# Chromosome-level assemblies of multiple *Arabidopsis thaliana* accessions reveal hotspots of genomic rearrangements

Wen-Biao Jiao and Korbinian Schneeberger*

Max Planck Institute for Plant Breeding Research, Department for Plant Developmental Biology, Carl-von-Linné-Weg 10, 50829 Cologne, Germany

* Author for correspondence: schneeberger@mpipz.mpg.de

**We report chromosome-level, reference-quality assemblies of seven *Arabidopsis thaliana* accessions selected across the geographic range of this model plant. Each genome assembly revealed between 13-17 Mb of rearranged and 5-6 Mb of novel sequence introducing copy-number changes in ~5000 genes, including ~1,900 genes which are not part of the current reference annotation. By analyzing the collinearity between the genomes, we identified ~350 hotspots of rearrangements covering ~4% of the euchromatic genome. Hotspots of rearrangements are characterized by accession-specific accumulation of tandem duplications and are enriched for genes implicated in disease resistance and secondary metabolite biosynthesis. Loss of meiotic recombination in hybrids within these regions is consistent with the accumulation of rare and deleterious alleles and incompatibility loci. Together this suggests that hotspots of rearrangements are governed by different evolutionary dynamics as compared to the rest of the genome and facilitate rapid responses to the ever-evolving challenges of biotic stress.**

The first complete assembly of a plant genome, the reference sequence of *A. thaliana* (Col-0), was based on a minimal tiling path of BACs sequenced with Sanger technology and was released in the year 2000[1]. Since then, this reference has been widely used for identification of genetic variants, which typically relied on short-read based resequencing or reference-guided assembly[2–8]. Although millions of small variants have been identified, the identification of large genomic rearrangements remained challenging. In contrast, reference-independent, chromosome-level assemblies promise accurate identification of all sequence differences independent of their complexity[9]. So-far, however, there are only few reports on

whole-genome *de-novo* assemblies for *A. thaliana* available and the assemblies have not been thoroughly compared to each other[10–13].

Using deep PacBio (45-71x) and Illumina (56-78x) whole-genome shotgun sequencing, we assembled the genomes of seven accessions from geographically diverse populations including An-1 (Antwerpen, Belgium), C24 (Coimbra, Portugal), Cvi-0 (Cape Verde Islands), Eri-1 (Eringsboda, Sweden), Kyo (Kyoto, Japan), L*er* (Gorzów Wielkopolski, Poland) and Sha (Shahdara, Tadjikistan) (Supplementary Table 1). The assembly of L*er* was already described in a recent study and used for the development of a whole-genome comparison tool[9], however, as it was generated in the same process it is integrated in this study as well. These accessions (together with the reference accession Col-0) were initially selected as the founder lines of Arabidopsis Multi-parent Recombination Inbreeding Lines (AMPRIL)[14] population. The contig assemblies featured N50 values from 4.8 - 11.2 Mb and chromosome-normalized L50 (CL50)[15] values of 1 or 2 indicating that nearly all chromosome arms were assembled into a few (typically one to five) contigs only (Table 1 and Supplementary Table 2). We generated chromosome-level assemblies based on the homology of these contigs to the reference sequence and validated the resultant contig order of two of the assemblies with three different genetic maps, where we did not find evidence for a single mis-placed contig (Supplementary Table 3). The seven chromosome-level assemblies consisted of 43 - 73 contigs and reached a total length of 117.7 - 118.8 Mb, which is very similar to the 119.1 Mb of the reference sequence (Fig. 1 and Table 1) and even included parts of the highly complex regions of centromeres, telomeres and rDNA clusters (Supplementary Table 4 and 5). The remaining unanchored contigs had a total length of 1.5 - 3.3 Mb and consisted almost entirely of repeats, which agrees with gaps between the contigs, which most of them (84.5% - 98.0%) were introduced due to repetitive regions (Supplementary Table 6).

Between 99.1% and 99.4% of the reference genes[16] could be aligned to each of the seven assemblies. Almost all of the remaining genes were truly missing in the genomes as we could confirm by mapping their Illumina short reads against the reference sequence (Supplementary Table 7), suggesting that the assemblies covered almost all of the genic regions. In agreement with this, we annotated 27,098 to 27,574 protein-coding genes in each of the assemblies, which is similar to the 27,445 genes annotated in the reference sequence[16] (Table1, Supplementary Table 8 and 9).

The lack of contiguity of short sequencing reads makes resequencing-based analyses mostly blind for large rearrangements. In contrast, the high contiguity of these new assemblies now enables the comprehensive description of complex structural rearrangements including inversions, translocations and duplications (Fig. 2a). By comparing each of the new assemblies against the reference sequence using the whole-genome comparison tool *SyRI*[9], we found between 102.2 and 106.6 Mb of syntenic regions and between 12.6 and 17.0 Mb of

rearranged regions in each of the genomes. The rearrangements included 1.5 - 4.2 Mb (33 - 46) inversions, 1.0 - 1.7 Mb (364 - 566) intra-chromosome translocations, and 0.9 - 1.3 Mb (365 - 626) inter-chromosome translocations. Apart from balanced structural variation, we also found 7.2 – 8.7 Mb (4,288 – 5,150) of duplication loss and gain variation (Fig. 2b and Supplementary Table 10). Similar to sequence variation, rearrangements were not evenly distributed along the chromosomes, but were enriched in pericentromeres (Supplementary Table 11). Their lengths ranged from a few dozen bp to hundreds of kb and even Mb scale (Fig. 2c). Among the largest differences were inversions including a 2.48 Mb inversion on chromosome 3 of Sha (Supplementary Fig. 1 and Supplementary Table 12), which is consistent with the earlier observation of suppressed meiotic recombination in this region in hybrids including the Sha haplotype[17,18]. Local sequence differences in rearranged regions were generally more frequent as compared to syntenic regions mostly due to an excess of sequences with different copy-number variation (Fig. 2d, Supplementary Fig. 2 and Supplementary Table 13). Overall, the allele frequencies of balanced rearrangements like inversions and translocations were generally lower as compared to duplications, where more than 50% of them were shared by at least two accessions suggesting differences in the selection pressures acting on balanced and non-balanced variation (Fig. 2e).

In each of the pairwise comparisons, we identified 5.1 - 6.5 Mb accession-specific sequence (Fig. 2b). Using these regions and their overlaps for a pan-genome analysis[19,20], we estimated a pan-genome size of ~135 Mb and a core-genome size of ~105 Mb (Fig. 2f). A similar analysis with genes modelled a gene pan-genome size of ~30,000 genes illustrating that one reference genome is not sufficient to capture the entire sequence diversity within *A. thaliana* (Supplementary Fig. 3).

