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

nature plants

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]*

Comparative genomics can unravel the genetic basis of species differences; however, successful reports on quantitative traits are still scarce. Here we present genome assemblies of 31 so-far unassembled Brassicaceae plant species and combine them with 16 previously published assemblies to establish the Brassicaceae Diversity Panel. Using a new interspecies association strategy for quantitative traits, we found a so-far unknown association between the unexpectedly high variation in CG to TG substitution rates in genes and the absence of *CHROMOMETHYLASE3 (CMT3)* orthologues. Low substitution rates were associated with the loss of *CMT3*, while species with conserved *CMT3* orthologues showed high substitution rates. Species without *CMT3* also lacked gene-body methylation (gbM), suggesting an evolutionary trade-off between the unknown function of gbM and low substitution rates in Brassicaceae, possibly due to low mutability of non-methylated cytosines.

Soon after the first bacterial genomes were assembled, comparative efforts to associate genomic differences with differences in the phenotypes were explored[1–3]. New high-throughput DNA sequencing technologies now allow us to decipher virtually every genome we are interested in, and it is tempting to also apply phenotype-to-genotype associations to the huge variation that can be found between eukaryotic species. However, common marker-based association mapping (for example, genome-wide association (GWA) methods) cannot be applied because such methods rely on shared variation and broad co-linearity of chromosomal sequences, which usually do not exst in the rearranged genomes of distinct species.

First attempts at interspecies association mapping were based on the idea of gene co-elimination, suggesting that all genes specific to a particular trait will be lost once the trait itself is lost[4]. Using the loss of a particular trait as phenotype, correlated patterns of gene loss were identified across distinct genomes. Successful studies include the search for the genetic basis of loss of vitamin C synthesis[5] and loss of vision[6] across different mammalian species, as well as the screens for genes recapitulating the loss of arbuscular mycorrhizal symbiosis[7] and nitrogen-fixing root nodule symbiosis[8] in plant genomes. Other studies exchanged the concept of gene loss by other genomic footprints of selection, as in genomic comparisons of different bird species that revealed genes with increased evolutionary rates specific to singing[9] or flightless birds[10]. Likewise, closely related tomato species were scanned for segregating variation, which was correlated to different environmental conditions (using an approach called phyloGWAS) assuming that ancestrally segregating variants confer advantages to variable conditions[11].

Most of these studies, however, were based on absence/presence phenotypes, even though most traits are actually more complex. In this study, we explored the potential of phylogenetic association mapping (PAM) of quantitative traits. For this we built up a set of 47 Brassicaceae species, which were selected throughout the entire phylogeny of this plant family to maximize sequence divergence between the species. For 31 of the species no reference sequences were available, and we therefore generated de novo whole-genome assemblies. We first validated the potential of phylogenetic association by mapping genes underlying differences in leaf shape complexity and identified a known key regulatory gene controlling this quantitative trait[12]. We then used the unexpectedly high variation in CG to TG dinucleotide substitution rates between the species and identified a DNA methyltransferase, *CHROMOMETHYLASE3* (*CMT3*), which was highly significantly linked to the variation in CG to TG substitution rates. The analysis of the DNA methylome of ten different species further supported that loss of *CMT3* is involved in the loss of gene-body methylation (gbM) as previously shown[13]. As methylated cytosines are known to be more mutable compared to non-methylated cytosines[14–16], this suggested *CMT3* was the prime candidate gene for the modification of CG to TG substitution rates and thereby could have the potential to introduce an evolutionary trade-off between the unknown function of gbM and lower substitution rates.

Results

Genome assemblies and their evolutionary relationships. We performed whole-genome shotgun sequencing to generate 31 genome assemblies of so-far unassembled Brassicaceae species. The species were selected in combination with 16 previously assembled genomes[17–31] to cover as much as possible from the phylogenetic width and genomic diversity of this family[32–35], in addition to one species represented by two individuals (Fig. 1). The newly assembled species include the edible watercress *Nasturtium officinale*, the traditional Chinese medical herb *Isatis tinctoria*, the resurrection plant 'Rose of Jericho' (*Anastatica hierochuntica*) and the ornamental plants 'wallflower' (*Erysimum cheiri*) and 'basket of gold' (*Alyssum saxatile*). Previous assemblies include Chinese cabbage

