Towards an unbiased characterization of genetic polymorphism

Anna A. Igolkina[1,†], Sebastian Vorbrugg[2,†], Fernando A. Rabanal[2,†], Hai-Jun Liu[1,†], Haim Ashkenazy[2,†], Aleksandra E. Kornienko[1,†], Joffrey Fitz[2], Max Collenberg[2], Christian Kubica[2], Almudena Mollá Morales[1], Benjamin Jaegle[1], Travis Wrightsman[2], Vitaly Voloshin[3], Victor Llaca[4], Viktoria Nizhynska[1], Ilka Reichardt[1], Christa Lanz[2], Felix Bemm[2], Pádraic J. Flood[5], Sileshi Nemomissa[6], Angela Hancock[5], Ya-Long Guo[7], Paul Kersey[3], Detlef Weigel[2,8]✉, Magnus Nordborg[1]✉

Our view of genetic polymorphism is shaped by methods that provide a limited and reference-biased picture. Long-read sequencing technologies, which are starting to provide nearly complete genome sequences for population samples, should solve the problem—except that characterizing and making sense of non-SNP variation is difficult even with perfect sequence data. Here, we analyze 27 genomes of *Arabidopsis thaliana* in an attempt to address these issues, and illustrate what can be learned by analyzing whole-genome polymorphism data in an unbiased manner. Estimated genome sizes range from 135 to 155 Mb, with differences almost entirely due to centromeric and rDNA repeats. The completely assembled chromosome arms comprise roughly 120 Mb in all accessions, but are full of structural variants, many of which are caused by insertions of transposable elements (TEs) and subsequent partial deletions of such insertions. Even with only 27 accessions, a pan-genome coordinate system that includes the resulting variation ends up being 40% larger than the size of any one genome. Our analysis reveals an incompletely annotated mobile-ome: our ability to predict what is actually moving is poor, and we detect several novel TE families. In contrast to this, the genic portion, or "gene-ome", is highly conserved. By annotating each genome using accession-specific transcriptome data, we find that 13% of all genes are segregating in our 27 accessions, but that most of these are transcriptionally silenced. Finally, we show that with short-read data we previously massively underestimated genetic variation of all kinds, including SNPs—mostly in regions where short reads could not be mapped reliably, but also where reads were mapped incorrectly. We demonstrate that SNP-calling errors can be biased by the choice of reference genome, and that RNA-seq and BS-seq results can be strongly affected by mapping reads to a reference genome rather than to the genome of the assayed individual. In conclusion, while whole-genome polymorphism data pose tremendous analytical challenges, they will ultimately revolutionize our understanding of genome evolution.

Introduction

The last 25 years have witnessed an explosion of genetic polymorphism data, fueled by the completion of the Human Genome Project and the availability of massively parallel technologies for sequencing and genotyping. Such data allow us to study population history, selection, and genetic architecture of traits—as well as the evolution of the genome itself. However, our current view of genetic polymorphism is shaped by technologies that attempt to align short sequence fragments to a reference genome in order to detect sites that differ. As a result, our knowledge of genome variation is incomplete and biased towards simple variants in regions that are easy to align—a small fraction of the genome in many species. A further source of bias arises from the use of a single reference genome.

All this is beginning to change now that long reads are making it possible to assemble high-quality, full-length chromosomal sequences from population samples. During the last couple of years, nearly complete genomes have been produced for several eukaryotic species, including yeast, animals (including humans), and many plants, including rice, tomato, soybean, grapevine, wheat, barley, maize, millet, *Brassica oleracea*, *Eucalyptus*, *Populus* and *Marchantia* sp.—as well as *A. thaliana*[1,22]. Impressive as these studies are, they have also highlighted how difficult it is to make sense of whole-genome polymorphism data, primarily because sequence alignment is not unambiguous. Pan-genome graphs[23,24] may provide elegant and computationally efficient ways of representing such data, but they do not solve this fundamental problem. To compare genomes and interpret their differences we require a modeling framework that reflects the mutational mechanisms and recombination history that gave rise to these differences, but such a framework is still largely missing.

Here we illustrate this problem by analyzing the genomes of 27 natural inbred accessions of *A. thaliana*, chosen to cover the global genetic diversity of the species. Our focus is on obtaining an unbiased and comprehensive picture of polymorphism and comparing it to existing data built over almost two decades[25–29]. To provide an unbiased picture of the "gene-ome", *i.e.*, the collection of genes across multiple genomes, we complement our genome assemblies with transcriptomes

[1]Gregor Mendel Institute, Austrian Academy of Sciences, Vienna, AT. [2]Max Planck Institute for Biology Tübingen, Tübingen, DE. [3]Royal Botanic Gardens Kew, London, UK. [4]Corteva Agriscience, Johnston, Iowa, US. [5]Max Planck Institute for Plant Breeding Research, Cologne, DE. [6]Addis Ababa University, Addis Ababa, ET. [7]Institute of Botany, Chinese Academy of Sciences, Beijing, CN. [8]Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, DE. [†]These authors contributed equally. ✉ Correspondence: weigel@tue.mpg.de; magnus.nordborg@gmi.oeaw.ac.at