Genomic rearrangements have the potential to delete, create or duplicate genes resulting in gene copy number variation (CNV). Based on gene family clustering of all genes in all eight accessions[21] we found 22,040 gene families with conserved copy number, while 4,957 gene families showed differences in gene copy numbers (Fig. 2g and Supplementary Table 14). Almost 99% of the copy-variable gene families had 5 or less copies, while only less than 10% of them showed more than two different copy numbers across the eight accessions (Supplementary Fig. 4). Among the copy-variable genes we found 1,941 non-reference gene families including 891 gene families present in at least two of the other accessions (Fig. 2h). Around 23% of the non-reference gene families featured orthologs in the closely related genome of *Arabidopsis lyrata* and, according to RNA-seq read mapping, 26%-40% of them showed evidence of expression in the individual accessions (Supplementary Table 15). The remaining 1,050 non-reference gene families, which were evenly distributed across the accessions (Fig. 2g), with the only exception of Cvi-0, where we found nearly twice as many (214) accession-specific genes which is in agreement with the divergent ancestry of this relict

accession[4,22].

The high contiguity of the assembled sequences enabled the first analysis of the conservation of collinearity between different Arabidopsis genomes. For this we introduced a new concept called *Synteny Diversity* $\pi_{syn}$, which is similar to *Nucleotide Diversity*[23], however, instead of measuring average sequence differences it measures average fraction of non-collinear genome pairs in a given population. $\pi_{syn}$ values can range from 0 to 1, where 1 refers to the complete absence of collinearity between any of the genomes and 0 to regions where all genomes are collinear. $\pi_{syn}$ can be calculated in any given region, however, the annotation of synteny still needs to established within the context of the whole genomes to avoid false assignments of homologous but non-allelic sequence (Methods).

We calculated $\pi_{syn}$ in 5-kb sliding windows across the genome using pair-wise comparisons of all eight accessions (Fig. 3a). As expected, $\pi_{syn}$ was generally high in pericentromeric regions and low in chromosome arms. Overall, this revealed around 90 Mb (76% of the genome) where all genomes were syntenic to each other, while for the remaining 29 Mb (24%) the collinearity between the genomes was not conserved. This, for example, included a region on chromosome 3 (ranging from Mb ~2.8 - 5.3), where $\pi_{syn}$ was increased to ~0.25 due to the 2.48 Mb inversion specific to Sha (Fig. 3a, arrow labelled with (a)).

Unexpectedly, however, some regions featured $\pi_{syn}$ values even larger than 0.5 indicating that not only two, but multiple independent, non-syntenic haplotypes segregate implying that these regions are more likely to undergo or conserve complex mutations as compared to the rest of the genomes. Overall, we found 576 such hotspots of rearrangements (HR) with a total size of 10.2 Mb including 351 regions in the gene-rich euchromatic chromosome arms with a total length of 4.1 Mb (or 4% of euchromatic genome) (Supplementary Table 16).

Even though HRs in euchromatic regions included more transposable elements and less genes as compared to conserved regions, they still contained significant numbers of genes, many of which occurred at high and variable copy-number between the accessions (Fig. 3b, c). For example, a HR on chromosome 4, which overlapped with the *RPP4/RPP5* R gene cluster[24,25], displayed five to 15 intact or truncated copies of the *RPP5* gene within the eight genomes (Fig. 3d and Supplementary Table 17). The different gene copies were primarily introduced by an accumulation of forward tandem duplications and large indels (Fig. 3e), a pattern, which was shared by many HR regions (Supplementary Fig. 5). While the individual duplications were not conserved between the haplotypes, the borders of HRs were typically well conserved (Fig. 3f). This suggested that either different selection regimes introduced clear-cut borders between highly HRs and their surrounding regions, or that increased tandem duplication rates were exactly limited to the HR regions. Such a local increase of mutation rates could potentially be enabled by non-allelic homologous recombination (NAHR) which

could be triggered by the high number of local repeats in these regions[26–28]. In any case the accession-specific accumulation of duplications suggested that the HR regions are only partially linked to their genomic vicinity. To test this, we analysed the linkage disequilibrium (LD) within 1,135 genomes of the 1001 Genomes Project[4] across the HR regions. LD was high in the regions around the HR and was also high within HRs, however, when calculated across the HR border LD was significantly lower supporting the idea that HRs are not linked to the surrounding haplotypes but that they are segregating as large non-recombining units (Fig. 4f). To test if HRs are indeed depleted for meiotic recombination, we overlapped them with 15,683 crossover (CO) sites previously identified within Col-0/Ler F2 progenies[29,30]. Only 64 of them partially overlapped with non-syntenic regions while all other COs co-located syntenic regions (Fig. 4b), suggesting that HRs are almost completely silenced for COs (P-value < 2e-16) and that they follow different evolutionary dynamics as compared to the rest of the (recombining) genome.

While lack of meiotic crossovers coupled with increased duplication rates can lead to rapid generation of new haplotypes, reduced meiotic recombination also has been linked to the accumulation of new (deleterious) mutations[31]. In agreement with this HRs also showed an accumulation of single nucleotide polymorphisms with low allele frequencies and potentially deleterious variation as compared to other regions in the genome (Fig. 4c, d and Supplementary Fig. 6). Moreover, reduced recombination combined with geographic isolation can provide the basis for the development of alleles, which are incompatible with distantly related haplotypes leading to intra-species incompatibilities[32]. To test this, we searched the location of nine recently reported genetic incompatible loci[33] (*DM1-9*) and found that all except of one overlapped with HRs, while *DM3*, the locus which did not overlap with HRs, was closely flanked by two HRs (Fig.5a and Supplementary Fig. 7-12). In addition, we also checked the locus of a recently published single-locus genetic incompatibility[34] and found that it was also residing in a HR region (Supplementary Fig. 13).

The excess of tandem duplications, accumulation of accession-specific gene-copy numbers and increased sequence diversity paired with the lack of meiotic recombination and the accumulation of deleterious alleles were reminiscent of the patterns that have been described for R gene clusters[35–42]. In fact, the 808 reference genes in HRs were significantly enriched for genes involved in defense response, signal transduction and secondary metabolite biosynthesis (Fig. 4e) suggesting a reoccurring role of HRs in the adaptation to biotic stress. As biotic challenges are constantly changing as pathogens evolve, it has been proposed that the accumulation of new gene duplicates could increase the genomic diversity and thereby enable rapid genomic changes supporting the genomic response of plants against pathogens[28,43–45].

The availability of these high-quality, chromosome-level genomes assemblies now

enables the simultaneous access to all these regions, which can now be studied not only locally, but in the context of entire genomes. For example, all but one of the 47 NB-LRR R-gene clusters were completely reconstructed in all of the seven new assemblies (Fig. 5, Supplementary Table 17)[46]. The completeness of the genomes allows an unbiased view on all rearrangements, what could be essential as many of the HRs were only covered with a few or even no allelic markers and could easily be missed with genome-wide selection or diversity-scans (Supplementary Fig. 14 and 15). The genome-wide estimation of *Synteny Diversity* might be one way to identify such regions which have the potential to act as the genetic origin of rapid responses to the ever-changing environmental challenges.

# Methods

### Plant material and whole genome sequencing

The seeds of all seven accession were received from Maarten Koornneef (MPI for Plant Breeding Research), and were grown under normal greenhouse conditions. DNA preparation and next generation sequencing was performed by the Max Planck Genome center. The DNA samples were prepared as previously described for the L*er*[9], and sequenced with a PacBio Sequel system. For each accession, data from two SMRT cells were generated. Besides, Illumina paired-end libraries were prepared and sequenced on the Illumina HiSeq2500 systems.

