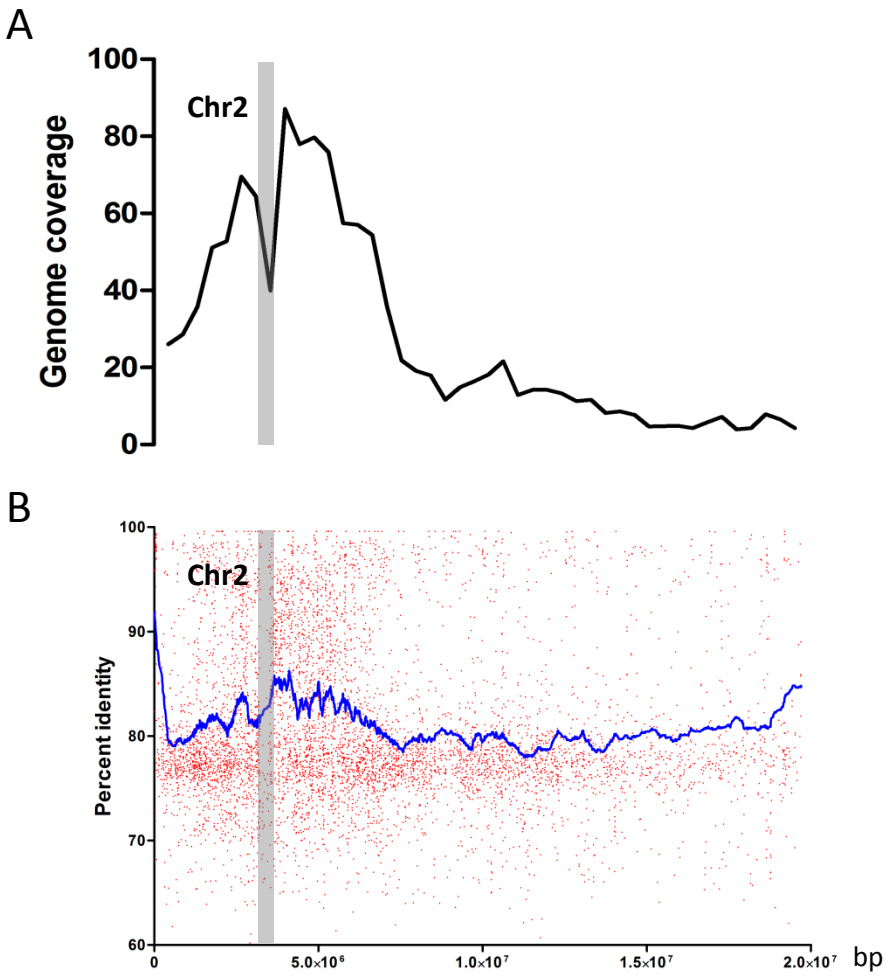
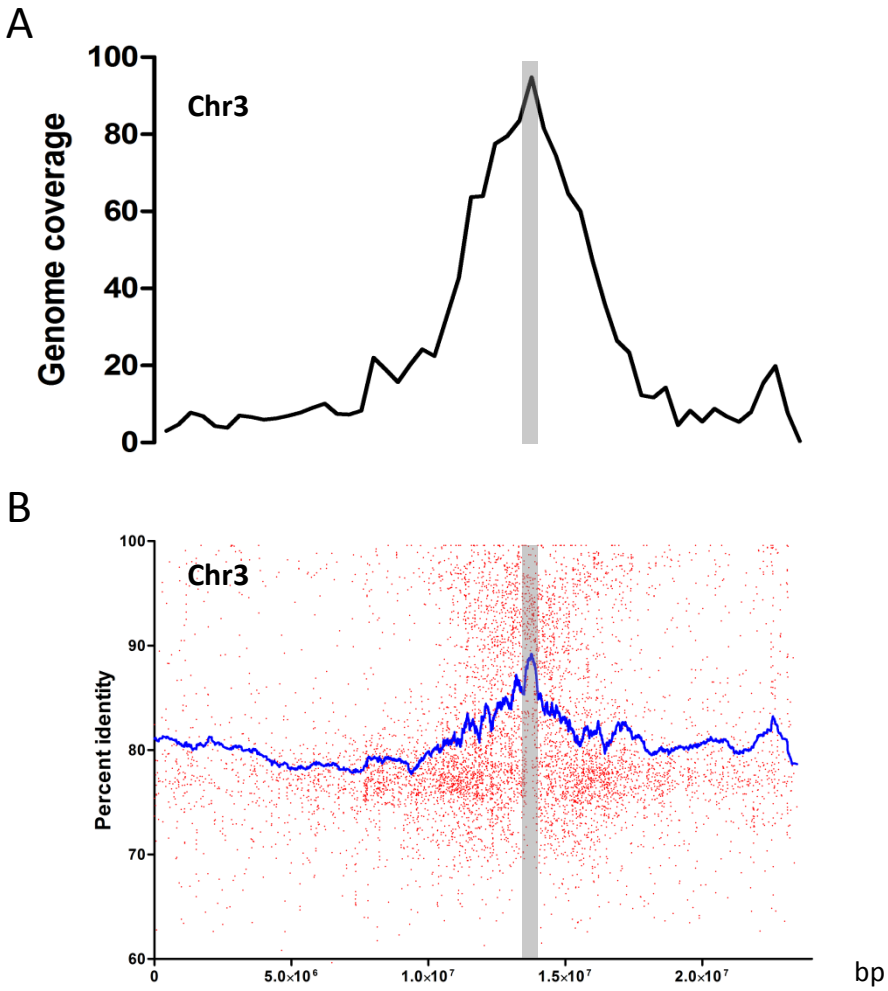
Supplementary Figure 1



Distribution of repeat density and identity along *A. thaliana* chromosome 2

(A) Percent repeat coverage per 500 kb bins. (B) plot (red dots) and smoothed curve (blue line) of the identities between genomic copies and consensus sequences. Grey shading indicates the centromere.

