nature genetics

Article

https://doi.org/10.1038/s41588-024-01715-9

A pan-genome of 69 *Arabidopsis thaliana* accessions reveals a conserved genome structure throughout the global species range

Received: 24 May 2023

Accepted: 11 March 2024

Published online: 11 April 2024

Check for updates

Qichao Lian[1], Bruno Huettel[2], Birgit Walkemeier[1], Baptiste Mayjonade[3], Céline Lopez-Roques[4], Lisa Gil[4], Fabrice Roux[3], Korbinian Schneeberger[1,5,6] & Raphael Mercier[1,6]

Although originally primarily a system for functional biology, *Arabidopsis thaliana* has, owing to its broad geographical distribution and adaptation to diverse environments, developed into a powerful model in population genomics. Here we present chromosome-level genome assemblies of 69 accessions from a global species range. We found that genomic colinearity is very conserved, even among geographically and genetically distant accessions. Along chromosome arms, megabase-scale rearrangements are rare and typically present only in a single accession. This indicates that the karyotype is quasi-fixed and that rearrangements in chromosome arms are counter-selected. Centromeric regions display higher structural dynamics, and divergences in core centromeres account for most of the genome size variations. Pan-genome analyses uncovered 32,986 distinct gene families, 60% being present in all accessions and 40% appearing to be dispensable, including 18% private to a single accession, indicating unexplored genic diversity. These 69 new *Arabidopsis thaliana* genome assemblies will empower future genetic research.

Genome rearrangements can dramatically impact genetic diversity, phenotypes[1,2], recombination[3–8] and thus local adaptation and evolution[9–12]. The whole-genome alignment of complete genome assemblies, which can be achieved using long-read Oxford Nanopore Technologies (ONT) and PacBio high-fidelity (HiFi) technologies, facilitates the identification of large and complex structural variants (SVs)[13,14]. Pan-genomes, which aggregate multiple genomes covering the diversity of a given species, provide greater insights into the overall genetic diversity compared to using a single reference and have allowed researchers to determine the natural phenotypic variation of a species[15–18], as recently shown in plant and animal species, including soybean[19], tomato[20,21], potato[22], rice[23–25], maize[26], barley[27], wheat[28], apple[29], silkworm[12] and human[30,31].

The first genome sequence of *Arabidopsis thaliana* (Col-0) was released in 2000 (ref. 32) and has greatly boosted plant biology and breeding research. This assembly was based on the sequencing of bacterial artificial chromosomes using Sanger technology[32] and (with some updates) has served as a reference genome until today. Based on

[1]Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Cologne, Germany. [2]Max Planck-Genome-centre Cologne, Max Planck Institute for Plant Breeding Research, Cologne, Germany. [3]Laboratoire des Interactions Plantes-Microbes-Environnement, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement, CNRS, Université de Toulouse, Castanet-Tolosan, France. [4]INRAE, GeT-PlaGe, Genotoul, Castanet-Tolosan, France. [5]Faculty of Biology, Ludwig-Maximilians-University Munich, Planegg-Martinsried, Germany. [6]Cluster of Excellence on Plant Sciences, Heinrich-Heine University, Düsseldorf, Germany. e-mail: schneeberger@mpipz.mpg.de; mercier@mpipz.mpg.de

 https://doi.org/10.1038/s41588-024-01715-9

this reference genome, several studies applying whole-genome resequencing arrays or short-read sequencing on thousands of worldwide natural accessions of *A. thaliana* (in particular, Africa, Eurasia and North America) have unraveled natural genomic variations, including single-nucleotide polymorphisms (SNPs), small indels and SVs. This, in turn, revealed the evolutionary history, divergence and adaptation of *A. thaliana* at the macro- and micro-evolutionary scales[33–40]. However, only a limited number of large and complex SVs were reported in *Arabidopsis*, including the well-characterized ~1.2 Mb inversion on chromosome 4 between Col-0 and L*er*[5,8,41] and a ~2.5 Mb inversion on chromosome 3 between Col-0 and Sha[3], which was captured through chromosome-level assemblies of a few *A. thaliana* accessions[3,5,42–45]. A first *Arabidopsis* pan-genome analysis used the whole-genome assembly of seven accessions based on PacBio CLR, providing a first glimpse into the diversity of this species[3]. Recently, a study using PacBio HiFi to assemble the genomes of 32 accessions showed that SVs can play an important role in local adaptation[46].

In this Article, we de novo assembled 69 reference-quality *Arabidopsis* genomes either using PacBio HiFi or Oxford Nanopore long-read technologies, including accessions from Europe, the Middle East, Asia, Africa, Madeira and North America. With such a geographic spread, we aim to capture and describe most of the diversity in *Arabidopsis* genomes worldwide and constitute a comprehensive resource for future studies bridging phenotypes and genotypes.

Results

Genomic relationship of 72 *A. thaliana* accessions

We selected 72 *Arabidopsis* accessions across the global species distribution (Fig. 1a and Supplementary Table 1) and examined their genetic diversity and relationships through SNP analysis. Principal component, phylogenetic tree and admixture analyses showed that global genetic relationships between the 72 genomes broadly mirrored their geographic origins, as previously shown in this species[35] (Fig. 1a,b). We identified four major genetic groups and named them after the geographic origin of the majority of their members: 'Europe' (35 accessions, including 4 accessions from recently colonized North America[47]), 'Africa' (13 accessions, including accessions from the Mediterranean rim), 'Madeira' (5 accessions exclusively from Madeira) and 'Asia' (16 accessions distributed from Eastern Europe to Japan). In addition, we identified three accessions (Nemrut-1, Dog-4 both from Turkey and Can-0 from Spain) that did not cluster in distinct groups but were labeled 'admixed' and originated from geographic regions between the distinct groups (Fig. 1a–c and Extended Data Fig. 1). We also found that Tsu-0, which is labeled as an accession from Japan, clustered together with the European accessions. This is consistent with previous reports[48,49] suggesting that Tsu-0 was mislabeled, and we treated it as belonging to the 'Europe' genetic group from thereon.

To infer the evolutionary relationships among *A. thaliana* accessions, we constructed a species tree of the accessions (see below for further detail in pan-genome analysis, Supplementary Fig. 1), which confirmed their evolutionary relationships and was consistent with previous reports highlighting the African populations as the most divergent and probably most ancient lineages[36].

Chromosome-level assemblies of 69 *A. thaliana* accessions

We generated genome assemblies for each of the 72 accessions using a combination of long-read (48 accessions with PacBio HiFi with a mean depth of 45×, and 24 accessions with Oxford Nanopore with a mean depth of 67×) and short-read sequencing, reference-guided scaffolding and manual curation (Methods). The quality of the assemblies was analyzed in six different aspects across the assemblies (Supplementary Tables 2–7, Extended Data Fig. 2 and Supplementary Figs. 2 and 3). Sixty-nine accessions were confirmed to be inbred lines, but Lu-1, Pa-1 and Istisu-1 showed signs of heterozygosity and thus were removed from the subsequent analysis (Methods). The remaining contig assemblies featured N50 values from 6.1 to 21.3 Mb with a mean of 13.3 Mb, and were scaffolded to the chromosome level (Fig. 1d and Supplementary Table 4).

The assembly sizes of the 69 accessions ranged from 128 to 148 Mb, with an average length of 135 Mb (Fig. 2a). Previous estimates of genome size variation in *A. thaliana* using flow cytometry were generally higher, ranging from 161 Mb to 184 Mb (ref. 50) (180 lines from Sweden, and the estimation for Col-0 was 166 Mb). However, genome size estimations based on flow cytometry are known to suffer from overestimation[51]. More recent estimates derived from *k*-mer analyses based on short read resequencing data of 89 accessions, indicated a range from 138 Mb to 175 Mb (ref. 51). This variation in genome sizes was found to be largely determined by 45S ribosomal DNA (rDNA) copy number variation[50]. We also used *k*-mers to estimate the genome sizes of our 69 accessions, and compared them to their assembly sizes (Fig. 2a). Some of the ONT read-based assemblies were substantially shorter than their estimated genome sizes, as their rDNA arrays and centromeres were not fully assembled (Fig. 2a and Extended Data Fig. 2).

With the aim of deciphering the underlying genomic features contributing to the variation in genome size, we selected the 46 most complete assemblies (42 accessions assembled with HiFi, and 4 accessions, Bur-0, Ge-0, Jea and Nok-1, assembled with ONT) based on both (i) the ratio of assembly and *k*-mer-based genome size estimation and (ii) the ratio of centromere repeat length and read coverage-based centromere size estimation (Fig. 2a and Extended Data Fig. 2). The assembly sizes of these accessions ranged from 130 to 148 Mb. Their centromeric repeat arrays were on average 14 Mb long (across all five chromosomes) ranging from 10 to 22 Mb and were highly correlated with assembly sizes (Pearson's correlation $r = 0.93$, $P < 2.2 \times 10^{-16}$, Fig. 2b, Extended Data Fig. 2 and Supplementary Table 7). This showed that the variation of the size of the centromeric arrays in *Arabidopsis*, as recently described by Wlodzimierz et al.[52], is a major contributor to the variation of genome size. Transposable elements (TEs) were annotated with a pan-TE library generated from initial TE annotations of the individual assemblies. The size of TE space (that is, genomic regions with similarity to TEs) was surprisingly similar between the genomes, with a mean length of 16.1 Mb, ranging from 15.2 to 17.6 Mb (Fig. 2d,e). Among them, long terminal repeat (LTR) retrotransposons (Copia, Gypsy and LTR unknown) and Helitrons made up the largest TE fractions and constituted 6.4% and 3.5% of the genome, respectively (Fig. 2d). Accordingly, the variation in the size of TE space between the accessions was moderately correlated with the total assembly size (Pearson's correlation $r = 0.42$, $P = 0.003$, Fig. 2e). This suggests that genome size variation in *Arabidopsis* (excluding the variation of 45S rDNA size) is mostly dominated by centromeric repeat length and that TEs are only minor contributors. This is in sharp contrast to the situation between plant species, where the main determinants of genome size variation are ploidy levels and TE content[53]. In species with high TE content, such as rice[23], TE space variation can contribute more largely to intraspecific variation in genome size. Interestingly, however, even though cumulative centromere size determines genome size in *A. thaliana*, the sizes of individual chromosomes were only weakly or not correlated in size (Pearson's correlation $r = 0.2$, $P = 0.171$, Fig. 2c and Supplementary Figs. 3 and 4), indicating that the sizes of individual centromeres evolve independently from each other.

A quasi-fixed karyotype across the *A. thaliana* species range

Chromosome-level genome assemblies allow accurate analysis of large-scale genomic rearrangements and genome colinearity[13]. Using pairwise whole-genome alignments, we found that the chromosome arms hardly contained any major rearrangements and were highly syntenic across all genomes, even when comparing genomes from distant parts of the world (Fig. 3, Supplementary Figs. 5–9 and Extended Data Fig. 3). Large insertion/deletion polymorphisms are absent from chromosome arms, the vast majority being smaller than 20 kb and the