### Genome assembly

PacBio reads were filtered for short (<50bp) or low quality (QV<80) reads using SMRTLink5 package. *De novo* assembly of each genome was initially performed using three different assembly tools including Falcon[47], Canu[48] and MECAT[49]. The resulting assemblies were polished with Arrow from the SMRTLink5 package and then further corrected with mapping of Illumina short reads using BWA[50] to remove small-scale assembly errors which were identified with SAMTools[51]. For each genome, the final assembly was based on the Falcon assembly as these assemblies always showed highest assembly contiguity. A few contigs were further connected or extended based on whole genome alignments between Falcon and Canu or MECAT assemblies. Contigs were labelled as organellar contigs if they showed alignment identity and coverage both larger than 95% when aligned against the mitochondrial or chloroplast reference sequences. A few of contigs aligned to multiple chromosomes and were split if no Illumina short read alignments supported the conflicting regions. Assembly contigs larger than 20kb were combined to pseudo-chromosomes according to their alignment positions when aligned against the reference sequence using MUMmer4[52]. Contigs with consecutive alignments were concatenated with a stretch of 500

Ns. To note, the assembly of the L*er* accession was already described in a recent study[9].

**Assembly evaluation**

We evaluated the assembly completeness by aligning the reference genes against each of the seven genomes using Blastn[53]. Reference genes which were not aligned or only partially aligned might reveal genes which were missed during the assembly. To examine whether they were really missed, we mapped Illumina short reads from each genome against the reference genome using the BWA[50] and checked the mapping coverage of these genes. The genes, which were missing in the assembly, should show fully alignment coverage (Supplementary Table 7).

Centromeric and telomeric tandem repeats were annotated by searching for the 178 bp tandem repeat unit[54] and the 7 bp tandem repeat unit of TTTAGGG[55]. rDNA clusters were annotated with Infernal version 1.1[56].

The assembly contiguity of Cvi-0 and L*er* were further tested using three previously published genetic maps[57–59] (Supplementary Table 3). For this we aligned the marker sequences against the chromosome-level assemblies and checked the order of the markers in the assembly versus their order in the genetic map. The ordering of contigs to chromosomes was perfectly supported by all three maps. Overall, only six (out of 1,156) markers showed conflicts between the genetic and physical map. In all six cases we found evidence that the conflict was likely caused by structural differences between the parental genomes.

**Gene annotation**

Protein-coding genes were annotated based on *ab initio* gene predications, protein sequence alignments and RNA-seq data. Three *ab initio* gene predication tools were used including Augustus[60], GlimmerHMM[61] and SNAP[62]. The reference protein sequences from the Araport 11[16] annotation were aligned to each genome assembly using exonerate[63] with the parameter setting "--percent 70 --minintron 10 --maxintron 60000". For five accessions (An-1, C24, Cvi-0, Ler-0, and Sha) we downloaded a total of 155 RNA-seq data sets from the NCBI SRA database (Supplementary Table 8). RNA-seq reads were mapped to the corresponding genome using HISAT2[64] and then assembled into transcripts using StringTie[65] (both with default parameters). All different evidences were integrated into consensus gene models using Evidence Modeler[66].

The resulting gene models were further evaluated and updated using the Araport 11[16] annotation. Firstly, for each of the seven genomes, the predicted gene and protein sequences were aligned to the reference sequence, while all reference gene and protein sequences were aligned to each of the other seven genomes using Blast[53]. Then, potentially mis-annotated

genes including mis-merged (two or more genes are annotated as a single gene), mis-split (one gene is annotated as two or more genes) and unannotated genes were identified based on the alignments using in-house python scripts. Mis-annotated or unannotated genes were corrected or added by incorporating the open reading frames generated by *ab initio* predications or protein sequence alignment using Scipio[67].

Noncoding genes were annotated by searching the Rfam database[68] using Infernal version 1.1[56]. Transposon elements were annotated with RepeatMasker (http://www.repeatmasker.org). Disease resistance genes were annotated using RGAugury[69]. NB-LRR R gene clusters were defined based on the annotation from a previous study[70].

**Pan-genome analysis**

Pan-genome analyses were performed at both sequence and gene level. To construct a pan-genome of sequences, we generated pair-wise whole genome sequence alignments of all possible pairs of the eight genomes using the nucmer in the software package MUMmer4[52]. A pan-genome was initiated by choosing one of the genomes, followed by iteratively adding the non-aligned sequence of one of the remaining genomes. Here, non-aligned sequences were required to be longer than 100bp without alignment with an identity of more than 90%. The core genome was defined as the sequence space shared by all sampled genomes. Like the pan-genome, the core-genome analysis was initiated with one genome. Then all other genomes were iteratively added, while excluding all those regions which were not aligned against each of the other genomes. The pan- and core-genome of genes was built in a similar way. The pan-genome of genes was constructed by selecting the whole protein coding gene set of one of the accessions followed by iteratively adding the genes of one of the remaining accessions. Likewise, the core-genome of genes was defined as the genes shared in all sampled genomes.

For each pan or core genomes analysis, all possible $(\sum_{n=1}^{8}\left(\frac{8!}{n!(8-n)}\right))$ combinations of integrating the eight genomes (or a subset of them) were evaluated. The exponential regression model $y = A\, e^{Bx} + C$ to was then used to model the pan-genome/core-genomes by fitting medians using the least square method implemented in the nls function of R.

**Analysis of structural rearrangements and gene CNV**

All assemblies were aligned to the reference sequence using nucmer from the MUMmer4[52] toolbox with parameter setting "-max -l 40 -g 90 -b 100 -c 200". The resulting alignments were further filtered for alignment length (>100) and identity (>90). Structural rearrangements and local variations were identified using SyRI[9]. The functional effects of sequence variation were annotated with snpEff[71]. The gene CNV were identified according to the gene family clustering using the tool OrthoFinder[21] based on all protein sequences from

286 the eight accessions.

287 **Synteny diversity, hotspots of rearrangements and diversity estimates**

288 *Synteny Diversity* was defined as the average fraction of non-syntenic sites found within
289 all pairwise genome comparisons within a given population. Here we denote *Synteny*
290 *Diversity* as

291 $$\pi_{syn} = \sum_{ij} x_i x_j \pi_{ij},$$

292 where $x_i$ and $x_j$ refer to the frequencies of sequence $i$ and $j$ and $\pi_{ij}$ to the average
293 probability of a position in $i$ to be non-syntenic. Note, $\pi_{syn}$ can be calculated in a given region
294 or for the entire genome. However even when calculated for small regions the annotation of
295 synteny still needs to established within the context of the whole genomes to avoid false
296 assignments of homologous but non-allelic sequence. Here we used the annotation of *SyRI*
297 to define syntenic regions. $\pi_{syn}$ values can range from 0 to 1, with higher values referring to a
298 higher average degree of non-syntenic regions between the genomes.

