DOI:

https://doi.org/10.1038/nplants.2014.23

Lack of symmetric CG methylation and long-lasting retrotransposon activity have shaped the genome of *Arabis alpina*

Eva-Maria Willing[1,*], Vimal Rawat[1,*], Terezie Mandáková[2,*], Florian Maumus[3], Geo Velikkakam James[1,10], Karl J.V. Nordström[1,10], Claude Becker[4], Norman Warthmann[4,5], Claudia Chica[6,10], Bogna Szarzynska[6,10], Matthias Zytnicki[3,10], Maria C. Albani[1,10], Christiane Kiefer[1], Sara Bergonzi[1,10], Loren Castaings[1,10], Julieta L. Mateos[1,10], Markus C. Berns[1], Nora Bujdoso[1,10], Thomas Piofczyk[1], Laura de Lorenzo[7,10], Cristina Barrero-Sicilia[8,10], Isabel Mateos[7,10], Mathieu Piednoël[1], Jörg Hagmann[4], Romy Chen-Min-Tao[6,10], Raquel Iglesias-Fernández[8], Stephan C. Schuster[9], Carlos Alonso-Blanco[7], François Roudier[6], Pilar Carbonero[8], Javier Paz-Ares[7], Seth J. Davis[1,10], Ales Pecinka[1], Hadi Quesneville[3], Vincent Colot[6], Martin A. Lysak[2], Detlef Weigel[4], George Coupland[1,‡], Korbinian Schneeberger[1,‡]

[*] These authors contributed equally.

[1] Max Planck Institute for Plant Breeding Research, Department of Plant Developmental Biology, Carl-von-Linné Weg 10, D-50829 Cologne, Germany.

[2] Research group Plant Cytogenomics, CEITEC – Central European Institute of Technology, Masaryk University, Brno, Czech Republic.

[3] INRA, UR1164 URGI—Research Unit in Genomics-Info, INRA de Versailles-Grignon, Route de Saint-Cyr, Versailles 78026, France.

[4] Max Planck Institute for Developmental Biology, Department of Molecular Biology, 72076 Tübingen, Germany.

[5] Research School of Biology, The Australian National University, Canberra, ACT 0200, Australia.

[6] *Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Centre National de la Recherche Scientifique (CNRS) UMR 8197 and Institut National de la Santé et de la Recherche Médicale (INSERM) U 1024, Paris, France.*

[7] *Department of Plant Molecular Genetics, Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas (CSIC), Campus de Cantoblanco, 28049 Madrid, Spain.*

[8] *Centro de Biotecnología y Genómica de Plantas (UPM-INIA). ETSI agrónomos, Universidad Politécnica de Madrid, Campus de Montegancedo, Pozuelo de Alarcón, 28223 Madrid, Spain*

[9] *Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, PA 16802, USA.*

[10] Present addresses:

Geo Velikkakam James: Rijk Zwaan R&D Fijnaart, Fijnaart, The Netherlands.

Karl J.V. Nordström: Laboratory of EpiGenetics, University of Saarland, Saarbrücken, Germany.

Claudia Chica: Departamento de Ciencias Biológicas, Universidad de los Andes, Carrera 1 Nº 18A- 12, Bogotá, Colombia.

Bogna Szarzynska-Erden: Center for Integrative Genomics, University of Lausanne, Genopode Building, CH-1015 Lausanne, Switzerland.

Matthias Zytnicki: INRA, MIAT UR-875, Castanet-Tolosan 31320, France.

Maria C. Albani: Botanical Institute, University of Cologne, Zülpicher Strasse 47B, D-50674 Cologne, Germany.

Sara Bergonzi: Laboratory of Plant Breeding, Department of Plant Sciences, Wageningen-UR, PO Box 386, 6700 AJ Wageningen, The Netherlands.

Loren Castaings: Biochimie et Physiologie Moléculaire des Plantes, Centre National de la Recherche Scientifique Unité Mixte de Recherche 5004, Institut de Biologie Intégrative des Plantes, 2 place Pierre Viala, F-34060 Montpellier Cedex 2, France.

Julieta Mateos: Fundación Instituto Leloir, Instituto de Investigaciones Bioquímicas de Buenos Aires-Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina, C1405BWE Buenos Aires, Argentina.

Nora Bujdoso: Currenta GmbH & Co. OHG, Department for Training CUR-BIL-BAN-ELB, Aprather Weg 18a, 42096 Wuppertal, Germany.

Laura de Lorenzo: Dept. Plant & Soil Sciences 301C Plant Science Building 1405 Veterans Drive University of Kentucky, Lexington, KY40546-0312 USA.

Cristina Barrero-Sicilia: Biological Chemistry and Crop Protection Department. Rothamsted Research, Harpenden, AL5 2JQ, U.K.

Isabel Mateos: Centro Hispano-Luso de Investigaciones Agrarias-CIALE, Salamanca, Spain.

Romy Chen-Min-Tao: Plateforme Bioinformatique, Institut Gustave Roussy, 94805 Villejuif, France.

Seth J. Davis: University of York, Department of Biology, York, UK

Despite evolutionary conserved mechanisms to silence transposable element (TE) activity, there are drastic differences in the abundance of TEs even among closely related plant species. We analysed the 375 Mb genome of the perennial model plant *Arabis alpina* and observed long-lasting as well as recent TE activity predominately driven by *Gypsy* long terminal repeat (LTR) retrotransposons. Their transposition extended the low-recombining peri-centromeres and transformed large and formerly euchromatic clusters of genes into repeat-rich peri-centromeric regions. This apparently reduced capacity for LTR retrotransposon silencing and removal in *A. alpina* co-occurs with unexpectedly low levels of DNA methylation. Most remarkable is the absence of symmetrical CG and CHG methylation suggesting strikingly reduced levels of DNA methylation maintenance in comparison to the related plant *Arabidopsis thaliana*. Phylogenetic reconstructions of genes in the DNA methylation pathways revealed species-specific patterns of evolution of the methylation maintenance machinery, in contrast to conserved family-wide patterns for *de novo* DNA methylation genes.