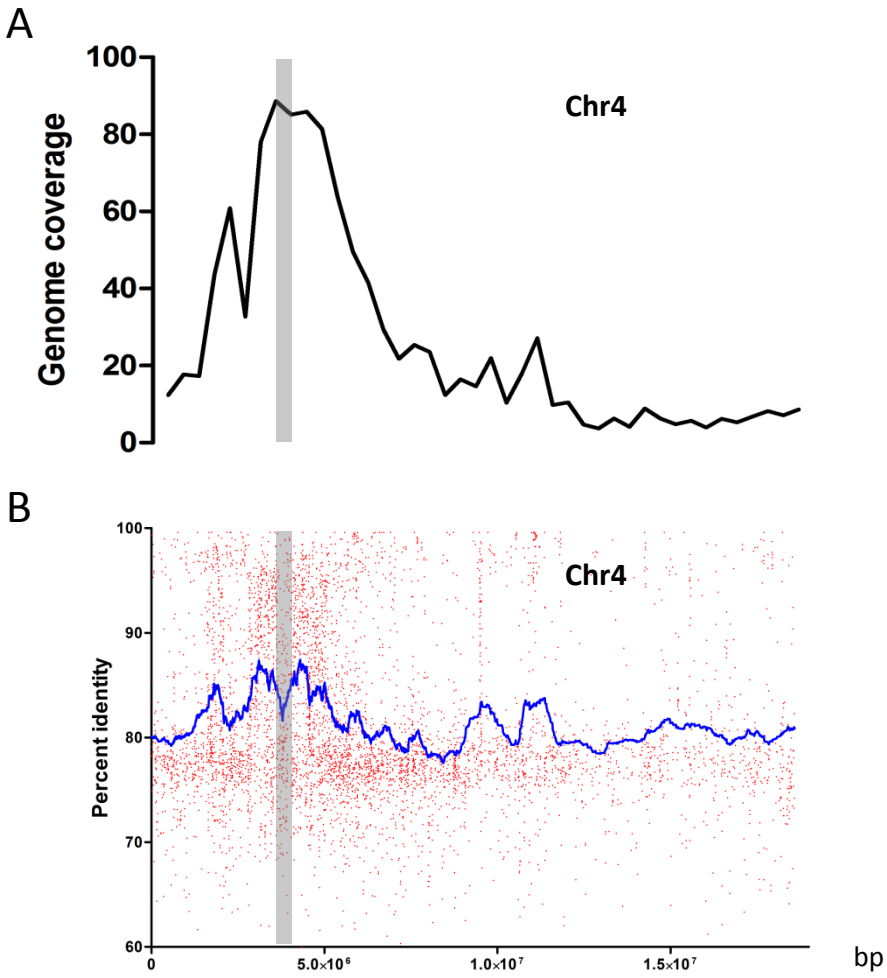
Supplementary Figure 2



Distribution of repeat density and identity along *A. thaliana* chromosome 3

(A) Percent repeat coverage per 500 kb bins. (B) plot (red dots) and smoothed curve (blue line) of the identities between genomic copies and consensus sequences. Grey shading indicates the centromere.

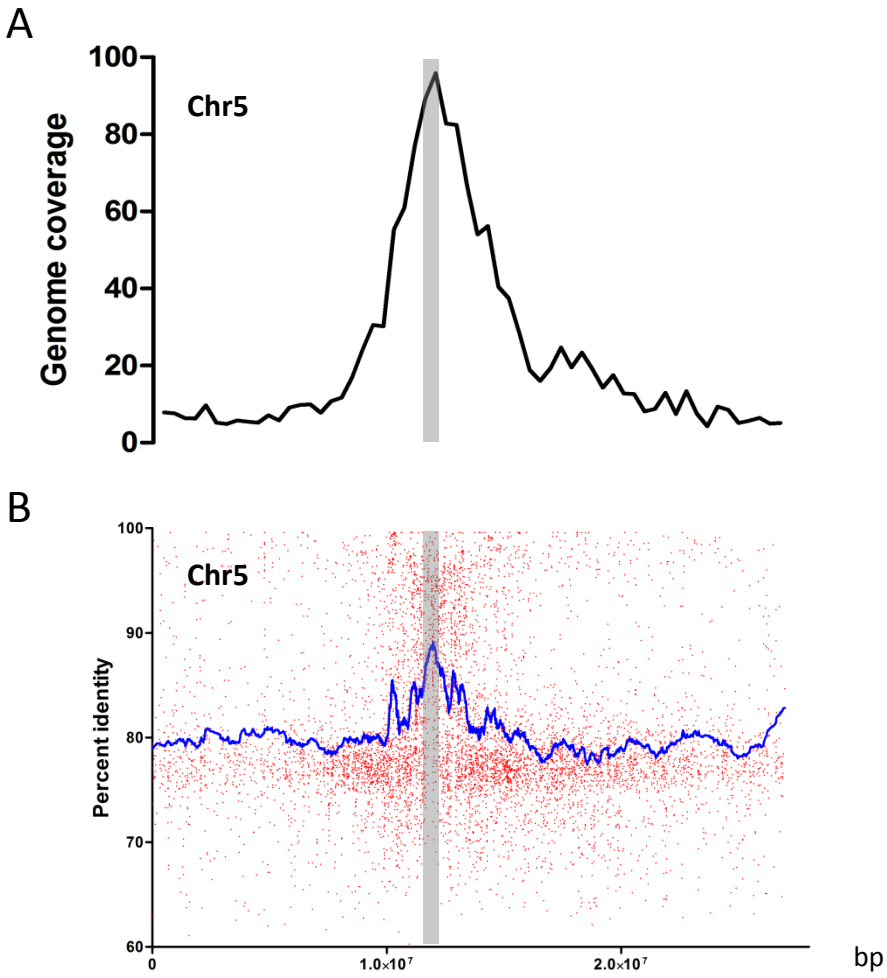
Supplementary Figure 3



Distribution of repeat density and identity along *A. thaliana* chromosome 4

(A) Percent repeat coverage per 500 kb bins. (B) plot (red dots) and smoothed curve (blue line) of the identities between genomic copies and consensus sequences. Grey shading indicates the centromere.