[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

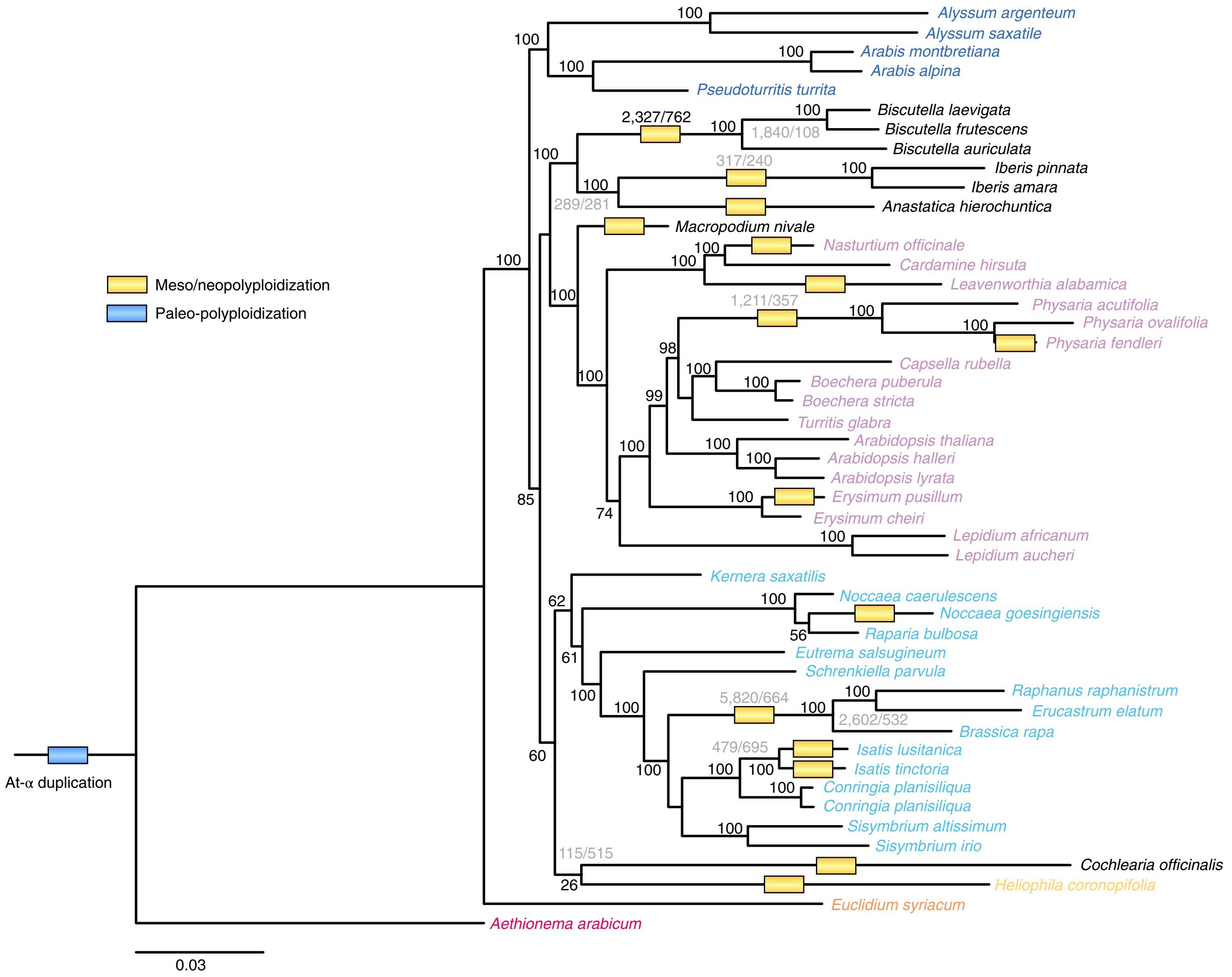


Fig. 1 | Phylogenetic tree of 47 Brassicaceae species including paleo- and meso/neopolyploidization events. Blue and yellow boxes indicate branches in the phylogenetic tree where genome duplication events occurred (blue, paleopolyploidization; yellow, meso/neopolyploidization). Numbers in grey indicate the number of gene trees that supported/did not support shared polyploidization events during the multi-taxon paleopolyploidy search (MAPS) analysis at the respective branches[43]. Black numbers indicate bootstrap values; the scale (0.03) describes the number of substitutions per nucleotide site. The phylogeny is largely congruent with the one presented by Nikolov et al.[40] (species names were coloured according to the respective grouping).

(*Brassica rapa*)[19], the plant model *Arabidopsis thaliana*[28] and models for studying perennial flowering[17,22] (*Arabis alpina*) and metal hypertolerance and accumulation[18] (*Arabidopsis halleri*).

With the exception of two species, which were assembled from PacBio long reads, all new assemblies were based on Illumina short reads. Average N50 contig sizes after the removal of alternative haplotypes were ~3.2 Mb for the high-quality assemblies and ~20 kb for the Illumina-only assemblies (Supplementary Fig. 1 and Supplementary Table 1). Genes were annotated using a homology-based method. Despite the lower contig contiguity, the gene sets annotated in the Illumina assemblies were as complete as those in the chromosome-level assemblies (Supplementary Fig. 1).

Orthologous groups (OG) were calculated with OrthoFinder[36]. Based on the OGs calculated in a set of eight core species, we defined 18,695 OGs that remained stable after adding all remaining species. To split OGs with gene duplications, which predated the divergence of the Brassicaceae or happened at the onset of the evolution of this family, we calculated gene trees for each OG and separated those subtrees that broadly recapitulated the species phylogeny into distinct OGs. This increased the number of OGs to 22,260 (now referred to as gene families).

A species phylogeny broadly recapitulated the major lineages of the Brassicaceae family (Fig. 1 and Supplementary Figs. 2–4)[32–35,37–40]. We estimated the split of the core Brassicaceae and its basal group represented by *Aethionema arabicum* to be 32.4 Ma (million years ago) and the split of *Lineage III* (represented by *Euclidium syriacum*) to be 27.3 Ma. The split of *A. thaliana* from *A. halleri* and *Arabidopsis lyrata* was dated to 7 Ma and the split of *N. officinale* and *Lepidium* happened 17.2 Ma. These estimates are in good agreement with previous reports[35].

To identify all meso/neopolyploids among the newly sequenced genomes we used a machine-learning-based approach. Based on 18 genome assemblies of species, which have been described as either diploid or meso- or neopolyploid[41,42], we trained a support vector machine with features that included the *Ks* (the number of synonymous substitutions per synonymous site) distribution calculated from the paranome of each species, average gene family size, gene duplication rate and the amount of single-copy genes and we