Whole-genome sequences of members of the *Brassicaceae* family[1-10] (including the plant model *A. thaliana*) are greatly expanding the scope for comparative genomics among closely related plant species (e.g. [5,8]). Despite their close phylogenetic relationship, *Brassicaceae* species have repeatedly evolved many differences in important life history traits, including the capacity for self-fertilization[11], senescence[12], as well as annual or perennial flowering behavior[13].

The reference accession of the perennial *A. alpina* Pajares was collected in the Cordillera Cantábrica mountains of Spain and was self-fertilized for six generations by single-seed descent[13]. We generated a high quality 309 Mb genome assembly from a mixture of 454, Illumina and Sanger BAC end sequences with a scaffold N50 of 788 and L50 of 160 kb (**Supplementary Fig. 1**, **Supplementary Table 1, 2** and **Supplementary Note**). By integrating comparative chromosome painting, we arranged more than 85% of the large scaffolds (>50 kb) into eight pseudo-molecules representing the eight chromosomes of *A. alpina* (**Supplementary Fig. 2-4** and **Supplementary Note**). During a manual annotation jamboree the structural annotation of 514 selected genes were curated from a total of 30,729 genes and 278,110 repetitive elements (**Supplementary Table 3, 4** and **Supplementary Note**). To resolve the phylogenetic placement of *A. alpina*, we calculated a whole genome-based phylogenetic consensus network for 10 *Brassicaceae* species with available reference assemblies (**Fig. 1a**). In addition to its topology, neutral variation and chromosome rearrangements specific to the karyotype of *A. alpina* argue for *A. alpina* being a member of a separate lineage, which diverged from the *Arabidopsis* lineage around 27±16 mya (**Fig. 1b**, **Supplementary Fig. 5**-**7** and **Supplementary Note**).

Within current annotations of *Brassicaceae* genomes, genes and introns account for 58 to 81 Mb, whereas TE content is much more variable and accounts for up to 148 Mb in *A. alpina* (**Fig. 1c**). By far the most abundant TE superfamily in *A. alpina* is the LTR retrotransposons superfamily*Ty3/Gypsy* (or *Gypsies*) (**Supplementary Table 5**). An increasing number of plant genomes have pointed to recent bursts of LTR retrotransposon transposition as a common phenomenon[2,14]. A hallmark of such recent transpositions are large amounts of young copies[9] (>95% sequence similarity). However, in *A. alpina* we found a large fraction of medium aged TEs (85% to 95% sequence similarity)[15]. The reduced amount of very young elements most likely indicates a recent reduction of *Gypsy* element activity as only small parts of this results from inefficiencies in the short read assembly. Rather large parts of the non-assembled sequence relates to simple sequence repeats (**Supplementary Fig. 8, 9** and **Supplementary Note**). Intriguingly, only the *Gypsies* contribute to the high number of medium-aged TEs in *A. alpina* (**Fig. 2a** and **Supplementary Fig. 10**). This suggests that *A. alpina Gypsy* elements proliferated over an extended period of time and that a large amount of these elements were retained and removed at only slow rates[16].

In order to analyse the degree to which these TEs may be epigenetically silenced we assayed four distinct chromatin marks in *A. alpina, A. thaliana* and *A. lyrata* including histone modifications H3K4me3, H3K27me3 and H3K27me1 assayed by ChIP-seq, and DNA methylation assayed by immunoprecipitation of methylated DNA coupled with high-throughput sequencing[17] (**Supplementary Note**). H3K27me1 and DNA methylation mark epigenetically silenced TEs, whereas H3K27me3 is a repressive mark specifically associated with genes and H3K4me3 is associated with regions that are actively transcribed[18]. As in *A. thaliana*[18], DNA methylation and H3K27me1 modification were mostly associated with TEs in *A. alpina* and *A. lyrata*, whereas H3K4me3 and H3K27me3 were preferentially associated with genes (**Fig. 2b** and **Supplementary Fig. 11**). In *A. alpina*, however, we found a three times larger proportion of TEs marked with H3K4me3 as in the other two species. *Gypsies* showed by far the largest fraction of elements marked with H3K4me3, whereas all other superfamilies did not show such a pronounced increase (**Fig. 2b** and **Supplementary Fig. 10, 11**)*.*

Even though *Gypsies* within genes were more likely to be marked with H3K4me3 and were consistently older in all three species (**Fig. 2c**), *A. alpina* showed only a slightly increased fraction (7.8%) of *Gypsies* in genes as compared to *A.*

thaliana (5.7%) and much less as compared to *A. lyrata* (15.1%) implying that elevated levels of H3K4me3 marking among *Gypsies* were not dominated by epigenetic states of genes. Moreover, H3K4me3-marked *Gypsy* elements outside of genes were drastically younger than those without this mark in *A. thaliana* and *A. lyrata* but not in *A. alpina* suggesting that many of these elements might have retained the ability to be transcribed over time. In fact, when analyzing the 1.5% of the RNA-seq reads that were not assigned to genes[9,19], we found that the *Gypsy* superfamily was more expressed than any of the other superfamilies in *A. alpina*, in contrast to *A. thaliana* in which Copias showed the highest fraction of RNA-seq reads (**Fig. 2d**, **Supplementary Fig. 12** and **Supplementary Note)**. Moreover, TEs with H3K4me3 mark were significantly enriched for expressed TEs as compared to TEs without this mark across all large superfamilies, even though this effect was less pronounced for *Gypsies* (**Fig. 2e**).

Two *Gypsy* families, *ATGPI* and *ATLANTYS2,* accounted for more than a fifth of all *Gypsy* elements in *A. alpina* (**Fig. 2f** and **Supplementary Note**). These two families showed an even more drastic increase in elements marked with H3K4me3, which was not apparent in any of the other *A. alpina* TE families and together with their increased copy number and age-distribution this suggests that the observed burst of transposition was mostly driven by this small group of TEs.