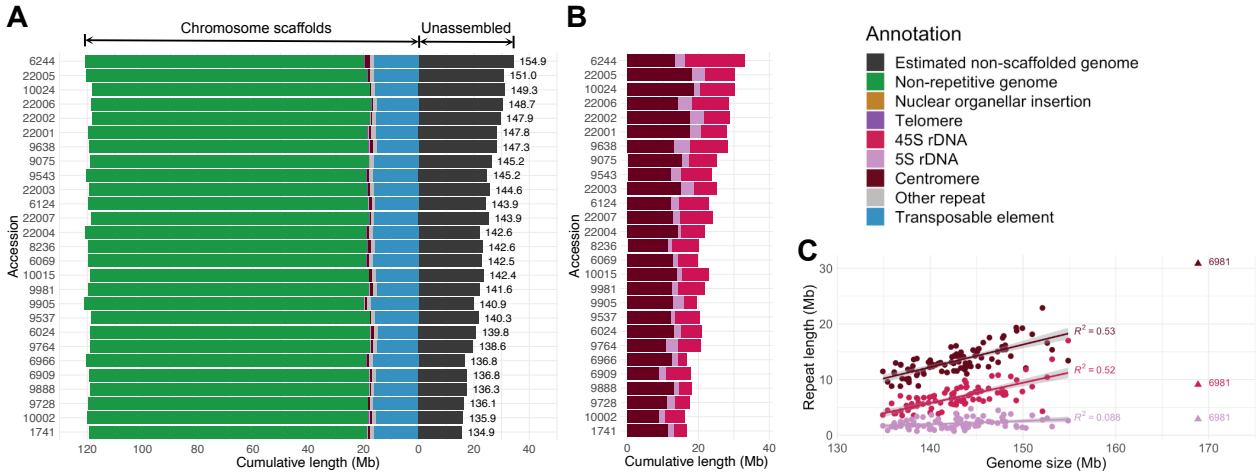


Figure 1 – Genome assemblies and size variation. A. Histogram of genome sizes estimated from k-mers in PCR-free short-reads with total assembly sizes superimposed (values to the left of "zero"). Assemblies are similar in size (∼120 Mb), with the majority of the variation in genome size attributable to unassembled regions (values to the right of "zero"). **B.** Histogram of the amount of centromeric, 5S rDNA and 45S rDNA repeats in the genome estimated from PCR-free short-reads with a BLAST-based approach. **C.** Correlation between genome size and each of the major satellite repeats. The estimated amounts of these three repeats jointly explain up to 92% ($P < 2.2 \times 10^{-16}$) of the estimated variation in genome size (see Supplementary Note 1.3 for details). The regression line and estimates in the figure exclude accession 6981 (Ws-2; indicated with a triangle), due to its extraordinarily high centromeric repeat content (∼8 Mb).

from multiple tissues for the entire sample. In addition we seek to lay the foundation for a community resource that will eventually comprise complete genomes for the thousands of natural inbred accessions that are publicly available for this model plant, thus connecting whole-genome polymorphism data to experimentally accessible germplasm and knowledge of a wide range of morphological, life-history, physiological and molecular phenotypes, as well as precise collection information (see 1001genomes.org).

Results

The organization of the *A. thaliana* genome

We selected 27 accessions to cover the global genetic variation of the species based on the original 1001 Genomes project[29], and additional samples from eastern Asia[30], Africa[31], and Madeira[32] (Extended Data Fig. 1). We sequenced their genomes with PacBio continuous long reads (CLRs) to high depth, assembled them into contigs with Canu[33], and polished these contigs with Illumina PCR-free short reads (for assembly statistics, see Supplementary Table 1). To reconstruct chromosomes, we generated hybrid scaffolds with individual Bionano optical maps for eight accessions, which allowed us to determine parameters for scaffolding[34] based on the TAIR10 reference genome[35]. Although this potentially introduces some reference bias, our focus in this paper is on the gene-dense chromosome arms, which are well-covered by large contigs, and appear to harbor few large-scale rearrangements (as we shall see below).

The assemblies reflect the organization of the genome. Like many plants, *A. thaliana* has experienced recent episodes of TE activity, leading to nearly identical sequences inserted across each genome[36]. These make short-read alignments very difficult, but the PacBio CLR technology used here produced reads long enough to bridge such insertions. However, extensive tracts of identical or near-identical tandem repeats, such as centromere satellites and 5S rDNAs, consistently break our assemblies (Extended Data Figs. 2-3). We note that centromeres can now be assembled with PacBio HiFi reads. While CLRs are less accurate, they are on average longer, hence marginally better for chromosome arms[15,37]. 45S rDNA clusters remain challenging regardless of technology[38].

Our 27 assemblies are all ∼120 Mb in size. This is noteworthy because we estimate that the full genomes range from 135 to 155 Mb (Fig. 1A). This variation is consistent with previous results[39,40], and must be due to the unassembled portions of the genomes. Using a BLAST-based approach, we estimate that centromeres and rDNA clusters alone can account for up to 92% of the estimated variation (Fig. 1B-C). That 45S rDNA plays a major role was already known[40], but the extent of the contribution of centromeric repeats is a new finding. Our results confirm that TEs do not contribute significantly to variation in genome size within *A. thaliana*[40]—contrary to the major role they play in inter-specific variation[41]. Individual TE families can vary greatly in size across accessions (Extended Data Fig. 4), but the cumulative effect of all TEs

on genome size variation appears to be small.

Detecting structural variation

Consistent with other studies[22], the chromosome arms are not only similar in length across accessions, but are also largely syntenic—as expected for a sexually reproducing organism with normal recombination (Fig. 2A). This said, thirteen Mb-scale rearrangements were readily apparent in whole-genome alignments. These include the previously described 1.2-Mb paracentric inversion associated with a heterochromatic knob on chromosome 4[42,43], which is accompanied in a subset of the knob-less accessions by an 877-kb inversion that is 100 kb closer to the centromere (Extended Data Fig. 5; Supplementary Table 2), and a very large reciprocal translocation in accession 22001 from the Yangtze River region, which has swapped the distal portions of chromosomes 3 and 5 (Supplementary Fig. 1). This latter rearrangement, which would presumably lead to decreased fertility in heterozygotes, appears to be rare, as the translocation was not detected in any of the other 117 accessions from the region that had previously been sequenced with Illumina short reads[30]. This rearrangement was treated in the analyses as if the part had remained in their original positions (see Supplementary Note 1.1).

Obvious large-scale rearrangements aside, a comprehensive characterization of structural variants (SVs, by which we mean any alteration that causes variation in length, orientation, or local context of sequence) remains challenging. While SVs, along with SNPs, can be identified in genome alignments, the characterization of SVs is a fundamentally different problem from SNP-calling. The latter can be viewed as a technical issue—how to distinguish single-nucleotide polymorphisms from sequencing errors—but SV-calling is challenging even with flawless chromosomal sequences. The reason is that the SVs identified between genomes depend on the alignment method and parameters used, and there is no obvious ground truth. Given these uncertainties, we pursued two distinct approaches.