 https://doi.org/10.1038/s41588-024-01715-9

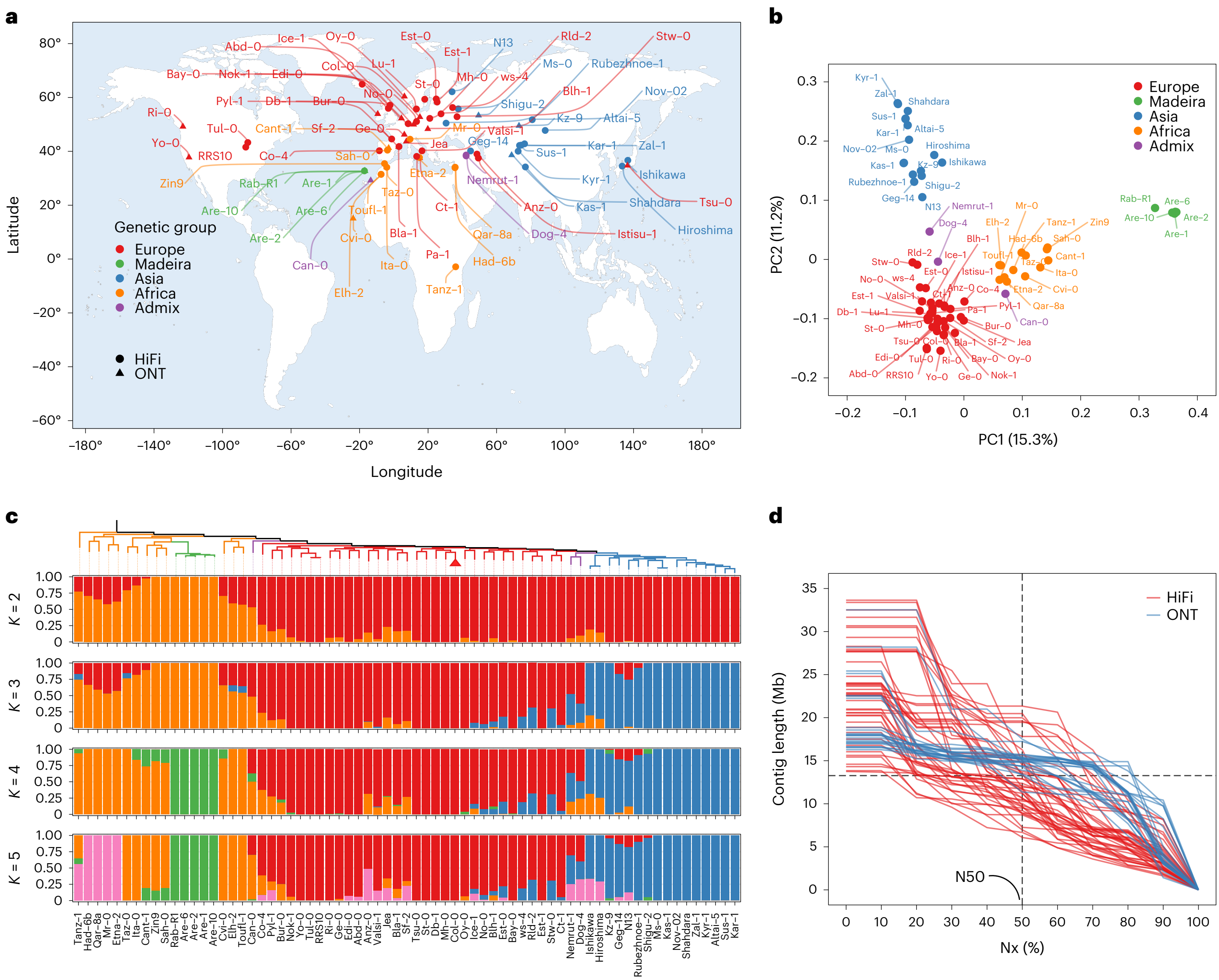


Fig. 1 | Geographic distribution and population analysis of 72 *A. thaliana* accessions. a, The geographic distribution of the 72 accessions in this study. The different sequencing technologies are indicated by differently shaped dots. **b**, A principal component analysis of the 72 accessions. Colored dots indicate the genetic classification of each accession; three accessions (Nemrut-1, Dog-4 and Can-0) were found to be admixed and are marked as purple dots. **c**, The phylogenetic tree and population structure of the 72 accessions with different numbers of ancestral kinships (*K* = 2, 3, 4 and 5). Each color represents one group. Each accession is represented by a vertical bar, and the length of each colored segment in each vertical bar represents the proportion contributed by ancestral populations. **d**, Assembly contiguity shown as the Nx (the length of the shortest contig that longer and equal length contigs represent x% of the assembly) plot of 69 accessions. Accessions sequenced by PacBio HiFi and Oxford Nanopore are colored in red and blue, respectively. The world map was generated in the ggplot2 package. Source data are provided as a source data file.

largest reaching ~55 kb (Extended Data Fig. 3). Inversions along chromosome arms are also rare, but are larger than insertions/deletions, with a few cases above one megabase. We identified a total of seven inversions on chromosome arms, almost all present in single accessions (Fig. 3, Supplementary Figs. 5–9 and Extended Data Fig. 3): a ~2.4 Mb inversion on chromosome 3 in Shahdara, a ~2.3 Mb inversion on chromosome 5 in Zal-1, a ~2.2 Mb inversion on chromosome 1 in N13, a ~1.8 Mb inversion on chromosome 4 in Ws-4, a ~1.2 Mb inversion on chromosome 4 in Stw-0 and a ~1 Mb inversion on chromosome 2 in Ge-0 (validated by long read alignment, Extended Data Fig. 4 and Supplementary Figs. 10–14). A notable exception to this was the well-described ~1.2 Mb inversion on chromosome 4 (refs. 5,41), which was observed in eight accessions (including Col-0) and which partially overlapped with the heterochromatic, pericentromeric regions. This inversion is found in the 'Europe' genetic group among geographically distant accessions (for example, Yo-0 from North America and Ws-4 from Belarus), thereby suggesting a long-lived segregation of this particular inversion.

The karyotype of *A. thaliana* is quite different to the estimated ancestral karyotype of the Brassicaceae, which is still conserved in the sister species *Arabidopsis lyrata*[54,55] and involved a species-specific deletion of three functional centromeres along with major chromosomal rearrangements and fusions[56]. The high structural similarity between the 69 genomes implied that the derived karyotype of *A. thaliana* arose during or shortly after speciation and was maintained virtually unchanged during the global spread of this species that colonized contrasted ecological habitats.

In contrast to the chromosome arms, large rearrangements of different types were highly abundant in and near the centromeres and as a result led to numerous different centromeric haplotypes (Fig. 3, Supplementary Figs. 5–9 and Extended Data Figs. 3 and 5).

https://doi.org/10.1038/s41588-024-01715-9

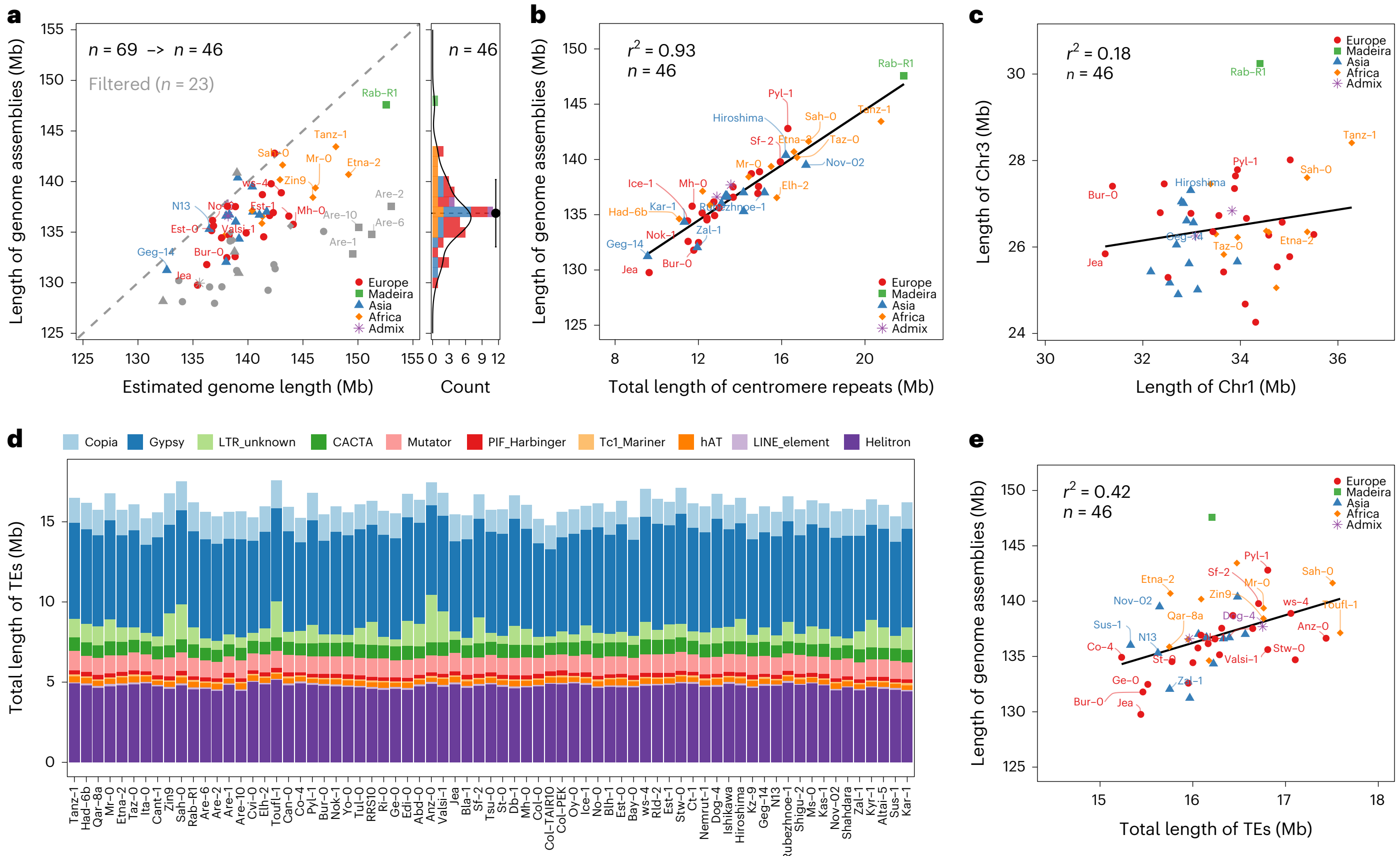


Fig. 2 | De novo assembly and annotation of 69 *A. thaliana* accessions. **a**, An assessment of the completeness of the 69 genome assemblies based on the comparison of estimated genome sizes and assembly features. The genome assemblies of 23 accessions were filtered out for genome size analyses based on the completeness ratio of genome and centromere assembly. **b**, A correlation analysis of genome assembly sizes and centromeric repeats. **c**, An example of a correlation analysis of assembly sizes of individual centromeres. **d**, The total length and composition of repetitive sequences of the genome assembly. **e**, A correlation analysis between the total lengths of genome assemblies and the TEs. Accessions are colored according to their genetic classification. Source data are provided as a source data file.

To quantify colinearity with higher resolution, we measured colinearity in sliding windows along the chromosomes and calculated average pairwise diversity in synteny of each of the newly assembled genomes against a recent genome assembly of Col-0 (ref. 57) (Col-PEK), which includes most of the centromeric regions (Fig. 4a). Average pairwise diversity in synteny ranges from 0 to 1, with 1 denoting complete absence of colinearity between the genome of a group and 0 indicating colinearity among all the genomes. Overall, for the 69 genomes, around 50% of the genome was highly colinear, with an average pairwise diversity in synteny lower than 0.2, which was exclusive to the chromosome arms (Fig. 4a and Supplementary Fig. 15). In contrast, ~33% of the genome was highly diverse with an average pairwise diversity in synteny of over 0.5, mostly including regions in and near the centromeres (Fig. 4a and Supplementary Fig. 15). We observed transitions between regions with very high and very low synteny in the peri-centromeres, which covered several Mb around each of the centromeres.

The broad colinearity in the chromosome arms showed a few interesting exceptions. The short arms of chromosomes 2 and 4, where rDNA clusters and nucleolus organizer regions are located, showed high levels of rearrangements, like the lower arms of chromosome 1 (~25 Mb) and 5 (~20 Mb) where large and highly diverse resistance *R* gene clusters reside (Fig. 4a)[58]. Also, the ~1.2 Mb inversion on chromosome 4 was marked by high diversity in synteny consistent with the fact that it segregates in several groups (Fig. 4a). In addition, we found individual spikes of high structural diversity in regions that were otherwise highly colinear. As previously described, such local hotspots of rearrangements were enriched for *R* gene clusters[3]. For example, at ~30 Mb on chromosome 1, synteny was broken by the high and variable copy number changes in a nucleotide-binding and leucine-rich repeat (NLR) gene cluster (Fig. 4b).

In addition, pairwise genome-wide colinear relationships reflected the genetic and geographical groups (Fig. 4c–e and Supplementary Fig. 16), implying that structural differences can recapitulate our genetic grouping based on SNPs (Extended Data Fig. 1 and Supplementary Fig. 17), which was also recently shown with read alignment-based SV calls[38]. Small subgroups of accessions exhibited increased colinearity. These corresponded to geographical clusters, including the Japanese accessions (Hiroshima and Ishikawa), the North American accessions (Yo-0, RRS10 and Tul-0) and the Madeiran accessions (Are-1, Are-2, Are-6, Are-10 and Rab-R1). Interestingly, we also found that the colinearity among African accessions was lower than the colinearity between European or Asian accessions (Fig. 4c–e), probably reflecting the higher genomic diversity in Africa[36].