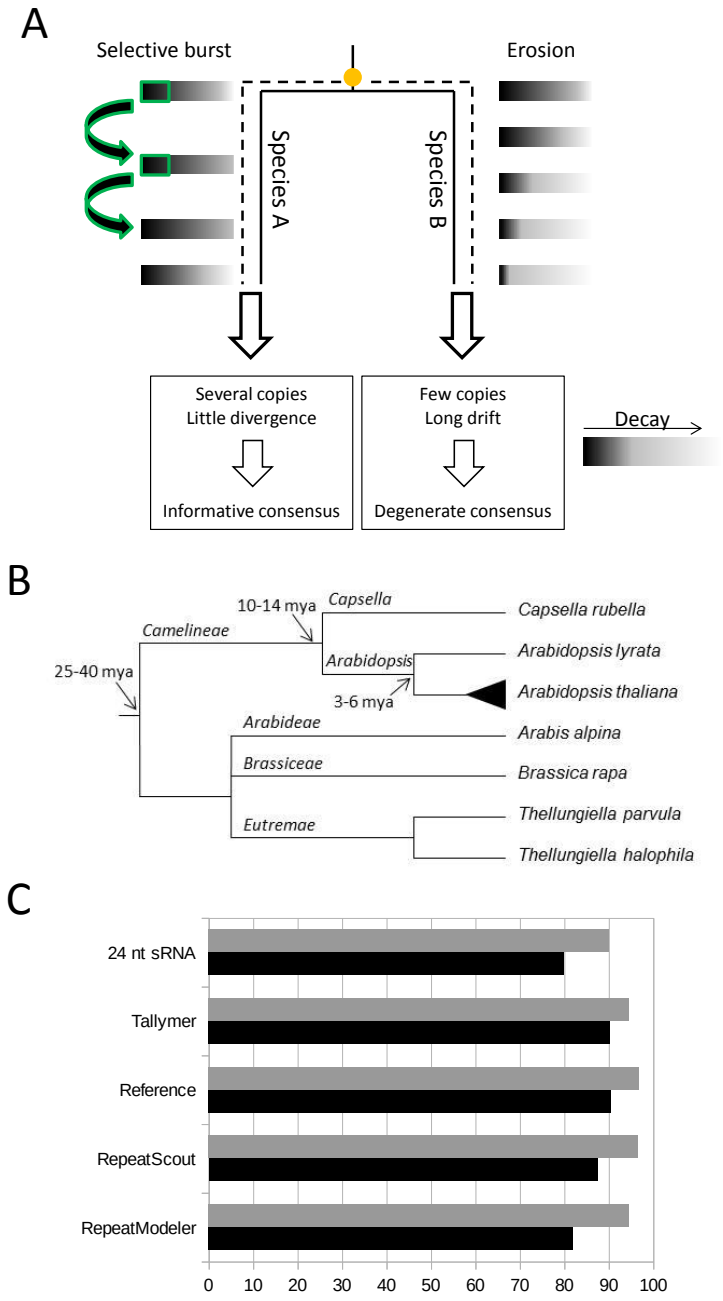
Supplementary Figure 4



Distribution of repeat density and identity along *A. thaliana* chromosome 5

(A) Percent repeat coverage per 500 kb bins. (B) plot (red dots) and smoothed curve (blue line) of the identities between genomic copies and consensus sequences. Grey shading indicates the centromere.

Supplementary Figure 5

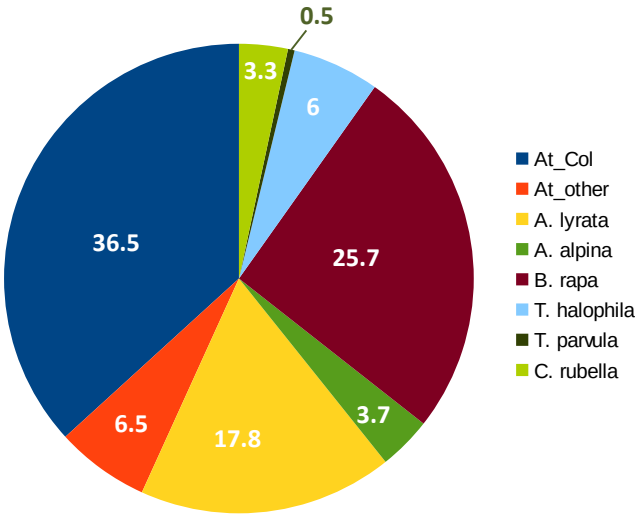


Rational, species, and sensitivity of the *Brassicaceae* annotation

(A) Schematic description of the selective burst hypothesis. From a common ancestor (orange dot), each family of autonomous elements can undergo one or more selective bursts (species A) or accumulate mutations and deletions (species B), leading to the differential conservation of ancestral information in modern species. Grey gradient boxes indicate the level of conservation

of the ancestral sequence among the copies of one repeat family. Green rectangles indicate the selection of the best conserved copies at each burst. (B) Cladogram representing the phylogenetic relationships between the *Brassicaceae* species used to construct the *Brassicaceae* library with TEdenovo. Arrows indicate branching dates as approximated from previous studies [1] [2]. (C) Coverage of different indicators of sensitivity by the annotations obtained using the Col-0 (black) and *Brassicaceae* (grey) libraries. The black triangle comprises five *A. thaliana* accessions: Col-0, Ler-1, Kro-0, Bur-0, and C24.

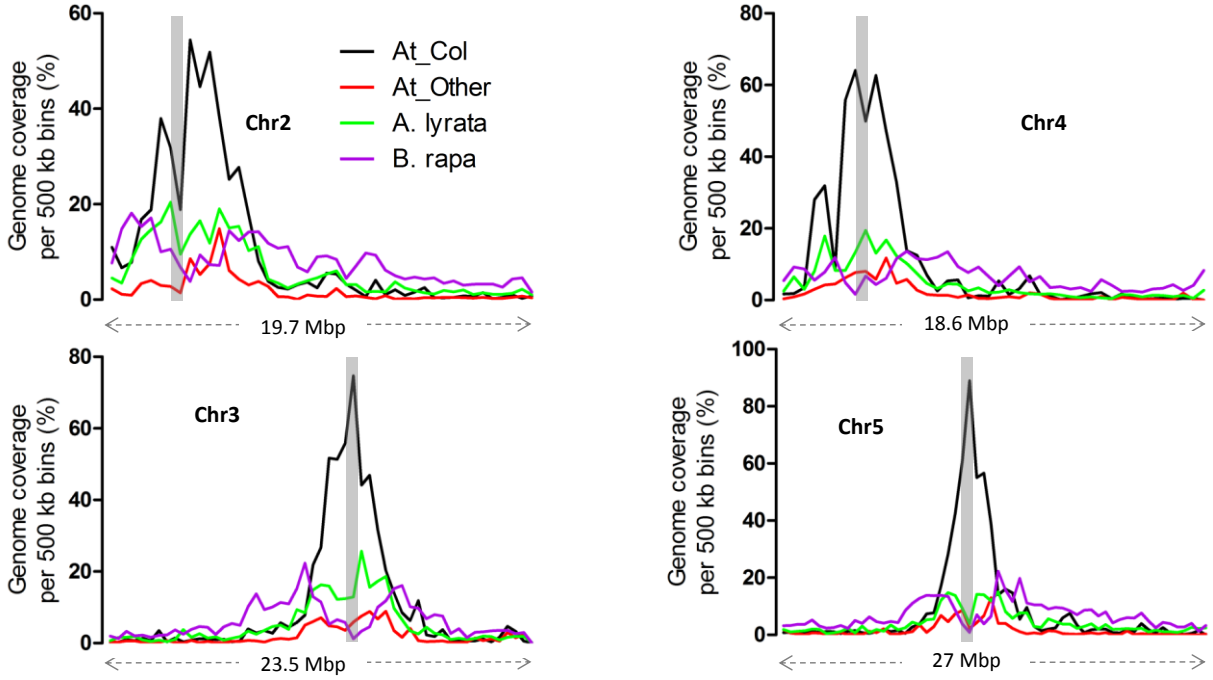
Supplementary Figure 6



Quantitative distribution of the *Brassicaceae* annotations in *A. thaliana*

Each copy in the *A. thaliana* genome is attributed (best score) to a consensus sequence from one of the *Brassicaceae* species. The pie chart indicates the contribution of copies attributed to each species in terms of percent coverage of the non-CDS *Brassicaceae* annotations. “At_other” represents the pool of Ler-1, Kro-0, Bur-0, and C24).