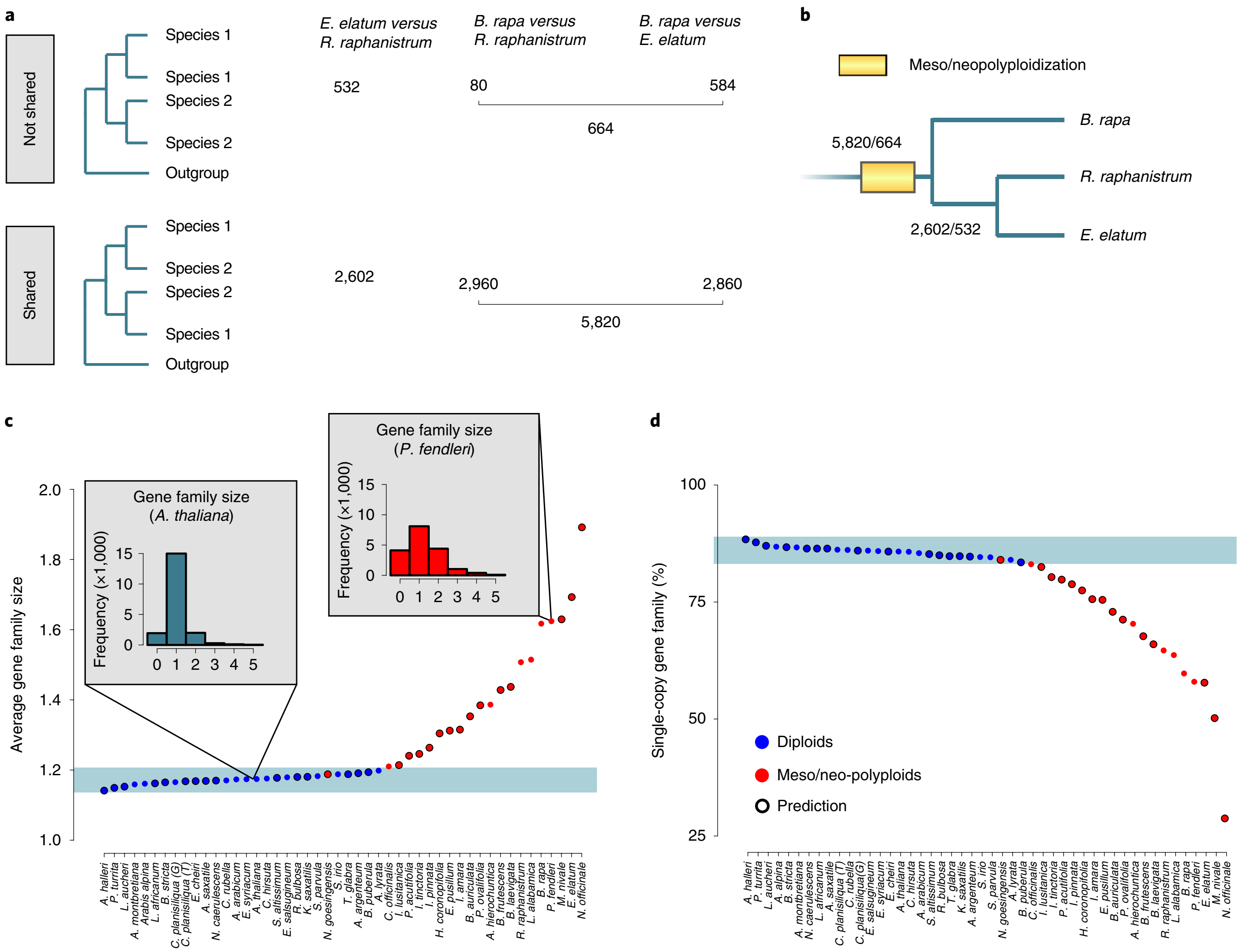


Fig. 2 | Polyploidization event identification, their phylogenetic placement and the impact of these events on the gene space. a, Example for the phylogenetic placement of polyploidization events. All three species, *Brassica rapa, Raphanus raphanistrum* and *Erucastrum elatum*, were annotated as polyploids. To infer whether these polyploidization events are shared between the species, gene trees were calculated and distinguished based on their topology either suggesting a 'Shared' or 'Not shared' polyploidization event (which is similar to MAPS[43]). Here, *E. elatum* and *R. raphanistrum* shared an event as many more gene trees (2,602) supported this scenario, as compared to the scenario of independent events (532). This event was also shared with *B. rapa* as the comparisons with the other two species revealed. **b**, The annotation of the meso-polyploidization event in the phylogenetic tree (as shown in Fig. 1 for the complete tree). **c**,**d**, Average gene family size (**c**) and percentage of single-copy gene families (**d**) across 22,260 gene families of 48 samples. Blue and red dots refer to diploids and meso/neopolyploids, respectively. Species without black circles were classified with the support vector machine.

predicted ploidy levels of the remaining genomes (Supplementary Figs. 5 and 6, Supplementary Table 2). This revealed 16 meso/neopolyploid species and 14 species without additional genome duplication. In total there were 21 meso/neopolyploid and 27 diploid samples in our set. Applying a modified version of MAPS[43], the 21 meso/neopolyploids could be traced back to 15 polyploidization events, including the well-described hexaploidization, whcih is common in the Brassica clade (Fig. 2a, Supplementary Fig. 7, Supplementary Table 3).

The 27 diploids featured ~40,000 (±~12,000) genes, while the 21 meso/neopolyploid genome assemblies included ~60,000 (±~19,000) genes on average (Supplementary Fig. 1). This difference was also reflected by the average gene family size, which was very uniform among the diploid species (from 1.14 to 1.20), with at least 84% of the gene families being single-copy genes (Fig. 2b and Supplementary Fig. 8). In contrast, the meso/neopolyploids featured average gene family sizes from 1.19 to 1.89.

Proof-of-concept of PAM. We developed a method for PAM to correlate the variation in gene copy number to differences in quantitative traits (Fig. 3a). For this we counted each individual member of a gene family independent of the ploidy of the genome in which the actual genes were found. PAM is based on a phylogenetic generalized least-squares approach[6,44–46], which is a modification of the generalized least-squares method and allows to correct for the covariance that is expected to occur solely due to the phylogenetic relationship between the species[47]. This covariance can lead to false associations if not corrected for, but it can be estimated by the shared branch length measured from the common ancestor to the respective species in the phylogenetic tree[44,45,47,48]. Once the regression is calculated, the significance of the association is estimated by testing whether the slope of the regression is significantly different from zero or not. To understand its performance, we performed a power evaluation of the interspecies association mapping based on the presence/absence gene patterns of 20,000 gene families that

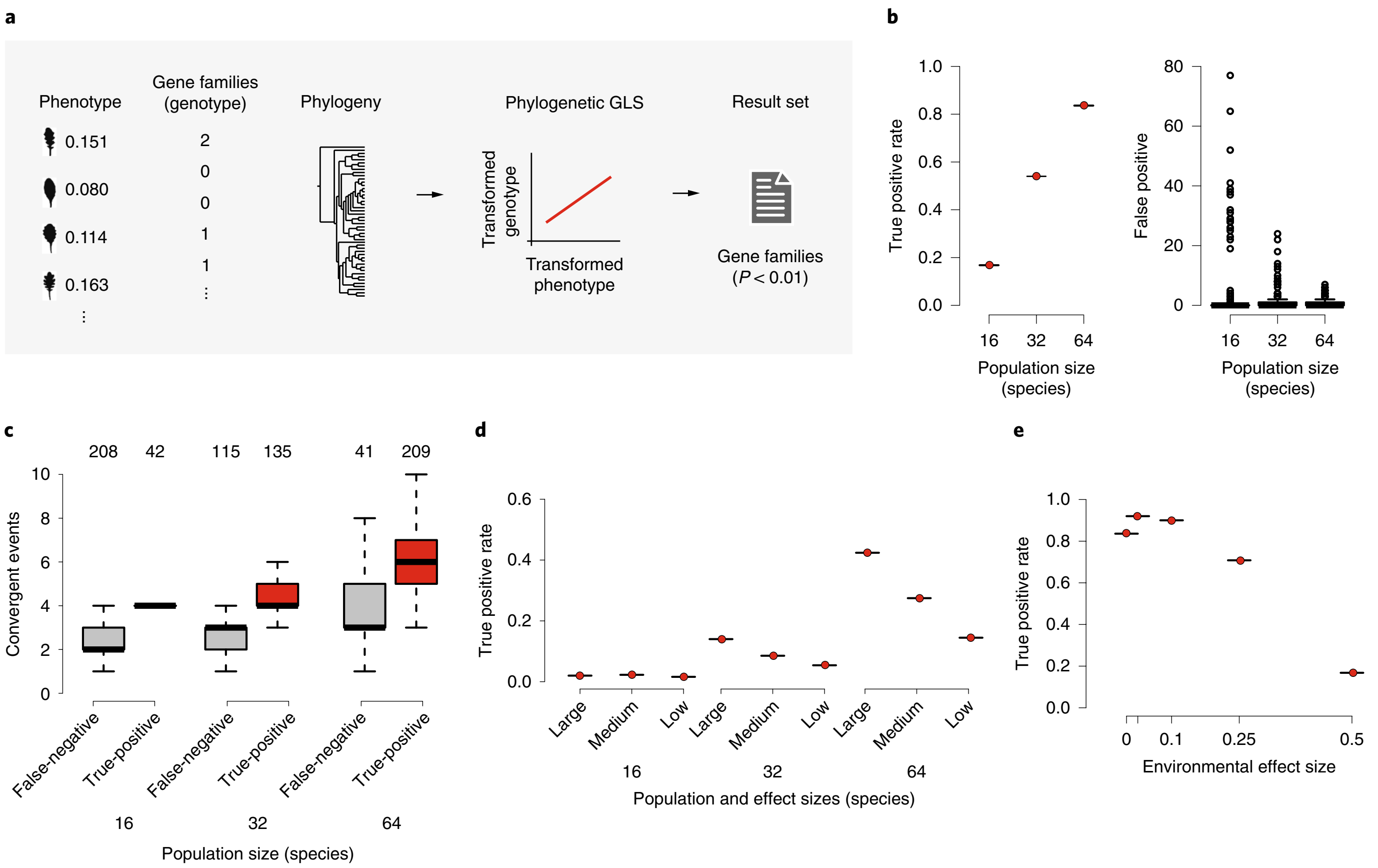


Fig. 3 | PAM and power assessment using simulations based on gene absence/presence patterns. **a**, PAM requires information about trait differences, genetic variation, and the phylogenetic relatedness of the species. Genotypes and phenotypes were correlated using a phylogenetically informed approach, which reveals highly significantly associated gene families. **b–e**, Power evaluation of the interspecies association mapping using gene families that were 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. **b**, Mapping of major effect genes: the phenotype was defined by the effect of a single gene family. Average true positive rate (left) and box plots of absolute false positive number (right) estimated in 250 individual simulations. **c**, The number of convergent gene loss events assessed separately for the gene families that resulted in true positives or false negatives in the simulations shown in **a**. The sample size of each box plot is shown along the top. **d**, Average true positive rates for phenotypes defined by the combined effects of ten gene families with variable effects (that is, large, medium and low) ($n = 250$ independent simulations per value). **e**, Average recall rates for the mapping of phenotypes with simple genetic architecture (as in **a**), which were additionally attributed to an increasing environmental (random) effect. Each value was estimated with 250 individual simulations. Box plots: 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).

