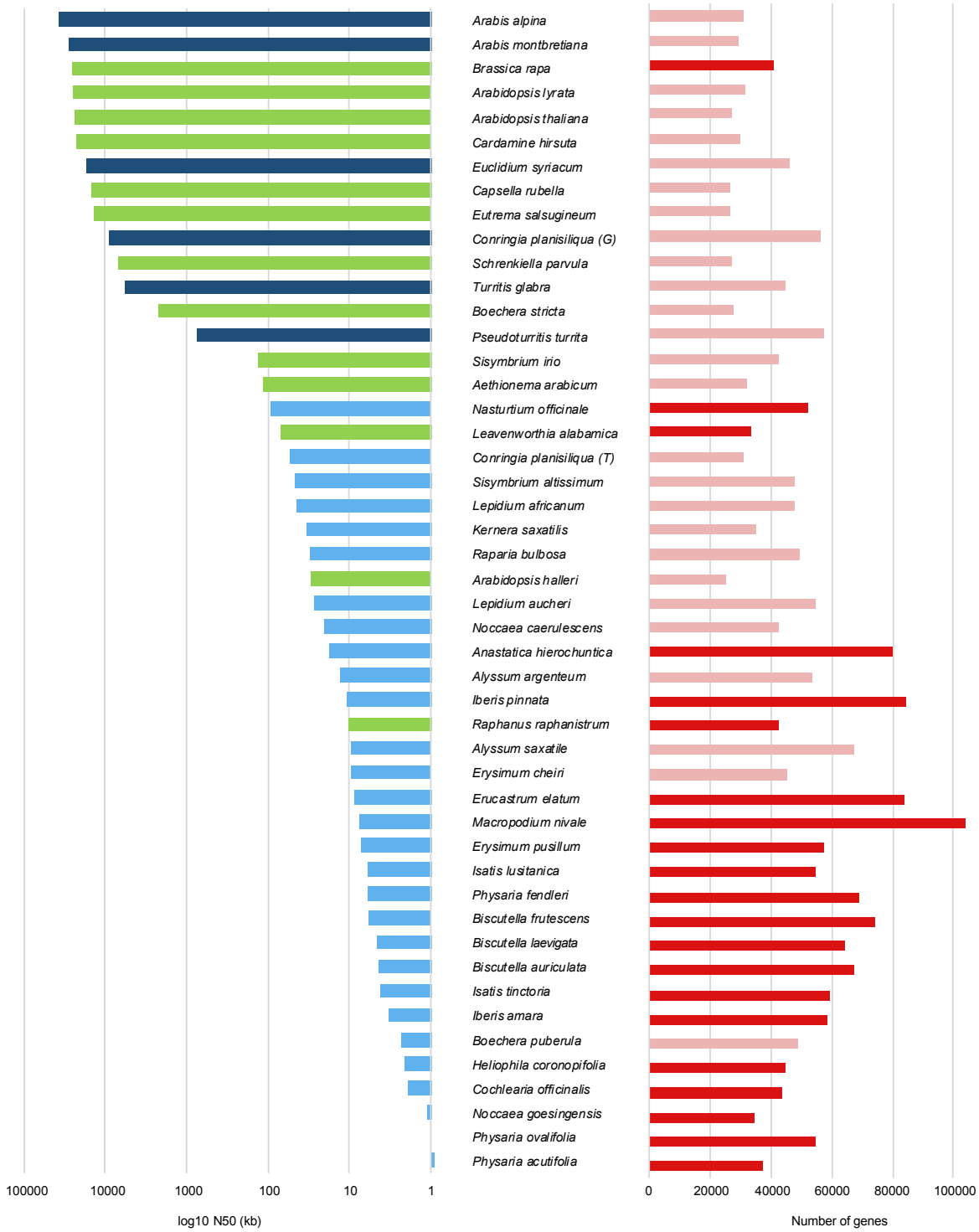
For each sample we estimated the degree of gene body methylation across the genome and correlated the degree to gene body methylation with the degree of gene body methylation as assessed with the respective replicate. As the degree of gene body methylation between all pairs of replicates were highly similar, we combined the information of both replicates for final metagene analyses.

Calculation and definition of gene-body methylation

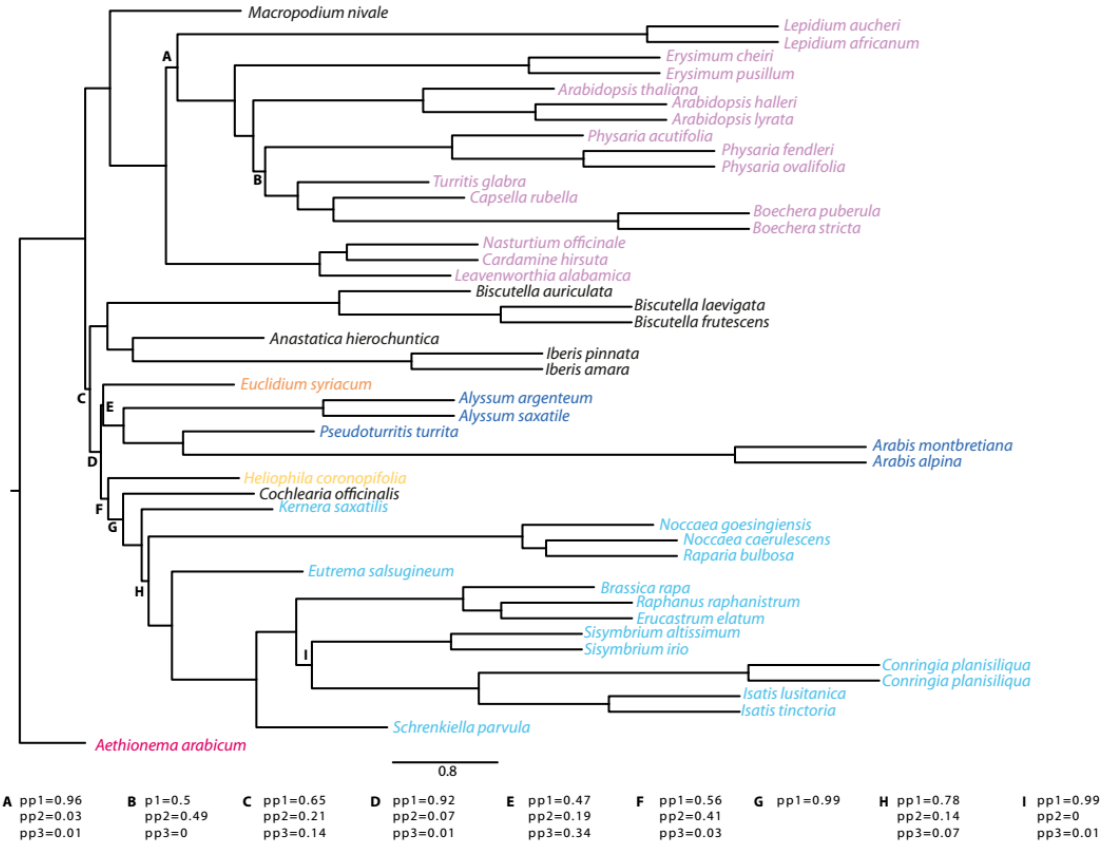
We calculated orthologous groups across the ten species for which we analyzed the methylomes using OrthoFinder[33]. This revealed 6,632 single-copy orthogroups. For each species and for each of these genes, we calculated average gene body methylation scores

in the following way. First, the level of cytosine methylation at each site was calculated as the fraction of reads that showed the footprint of methylation as compared to those read alignments that did not show a footprint of methylation (using both replicates of each species). This pattern was further tested for significance (alpha-level of 0.05) using a binomial test[62]. The CG (or gbM) methylation status of a gene was then calculated as the average methylation level across each of the sites that were sufficiently covered and significantly methylated (if methylated). To remove genes with unreliable patterns, we filtered for genes where (in each species) at least 30% of the cytosines in CG context (with a minimum of 20 cytosines) were covered by more than five reads and where the methylation (if present) was significant. This revealed 1,772 genes with high quality methylation level states across all ten species (values shown in Fig. 4F). To assess genome-wide values, we calculated the average of these 1,772 gene methylation states across each species (shown in Fig. 4E). We further used a hard cutoff of 0.1 for the methylation status to define unmethylated genes (grey bar in Fig. 4F).