First, we developed a whole-genome multiple-alignment pipeline (Pannagram), which produces an intuitive representation for genome-browser visualization. This approach can be considered an extension of pairwise-alignment methods[44] to handle multiple genomes in a reference-free manner. We developed a pan-genome coordinate system based on the resulting alignment and anchored it to the TAIR10 reference genome. Second, we used the Pan-Genome Graph Builder (PGGB) pipeline[24], which yields a graph representation of multiple genomes by collapsing identical sequences into single nodes, connected by multiple paths forming bubbles where the genomes differ. The resulting object is computationally efficient for genotyping, but it can also be highly complex, with neither nodes nor bubbles having an obvious biological interpretation.

Comparing the SVs identified by these approaches proved to be challenging. In terms of base-to-base overlap, we found

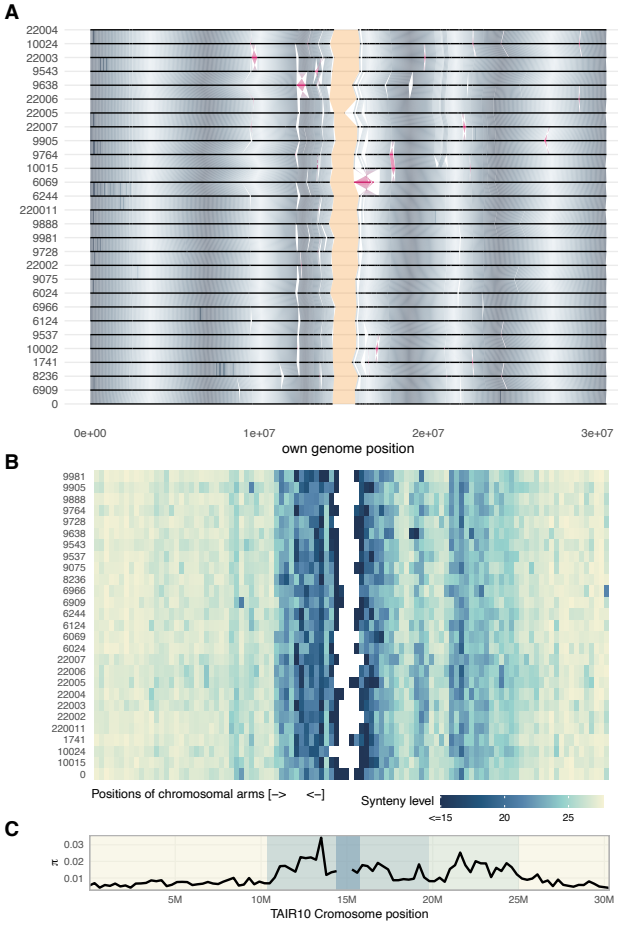


Figure 2 – Distribution of variation on chromosome 1. A. Whole-genome alignment of the assembled parts illustrates that the genome is generally structurally conserved away from centromeric regions. The chromosomes are aligned from both ends to emphasize the contiguity of the arms, with yellow color filling the unassembled centromeric regions. Inversions are shown in pink. Oscillating gray shades highlight homologous regions (the periodicity of the gradients reflects repeated use of the color scheme, and has no biological meaning). For chromosomes 2-5, see Extended Data Fig. 5. **B.** The density of pan-genome graph bubbles, reflecting higher diversity in pericentromeric regions and at ∼20 Mb, where an ancestral centromere was lost through chromosome fusion since the divergence of *A. thaliana* from the rest of the genus[40]. Synteny level refers to average sharing of links (1-27) between nodes across the 27 genomes in 300 kb blocks. **C.** Distribution of nucleotide diversity based on SNPs called from the multiple alignment. The dark blue zone corresponds to the centromeric region, and light blue highlights the pericentromeric area as well as the the old centromere[40].

that the variation captured by the PGGB encompassed almost all variation identified by Pannagram and covered nearly twice as much sequence (see Extended Data Fig. 7A). This is most pronounced in pericentromeric regions, which generally harbor more polymorphism (Fig. 2B-C), but the phenomenon is genome-wide (Extended Data Fig. 7B). We identified three different reasons for this.

First, the presence of physically distant but closely related sequences (*e.g.*, reflecting recent TE activity) often leads to large loops in the graph that may or may not reflect actual SVs. Masking repetitive sequences will reduce this problem, but requires good repeat annotation—and would also make it impossible to study genome-variation comprehensively (one of our goals in this paper).

Second, even in the case of short-range duplications, the graph combines duplicated sequences into a single node and hence detects all loci as SVs, even if only one of them is variable (Extended Data Fig. 8).

Third, PGGB and Pannagram rely on different alignment parameters. In order to reduce the number of uninterpretable nodes, PGGB utilizes strict similarity criteria, whereas Pannagram uses more relaxed cutoffs to maximize homology detection. The *A. thaliana* genome contains many highly diverged regions[45], and these tend to be treated by PGGB as long SVs, whereas Pannagram tries to align them, often identifying a number of shorter variants. Whether Pannagram or PGGB results are more biologically relevant ultimately depends on the origin of the region and the cause of the high divergence (Extended Data Fig. 8).

A final difference in SV-calling between PGGB and Pannagram lies in how nested length variants are represented. PGGB shows these as easily interpreted loops-within-loops, whereas Pannagram treats them as complex SV regions (Extended Data Fig. 8). This, notably, does not affect the regions covered by SVs, but causes differences in the number of variants called.

Since Pannagram provides straightforward SVs, convenient pan-genome coordinates, and all its structural variants are covered by PGGB's SVs, we decided to base our analysis on Pannagram's SVs.