299 For the analyses, we calculated $\pi_{syn}$ in 5-kb sliding windows with 1kb step-size across the
300 entire genome. HR regions were defined as regions with $\pi_{syn}$ larger than 0.5. Neighboring
301 regions were merged into one HR if their distance was shorter than 2kb.

302 The nucleotide and haplotype diversity were calculated with the R package PopGenome[72]
303 using SNP markers (with MAF > 0.05) from 1001 Genomes Project[4]. LD were calculated as
304 correlation coefficients $r^2$ using SNP markers with MAF > 0.05. GO enrichment analysis was
305 performed using the webtool DAVID[73,74].

## Data availability

All raw sequencing data, assemblies and annotations can be accessed in the European Nucleotide Archive under the project accession number PRJEB31147.

## Code availability

Custom code used in this study can be found online at *https://github.com/schneebergerlab/AMPRIL-genomes*.

## Acknowledgements

The authors would like to thank Beth R. Rowan (UC Davis) for providing the CO breakpoint list, Bruno Hüttel (Max Planck Genome center) for support in genome sequencing, Sigi Effgen and Maarten Koornneef (Max Planck Institute for Plant Breeding Research) for providing seeds, Onur Dogan (Max Planck Institute for Plant Breeding Research) for help in the greenhouse, Angela M. Hancock (Max Planck Institute for Plant Breeding Research) for helpful discussions, and Raphael Mercier and Padraic J. Flood (Max Planck Institute for Plant Breeding Research) for helpful comments on the manuscript and the interpretation of HR regions.

## Authors contributions

W.-B.J. and K.S. designed the study. W.-B.J. performed all analysis. K.S. supervised the study. W.-B.J. and K.S. wrote the manuscript. All authors read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

# References


1. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796–815 (2000).
2. Cao, J. *et al.* Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.* **43**, 956–965 (2011).
3. Long, Q. *et al.* Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat. Genet.* **45**, 884–890 (2013).
4. Alonso-Blanco, C. *et al.* 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**, 481–491 (2016).
5. Schneeberger, K. *et al.* Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *PNAS* **108**, 10249–10254 (2011).
6. Gan, X. *et al.* Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* **477**, 419–423 (2011).
7. Schneeberger, K. *et al.* Simultaneous alignment of short reads against multiple genomes. *Genome Biol.* (2009). doi:10.1186/gb-2009-10-9-r98
8. Schmitz, R. J. *et al.* Patterns of population epigenomic diversity. *Nature* (2013). doi:10.1038/nature11968
9. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: identification of syntenic and rearranged regions from whole-genome assemblies. *bioRxiv* (2019). doi:10.1101/546622
10. Zapata, L. *et al.* Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. *Proc. Natl. Acad. Sci.* **113**, E4052–E4060 (2016).
11. Michael, T. P. *et al.* High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nat. Commun.* **9**, 1–8 (2018).
12. 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* (2019). doi:10.1371/journal.pone.0216233
13. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Meth* **advance on**, (2016).
14. Huang, X. *et al.* Analysis of natural allelic variation in *Arabidopsis* using a multiparent recombinant inbred line population. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4488–4493 (2011).
15. Jiao, W.-B. *et al.* Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome

conformation capture data. *Genome Res.* **27**, 778–786 (2017).

16. Cheng, C. Y. *et al.* Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *Plant J.* (2017). doi:10.1111/tpj.13415

17. 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).

18. 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).

19. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial 'pan-genome'. *Proc. Natl. Acad. Sci.* **102**, 13950–13955 (2005).

20. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Current Opinion in Genetics and Development* (2005). doi:10.1016/j.gde.2005.09.006

21. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* (2015). doi:10.1186/s13059-015-0721-2

22. Durvasula, A. *et al.* African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci.* **114**, 5213–5218 (2017).

23. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci.* (1979). doi:10.1073/pnas.76.10.5269

24. Parker, J. E. The Arabidopsis Downy Mildew Resistance Gene RPP5 Shares Similarity to the Toll and Interleukin-1 Receptors with N and L6. *PLANT CELL ONLINE* (1997). doi:10.1105/tpc.9.6.879

25. Van Der Biezen, E. A., Freddie, C. T., Kahn, K., Parker, J. E. & Jones, J. D. G. Arabidopsis RPP4 is a member of the RPP5 multigene family of TIR-NB-LRR genes and confers downy mildew resistance through multiple signalling components. *Plant J.* **29**, 439–451 (2002).

26. Wicker, T., Yahiaoui, N. & Keller, B. Illegitimate recombination is a major evolutionary mechanism for initiating size variation in plant resistance genes. *Plant J.* (2007). doi:10.1111/j.1365-313X.2007.03164.x

27. Nagy, E. D. & Bennetzen, J. L. Pathogen corruption and site-directed recombination at a plant disease resistance gene cluster. *Genome Res.* (2008). doi:10.1101/gr.078766.108

28. Leister, D. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends in Genetics* (2004).

doi:10.1016/j.tig.2004.01.007

29. Serra, H. *et al.* Massive crossover elevation via combination of HEI10 and recq4a recq4b during Arabidopsis meiosis . *Proc. Natl. Acad. Sci.* (2018). doi:10.1073/pnas.1713071115

30. Rowan, B. A. *et al.* An ultra high-density Arabidopsis thaliana crossover map that refines the influences of structural variation and epigenetic features. *bioRxiv* (2019). doi:10.1101/665083

31. Kondrashov, A. S. Deleterious mutations and the evolution of sexual reproduction. *Nature* (1988). doi:10.1038/336435a0

32. Bomblies, K. & Weigel, D. Hybrid necrosis: Autoimmunity as a potential gene-flow barrier in plant species. *Nat. Rev. Genet.* **8**, 382–393 (2007).

33. Chae, E. *et al.* Species-wide genetic incompatibility analysis identifies immune genes as hot spots of deleterious epistasis. *Cell* **159**, 1341–1351 (2014).

34. 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**, (2011).

35. Michelmore, R. W. & Meyers, B. C. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Research* (1998). doi:10.1101/gr.8.11.1113

36. Meyers, B. C., Shen, K. A., Rohani, P., Gaut, B. S. & Michelmore, R. W. Receptor-like genes in the major resistance locus of lettuce are subject to divergent selection. *Plant Cell* (1998). doi:10.1105/tpc.10.11.1833

37. Noel, L. Pronounced Intraspecific Haplotype Divergence at the RPP5 Complex Disease Resistance Locus of Arabidopsis. *Plant Cell Online* **11**, 2099–2112 (1999).

38. McDowell, J. M. *et al.* Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the RPP8 locus of arabidopsis. *Plant Cell* (1998). doi:10.1105/tpc.10.11.1861

39. Botella, M. A. *et al.* Three genes of the arabidopsis RPP1 complex resistance locus recognize distinct Peronospora parasitica avirulence determinants. *Plant Cell* (1998). doi:10.1105/tpc.10.11.1847

40. Barragan, C. A. *et al.* RPW8/HR repeats control NLR activation in Arabidopsis thaliana. *PLOS Genet.* (2019). doi:10.1371/journal.pgen.1008313

41. Bakker, E. G. A Genome-Wide Survey of R Gene Polymorphisms in Arabidopsis. *Plant Cell Online* **18**, 1803–1818 (2006).