In *A. alpina, A. thaliana* and *A. lyrata*, TE density increases towards the centromeres[1,2] (**Fig. 3a** and **Supplementary Fig. 13, 14**). Typically these repeat-rich regions overlap with heterochromatic peri-centromeres. Here we defined peri-centromeres as regions with high amounts of H3K27me1 surrounding the centromeres (**Supplementary Table 6** and **Supplementary Note**). Peri-centromeres in *A. alpina* were drastically larger (average length: 14.9 Mb in *A. alpina*, 3.9 Mb in *A. thaliana*[1], 10.3 Mb in *A. lyrata*[2]) and included many more genes than the other two species (**Fig. 3b**). *Gypsy* elements are significantly enriched among the repeats in peri-centromeres (U-Test, p-value < 2e-16) and account for large parts of the size differences of peri-centromeres (**Fig. 3c**). In *A. thaliana* and *A. lyrata* H3K4me3 markings are strongly correlated with gene density throughout the chromosomes. In *A. alpina,* however, this correlation was weak and even entirely missing in peri-centromeres, where H3K4me3 was slightly correlated with Gypsy element density instead, suggesting that Gypsies are epigenetically active even in the heterochromatic peri-centromere in *A. alpina* (**Fig. 3a** and **Supplementary Fig. 15-17**).

Genome-size differences between *Brassicaceae* species have previously been attributed to peri-centromere expansion[6,20], but the causes and functional consequences have remained unclear. Centromeres in many species suppress crossover (CO) recombination during meiosis, a phenomenon that usually extends into heterochromatic regions near the centromere[21]. CO frequencies along seven investigated chromosomes of *A. alpina* revealed for each chromosome a region with suppressed COs (**Supplementary Table 7, 8** and **Supplementary Note**). These regions co-localize with large parts of the peri-centromeres implying that the extent of non-recombining DNA in *A. alpina* is greatly increased compared to *A. thaliana* and *A. lyrata*.

Earlier analyses reported differences in gene content in peri-centromeres of *Brassicaceae*[20], but were complicated by the lack of whole-genome sequences. Reconstruction of ancestral chromosomal rearrangements of *A. alpina* revealed a single homeologous paleocentromere (chromosome 2) with *A. lyrata*[2] (**Supplementary Fig. 2**, **3**). The assembly of the long arm of chromosome 2 shows a clear transition between gene- and repeat-rich regions in both species (**Fig. 3d**). Near the transition zone, there are 207 orthologs that reside in the repeat-dense regions in *A. alpina*, but outside the peri-centromere in gene-rich regions in *A. lyrata.* Comparing two sparse genetic maps of these species suggested that the repeat-rich region in *A. alpina* shows more strongly suppressed recombination than the orthologous regions in *A. lyrata* (**Fig. 3d** and **Supplementary Note**)[22]. This implies that upon expansion of the repeat-rich peri-centromeric regions in *A. alpina*, genes in formerly gene-rich regions became incorporated into the peri-centromere, with the consequence that large clusters of genes experience very little meiotic recombination in *A. alpina*. Although we cannot fully exclude the possibility of accelerated loss of TEs and peri-centromere shrinkage in *A. lyrata*, we found no evidence for large numbers of solo-LTRs that would indicate on-going loss through unequal homologous deletions in this particular genomic region (**Fig. 3d** and **Supplementary Note**).

Increased TE activity as well as gain of H3K4me3 has been linked to reduction in DNA methylation at TEs in *A. thaliana*[23]. To further examine DNA methylation in *A. alpina*, we performed whole-genome bisulfite sequencing[24] using leaf material and compared it to analogous data previously generated for *A. thaliana*[25] (**Supplementary Note**). Though this revealed similar amounts of methylated cytosines in *A. alpina* (19%) and *A. thaliana* (16%)[25] and similar methylation profiles along genes and TEs in both species[24] (**Supplementary Fig. 18**), the position-wise frequency of CG methylation was strikingly different. Whereas most methylated CGs showed 80-100% methylation in *A. thaliana*, these levels tended to be much lower in *A. alpina* irrespective of sequence annotation (**Fig. 4a, b** and **Supplementary Fig. 19**). In contrast, the distribution of methylation levels at CHGs was only slightly shifted towards lower values in *A. alpina*, and was very similar for CHH sites. In *A. thaliana*, CG and CHG methylation typically occurs on both Cs of the opposite strands of these palindromes, indicative of methylation copying via the maintenance machinery during replication[24,26]. Surprisingly, the two strands are essentially uncorrelated in their methylation levels at CG sites, and much more weakly correlated at CHG sites in *A. alpina* throughout the entire genome suggesting that methylation maintenance is much less pervasive (**Fig. 4c** and **Supplementary Fig. 20**).

Given these fundamental differences, we suspected that the DNA methylation maintenance machinery might function differently in *A. alpina*. To explore this possibility, we examined the *Brassicaceae* genomes for intact homologs of the five major gene families involved in DNA methylation[26] (**Fig. 4d**). At least one homolog of each family was found and showed expression in *A. alpina* (**Supplementary Table 9**). The phylogenies of *DDM1*, required for CG methylation maintenance, *CMT3*, required for CHG methylation maintenance, and *DRM2*, involved in *de novo* methylation in all contexts, broadly recapitulated the family phylogeny (**Fig. 4e, f, g**). In contrast, the homologs of *MET1* and *VIM1*, which in addition to *DDM1* are essential for CG methylation maintenance in *A. thaliana*[26], clustered in a species- and lineage-specific manner (**Fig. 4h, i**). This implies that all species outside of the *Arabidopsis* lineage, lacked clear orthologs for *MET1* and *VIM1* genes, which was also apparent from the lack of synteny of these genes with any of their homologs outside of this lineage. Moreover, d_N/d_S calculated for each gene family revealed values highly similar to a genome-wide background distribution, except for *MET1* family members with consistently enriched values suggesting that less purifying selection pressure acts on *MET1* (**Fig. 4j**).

Although *MET1* and *VIM1* homologs are present in *A. alpina*, it remains possible that the lineage-specific evolution of these genes might relate to the differences in CG methylation maintenance, as homologs of the main methylation genes are present in other species with strong differences in DNA methylation[27]. However more complex changes in other methylation pathways might need to be considered to reveal the basis of DNA methylation differences between *A. alpina* and *A. thaliana*. As the absence of symmetrical CG methylation levels did not correlate with an overall lower amount of at least partially methylated cytosines, *de novo* DNA methylation probably compensates for the lack of DNA methylation maintenance[28] underlining the high importance of *de novo* DNA methylation in *A. alpina*.