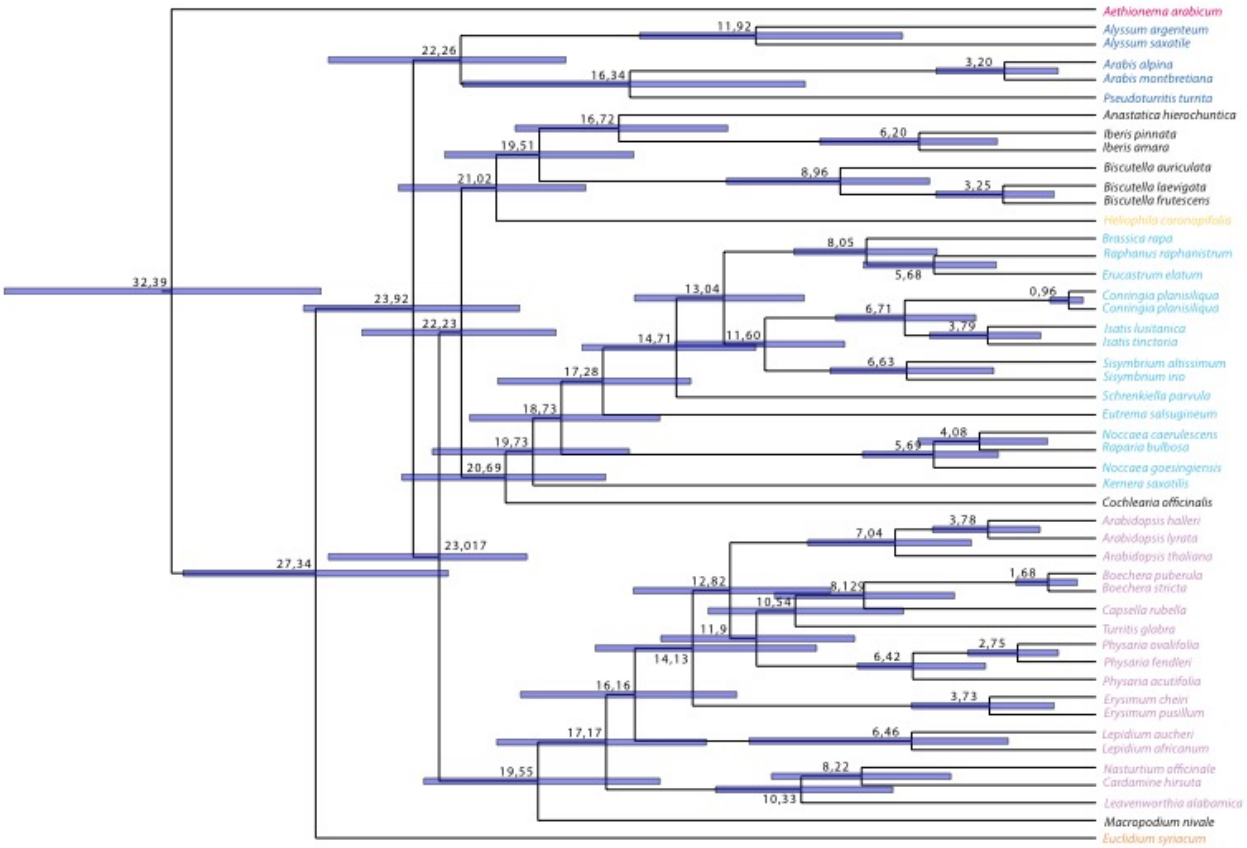
Supplementary Figures



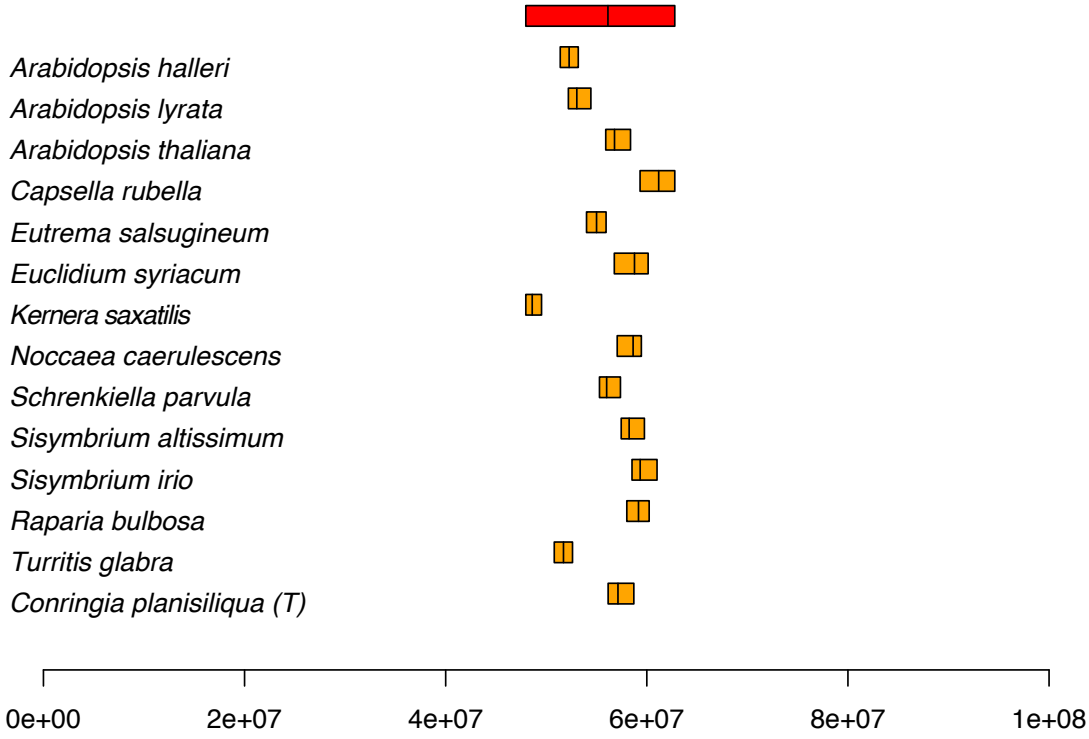
Supplementary Figure S1. Assembly contiguity (N50) and number of genes. Left side shows the N50 values for each genome assembly (public reference sequences (green); high quality assemblies generated by us for this and other projects[3,64] (*A. montbretiana* assembly prior to release) (dark blue); short read assemblies generated by us (light blue)). Right side shows the number of genes annotated for each species (diploid species (light red); meso/neo-polyploids (dark red)).