were simulated after an evolutionary framework based on 16, 32 or 64 species (Fig. 3b–e). Gene duplication or gene-loss events were randomly introduced at each branch of the phylogenies. Major effect genes could be found at high recall rates (>80%) in large populations, but were rarely found in smaller populations. False positives were very generally very low (0–3 false positives), with the exception that some of the individual simulations showed clearly enriched numbers of false positives. Such outliers were more abundant and more extreme in smaller populations. Similar to population size, the number of convergent evolutionary events, which underlay the gene families with effect on the phenotype, was an important factor for finding genes. Even in large populations, phenotypes that did not rely on multiple convergent events could not be identified. The complexity of phenotypes (defined by different genes with variable effects) affected the recall rate. In large populations, major effect genes of complex phenotypes could be identified in around half the cases compared to phenotypes that relied solely on a single major effect gene. Environmental influence on the variation in the phenotype (that is, randomness) had surprisingly little effect on the recall rate, and only when as much as 50% of the phenotypic values resulted from environment, were the recall rates severely affected. Similar simulations based on gene copy numbers resulted in comparable conclusions, with the major differences that high recall rates could be achieved with small populations, which were influenced by drastically increased false positive rates (Supplementary Fig. 9).

As a proof-of-concept, we analysed the genetic basis of differences in leaf shape complexity between the Brassicaceae species. This is a well-studied trait between species and depends on a key player of leaf shape complexity, *REDUCED COMPLEXITY* (*RCO*), which evolved around ~26–44 Ma by a local duplication of the *LMI1* gene[12]. Combined changes in the expression domain and protein sequence allowed *RCO* to contribute to leaf complexity by sculpting developing leaflets by repressing growth at their flanks[12,49]. Comparative work across different Brassicaceae species showed that species that never evolved or lost *RCO* developed simple leaves[12]. Augmenting the species phylogeny with the absence/presence patterns of *RCO* homologues suggested that *RCO* was independently lost at least 12 times during the evolution of the Brassicaceae.

We measured leaf shape complexity of 39 of the 48 samples of our panel. Complexity was quantified as the percentage of pixels that did not correspond to leaves within a convex hull drawn around

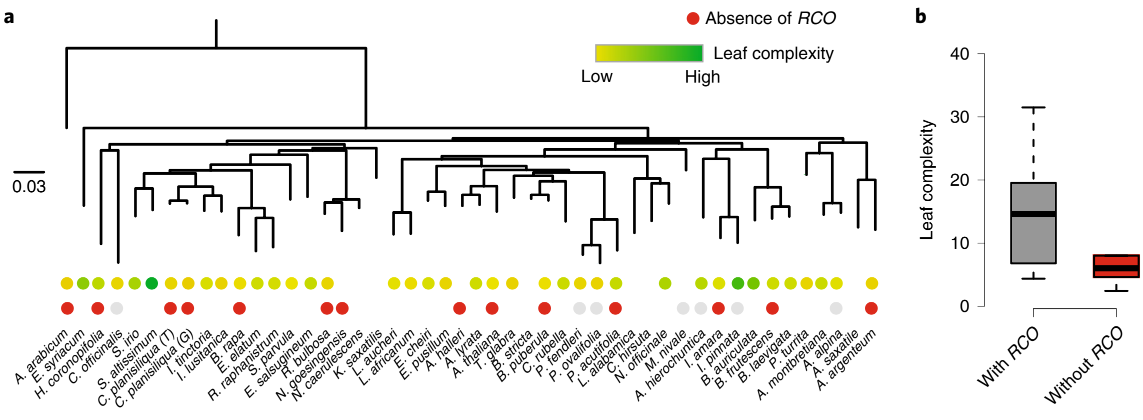


Fig. 4 | PAM of leaf shape complexity. a, Leaf shape complexity for each species is shown from yellow (low complexity) to green (high complexity) above each species names. Red dots indicate the absence of an orthologue of *REDUCED COMPLEXITY* (*RCO*), grey dots indicate species without clear signal of absence/presence of *RCO*. The scale (0.03) describes the number of substitutions per nucleotide site. **b**, Species without *RCO* are simple leaves ($n=12$). Species with *RCO* express a wide range of leaf complexities ($n=21$). Box plots: 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).

a digital image of each leaf at the time the plants started to flower (Fig. 4 and Supplementary Table 4).

We applied PAM to associate the gene copy number and absence/presence patterns of 22,260 gene families to leaf shape complexity. This resulted in 82 and 71 highly significantly associated gene families, respectively ($P<0.01$; Supplementary Table 5). The two result sets were not overlapping because the associations of the gene copy number association were driven by copy differences rather than by absence/presence patterns (Supplementary Fig. 10). *RCO* was found among the 71 gene families identified with absence/presence patterns and was ranked 16th ($P=0.0015$). This is in agreement with the observation that complex leaves rely on the presence of *RCO* and that additional copies of *RCO* do not increase complexity further[12].

Although the remaining genes in the gene absence/presence association did not include any other obvious candidate genes, we found that the majority of them were highly correlated to the absence/presence patterns of *RCO*. A pairwise comparison showed striking similarities between these gene families, which could explain the association of these gene families without an actual effect on leaf complexity (Supplementary Fig. 11). This type of linkage between gene families can be introduced not only by physical linkage, but also by shared whole-genome duplication events, which will lead to similar evolutionary trajectories (and thus similar association signals) of otherwise independent gene families.

To gain insights into the robustness of these associations, we randomly introduced errors in the gene families, and also in the species tree, and we repeated the association to leaf shape complexity. This showed that errors in gene families can lead to drastically increased numbers of false positives and hinder the identification of real gene families. This was observed when introducing errors in the species tree, while the errors in the species tree did not lead to drastically increased numbers of false positives (Supplementary Figs. 12 and 13).