Characterization of structural variants

SVs come in many types and sizes, reflecting diverse mutational mechanisms. In what follows, we will focus on length variants, which are by far the most numerous (we identified fewer than a hundred inversions, for example). We further divide them into bi-allelic presence/absence polymorphisms, consistent with *simple* insertion/deletion (indel) mutations (sSV), and more *complex* multi-allelic polymorphisms (cSV) (Extended Data Fig. 6). We primarily consider the former, as they are easier to interpret, and also constitute the majority (Fig. 3A), especially in the chromosome arms—there is a pronounced increase in the fraction of cSVs near the centromeres (Fig. 3B), in agreement with the overall higher polymorphism levels in these regions (Fig. 2B-C).

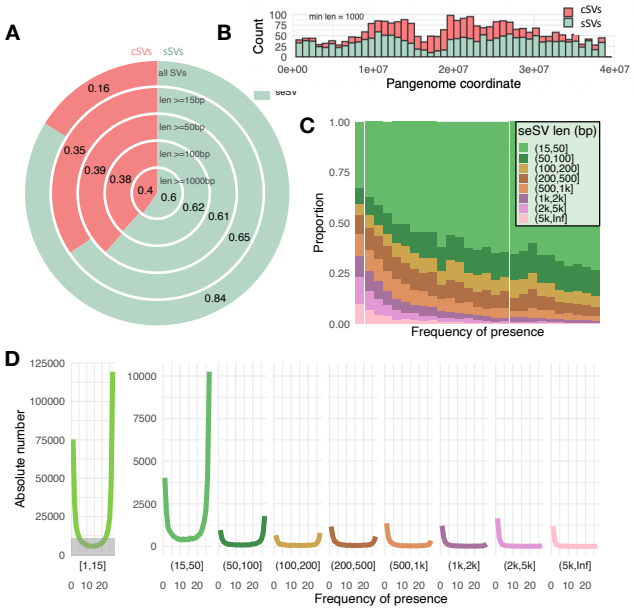


Figure 3 – Structural variation. A. Fraction of sSVs and cSVs by length. The prevalence of short sSVs is at least partly explained by the low probability of the multiple events required for cSVs in small DNA segments. **B.** The distribution of sSVs and cSVs across chromosome 5—the latter are enriched in pericentromeric regions. **C.** The proportion of length classes of sSVs in each frequency bin. **D.** Frequency of the presence allele of all sSVs with length $\geq$ 15 bp, by length of variant. The height of the gray block in the left panel is equivalent to the height of the complete panel on the right.

We identified 557,473 sSVs, affecting over 37 Mb in total—although the length-distribution is heavily skewed towards short variants (Fig. 3C-D). To gain further insight into the nature of these polymorphisms, we consider sSVs of length between 15 bp and 20 kb, because shorter variants are sensitive to details of alignment, and larger ones are too few for statistical treatment. Determining the ancestral state of these variants using an outgroup species was in general impossible, as the homologous regions are typically intergenic and difficult to align. However, the frequency distribution for the presence-allele of sSVs was consistent with sSVs mostly being due to rare insertions or rare deletions (Fig. 3D). Specifically, long presence alleles tend to be rare, while short ones tend to be common—suggesting that sSVs are mainly caused by long insertions and short deletions. As noted, both types are more common in intergenic than in genic regions, and they are also more common in introns than in exons, consistent with purifying selection removing these kinds of variants (Extended Data Fig. 9).

Finally, rare nuclear insertions of organellar DNA are easy to spot. We found 108 such insertions, almost all of them singletons or doubletons, and ranging from a few hundred bp

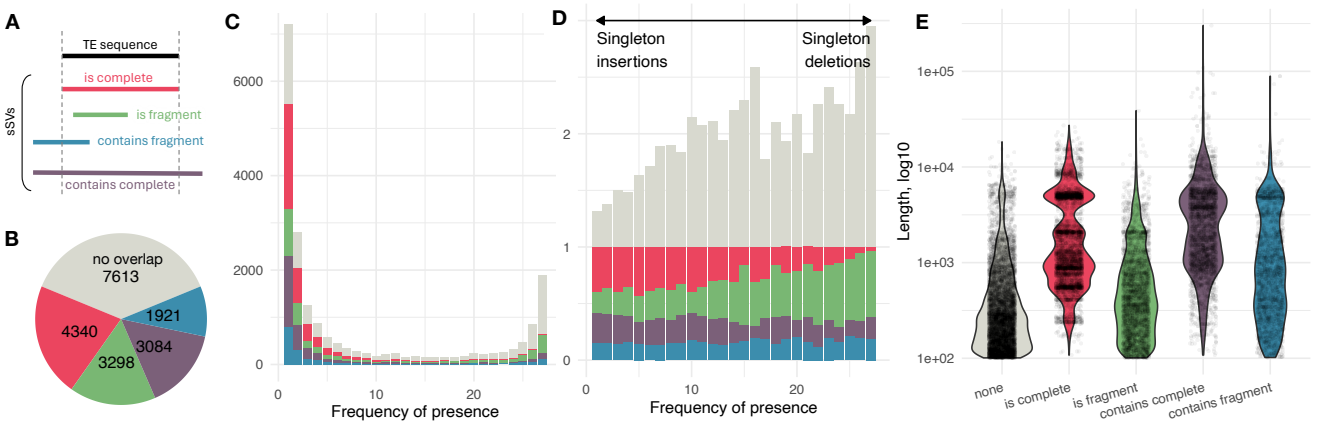


Figure 4 – The role of TEs in bi-allelic indels. A. Categories of overlap between presence allele of sSVs and TE sequence from annotation. **B.** Total annotated TE content in sSVs. **C-D.** Annotated TE content as a function of allele frequency. **E.** Distribution of the length of presence-alleles in different categories.

to entire organellar genomes (Extended Data Fig. 10, Supplementary Table 3). None of our genomes, other than 6909 (corresponding to the reference accessions Col-0) shared the large nuclear insertion of mitochondrial DNA in chromosome 2 of the TAIR10 reference genome[37,46,47].