42. Guo, Y.-L. *et al.* Genome-Wide Comparison of Nucleotide-Binding Site-Leucine-Rich Repeat-Encoding Genes in Arabidopsis. *Plant Physiol.* **157**, 757–769 (2011).

43. Dangl, J. L. & Jones, J. D. G. Plant pathogens and integrated defence responses to

infection. *Nature* (2001). doi:10.1038/35081161

44. Boller, T. & He, S. Y. Innate immunity in plants: An arms race between pattern recognition receptors in plants and effectors in microbial pathogens. *Science* (2009). doi:10.1126/science.1171647

45. Kondrashov, F. A. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the Royal Society B: Biological Sciences* (2012). doi:10.1098/rspb.2012.1108

46. Choi, K. *et al.* Recombination Rate Heterogeneity within Arabidopsis Disease Resistance Genes. *PLoS Genet.* (2016). doi:10.1371/journal.pgen.1006179

47. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**, 1050–1054 (2016).

48. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* (2017). doi:10.1101/gr.215087.116

49. Xiao, C. Le *et al.* MECAT: Fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat. Methods* **14**, 1072–1074 (2017).

50. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

51. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

52. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* (2018). doi:10.1371/journal.pcbi.1005944

53. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* (1990). doi:10.1016/S0022-2836(05)80360-2

54. Heslop-Harrison, J. S., Murata, M., Ogura, Y., Schwarzacher, T. & Motoyoshi, F. Polymorphisms and Genomic Organization of Repetitive DNA from Centromeric Regions of Arabidopsis Chromosomes. *Plant Cell* **11**, 31 LP – 42 (1999).

55. Richards, E. J. & Ausubel, F. M. Isolation of a higher eukaryotic telomere from Arabidopsis thaliana. *Cell* (1988). doi:10.1016/0092-8674(88)90494-1

56. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* (2013). doi:10.1093/bioinformatics/btt509

57. 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* (2008). doi:10.1534/genetics.107.083899

58. Singer, T. *et al.* A high-resolution map of Arabidopsis recombinant inbred lines by whole-genome exon array hybridization. *PLoS Genet.* (2006). doi:10.1371/journal.pgen.0020144

59. Giraut, L. *et al.* Genome-wide crossover distribution in Arabidopsis thaliana meiosis

reveals sex-specific patterns along chromosomes. *PLoS Genet.* (2011). doi:10.1371/journal.pgen.1002354

60. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, (2003).

61. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics* (2004). doi:10.1093/bioinformatics/bth315

62. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* (2004). doi:10.1186/1471-2105-5-59

63. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).

64. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* (2015). doi:10.1038/nmeth.3317

65. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* (2015). doi:10.1038/nbt.3122

66. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).

67. Keller, O., Odronitz, F., Stanke, M., Kollmar, M. & Waack, S. Scipio: Using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC Bioinformatics* **9**, 1–12 (2008).

68. Kalvari, I. *et al.* Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gkx1038

69. Li, P. *et al.* RGAugury: A pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics* (2016). doi:10.1186/s12864-016-3197-x

70. Choi, K. *et al.* Recombination Rate Heterogeneity within Arabidopsis Disease Resistance Genes. *PLoS Genet.* **12**, 1–30 (2016).

71. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin).* (2012). doi:10.4161/fly.19695

72. Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: An efficient swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* **31**, 1929–1936 (2014).

73. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* (2009). doi:10.1038/nprot.2008.211

74. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: Paths

toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* (2009). doi:10.1093/nar/gkn923

# Tables

**Table 1. Genome assembly and annotation of eight *A. thaliana* accessions**

| | Col-0 | An-1 | C24 | Cvi-0 | Eri-1 | Kyo | L*er* | Sha |
|---|---|---|---|---|---|---|---|---|
| Total sequences | 5 | 151 | 167 | 140 | 200 | 230 | 149 | 143 |
| Total length (Mbp) | 119.1 | 120.1 | 119.2 | 119.7 | 120.8 | 122.2 | 120.3 | 120.3 |
| Contig N50 (Mbp) | - | 8.2 | 4.8 | 7.4 | 4.8 | 9.1 | 11.2 | 7.0 |
| Contig L50 | - | 6 | 8 | 7 | 8 | 6 | 5 | 6 |
| Contig N90 (Mbp) | - | 1.3 | 0.6 | 2.0 | 0.7 | 0.9 | 0.8 | 1.0 |
| Contig L90 | - | 17 | 34 | 17 | 34 | 19 | 18 | 23 |
| CL50 | - | 2 | 2 | 2 | 2 | 2 | 1 | 1 |
| Chr. Length (Mbp) | 119.1 | 118.4 | 117.7 | 118.3 | 117.7 | 118.8 | 118.5 | 118.4 |
| Genes | 27,445 | 27,342 | 27,214 | 27,098 | 27,285 | 27,574 | 27,376 | 27,293 |