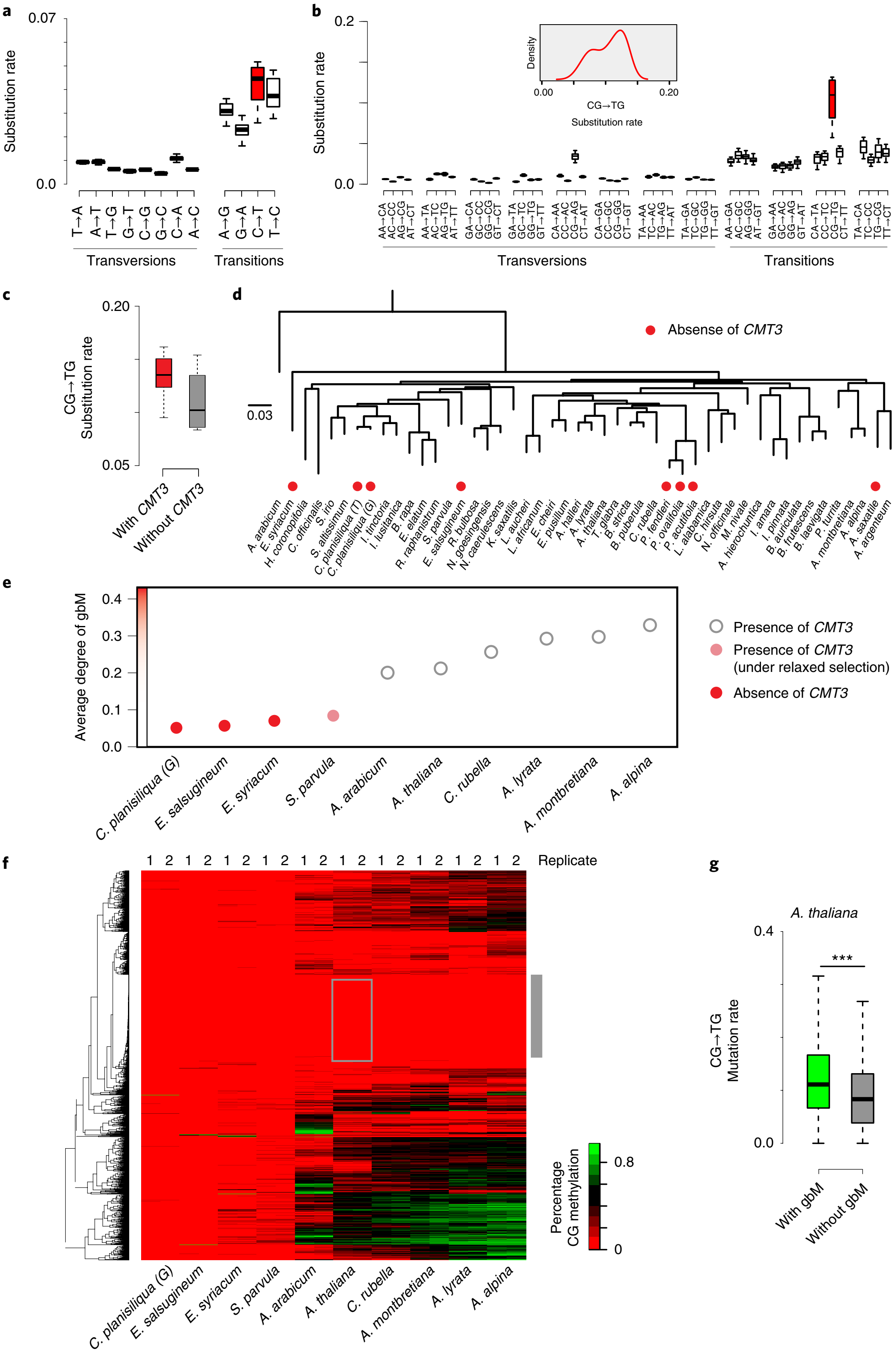
CG to TG substitution rates are linked to *CMT3*. Assessing quantitative phenotypes between species needs to be performed in the presence of different developmental processes and different responses to controlled conditions. This can complicate the assessment of comparable phenotypes if the heritability of a trait is low. In contrast, molecular phenotypes, such as mutation or substitution rates, can be estimated from genomic data alone and do not need to be tested in specific conditions, and are therefore particularly suitable for interspecies associations.

For example, differences in mutation rates can be introduced over time by mutations in genes that contribute to DNA integrity[50,51]. If these changes are small, evolutionary theory suggested that they would not be removed by natural selection, which, in turn, suggests that mutation rates can vary within and between species[50,52]. In fact, human populations carry significant differences in their mutational spectrum, which are even higher when compared to the mutational spectra of close relatives[50,51,53]. Moreover, a recent study in yeast showed that mutation rate varies among yeast strains and it illustrated that these differences can be used to map genes that act as mutational modifier genes[54].

To investigate whether such mutational differences also exist in our panel of plants, and to ultimately map underlying mutational modifiers, we analysed the substitution spectra of each species as an approximation of their mutational spectra. For this, we assessed the rate of derived single-nucleotide substitutions in

Fig. 5 | Single-nucleotide mutation rates, their association to variation of *CMT3* and differences in gbM. a, Derived single-nucleotide substitutions were used to estimate species-specific mutation rates. The box plots show the mutation rate variation of single-nucleotide mutation across $n=25$ diploid species. C to T mutation rates were highest and also showed the largest variation between the species (red box plot). **b**, Same data as in **a** but mutation types were separated according to their 3′ neighbouring nucleotide ($n=25$). CG to TG showed the highest and most variable mutation rates. **c**, CG to TG mutation rates in species with (red, $n=20$) and without (grey, $n=5$) an orthologue of *CMT3*. **d**, Phylogenetic tree showing species without *CMT3* orthologue. **e**, The average degree of gbM across ten Brassicaceae species calculated as the average number of methylated CG sites across all CG sites in ~1,700 orthologous genes (*y*-axis). The scale (0.03) describes the number of substitutions per nucleotide site. **f**, Cytosine methylation in each of the genes (and replicates) separately shown (red: low levels; green: high levels of methylation). Grey bar indicates genes that are not significantly methylated in any of the species. **g**, Box plot comparing *A. thaliana* genes without patterns of gbM ($n=1,228$; grey box in **f**) to genes with gbM ($n=393$). Mutation rates in genes with gbM are significantly higher compared to the genes without gbM (Kolmogorov–Smirnov, $^{***}P=1.541\times10^{-8}$). Box plots: 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).

pairwise alignments of orthologous genes of 25 diploid species (using *A. arabicum* and *E. syriacum* as outgroups). This allowed us to assign orthologous nucleotides, which was not possible in highly divergent intergenic regions. As expected, transversions were less frequent than transitions (Fig. 5a). Reciprocal transversions (for example, T to A and A to T substitutions) showed similar rates, with the exception of C⇔A substitutions, where C to A substitutions were consistently more frequent than A to C substitutions.



As well as C⇔T substitutions, where C to T substitutions were more frequent than T to C substitutions.

We also found a very strong influence of the 3′ nucleotide on non-symmetrical substitutions rates. For example, CG to AG substitutions were clearly enriched compared to all other C to A substitutions and fully accounted for the enrichment of C to A transversions. Similarly, the increased C to T transition rate was only present in CG context, where CG to TG substitutions were much more frequent than any other type of C to T substitutions (Fig. 5b).

Intriguingly, in contrast to CG to AG, the increase of CG to TG substitution rates showed great variation among the diploid species, ranging from 5.7–13.1% (inlay in Fig. 5b). This degree of variation could not be observed for any other type of substitutions (independent of its actual frequency).

Using these highly variable CG to TG substitutions rates as phenotypes, we applied PAM to identify candidates for mutational modifiers that could be responsible for these differences. Across all 22,260 gene families, PAM resulted in 55 gene families where the pattern of copy-number differences between the species were highly significantly associated with the variation in CG to TG substitutions across the 25 diploid species.