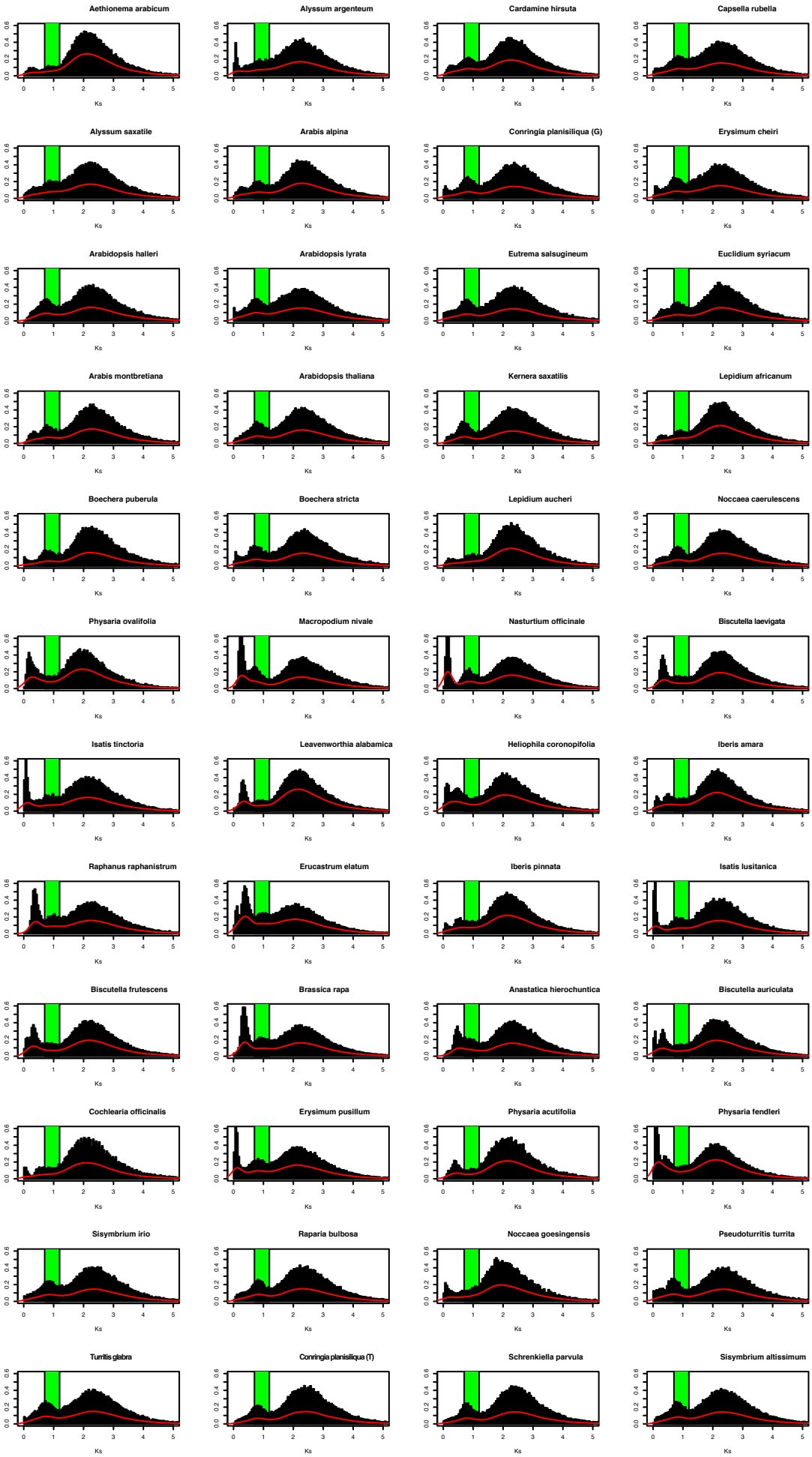
Supplementary Figure S2. Phylogeny of 47 Brassicaceae species. To obtain a phylogenetic tree reconstructed by a different method, we performed a second phylogenetic reconstruction using the quartet method as implemented in ASTRAL5.6.340. This second phylogeny was largely congruent with the first reconstruction obtained by RAxML (see Fig. 1). Numbers at the branching points indicate the probabilities of the quartets.



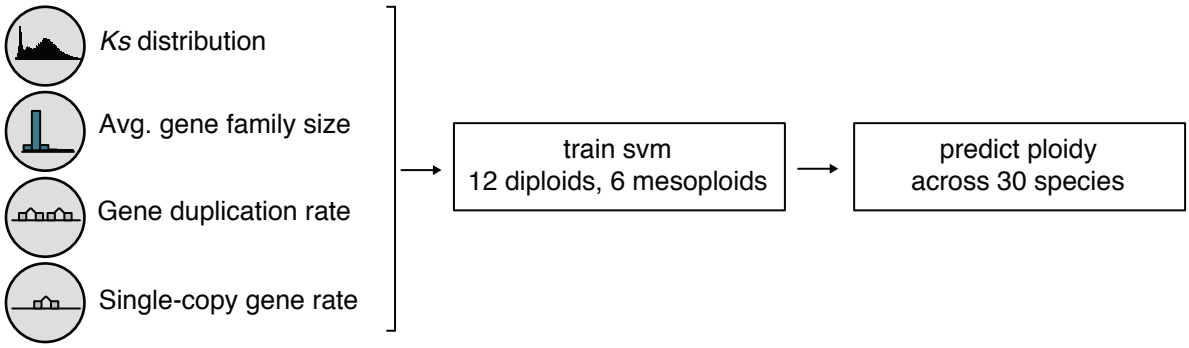
Supplementary Figure S3. BEAST analysis of 47 Brassicaceae species. BEAST was used for dating the branching points of the phylogenetic reconstructions generated with RAxML and ASTRAL. The dates were largely congruent with published estimates, like Hohmann et al42. Species names were colored according to the grouping defined by Nikolov et al65.



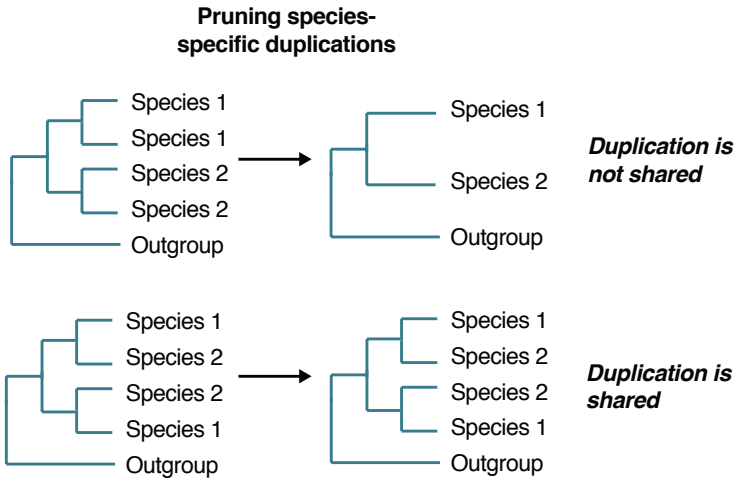
Supplementary Figure S4. Dating of the At-α duplication event using 14 diploid species. A duplication event dating was independently performed for each species (see Supplementary Methods). The final estimate is an average value of the individual estimates (56 mya) while the final confidence interval was selected to include all individually estimated confidence intervals (47.9-62.8 mya). (y-axis: mya).