# Figure legends

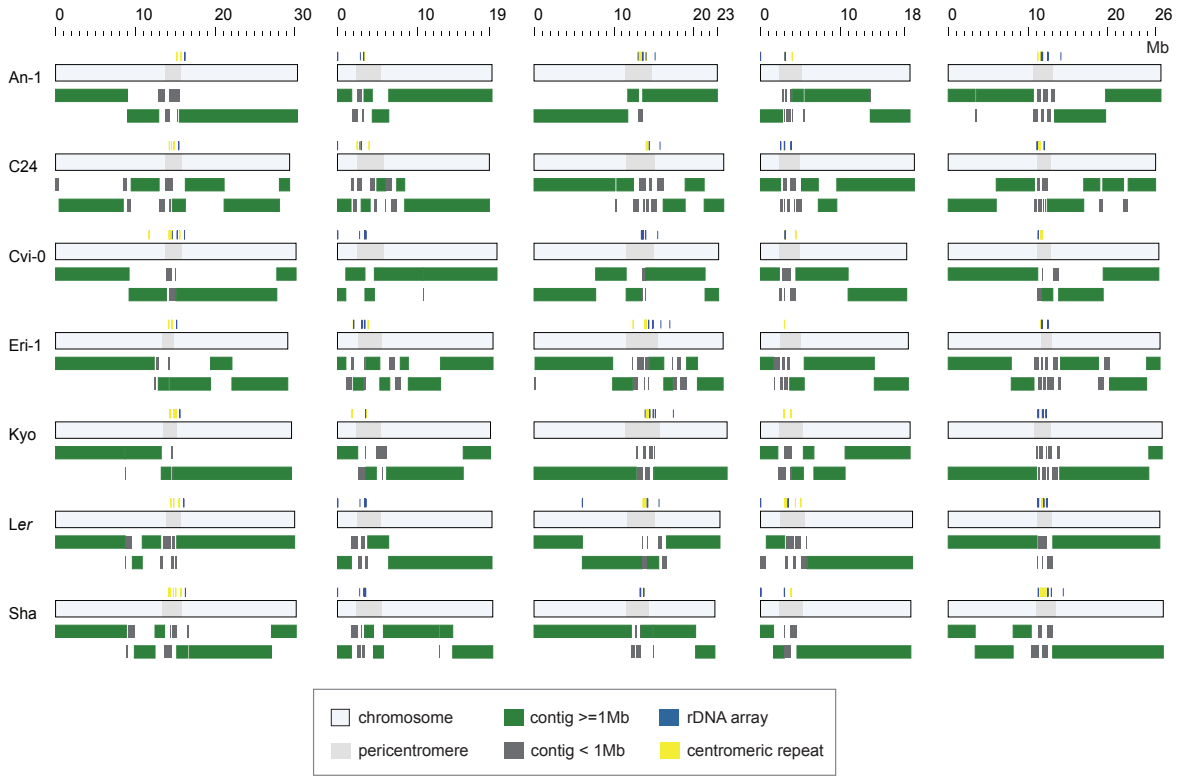



**Figure 1. Chromosome-level genome assemblies of seven *A. thaliana* accessions.** The arrangement of contigs to chromosomes for seven assemblies are shown with green (>1 Mb) and dark grey (<1 Mb) boxes. The full range of each chromosomes is shown with a light blue bar, where the gray inlays outline the extend of the pericentromeric regions. The location of centromeric tandem repeat arrays and rDNA clusters within the assemblies are marked by yellow and blue bars above each of the chromosomes.

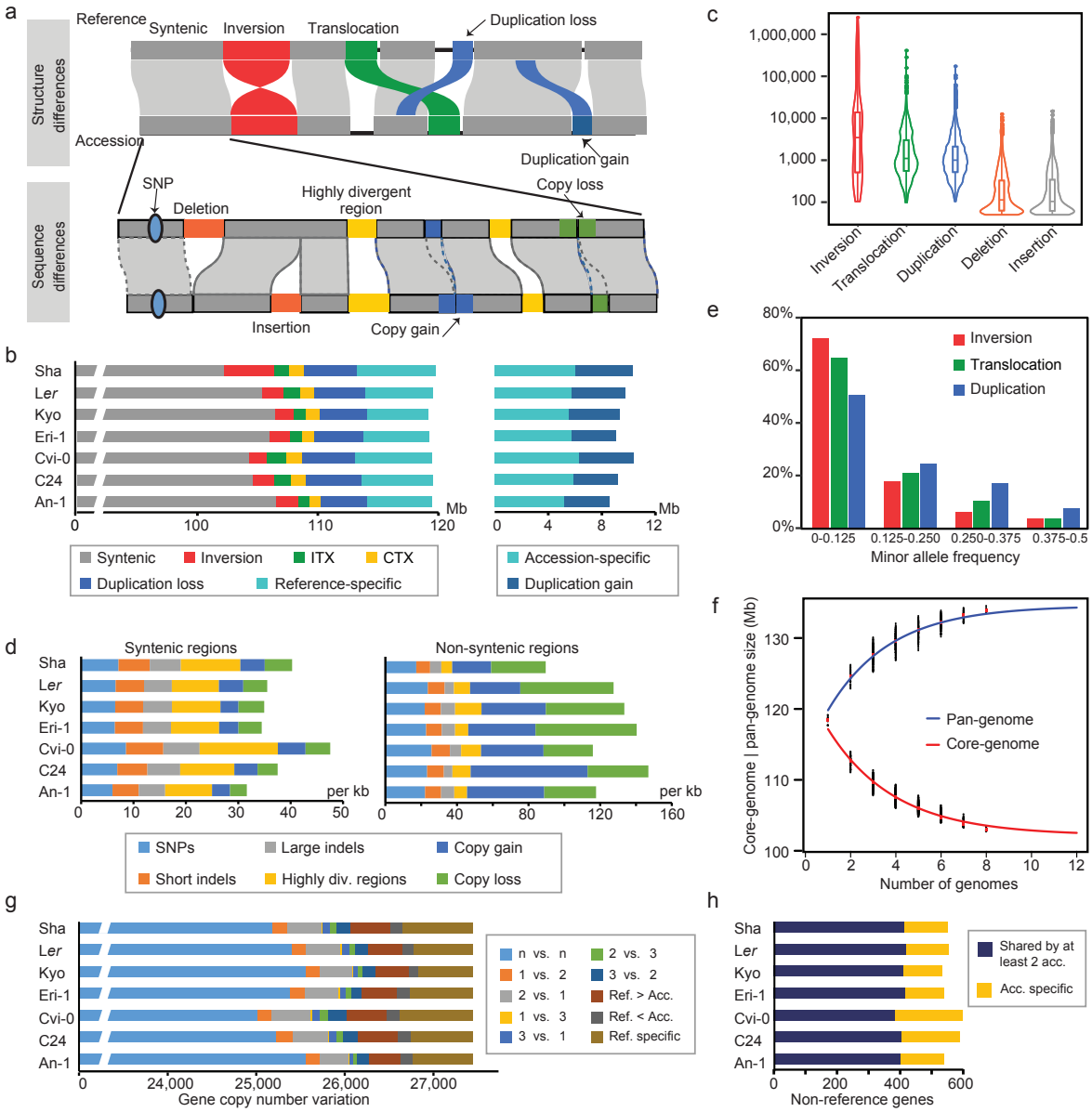
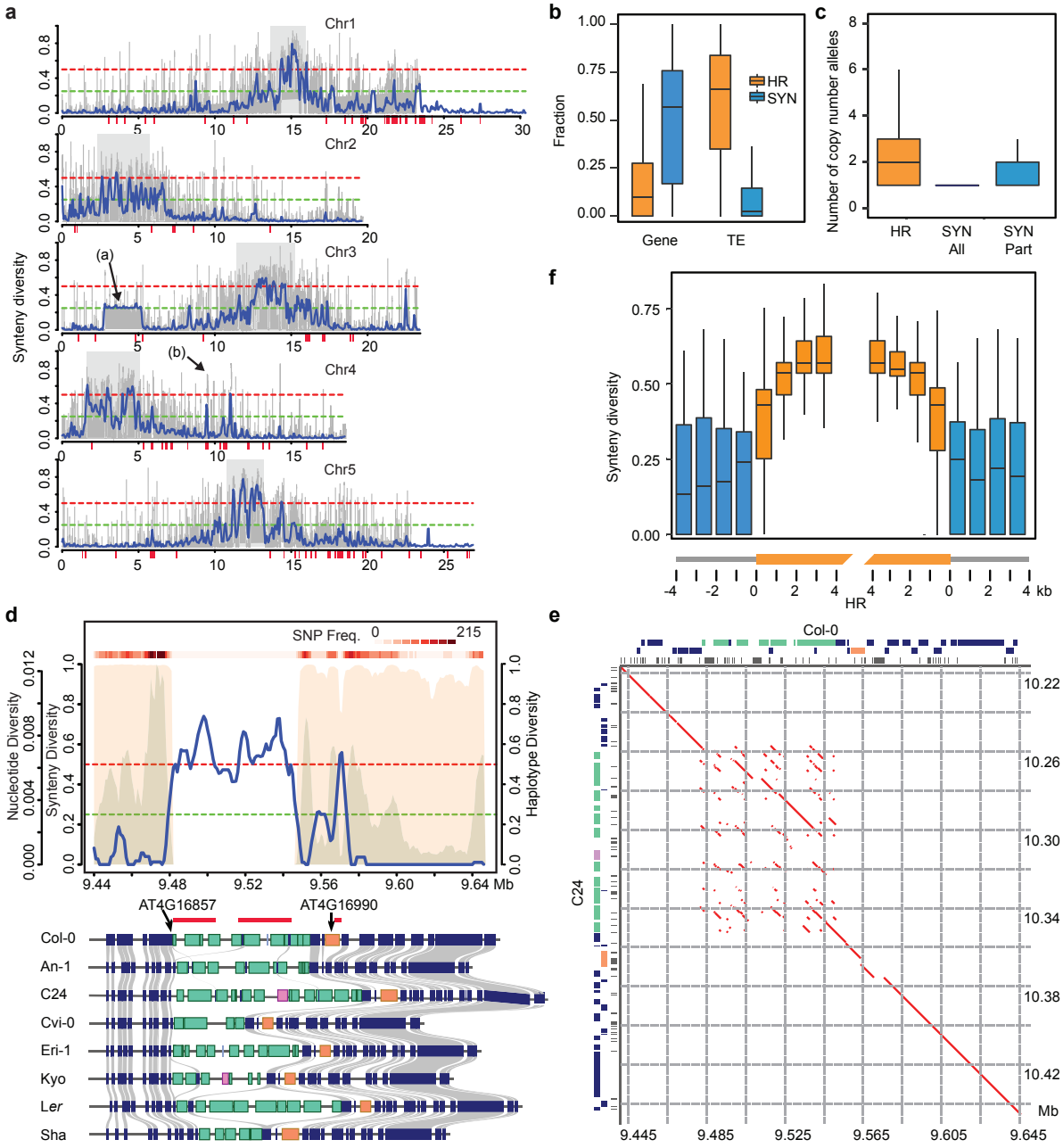