The pan-genome of *A. thaliana*

We annotated 27,246 to 28,989 protein-coding genes in each of the 69 assemblies (Fig. 5a), compared to 27,445 genes in the reference sequence[59]. To unravel the gene repertoire of *A. thaliana*, we clustered all 1,928,005 genes combined with the gene sets of the reference sequence (Col-0 and Araport11), an additional recent telomere-to-telomere (T2T) Col-0 assembly (Col-PEK) and the reference sequences of the sister species *A. lyrata* and *Capsella rubella*, which served as outgroups. In

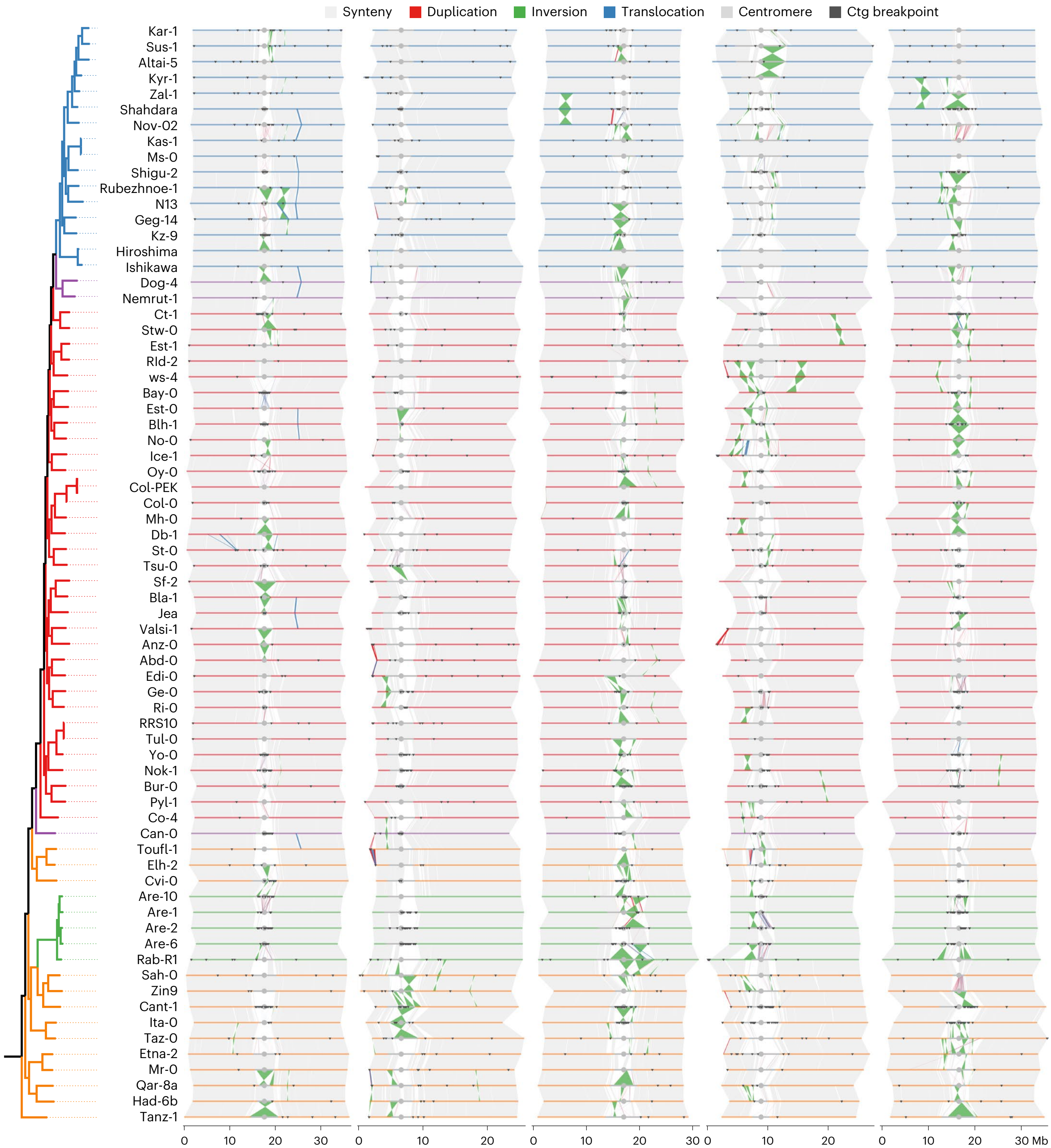


Fig. 3 | Whole-genome alignments of 69 *A. thaliana* accessions. The chromosomes of each accession were represented by segments, which are colored in line with the genetic classification. The gray segments represent the centromeric regions, and the middle points are indicated by gray circles. The syntenic regions between chromosomes are shown in light gray. The structural rearrangements, including duplications, inversions and translocations, are colored red, green and blue, respectively. Contig (ctg) breakpoints are marked by black triangles. Source data are provided as a source data file.

total, we identified 36,991 gene families across the 73 genomes, including 13,328 single-copy gene families that were used for constructing the phylogeny of the *A. thaliana* accessions (Supplementary Fig. 1).

Excluding the genes from the reference genome, the T2T Col-0 accession, *A. lyrata* and *C. rubella*, we found that 32,986 gene families included genes from at least one of the 69 *A. thaliana* genomes. Of those, 19,721 (60%) were present in all 69 genomes and were defined as core gene families. Among the gene families that appear nonessential, 1,613 (5% of total) were present in 63–69 accessions (>90% of the accessions), defined as softcore; 5,582 (17%) were present in 2–62

https://doi.org/10.1038/s41588-024-01715-9

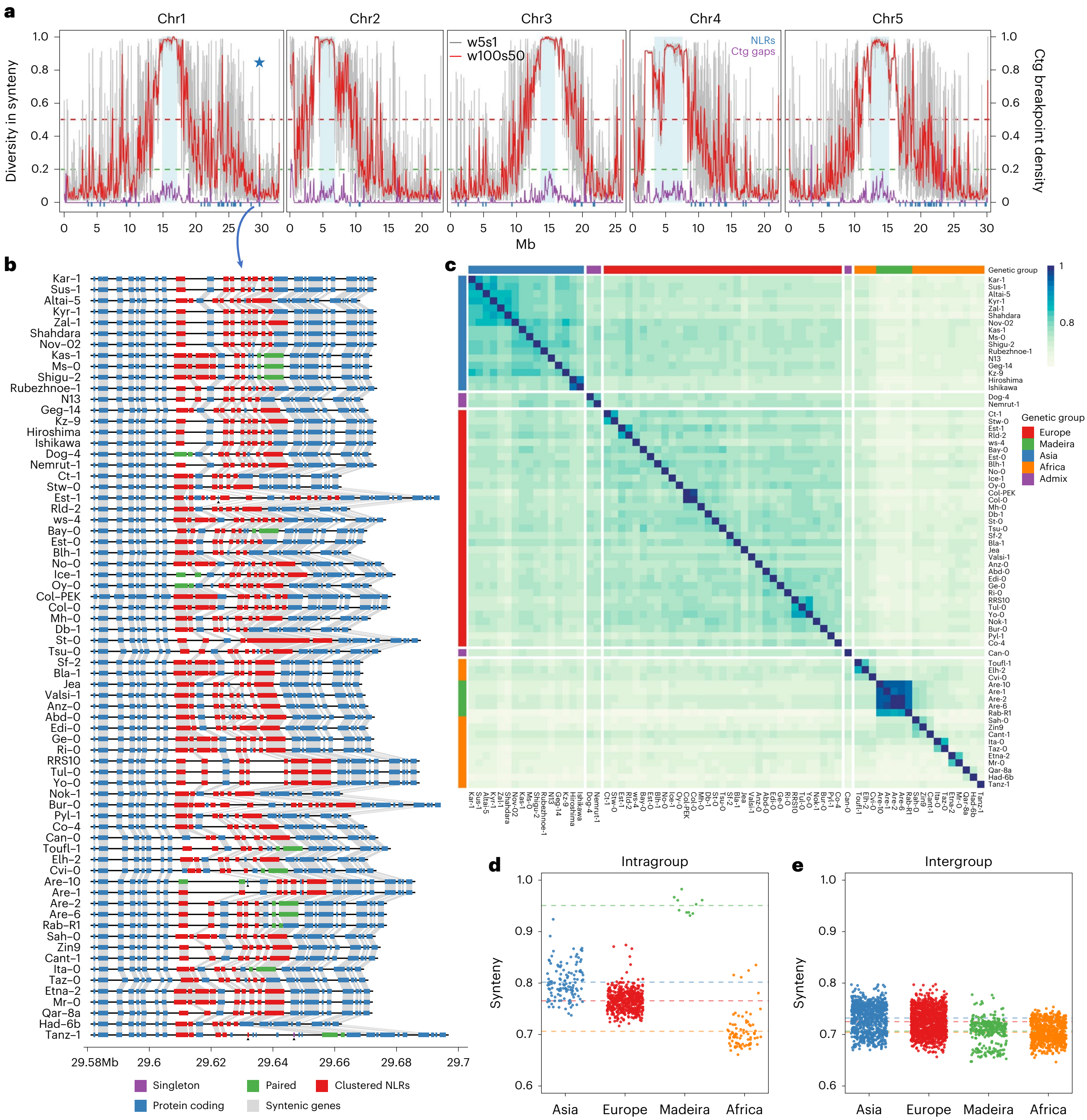


Fig. 4 | Characterization of the synteny landscape of 69 *A. thaliana* accessions. a, The diversity in synteny and density of contig (ctg) breakpoints along the five chromosomes. The diversity of synteny is shown at two resolutions, in red and gray (sliding window: 100 kb window size with 50 kb step size in red; 5 kb window size with 1 kb step size in gray). The density of contig breakpoints is shown in purple (100 kb window size with 50 kb step size). The blue bar indicates NLR genes. The horizontal dashed green and red lines indicate thresholds for synteny diversity values of 0.25 and 0.50. Mb, megabases. **b**, The local syntenic gene order in a highly divergent region of chromosome 1. The protein-coding genes and NLRs (singleton, pair and cluster) are presented by blue, purple, green and red rectangles. The gray links between the rectangles indicate homologous relationships. The private genes were marked by black triangles. **c**, Pairwise synteny relationships measurement along chromosome arms. **d**,**e**, A comparison of synteny relationships within (**d**) and between (**e**) groups. Source data are provided as a source data file.