Aethionema arabicum
Alyssum argenteum
Cardamine hirsuta
Capsella rubella
Alyssum saxatile
Arabis alpina
Conringia planisiliqua (G)
Erysimum cheiri
Arabidopsis halleri
Arabidopsis lyrata
Eutrema salsugineum
Euclidium syriacum
Arabis montbretiana
Arabidopsis thaliana
Kernera saxatilis
Lepidium africanum
Boechera puberula
Boechera stricta
Lepidium aucheri
Noccaea caerulescens
Physaria ovalifolia
Macropodium nivale
Nasturtium officinale
Biscutella laevigata
Isatis tinctoria
Leavenworthia alabamica
Heliophila coronopifolia
Iberis amara
Raphanus raphanistrum
Erucastrum elatum
Iberis pinnata
Isatis lusitanica
Biscutella frutescens
Brassica rapa
Anastatica hierochuntica
Biscutella auriculata
Cochlearia officinalis
Erysimum pusillum
Physaria acutifolia
Physaria fendleri
Sisymbrium irio
Raparia bulbosa
Noccaea goesingensis
Pseudoturritis turrita
Turritis glabra
Conringia planisiliqua (T)
Schrenkiella parvula
Sisymbrium altissimum
Ks

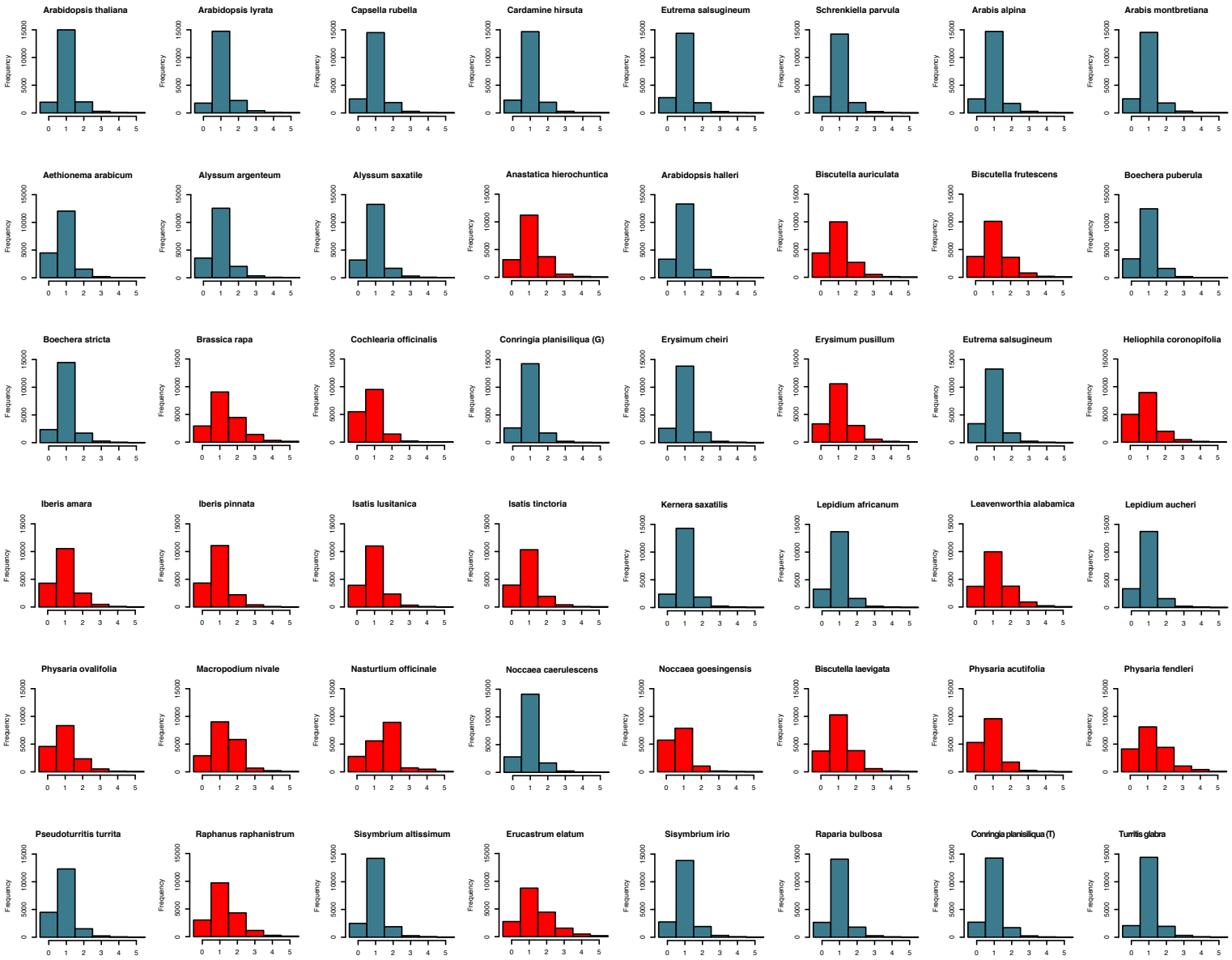
Supplementary Figure S5. Ks distributions for each species. The distributions were estimated from the *Ks* values generated from pairwise alignments of paralogs from each gene family. The green bar indicates the regions in which the peaks related to the At-α duplication are expected. Peaks before (i.e. to the left) relate to later whole-genome duplication events, or to the assembly of alternative haplotypes in heterozygous individuals. The peak after the At-α peak (i.e. to the right) reveals the duplicates which originated from the At-β duplication event.



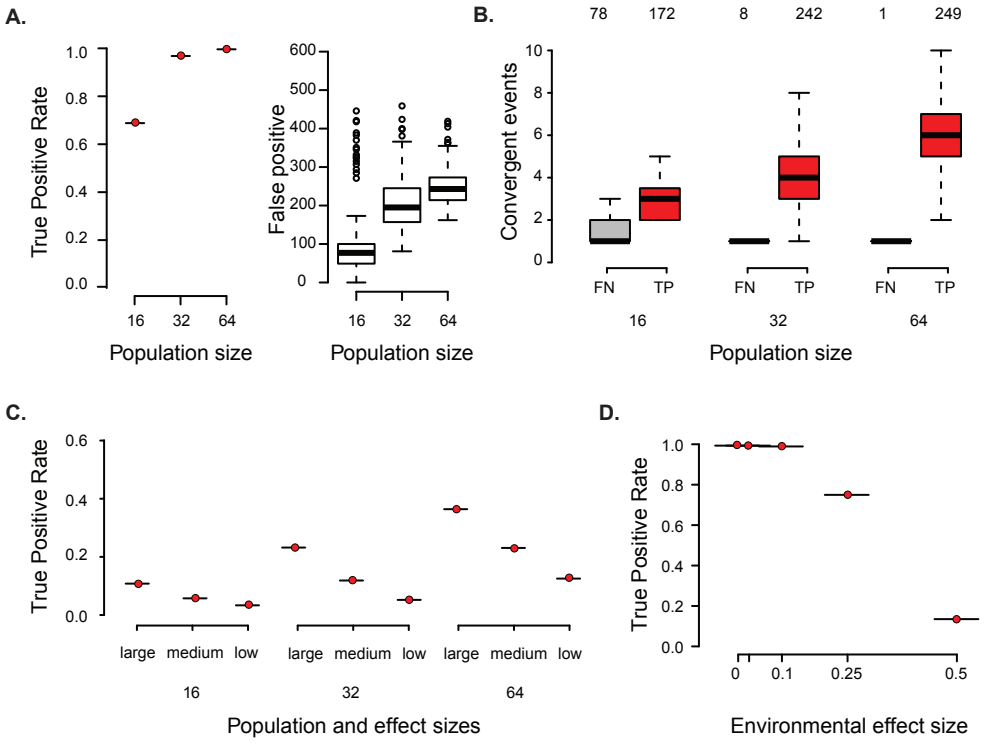
Supplementary Figure S6. Schematic of the support vector machine approach for the identification of meso/neopolyploids. We trained a support vector machine (svm) with 18 genomes for which the ploidy level was known. The features used included *Ks* distributions of the paranomes, gene family sizes, gene duplication and single-copy gene rates. The trained svm was then used to predict the ploidy level in the remaining species.



Supplementary Figure S7. Gene tree analysis for shared whole-genome duplication identification. Gene trees calculated from multiple sequence alignments were screened for two types of topologies after the removal of species-specific duplications. The first topology (upper row) separated the gene copies of a species into one clade. In contrast, if a whole-genome duplication was shared between the species, individual gene copies between the species tend to be more similar as compared to their paralogs (as the duplication predates speciation). This leads to gene trees with a different topology (lower row).