**Figure 2. Comprehensive catalogue of structural variation**

**(a)** Schematic of the structural differences (upper panel) and local sequence variation (lower panel) that can be identified by whole-genome comparisons using SyRI[9]. Note, local sequence variation can reside in syntenic as well as in rearranged regions. **(b)** The total span of syntenic and rearranged regions between the reference and each of other accessions. The left plot shows the sequence span in respect to the reference sequence, while the right plot shows the sequence space, which is specific to each of the accessions. **(c)** Size distributions of different types of structural variation. **(d)** Local variation (per kb) in syntenic (left) and rearranged (right) regions between the reference and each of other accessions. **(e)** Minor allele frequencies of all three types of rearranged regions. **(f)** Pan-genome and core-genome estimations based on all pairwise whole-genome comparisons across all eight accessions. Each point corresponds to a pan- or core-genome size estimated with a particular

541 combination of genomes. The red points indicate median values for each combination with
542 the same number of genomes. Pan-genome (blue) and core-genome (red) estimations were
543 fitted using an exponential model. **(g)** Gene copy number variations (CNV) between the
544 reference and each of the accessions group by differences in the gene families: n *vs*. n: both
545 the reference and the query genomes have the same number of genes in a gene family.
546 Others categories (1 *vs*. 2, 2 *vs*. 1, 1 *vs*. 3, 3 *vs*. 1, 2 *vs*. 3 and 3 *vs* 2) indicate the number of
547 reference and accession genes in the gene families with the respective size differences. “Ref. >
548 Acc.” and “Ref. < Acc.” refer to all remaining gene families where either the reference or the
549 accession has more genes. “Ref. specific” refers to gene families which are only present in
550 the reference genome. **(h)** The number of non-reference genes found in at least two
551 accessions (blue), or found to be specific to an accession genome (yellow).




552
553

554 **Figure 3. Hotspots of rearrangements revealed by *Synteny Diversity***

555 **(a)** Synteny Diversity calculated along each chromosome; in blue: 100kb sliding windows with
556 a step-size of 50kb; in grey: 5kb sliding windows with a step-size of 1kb. The red bars under
557 the x-axes indicate the location of R gene clusters. Gray rectangles indicate the location of
558 centromeric regions. The dashed green and red lines indicate thresholds for Synteny Diversity
559 values of 0.25 and 0.50 indicative for the segregation of two (0.25) or three (0.50) non-
560 syntenic haplotypes (in a population of eight genomes). The arrow labelled with "(a)" indicates
561 a region of a 2.48 Mb inversion in the Sha genome. The arrow labelled with "(b)" indicates the
562 location of the example region show in (d). **(b)** Gene and TE densities in syntenic (SYN) and
563 hotspots of rearrangements (HR) regions. **(c)** Distribution of different copy-number alleles in

564 syntenic regions (SYN-All), hotspots of rearrangements (HR) and the remaining, partially
565 syntenic regions (SYN-Part)**. (d)** An example of an HR region which includes the *RPP4/RPP5*
566 R gene cluster. The upper panel shows the distribution of Synteny Diversity (blue curve),
567 nucleotide diversity (gray background) and haplotype diversity (pink background) in a 5kb
568 sliding window with a step-size of 1kb. Both the nucleotide diversity and the haplotype
569 diversity were calculated based on the informative markers (MAF >= 0.05, missing rate < 0.2)
570 from the 1001 Genomes Project[4]. The green and red dashed lines indicate the value 0.25 and
571 0.50 of synteny diversity, respectively. The marker density is shown as the heatmap on top.
572 The schematic in the lower part shows the location of the protein-coding genes (rectangles)
573 annotated in each of the eight genomes. In blue, gene without function implicated in disease
574 resistance. Other colored rectangles represent the resistance genes, where genes with the
575 same color belong to the same gene family. The gray links between the rectangles indicate
576 the homolog relationship of non-resistance genes. The red lines indicate the positions of HRs**.**
577 **(e)** A dot plot of Col-0 and C24 from the HR region shown in (d), where the red lines show the
578 regions with substantial homology between the two genomes. The three rows on top and the
579 three columns on the right show the location of genes on the forward strand (top), on the
580 reverse strand (middle) and the repeat regions (bottom). Genes are colored as in (d). **(f)** The
581 distribution of Synteny Diversity values in 1kb sliding windows around and in all HRs.
582