Even though co-occurrence of expanded TE content and DNA methylation maintenance deficiency in *A. alpina* does not necessarily imply a causal relationship, it nevertheless remains an attractive possibility, that apparent methylation deficiency may have contributed to the elevated numbers of *Gypsy* elements, possibly due to reduced silencing of specific TE families, as was shown for DNA methylation maintenance deficient mutants in *A. thaliana*[29,30].

Online Methods

Materials and methods are described in detail in the supplementary material.

Correspondence and requests for materials

Correspondence and requests for materials should be addressed to K.S. (schneeberger@mpipz.mpg.de) and G.C. (coupland@mpipz.mpg.de).

Acknowledgements

We thank José Jiménez-Gómez for helpful discussion throughout the entire project and Maarten Koornneef as well as Miltos Tsiantis for critical reading of the manuscript. We apologize to all colleagues whose work has not been cited here because of space constraints. The TRANSNET consortium was initiated based on the funding of PLANT KBBE "Transcriptional networks and their evolution in the Brassicaceae (TRANSNET)". B.S. was supported by a postdoctoral fellowship from the TRANSNET consortium. C.C. was supported by a postdoctoral fellowship first from Investissements d'Avenir ANR-10-LABX-54 MEMO LIFE and then from the European Union EpiGeneSys FP7 Network of Excellence (number 257082). R.C.M.T. was supported by Investissements d'Avenir ANR-10-LABX-54 MEMO LIFE. J.L.M was supported by an Alexander von Humboldt Postdoctoral Fellowship. C.B.S was supported by a postdoctoral fellowship from the TRANSNET consortium. R.I.F. was supported by a post-doctoral Juan de la Cierva contract (JCI-2010-07909) from Ministerio de Ciencia e Innovación (MICINN). A.P. was supported by the research grant PE-1853/2 from German Research Foundation. Work in the Carbonero's group was supported by the Spanish grants BFU2009-11809 and Consolider CSD2007-00057 from Ministerio de Ciencia e Innovación (MICINN). C.A.-B. laboratory was funded by grant BIO2013-45407-P from the Ministerio de Economía y Competitividad of Spain. Work in the Colot group was supported by the Agence Nationale de la Recherche (Investissements d'Avenir ANR-10-LABX-54 MEMO LIFE and ANR-11-IDEX-0001-02 PSL* Research University) and the European Union (EpiGeneSys FP7 Network of Excellence number 257082). Work in the Weigel group was supported by an FP7 project AENEAS and the TRANSNET consortium. Work in the Quesneville group was supported by the TRANSNET consortium and by the French national research agency (ANR-08-KBBE-012). Work in the Lysak group was supported by a research grant from the Czech Science Foundation (P501/12/G090) and by the European Social Fund (projects CZ.1.07/2.3.00/30.0037 and CZ.1.07/2.3.00/20.0189). Work on *A. alpina* in the Coupland laboratory is partly funded by the Cluster of Excellence on Plant Sciences (CEPLAS). Work in the Paz-Ares laboratory has been supported by the French-German-Spanish Trilateral program on Plant Genomics (grant TRANSNET) and Spanish Ministry of Economy competiveness (CONSOLIDER 2007-28317, BIO2011-30546). Schneeberger, Coupland, Weigel and Pecinka groups were supported by the Max Planck Society.

Author information

E.M.W., S.C.S., C.A.-B., F.R., P.C., J.P.A., S.J.D., A.P., H.Q., V.C., M.A.L., D.W., G.C. and K.S. conceived this study and supervised experiments and analyses. L.C., M.C.A., B.S., S.B., L.C., J.L.M., M.C.B., N.B., T.P., L.D.L., I.M. and C.B.S. prepared samples for DNA and RNA sequencing. B.S. prepared samples and performed ChIP and MeDIP experiments. C.B. prepared samples and performed bisulfite experiments. N.W., M.C.A. and C.A.-B. constructed genetic maps. T.M. and M.A.L. conducted FISH, chromosome painting and karyotype evolution analysis.

K.J.V.N., E.M.W. and N.W. conducted de novo assembly of *A. alpina*. C.K. conducted de novo assembly of *A. montbretiana*. G.V.J., E.M.W., C.B.S. and M.Z. performed genome annotations of *A. alpina* together with all participants of the annotation jamboree held in 2012 in Paris, France. E.M.W. performed genome annotations of *A. montbretiana*. E.M.W., V.R. and C.C. conducted expression analyses. F.M. performed transposon annotations. E.M.W., F.M., M.P. and K.S. performed transposon analysis. E.M.W., C.C., R.C.M.T. and K.S. performed analysis of ChIP-seq and MeDIP-seq data. E.M.W., C.B., J.H. and K.S. performed BS-seq analysis. E.M.W., V.R. and K.S. conducted comparative genomic analyses. E.M.W. and K.S. wrote the paper with contributions from all authors.

Competing financial interests

The authors declare that no competing interests exist.

Data availability

All data from this study have been deposited at the NCBI Sequence Read Archive (SRA) under BioProject PRJNA241291. The whole genome assembly of *A. alpina* has been deposited at DDBJ/EMBL/GenBank under the accession JNGA00000000. The version described in this manuscript is version JNGA01000000.

Figure legends

Fig. 1 | Phylogenetic reconstruction and karyotype evolution support a distinct phylogenetic placement of *A. alpina*. (a) Consensus network of the *Brassicaceae* phylogeny based on 1,787 single-copy COGs. Its topology did not unambiguously place the *Arabis* species with Lineage II as proposed earlier. (Lineage I species (red); Lineage II species (blue); *Arabis* species (green)). **(b)** Karyotype evolution at the base of *Brassicaceae* evolution. Reconstruction of the chromosome evolution from the Ancestral Crucifer Karyotype (ACK) to the *A. alpina* karyotype (KAA) suggested nine chromosomal rearrangements, which are different from the rearrangements that occurred in the evolution from the ACK to the Proto-Calepineae Karyotype (PCK), which is ancestral to Lineage II. **(c)** *Brassicaceae* genome compositions.