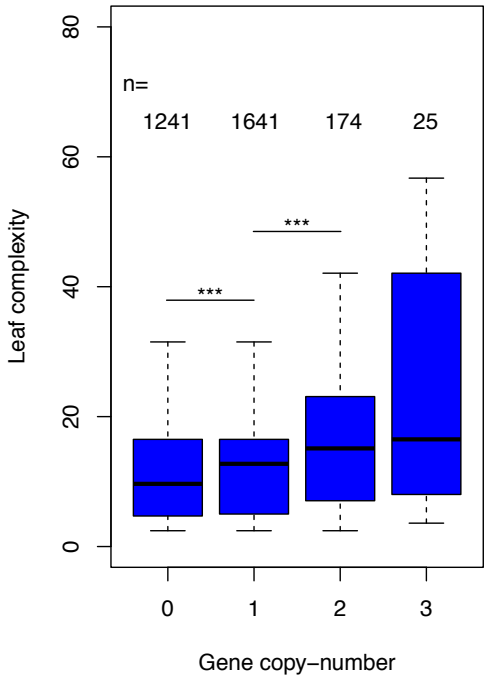

Arabidopsis thaliana
Arabidopsis lyrata
Capsella rubella
Cardamine hirsuta
Eutrema salsugineum
Schrenkiella parvula
Arabis alpina
Arabis montbretiana
Aethionema arabicum
Alyssum argenteum
Alyssum saxatile
Anastatica hierochuntica
Arabidopsis halleri
Biscutella auriculata
Biscutella frutescens
Boechera puberula
Boechera stricta
Brassica rapa
Cochlearia officinalis
Conringia planisiliqua (G)
Erysimum cheiri
Erysimum pusillum
Eutrema salsugineum
Heliophila coronopifolia
Iberis amara
Iberis pinnata
Isatis lusitanica
Isatis tinctoria
Kernera saxatilis
Lepidium africanum
Leavenworthia alabamica
Lepidium aucheri
Physaria ovalifolia
Macropodium nivale
Nasturtium officinale
Noccaea caerulescens
Noccaea goesingensis
Biscutella laevigata
Physaria acutifolia
Physaria fendleri
Pseudoturritis turrita
Raphanus raphanistrum
Sisymbrium altissimum
Erucastrum elatum
Sisymbrium irio
Raparia bulbosa
Conringia planisiliqua (T)
Turritis glabra
Frequency
0
5000
10000
15000
0 1 2 3 4 5

Supplementary Figure S8. Gene family size distributions. For each species the amount of gene families according to their different sizes are shown. This includes those gene families, which have been identified in the complete set of species, but are not present in the individual species (first bar of each histogram). Red histograms indicate meso/neo-polyploid species, blue histograms indicate diploid species.

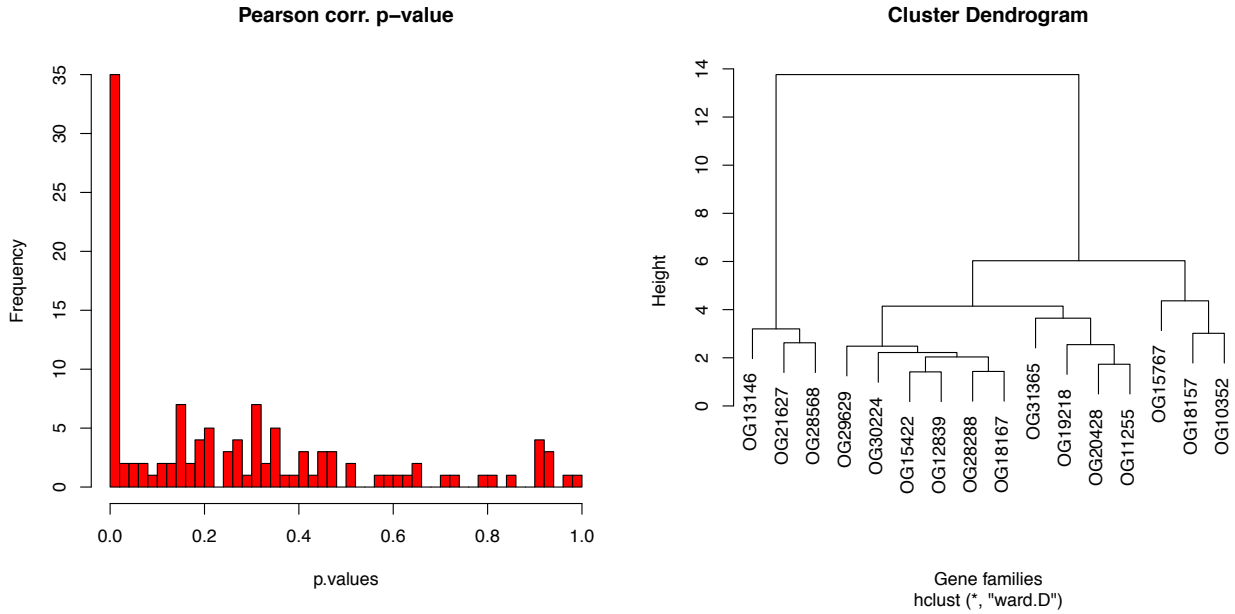


Supplementary Figure S9. Phylogenetic association mapping and power assessment using simulations using gene-copy number. (A-D) Power evaluation of the inter-species association mapping using gene families that are simulated after an evolutionary framework with 16, 32 or 64 species. Gene duplication or gene loss events were randomly introduced at each branch of the phylogenies. The genotype used for the association was the actual gene-copy number in each gene family (A) Mapping of major effect genes: the phenotype was defined by the effect of a single gene family. Average true positive rate (left) and boxplots of absolute false positive number (right) estimated in 250 individual simulations. (B) The number of convergent gene loss events separately assessed for the gene families that resulted in true positives or false negatives in the simulations shown in (A). The sample size of each boxplot is shown on top. (C) Average true positive rates for phenotypes defined by the combined effects of ten gene families with variable effects (i.e. large, medium and small) (n=250 independent simulation per value). (D) Average recall rates for the mapping of phenotypes with simple genetic architecture (like in A), which how were attributed to an increasing environmental (random) influence. Each value was estimated with 250 individual

simulations. Boxplots in this figure: black bars describe the median, boxes refer to quartile groups 2 and 3, whiskers show quartile group 1 and 4 and dots refer to outliers (if shown).

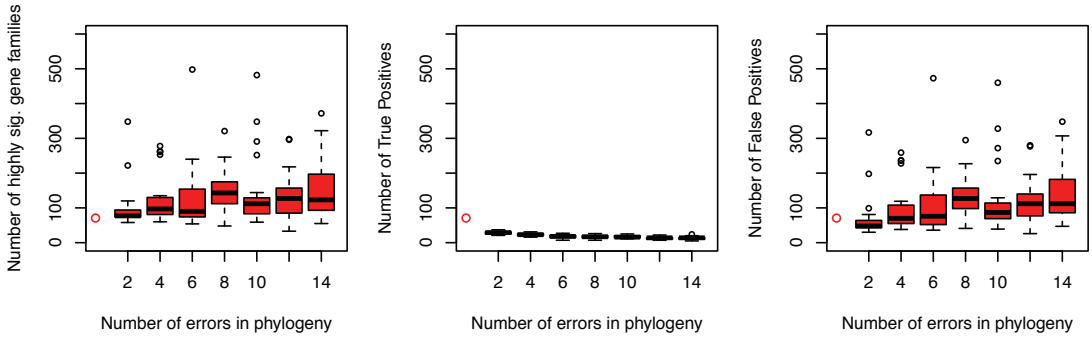


Supplementary Figure S10. Leaf complexity values separately shown for species with different gene-copy number found in the RCO gene family. Leaf complexity was not only defined by the presence of RCO (i.e. species without RCO showed the simplest leaves), but gene-copy number had a significant impact on leaf complexity (i.e. species with two or more RCO gene copies showed more complex leaves than species with only one RCO copy). Note, gene-copies were separately assessed from the ancestral paralog *LMI1*, which was assessed in a different gene family. Boxplots in this figure: black bars describe the median, boxes refer to quartile groups 2 and 3, whiskers show quartile group 1 and 4 and dots refer to outliers (if shown).