genomes, defined as dispensable; and the remaining 6,070 (18%) gene families were present in only one genome, defined as private gene families (Fig. 5b and Extended Data Fig. 6). On average, a single genome consisted of 86.8% core, 6.7% softcore, 6.1% dispensable and 0.4% private genes (Fig. 5a). As expected, the increased size of our collection reduced the number of core gene families as compared to a previous estimation, which was based on eight assemblies only[3]. However, our core gene family number estimate was similar

 https://doi.org/10.1038/s41588-024-01715-9

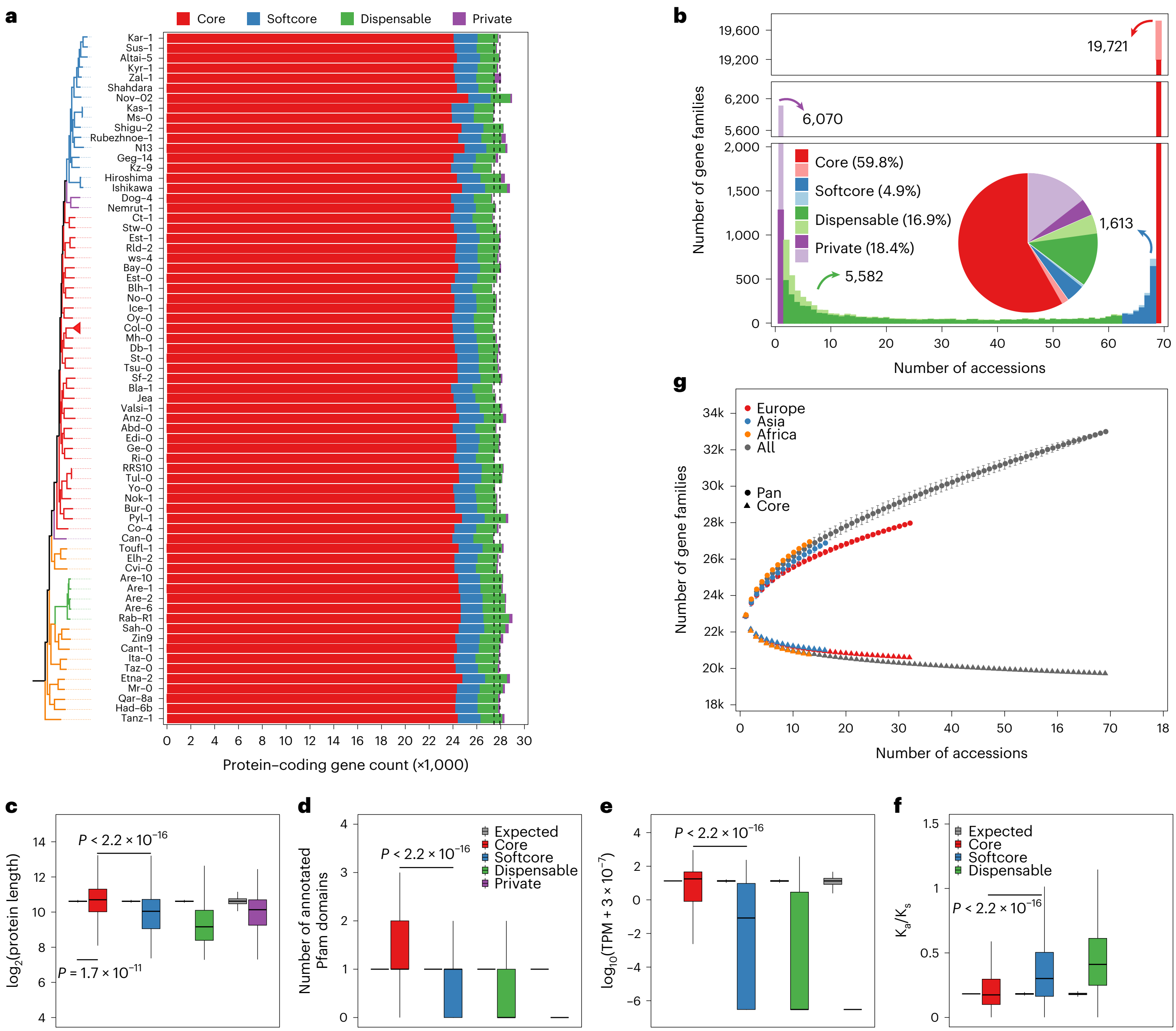


Fig. 5 | Pan-genome analysis of 69 *A. thaliana* accessions. a, The annotated protein-coding genes, and composition of core (red), softcore (blue), dispensable (green), and private (purple) genes in each individual accession. **b**, The number and proportion of core (red), softcore (blue), dispensable (green) and private (purple) gene families in the pan-genome. The split–merge cases are indicated by light colors. **c**–**f**, The protein length (**c**), number of annotated Pfam domains (**d**), gene expression landscape (**e**) and K_a/K_s (**f**) of core (red), softcore (blue), dispensable (green) and private (purple) genes of Col-0. The expression level of the gene is defined as the median value in the 79 organs and developmental stages. The expected dataset was made by 1,000 simulations with the same data size of the testing dataset. The two-sided Mann–Whitney test was performed, and the *P* value is indicated respectively. Intervals for boxplots: center, median (50th percentile); lower bounds of box, 25th percentile (Q1); upper bounds of box, 75th percentile (Q3); lower whisker, maximum of (minima, Q1 − 1.5 × IQR); upper whisker, minimum of (maxima, Q3 + 1.5 × IQR). IQR, interquartile range (range of Q1 to Q3). **g**, The increase of the pan-genome size and the decrease of core-genome size in the whole population (gray), Europe (red), Asia (blue) and Africa (orange) groups. Accessions were sampled as 2,000 random combinations of each given number of accessions ranging from 2 to 67. The mean number of gene families is shown with the standard deviation. Source data are provided as a source data file.

to one of a recent study where 21,545 gene families were annotated as core gene families among 32 accessions[46], indicating that our estimation probably reflects the actual core gene set of *A. thaliana*. In contrast, however, we found a much higher number of private genes in the genomes (18.4% of all gene families) as compared to the previous study. To understand the origin of the private genes, we explored local sequence homology and structure of all gene families, and found that 6,954 gene families were formed by differences in annotation, where gene models are split or merged differently in association with local polymorphisms (Methods). The split–merge cases represent 2.7%, 11%, 26.2% and 78.9% of the core, softcore, dispensable and private gene families, respectively (Fig. 5b, pastel colors). Even though these gene families could result from errors in the annotation, split–merge cases may also represent differences in the transcriptomes with functional consequences. For all the remaining 1,281 private gene families, no sequence similarity can be detected in the other accessions, suggesting that they represent de novo evolved genes.

We found that the length of the encoded protein sequences of the core genes was longer and more often matched Pfam domains as compared to other types of gene (Mann–Whitney test between core and softcore genes, $P < 2.2 \times 10^{-16}$, Fig. 5c-d). Gene expression across 79 organs and developmental stages (measured in Col-0 (median ≥1 transcript per million (TPM))), revealed a much higher fraction of expressed genes with significantly higher expression levels in the core and softcore gene families as compared to the private genes (even though 30.8% of the Col-0 dispensable genes (488/1,584) and 10% of the Col-0 private genes (3/31) were still expressed) (Mann–Whitney test between core and softcore genes, $P < 2.2 \times 10^{-16}$, Fig. 5e). Moreover, core genes showed significantly lower nonsynonymous/synonymous substitution ratios (K_a/K_s) than accessory genes, suggesting that core genes were more functionally constrained than accessory genes (Mann–Whitney test between core and softcore genes, $P < 2.2 \times 10^{-16}$, Fig. 5f). This is corroborated by the functional categories of the different types of gene family that were analyzed with a Gene Ontology (GO) enrichment analysis. Core genes were enriched for basic and essential biological processes, including metabolic, cellular, developmental processes, reproduction and regulation of biological process (Supplementary Fig. 18 and Supplementary Table 8). Among them, the 27 meiosis-specific genes[60] were all highly conserved, and were present in the same copy numbers (26 in one copy, and 1 in two copies) across the 69 accessions (Supplementary Table 9). Accessory genes (softcore, dispensable and private categories, analyzed independently or together) were enriched for biological processes such as cell killing and defense response (Supplementary Fig. 19 and Supplementary Table 10). Altogether, while this suggests that the accessory genes are substantially different from the core genes, it also suggests that a fraction of those have functional features and could contribute to phenotypic diversity and adaptation.

Finally, we measured the sizes of the pan-gene sets of the 69 *A. thaliana* accessions by subsetting the genomes (Fig. 5g and Supplementary Fig. 20). Even after adding all genomes, the pan-genome gene set did not reach a plateau, indicating that our set of accessions did not capture the entire complement of diverse gene families in this species. The reason for this was the high number of gene families that were only present in one or a few of the accessions ('dispensable' and 'private' gene families), hinting at an untapped genetic diversity.

Discussion

In this study, we have generated 69 reference-quality genome assemblies, which capture a large degree of the genetic diversity in *A. thaliana*. These assemblies were generated from accessions selected from Central Africa to Iceland and from North America to Japan, but despite these huge geographical distances, genome structure was highly conserved among plants. The assemblies also revealed a total of 10,420 novel protein-coding gene clusters, which are absent from the reference genome (Col-0 and Araport11) and provide a very powerful resource to study the genetic basis of hitherto undescribed variation. In addition, our collection of genome assemblies contains the parental strains of powerful and publicly available material such as recombinant inbred lines[61,62]. Genome assemblies will help to unravel the genetic basis of important traits relying on complex structural variations[63–66]. Finally, these 69 genomes, together with others, provide a great resource to study the mechanisms of genome dynamics, including recombination. These resources pave the way for further functional genomic investigation.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-01715-9.

References

1. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
2. Alonge, M. et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145–161 e23 (2020).
3. Jiao, W. B. & Schneeberger, K. Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat. Commun.* **11**, 989 (2020).
4. Lian, Q. et al. The megabase-scale crossover landscape is largely independent of sequence divergence. *Nat. Commun.* **13**, 3828 (2022).
5. Zapata, L. et al. Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms. *Proc. Natl Acad. Sci. USA* **113**, E4052–E4060 (2016).
6. Capilla-Perez, L. et al. The synaptonemal complex imposes crossover interference and heterochiasmy in Arabidopsis. *Proc. Natl Acad. Sci. USA* **118**, e2023613118 (2021).
7. Durand, S. et al. Joint control of meiotic crossover patterning by the synaptonemal complex and HEI10 dosage. *Nat. Commun.* **13**, 5999 (2022).
8. Schmidt, C. et al. Changing local recombination patterns in *Arabidopsis* by CRISPR/Cas mediated chromosome engineering. *Nat. Commun.* **11**, 4418 (2020).
9. Lowry, D. B. & Willis, J. H. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol.* **8**, e1000500 (2010).
10. Lamichhaney, S. et al. Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). *Nat. Genet.* **48**, 84–88 (2016).
11. Harringmeyer, O. S. & Hoekstra, H. E. Chromosomal inversion polymorphisms shape the genomic landscape of deer mice. *Nat. Ecol. Evol.* **6**, 1965–1979 (2022).
12. Tong, X. et al. High-resolution silkworm pan-genome provides genetic insights into artificial selection and ecological adaptation. *Nat. Commun.* **13**, 5619 (2022).
13. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
14. Nattestad, M. & Schatz, M. C. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021–3023 (2016).
15. Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J. & Edwards, D. Plant pan-genomes are the new reference. *Nat. Plants* **6**, 914–920 (2020).
16. De Coster, W., Weissensteiner, M. H. & Sedlazeck, F. J. Towards population-scale long-read sequencing. *Nat. Rev. Genet.* **22**, 572–587 (2021).
17. Della Coletta, R., Qiu, Y., Ou, S., Hufford, M. B. & Hirsch, C. N. How the pan-genome is changing crop genomics and improvement. *Genome Biol.* **22**, 3 (2021).
18. Jayakodi, M., Schreiber, M., Stein, N. & Mascher, M. Building pan-genome infrastructures for crop plants and their use in association genetics. *DNA Res.* **28**, dsaa030 (2021).
19. Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176 e13 (2020).
20. Gao, L. et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* **51**, 1044–1051 (2019).
21. Zhou, Y. et al. Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* **606**, 527–534 (2022).
22. Tang, D. et al. Genome evolution and diversity of wild and cultivated potatoes. *Nature* **606**, 535–541 (2022).