To focus on likely candidate genes, we screened the *A. thaliana* members of the gene families for function related to DNA integrity. We identified two genes, one of them with function in DNA damage response, *DNA-DAMAGE-REPAIR/TOLERATION 2* (*DRT102*; $P=0.007$; rank = 37), which is involved in DNA repair of ultraviolet light-induced damage[55]. Even though ultraviolet light induces predominately C to T mutations, the ultraviolet mutation-prone sequence context is TC to TT and is therefore different to the CG to TG sequences, which are enriched in the substitutions[14].

The second gene with function related to DNA integrity was *CMT3* ($P=0.007$; rank = 38), which is involved in catalysing CHG methylation[56] (where H = A, C or T). *CMT3* is also required for gbM[13], which occurs exclusively in CG context and thus matches the context that showed increased substitution rates. Moreover, DNA methylation has previously been connected to increased mutation rates[14,15]. For example, DNA methylation increased the mutation rates in *A. thaliana* mutation accumulation lines grown under controlled conditions[16], and similar patterns were observed in the genomes of nearly identical *A. thaliana* accessions from North America[15]. In agreement with this, species without *CMT3* orthologues showed reduced levels of CG to TG substitutions compared to species with *CMT3*, suggesting that reduced methylation rates might lead to reduced mutation rates (Fig. 5c). Overall, *CMT3* was lost in eight of the 48 samples, probably from five independent gene loss events (Fig. 5d), while orthologues of the two most closely related CHROMOMETHYLASES to *CMT3*—*CMT2* and *DMT4*—were included in all diploid species.

Variation in gbM correlates with CG to TG mutation rates. Recent work described the absence of *CMT3* in two Brassicaceae species, *Eutrema salsugineum* and *Cassia planisiliqua*, and the associated loss of genic CG methylation, which is the hallmark of gbM[13]. As methylated cytosines are known to be more mutation-prone compared to non-methylated cytosines, it is possible that the loss of *CMT3* leads to reduced mutation rates via its associated loss of gbM.

To further corroborate the role of *CMT3* during the loss of gbM, we analysed the genome-wide DNA methylome of ten of the diploid Brassicaceae species, including three species without *CMT3*. These three species included two previously analysed species without *CMT3*, as well as *E. syriacum*, which experienced an independent loss event of *CMT3* and had not been analysed before, and thus can act as an additional line of evidence.

For each species, we assessed the percentage of methylated CG sites within 1,772 single-copy gene families shared by each of the ten species. DNA methylation of CG sites was symmetrical in all samples, including *A. alpina* for which the absence of symmetrical DNA methylation has been described before[17]. Such differences in symmetrical methylation might be explained by environmental factors, such as different growth conditions, or spontaneous genetic variation in otherwise isogenic material[57]. In contrast, gbM was variable between the species and absent in all three species without *CMT3*, including the newly assessed *E. syriacum*. All other genomes with *CMT3* showed the typical footprints of gbM, with the exception of *Schrenkiella parvula* where only a few genes were methylated (Fig. 5e and Supplementary Fig. 14). The reduced levels of gbM in *S. parvula* have been described before and were addressed to reduce selective constraints of the *CMT3* gene[58,59]. The relaxation of sequence conservation might have allowed for the introduction of variation that could have affected the function of *CMT3* within this clade and thereby affected gbM[59]. This would explain the reduced gbM in *S. parvula* despite the presence of a *CMT3* orthologue. In addition to the differences in gbM, it has also been reported that the CHG methylation levels in repetitive regions were significantly lower in species without *CMT3* (ref. [13].). In agreement with this we found that the methylation levels of methylated CHG sites were drastically reduced in those species with low gbM (Supplementary Fig. 15).

During the analysis of the methylomes, we observed a set of 393 genes that was consistently unmethylated across all genomes, even in those with otherwise clear evidence of gbM (grey bar in Fig. 5f). This conserved set of single-copy orthologues, which have not been methylated for long in evolutionary terms, allowed us to test whether differences in gbM lead to differences in substitution rates. Such evolutionary unmethylated genes should show lower CG to TG substitution rates, even when compared to methylated genes within the same species where other transfactors can be excluded for putative mutation rate differences. To test this, we analysed the substitution rates of evolutionary methylated genes versus unmethylated genes in *A. thaliana* (blue box in Fig. 5f). Intriguingly, the genes with gbM showed a significantly increased CT to TG substitution rate compared to the mutation rates in genes without gbM (Kolmogorov–Smirnov, $P<1.6\times10^{-8}$; Fig. 5g). Even though we cannot exclude other evolutionary constraints that could also contribute to these differences, this further supports our suggestion that variation in gbM can lead to differences in substitution rates.

Discussion

Here we presented a set of representative individuals of 47 Brassicaceae species along with genome assemblies for each individual, which we refer to as the Brassicaceae Diversity Panel. We used this set to perform interspecies association mapping of leaf shape complexity and CG to TG substitution rate differences to their underlying genetic architectures. This was established as an association of quantitative measurements to gene copy numbers and gene absence/presence patterns. Marker-associated methods would not work for such interspecies associations due to the highly disturbed linkage within plant genomes[60]. As a consequence, the genetic variation associated with the phenotype needs to contribute directly to the phenotypic variation and it is therefore necessary to select genetic variation that explains part of the phenotypic variation.

Differences in the gene content is an obvious choice for this because there is growing evidence that gene number changes (in particular gene loss) frequently drives species differences[61]. Trait reversal can be achieved easily by loss of function mutations in genes, and thereby facilitate subsequent diversification through the release of vestigial traits. This, however, does not exclude other types of variation; more complex genomic differences or epistatic effects could also contribute to the phenotypes.

Using this interspecies quantitative association mapping to correlate genomic differences to differences in mutation rates revealed a correlation between CG to TG substitutions in genes and the loss of *CMT3*, which was proposed to be necessary for gbM[13]. The fact that this correlation can be identified in a quantitative PAM study of only 25 diploid species points to a major contribution of gbM variation to CG to TG substitution rates in genes. As a result, this suggests an evolutionary trade-off between the unknown function of gbM and reduced CG to TG substitution rates and points to *CMT3* as a potential candidate as a modifier of CG to TG mutation rates in the Brassicaceae species.

In addition to CG to TG substitutions, CG to AG substitutions also consistently increased across all 47 species in our panel. This recapitulates the increase of mutations at methylated sites in natural accessions of *A. thaliana*, which not only increased C to T, but also C to G/A mutations[15]. However, in contrast to CG to TG, CG to AG rates only showed minor variation across species, which was not significantly correlated to the presence or absence of *CMT3*. This supports the idea that the general increase of cytosine mutations cannot be explained by DNA methylation alone[16].

In summary, natural phenotypic variation between closely related species is a rich, but still under-explored resource. Here we introduced the Brassicaceae Diversity Panel, which consisted of 47 species, and introduced a quantitative association mapping method that correlated quantitative phenotypic differences to genes with major effects. To gain a deeper insight into the differences between species, we are currently working on expanding and inbreeding more Brassicaceae species to improve the Brassicaceae Diversity Panel.