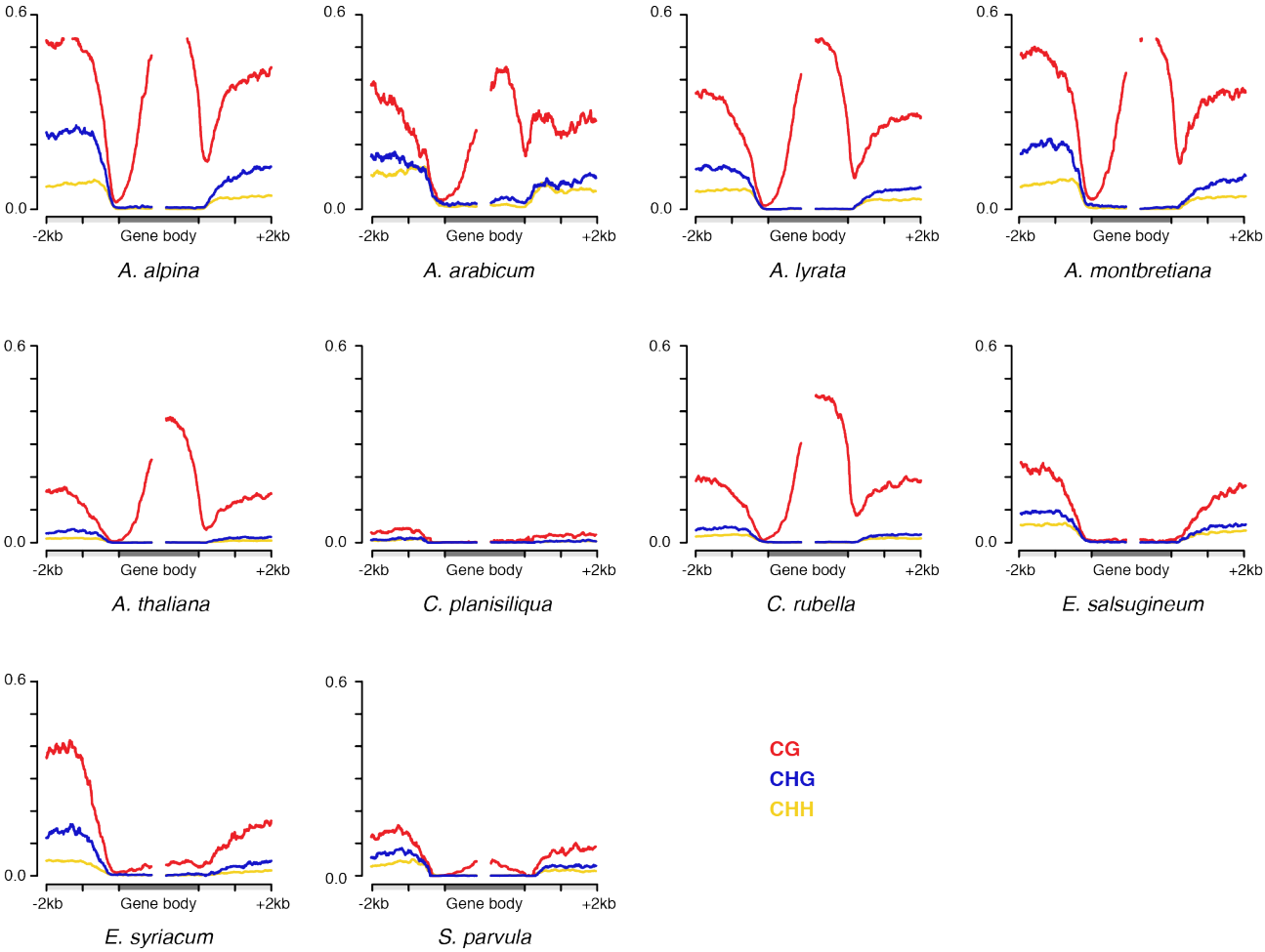


Supplementary Figure S11. Comparison of gene families highly significantly associated to leaf complexity. (A) Histogram of p-values of n=120 correlation tests of all pairwise comparisons of 16 gene families with the most significant associations to leaf shape complexity. The extensive correlation between the gene families indicates high levels of similarity between them. **(B)** Clustering of the 16 gene families based on Euclidean distances between their absence/presence patterns revealed two main clusters. The gene family of *RCO* is among the larger cluster (OG10352). Independent of their putative impact on leaf shape complexity, the correlated gene families are likely to correlate to leaf shape complexity due to their similarity to the RCO gene family.

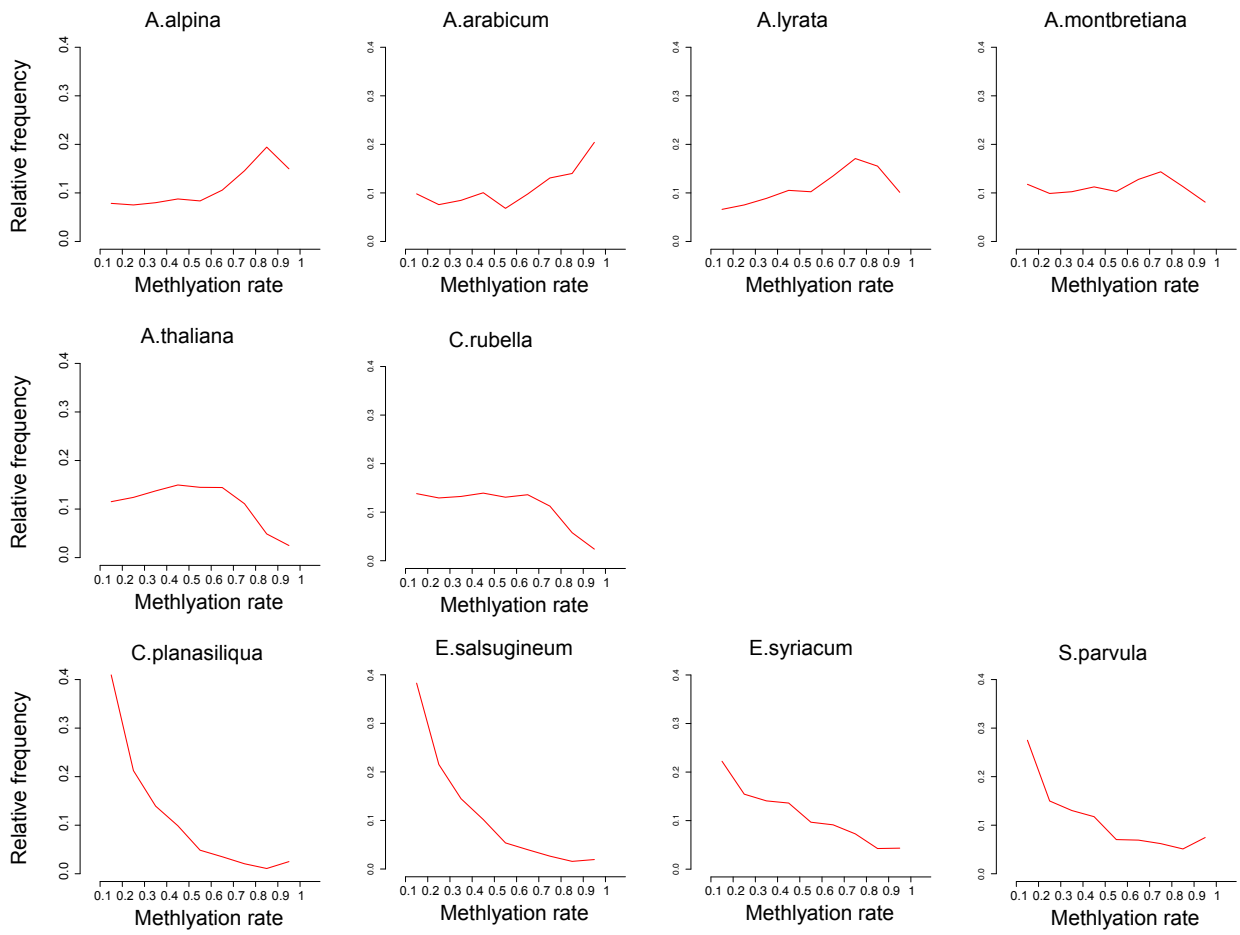
Supplementary Figure S12. Number of highly significantly associated gene families in the presence of errors in the gene families. We introduced an increasing number of errors in the gene families by increasing the number of gene families with false (randomly increased or decreased) gene copy numbers (each with n=25 independent iterations). The red dots indicate the number highly significantly associated gene families to leaf complexity in the original data. The boxplots show the distribution of gene family numbers for each of the n=25 iterations. The increasing number of errors constantly increased the number of significantly associated gene families until it reached a plateau of around 1500 gene families when introducing 12 or more errors in each gene family. At this error level, there were virtually no true positives (middle) present in the highly significant of significant results sets anymore, while the false positives (right) represented the entire result set. Boxplots in this figure: black bars describe the median, boxes refer to quartile groups 2 and 3, whiskers show quartile group 1 and 4 and dots refer to outliers (if shown).