23. Shang, L. et al. A super pan-genomic landscape of rice. *Cell Res.* **32**, 878–896 (2022).
24. Zhang, F. et al. Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes. *Genome Res.* **32**, 853–863 (2022).
25. Qin, P. et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542–3558 e16 (2021).
26. Hufford, M. B. et al. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, 655–662 (2021).
27. Jayakodi, M. et al. The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature* **588**, 284–289 (2020).
28. Walkowiak, S. et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature* **588**, 277–283 (2020).
29. Sun, X. et al. Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nat. Genet.* **52**, 1423–1432 (2020).
30. Liao, W. W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
31. Vollger, M. R. et al. Increased mutation and gene conversion within human segmental duplications. *Nature* **617**, 325–334 (2023).
32. Initiative, A. G. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
33. Cao, J. et al. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43**, 956–963 (2011).
34. Gan, X. et al. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **477**, 419–423 (2011).
35. The 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**, 481–491 (2016).
36. Durvasula, A. et al. African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **114**, 5213–5218 (2017).
37. Zou, Y. P. et al. Adaptation of *Arabidopsis thaliana* to the Yangtze River basin. *Genome Biol.* **18**, 239 (2017).
38. Goktay, M., Fulgione, A. & Hancock, A. M. A new catalog of structural variants in 1,301 *A. thaliana* lines from Africa, Eurasia, and North America reveals a signature of balancing selection at defense response genes. *Mol. Biol. Evol.* **38**, 1498–1511 (2021).
39. Horton, M. W. et al. Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat. Genet.* **44**, 212–216 (2012).
40. Frachon, L. et al. Intermediate degrees of synergistic pleiotropy drive adaptive evolution in ecological time. *Nat. Ecol. Evol.* **1**, 1551–1561 (2017).
41. Fransz, P. et al. Molecular, genetic and evolutionary analysis of a paracentric inversion in *Arabidopsis thaliana*. *Plant J.* **88**, 159–178 (2016).
42. Barragan, A. C. et al. A truncated singleton NLR causes hybrid necrosis in *Arabidopsis thaliana*. *Mol. Biol. Evol.* **38**, 557–574 (2021).
43. Michael, T. P. et al. High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *Nat. Commun.* **9**, 541 (2018).
44. Pucker, B. et al. A chromosome-level sequence assembly reveals the structure of the *Arabidopsis thaliana* Nd-1 genome and its gene set. *PLoS ONE* **14**, e0216233 (2019).
45. Rabanal, F. A. et al. Pushing the limits of HiFi assemblies reveals centromere diversity between two *Arabidopsis thaliana* genomes. *Nucleic Acids Res.* **50**, 12309–12327 (2022).
46. Kang, M. et al. The pan-genome and local adaptation of *Arabidopsis thaliana*. *Nat. Commun.* **14**, 6259 (2023).
47. Hagmann, J. et al. Century-scale methylome stability in a recently diverged *Arabidopsis thaliana* lineage. *PLoS Genet.* **11**, e1004920 (2015).
48. Anastasio, A. E. et al. Source verification of mis-identified *Arabidopsis thaliana* accessions. *Plant J.* **67**, 554–566 (2011).
49. Simon, M. et al. DNA fingerprinting and new tools for fine-scale discrimination of *Arabidopsis thaliana* accessions. *Plant J.* **69**, 1094–1101 (2012).
50. Long, Q. et al. Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nat. Genet.* **45**, 884–890 (2013).
51. Sun, H., Ding, J., Piednoel, M. & Schneeberger, K. findGSE: estimating genome size variation within human and *Arabidopsis* using *k*-mer frequencies. *Bioinformatics* **34**, 550–557 (2018).
52. Wlodzimierz, P. et al. Cycles of satellite and transposon evolution in *Arabidopsis* centromeres. *Nature* **618**, 557–565 (2023).
53. Willing, E. M. et al. Genome expansion of *Arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation. *Nat. Plants* **1**, 14023 (2015).
54. Murat, F. et al. Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome Biol.* **16**, 262 (2015).
55. Schranz, M. E., Lysak, M. A. & Mitchell-Olds, T. The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. *Trends Plant Sci.* **11**, 535–542 (2006).
56. Hu, T. T. et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481 (2011).
57. Hou, X., Wang, D., Cheng, Z., Wang, Y. & Jiao, Y. A near-complete assembly of an *Arabidopsis thaliana* genome. *Mol. Plant* **15**, 1247–1250 (2022).
58. Van de Weyer, A. L. et al. A species-wide inventory of NLR genes and alleles in *Arabidopsis thaliana*. *Cell* **178**, 1260–1272 e14 (2019).
59. Cheng, C. Y. et al. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* **89**, 789–804 (2017).
60. Thangavel, G., Hofstatter, P. G., Mercier, R. & Marques, A. Tracing the evolution of the plant meiotic molecular machinery. *Plant Reprod.* **36**, 73–95 (2023).
61. Simon, M. et al. Quantitative trait loci mapping in five new large recombinant inbred line populations of *Arabidopsis thaliana* genotyped with consensus single-nucleotide polymorphism markers. *Genetics* **178**, 2253–2264 (2008).
62. Loudet, O., Chaillou, S., Camilleri, C., Bouchez, D. & Daniel-Vedele, F. Bay-0 x Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in *Arabidopsis*. *Theor. Appl. Genet.* **104**, 1173–1184 (2002).
63. Durand, S., Bouche, N., Perez Strand, E., Loudet, O. & Camilleri, C. Rapid establishment of genetic incompatibility through natural epigenetic variation. *Curr. Biol.* **22**, 326–331 (2012).
64. Bikard, D. et al. Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science* **323**, 623–626 (2009).
65. Smith, L. M., Bomblies, K. & Weigel, D. Complex evolutionary events at a tandem cluster of *Arabidopsis thaliana* genes resulting in a single-locus genetic incompatibility. *PLoS Genet.* **7**, e1002164 (2011).
66. Demirjian, C. et al. An atypical NLR gene confers bacterial wilt susceptibility in *Arabidopsis*. *Plant Commun.* **4**, 100607 (2023).

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

https://doi.org/10.1038/s41588-024-01715-9

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2024

Methods

Plant material and whole-genome sequencing

We ordered the seeds of 66 accessions from the Versailles Arabidopsis Stock Center at Jean-Pierre Bourgin Institute and received the seeds of the remaining 6 accessions (Elh-2, Ice-1, Rab-R1, Tanz-1, Taz-0 and Zin9) from Angela Hancock (MPIPZ). The accession numbers of seeds from Versailles are presented in Supplementary Table 1. Plants were grown in greenhouses or growth chambers.

For the 48 accessions that were sequenced with PacBio HiFi, high-molecular-weight (HMW) DNA was prepared from a pool of 30–40 4-week-old plants using the NucleoBond HMW DNA kit. The HiFi libraries were prepared using the SMRTbell prep kit 3.0 and the BluePippin cartridge for enriching fragments greater than 9 kb up to 50 kb. Finally, HiFi sequencing was performed on the Sequel IIe platform at the Max Planck Genome-centre Cologne. SMRTlink software (PacBio) was used to demultiplex and extract HiFi datasets. DNA form a single plant leave was used to prepare PCR-free short-read libraries according to the protocol of the NEBNext Ultra II DNA PCR-free Library Prep Kit for Illumina (New England Biolabs). Libraries were then sequenced with 2 × 150 paired-end reads on the NextSeq 2000 platform.

As previously described[67], for the 24 accessions that were sequenced with Oxford Nanopore, HMW DNA was extracted from 3-week-old plants according to the protocol described by Russo et al.[68]. The subsequent library preparation and sequencing were performed at the GeT-PlaGe core facility, INRAE Toulouse. ONT libraries were prepared using the EXP-NBD103 and SQK-LSK109 kits according to the manufacturer's instructions and using 4 µg of 40 kb-sheared DNA (Megaruptor, Diagenode) as input. Pools of six samples were sequenced on one R9.4.1 flowcell. Between 14 and 20 fmol of library was loaded on each flowcell and sequenced on a PromethION instrument. Illumina libraries were prepared using the Illumina TruSeq Nano DNA HT Library Prep Kit. Libraries were then sequenced with 2 × 150 bp paired-end reads on the Hiseq3000 platform.

De novo genome assembly

Genome heterozygosity and size were estimated using Jellyfish v2.2.6 (ref. 69) and findGSE[51].

For the 48 accessions that were sequenced with PacBio HiFi, the initial de novo assembly was performed using three different assembly tools: Canu v2.1.1 (ref. 70), Flye v2.7 (ref. 71) and Hifiasm v0.16.1 (ref. 72). Then, purge_dups v1.2.5 (ref. 73) was used to purge haplotigs and overlaps in the assemblies from Canu and Hifiasm. To improve contiguity, we combined the assemblies derived from the three assemblers using quickmerge v0.3 (ref. 74). First, the assembly with the best quality (longest contiguity measured as N50, assembly size and correctness, centromere coverage and so on) was selected as query, and the assembly with second longest N50 was used as reference to join contigs in the query assembly (Supplementary Table 2). The resulting assembly was further improved by using the third assembly as reference. Then, we used a homology-based scaffolding tool, RaGOO v1.1 (ref. 75), to order and orient contigs on the basis of whole-genome alignments to the Col-CEN genome[76]. Manual evaluation and correction were performed based on the whole-genome alignment and position of centromeric repeats. Finally, to close the gaps in scaffolds, we ran four rounds of MaSuRCA v4.1.0 (ref. 77) by using the HiFi reads and the three assemblies individually.

For the 24 accessions that were sequenced with Oxford Nanopore, the long reads were filtered for adapters, short (<1 kb) or low-quality reads (mean quality >70) using Porechop v0.2.4 (https://github.com/rrwick/Porechop) and Filtlong v0.2.1 (https://github.com/rrwick/Filtlong). De novo assembly of each genome was initially performed using SMARTdenovo (https://github.com/ruanjue/smartdenovo)[78] and Flye. To fix base errors in the initial assemblies, we polished the genome by running three rounds of Racon v1.4.10 (ref. 79) with long reads, and four rounds of NextPolish v1.3.1 (ref. 80) with short reads. Then, assemblies were purged using purge_dups. Similar as for the process described above for HiFi datasets, assemblies were further improved, corrected and scaffolded (Supplementary Table 3). Finally, scaffolds were polished using NextPolish.

Genome assembly evaluation

To evaluate the completeness of each genome assembly, compleasm v0.2.2 (ref. 81) was used with the OrthoDB database brassicales_odb10 (brassicales, 2020-08-05). We evaluated the consensus quality value (QV) and completeness of genome assemblies on the basis of the *k*-mers spectrum of Illumina whole-genome sequencing reads, using Merqury v1.3 (ref. 82) with default parameters, to estimate the goodness and completeness of the reference protein-coding genes (TAIR10 and Araport11) in each genome assembly. The reference genes were aligned against the genome assemblies using blastn[83], and reference genes were considered well assembled in genome assemblies which were aligned with identity ≥80 and coverage ≥0.9. We also used Liftoff v1.6.3 (ref. 84) to 'lift over' the reference genes to the 69 genome assemblies, with parameters '-copies -sc 0.90 -polish'. Additionally, to evaluate the assembly continuity, LTR Assembly Index (LAI) was calculated for each genome assembly using LTR_retriever v2.9.0 (ref. 85).

Centromeric and telomeric repeats were annotated using Bowtie2 v2.4.4 (ref. 86) (-a–very-sensitive) to search for the consensus sequence of the 178-bp and 7-bp repeat motifs in each genome assembly, respectively. To estimate the copy number and length of centromeric repeats, we compared the sequencing depths obtained from aligning Illumina short reads against the genome assembly and concatenated sequence of four copies of the repeat motif, using Bowtie2 (-k 1), separately. The sequencing depth was calculated using samtools v1.9 (ref. 87) with the parameter setting '-Q 1 -d 0 -a'.

We evaluated the assembly quality in the following six aspects, and found that (1) the assembled genome size is comparable to the length of T2T assembly of Col-0 reported by recent studies[57,76,88] (Fig. 2a, Supplementary Figs. 3 and 21, and Supplementary Table 4); (2) the completeness estimated by Benchmarking Universal Single-Copy Orthologs was 99.8%, which is comparable to the Col-0 reference and T2T genomes (Supplementary Fig. 2 and Supplementary Table 4); (3) on the basis of the *k*-mer-based estimation, the assembled genomes showed a mean of 53.4 QV and 98.5% completeness (Extended Data Fig. 2 and Supplementary Table 4); (4) the analysis of completeness of reference protein-coding genes by homology search against the assembled genomes (BLASTP and Liftoff), a mean of 97.3% and 98.4% were successfully assembled (Supplementary Table 5); (5) on the basis of the LAI (mean of 22), which reached the 'gold standard' level (LAI >20) (Supplementary Table 6); (6) on the basis of the completeness of centromere repeats (mean of 96%), estimated using the Illumina short-read dataset (Extended Data Fig. 2 and Supplementary Table 7). Three accessions, Lu-1, Pa-1 and Istisu-1, had lower values in the measurement of QV and *k*-mer completeness. Further alternative allele frequency analysis indicated the presence of heterozygosity, which are unexpected in a pure line, probably due to the mixture of two distinct lineages or segregation of polymorphism in the population. These three accessions were ignored in subsequent analyses (Supplemental Fig. 28). These results suggested that the quality of all 69 genome assemblies was comparable to that achieved by the Col-0 reference and T2T genome assemblies, indicating high continuity and completeness.

Annotation of repetitive elements

To annotate the TEs in the 69 genome assemblies, we first generated a nonredundant TE library for each accession, using the Extensive de novo TE Annotator (EDTA) v2.0.1 (ref. 89) with parameters '–overwrite 1–sensitive 1–anno 1–evaluate 1'. Then, all the individual TE libraries were combined to construct a pan-TE library using panEDTA[90]. RepeatMasker v4.1.1 (http://www.repeatmasker.org) was further

https://doi.org/10.1038/s41588-024-01715-9

employed to re-annotate the repeat regions with parameters '-q -div 40 -cutoff 225' and the pan-TE library.