Methods

Sequencing, assembly, annotation and analysis of genomes. In total, we selected 48 samples based on their placement across the Brassicaceae phylogeny representing 47 species (from 23 of the 51 tribes), most of which were retrieved from botanic gardens. For 17 species, high-quality reference sequences were available. The genomes of the remaining 31 species were de novo assembled from long-read technologies for two species and short-read technologies for 29 species using FALCON or SOAP assembly tools[62,63]. Long-read assemblies were polished with BWA (Burrows–Wheeler Aligner) and SAMtools[64,65]. For short-read assembly, we assessed the best *k*-mer size using SPAdes before removing redundant contigs using Redundans[66,67]. Genome sizes were estimated with findGSE[68]. Genes were annotated with a series of de novo and evidence-based gene prediction tools[69–73] and tested for completeness using BUSCO[74]. To identify repeats, we used RepeatMasker (http://www.repeatmasker.org) with a customized repeat database, including the P_MITE, PlantSat, TIGR, Repbase and PGSB databases. We also used dnaPipeTE[75–80] to identify de novo repeats. OGs were calculated with OrthoFinder[36] and refined using gene trees to separate ancient paralogs from actual orthologues.

Methylome sequencing data was analysed using the BS-Seeker2 pipeline[81].

Phylogeny and age estimates. Phylogenetic reconstruction was performed by RAxML-ng[82] using a concatenated, partitioned (using PartitionFinder[83–85]) alignment of 354 genes, and also by ASTRAL v.5.6.3 (refs. [86,87]) combining 352 individually reconstructed trees (RAxML-ng[82]). Age estimates were calculated using the program BEAST v.2.5.1 (ref. [87]).

Identification and phylogenetic placement of meso/neopolyploids. We used a support vector machine (R library 'e1071') to identify meso/neopolyploid species. For training we used 18 species for which the ploidy was known. The features included peaks in the *Ks* distribution, duplicated genes, single-copy genes and average gene family size. The performance of the support vector machine was tested following a leave-one-out strategy.

We used a modified version of MAPS tool to find shared polyploidization events. Closely related pairs of polyploid species were tested using gene families with two gene copies in both focal species. We decided whether an event was shared based on the number of gene trees that supported a duplication event versus the number of gene trees that did not support a common duplication. This iterative method was used to test all branches of the phylogeny, which potentially could reveal shared events. To come close to the time of the At-α polyploidization event we used a recently published approach[88,89].

Methylation analysis. The data sets consisted of four previously analysed data sets (*C. planisiliqua*[22], *A. thaliana*, *A. lyrata* and *Capsella rubella*[90]) and six newly generated methylomes with two replicates for each species.

PAM method. To account for the different relatedness of the species we used a phylogenetic framework for interspecies association (R library 'caper')[91]. As genotype we used the species-specific sizes of 22,260 gene families and correlated them to leaf shape complexity and CG to TG substitution rates. The *P* values of the slopes of the regressions were calculated with analysis of variance and were taken as significance values for the associations.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All sequencing data, assemblies and gene annotations generated in this project can be found in the European Nucleotide Archive under the project accession number PRJEB26555.

Code availability

The code used to generate associations, the species tree, and the genotypic and phenotypic data can be accessed at https://github.com/schneebergerlab/Scripts_supporting_Kiefer_Willing_et_al_2019.

Received: 15 August 2018; Accepted: 25 June 2019;
Published online: 29 July 2019

References

1. Huynen, M., Dandekar, T. & Bork, P. Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. *FEBS Lett.* **426**, 1–5 (1998).
2. Gaasterland, T. & Ragan, M. A. Constructing multigenome views of whole microbial genomes. *Microb. Comp. Genomics* **3**, 177–192 (1998).
3. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA* **96**, 4285–4288 (1999).
4. Aravind, L., Watanabe, H., Lipman, D. J. & Koonin, E. V. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl Acad. Sci. USA* **97**, 11319–11324 (2000).
5. Hiller, M. et al. A 'forward genomics' approach links genotype to phenotype using independent phenotypic losses among related species. *Cell Rep.* **2**, 817–823 (2012).
6. Prudent, X., Parra, G., Schwede, P., Roscito, J. G. & Hiller, M. Controlling for phylogenetic relatedness and evolutionary rates improves the discovery of associations between species' phenotypic and genomic differences. *Mol. Biol. Evol.* **33**, 2135–2150 (2016).
7. Delaux, P.-M. et al. Comparative phylogenomics uncovers the impact of symbiotic associations on host genome evolution. *PLoS Genet.* **10**, e1004487 (2014).
8. Griesmann, M. et al. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science* **361**, eaat1743 (2018).
9. Zhang, G. et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).
10. Burga, A. et al. A genetic signature of the evolution of loss of flight in the Galapagos cormorant. *Science* **356**, eaal3345 (2017).
11. Pease, J. B., Haak, D. C., Hahn, M. W. & Moyle, L. C. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biol.* **14**, e1002379 (2016).
12. Vlad, D. et al. Leaf shape evolution through duplication, regulatory diversification, and loss of a homeobox gene. *Science* **343**, 780–783 (2014).
13. Bewick, A. J. et al. On the origin and evolutionary consequences of gene body DNA methylation. *Proc. Natl Acad. Sci. USA* **113**, 9111–9116 (2016).
14. Willing, E.-M. et al. UVR2 ensures transgenerational genome stability under simulated natural UV-B in *Arabidopsis thaliana*. *Nat. Commun.* **7**, 13522 (2016).
15. Exposito-Alonso, M. et al. The rate and potential relevance of new mutations in a colonizing plant lineage. *PLoS Genet.* **14**, e1007155 (2018).
16. Ossowski, S. et al. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94 (2010).
17. Willing, E.-M. et al. Genome expansion of *Arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation. *Nat. Plants* **1**, 14023 (2015).
18. 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).
19. The *Brassica rapa* Genome Sequencing Project Consortium et al. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* **43**, 1035–1039 (2011).
20. Yang, J. et al. The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. *Nat. Genet.* **48**, 1225–1232 (2016).
21. Chalhoub, B. et al. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* **345**, 950–953 (2014).
22. 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.* **27**, 778–786 (2017).