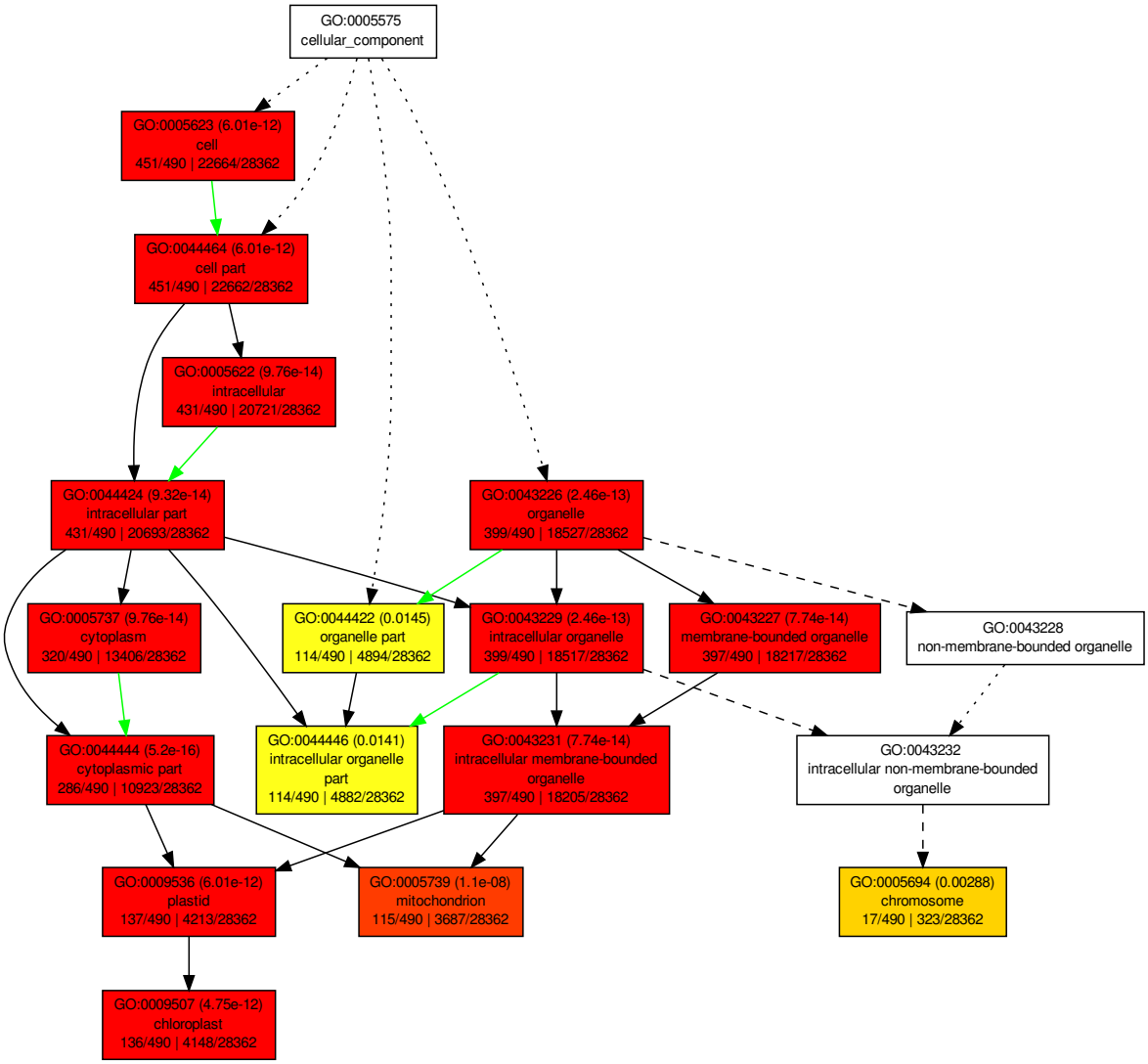
Supplementary Figure S13. Number of highly significantly associated gene families in the presence of errors in the species tree. We introduced an increasing number of errors in the species tree by swapping the phylogenetic location of a random pair of species (each with n=25 independent iterations). The red dots indicate the number highly significantly associated gene families to leaf complexity in the original data. The boxplots show the distribution of gene family numbers for each of the n=25 iterations. The increasing number of errors constantly decreased the number of significantly associated gene families (true positives) until there were no gene families left which were highly significant associated given the original tree at around 8 random swaps. In contrast, the number of false associations did not drastically increase even when introducing a huge number of errors in the phylogeny. Boxplots in this figure: black bars describe the median, boxes refer to quartile groups 2 and 3, whiskers show quartile group 1 and 4 and dots refer to outliers (if shown).



Supplementary Figure S14. Metagene plots summarizing gene methylation in ten Brassicaceae for species. Colored lines show the average methylation rates across all genes in each species for each of the three DNA methylation contexts. For example, the metagene plot of *A. thaliana* shows a clear enrichment for CG methylation within the gene bodies, while methylation of C in other context is mostly absent. This signal is almost absent in species without *CMT3*.



Supplementary Figure 15. Level of CHG methylation. The histograms show the fraction of sites with different degrees of methylation (across all methylated sites in the respective species). The four species in the bottom row have either lost CMT3 or contain a CMT3 allele with relaxed selection patterns. All four species show an increased fraction of nucleotides which are methylated with low levels only.



Supplementary Figure S16. GO analysis ("Cellular component"). Results are shown for the analysis of *A. thaliana* members of the gene families, which are absent in at least 20 of the 21 polyploid species.

Supplementary Tables

Supplementary tables can be found in associated excel sheets.

Supplementary Table S1. Selected species, generated sequencing data and assembly quality

Supplementary Table S2. Meso-polyploidization prediction (lower part) trained on genomes with known meso-polyploidizations

Supplementary Table S3. MAPS analysis of species with putatively shared WGD events.