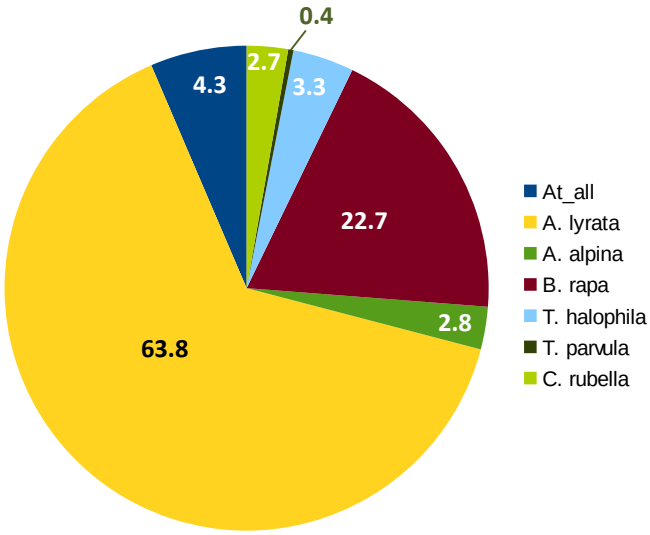
Supplementary Figure 7



Stratigraphy of the *A. thaliana* repeatome

Distribution along the Col-0 chromosomes of the contributions of the annotations attributed to consensus sequences from different *Brassicaceae* species and *A. thaliana* ecotypes ("At_other" represents the pool of Ler-1, Kro-0, Bur-0, and C24 accessions). Grey shading indicates the centromeres.

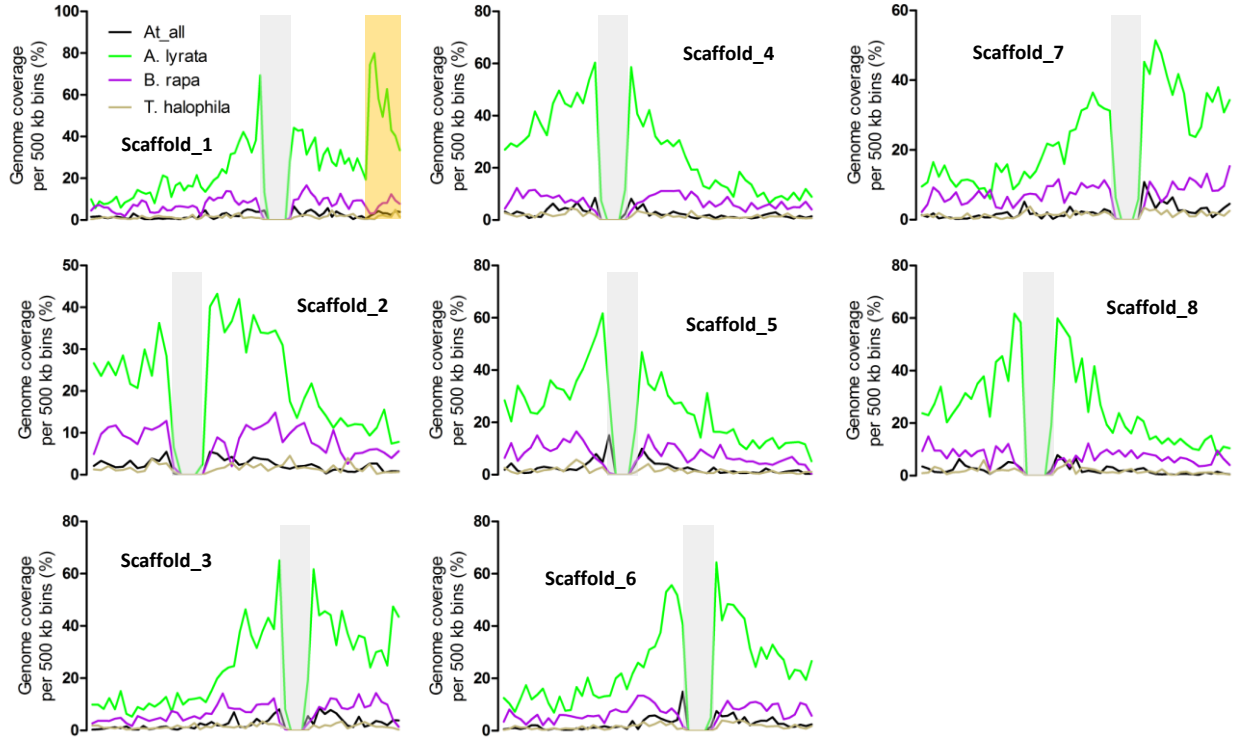
Supplementary Figure 8



Quantitative distribution of the *Brassicaceae* annotations in *A. lyrata*

Each copy in the *A. lyrata* genome is attributed (best score) to a consensus sequence from one of the *Brassicaceae* species. The pie chart indicates the contribution of copies attributed to each species in terms of percent coverage of the non-CDS *Brassicaceae* annotations. “At_all” represents the pool of Col-0, Ler-1, Kro-0, Bur-0, and C24 accessions).