Fig. 2 | Genome size variation and differences in transposable element content. **(a)** TE-age spectra based on similarity between TE copies and consensus sequence. *A. alpina* shows an unique increase of medium-aged *Gypsy* elements*.* **(b)** Fraction of genes and TEs marked with H3K4me3. The three largest superfamilies are shown separately. **(c)** Age distribution of *Gypsies* inside and outside of genes separated by their different H3K4me3 markings. **(e)** TE superfamily expression estimated by the amount of non-genic RNA-seq reads. **(f)** Fraction of expressed TEs with and without H3K4me3 markings. **(g)** Size of the ten largest TE families in *A. alpina* along with their family-wide fraction of H3K4me3 marks within all three species.

Fig. 3 | Differences in the distribution of genes, TEs and chromatin marks between *A. thaliana*, *A. lyrata* and *A. alpina*. (a) Gene, TE and histone mark density, along orthologous chromosomes (missing sequence marked in grey). **(b)** Genomic fraction and gene space in chromosome arms and peri-centromeres. **(c)** Genome coverage of the three largest TE superfamilies. **(d)** Comparison of *A. alpina* and *A. lyrata* chromosome 2 sharing the same ancestral centromere. Grey lines connect single-gene orthologs. Orthologs that reside in peri-centromeric regions in *A. alpina*, but are outside these regions in *A. lyrata*, are indicated by dark grey lines. Locations of solo-LTRs indicated by grey crosses. (Gene and TE densities as in (a), CO frequency (red), peri-centromeres (dark brown)).

Fig. 4 | Species-specific differences in DNA methylation. (a) Position-wise DNA methylation frequencies. **(b)** DNA methylation frequencies in *A. alpina* separated by genomic regions. **(c)** Correlation of methylation frequency on Watson and Crick strand at symmetrical CG and CHG sites (Aa, *A. alpina*; At, *A. thaliana*). **(d)** Gene family sizes of DNA methylation genes. **(e)** – **(i)** Gene family phylogenies (Aa, *A. alpina*; Al, *A. lyrata*, At, *A. thaliana*; Br, *B. rapa*; Cp, *C. papaya*, Cr, *C. rubella*; Es, *E. salsugineum*, Sp, *S. parvula*). **(j)** d_N/d_S values for orthologous genes pairs between *A. alpina* and *A. thaliana* (light blue) and d_N/d_S values of each methylation gene family (coloured dots).

References

1. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796-815 (2000).

2. Hu, T.T. et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet* **43,** 476-81 (2011).

3. Wang, X. et al. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat Genet* **43,** 1035-9 (2011).

4. Dassanayake, M. et al. The genome of the extremophile crucifer *Thellungiella parvula*. *Nat Genet* **43,** 913-8 (2011).

5. Wu, H.-J. et al. Insights into salt tolerance from the genome of *Thellungiella salsuginea*. *Proc Natl Acad Sci U S A* **109,** 12219-24 (2012).

6. Yang, R. et al. The Reference Genome of the Halophytic Plant *Eutrema salsugineum*. *Front Plant Sci* **4,** 46 (2013).

7. Cheng, S. et al. The *Tarenaya hassleriana* genome provides insight into reproductive trait and genome evolution of crucifers. *Plant Cell* **25,** 2813-30 (2013).

8. Haudry, A. et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet* **45,** 891-8 (2013).

9. Slotte, T. et al. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat Genet* **45,** 831-5 (2013).

10. Liu, S. et al. The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nat Commun* **5,** 3930 (2014).

11. Tedder, A., Ansell, S.W., Lao, X., Vogel, J.C. & Mable, B.K. Sporophytic self-incompatibility genes and mating system variation in *Arabis alpina*. *Ann Bot* **108,** 699-713 (2011).

12. Wingler, A., Stangberg, E.J., Saxena, T. & Mistry, R. Interactions between temperature and sugars in the regulation of leaf senescence in the perennial herb *Arabis alpina* L. *J Integr Plant Biol* **54,** 595-605 (2012).

13. Wang, R. et al. PEP1 regulates perennial flowering in *Arabis alpina*. *Nature* **459,** 423-7 (2009).

14. Sanmiguel, P. & Bennetzen, J.L. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Annals of Botany* **82,** 37-44 (1998).

15. Maumus, F. & Quesneville, H. Ancestral repeats have shaped epigenome and genome composition for millions of years in *Arabidopsis thaliana*. *Nat Commun* **5,** 4104 (2014).

16. Ma, J. & Bennetzen, J.L. Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci U S A* **101,** 12404-10 (2004).

17. Zhang, X. et al. Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell* **126,** 1189-201 (2006).

18. Roudier, F. et al. Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *EMBO J* **30,** 1928-38 (2011).

19. Hollister, J.D., Smith, L.M., Guo, Y.-L., Ott, F., Weigel, D. & Gaut, B.S. Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Proc Natl Acad Sci U S A* **108,** 2322-7 (2011).

20. Hall, A.E., Kettler, G.C. & Preuss, D. Dynamic evolution at pericentromeres. *Genome Res* **16,** 355-64 (2006).

21. Wijnker, E. et al. The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. *Elife* **2,** e01426 (2013).

22. Kuittinen, H. et al. Comparing the linkage maps of the close relatives *Arabidopsis lyrata* and *A. thaliana*. *Genetics* **168,** 1575-84 (2004).

23. Zhang, X., Bernatavichute, Y.V., Cokus, S., Pellegrini, M. & Jacobsen, S.E. Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in *Arabidopsis thaliana*. *Genome Biol* **10,** R62 (2009).

24. Cokus, S.J. et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452,** 215-9 (2008).

25. Becker, C. et al. Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **480,** 245-9 (2011).

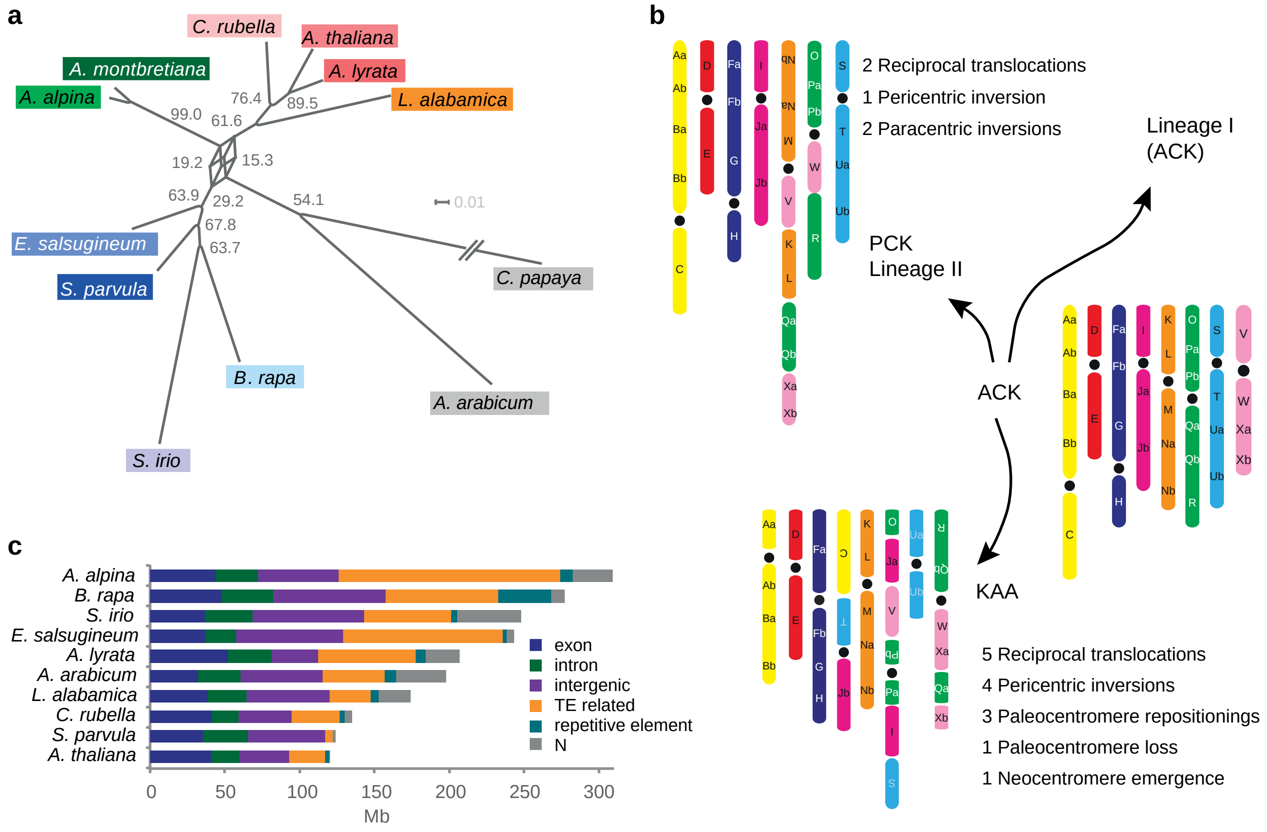
26. Law, J.A. & Jacobsen, S.E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* **11,** 204-20 (2010).

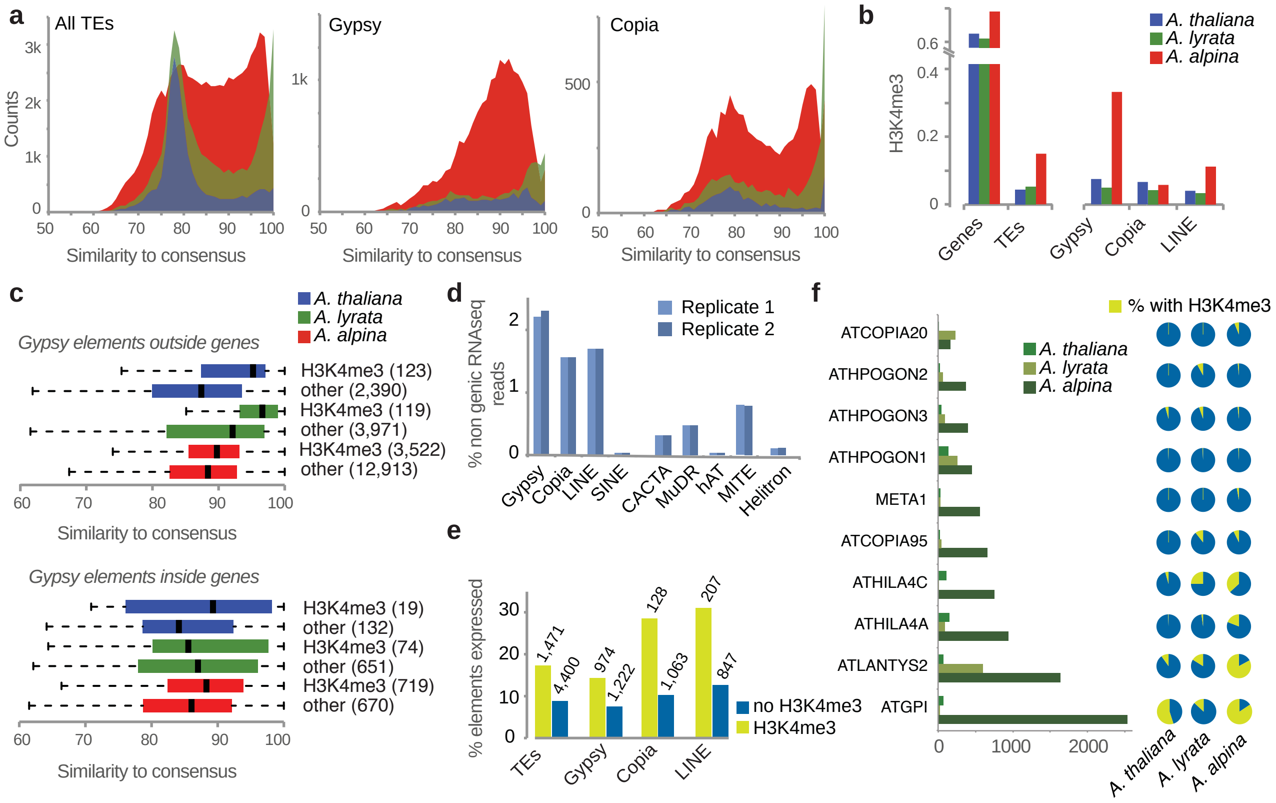
27. Zemach, A., McDaniel, I.E., Silva, P. & Zilberman, D. Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation. *Science* **328,** 916-9 (2010).