Gene prediction and annotation

Protein-coding genes were annotated on the basis of a strategy that integrated ab initio gene prediction, transcriptome-based de novo transcript assembly and homologous protein sequence alignment. First, four ab initio gene prediction tools were used: Augustus v3.3.3 (ref. 91), GeneMark v4.62 (ref. 92), GlimmerHMM v3.0.4 (ref. 93) and SNAP (version 2006-07-28)[94]. Second, we collected a list of 308 public RNA sequencing (RNA-seq) datasets for 28 accessions from the NCBI SRA database (Supplementary Table 11). The quality of short reads was checked with FastQC. Trimmomatic v0.39 (ref. 95) was used to remove potential adapter and low-quality sequences, with parameters 'LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36' for single-end reads and 'LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36' for paired-end reads. To obtain the protein sequence of the transcript, the reads were then processed with Trinity v2.14.0 (ref. 96) to assembly transcript sequences, and TransDecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder) to identify candidate coding regions. The predicted longest open reading frames were searched against the UniPort database[97] using DIAMOND v2.0.4 (ref. 98) with the parameter setting '-k 1 -f 6 -e 1e-5–ultra-sensitive', and Pfam database[99] using hmmsearch from HMMER v3.1b2 (http://hmmer.org). Then, we merged all the predicted protein sequences to generate the pan-pep library and selected representative sequences using CD-HIT v4.6.8 (refs. 100,101) (-c 0.98). Third, protein sequences of *A. thaliana* (447_Araport11), *A. lyrata* (384_v2.1), *Oryza sativa* (323_v7.0) and *Solanum lycopersicum* (514_ITAG3.2), which were downloaded from Phytozome v13 database[102] and the pan-pep library, were aligned to each genome assembly using Exonerate v2.2.0 (ref. 103) (–percent 70–minintron 10–maxintron 60000). Finally, all different evidences of gene models were integrated using EVidenceModeler v1.1.1 (ref. 104). The resulting gene models, especially for those from nonscaffold contigs, were further evaluated by comparing to the National Center for Biotechnology Information (NCBI) nonredundant (NR) database using DIAMOND, outside which Brassicaceae proteins were excluded.

Noncoding genes were annotated by integrating the predictions from Barrnap v0.9 (https://github.com/tseemann/barrnap), Infernal v1.1.4 (ref. 105) (Rfam database v14.8) and tRNAscan-SE v2.0.9 (ref. 106). Noncoding RNA and TE-related genes were identified by checking alignment/overlap between predicted gene models and TE sequences (TAIR10), representative gene models (Araport11), TE genes (Araport11), TE and noncoding RNA annotations of each assembly.

Disease resistance genes were identified by using NLR-Annotator v2.1 (ref. 107) and RGAugury[108]. NLRs have been reported to be mostly present in pairs or cluster[23,58]. Pair NLRs were defined as fewer than two non-NLR genes between the two NLRs. Cluster NLRs were defined as more than two NLRs with fewer than two non-NLR genes between any two NLRs. The remaining NLRs were labeled as singletons.

The resulting gene models were further annotated functionally using InterProScan v5.59-91.0 (ref. 109) (parameters: -f TSV -t p -iprlookup -goterms -pa). The GO enrichment analysis was performed using AgriGO v2.0 (ref. 110).

Gene-based pan-genome construction and analysis

All the protein-coding genes from the 69 assembled genomes, representative protein-coding genes from Col-0 TAIR10 (refs. 32,111), Col-PEK[57] and two out-species, *A. lyrata* (384_v2.1) and *C. rubella* (474_v1.1)[102], were clustered using OrthoFinder v2.5.4 (ref. 112) with the parameter setting '-S diamond_ultra_sens', resulting 37,921 gene clusters for 73 genomes. We used Liftoff to 'lift over' all the predicted protein-coding genes (including genes from the Col-0 TAIR10) to each of the 69 genome assemblies, with parameters '-copies -sc 0.90 -polish'. The gene locus, that is, low allele frequency, in each genome was checked for the presence of homologous genes (95% coverage for both query and hit genes, and reciprocal best hit) from the other accessions, and then, the related orthogroups were fused. Furthermore, the potential split–merge cases were also evaluated for the alignment and coverage between the query gene and representative gene (across the 69 accessions) or TAIR10 reference genes. After correction, we obtained 36,991 gene clusters for 73 genomes, and 32,986 gene clusters for the 69 assembled genomes. The orthologous groups were then classified into four categories: core gene clusters were defined as the genes shared between all the 69 genomes; softcore gene clusters were present in more than 90% of accessions (63–68); dispensable gene clusters were found in more than one accession (2–62); and private gene clusters that were accession specific. To estimate the pan-genome and core-genome size (the number of gene families defined by OrthoFinder), we carried out 2,000 random samplings of accessions for each number of sample size (ranging from 2 to 67) from the 69 accessions.

Detection of SNPs and indels

The Illumina whole-genome sequencing short-reads of the 72 accessions (with mean depths of 43×) were aligned against the Col-PEK genome by BWA v0.7.15-r1140 with default parameters, and duplicated reads were removed using samtools. Then, SNPs and small indels (ranging from 1 to 20 bp) were detected and filtered for each genome assembly and merged by inGAP-family[113]. The resulting variants were further processed by VCFtools v0.1.16 (ref. 114) to obtain the high-quality and informative SNP list (parameters: '–maf 0.05–max-missing 0.2–min-alleles 2–max-alleles 2–min-meanDP 6–max-meanDP 226'). SNPs that were located in the region of centromeric repeats and TEs were removed. We found a total of 7,056,033 SNPs, among which 2,254,527 were common and located in noncentromeric and non-TE regions, with a density of one SNP per 59 bp.

SV detection and analysis

To fully take advantage of the 69 high-contiguity genome assemblies, we performed the whole-genome alignment against the Col-PEK genome using minimap2 v2.21-r1071 (refs. 115,116) (parameters: -ax asm5–eqx), and SyRI v1.6 (ref. 13) was applied to identify SVs with default parameters. SVs (20 bp to 10 kb) from the 69 genome assemblies were merged by SURVIVOR with the parameters '1000 1 1 0 0 1'. SVs longer than 10 kb from 69 accessions detected by SyRI were retained and merged by SURVIVOR (parameters: '20000 1 1 0 0 1').

Population and phylogenetic analysis

For the SNPs in the 72 *A. thaliana* accessions, linkage pruning was performed by PLINK v1.90b6.18 (ref. 117), with the parameters '–indep-pairwise 50 10 0.1'. Then, principal component analysis (PCA) was conducted by PLINK, which showed that PC1 splits Madeira and Africa accessions from the North America–Europe–Asia group, while PC2 further divides Asia accessions from the others (Fig. 1b). We performed population structure inference using ADMIXTURE v1.3.0 (ref. 118), with the number of population settings ranging from 2 to 5. For $K = 2$, we found a division between Africa and the other accessions. When $K = 4$, we saw a new subgroup (Madeira) within the Africa group. When $K = 5$, the subgroup Sicily and Lebanon emerged within the Africa group (Fig. 1c).

To build the phylogenetic tree of the 69 accessions, a total of 13,328 single-copy orthologous gene clusters were selected and used for generating the amino acid alignment using MUSCLE v3.8.31 (ref. 119) with default parameters. For the amino acid alignment of each ortholog group, the nucleotide sequence that corresponds with the amino acid sequence was extracted by seqkit v2.3.0 (ref. 120) with default parameters, and then the coding sequence (CDS) alignment was generated by PAL2NAL v14 (ref. 121) with the parameter '-nomismatch'. Then, a concatenated super matrix of the 13,328 ortholog-based CDS alignment was constructed with different partitions defined corresponding to different gene clusters. The super matrix and partition definition were used for building a maximal likelihood tree using IQ-TREE v1.6.12 (ref. 122) (parameters: -m MFP -bb 1000 -alrt 1000 -redo -safe). The SNP list of the 72 accessions was converted into FASTA format using vcf2phylip v2.8 (https://github.com/edgardomortiz/vcf2phylip), and then was taken by IQ-TREE to reconstruct the evolutionary tree (parameters: -m GTR + ASC -bb 1000 -alrt 1000 -redo -safe). The gene presence–absence variation matrix (phylip format) was generated and used for tree building by IQ-TREE (parameters: -st MORPH -m MK + ASC -bb 1000 -alrt 1000 -redo -safe).

To estimate population recombination rates ($\rho = 4N_e r$, where N_e is the effective population size and r is the recombination rate of the window), we used FastEPRR v2.0 (ref. 123) with 100-kb nonoverlapping window size. The nucleotide diversity of each site was calculated by VCFtools.

To calculate the K_a, K_s and K_a/K_s of each orthologous gene pair (*A. lyrata* as the reference), the amino acid sequences were aligned using MUSCLE, transformed into CDS alignments with PAL2NAL, and then fed into the KaKs_Calculator v2.0 (refs. 124,125).

Gene expression analysis

The RNA-seq dataset from a previous study[126], including 79 organs and developmental stages of *A. thaliana*, was downloaded from the NCBI SRA database. First, the potential adapter and low-quality sequences were identified and removed by Trimmomatic v0.39 (ref. 95), with parameters 'LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36'. Then, HISAT2 v2.1.0 (refs. 127,128) was used to align the clean reads against the Col-PEK genome. Gene expression was normalized as TPM, which was calculated by StringTie v2.0.6 (ref. 129) with default parameters.

Reporting summary

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

Data availability

The raw Illumina, PacBio HiFi and Oxford Nanopore sequencing data of the 72 accessions can be accessed in EMBL-ENA under the accession number PRJEB62038. The genome assemblies can be accessed in NCBI under the accession number PRJNA1033522. The data generated in this study can be accessed in Edmond (the Open Research Data Repository of the Max Planck Society, https://doi.org/10.17617/3.AEOJBL) (ref. 130), including genome assemblies (including Lu-1, Pa-1 and Istisu-1), gene and TE annotations, SNPs and SVs, pan-genome matrix and othrogroups. The RNA-seq dataset used in this study is downloaded from the NCBI SRA database (the accession numbers are included in Supplementary Table 11). The databases used in this study, including OrthoDB brassicales_odb10 (brassicales, 2020-08-05), NCBI NR database and Rfam database v14.8, are all public available. Source data are provided with this paper.

Code availability

The related code is available at GitHub (https://github.com/qclian/Pan_Ath) and Zenodo (https://doi.org/10.5281/zenodo.10567419) (ref. 131). All software used in the study are publicly available from the Internet as described in Methods and Reporting Summary.

References

67. Simon, M. et al. APOK3, a pollen killer antidote in *Arabidopsis thaliana*. *Genetics* **221**, iyac089 (2022).
68. Russo, A. et al. Low-input high-molecular-weight DNA extraction for long-read sequencing from plants of diverse families. *Front. Plant Sci.* **13**, 883897 (2022).
69. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
70. Nurk, S. et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
71. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
72. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
73. Guan, D. et al. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
74. Chakraborty, M., Baldwin-Brown, J. G., Long, A. D. & Emerson, J. J. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.* **44**, e147 (2016).
75. Alonge, M. et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* **20**, 224 (2019).
76. Naish, M. et al. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. *Science* **374**, eabi7489 (2021).
77. Zimin, A. V. et al. The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
78. Liu, H., Wu, S., Li, A. & Ruan, J. SMARTdenovo: a de novo assembler using long noisy reads. *GigaByte* **2021**, gigabyte15 (2021).
79. Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
80. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
81. Huang, N. & Li, H. compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics*, **39**, btad595 (2023).
82. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
83. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
84. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2020).
85. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
86. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
87. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
88. Wang, B. et al. High-quality *Arabidopsis thaliana* genome assembly with nanopore and HiFi long reads. *Genomics Proteom. Bioinform.* **20**, 4–13 (2022).
89. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
90. Ou, S. et al. Differences in activity and stability drive transposable element variation in tropical and temperate maize. Preprint at *bioRxiv* https://doi.org/10.1101/2022.10.09.511471 (2022).
91. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
92. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O. & Borodovsky, M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990 (2008).

93. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
94. Korf, I. Gene finding in novel genomes. *BMC Bioinform.* **5**, 59 (2004).
95. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
96. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
97. UniProt, C. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
98. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
99. Finn, R. D. et al. The Pfam protein families database. *Nucleic Acids Res.* **38**, D211–D222 (2010).
100. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
101. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
102. Goodstein, D. M. et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186 (2012).
103. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinform.* **6**, 31 (2005).
104. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
105. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
106. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).
107. Steuernagel, B. et al. The NLR-Annotator Tool enables annotation of the intracellular immune receptor repertoire. *Plant Physiol.* **183**, 468–482 (2020).
108. Li, P. et al. RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics* **17**, 852 (2016).
109. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
110. Tian, T. et al. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* **45**, W122–W129 (2017).
111. Lamesch, P. et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210 (2012).
112. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
113. Lian, Q., Chen, Y., Chang, F., Fu, Y. & Qi, J. inGAP-family: accurate detection of meiotic recombination loci and causal mutations by filtering out artificial variants due to genome complexities. *Genomics Proteom. Bioinform.* **20**, 524–535 (2022).
114. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
115. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
116. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
117. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
118. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655–1664 (2009).
119. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
120. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE* **11**, e0163962 (2016).
121. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).
122. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
123. Gao, F., Ming, C., Hu, W. & Li, H. New software for the fast estimation of population recombination rates (FastEPRR) in the genomic era. *G3* **6**, 1563–1571 (2016).
124. Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteom. Bioinform.* **8**, 77–80 (2010).
125. Zhang, Z. KaKs_Calculator 3.0: calculating selective pressure on coding and non-coding sequences. *Genomics Proteom. Bioinform.* **20**, 536–540 (2022).
126. Klepikova, A. V., Kasianov, A. S., Gerasimov, E. S., Logacheva, M. D. & Penin, A. A. A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *Plant J.* **88**, 1058–1070 (2016).
127. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
128. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
129. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
130. A pan-genome of 69 *Arabidopsis thaliana* accessions reveals a conserved genome structure throughout the global species range. *Edmond* https://doi.org/10.17617/3.AEOJBL (2024).
131. Lian, Q. The related code for a pan-genome of 69 *Arabidopsis thaliana* accessions. *Zenodo* https://doi.org/10.5281/zenodo.10567419 (2024).

Acknowledgements

We thank the Max Planck-Genome-centre Cologne (MP-GC) for DNA extraction, library preparation and sequencing, H. Sun and F. A. Rabanal for helpful discussions and comments, and N. Donnelly for proofreading the manuscript. We thank A. Hancock and The Versailles Arabidopsis Stock Center for providing genetic material. This work was supported by core funding from the Max Planck Society and an Alexander von Humboldt Fellowship to Q.L. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2048/1–390686111 (to K.S. and R.M.) and TRR 341/1–456082119 to K.S., as well as by the European Research Council (ERC) with the grant 'INTERACT' (802629) to K.S.

Author contributions

Q.L., F.R., K.S. and R.M. designed the research. Q.L., K.S. and R.M. analyzed the data. B.W. and B.M. generated plant materials. B.H., C.-L.R. and L.G. supervised the whole-genome sequencing work. Q.L., K.S. and R.M. wrote the article with input from the other authors.

Funding

Open access funding provided by Max Planck Society.

 https://doi.org/10.1038/s41588-024-01715-9

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-024-01715-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01715-9.

Correspondence and requests for materials should be addressed to Korbinian Schneeberger or Raphael Mercier.

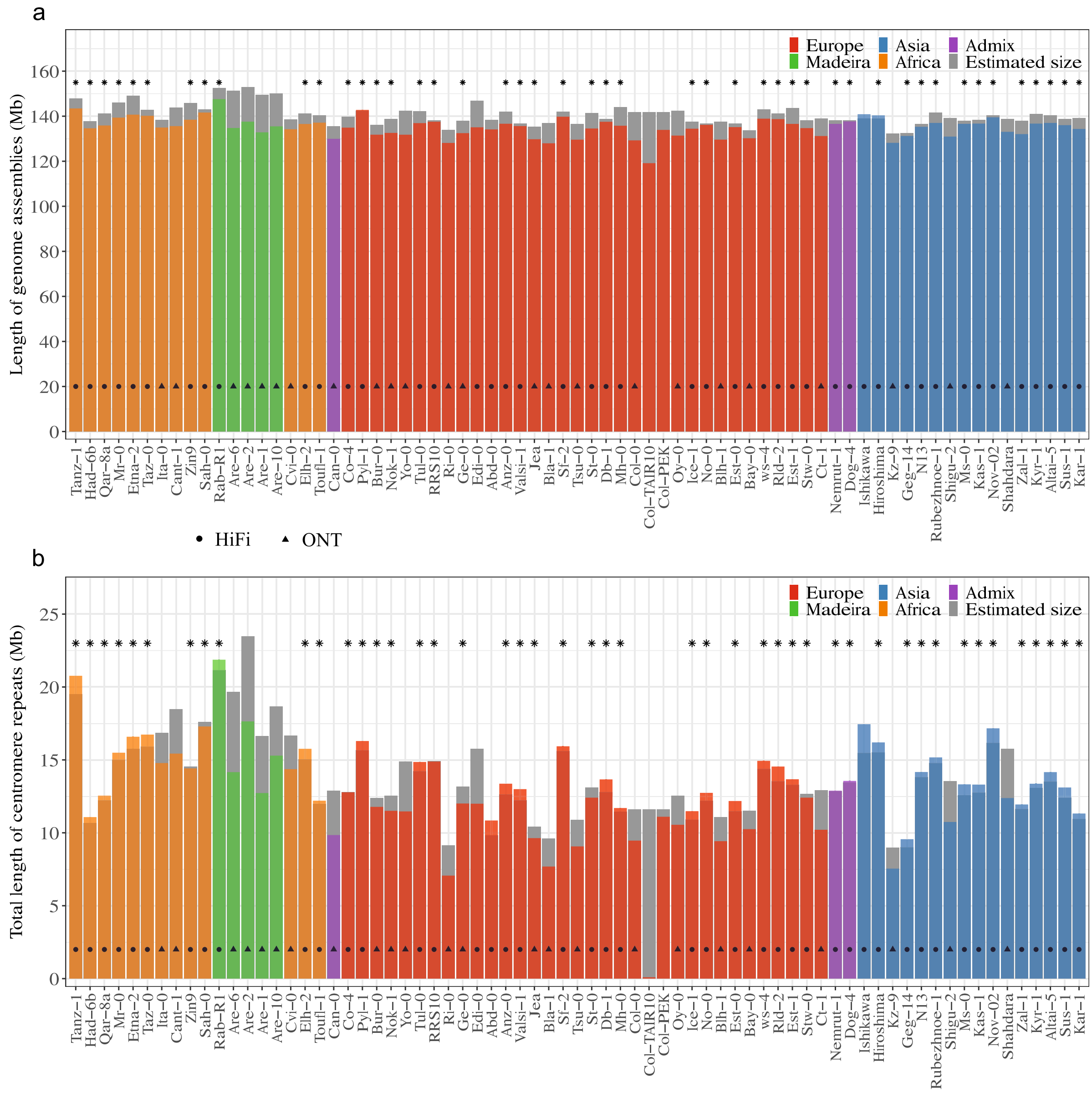
Peer review information *Nature Genetics* thanks Grey Monroe, Kentaro Shimizu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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

https://doi.org/10.1038/s41588-024-01715-9

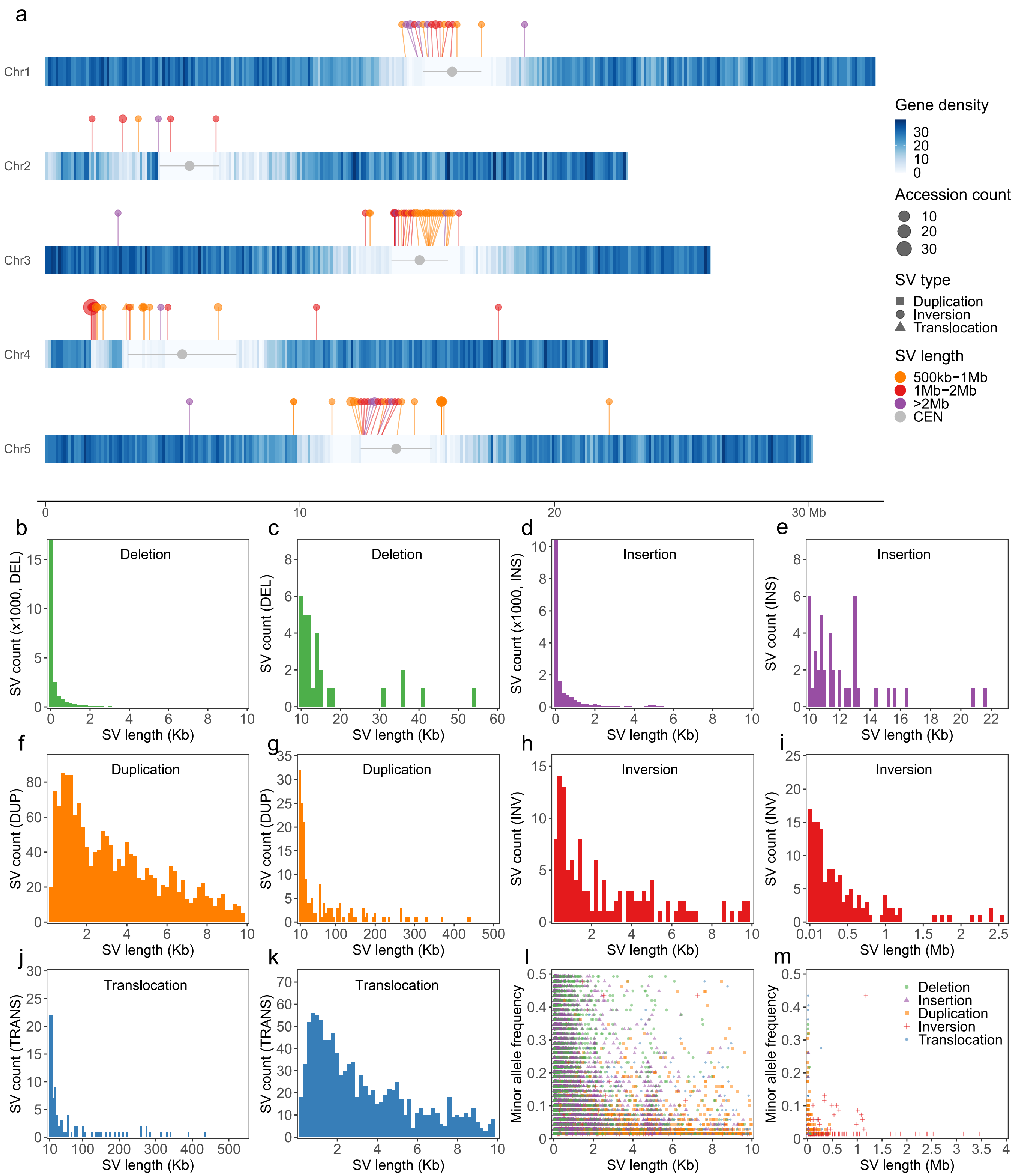
Extended Data Fig. 1 | Phylogenetic tree based on SNPs in the 72 *A. thaliana* genomes. Tree branches (accessions) are coloured according to the genetic classification. Europe (red), Asia (blue), Madeira (green), Africa (orange) and admixture (purple).

 https://doi.org/10.1038/s41588-024-01715-9



Extended Data Fig. 2 | Comparison of the assembly and estimated lengths of genomes and centromeres. (**a**) Comparison of the assembly and genome sizes of the 69 accessions. The genome sizes were estimated based on k-mer from Illumina reads. (**b**) Comparison of the assembly and centromere sizes of the 69 accessions. The centromere sizes were estimated based on sequencing depths of Illumina reads. The estimated sizes are shown by grey bars. The assembly size of each accession are coloured according to its genetic classification. The most complete 46 genomes are marked by stars.