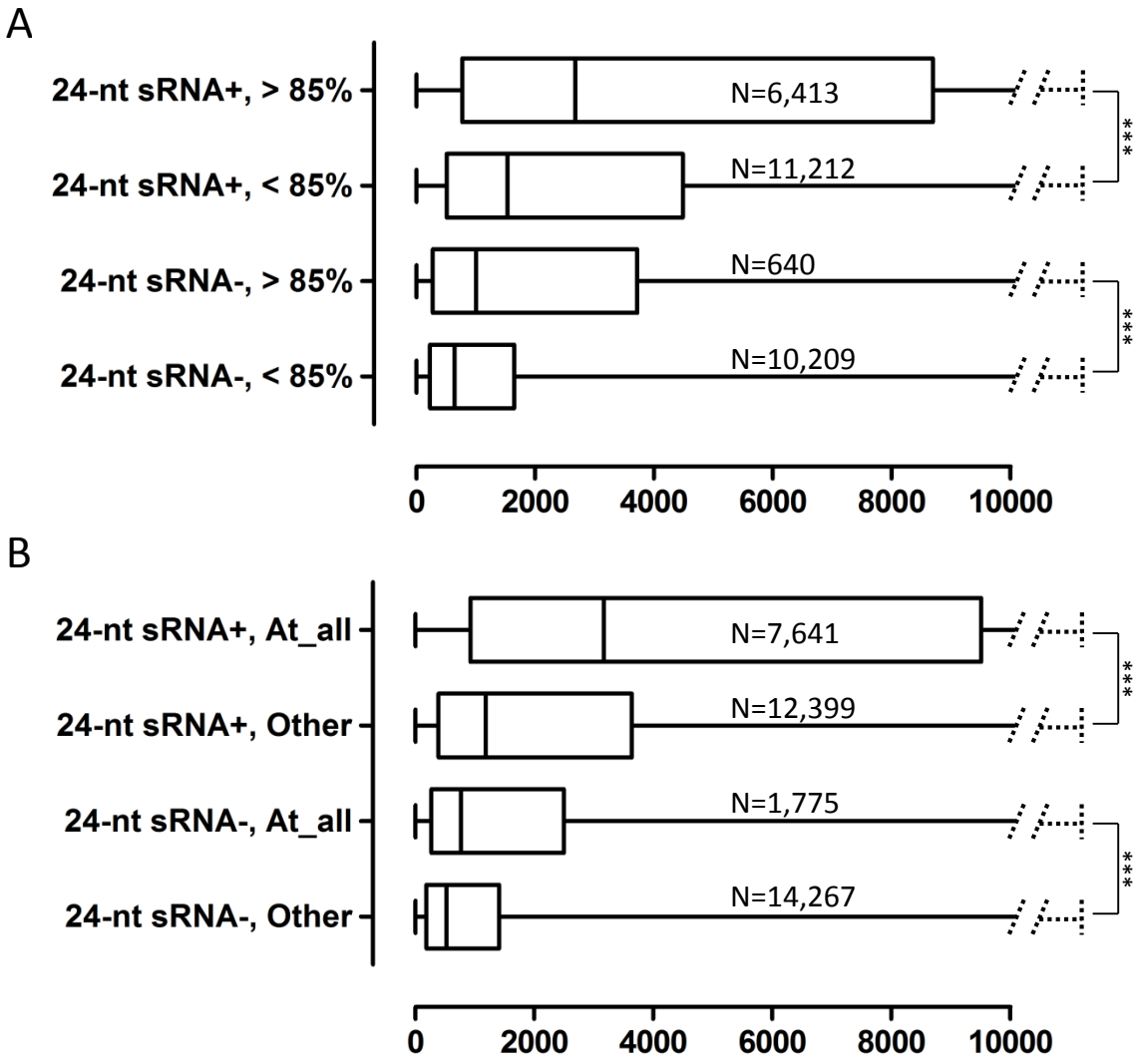
Supplementary Figure 9



Stratigraphy of the *A. lyrata* repeatome

Distribution along the *A. lyrata* chromosomes of the contributions of the annotations attributed to consensus sequences from different *Brassicaceae* species and *A. thaliana* ecotypes ("At_other" represents the pool of Ler-1, Kro-0, Bur-0, and C24 accessions). Grey shading indicates the centromeres. Orange shading in scaffold 1 indicates a probable error in the assembly of the *A. lyrata* genome: this region most likely corresponds to an unplaced pericentromeric region [3].

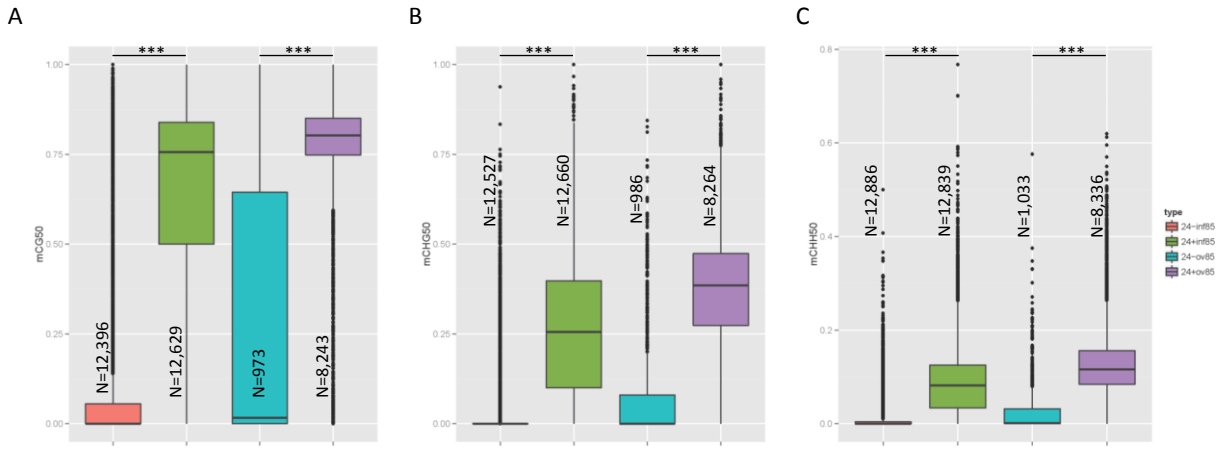
Supplementary Figure 10



Distance to genes according to 24-nt sRNA targeting

Distance of the repeat copies to the nearest CDS with respect to overlap with at least one 24-nt sRNA and copy age as approximated by percent identity with cognate consensus sequence (A) or attribution to a consensus sequence from *A. thaliana* (including all accessions) or not following the annotation using the *Brassicaceae* library (B). For all samples, error bars are defined as standard error of mean and extend to the maximum distance considered that was fixed at 25 kb for this analysis. *** indicates statistically supported differences (MWU P value < 0.0001).