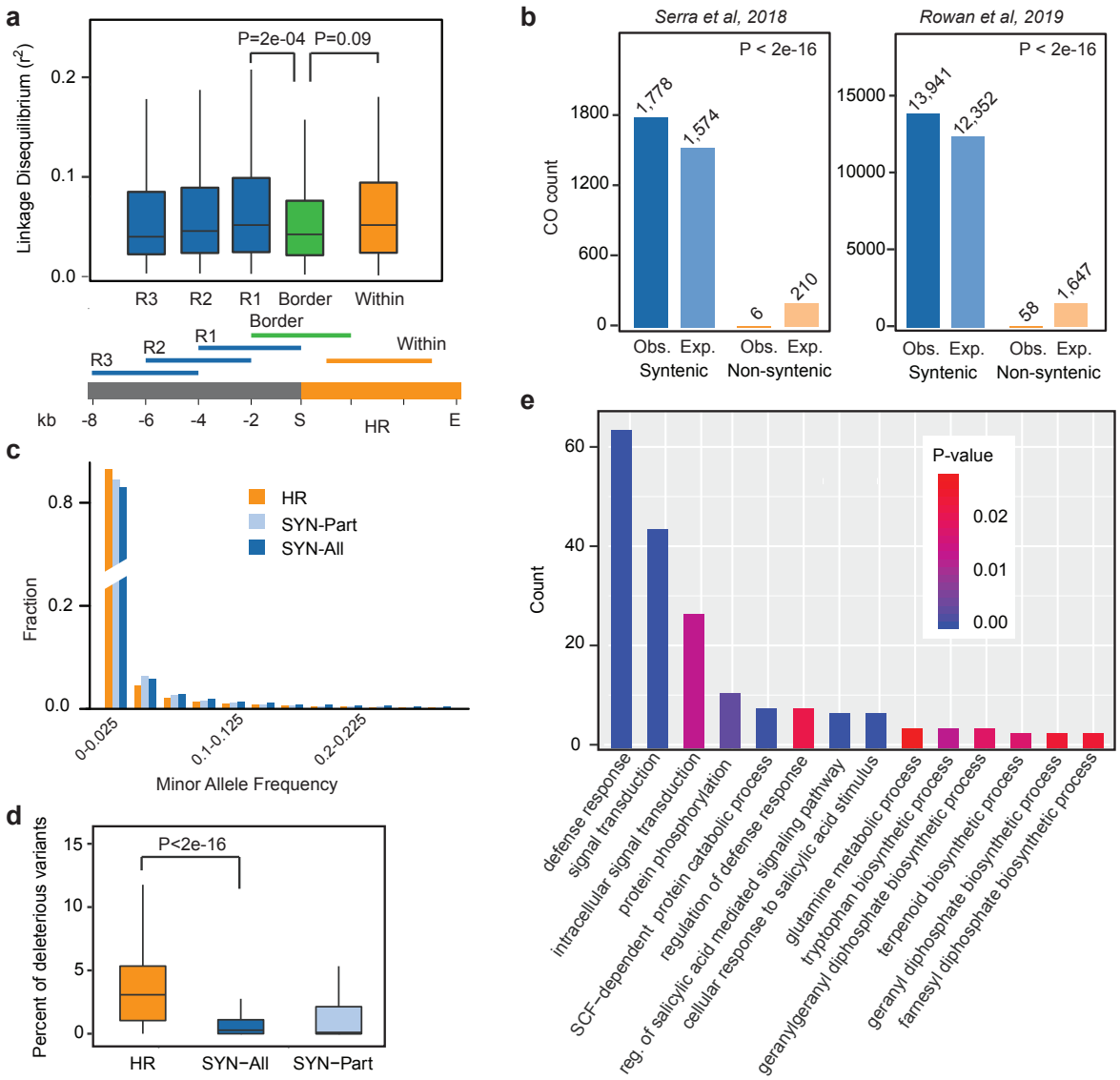



**Figure 4. Evolutionary implications of hotspot of rearranged regions**

**(a)** Linkage disequilibrium (LD) calculated in 4kb windows in and around each of the HRs. R1, R2 and R3 refer to 4kb windows up-/down-stream of each HR; "Within" refers to the 4kb in the center of the HRs; "Border" refers to the 4kb windows centered on each of the two borders of each HR. LD was calculated as the correlation coefficient ($r^2$) based on informative SNP markers (MAF > 0.05, missing rate < 0.2) from 1001 Genomes Project[4]. **(b)** Recently assessed crossover (CO) breakpoint sites[29,30] in syntenic and non-syntenic regions. Only unique CO intervals smaller than 5kb were used. Chi-square test was applied. **(c)** Fraction of SNP markers from SYN-All, SYN-Part and HRs regions across different minor allele frequency bins. The SNP markers (MAF >0.005, missing rate < 0.2) from 1001 Genomes Project were used. **(d)** Frequency of deleterious mutations in syntenic regions (SYN-All), hotspots of rearrangements (HR) and the remaining, partially syntenic regions (SYN-Part). Deleterious mutations include SNPs and small indels that introduce premature stop codons,

598 loss of start or stop codons, frameshifts, splicing sites mutations or deletions of exons. (e) GO
599 term enrichment analysis of protein-coding genes in HR regions (P-values cut-off = 0.05).
600

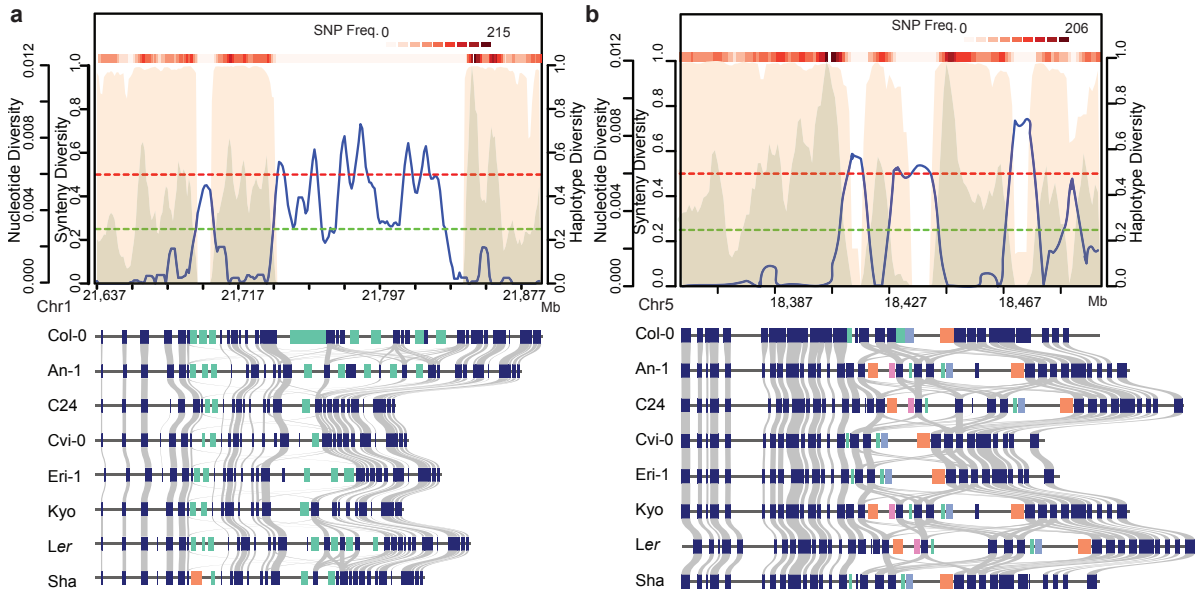



**Figure 5. Two examples for R gene clusters identified as hotspots of rearrangements.** Visualization of **(a)** the *DM6* locus (*RPP7*)[33] and **(b)** an unnamed R gene cluster on chromosome 5. Descriptions for the plots can be found in the legend of Fig. 3d.