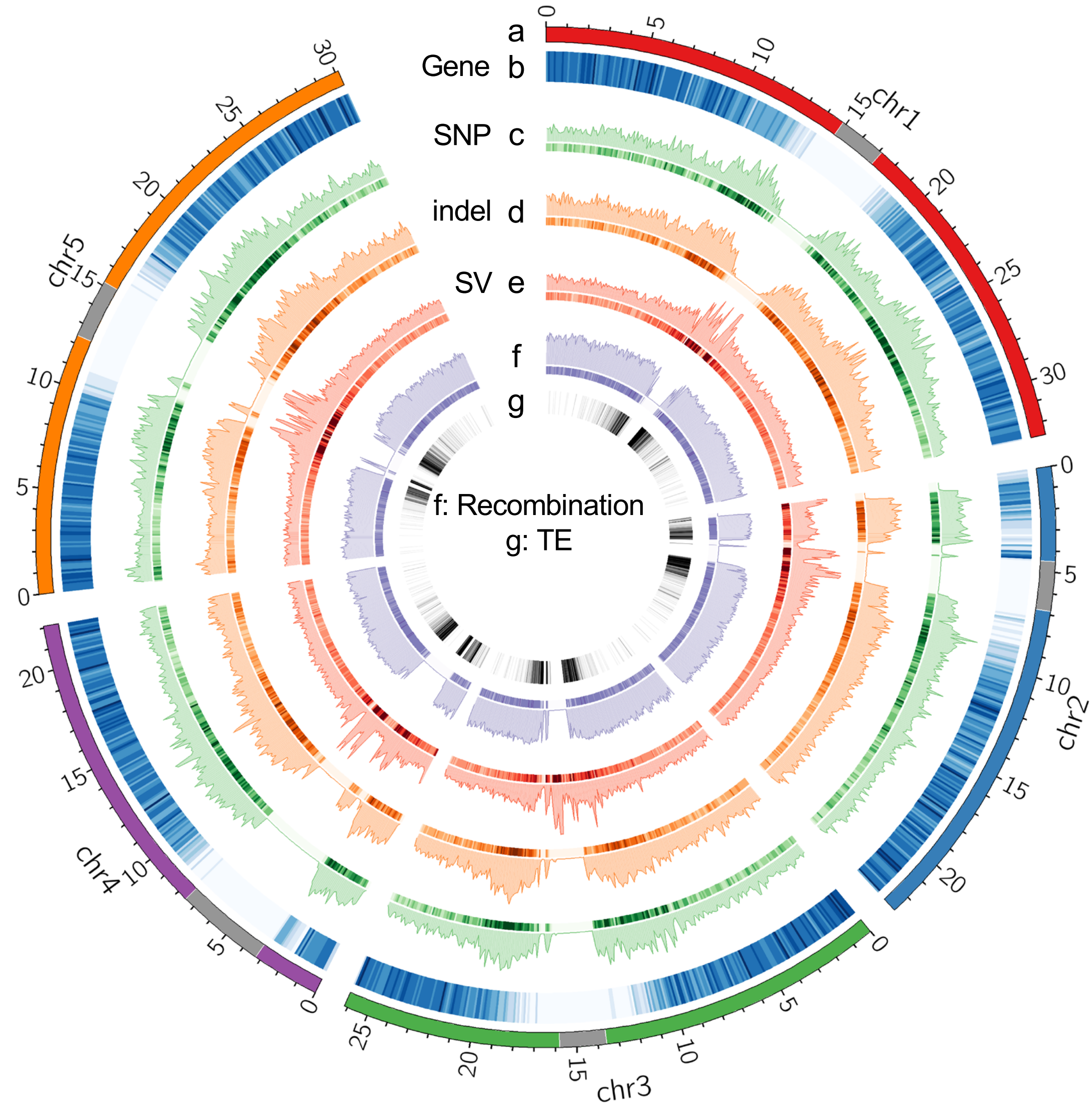
 https://doi.org/10.1038/s41588-024-01715-9



Extended Data Fig. 3 | The distribution of length and minor allele frequency of SVs in the 69 genomes. (**a**) The chromosomal distribution of SVs (> 500 kb) in the 69 genomes along chromosomes of Col-PEK. The heatmap indicates the gene density. The centromeres are indicated by grey circles and segments. The colour, shape and size of the individual points represent the size, type and allele frequency of the SVs. (b-k) Length distributions of SVs (deletions, insertions, duplications, inversions and translocations) with size from 20 bp to 10 kb and longer than 10 kb, separately. (l-m) The minor allele distribution of SVs with sizes from 20 bp to 10 kb and longer than 10 kb, separately. The colour and shape of points represent the type of the SVs.

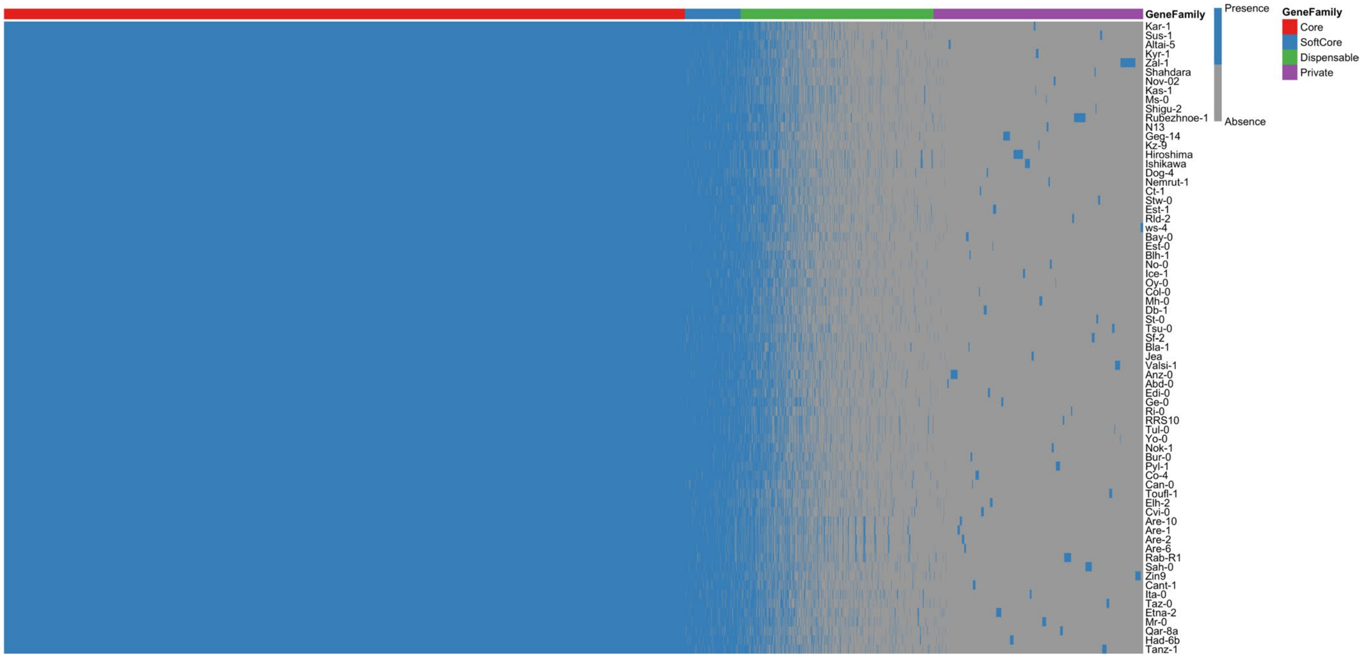
Extended Data Fig. 4 | Alignment of long-reads in the breakpoint regions of the inversion in Stw-0. The top, middle and bottom panels show the alignment of long-reads from Stw-0 against Col-PEK, the closest accession Ct-1, and the Stw-0 assemblies, with a window of 30 kb covering the left and right breakpoints of the detected inversion, respectively.

https://doi.org/10.1038/s41588-024-01715-9



Extended Data Fig. 5 | The distribution of genetic variation in the 72 genomes along chromosomes of Col-PEK. Distribution map of genetic variation in the 72 genomes along the chromosomes profiled in 100 kb windows. a: Chromosomes, centromeres are marked by grey; b: Gene density; c: SNP density; d: small indel density; e: SV density (69 accessions); f: historical recombination map (4Ner per kb), centromeres masked; g: TE density.

 https://doi.org/10.1038/s41588-024-01715-9



Extended Data Fig. 6 | Presence and absence of pan gene families in the 69 *A. thaliana* genomes. Each column indicates a non-redundant gene (gene family), and each row indicates an accession. The gene families were grouped by their categories including core, softcore, dispensable and private. Accessions are ordered according to their genetic relationship.

nature portfolio

Corresponding author(s): Raphael Mercier

Last updated by author(s): Jan 22, 2024

Reporting Summary

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

Statistics

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

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

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

Software and code

Policy information about availability of computer code

Data collection	No software and code used for data collection.
Data analysis	All analysis were performed by using software and code publicly available, details as described in the methods, including PacBio SMRT Link V10, Jellyfish v2.2.6, findGSE v1.0, Canu v2.1.1, Flye v2.7 and Hifiasm v0.16.1, purge_dups v1.2.5, quickmerge v0.3, RaGOO v1.1, MaSuRCA v4.1.0, Porechop v0.2.4, Filtlong v0.2.1, SMARTdenovo (version, 2018-02-19), Racon v1.4.10, NextPolish v1.3.1, complesam v0.2.2, Merqury v1.3, blastn v2.14.1, Liftoff v1.6.3, LTR_retriever v2.9.0, Bowtie2 v2.4.4, samtools v1.9, EDTA v2.0.1, RepeatMasker v4.1.1, Augustus v3.3.3, GeneMark v4.62, GlimmerHMM v3.0.4, SNAP (version 2006-07-28), FastQC v0.11.9, Trimmomatic v0.39, Trinity v2.14.0, TransDecoder v5.5.0, DIAMOND v2.0.4, HMMER v3.1b2, CD-HIT v4.6.8, Exonerate v2.2.0, EVidenceModeler v1.1.1, Barrnap v0.9, Infernal v1.1.4, tRNAscan-SE v2.0.9, NLR-Annotator v2.1, RGAugury v2.0, InterProScan v5.59-91.0, AgriGO v2.0, OrthoFinder v2.5.4, BWA v0.7.15-r1140, inGAP-family v1.0, VCFtools v0.1.16, minimap2 v2.21-r1071, SyRI v1.6, SURVIVOR, PLINK v1.90b6.18, ADMIXTURE v1.3.0, MUSCLE v3.8.31, seqkit v2.3.0, PAL2NAL v14, IQ-TREE v1.6.12, vcf2phylip v2.8, FastEPRR v2.0, KaKs_Calculator v2.0, HISAT2 v2.1.0, StringTie v2.0.6. The related code is available at GitHub (https://github.com/qclian/Pan_Ath) and Zenodo (https://doi.org/10.5281/zenodo.10567419).

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The raw Illumina, PacBio HiFi and Oxford Nanopore sequencing data of the 72 accessions can be accessed in EMBL-ENA under the accession numbers PRJEB62038. The genome assemblies can be accessed in NCBI under the accession numbers PRJNA1033522. The data generated in this study can be accessed in Edmond (the Open Research Data Repository of the Max Planck Society, https://doi.org/10.17617/3.AEOJBL), including genome assemblies (including Lu-1, Pa-1 and Istisu-1), gene and TE annotations, SNPs and SVs, pan-genome matrix, othrogroups.
The RNA-seq dataset used in this study are downloaded from the NCBI SRA database (the accession numbers are included in the Supplementary Table 11).
The database used in this study, including OrthoDB brassicales_odb10 (brassicales, 2020-08-05), NCBI NR database, and Rfam database v14.8, are all public available.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	72 accessions were selected for PacBio HiFi or Oxford Nanopore, and whole genome Illumina sequencing.
Data exclusions	The assembly of three accessions were excluded from the analysis, as they were not inbred as described in the manuscript.
Replication	Genome assembly from long-read data was examined by short-read data. All attempts of replication were successful.
Randomization	This not relevant to this study, as it is about assembly and analysis of 72 Arabidopsis thaliana genomes.
Blinding	The investigators were not blinded, as it it not relevant for this study.

Behavioural & social sciences study design

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

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic*

Research sample	*information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

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

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access & import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*

Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	*State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	*For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Reporting on sex	*Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☐	☐	Public health
☐	☐	National security
☐	☐	Crops and/or livestock
☐	☐	Ecosystems
☐	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks	The stock center accesion numbers are provided a a supplementary table
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
Normalization template	*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
Noise and artifact removal	*Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring | *Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings | *Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*

Effect(s) tested | *Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference | *Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

(See Eklund et al. 2016)

Correction | *Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity | *Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*

Graph analysis | *Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*

Multivariate modeling and predictive analysis | *Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*