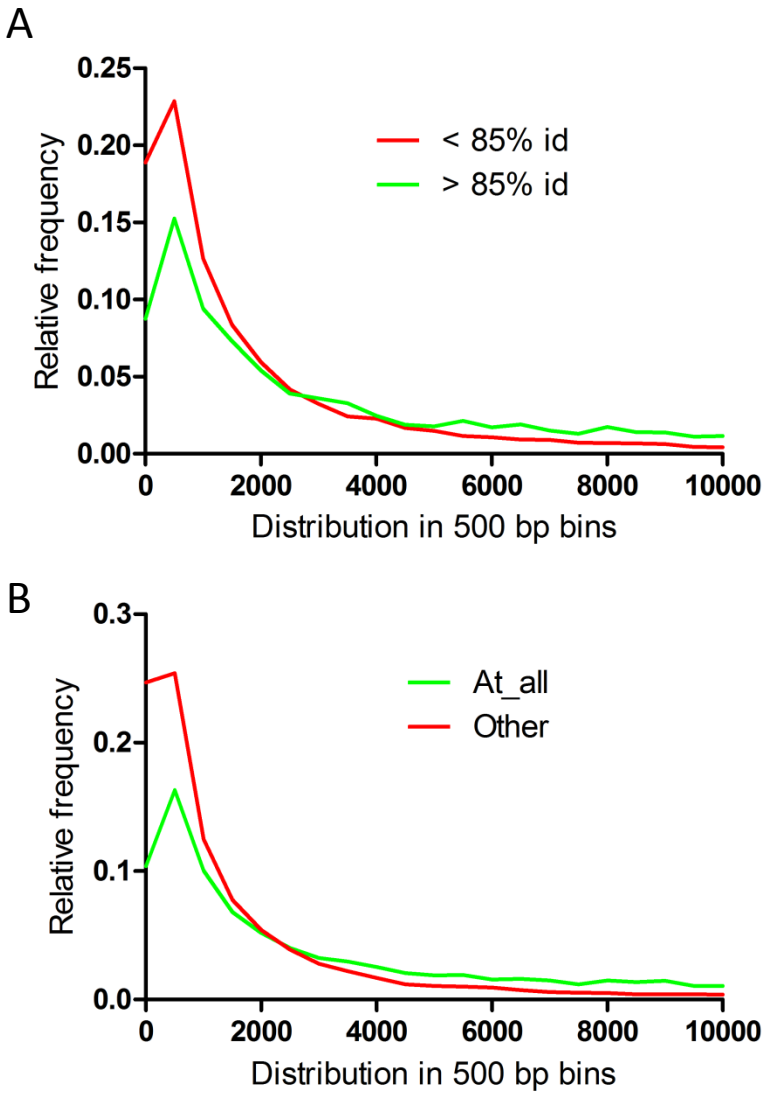
Supplementary Figure 11



Methylation levels in old and young repeats

CG50 (A), CHG50 (B) , and CHH50 (C) indicate methylation levels from 50-bp genomic windows over repeat copies regarding whether they are old 24-nt sRNA- (red), old 24-nt sRNA+ (green), young 24-nt sRNA- (blue), or young 24-nt sRNA+ (purple) . *** indicates MWU P value < 0.0001. The bands inside the boxes represent median values. The bottom and top of the boxes indicate the first and the third quartiles, respectively. The ends of the whiskers represent the lowest data still within 1.5 x [interquartile range] of the lower quartile, and the highest data still within 1.5 x [interquartile range] of the upper quartile. Data beyond the end of the whiskers are plotted as points.

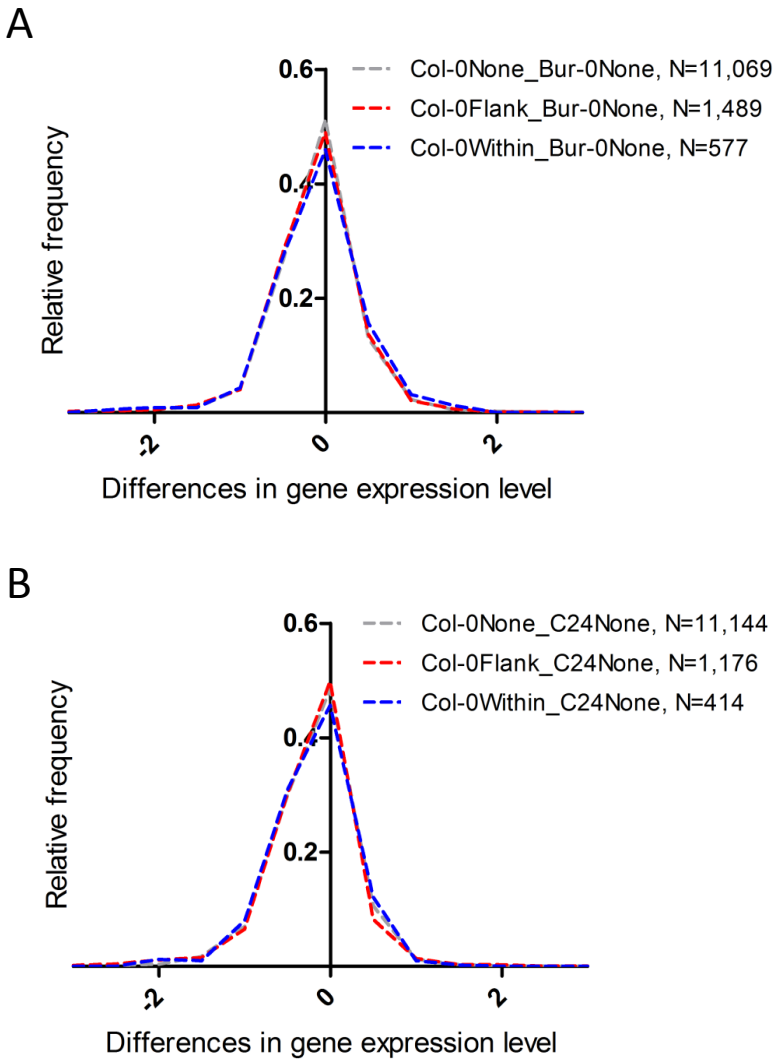
Supplementary Figure 12



Distance from repeats to genes

Distribution in 500 bp bins of the distance between each repeat and the nearest gene with respect to the identity with cognate consensus using the Col-0 library (>=85% or < 85%) (A) or to their attribution to consensus sequences obtained from *A. thaliana* ("At_all" refers to the pool of the accessions Col-0, Ler-1, Kro-0, Bur-0, and C24) or to consensus sequences obtained from other *Brassicaceae* species ("Other").