Structural variants and annotated TEs

TEs are responsible for structural variants of diverse nature, generating insertions and deletions when active, and by serving as templates for various mutational processes *e.g.*, double-stranded break repair[48]. To investigate the role of TEs in generating structural variation, we classified the presence alleles of 20,256 sSVs of length $\geq$ 100 bp with respect to their overlap with the sequences of $\sim$35k annotated TEs, covering $\sim$15% of the *A. thaliana* reference genome (*cf.* Fig. 1A). We defined different categories reflecting the extent of overlap (see Fig. 4A; for an analysis broken down by TE family, see Supplementary Note 3). In total over 60% of sSVs showed some overlap, confirming a strong connection between our sSVs and TEs (Fig. 4B).

We showed above that putative insertions (presence allele in 1-3 accessions) tend to be longer. They also more likely correspond to complete TE sequences, supporting their being recent TE insertions (Fig. 4C-D). At the other end of the frequency distribution, putative deletions (absence allele in 1-3 accessions) are more likely to correspond to fragments of TEs, suggesting that they reflect decaying TEs (Fig. 4C-D). We tested this latter claim by examining the flanking regions of sSVs similar to fragments of TEs: 73% of putative deletions were part of a TE, while the same was true for only 35% of putative insertions.

As expected, sSVs corresponding to complete TEs are enriched for particular lengths (specifically around 5 kb), reflecting active TE families (Fig. 4E). Similar patterns of enrichment were found for sSVs in all other categories except those corresponding to fragments of TEs. This supports recent reports that active TEs are far from perfectly annotated[49,50].

The mobile-ome

Our species-wide sample of genomes makes it possible to look directly for genetic elements that are actually mobile—the mobile-ome—without relying on annotation. Actively mobile elements will generate insertions and deletions with similar sequence in different locations throughout the genomes. To identify these, we clustered all presence-alleles from sSVs based on sequence similarity, and represented the output as a graph of nestedness, where nodes represent sequences. The graph consists of many connected components, which we expected to correspond to distinct TE families or subfamilies (see Supplementary Note 3).

Almost all sSVs that corresponded to complete TEs are part of the graph, consistent with their being part of the mobile-ome (Fig. 5A). Among other TE-related sSVs, approximately 70% are linked to the graph, and are thus also related to the mobile-ome, presumably reflecting a mixture incompletely annotated TE families, complex insertion behaviors, and deletions within active TE families. The remaining 30% could result from low-activity TEs or decay of inactive TEs.

To understand the relationship between annotated TEs, we filtered out sequences without TE content, and merged sequences with high similarity into larger nodes. The resulting subgraph consists of one dominant component along with numerous smaller ones (Fig. 5B). As expected, there are multiple large nodes corresponding to complete TEs, but we also see large nodes corresponding to sequences ostensibly containing complete TEs or containing TE fragments, demon-