23. Haudry, A. et al. An atlas of over 90,000 conserved non-coding sequences provides insight into crucifer regulatory regions. *Nat. Genet.* **45**, 891–898 (2013).
24. 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* **26**, 1925–1937 (2014).
25. Dassanayake, M. et al. The genome of the extremophile crucifer *Thellungiella parvula*. *Nat. Genet.* **43**, 913–918 (2011).
26. Yang, R. et al. The reference genome of the halophytic plant *Eutrema salsugineum*. *Front. Plant Sci.* **4**, 46 (2013).
27. Slotte, T. et al. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.* **45**, 831–835 (2013).
28. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
29. Hu, T. T. et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481 (2011).
30. Gan, X. et al. The *Cardamine hirsuta* genome offers insight into the evolution of morphological diversity. *Nat. Plants* **2**, 16167 (2016).
31. Lee, C.-R. et al. Young inversion with multiple linked QTLs under selection in a hybrid zone. *Nat. Ecol. Evol.* **1**, 0119 (2017).
32. Kiefer, M. et al. BrassiBase: introduction to a novel knowledge database on Brassicaceae evolution. *Plant Cell Physiol.* **55**, e3 (2014).
33. Koch, M. A., German, D. A., Kiefer, M. & Franzke, A. Database taxonomics as key to modern plant biology. *Trends Plant Sci.* **23**, 4–6 (2018).
34. Guo, X. et al. Plastome phylogeny and early diversification of Brassicaceae. *BMC Genomics* **18**, 176 (2017).
35. 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).
36. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
37. Duarte, J. M. et al. Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evol. Biol.* **10**, 61 (2010).
38. 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).
39. Huang, C.-H. et al. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Mol. Biol. Evol.* **33**, 394–412 (2016).
40. Nikolov, L. A. et al. Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. *New Phytol.* **222**, 1638–1651 (2019).
41. Kagale, S. et al. Polyploid evolution of the Brassicaceae during the Cenozoic Era. *Plant Cell* **26**, 2777–2791 (2014).
42. 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).
43. Li, Z. et al. Early genome duplications in conifers and other seed plants. *Sci. Adv.* **1**, e1501084 (2015).
44. Pagel, M. Inferring evolutionary processes from phylogenies. *Zool. Scr.* **26**, 331–348 (1997).
45. Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884 (1999).
46. Grafen, A. The phylogenetic regression. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **326**, 119–157 (1989).
47. Felsenstein, J. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.* **25**, 471–492 (1973).
48. Felsenstein, J. Phylogenies and quantitative characters. *Annu. Rev. Ecol. Syst.* **19**, 445–471 (1988).
49. Vuolo, F. et al. Coupled enhancer and coding sequence evolution of a homeobox gene shaped leaf diversity. *Genes Dev.* **30**, 2370–2375 (2016).
50. Harris, K. & Pritchard, J. K. Rapid evolution of the human mutation spectrum. *eLife* **6**, e24284 (2017).
51. Harris, K. Evidence for recent, population-specific evolution of the human mutation rate. *Proc. Natl Acad. Sci. USA* **112**, 3439–3444 (2015).
52. Lynch, M. Evolution of the mutation rate. *Trends Genet.* **26**, 345–352 (2010).
53. Mathieson, I. & Reich, D. Differences in the rare variant spectrum among human populations. *PLoS Genet.* **13**, e1006581 (2017).
54. Gou, L., Bloom, J. S. & Kruglyak, L. The genetic basis of mutation rate variation in yeast. *Genetics* **211**, 731–740 (2019).
55. Pang, Q., Hays, J. B., Rajagopal, I. & Schaefer, T. S. Selection of *Arabidopsis* cDNAs that partially correct phenotypes of *Escherichia coli* DNA-damage-sensitive mutants and analysis of two plant cDNAs that appear to express UV-specific dark repair activities. *Plant Mol. Biol.* **22**, 411–426 (1993).
56. Du, J., Johnson, L. M., Jacobsen, S. E. & Patel, D. J. DNA methylation pathways and their crosstalk with histone methylation. *Nat. Rev. Mol. Cell Biol.* **16**, 519–532 (2015).
57. Becker, C. et al. Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **480**, 245–249 (2011).
58. Niederhuth, C. E. et al. Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* **17**, 194 (2016).
59. Bewick, A. J. et al. The evolution of CHROMOMETHYLASES and gene body DNA methylation in plants. *Genome Biol.* **18**, 65 (2017).
60. Zhao, T. & Schranz, M. E. Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. *Proc. Natl Acad. Sci. USA* **116**, 2165–2174 (2019).
61. Albalat, R. & Cañestro, C. Evolution by gene loss. *Nat. Rev. Genet.* **17**, 379–391 (2016).
62. Chin, C.-S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
63. Li, R. et al. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
64. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
65. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
66. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
67. Pryszcz, L. P. & Gabaldón, T. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* **44**, e113 (2016).
68. Sun, H., Ding, J., Piednoël, M., Schneeberger, K. & Birol, I. findGSE: estimating genome size variation within human and *Arabidopsis* using k-mer frequencies. *Bioinformatics* **34**, 550–557 (2018).
69. 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).
70. 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).
71. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
72. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
73. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
74. 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).
75. 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).
76. Macas, J., Mészáros, T. & Nouzová, M. PlantSat: a specialized database for plant satellite repeats. *Bioinformatics* **18**, 28–35 (2002).
77. 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–D363 (2004).
78. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
79. Nussbaumer, T. et al. MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic Acids Res.* **41**, D1144–D1151 (2013).
80. 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).
81. Guo, W. et al. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics* **14**, 774 (2013).
82. 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. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btz305 (2019).
83. 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).
84. 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).
85. 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).
86. 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).

87. Bouckaert, R. et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e1006650 (2019).
88. Vanneste, K., Van de Peer, Y. & Maere, S. Inference of genome duplications from age distributions revisited. *Mol. Biol. Evol.* **30**, 177–190 (2013).
89. 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).
90. 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).
91. Orme, D. The caper package: comparative analysis of phylogenetics and evolution in R (CRAN, 2018); https://cran.r-project.org/web/packages/caper/vignettes/caper.pdf

Acknowledgements

We would like to thank Bruno Hüttel (Max Planck Genome Centre, Cologne, Germany). This work was supported by the ERC STG Grant INTERACT 802629 (K.S.). Work on *BrassiBase* was funded within the framework of DFG-SPP 1529 (KO2302-13-1/2) (M.A.K.).

Authors contributions

K.S., M.A.K., E.-M.W. and C.K. designed research. C.K., E.-M.W., W.-B.J., H.S., M.P. and K.S. analysed the data. C.K., U.H. and B.H. performed the plant work. K.S. wrote the paper with the help of all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41477-019-0486-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to K.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

natureresearch

Corresponding author(s): Korbinian Schneeberger

Last updated by author(s): Jun 12, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ ☒ A description of all covariates tested

☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection: No software was used to collect data.

Data analysis: Genome assembly: soap (version 2.04), SPAdes (version 3.5.0), Redundans (version 0.12c), SMRT Analysis software (version 2.3), FALCON (version 3.0), Quiver (version 2.1.0), bwa (v0.7.12), SAMtools (not known)
Gene annotation: Augustus (version 3.0), glimmer (version 3.0.3), snap (version 2013-11-29), scipio(not known), EVM (version 2012-06-25), BUSCO (version 3), RepeatMasker (version 4.0.7), dnaPipeTE (version 1.3).
Orthologs/Phylogeny: OrthoFinder (version 1.1.4), agriGO (version 2.0), PartitionFinder (not known), RAxML (not known), ASTRAL (5.6.3), BEAST (2.5.1), Tracer (1.7.1)
Misc: MCL (version 14-137), R library “e1071”(version 1.7-0), ImageJ (not known), R library "caper" (version 1.0.1), BS-Seeker2 (version 2.0.8)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All sequencing data, assemblies and gene annotations generated in this project can be found in the European Nucleotide Archive under the project accession number PRJEB26555. We requested publishing of the data on June 11, 2019.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size of the association mapping is 48. Small size was chosen as a pragmatic compromise between maximizing species number and feasability. Simulations on such associations and the results of the association should that the size was sufficient for this study.
Data exclusions	No data was excluded from the study.
Replication	Replications were performed for the analysis of the DNA methylomes of 10 species. For each species the replication was highly successful.
Randomization	Samples were not selected randomly, but according to their phylogenetic placements. This has been taken care of in the association study by correcting for the expected covariation due to different relatedness.
Blinding	Blinding was not performed. The assessment of the phenotypes happens before the association. As the association are complex there is no chance to known on the influence of the assessed values. In this sense blinding was immanent in the experiment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging