LBNL Title Page

The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change

Tina T. Hu[1¶]*, Pedro Pattyn[2]*, Erica G. Bakker[3,4¶], Jun Cao[5], Jan-Fang Cheng[6], Richard M. Clark[5¶], Noah Fahlgren[7], Jeffrey A. Fawcett[2], Jane Grimwood[6,8], Heidrun Gundlach[9], Georg Haberer[9], Jesse D. Hollister[10¶], Stephan Ossowski[5], Robert P. Ottilar[6], Asaf A. Salamov[6], Korbinian Schneeberger[5], Manuel Spannagl[9], Xi Wang[9], Liang Yang[10], Mikhail E. Nasrallah[11], Joy Bergelson[3], James C. Carrington[7], Brandon S. Gaut[10], Jeremy Schmutz[6,8], Klaus F. X. Mayer[9], Yves Van de Peer[2], Igor V. Grigoriev[6], Magnus Nordborg[1,12], Detlef Weigel[5§] and Ya-Long Guo[5§]

[1]Molecular and Computational Biology, University of Southern California, Los Angeles, California 90089, USA. [2]Department of Plant Systems Biology, Ghent University, VIB, 9052 Ghent, Belgium. [3]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA. [4]Center for Genome Research and Biocomputing, and Department of Horticulture, Oregon State University, Corvallis, Oregon 97331, USA. [5]Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany. [6]US Department of Energy Joint Genome Institute, Walnut Creek, California 94598, USA. [7]Center for Genome Research and Biocomputing, and Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon 97331, USA. [8]Hudson Alpha Genome Sequencing Center, Hudson Alpha Institute for Biotechnology, Huntsville, Alabama 35806, USA. [9]MIPS/IBIS Institute for Bioinformatics and Systems Biology, Helmholtz Center Munich, 85764 Neuherberg, Germany. [10]Department of Ecology and Evolutionary Biology, University of California Irvine, Irvine, California 92697, USA. [11]Department of Plant Biology, Cornell University, Ithaca, New York 14853, USA. [12]Gregor Mendel Institute, 1030 Vienna, Austria

**To whom correspondence may be addressed. E-mail:* To weigel@weigelworld.org (D.W.); yalong_guo@hotmail.com (Y.-L.G.)

11 May 2011

ACKNOWLEDGMENTS:

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

DISCLAIMER:

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change

Tina T. Hu[1¶]*, Pedro Pattyn[2]*, Erica G. Bakker[3,4¶], Jun Cao[5], Jan-Fang Cheng[6], Richard M. Clark[5¶], Noah Fahlgren[7], Jeffrey A. Fawcett[2], Jane Grimwood[6,8], Heidrun Gundlach[9], Georg Haberer[9], Jesse D. Hollister[10¶], Stephan Ossowski[5], Robert P. Ottilar[6], Asaf A. Salamov[6], Korbinian Schneeberger[5], Manuel Spannagl[9], Xi Wang[9], Liang Yang[10], Mikhail E. Nasrallah[11], Joy Bergelson[3], James C. Carrington[7], Brandon S. Gaut[10], Jeremy Schmutz[6,8], Klaus F. X. Mayer[9], Yves Van de Peer[2], Igor V. Grigoriev[6], Magnus Nordborg[1,12], Detlef Weigel[5§] and Ya-Long Guo[5§]

[1]Molecular and Computational Biology, University of Southern California, Los Angeles, California 90089, USA. [2]Department of Plant Systems Biology, Ghent University, VIB, 9052 Ghent, Belgium. [3]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA. [4]Center for Genome Research and Biocomputing, and Department of Horticulture, Oregon State University, Corvallis, Oregon 97331, USA. [5]Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany. [6]US Department of Energy Joint Genome Institute, Walnut Creek, California 94598, USA. [7]Center for Genome Research and Biocomputing, and Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon 97331, USA. [8]Hudson Alpha Genome Sequencing Center, Hudson Alpha Institute for Biotechnology, Huntsville, Alabama 35806, USA. [9]MIPS/IBIS Institute for Bioinformatics and Systems Biology, Helmholtz Center Munich, 85764 Neuherberg, Germany. [10]Department of Ecology and Evolutionary Biology, University of California Irvine, Irvine, California 92697, USA. [11]Department of Plant Biology, Cornell University, Ithaca, New York 14853, USA. [12]Gregor Mendel Institute, 1030 Vienna, Austria.

[§]To whom correspondence should be addressed. E-mail: weigel@weigelworld.org (D.W.); yalong_guo@hotmail.com (Y.-L.G.).

*These authors contributed equally to this work.

¶Present addresses: Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544, USA (T.T.H.); Dow AgroSciences, Portland, Oregon 97224, USA (E.G.B.); Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA (J.D.H.).

We present the 207 Mb genome sequence of the outcrosser *Arabidopsis lyrata*, which diverged from the self-fertilizing species *A. thaliana* about 10 million years ago. It is generally assumed that the much smaller *A. thaliana* genome, which is only 125 Mb, constitutes the derived state for the family. Apparent genome reduction in this genus can be partially attributed to the loss of DNA from large-scale rearrangements, but the main cause lies in the hundreds of thousands of small deletions found throughout the genome. These occurred primarily in non-coding DNA and transposons, but protein-coding multi-gene families are smaller in *A. thaliana* as well. Analysis of deletions and insertions still segregating in *A. thaliana* indicates that the process of DNA loss is ongoing, suggesting pervasive selection for a smaller genome.

The size of angiosperm genomes ranges from merely 64 Mb in the corkscrew plant *Genlisea*[1] to an enormous 124 Gb in *Fritillaria* lilies[2,3]. Two major processes contribute to an increase in genome size: polyploidization and the proliferation of transposable elements (TEs). Processes that counteract genome expansion have received less attention, but include the loss of entire chromosomes as well as deletion-biased mutations due to unequal homologous recombination and illegitimate recombination[4-6]. Recent work comparing two cereal species, rice and sorghum, has begun to shed light on the processes behind genome expansion and contraction on the whole-genome level[7]. However, these species are separated by 60 to 70 million years, making it difficult to disentangle the different evolutionary forces at work.

An exciting opportunity to understand what drives differences in genome size over shorter time scales is offered by the genus *Arabidopsis* in the Brassicaceae family. The genome of the self-incompatible perennial *A. lyrata* is larger than 200 Mb, which is near the average for the family[8,9], while the self-compatible annual *A. thaliana* has one of the smallest angiosperm genomes, at about 125 Mb, even though the two species diverged only about 10 million years ago[10-12].

RESULTS

Genome assembly and annotation

A high-quality genome sequence for the partially inbred *A. lyrata* strain MN47 was assembled from approximately 8.3x coverage of paired-end dideoxy sequencing reads using libraries with different insert sizes, making use of information from genetic maps and chromosome painting[13-16] (Supplementary Information Online). The final assembly included 206.7 Mb of sequence, 90% of which are included in eight large scaffolds covering the majority of each of the eight chromosomes, and another large scaffold of 1.9 Mb representing one of the centromeres. Based on cytological observations[17], the centromeric gaps were estimated to span 17.2 Mb. A combination of de novo predictions, homology to *A. thaliana* genomic and expressed sequences, and short-read sequencing of polyA$^+$-RNA and small RNAs was used to annotate the genome . In *A. lyrata*, we predicted 32,670 protein-coding genes, compared to 27,025 genes in *A. thaliana*[18].

Synteny and rearrangements

Since overall sequence identity between *A. lyrata* and *A. thaliana* is greater than 80% (Supplementary Fig. 1a-c), the two genomes could be easily aligned to each other (Fig. 1a). Genetic mapping[13,15,16] has revealed 10 major rearrangements, including two reciprocal translocations and three chromosomal fusions, that led to the *A. thaliana* karyotype of five chromosomes compared to the ancestral state of eight, as found in *A. lyrata* and other Brassicaceae. Although centromeric regions are difficult to assemble, we could identify the syntenic region in *A. thaliana* that corresponds to the chromosome 4 centromere of *A. lyrata*. The entire centromere has been lost, with only two remnants of satellite repeats in the 1.4 kb intergenic region between the two protein coding genes *At2g26570* and *At2g26580* (Supplementary Fig. 2).

Apart from chromosomal-scale changes, approximately 90% of the *A. thaliana* and *A. lyrata* genomes have remained syntenic, with the great majority in highly conserved colinear arrangements (Fig. 1b and Supplementary Fig. 1d). The run length distribution of colinear gene pairs is bimodal, with a first peak of fragments of five or fewer colinear gene pairs (Fig. 1c), reflecting an abundance of small-scale rearrangements (<10 kb), including single gene transpositions. Windows containing a breakpoint in colinearity are enriched for TEs and other repeats (Supplementary Table 1), in agreement with repetitive elements often being associated with chromosomal rearrangements and transposed genes[19-24], although they might not necessarily be causal[25]. Two thirds of the 154 inversions identified between the two species are flanked by inverted repeats.

Sequence content

Despite this overall similarity in gene arrangement, the two genomes are strikingly different in size. A whole-genome alignment reveals that more than 50% (~114 Mb) of the *A. lyrata* genome appears to be missing from the *A. thaliana* reference genome. In contrast, only about 25% (~30 Mb) of the *A. thaliana* genome is absent from *A. lyrata* (Fig. 1d and Supplementary Table 2). Nevertheless, the distribution across different sequence classes is very similar in both cases, with about half of unalignable sequences being in TEs, and about a quarter in intergenic regions. The net effect of these changes is that the *A. thaliana* genome is ~80 Mb smaller than the *A. lyrata* genome, with a much higher fraction of genic sequences, 42% instead of 29%, even though the total number of genes is smaller (Fig. 1e). The apparent shrinkage of the *A. thaliana* genome is not simply due to a few chromosomal level changes: only about 10% of the difference in genome size may be attributed to the three missing centromeres; the rest is due to hundreds of thousands of smaller insertions and deletions, spanning all classes of sites. Strikingly, while large differences much more often correspond to sequences only found in *A. lyrata*, this is not true for very small insertions and deletions (Fig. 2). This is in stark contrast to other genomes from other closely related species, but with similarly sized genomes, such as chimpanzee and human[26].

Although rearrangements are correlated with genome shrinkage (rearranged regions are on average shorter in *A. thaliana* relative to *A. lyrata* than are colinear regions; Fig. 3 and Fig. 4a), unalignable sequences are found throughout the genome. An analysis of colinear gene pairs confirmed that in most cases, intergenic regions in *A. lyrata* are longer than their counterparts in *A. thaliana* (Fig. 4b). The same is true for introns, although the difference is smaller[11] (tables S3 and S4).

The gene content of *A. thaliana* is modestly lower than that of *A. lyrata*, without major differences in Gene Ontology (GO) distribution (Supplementary Table 5 and Supplementary Fig. 3). Similarly, divergence patterns for different gene families between the two species mirror those of within-*Arabidopsis thaliana* polymorphism levels[27,28] (Supplementary Fig. 4). The combined gene sets of *A. lyrata* and *A. thaliana* result in 12,951 MCL (ref. 29) clusters, with fewer singletons in *A. thaliana* (Fig. 4c). Among the 8,794 shared multi-gene MCL clusters, there is a tendency to smaller size in *A. thaliana* (Fig. 2d), with clusters that are smaller in *A. thaliana* outnumbering those that are smaller in *A. lyrata* (1,797 to 612). We examined F-box and NB-LRR genes as two examples of highly variable gene families with particularly high birth and death rates in plants[27,28,30-32]. *A. lyrata* has 596 F-box and 187 NB-LRR genes, compared to 502 and 159, respectively, in *A. thaliana* (Supplementary Fig. 5). The trend of fewer genes in *A. thaliana* is supported by a broader comparison of the *Arabidopsis* gene set with those of two other dicots, poplar and grapevine[33-35]. *Arabidopsis lyrata* has 114 ortholog clusters[36] shared with poplar and grapevine but not *A. thaliana*, while *A. thaliana* has only 45 clusters found in poplar and grapevine but not *A. lyrata*. Similarly, *A. lyrata* has 875 clusters not detected in any of the other three species, while *A. thaliana* has only 156 species-specific clusters (Supplementary Table 6 and Supplementary Fig. 6).

As in other taxa, TEs make an important contribution to the change in genome size (Fig. 1d). TEs comprise a larger fraction of the *A. lyrata* genome, compared to the smaller *A. thaliana* genome (Fig. 1e), and much of this difference is apparently due to the higher and/or more recent activity of TEs in *A. lyrata*. Without an outgroup, one cannot infer directly how much differential purging of ancestral TEs or different TE activity levels since speciation have shaped these patterns, but one can exploit the molecular clock to estimate the average age of long terminal repeat (LTR) retrotransposons[37] (Fig. 1e). Using the experimentally determined mutation rate in *A. thaliana*[12], we calculated the mean and median age in *A. thaliana* to be 3.1 and 2.1 million years, respectively, compared to a more recent age of 1.1 and 0.6 million years in *A. lyrata* (Fig. 5a), which is in agreement with previous estimates[38]. Assuming that LTR retrotransposons have been removed at constant rates in each species[39,40], we furthermore find that the deduced half-life of LTR retrotransposons in *A. thaliana* is greater than in *A. lyrata*, at 2.0 compared to 0.6 million years. Thus, it does not appear that *A. thaliana* eliminates LTR retrotransposons more efficiently, but rather that their activity is reduced [6]. A phylogenetic analysis also supports a greater expansion of specific LTR retrotransposon clades in *A. lyrata* (Fig. 5b).

Furthermore, TEs are differently distributed in the two species, with *A. lyrata* having a higher proportion of genes with a TE nearby than *A. thaliana* (Fig. 5c), and this distance is skewed towards larger values in *A. thaliana* (Supplementary Table 7 and Supplementary Fig. 7). Together, these observations are consistent with a model under which selection purges TEs with deleterious effects on adjacent genes, such that TEs more distant from genes preferentially survive[41], with TE elimination having been more efficient in *A. thaliana*.

Evidence for ongoing genome shrinkage in *A. thaliana*

The evidence presented so far points to *A. thaliana* having evolved a smaller genome through a large number of deletions throughout its genome. We can use within-species polymorphisms to shed light on the process by which this has happened. If the *A. thaliana* genome continues to shrink, we would expect fewer segregating insertions than deletions. Using the *A. lyrata* genome to determine the derived state among a set of insertion and deletion polymorphisms found throughout the genome of 95 *A. thaliana* individuals[42], we find 2,685 fixed and 852 segregating deletions, compared to 1,941 fixed and 106 segregating insertions, a clear excess of deletions over insertions. Furthermore, among the fixed differences, deletions are on average longer than insertions (Fig. 6a). If no selection were involved, and if this pattern were only due to mutational bias favoring deletions[43,44], deletion and insertion polymorphisms should have similar allele frequencies in the *A. thaliana* population. However, segregating insertions are on average found in fewer individuals than are deletions or single-nucleotide polymorphisms. Deletions are often found in the majority of individuals, and many are approaching fixation in *A. thaliana* (Fig. 6b). This pattern suggests that deletions are favored over insertions because of selection, rather than simple mutational bias, thus leading to a smaller genome. Insertions appear to be under both purifying and positive selection.

The pattern of divergence between the two genomes supports this hypothesis. While more deletions have occurred on the *A. thaliana* than on the *A. lyrata* lineage, the bias towards deletions in *A. thaliana* becomes stronger the longer the missing sequence, and it is absent for sequences shorter than 5 bp or so (Fig. 2). This is consistent with a model where long deletions are selectively favored in *A. thaliana*, whereas short deletions are not.

DISCUSSION

We have presented a high-quality reference genome sequence for *A. lyrata*, which will be a valuable resource for functional, evolutionary and ecological studies in the genus *Arabidopsis*. We found that several processes contribute to the difference in genome size between the predominantly selfing *A. thaliana* and the outcrossing *A. lyrata*. In just a few million generations, numerous chromosomal rearrangements have occurred, consistent with theoretical predictions of rearrangements that reduce fitness in heterozygotes being fixed much more easily in strongly selfing species[45]. Much of the genome size difference seems to be due to reduced TE activity and/or more efficient TE elimination in *A. thaliana*, especially near genes, but both non-TE intergenic sequences and introns are shorter in *A. thaliana* as well. Specifically, we find that segregating deletions at non-coding sites in *A. thaliana* are skewed towards higher allele frequencies, and that both fixed and polymorphic deletions are more common than insertions. Together, this suggests pervasive selection for a smaller genome in *A. thaliana*. Apart from apparent advantages for species with smaller genomes that have been inferred from meta analyses[46], the transition to selfing might be an important factor in this process[45].

What role, if any, genome expansion might play in *A. lyrata* can be addressed once detailed *A. lyrata* polymorphism information as well as closely related outgroup genomes become available, such as the one from *Capsella rubella*, which is currently being assembled. A complete understanding of the processes behind genome contraction and expansion over short time scales will in addition require better knowledge of mutational events, and a deeper understanding of the distribution of, and selection on, non-coding regulatory sequences[41]. For both, high-quality whole genome sequences of additional *Arabidopsis* relatives will be an important tool.

METHODS

Complete methods are available as Supplementary Information Online.

Sequencing and assembly. *Arabidopsis lyrata* strain MN47 was derived by forced selfing from material collected in Michigan, USA, by Dr. Charles Langley (UC Davis). It was inbred six times before extracting DNA for sequencing. Libraries with various insert sizes including fosmids and BACs were dideoxy sequenced on ABI 3730XL capillary sequencers. Reads were assembled with Arachne[47], and colinearity information was integrated with marker information from genetic maps[13,15,16] to reconstruct the eight linkage groups.

Annotation. The genome was annotated using *ab initio* and homology-based gene predictors along with RNA-seq data (Supplementary Information Online).

URLs. The assembly and annotation is available from JGI (http://www.phytozome.net/alyrata.php).

Accession numbers. The annotated genome assembly has been deposited with GenBank under accession number ADBK00000000. Seeds of the MN47 strain have been deposited with the Arabidopsis Biological Resource Center under accession number CS22696.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

The U.S. Department of Energy Joint Genome Institute (JGI) provided sequencing and analyses under the Community Sequencing Program supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We are particularly grateful to Dan Rokhsar and Kerrie Barry for providing leadership for the project at JGI. We thank Justin Borevitz, Anne Hall, Charles Langley, June Nasrallah, Barbara Neuffer, Outi Savolainen and Stephen Wright for contributing to the initial sequencing proposal submitted to the Community Sequencing Program at JGI, Christa Lanz and Kenneth Lett for technical assistance, and Peter Andolfatto and Rod Wing for comments on the manuscript. This work was supported by NSF DEB-0723860 (B.S.G.), NSF DEB-0723935 (M.N.), NSF MCB-0618433 (J.C.C.), NSF IOS-0744579 (M.E.N.), NIH GM057994 (J.B.), grant GABI-DUPLO 0315055 of the German Federal Ministry of Education and Research (K.F.X.M.), ERA-PG grant ARelatives from the Deutsche Forschungsgemeinschaft (D.W.) and Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT) and the Inter-University Network for Fundamental Research (P6/25, BioMaGNet) (Y.V.d.P.), a Gottfried Wilhelm Leibniz Award of the DFG (D.W.), the Austria Academy of Sciences (M.N.), and the Max Planck Society (D.W. and Y.-L.G.).

AUTHOR CONTRIBUTIONS

J.B., J.C.C., B.S.G., I.V.G., Y.-L.G., K.F.X.M., M.N., Y.V.d.P. and D.W. conceived the study; M.E.N. provided the biological material; J.C., J.-F.C., R.M.C., N.F., J.G. and Y.-L.G. performed the experiments; E.G.B., J.A.F., N.F., H.G., Y.-L.G., G.H., J.D.H., T.T.H., R.P.O., S.O., P.P., A.S., J.S., K.S., M.S., X.W., and L.Y. analyzed the data; and Y.-L.G., T.T.H., M.N. and D.W. wrote the paper with contributions from all authors.

COMPETING INTEREST STATEMENT

The authors declare no competing financial interests.

1. Greilhuber, J. et al. Smallest angiosperm genomes found in Lentibulariaceae, with chromosomes of bacterial size. *Plant Biol.* **8**, 770-7 (2006).
2. Gregory, T.R. et al. Eukaryotic genome size databases. *Nucleic Acids Res.* **35**, D332-8 (2007).
3. Gaut, B.S. & Ross-Ibarra, J. Selection on major components of angiosperm genomes. *Science* **320**, 484-6 (2008).
4. Bennetzen, J.L., Ma, J. & Devos, K.M. Mechanisms of recent genome size variation in flowering plants. *Ann. Bot.* **95**, 127-32 (2005).
5. Piegu, B. et al. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* **16**, 1262-9 (2006).
6. Vitte, C., Panaud, O. & Quesneville, H. LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. *BMC Genomics* **8**, 218 (2007).
7. Paterson, A.H. et al. The *Sorghum bicolor* genome and the diversification of grasses.

Nature **457**, 551-6 (2009).

8. Johnston, J.S. et al. Evolution of genome size in Brassicaceae. *Ann. Bot.* **95**, 229-35 (2005).

9. Oyama, R.K. et al. The shrunken genome of *Arabidopsis thaliana*. *Plant Systemat. Evol.* **273**, 257-271 (2008).

10. Koch, M.A., Haubold, B. & Mitchell-Olds, T. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Mol. Biol. Evol.* **17**, 1483-1498 (2000).

11. Wright, S.I., Lauga, B. & Charlesworth, D. Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. *Mol. Biol. Evol.* **19**, 1407-20 (2002).

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

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

14. Koch, M.A. & Kiefer, M. Genome evolution among cruciferous plants: a lecture from the comparison of the genetic maps of three diplod species–*Capsella rubella*, *Arabidopsis lyrata* subsp. *petraea*, and *A. thaliana*. *Am. J. Bot.* **92**, 761-767 (2005).

15. Yogeeswaran, K. et al. Comparative genome analyses of *Arabidopsis* spp.: inferring chromosomal rearrangement events in the evolutionary history of *A. thaliana*. *Genome Res.* **15**, 505-15 (2005).

16. Lysak, M.A. et al. Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. *Proc. Natl. Acad. Sci. USA* **103**, 5224-9 (2006).

17. Berr, A. et al. Chromosome arrangement and nuclear architecture but not centromeric sequences are conserved between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Plant*

J. **48**, 771-83 (2006).

18. Swarbreck, D. et al. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* (2007).

19. Lim, J.K. & Simmons, M.J. Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. *Bioessays* **16**, 269-75 (1994).

20. Stankiewicz, P. et al. Genome architecture catalyzes nonrecurrent chromosomal rearrangements. *Am. J. Hum. Genet.* **72**, 1101-16 (2003).

21. Korbel, J.O. et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420-6 (2007).

22. Lee, J., Han, K., Meyer, T.J., Kim, H.S. & Batzer, M.A. Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. *PLoS ONE* **3**, e4047 (2008).

23. Braumann, I., van den Berg, M.A. & Kempken, F. Strain-specific retrotransposon-mediated recombination in commercially used *Aspergillus niger* strain. *Mol. Genet. Genomics* **280**, 319-25 (2008).

24. Woodhouse, M.R., Pedersen, B. & Freeling, M. Transposed genes in *Arabidopsis* are often associated with flanking repeats. *PLoS Genet.* **6**, e1000949.

25. Ranz, J.M. et al. Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS Biol.* **5**, e152 (2007).

26. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87 (2005).

27. Clark, R.M. et al. Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317**, 338-42 (2007).

28. Borevitz, J.O. et al. Genome-wide patterns of single-feature polymorphism in

Arabidopsis thaliana. *Proc. Natl. Acad. Sci. USA* **104**, 12057-62 (2007).

29. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575-84 (2002).

30. Michelmore, R.W. & Meyers, B.C. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.* **8**, 1113-30 (1998).

31. Thomas, J.H. Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants. *Genome Res.* **16**, 1017-30 (2006).

32. Yang, X. et al. The F-box gene family is expanded in herbaceous annual plants relative to woody perennial plants. *Plant Physiol.* **148**, 1189-200 (2008).

33. Tuskan, G.A. et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596-604 (2006).

34. Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463-7 (2007).

35. Velasco, R. et al. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* **2**, e1326 (2007).

36. Li, L., Stoeckert, C.J., Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178-89 (2003).

37. SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. & Bennetzen, J.L. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43-5 (1998).

38. Devos, K.M., Brown, J.K. & Bennetzen, J.L. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.* **12**, 1075-9. (2002).

39. Pereira, V. Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. *Genome Biol.* **5**, R79 (2004).

40. Wicker, T. et al. A unified classification system for eukaryotic transposable elements.

Nat. Rev. Genet. **8**, 973-982 (2007).

41. Hollister, J.D. & Gaut, B.S. Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.* **19**, 1419-28 (2009).

42. Nordborg, M. et al. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3**, e196 (2005).

43. Petrov, D.A., Sangster, T.A., Johnston, J.S., Hartl, D.L. & Shaw, K.L. Evidence for DNA loss as a determinant of genome size. *Science* **287**, 1060-2 (2000).

44. Petrov, D.A., Lozovskaya, E.R. & Hartl, D.L. High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384**, 346-9 (1996).

45. Charlesworth, B. Evolutionary rates in partially self-fertilizing species. *Am. Nat.* **140**, 126-48 (1992).

46. Knight, C.A., Molinari, N.A. & Petrov, D.A. The large genome constraint hypothesis: evolution, ecology and phenotype. *Ann Bot.* **95**, 177-90 (2005).

47. Jaffe, D.B. et al. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91-6 (2003).

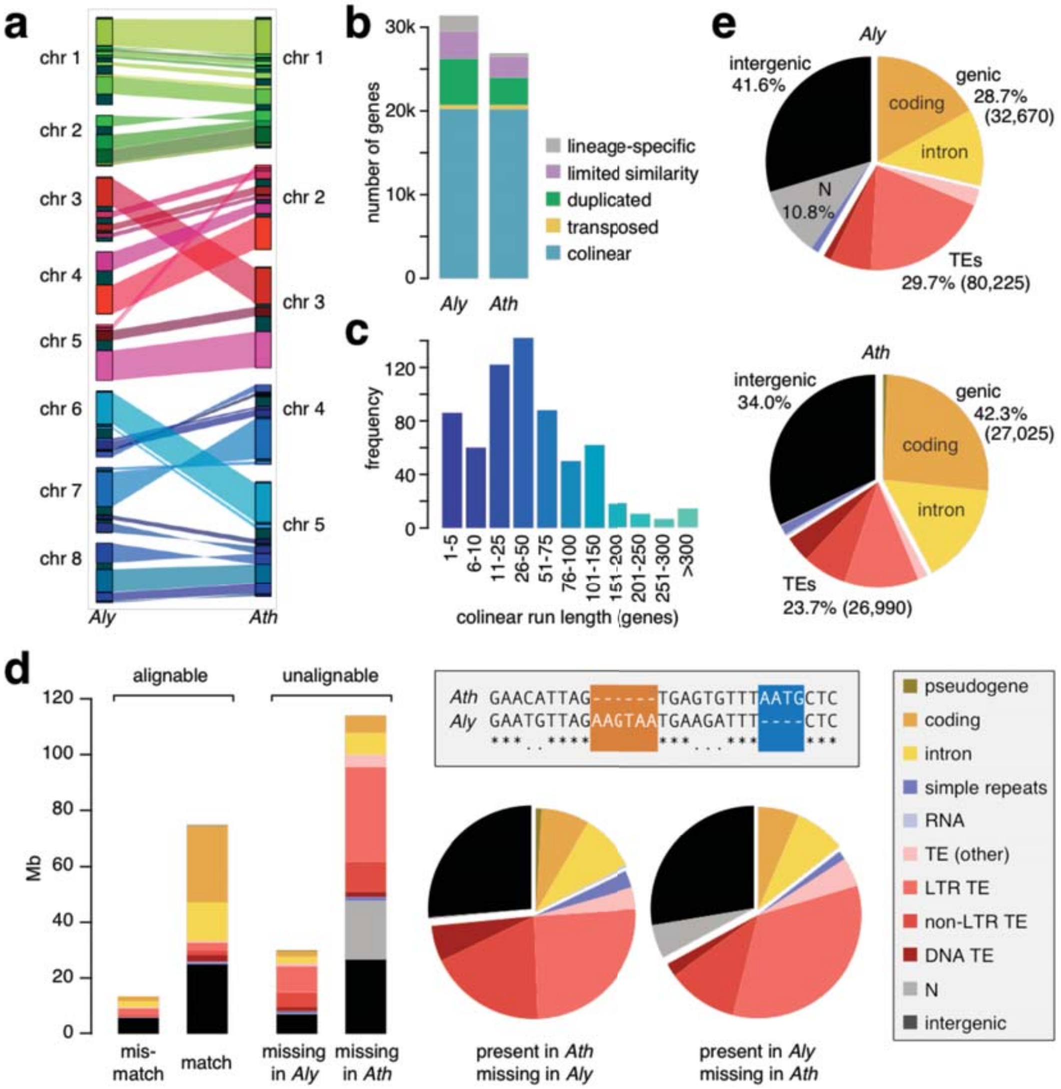


Figure 1 Comparison of *A. lyrata* and *A. thaliana* genomes. (**a**) Alignment of *A. lyrata* (*Aly*) and *A. thaliana* (*Ath*) chromosomes. Genomes are scaled to equal size. Only syntenic blocks of at least 500 kb are connected. (**b**) Orthology classification of genes. (**c**) Distribution of run lengths of colinear genes. The mode at 1-5 reflects frequent single-gene transpositions. (**d**) Unalignable sites can be considered as present in one species and absent in the other, as shown in the boxed sequence diagram; matches are indicated by asterisks, and mismatches by periods. The histogram on the left indicates the absolute number of unalignable sites, and the pie charts in the middle compare their relative distribution over different genomic features. See also Supplementary Table 2. (**e**) Genome composition (number of elements in parentheses).

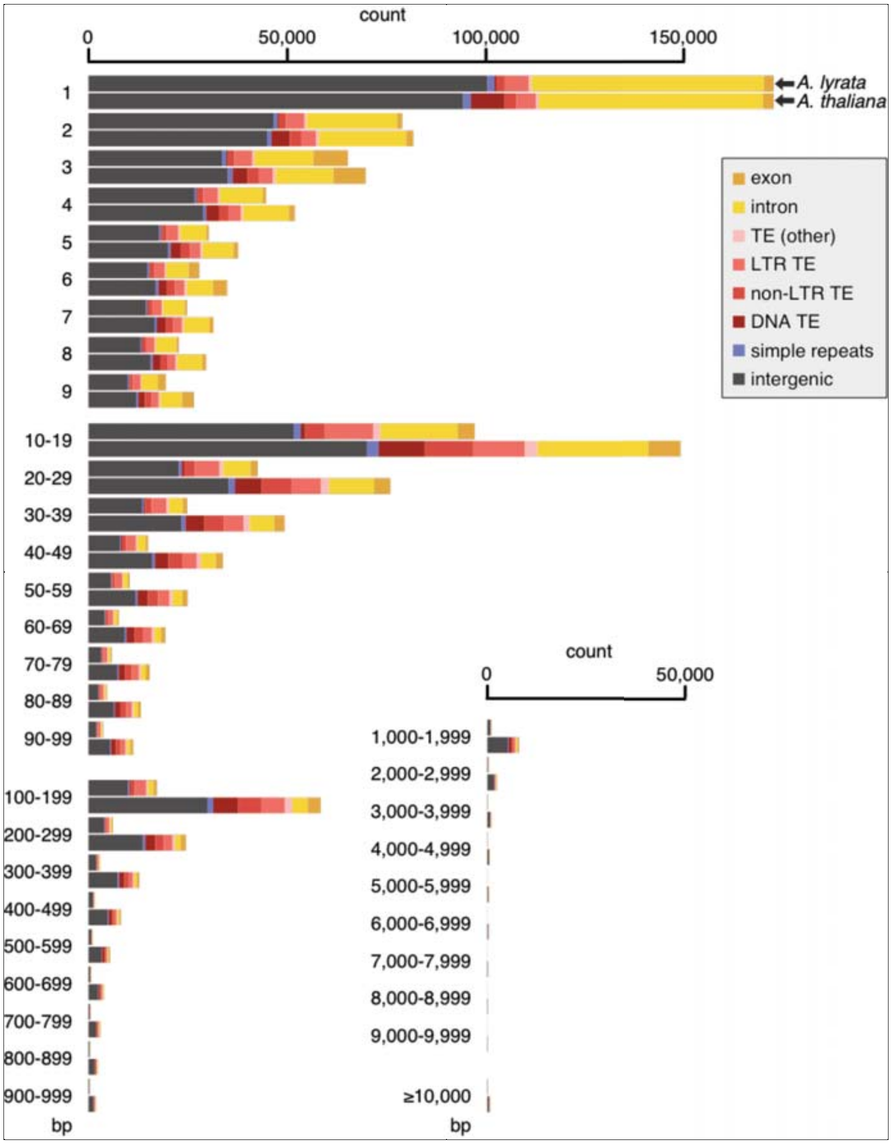


Figure 2 Apparent deletions by size and annotation. *A. lyrata* is always shown on top, *A. thaliana* on bottom.

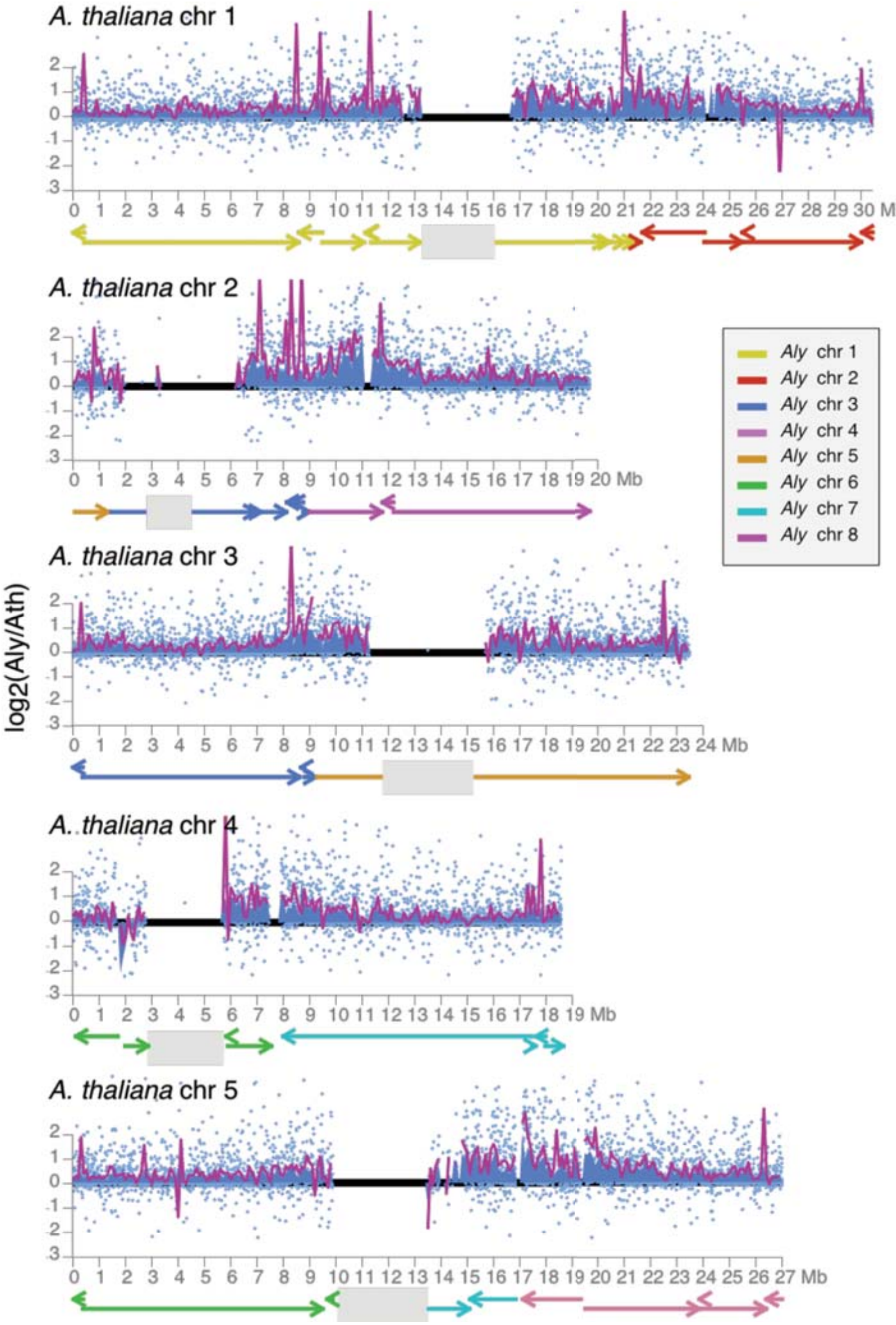


Figure 3 Changes in genomic intervals along the *A. thaliana* genome. Mean ratios for all colinear gene pairs in each 100 kb window are shaded in blue, with individual values shown as light blue dots. The ratio of the absolute length of each non-overlapping 100 kb window is shown as a dark purple line. Centromeres are indicated as grey boxes.

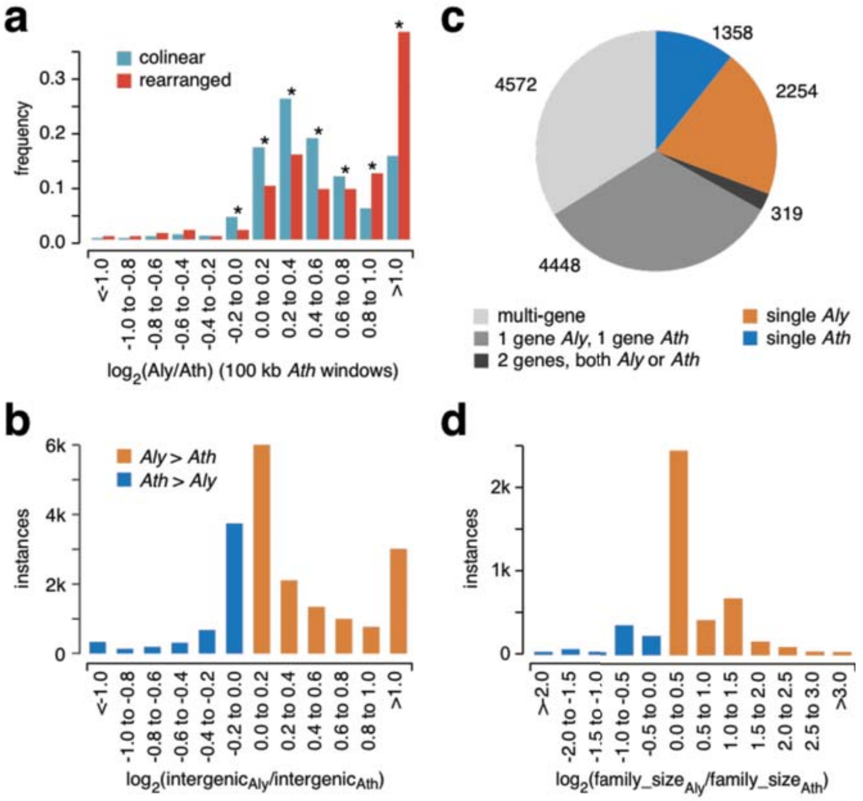


Figure 4 Change in size of colinear and rearranged regions, intergenic regions and gene families. (**a**) Size comparison of colinear regions, relative to 100 kb windows in *A. thaliana*. Asterisks indicate significant differences (binomial test, p<0.001). (**b**) Relative size of intergenic regions. (**c**) MCL clusters. (**d**) Relative size of gene families.

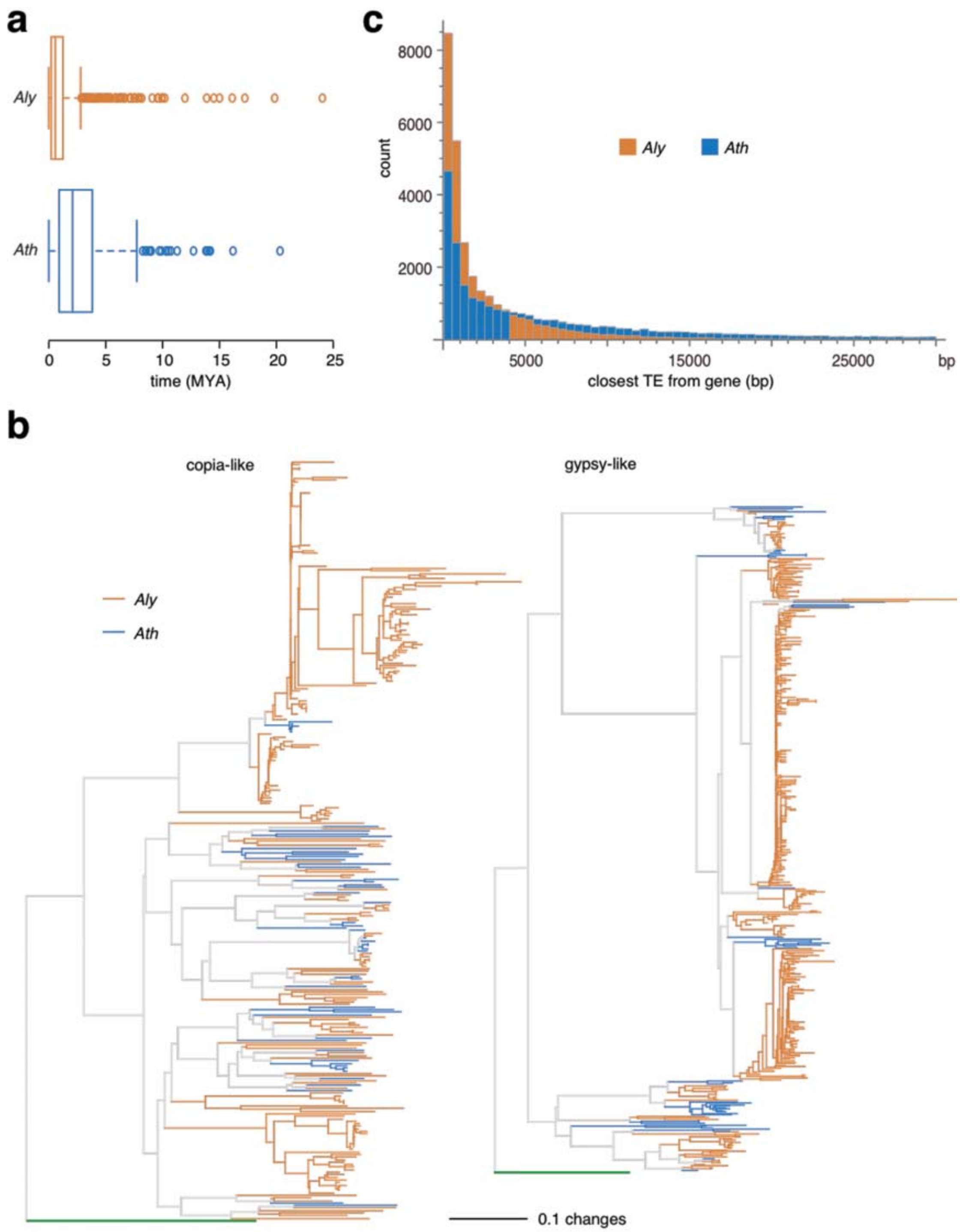


Figure 5 Comparison of transposable elements. (**a**) Estimated insertion times of LTR retrotransposons, based on the experimentally determined mutation rate for *A. thaliana*. The whiskers indicate values up to 1.5 times the interquartile range. The difference between the species is highly significant (Wilcoxon rank sum test, $p<2.2x10^{-16}$). (**b**) Phylogeny of

Ty1/copia-like and Ty3/gypsy-like LTR retrotransposons. *S. cerevisiae* Ty1 and Ty3 used as outgroups are indicated in green. (**c**) Distances of nearest TE from each gene. The difference between the two species is not simply due to fewer transposable elements in the *A. thaliana* genome (Supplementary Table 7 and Supplementary Fig. 7).

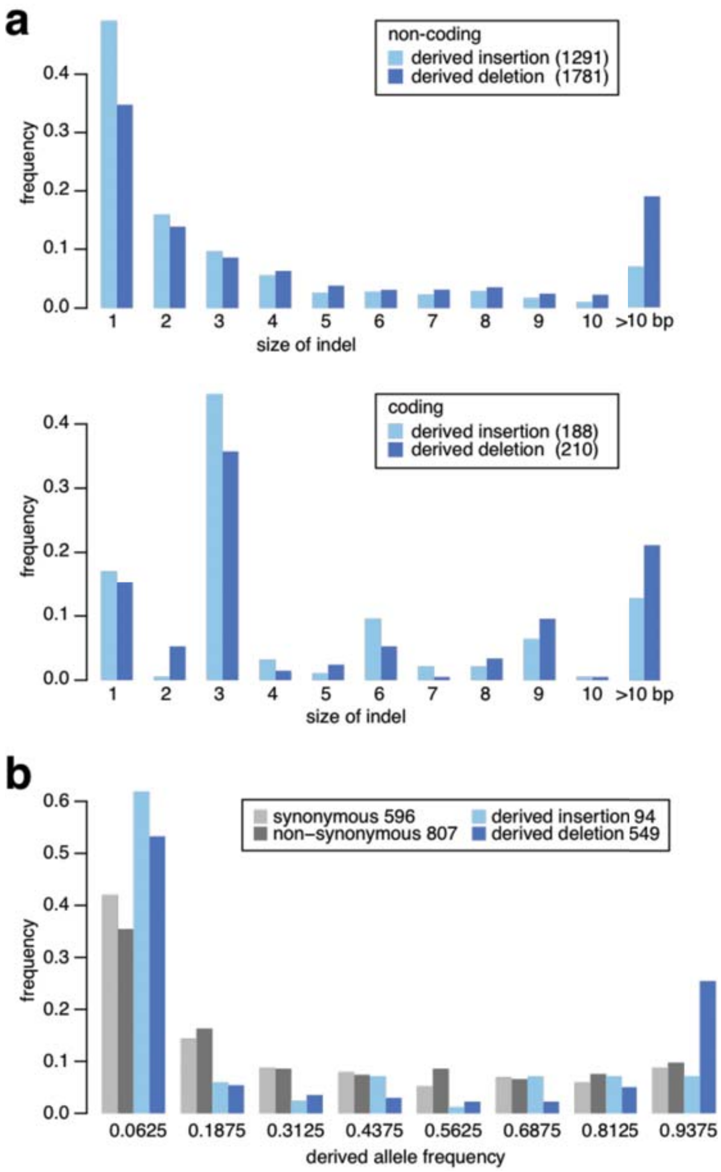


Figure 6 Sizes and allele frequency distribution of derived insertions and deletions that are either fixed or still segregating in 95 *A. thaliana* individuals[42]. (**a**) Size distribution of fixed insertions and deletions. Insertions and deletions that are multiples of a single codon (3 bp) are overrepresented in coding regions. (**b**) Allele frequency of segregating non-coding insertion and deletion frequencies compared to that of synonymous and non-synonymous polymorphisms.

Supplementary Online Material

The *Arabidopsis lyrata* genome and the basis of rapid genome size change

Supplementary Methods

Plant samples and DNA extraction

Sequencing samples. *Arabidopsis lyrata* is a perennial, outbreeding, rosette-forming herb found between sea level and <1500 m in elevation in cold-temperate to mild climatic regions of the Northern Hemisphere[1]. Within diploid *A. lyrata*, two major subspecies are recognized, including *A. lyrata* ssp. *petraea* in central and northern Eurasia, *A. lyrata* ssp. *lyrata* in North America. The material used here was from North America. As an outcrosser, *A. lyrata* is highly heterozygous. We therefore sequenced the genome of an inbred strain, MN47, which was derived by forced selfing from material collected in Michigan, USA, by Dr. Charles Langley (UC Davis). This strain was selfed six times before extracting DNA for sequencing. Seeds have been deposited with the Arabidopsis Biological Resource Center, accession number CS22696.

DNA extraction. Nuclear DNA was extracted from young leaves, using a protocol (http://my.jgi.doe.gov/general/protocols/DNA_Isolation_of_Nuclear_DNA_of_Arabidopsis_lyrata.pdf) modified from published methods[2,3]. The final cpDNA contamination, monitored by real-time PCR, was in the range of a few percent.

Sequencing and assembly

Sequencing. Five shotgun libraries were analyzed on ABI 3730XL capillary sequencers, sequencing both ends. Three different sized libraries were used as templates for plasmid shotgun clones, and one each for fosmids and BACs. The majority of reads was collected at the DOE Joint Genome Institute; BAC end sequences were obtained at the Hudson Alpha Institute. Insert sizes and sequence coverages of individual libraries were as follows (calculated from final assembly):

Library	Clones	Avg. insert size (bp)	Std. deviation (bp)	Coverage (x)
Plasmid	1,392,278	3,313	364	3.69
Plasmid	846,313	6,553	574	2.35
Plasmid	494,332	6,843	703	1.29
Fosmid	338,878	35,255	4,228	0.87
BAC	61,440	155,619	30,277	0.14

Genome assembly and chromosome anchoring. Sequence reads were assembled using a modified version of Arachne v.20071016 (ref. 4) with parameters maxcliq1=100, correct1_passes=0 and BINGE_AND_PURGE=True. This produced 1,309 scaffold sequences, with N50 of 5.2 Mb, 74 scaffolds larger than 100 kb, and a total scaffold size of 194.0 Mb. The following table provides summary statistics of the output of the whole genome shotgun assembly, before breaking and constructing chromosome scale pieces.

Minimum scaffold size (bp)	Scaffolds		Included contigs		Gaps (%)
	#	Total sequence (bp)	#	Total sequence (bp)	
5,000,000	13	97,771,728	879	96,820,301	0.97
2,500,000	26	144,179,466	1,450	142,492,702	1.17
1,000,000	41	170,675,198	1,814	168,368,678	1.35
500,000	49	176,531,335	2,048	173,758,223	1.57
250,000	57	179,154,970	2,214	175,966,556	1.78
100,000	74	181,634,867	2,366	177,785,913	2.12
50,000	115	184,341,871	2,650	179,796,156	2.47
25,000	156	185,906,379	2,845	180,806,032	2.74
10,000	396	189,386,567	3,381	184,164,675	2.76
5,000	827	192,418,037	4,188	187,009,633	2.81

2,500	1,196	193,937,079	4,855	188,367,919	2.87
1,000	1,221	193,975,355	4,888	188,401,134	2.87
0	1,309	194,022,879	4,976	188,448,658	2.87

Four genetic maps (two for the entire genome, and two of higher resolution for linkage groups 1, 2, 6 and 7) were used to assign the scaffolds to the eight *A. lyrata* linkage groups[5-8]. One of the maps[5] was from a cross with the *lyrata* subspecies, which was sequenced, as one parent; the three other maps had the European *petraea* subspecies as parent. Combining these maps resulted in 255 non-redundant markers that could be used for scaffolding.

To assign scaffolds to linkage groups, colinearity between the *A. thaliana* pseudomolecules[9] and the *A. lyrata* scaffolds was established. Colinearity information was then integrated with marker information from the genetic maps, to reconstruct the eight linkage groups. Despite the high quality of the Arachne assembly, a number of obvious false joins between the sequenced reads remained. To resolve these incoherencies, a most parsimonious solution was searched. Suppose that a first group of markers assign a scaffold to chromosome A and another group to chromosome B. Each marker contributed to a combined score, and the scaffold was assigned to the chromosome with the highest scoring marker group. Markers from the sequenced *lyrata* subspecies conferred a score twice that of markers from the *petraea* subspecies. In addition, distortion of the genetic map [5] was taken into account. Likewise, the specific order of a subset of markers from a linkage group sometimes displayed inconsistencies with one or more scaffolds, again implying false joins. All areas containing putative false joins were identified and rechecked in the original assembly. All of ~30 putative false joins indeed turned out to be suspicious and were broken accordingly. In a second phase, scaffolds without markers, but falling by parsimony in between successfully fixed scaffolds, were assigned to the corresponding chromosome.

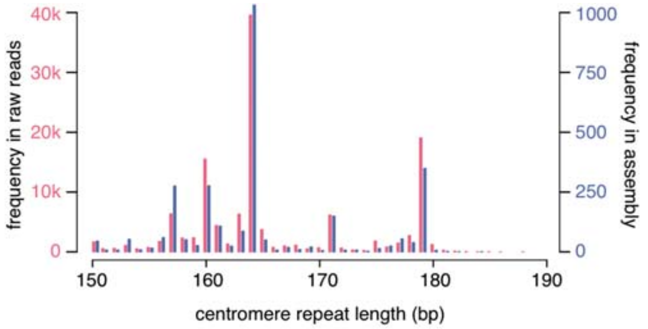
The final assembly included eight large scaffolds covering the majority of each of the eight chromosomes, with another large scaffold of about 1.9 Mb representing the centromere of either chromosome six or seven; together, they covered 196.1 Mb.

Scaffold	Length (bp)	GC content (%)	Centromere gaps (bp)*
1	33,132,539	37.49	2,739,197
2	19,320,864	37.40	2,156,401
3	24,464,547	37.47	1,909,399
4	23,328,337	37.55	2,158,401
5	21,221,946	37.41	1,651,689
6	25,113,588	37.57	2,325,319
7	24,649,197	37.27	2,156,401
8	22,951,293	37.78	2,156,401
9	1,906,741	36.86	N/A

*Estimated by comparing length of assembled sequence and cytological size estimates [10].

The remaining scaffolds were classified depending on sequence content. Ninety scaffolds were identified as prokaryotic using megablast against GenBank nr and blastp against a set of microbial proteins. Additional scaffolds included 54 unanchored rDNA scaffolds, one mitochondrial scaffold, and six chloroplast scaffolds. There were 343 small unanchored repetitive scaffolds, as defined by 95% of the 24mers occurring at least five times in the large scaffolds. Sixteen scaffolds shorter than 1 kb were removed as well. In addition to the nine major scaffolds, 686 scaffolds were retained; together they added up to 206.7 Mb. Main genome scaffold N50 (L50) was 4 (24.5 Mb), and main genome contig N50 (L50) was 247 (227.4 kb). The number of scaffolds >50 kb was 49, covering 97.2% of the genome. The total number of contigs was 3,648.

To evaluate the quality of the assembly, we compared the frequency of differently sized centromere satellite repeats, which are particularly difficult to assemble, in both the raw sequence reads and the assembly. The results were very similar, indicating that the final assembly presents an accurate picture of the genome:



Comparison of gene content and size

Gene prediction. The genome was annotated using several gene prediction, annotation and analysis tools. Gene models were predicted using the Fgenesh package of *ab initio* and homology-based gene predictors[11], EuGene[12], which exploits both intrinsic and extrinsic information, and GeneID[13] (http://genome.imim.es/software/geneid/) applying dicot and *A. thaliana* specific matrices. *A. thaliana* proteins were mapped onto the *A. lyrata* genome with Genomethreader[14], and these were used to build *A. lyrata*-specific Interpolated Markov Models (IMMs), for training of EuGene. Splice sites were predicted by SpliceMachine[15] with the *A. thaliana* model. Swissprot and TAIR8 proteins were incorporated as homology information based on BlastX results, and *Arabidopsis* EST and cDNAs (mainly *A. thaliana*) from NCBI mapped on to the genome were used to assist the gene prediction. Homology information was enhanced by GenomeThreader mapping of TAIR8 proteins with a more stringent setting, which was given especially high weight so as to capture the information of well-conserved homologous genes as much as possible. TblastX was run against the genomic sequences of *A. thaliana* and also *Populus trichocarpa* with a specific matrix that assigns a high penalty to stop codons in order to detect conserved stretches of sequences without frameshifts, which may include open reading frames that had not been previously annotated. The output was integrated into the EuGene prediction.

Heterologous transcriptional evidence included EST transcript assemblies[16] of the following species: *Arabidopsis thaliana* TAIR8 annotation[17], *Brassica napus*, *B. rapa*, *B.*

oleraceae, *Solanum tuberosum*, *Nicotiana tabacum*, *Helianthus annuus*, *Glycine max*, *Vitis vinifera* and *Malus domestica*. For each locus, the model combination with the best fit to reference proteins (SWISSPROT and all proteins of *Oryza sativa* and *A. thaliana* that are supported by full length cDNAs) was reported as consensus model. This non-redundant model set was used as the primary gene set for further analyses. The majority, over 90%, of predicted gene models was supported by sequence similarity of GenBank nrprot.

Additional evidence came from RNA-seq analysis, generated by sequencing double-stranded cDNA prepared from RNA extracted from mixed stage-flowers on the Illumina GA platform. A total of 21,875,086 reads of length 38 bp, for a total of 832 Mb, were aligned to the genome using SHORE/GenomeMapper[18,19]. Over 120,000 expressed segments, equivalent to exons or exon fragments, were identified, and these provided support for 37% of predicted genes. Seventy percent of predicted genes were supported by similarity to *A. thaliana* ESTs.

In total, 32,670 genes were predicted. Characteristics of these gene models are summarized in the following table.

***A. lyrata* gene model characteristics**

Average gene length (bp)	2,080
Average transcript length (bp)	1,870
Exon length (bp)	223
Intron length (bp)	202
Protein length (amino acids)	361
Exons per gene	5.33
Genes per Mbp	158

tRNAscan-SE version 1.21 (ref. 20) was used to detect transfer RNA genes in the genome sequences with default parameters. In *A. thaliana* 639 tRNA genes (including 8 pseudogenes) were predicted, with 83 tRNA genes having introns. In *A. lyrata*, 639 tRNA genes (including 12 pseudogenes) were predicted, with 58 tRNA genes having introns.

Genome position and miRBase[21] were used to identify orthologous or paralogous miRNAs. Those with three or fewer mismatches to the *A. thaliana* set of mature miRNAs were defined as conserved, and those with more as diverged. Unique miRNA loci had no apparent ortholog in the other species. In the miRBase set, miR775 was labeled as unique, but a de novo *A. lyrata* miRNA search revealed a miRNA that was diverged from, but still clearly related to miR775. Small RNA libraries were prepared and sequenced as described[22]. A single read from any of seven *A. thaliana* or three *A. lyrata* small RNA libraries was considered as evidence for expression. The final count of miRNA loci was a follows:

	Expression in *A. thaliana*		**Expression in *A. lyrata***	
	Yes	**No**	**Yes**	**No**
miRNA conserved	134	0	120	26
miRNA diverged	7	0	3	4
Unique miRNA	26	0	0	0

Annotation and GO categorization. Using the JGI Annotation pipeline, predicted gene models were functionally annotated by sequence similarity to TAIR8 gene models[17] and GenBank nrprot, and classified according to Gene Ontology [23], eukaryotic orthologous groups (KOGs)[24], and KEGG metabolic pathways[25]. InterPro and PFAM domains were predicted in 21,779 (67%) and 19,661 (60%) *A. lyrata* genes, respectively, which corresponded to 3,422 distinct InterPro and 2,351 Pfam domains. A total of 807 different EC numbers were assigned to 3,380 genes mapped to KEGG pathways. The majority of genes, 19,511 and 15,639, were described by 3,453 KOG functions and 2,325 GO terms, respectively. Additional Gene ontology (GO) labels were assigned using Blast2GO[26], with nrprot as reference database (E-value $\leq 1x10^{-4}$). For each query, only the twenty best subjects were considered as input for Blast2GO. Due to the relatively recent speciation between the two *Arabidopsis* species, it was further assumed each GO label should be present for all members of an orthologous group. A GO label/domain was considered characteristic for an orthologous group when it was shared by at least 30 % of members and a significant enrichment could be shown by a hypergeometric test with Benjamini-Hochberg correction ($P < 0.05$).

MCL cluster analyses. MCL (mcl-06-058 package; http://micans.org/mcl/src/) was used with default parameters (-I 2, -S 6) based on clustering of hits with E-value $\leq 10^{-5}$. MCL uses a Markov cluster algorithm that attempts to overcome many of the difficulties with protein sequence clustering, such as the presence of multi-domain proteins, peptide fragments and proteins with very common domains. The method has been used for a variety of animal genomes[27-29].

OrthoMCL analysis. Orthologous gene clusters were computed from OrthoMCL comparisons[30] of four dicotyledonous species with finished genomes: *A. thaliana* and *A. lyrata*, *Populus trichocarpa*[31] and *Vitis vinifera*[32,33]. A search for potentially missed genes in both *Arabidopsis* genomes resulted in minor adjustments of the OrthoMCL clusters. Instead of 10,573, 10,878 clusters now contained at least one gene of each the four species, and instead of 5,699, 5,800 clusters were *Arabidopsis*-specific. To determine deleted or newly generated orthologs (by OrthoMCL definition) between the two species, we focused on clusters specific for either *A. lyrata* or *A. thaliana*. For both species, there are two cluster types, those that are supported by members in *P. trichocarpa* and/or *V. vinifera* (supported specific cluster, SSC), and clusters exclusively found in one of the *Arabidopsis* species (exclusive specific cluster, ESC). We did not consider 2,939 and 6,103 unclustered genes (singletons) in *A. thaliana* and *A. lyrata*, respectively.

In our initial analysis, we detected 354 SSCs and 161 ESCs for *A. thaliana*, and 168 SSCs and 833 ESCs for *A. lyrata*. Whole genome projects, however, may contain false positive as well as missed or incomplete/partial gene calls that impose difficulties for OrthoMCL to detect orthologous relationships. To ensure that genes from the previously detected SSCs were indeed specific for one of the *Arabidopsis* species, we re-evaluated absence or presence of specific gene calls in the two genome sequences. Previously missed genes detected by GenomeThreader were added to each of the gene sets and the OrthoMCL analysis was repeated.

F-box and NB-LRR gene analysis. Using F-box PF00646.hmm as HMM profile with hmmsearch (E-value ≤ 10^{-5}), 394 hits were found from in *A. thaliana* and 461 hits in *A. lyrata*. Alignment of these sequences was optimized with the PF00646 seed using ClustalX 2.0 (ref. 34). The final alignment was produced by aligning with hmmalign against PF00646.hmm, to construct an *Arabidopsis* specific HMM F box profile. With this HMM profile, 502 hits were found in *A. thaliana*, and 596 hits in *A. lyrata*. hmmalign was used to align all of these against PF00646.hmm.

A blastp search (E-value ≤ 10^{-10}) performed with the NB domain (based on HMMEMIT, from http://niblrrs.ucdavis.edu/At_RGenes/). The NB domains of the retrieved proteins, 142 in *A. thaliana* and 162 in *A. lyrata*, were aligned using ClustalX (ref. 34). This alignment was used to develop an *Arabidopsis*-specific HMM profile, which was used to search the complete set of proteins encoded by both the two genomes (cut off E ≤ 10^{-5}).

PAUP* version 4.0b10 (ref. 35) was used to reconstruct phylogenetic trees with neighbor-joining method.

RepeatMasker analyses. To develop *de novo* repeat libraries for both species, we used RepeatModeler (version Beta 1.0.3, http://www.repeatmasker.org/RepeatModeler.html). To reduce false positives, unclassified repeats were compared to annotated genes, eliminating all that had at least 80% identity to annotated genes over at least 80 bp (GenBank: Green plant GB all [protein]; blastx with E-value ≤ 10^{-10}). The remaining RepeatModeler predictions were classified with the 80-80-80 rule[36], grouping repeats if they shared at least 80% identity over at least 80% of the aligned sequence, which had to be at least 80 bp long. The identified repeats were appended to RepBase (*Arabidopsis* library - RM database version 20080611), resulting in a final library with 1,152 repeat units.

LTR retrotransposons. Intact LTR retrotransposons were identified de novo using LTR_STRUC (ref. 37) with default parameters. Based on the sequence divergence between the two LTRs of the same element, insertion times were estimated. All LTR pairs were aligned using MUSCLE (ref. 38), and the distance *K* between them calculated with the Kimura two-parameter model using the distmat program implemented in the EMBOSS package (http://emboss.sourceforge.net/). The insertion time *T* was calculated as $T = K/(2)(r)$, with *r* as the rate of nucleotide substitution. The molecular clock was set based on the observed mutation rate of 7 x 10^{-9} per site per generation (assumed to equal one year)[39].

Classification and phylogeny of LTR retrotransposons. LTR retrotransposons can be classified into Ty1/copia-like and Ty3/gypsy-like elements [40]. We classified repeats using RepBase (version 13.08, http://www.girinst.org/server/RepBase/) and blastn (E-value ≤ 10^{-10}), and by direct comparison against the JCVI/TIGR plant repeat database (http://blast.jcvi.org/euk-blast/index.cgi?project=plant). All intact LTR retrotransposons were compared with blastx (E-value ≤ 10^{-10}) against a conserved 156 amino acid segment corresponding to the reverse transcriptase domain[41] of Ty1/copia-like and Ty3/gypsy-like sequences, and this segment was then used for phylogenetic reconstruction, with PAUP* version 4.0b10 (ref. 35) and neighbor-joining method. As outgroup sequence, we used yeast the reverse transcriptase domain from yeast Ty1 and Ty3 elements, respectively[41].

Genome organization

Detection and analysis of chromosomal breakpoints. Genome wide colinearity was detected by running i-ADHoRe (ref. 42) on the core-orthologous genes, allowing the identification of breakpoints including inversions and nested inversions. For each inversion, 10 kb up- and downstream of the delimiting breakpoints were compared to each other using blastn (word size 4), tblastx (word size 1) and SSEARCH (refs. 43-45). Tblastx outperforms blastn for coding regions. In non-coding regions, SSEARCH is more sensitive than blastn, but computationally less efficient, and hence most useful for comparison of shorter sequences. Only one hit per strand was reported. Therefore, for each pair of inversion flanking regions, all combinations of repeats and protein coding genes were evaluated. Default settings were used for gap penalties. An E-value of ≤ 0.01 was considered as indicating similarity between the up- and downstream regions.

Similarity of syntenic regions. To investigate nucleotide divergence of intergenic regions around coding genes (Supplementary Figure 1b), we extracted for each syntenic gene pair the 2 kb sequences 5' of the start codon and 3' from the stop codon. If the neighboring gene was closer than 2 kb, the extracted sequence was accordingly trimmed. Coding sequences of syntenic genes were also analyzed. Global alignments of syntenic sequences were generated using the Needleman-Wunsch algorithm as implemented in the EMBOSS package 5.0 (default parameters). Sequence identity of coding regions was measured over the full-length alignment. To investigate whether divergence of intergenic sequence is affected by relative orientation to neighboring genes, upstream sequences were split into head-to-tail and head-to-head groups, and downstream sequences into tail-to-head and tail-to-tail groups.

Fixed insertions and deletions. To identify fixed insertions and deletions among 1,238 fragments that had been amplified by PCR and sequenced in 95 *A. thaliana* individuals[46], two representative sequences for each fragment were first constructed to represent the insertion and deletion states among all segregating indels. The representative sequence containing insertions was then queried against the *A. lyrata* genome with both BLAT (-maxGap=100 -extendThroughN -minIdentity=80) and BLAST (-e .00001 -F F -G -5 -E -1) (ref. 47,48). From the union of hits obtained by both methods, the representative sequences for each alignment were profile-aligned with the *A. lyrata* allele with MAFFT (ref. 49). Fixed insertions and deletions were identified in the resulting alignment.

Segregating insertions/deletions. A similar procedure to that described above was used to identify the *A. lyrata* allele (presumed ancestral state) for each polymorphic indel in *A. thaliana*. Instead of querying the entire fragment, we queried each insertion allele along with 25 bps flanking each side, against the *A. lyrata* genome using BLAT. We then filtered for hits that spanned both sides of the indel site (by at least 3 bps) and reported each indel as either a derived insertion (if the *A. lyrata* allele was a deletion in the resulting profile alignment) or a derived deletion (if the *A. lyrata* allele was not a deletion).

Data and seed availability

The assembly and annotation (Entrez Genome Project ID 41137) are available from GenBank (accession number ADBK00000000) and from JGI's PHYTOZOME portal (http://www.phytozome.net/alyrata.php). Seeds of the MN47 strain have been deposited with the Arabidopsis Biological Resource Center under accession number CS22696.

Supplementary References

1. Clauss, M.J. & Koch, M.A. Poorly known relatives of *Arabidopsis thaliana*. *Trends Plant Sci.* **11**, 449-59 (2006).

2. Peterson, D.G., Boehm, K.S. & Stack, S.M. Isolation of milligram quantities of DNA from tomato (*Lycopersicon esculentum*), plant containing high levels of polyphenolic compounds. *Plant Mol. Biol. Reptr.* **15**, 148-153 (1997).

3. Peterson, D.G., Tomkins, J.P., Frisch, D.A., Wing, R.A. & Paterson, A.H. Construction of plant bacterial artificial chromosome (BAC) libraries: An illustrated guide. *J. Agric. Genomics* **5**, www.ncgr.org/research/jag (2000).

4. Jaffe, D.B. et al. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91-6 (2003).

5. Yogeeswaran, K. et al. Comparative genome analyses of *Arabidopsis* spp.: inferring chromosomal rearrangement events in the evolutionary history of *A. thaliana*. *Genome Res.* **15**, 505-15 (2005).

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

7. Hansson, B., Kawabe, A., Preuss, S., Kuittinen, H. & Charlesworth, D. Comparative gene mapping in *Arabidopsis lyrata* chromosomes 1 and 2 and the corresponding *A. thaliana* chromosome 1: recombination rates, rearrangements and centromere location. *Genet. Res.* **87**, 75-85 (2006).

8. Kawabe, A., Hansson, B., Hagenblad, J., Forrest, A. & Charlesworth, D. Centromere locations and associated chromosome rearrangements in *Arabidopsis lyrata* and *A. thaliana*. *Genetics* **173**, 1613-9 (2006).

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

10. Berr, A. et al. Chromosome arrangement and nuclear architecture but not centromeric sequences are conserved between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Plant J.* **48**, 771-83 (2006).

11. Salamov, A.A. & Solovyev, V.V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516-22 (2000).

12. Foissac, S. et al. Genome annotation in plants and fungi: EuGene as a model platform.

Curr. Bioinformatics **3**, 87-97 (2008).

13. Parra, G., Blanco, E. & Guigo, R. GeneID in *Drosophila*. *Genome Res.* **10**, 511-5 (2000).

14. Gremme, G., Brendel, V., Sparks, M. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Inform. Software Tech.* **47**, 965-978 (2005).

15. Degroeve, S., Saeys, Y., De Baets, B., Rouze, P. & Van de Peer, Y. SpliceMachine: predicting splice sites from high-dimensional local context representations. *Bioinformatics* **21**, 1332-8 (2005).

16. Childs, K.L. et al. The TIGR Plant Transcript Assemblies database. *Nucleic Acids Res.* **35**, D846-51 (2007).

17. Swarbreck, D. et al. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* **36**, D1009-14 (2008).

18. Ossowski, S. et al. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* **18**, 2024-2033 (2008).

19. Schneeberger, K. et al. Simultaneous alignment of short reads against multiple genomes. *Genome Biol.* **10**, R98 (2009).

20. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955-64 (1997).

21. Griffiths-Jones, S., Saini, H.K., van Dongen, S. & Enright, A.J. miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**, D154-8 (2008).

22. Fahlgren, N. et al. Computational and analytical framework for small RNA profiling by high-throughput sequencing. *RNA* **15**, 992-1002 (2009).

23. Harris, M.A. et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258-61 (2004).

24. Koonin, E.V. et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **5**, R7 (2004).

25. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277-80 (2004).

26. Conesa, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674-6 (2005).

27. Demuth, J.P., De Bie, T., Stajich, J.E., Cristianini, N. & Hahn, M.W. The evolution of mammalian gene families. *PLoS ONE* **1**, e85 (2006).

28. Prachumwat, A. & Li, W.H. Gene number expansion and contraction in vertebrate genomes with respect to invertebrate genomes. *Genome Res.* **18**, 221-32 (2008).

29. Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**, 203-218 (2007).

30. Li, L., Stoeckert, C.J., Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178-89 (2003).

31. Tuskan, G.A. et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596-604 (2006).

32. Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463-7 (2007).

33. Velasco, R. et al. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* **2**, e1326 (2007).

34. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. & Higgins, D.G. The CLUSTAL-X Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876-4882 (1997).

35. Swofford, D.L. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods): Version 4. (Sinauer Associates, Sunderland, Massachusetts, 2003).

36. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973-982 (2007).

37. McCarthy, E.M. & McDonald, J.F. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* **19**, 362-7 (2003).

38. Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

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

40. Xiong, Y. & Eickbush, T.H. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9**, 3353-62 (1990).

41. Zhang, X. & Wessler, S.R. Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and *Brassica oleracea*. *Proc. Natl. Acad. Sci. USA* **101**, 5589-94 (2004).

42. Simillion, C., Vandepoele, K., Saeys, Y. & Van de Peer, Y. Building genomic profiles for uncovering segmental homology in the twilight zone. *Genome Res.* **14**, 1095-106 (2004).

43. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-402 (1997).

44. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-7 (1981).

45. Pearson, W.R. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* **11**, 635-50 (1991).

46. Nordborg, M. et al. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3**, e196 (2005).

47. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-64 (2002).

48. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).

49. Katoh, K., Asimenos, G. & Toh, H. Multiple alignment of DNA sequences with MAFFT. *Methods Mol Biol* **537**, 39-64 (2009).

Supplementary Tables

Supplementary Table 1 Enrichment of genomic features in 204 chromosomal breakpoint regions.

	Results of two sample z test[†]	
Feature class	***A. lyrata***	***A. thaliana***
DNA transposons	$2.28\text{x}10^{-114}$*	$1.10\text{x}10^{-155}$*
LTR retrotransposons	$6.83\text{x}10^{-159}$*	$3.75\text{x}10^{-17}$*
Non-LTR retrotransposons	$3.70\text{x}10^{-26}$*	$3.55\text{x}10^{-82}$*
rRNA	1.00	1.00
Satellites	$5.49\text{x}10^{-33}$*	0.99
Simple repeats	$2.56\text{x}10^{-48}$*	$1.69\text{x}10^{-118}$*
Protein coding genes	1.00	1.00

[†] The fraction of nucleotides occupied by each feature class was counted in sliding windows, with window size 100 kb and step size 1 kb. Two hundred and four breakpoints were defined as the borders of the non-colinear regions at the breakpoints. Windows containing one or more breakpoints were considered as breakpoint associated. Enrichment in the breakpoint windows was evaluated using two sample z tests.

*p < 0.005

Supplementary Table 2 Annotation of unalignable sites (see Fig. 1d in main text), excluding Ns and centromeric regions.

Annotation	***A. lyrata***	***A. thaliana***
Genic	17%	15%
Exon	8%	7%
Intron	9%	8%
Intergenic	26%	29%
TEs and repeats	56%	56%
DNA transposon	18%	12%
LTR retrotransposon	26%	35%
non-LTR retrotransposon	3%	5%
Simple repeats	3%	1%

Supplementary Table 3 Average CDS exon and intron sizes (bp).

	A. lyrata		***A. thaliana***	
	exon	**intron**	**exon**	**intron**
All genes	223.3	201.5	230.0	164.0
Genes ≥3 exons	193.6	193.3	188.3	159.7
Orthologs	207.0	173.0	207.0	155.5

Supplementary Table 4 Detailed intron size comparison.

Length difference (bp)	***A. lyrata* >** ***A. thaliana***	***A. thaliana* >** ***A. lyrata***	***p*** [*]
1-10	21,525	22,610	0.0002
11-20	6,988	4,885	2.1×10^{-42}
21-50	6,384	4,007	4.4×10^{-61}
51-100	2,606	1,437	1.2×10^{-38}
101-500	2,109	996	2.7×10^{-45}
501-1000	177	60	7.7×10^{-8}
>1000	189	23	7.5×10^{-16}
All[†]	39,978	34,018	3.9×10^{-54}

[*] p-value of an exact test under the null hypothesis that the number of introns larger than in the other species is the same for *A. thaliana* and *A. lyrata.*

[†] 11,157 introns had the same length, and were not included.

Supplementary Table 5 Genes in GO categories.

GO Category	***A. lyrata***	***A. thaliana***
Molecular function	18,773 (59.3%)	16,487 (61.0%)
Biological process	16,708 (52.7%)	14,708 (54.4%)
Cellular component	19,254 (60.8%)	17,847 (66.0%)

Supplementary Table 6 Supported species-specific (SSC) and exclusive species-specific (ESC) OrthoMCL clusters, with the total number of genes in parentheses.

Species	Type*	N	Within 10 kb of breakpoints	In tandem arrays
A. lyrata	**SSC**	114 (146)	17	127
	ESC	875 (2968)	466	689
A. thaliana	**SSC**	45 (54)	12	45
	ESC	156 (465)	63	264

*SSCs may be interpreted under maximum parsimony as deletion of genes that are conserved in *P. trichocarpa* and/or *V. vinifera* in one of the *Arabidopsis* species, while ESCs, which are not found outside the *Arabidopsis* genus, likely contain newly duplicated or *de novo* generated genes.

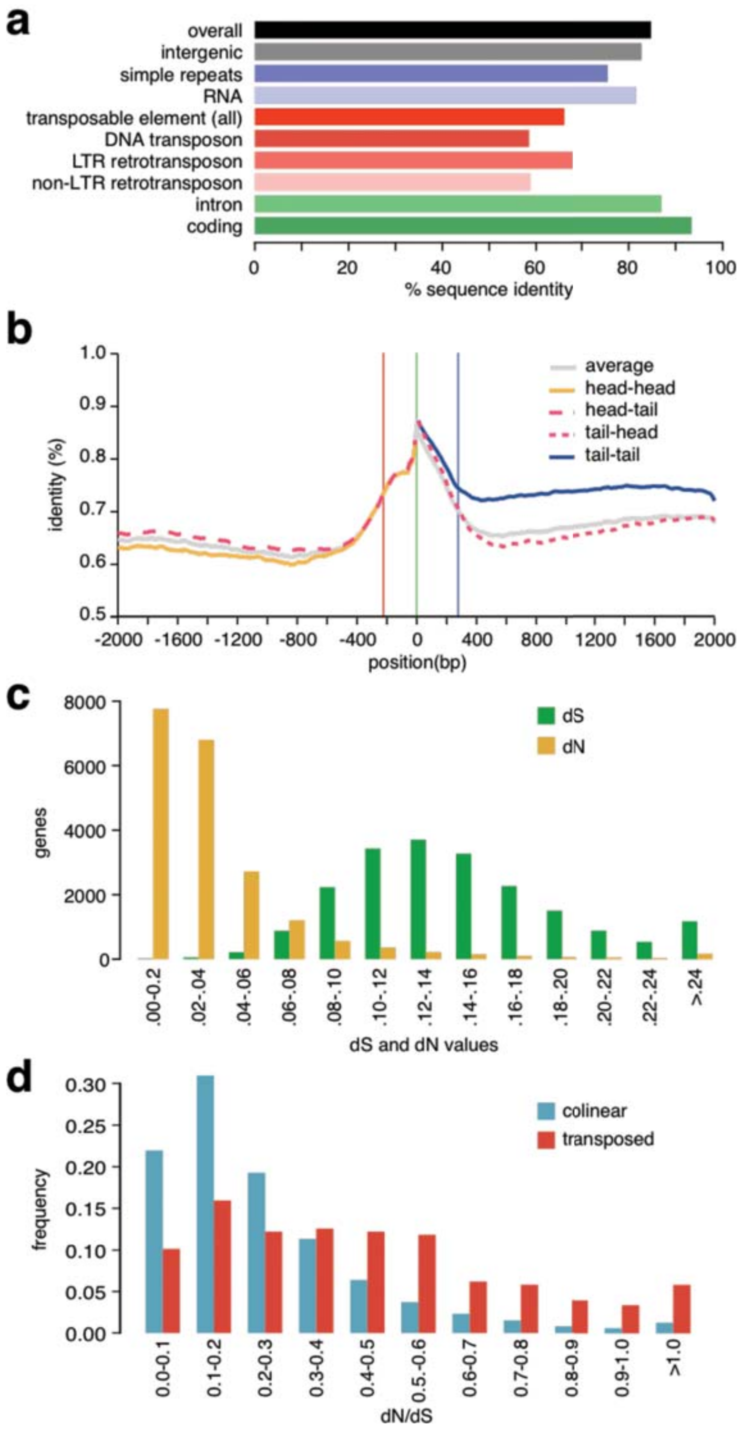
Supplementary Table 7 Average distance in bp of the closest TE from each gene[†].

	A. thaliana			***A. lyrata***		
	observed	**random**	**difference***	**observed**	**random**	**difference***
Median	3,652	1,784	1,868	999	1,005	-6
Mean	7,394	2,581	4,813	2,411	1,451	960

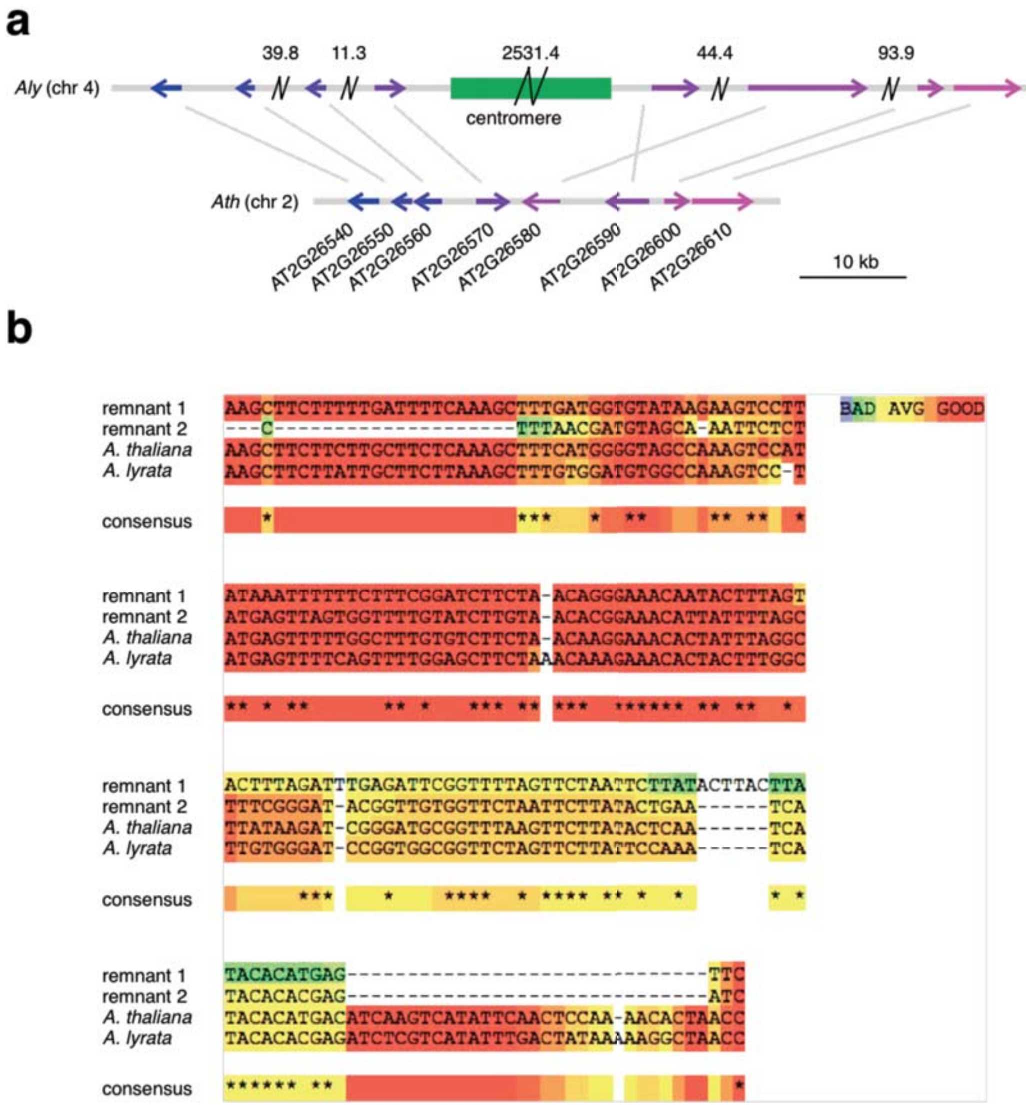
[†] Centromeres were excluded, and only intergenic sites were considered. The random data are from 1,000 simulated datasets. See Supplementary Figure 7.

*The difference between observed and simulated data was always significant at the $p < 0.05$ level after Bonferroni correction for *A. thaliana*, but only in 752 of the 1,000 datasets for *A. lyrata*.

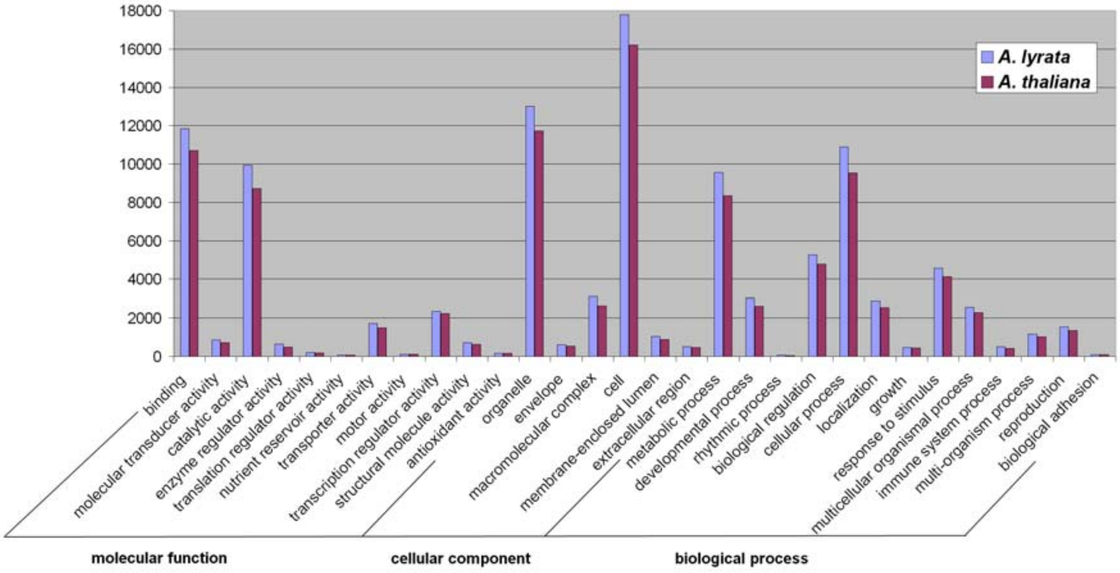
Supplementary Figures



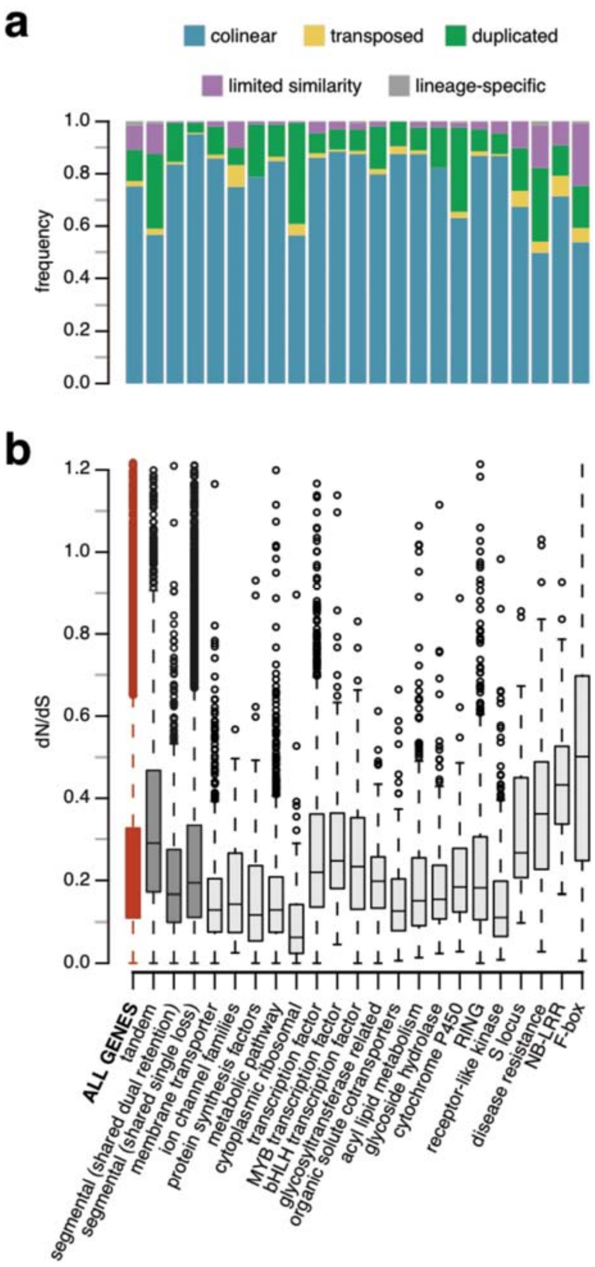
Supplementary Figure 1 Sequence divergence. Sequence identity in different classes of alignable sites (**a**) and up- and downstream of orthologs (window size 50 bp, step size 10 bp). (**b**) Red and blue lines depict average lengths of 5' and 3' UTRs (223 and 277 bp, respectively) in *A. thaliana*[17]. (**c**) Divergence at synonymous and non-synonymous sites between orthologs. (**d**) dN/dS between colinear and transposed genes.


a
39.8
11.3
2531.4
44.4
93.9
kb
Aly (chr 4)
centromere
Ath (chr 2)
AT2G26540
AT2G26550
AT2G26560
AT2G26570
AT2G26580
AT2G26590
AT2G26600
AT2G26610
10 kb
b
BAD AVG GOOD
remnant 1 AAGCTTCTTTTTGATTTTCAAAGCTTTGATGGTGTATAAGAAGTCCTT
remnant 2 ---C--------------------TTTAACGATGTAGCA-AATTCTCT
A. thaliana AAGCTTCTTCTTGCTTCTCAAAGCTTTCATGGGGTAGCCAAAGTCCAT
A. lyrata AAGCTTCTTATTGCTTCTTAAAGCTTTGTGGATGTGGCCAAAGTCC-T
consensus * *** * ** ** ** *
remnant 1 ATAAATTTTTTCTTTCGGATCTTCTA-ACAGGGAAACAATACTTTAGT
remnant 2 ATGAGTTAGTGGTTTTGTATCTTGTA-ACACGGAAACATTATTTTAGC
A. thaliana ATGAGTTTTTGGCTTTGTGTCTTCTA-ACAAGGAAACACTATTTAGGC
A. lyrata ATGAGTTTTCAGTTTTGGAGCTTCTAAACAAAGAAACACTACTTTGGC
consensus ** * ** ** * *** ** *** ****** ** ** *
remnant 1 ACTTTAGATTTGAGATTCGGTTTTAGTTCTAATTCTTATACTTACTTA
remnant 2 TTTCGGGAT-ACGGTTGTGGTTCTAATTCTTATACTGAA------TCA
A. thaliana TTATAAGAT-CGGGATGCGGTTTAAGTTCTTATACTCAA------TCA
A. lyrata TTGTGGGAT-CCGGTGGCGGTTCTAGTTCTTATTCCAAA------TCA
consensus *** * **** * **** ** * * * *
remnant 1 TACACATGAG------------------------------TTC
remnant 2 TACACACGAG------------------------------ATC
A. thaliana TACACATGACATCAAGTCATATTCAACTCCAA-AACACTAACC
A. lyrata TACACACGAGATCTCGTCATATTTGACTATAAAAAGGCTAACC
consensus ****** ** *

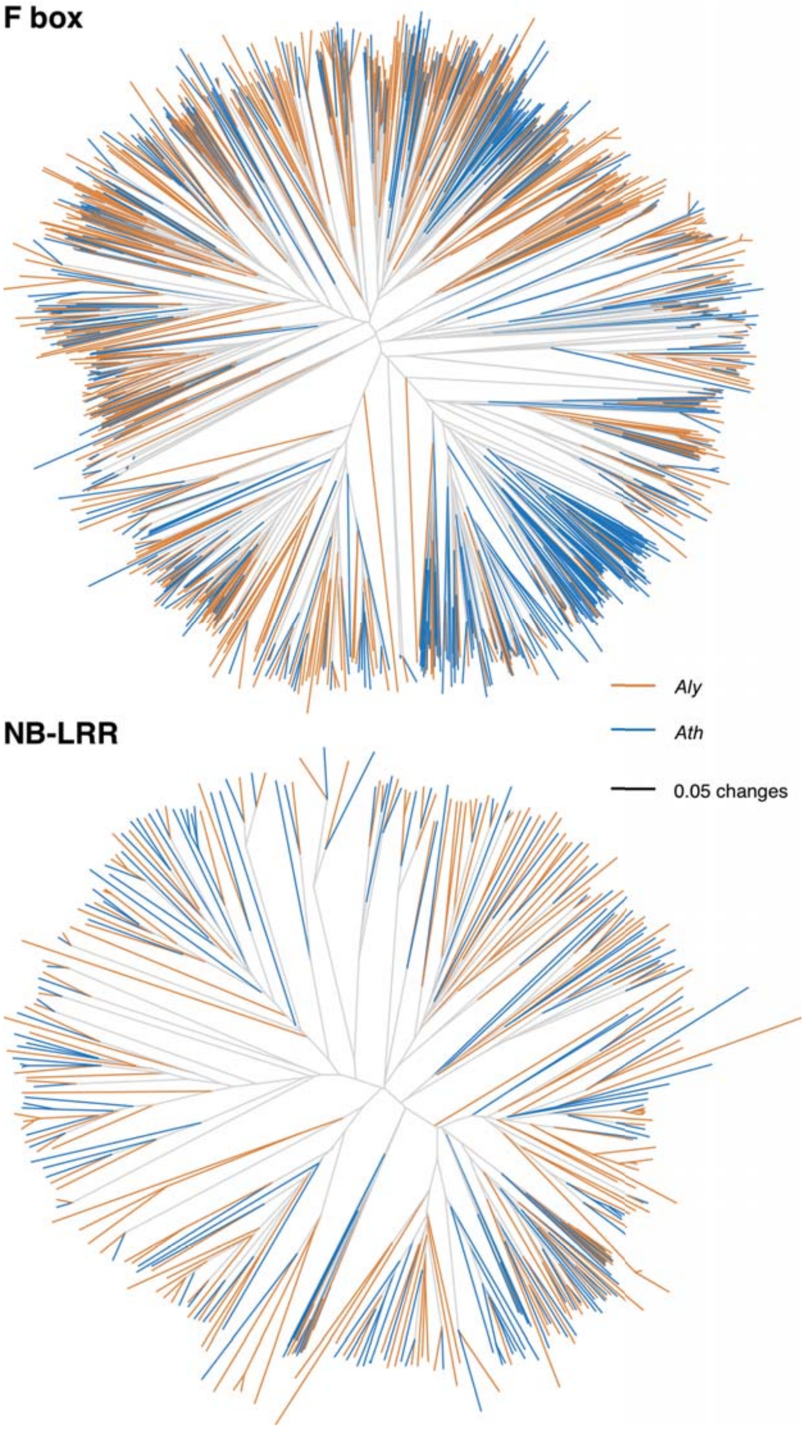
Supplementary Figure 2 Centromere loss in *A. thaliana*. (**a**) Comparison of centromere 4 of *A. lyrata* with a region on chromosome 2 of *A. thaliana*. Numbers on top give intergenic distances not drawn to scale. There are 16 annotated genes within the Al-CHR4 region that are not present in *A. thaliana*. Two of these are strongly supported by similarity to genes in *A. thaliana* or other plants. (**b**) Alignment of two remnants of satellite repeat-like sequences in this region from *A. thaliana* with canonical centromeric repeats in *A. thaliana* and *A. lyrata*.



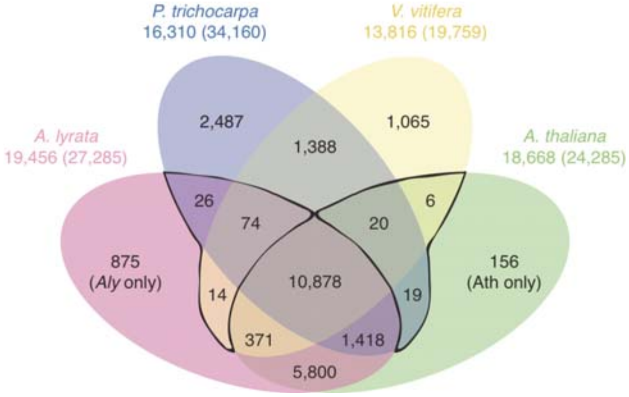
Supplementary Figure 3 Gene counts in major GO categories.



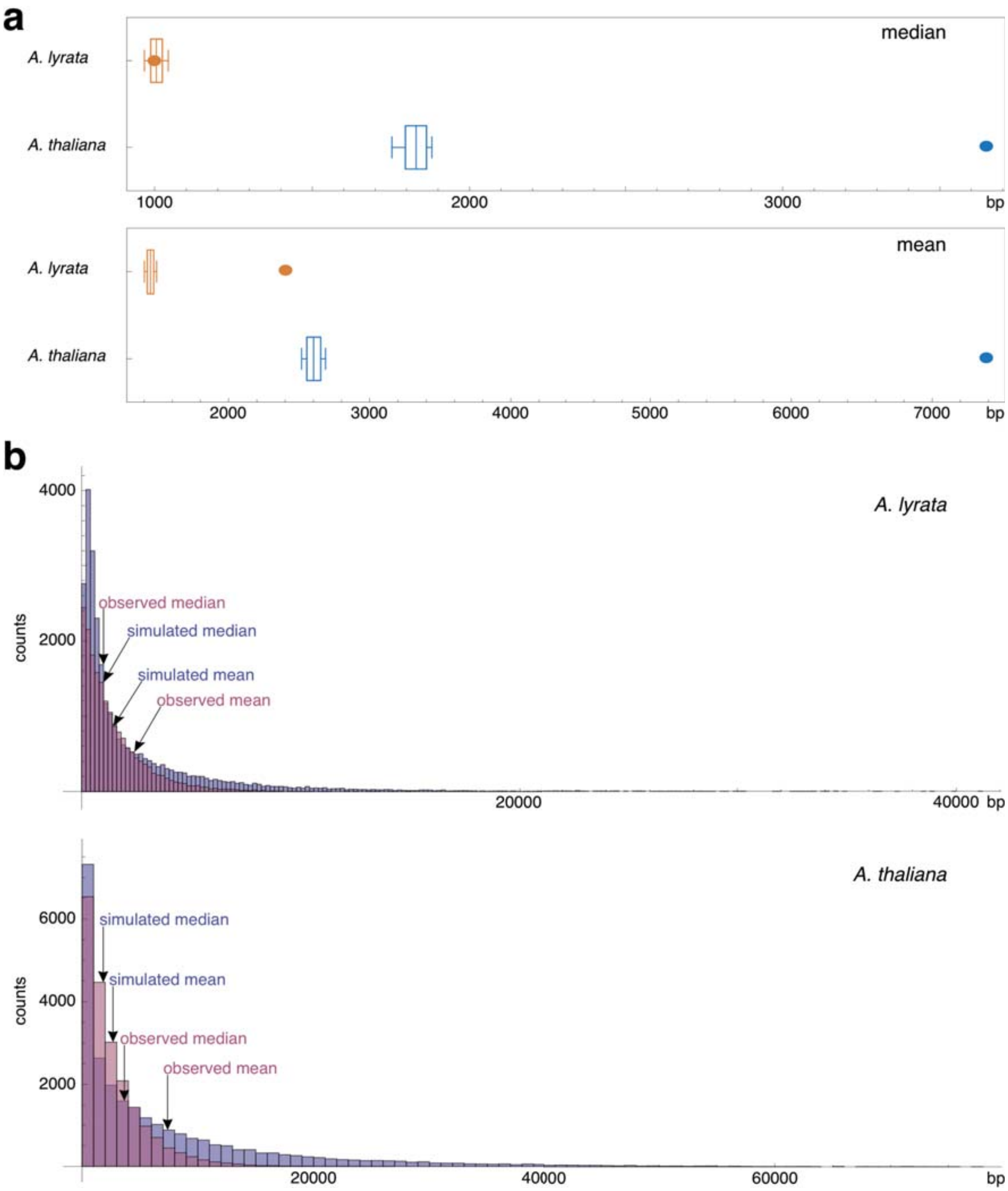
Supplementary Figure 4 Orthology status and dN/dS by syntenic position, annotation and gene family. (**a**) The distribution by orthology status of members to the gene family. (**b**) Boxes in box plot cover the first and third quartile, and the whiskers represent values that are not more than 1.5 times the interquartile range. Annotation at the bottom is for both panels.



Supplementary Figure 5 Phylogenetic trees of F-box and NB-LRR genes.



Supplementary Figure 6 OrthoMCL clusters in *Arabidopsis* and two other dicots. The number of OrthoMCL (ref. 30) clusters, with the number of genes that contribute to these clusters in parentheses, is given for each species. Black borders indicate clusters shared by all species except one of the *Arabidopsis* species (SSCs in Supplementary Table 6). The "*Aly* only" and "*Ath* only" groups are also indicated (ESCs in Supplementary Table 6).



Supplementary Figure 7 Comparison of observed with random distance of closest TE from each gene. (**a**) Box plots showing the distribution of the distance from genes to the nearest TE insertion in 1,000 random permutations. The boxes cover 90% of the data and the whiskers indicate the full range across the 1,000 simulations. The solid dots show the observed values. For the median distances, the observed value in *A. lyrata* is very close to the average of simulated values, whereas the distance in *A. thaliana* is much greater than expected from a random distribution. For the mean distances, the observed values in both species are greater than expected by chance, but the deviation is much greater for *A. thaliana*: 164 standard deviations from the simulation mean, compared to 71 for *A. lyrata*. (**b**) Distribution of observed (pink) and simulated distances in one random permutation (blue).