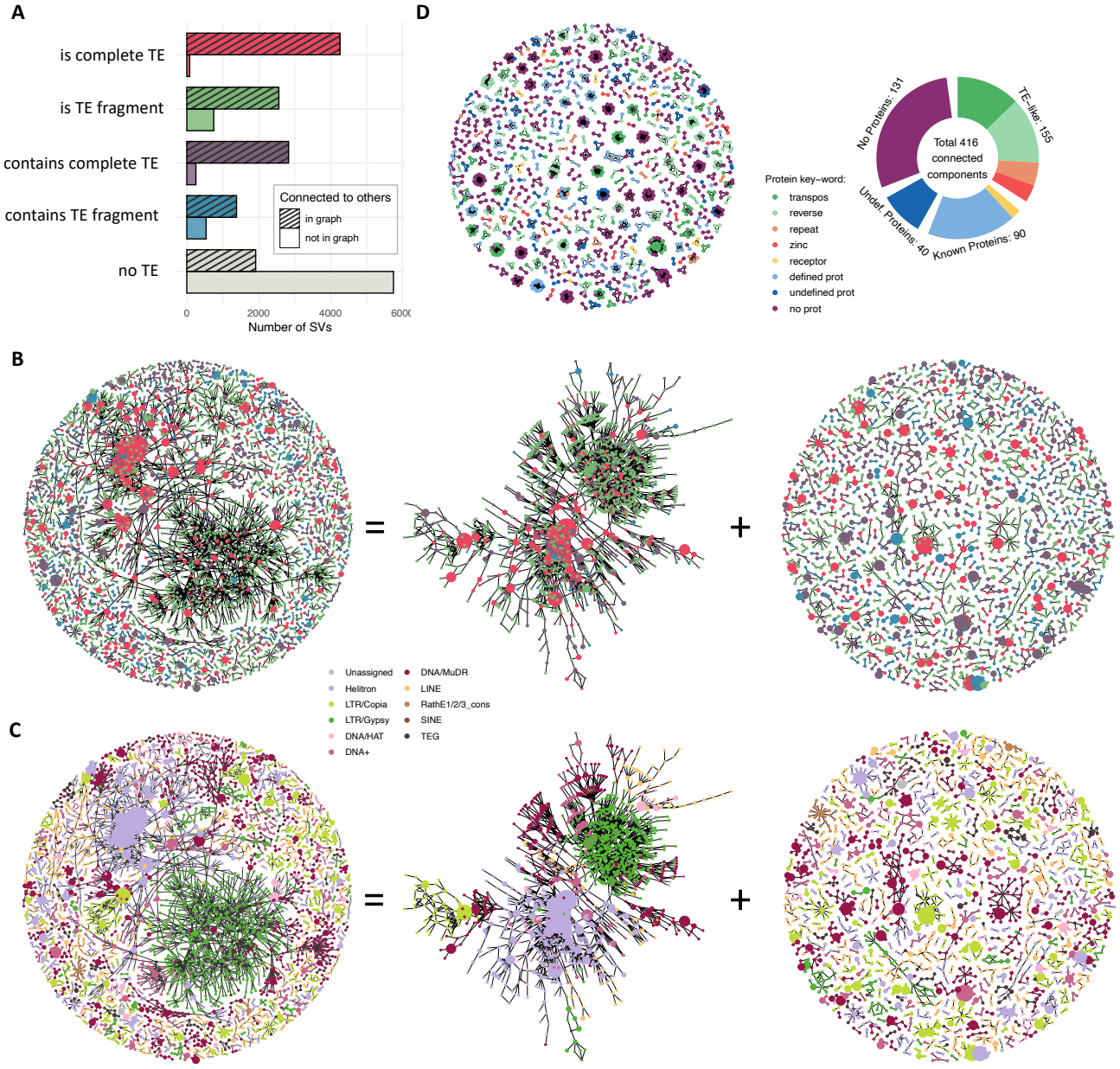


Figure 5 – The graph structure of sSVs. A. Comparison of the number of sSVs that is included with those not included in the graph of nestedness, by TE content. **B.** Graph illustrating the nestedness of sSVs, including only presence alleles with non-zero TE content. Each node represents a set of mutually nested sequences, using a similarity threshold 0.85, with node size reflecting the number of sequences included. Node are colored by TE content. **C.** The same graph as in B, but colored by TE families. Note that the algorithm used no prior knowledge of TE families. A large, dominant component connecting all TE families is also seen in a graph built using the entire *A. thaliana* TE annotation, suggesting that our sSVs and the known annotation have comparable properties (Extended Data Fig. 12). **D.** A graph of nestedness for sSVs without TE content, colored by ORF content based on protein BLAST.

strating that many of these sequences are in fact part of the mobile-ome (Extended Data Fig. 11). Coloring the nodes by TE family reveals that, while the smaller components mostly can be attributed to individual TE families, the dominant component contains members of all known TE families (Fig. 5C). Possible reasons include TEs inserting into existing TEs, and, of course, incorrect TE annotation (*e.g.*, Extended Data Fig. 13). Further evidence for the incompleteness of the TE annotation comes from small graph components that contain large nodes of longer elements that contain only some TE content. These nodes presumably correspond to un- or mis-annotated TE sub-families (Extended Data Figs. 14–15).

In the set of sSVs without TE content, the majority (75%) is not connected in the graph, presumably reflecting unique events. Some are gene duplications; some may also be un-annotated TEs that are too rare in our limited set of genomes to be detectable by our method. Among the remaining 25%, we observed 416 connected components, similar in structure to small components corresponding to single TE sub-families in the analysis above (Fig. 5D). We hypothesized that these components contain previously un-annotated TEs. Protein BLAST analysis of ORFs from the sequences of these components revealed that 155 of them are indeed TE-like in that they are similar to TE proteins in *A. thaliana* or other species. Among these, we observed TE subfamilies that showed evidence of horizontal gene transfer, as they have greater protein similarity to species outside a panel of five Brassicaceae species (Extended Data Fig. 16). We also found potentially new mobile element families lacking clear protein-coding potential. They form relatively large components in the graph (Fig. 5D, purple islands), composed of multiple sequences of similar length. Theey are not low-complexity and they are exclusive to *A. thaliana* (Extended Data Fig. 17).

Annotated TEs in *A. thaliana* are generally methylated and transcriptionally silenced. However, most published results have relied on a single reference genome, without distinguishing between active and inactive TEs. Our data allow us to begin to address the distinction between active and inactive TEs, and also to consider the age of insertions, which which should be proportional to their population frequency. All sSVs that are part of the graph of nestedness are highly methylated, consistent with their being part of the mobile-ome (Extended Data Fig. 18). Methylation levels on sSVs with TE content generally increase with the frequency of the presence allele in the population, indicating that older insertion are more likely to be methylated. Methylation levels on sSVs without annotated TE content show the opposite pattern, which is expected given that the subset of these sSVs that are nevertheless part of the mobile-ome (and hence highly methylated), are mostly rare in the sampled genomes. Levels of methylation are highest on sSVs corresponding to complete TE sequences or TE fragments, followed by those that contain complete TE sequences or contain TE fragments. The methylation on sSVs without TE content is bimodal, because this category contains a mixture of unmethylated unique sSVs (some of which are genes) and heavily methylated un-annotated members of the mobile-ome. Expression levels of all putative TEs, annotated or not, are low relative to sSVs without TE content (Extended Data Fig. 19).

The gene-ome

To investigate how much protein-coding genes vary among accessions while minimizing reference-bias, we annotated each genome independently with sequence-based gene modeling and RNA-seq data from four different tissues[51]. To connect the annotations, we used the pan-genome coordinate system (rather than matching orthologs[22], which is the only possibility for distantly related genomes). Despite 83% of genetic loci having a one-to-one correspondence in the aligned genomes, gene model predictions varied considerably between accessions. While some of these differences likely represent genuine genetic variation, we lacked data to distinguish this from artifacts, and thus adopted a majority voting approach to harmonize annotations (Methods and Supplementary Note 4.1). After including 1,596 TAIR10 genes that were not re-identified in our study (unsurprisingly given that we used only four tissues[35,52]), the final annotation contained a total 34,819 putative genes, ∼25,000 of which are *bona fide* protein-coding genes (Fig. 6A; Supplementary Note 4.2). Of these, 2,265 were not previously annotated (for a detailed analysis, see Supplementary Note 4.3).

Focusing on presence-absence variation of genomic loci rather than annotated gene models, we found that 13% of loci were segregating in the sample (Fig. 6B). This variation could reflect deletions (in some accessions) of ancestral genes, or segregating insertions specific to *A. thaliana*. To resolve this, we compared our genes to those in *A. lyrata* in order to determine whether they were present and, if so, whether they were syntenic (Fig. 6C). This analysis revealed a striking difference between TAIR10 and the previously un-annotated genes: while most TAIR10 genes are ancestral, most newly identified genes are not (Fig. 6D). There was also a notable difference between segregating and fixed genes: as expected, most of the latter are ancestral, yet, surprisingly, the vast majority of the segregating genes is not (Fig. 6E). Although it is likely that a fraction of ancestral genes is underestimated due our reliance on the *A. lyrata* gene annotation, our results strongly suggest that segregating ancestral deletions are rare.