28. Matzke, M., Kanno, T., Daxinger, L., Huettel, B. & Matzke, A.J.M. RNA-mediated chromatin-based silencing in plants. *Curr Opin Cell Biol* **21,** 367-76 (2009).

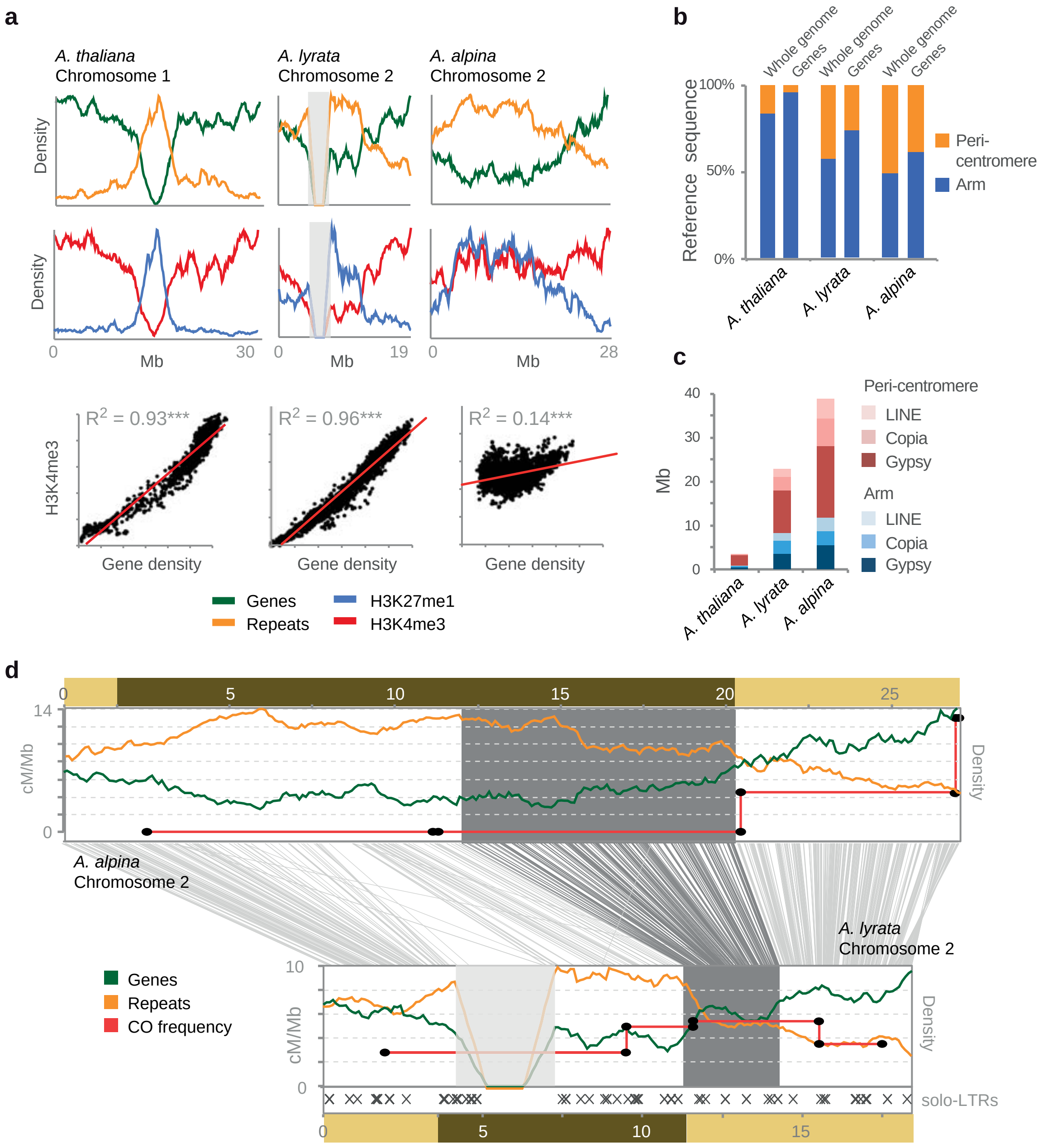
29. Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A. & Kakutani, T. Bursts of retrotransposition reproduced in Arabidopsis. *Nature* **461,** 423-6 (2009).

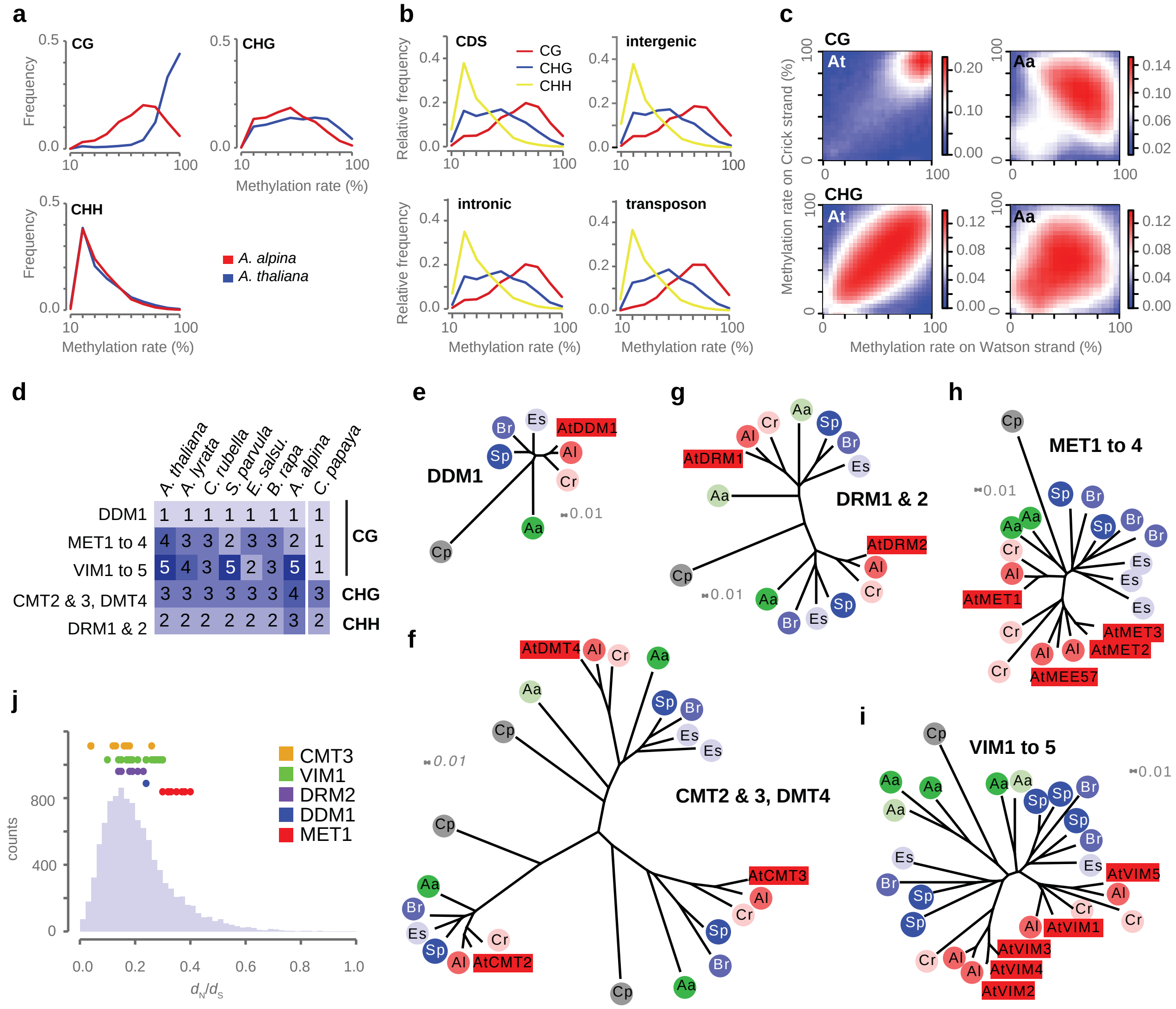
30. Mirouze, M. et al. Selective epigenetic control of retrotransposition in Arabidopsis. *Nature* **461,** 427-30 (2009).