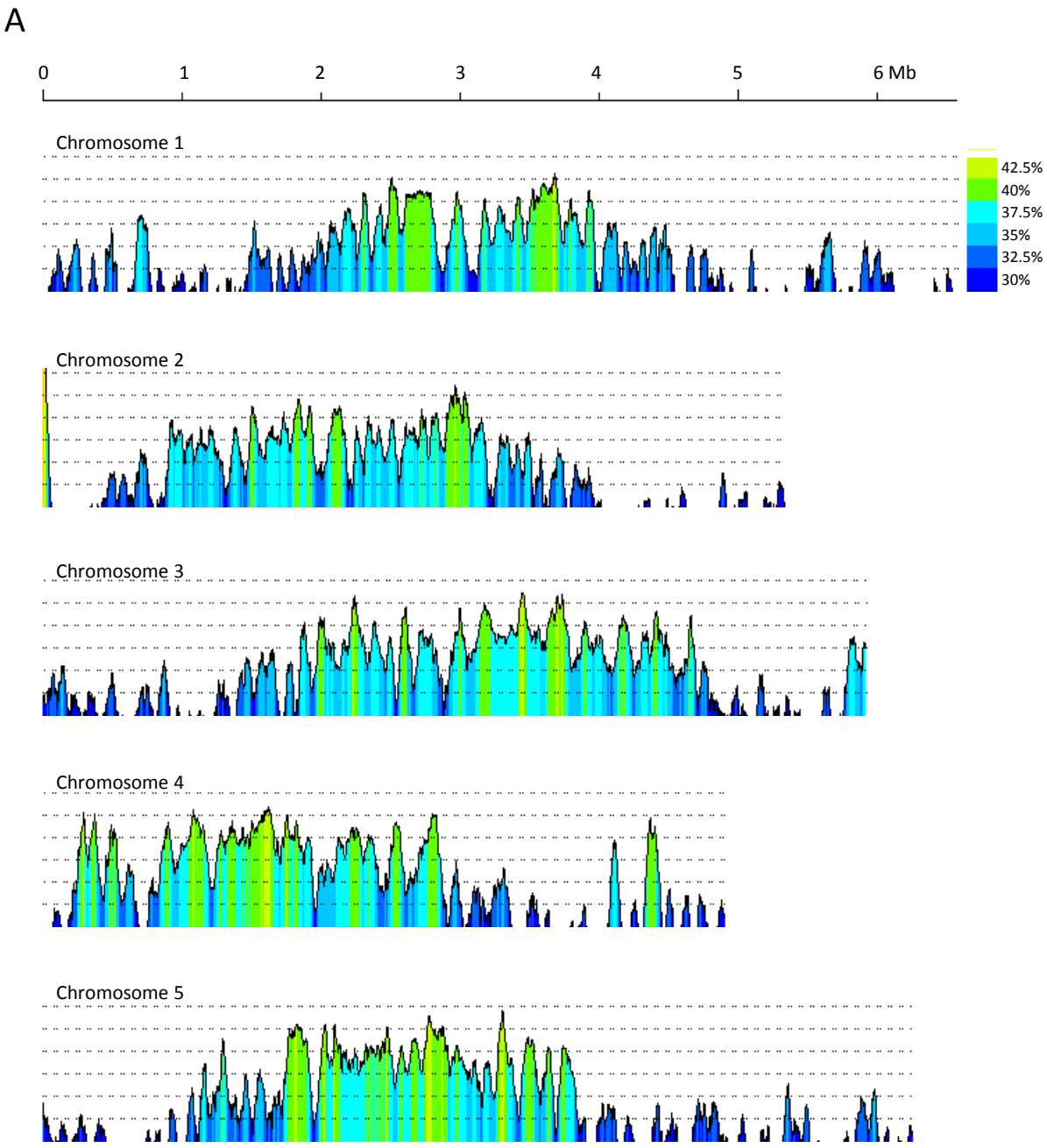
Supplementary Figure 13

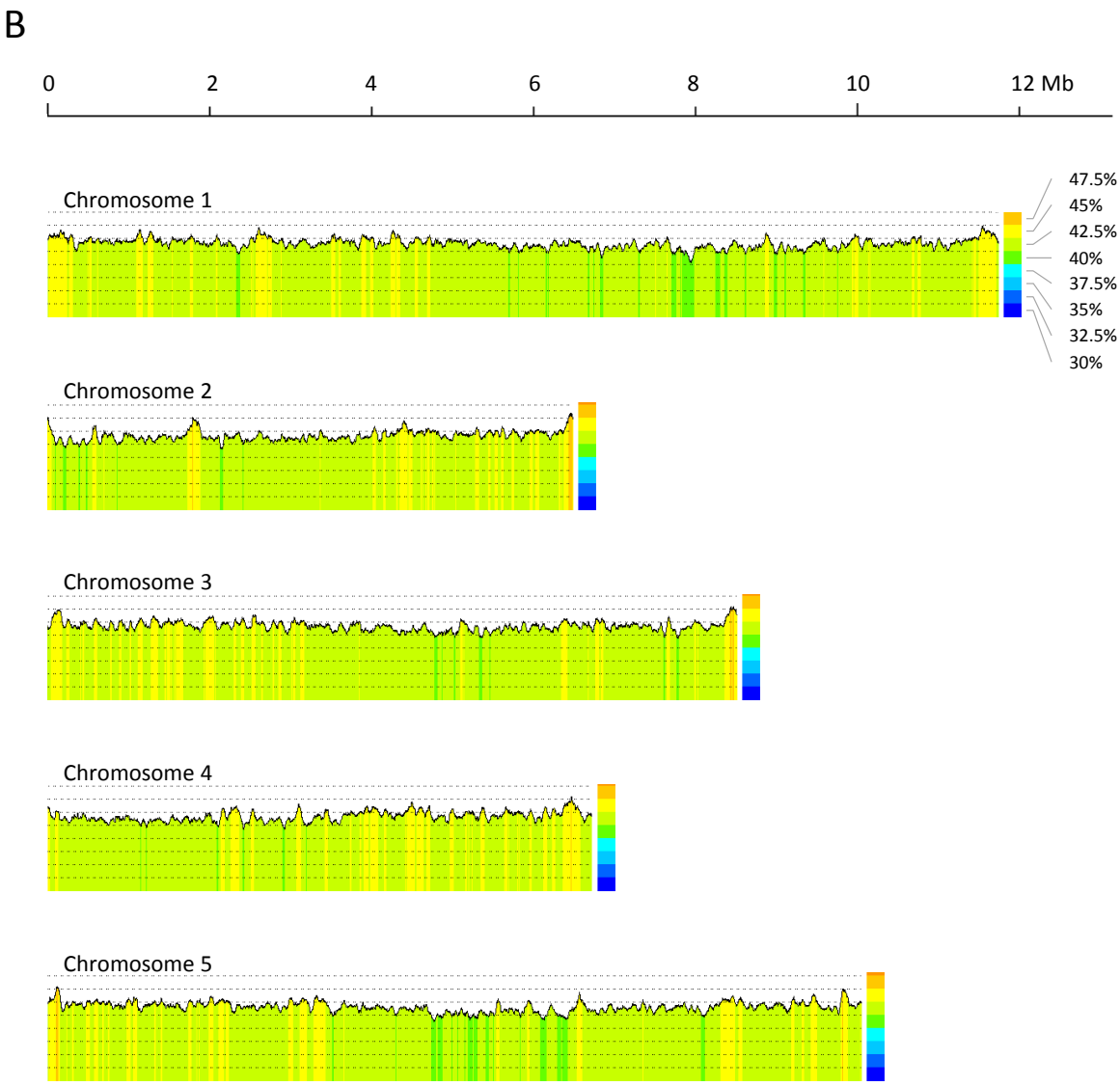


Differences in gene expression levels between accessions

(A) Differences in gene expression levels between three sets of Col-0/Bur-0 pairs of orthologous genes. The first set includes orthologs that are repeat-free in both accessions (grey), the second comprises genes with a flanking repeat in Col-0 and repeat-free orthologs in Bur-0 (red), the third set includes genes carrying a repeat in Col-0 and repeat-free orthologs in Bur-0 (blue). (B) Differences in gene expression levels between three sets of Col-0/C24 pairs of orthologous genes. Set of orthologous genes are defined as in (A) but C24 replaces Bur-0.

Supplementary Figure 14

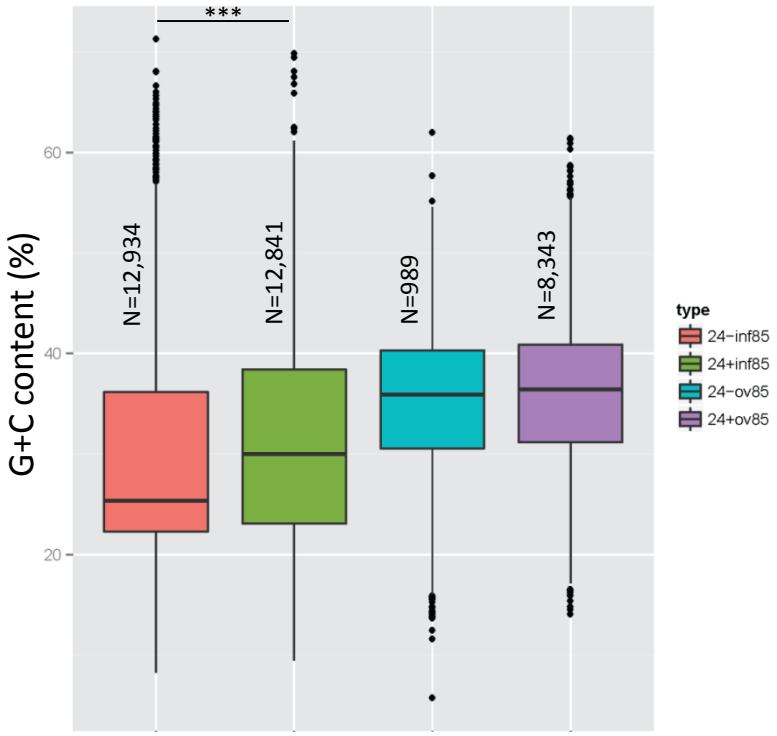




G+C landscape of the *A. thaliana* repeatome and CDS space

(A) Plots showing the G+C content along the concatenated repeat copies detected on each *A. thaliana* chromosome using the Col-0 library. Percentages color code indicates G+C content and gaps indicate values below 30%. (B) Plots showing the G+C content along the concatenated CDS from each *A. thaliana* chromosome.

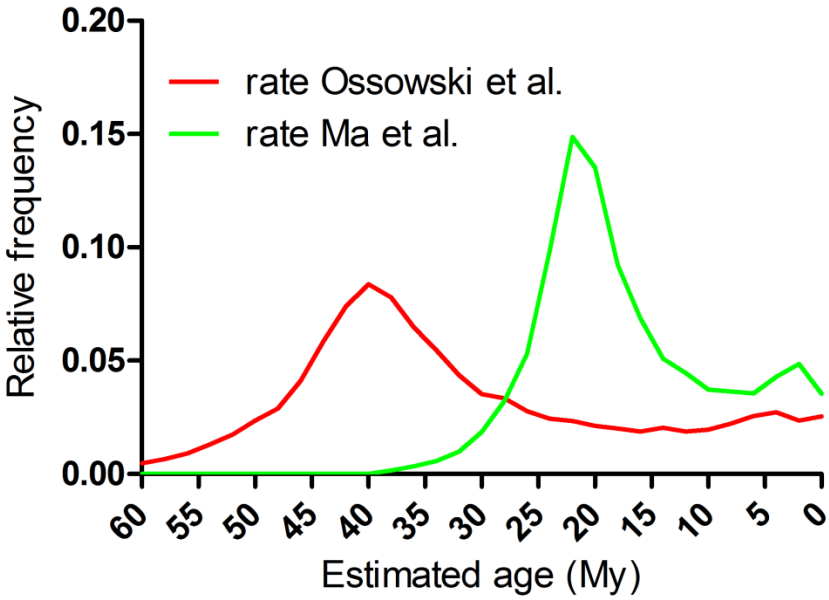
Supplementary Figure 15



Relationship between G+C content and 24-nt sRNA targeting