Segregating genes are somewhat enriched near centromeres, while fixed genes are clearly more common in the arms (Fig. 6F). Syntenic ancestral genes are also enriched in the arms, while other categories are evenly distributed across the genome. Not unexpectedly, TE genes of higher frequency are more likely to be found near centromeres, consistent with selection against TE insertions in the arms[36].

Finally, there is a striking difference in expression between fixed and segregating genes: even genes that are only absent in one or two accessions tend to be almost silent in the remaining accessions (Fig. 6G). Consistent with this, epigenetic

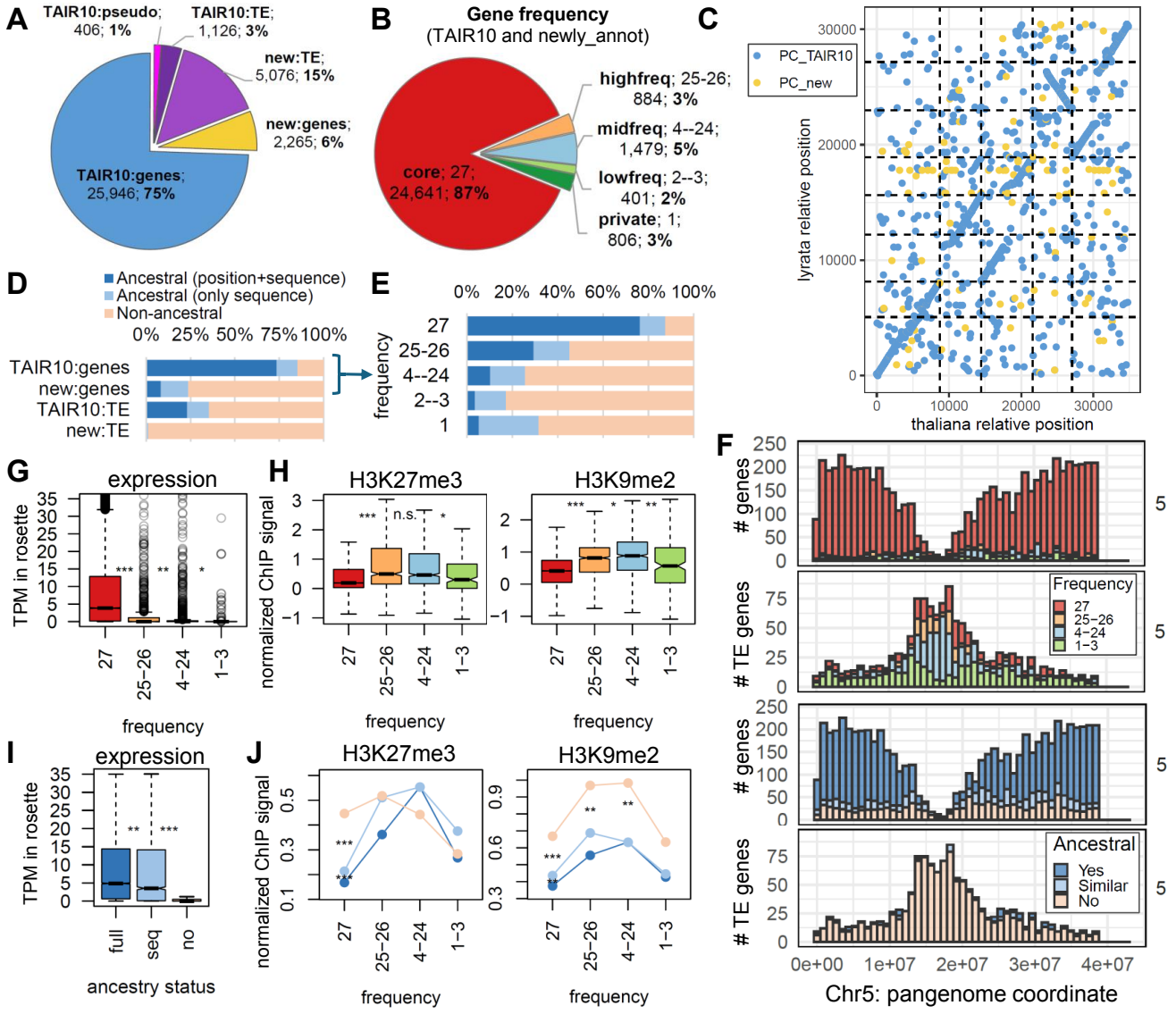


Figure 6 – Analysis of the gene-ome. A. The 34,819 annotated genes categorized based on whether they correspond to the TAIR10 annotation (TAIR10:genes, TAIR10:TE, TAIR10:pseudo), or not (new:genes, new:TE). Separation into genes and TEs is based on either high nucleotide similarity to annotated TEs, or amino acid similarity to proteins involved in TE function. **B.** Genes categorized based on the frequency of the gene locus across 27 genomes. For TE gene frequency distribution and the comparison of gene locus to the frequency of gene model presence, as well as to previously published results[22], see Extended Data Fig. 20A–B. **C.** Synteny analysis using *A. lyrata*. Genes are plotted not by their chromosomal position but by their consecutive gene IDs along the chromosomes. For TE genes, see Extended Data Fig. 20C-D. **D.** Ancestral status of TAIR10-genes and newly annotated genes. **E.** Ancestral status of genes by frequency of presence. **F.** Chromosome 1 distribution of genes (TAIR10:genes and new:genes) and TE genes (new:TE) by frequency (top) and ancestral status (bottom). For other chromosomes, see Extended Data Fig. 21. **G.** Gene expression by frequency in 9-leaf rosette data[51]. Results for accession 6024 are shown; for other accessions see Extended Data Fig. 22. Only genes present in each accession are plotted. **H.** H3K27me3 and H3K9me2 silencing for genes of different frequencies of presence allele. Quantile-normalized H3K27me3 and H3K9me2 levels in mature leaves of accession 6024 are shown[51]. For DNA methylation levels, 24 nt sRNA-coverage, and for the summary of the trends for all accessions, see Extended Data Fig. 22. **I.** Expression of genes by ancestral status. Only genes present in accession 6024 are plotted; for other accessions see Extended Data Fig. 22. **J.** Median H3K27me3 and H3K9me2 levels for genes by ancestral status and frequency of presence allele. Colors as in panel D. All significance levels were estimated using the Mann-Whitney test (***: $P < 10^{-10}$, **: $P < 10^{-5}$, *: $P < 10^{-2}$, n.s.: $P \geq 10^{-2}$).