a
C. rubella
A. thaliana
A. lyrata
A. montbretiana
A. alpina
L. alabamica
76.4
99.0
61.6
89.5
19.2
15.3
63.9
29.2
54.1
0.01
67.8
63.7
E. salsugineum
S. parvula
C. papaya
B. rapa
A. arabicum
S. irio
b
2 Reciprocal translocations
1 Pericentric inversion
2 Paracentric inversions
Lineage I
(ACK)
PCK
Lineage II
ACK
KAA
5 Reciprocal translocations
4 Pericentric inversions
3 Paleocentromere repositionings
1 Paleocentromere loss
1 Neocentromere emergence
Aa
Ab
Ba
Bb
C
D
E
Fa
Fb
G
H
I
Ja
Jb
K
L
M
Na
Nb
O
Pa
Pb
Qa
Qb
R
S
T
Ua
Ub
V
W
Xa
Xb
c
A. alpina
B. rapa
S. irio
E. salsugineum
A. lyrata
A. arabicum
L. alabamica
C. rubella
S. parvula
A. thaliana
exon
intron
intergenic
TE related
repetitive element
N
0
50
100
150
200
250
300
Mb



a
All TEs
Counts
3k
2k
1k
0
50 60 70 80 90 100
Similarity to consensus
Gypsy
1k
0
50 60 70 80 90 100
Similarity to consensus
Copia
500
0
50 60 70 80 90 100
Similarity to consensus
b
H3K4me3
0.6
0.4
0.2
0
Genes
TEs
Gypsy
Copia
LINE
A. thaliana
A. lyrata
A. alpina
c
A. thaliana
A. lyrata
A. alpina
Gypsy elements outside genes
H3K4me3 (123)
other (2,390)
H3K4me3 (119)
other (3,971)
H3K4me3 (3,522)
other (12,913)
60 70 80 90 100
Similarity to consensus
Gypsy elements inside genes
H3K4me3 (19)
other (132)
H3K4me3 (74)
other (651)
H3K4me3 (719)
other (670)
60 70 80 90 100
Similarity to consensus
d
% non genic RNAseq reads
2
1
0
Replicate 1
Replicate 2
Gypsy
Copia
LINE
SINE
CACTA
MuDR
hAT
MITE
Helitron
e
% elements expressed
30
20
10
0
1,471
4,400
974
1,222
128
1,063
207
847
no H3K4me3
H3K4me3
TEs
Gypsy
Copia
LINE
f
% with H3K4me3
A. thaliana
A. lyrata
A. alpina
ATCOPIA20
ATHPOGON2
ATHPOGON3
ATHPOGON1
META1
ATCOPIA95
ATHILA4C
ATHILA4A
ATLANTYS2
ATGPI
0
1000
2000
A. thaliana
A. lyrata
A. alpina


a
A. thaliana
Chromosome 1
A. lyrata
Chromosome 2
A. alpina
Chromosome 2
Density
Density
0
Mb
30
0
Mb
19
0
Mb
28
H3K4me3
R^2 = 0.93***
R^2 = 0.96***
R^2 = 0.14***
Gene density
Gene density
Gene density
Genes
Repeats
H3K27me1
H3K4me3
b
Whole genome
Genes
Whole genome
Genes
Whole genome
Genes
Reference sequence
100%
50%
0%
Peri-centromere
Arm
A. thaliana
A. lyrata
A. alpina
c
Mb
40
30
20
10
0
Peri-centromere
LINE
Copia
Gypsy
Arm
LINE
Copia
Gypsy
A. thaliana
A. lyrata
A. alpina
d
0
5
10
15
20
25
14
0
cM/Mb
Density
A. alpina
Chromosome 2
A. lyrata
Chromosome 2
Genes
Repeats
CO frequency
10
0
cM/Mb
Density
solo-LTRs
0
5
10
15



	A. thaliana	A. lyrata	C. rubella	S. parvula	E. salsu.	B. rapa	A. alpina	C. papaya	
DDM1	1	1	1	1	1	1	1	1	CG
MET1 to 4	4	3	3	2	3	3	2	1	
VIM1 to 5	5	4	3	5	2	3	5	1	
CMT2 & 3, DMT4	3	3	3	3	3	3	4	3	CHG
DRM1 & 2	2	2	2	2	2	2	3	2	CHH