Supplementary Table S4. Leaf shape complexity values

Supplementary Table S5. Orthologous groups(OG) with highly significant associations to leaf shape complexity

References

1. Koch, M. A., Karl, R., German, D. A. & Al-Shehbaz, I. A. Systematics, taxonomy and biogeography of three new Asian genera of Brassicaceae tribe Arabideae: An ancient distribution circle around the Asian high mountains. Taxon 61, 955–969 (2012).
2. Al-Shehbaz, I. A., German, D. A., Karl, R., Jordon-Thaden, I. & Koch, M. A. Nomenclatural adjustments in the tribe Arabideae (Brassicaceae). Plant Divers. Evol. 129, 71–76 (2011).
3. Jiao, W.-B. et al. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. Genome Res. gr.213652.116 (2017). doi:10.1101/gr.213652.116
4. The Brassica rapa Genome Sequencing Project Consortium et al. The genome of the mesopolyploid crop species Brassica rapa. Nat. Genet. 43, 1035–1039 (2011).

5. Hu, T. T. et al. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat. Genet. 43, 476–481 (2011).
6. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815 (2000).
7. Gan, X. et al. The Cardamine hirsuta genome offers insight into the evolution of morphological diversity. Nat. Plants 2, 16167 (2016).
8. Slotte, T. et al. The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat. Genet. 45, 831–835 (2013).
9. Yang, R. et al. The Reference Genome of the Halophytic Plant Eutrema salsugineum. Front. Plant Sci. 4, 46 (2013).
10. Dassanayake, M. et al. The genome of the extremophile crucifer Thellungiella parvula. Nat. Genet. 43, 913–918 (2011).
11. Lee, C.-R. et al. Young inversion with multiple linked QTLs under selection in a hybrid zone. Nat. Ecol. Evol. 1, 0119 (2017).
12. Haudry, A. et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat. Genet. 45, 891–898 (2013).
13. Briskine, R. V. et al. Genome assembly and annotation of Arabidopsis halleri, a model for heavy metal hyperaccumulation and evolutionary ecology. Mol. Ecol. Resour. 17, 1025–1036 (2017).
14. Moghe, G. D. et al. Consequences of Whole-Genome Triplication as Revealed by Comparative Genomic Analyses of the Wild Radish Raphanus raphanistrum and Three Other Brassicaceae Species. Plant Cell Online 26, 1925–1937 (2014).
15. Li, R. et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25, 1966–1967 (2009).
16. Bankevich, A. et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. 19, 455–477 (2012).

17. Pryszcz, L. P. & Gabaldón, T. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res. 44, e113 (2016).
18. Chin, C.-S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods 13, 1050–1054 (2016).
19. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
20. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
21. Keller, O., Odronitz, F., Stanke, M., Kollmar, M. & Waack, S. Scipio: Using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. BMC Bioinformatics 9, 278 (2008).
22. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 33, W465–W467 (2005).
23. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20, 2878–2879 (2004).
24. Korf, I. Gene finding in novel genomes. BMC Bioinformatics 5, 59 (2004).
25. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9, R7 (2008).
26. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212 (2015).
27. Chen, J., Hu, Q., Zhang, Y., Lu, C. & Kuang, H. P-MITE: a database for plant miniature inverted-repeat transposable elements. Nucleic Acids Res. 42, D1176–D1181 (2014).
28. Macas, J., Mészáros, T. & Nouzová, M. PlantSat: a specialized database for plant

satellite repeats. Bioinforma. Oxf. Engl. 18, 28–35 (2002).

29. Ouyang, S. & Buell, C. R. The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res. 32, D360–363 (2004).
30. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6, (2015).
31. Nussbaumer, T. et al. MIPS PlantsDB: a database framework for comparative plant genome research. Nucleic Acids Res. 41, D1144–1151 (2013).
32. Goubert, C. et al. De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). Genome Biol. Evol. 7, 1192–1205 (2015).
33. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16, (2015).
34. Zhang, N., Zeng, L., Shan, H. & Ma, H. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. New Phytol. 195, 923–937 (2012).
35. Duarte, J. M. et al. Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryzaand their phylogenetic utility across various taxonomic levels. BMC Evol. Biol. 10, 61 (2010).
36. Guindon, S. et al. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst. Biol. 59, 307–321 (2010).
37. Lanfear, R., Calcott, B., Ho, S. Y. W. & Guindon, S. PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses. Mol. Biol. Evol. 29, 1695–1701 (2012).

38. Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T. & Calcott, B. PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. Mol. Biol. Evol. 34, 772–773 (2017).
39. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. bioRxiv 447110 (2019). doi:10.1101/447110
40. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19, 153 (2018).
41. Bouckaert, R. et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLOS Comput. Biol. 15, e1006650 (2019).
42. Hohmann, N., Wolf, E. M., Lysak, M. A. & Koch, M. A. A Time-Calibrated Road Map of Brassicaceae Species Radiation and Evolutionary History. Plant Cell 27, 2770–2784 (2015).
43. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst. Biol. 67, 901–904 (2018).
44. Vision, T. J., Brown, D. G. & Tanksley, S. D. The Origins of Genomic Duplications in Arabidopsis. Science 290, 2114–2117 (2000).
45. Bowers, J. E., Chapman, B. A., Rong, J. & Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422, 433–438 (2003).
46. Schranz, M. E. & Mitchell-Olds, T. Independent Ancient Polyploidy Events in the Sister Families Brassicaceae and Cleomaceae. Plant Cell 18, 1152–1165 (2006).
47. Barker, M. S., Vogel, H. & Schranz, M. E. Paleopolyploidy in the Brassicales: Analyses of the Cleome Transcriptome Elucidate the History of Genome Duplications in

Arabidopsis and Other Brassicales. Genome Biol. Evol. 1, 391–399 (2009).

48. Vanneste, K., Van de Peer, Y. & Maere, S. Inference of Genome Duplications from Age Distributions Revisited. Mol. Biol. Evol. 30, 177–190 (2013).
49. Vanneste, K., Baele, G., Maere, S. & Van de Peer, Y. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. Genome Res. 24, 1334–1347 (2014).
50. Dongen, S. van & Abreu-Goodger, C. Using MCL to Extract Clusters from Networks. in Bacterial Molecular Networks 281–295 (Springer, New York, NY, 2012).
51. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30, 1575–1584 (2002).
52. Ossowski, S. et al. The Rate and Molecular Spectrum of Spontaneous Mutations in Arabidopsis thaliana. Science 327, 92–94 (2010).
53. Jordon-Thaden, I. & Koch, M. Species richness and polyploid patterns in the genus Draba (Brassicaceae): a first global perspective. Plant Ecol. Divers. 1, 255–263 (2008).
54. Mandáková, T., Li, Z., Barker, M. S. & Lysak, M. A. Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. Plant J. 91, 3–21 (2017).
55. Kagale, S. et al. Polyploid Evolution of the Brassicaceae during the Cenozoic Era. Plant Cell Online 26, 2777–2791 (2014).
56. Li, Z. et al. Early genome duplications in conifers and other seed plants. Sci. Adv. 1, e1501084 (2015).
57. Smet, R. D. et al. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc. Natl. Acad. Sci. 110, 2898–2903 (2013).
58. Tian, T. et al. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res. 45, W122–W129 (2017).
59. Felsenstein, J. Maximum-likelihood estimation of evolutionary trees from continuous

characters. Am. J. Hum. Genet. 25, 471–492 (1973).

60. Orme, D. The caper package: comparative analysis of phylogenetics and evolution in R. 36

61. Seymour, D. K., Koenig, D., Hagmann, J., Becker, C. & Weigel, D. Evolution of DNA Methylation Patterns in the Brassicaceae is Driven by Differences in Genome Organization. PLOS Genet. 10, e1004785 (2014).

62. Bewick, A. J. et al. On the origin and evolutionary consequences of gene body DNA methylation. Proc. Natl. Acad. Sci. 113, 9111–9116 (2016).

63. Guo, W. et al. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics 14, 774 (2013).

64. Willing, E.-M. et al. Genome expansion of Arabis alpina linked with retrotransposition and reduced symmetric DNA methylation. Nat. Plants 1, 14023 (2015).

65. Nikolov, L. A. et al. Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. New Phytol. 222, 1638–1651 (2019).