Plot indicating G+C content in four different sets of repeat copies i.e. old 24-nt sRNA- (red), old 24-nt sRNA+ (green), young 24-nt sRNA- (blue), and young 24-nt sRNA+ (purple). *** indicates statistically supported differences (MWU P value < 0.0001). The bands inside the boxes represent median values. The bottom and top of the boxes indicate the first and the third quartiles, respectively. The ends of the whiskers represent the lowest data still within 1.5 x [interquartile range] of the lower quartile, and the highest data still within 1.5 x [interquartile range] of the upper quartile. Data beyond the end of the whiskers are plotted as points.

Supplementary Figure 16



Age approximations of the *A. thaliana* repeats

Distribution in 2 million years (my) bins of copy ages estimated using two different rates from references [4 5].

SUPPLEMENTARY REFERENCES

1. Clauss MJ, Koch MA. Poorly known relatives of Arabidopsis thaliana. *Trends Plant Sci* **11**, 449-459 (2006).

2. Franzke A, Lysak MA, Al-Shehbaz IA, Koch MA, Mummenhoff K. Cabbage family affairs: the evolutionary history of Brassicaceae. *Trends Plant Sci* **16**, 108-116 (2011).

3. Slotte T*, et al.* The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat Genet*, (2013).

4. Ossowski S*, et al.* The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science* **327**, 92-94 (2010).

5. Ma J, Devos KM, Bennetzen JL. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res* **14**, 860-869 (2004).