silencing is a function of population frequency: both gene-like Polycomb silencing (H3K27me3) and TE-like silencing (H3K9me2 and DNA methylation) are highest for genes of intermediate frequency (Fig. 6H). Likewise, non-ancestral genes show a drop in expression and increased silencing, with H3K9me2 elevated for all gene frequencies (Fig. 6I–J).

The pan-genome

The term "pan-genome" is currently used with a variety of meanings, from the original use in prokaryotes to describe the observation that genomes from different strains of the same species vary enormously in gene content[53], to the human genetics ambition of representing all polymorphism in a "pan-genome graph"[14,23]. We will discuss the utility of these concepts further below—here we simply consider how the pan-genome grows with sample size, and why. As shown in Figure 7A, all components of variation considered in this paper grow with sample size, but at different rates: the mobile-ome grows faster than the full genome and the gene-ome more slowly, consistent with the former being under stronger purifying selection. All components increase faster than the logarithmic growth expected under neutrality in a constant population, and more slowly than the linear growth expected in an exponentially growing population[54].

The growth of the pan-genome is reflected in the coordinate system, and is not uniform along each chromosome, because most of the variation is found in centromeric regions (Fig. 7B). Already with 27 accessions, the pan-genome is 40% longer than the TAIR10 reference genome.

Missing polymorphism

As part of the 1001 Genomes Project, we previously "re-sequenced" 1,135 accessions using short reads[29]. We were well aware that the data were both incomplete and error-prone: we only called SNPs and short SVs, and only an average of 84% of the reference genome was covered by short reads from any particular accession.

With our whole-genome polymorphism data, we are now in a position to assess how much variation was previously missed. In the 1001 Genomes data, a pair of accessions differed, on average, at ∼440k SNPs. In our whole-genome alignment, the corresponding number is 600k–800k SNPs, depending on how SNPs are defined (see Methods). In other words, we previously missed 25–45% of the SNP variation. In addition, whole-genome alignments of two genomes reveal on average ∼190k SVs (of length < 10k bp) covering a total over ∼11.6 Mb of sequence—nearly 10% of the assembled genome (Fig. 7C).

We investigated the causes of the missing SNPs by calling SNPs for our 27 genomes using the PCR-free, high-coverage short reads that were generated for this study, and comparing the results to those from whole-genome alignment (Supplementary Note 5). We found the main reason for missing

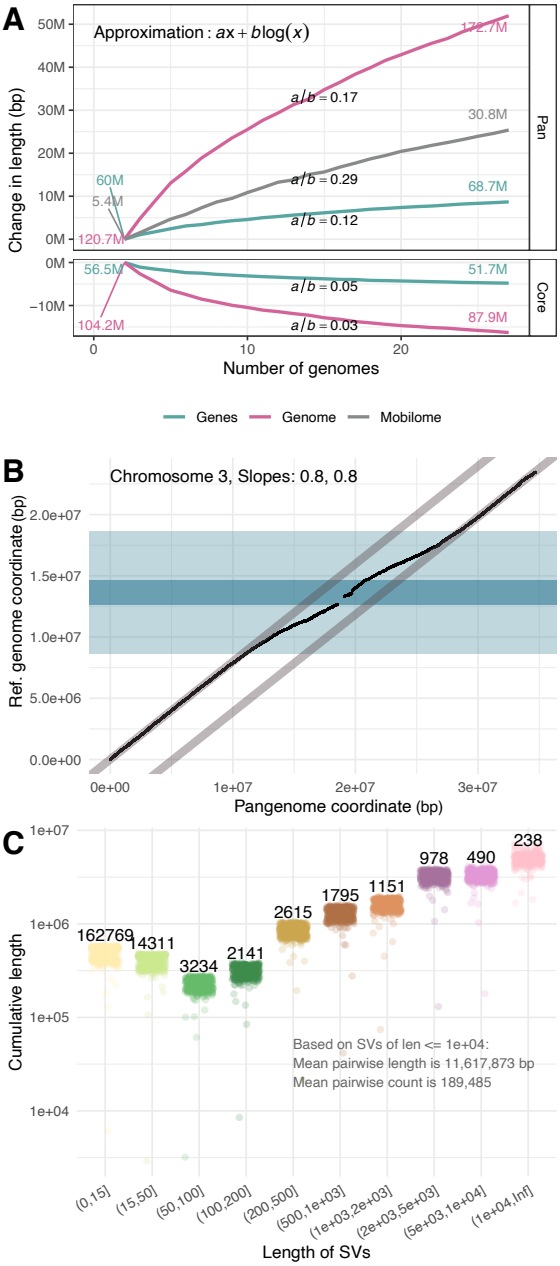


Figure 7 – The growth of the pan-genome. A. The dependence on sample size for the union ("pan") and intersection ("core") sequence length, separately for the full genome, the mobile-ome, and the gene-ome (numbers and normalized views are in Extended Data Figs. 23–24). **B.** Pan-genome vs. reference genome coordinates. The pericentromeric region (light blue) shows a higher "dilution" of the spatial coordinates due to the increased number of SVs in this region, but note that the slope is never 1, reflecting their ubiquity. The centromeric region is dark blue (for other chromosomes, see Extended Data Fig. 25). **C.** Contribution of sSVs to differences between pairs of genomes, by length of the sSV. Numbers are counts of sSVs. Shorter sSVs are more numerous, but longer and rarer sSVs contribute more to length variation.

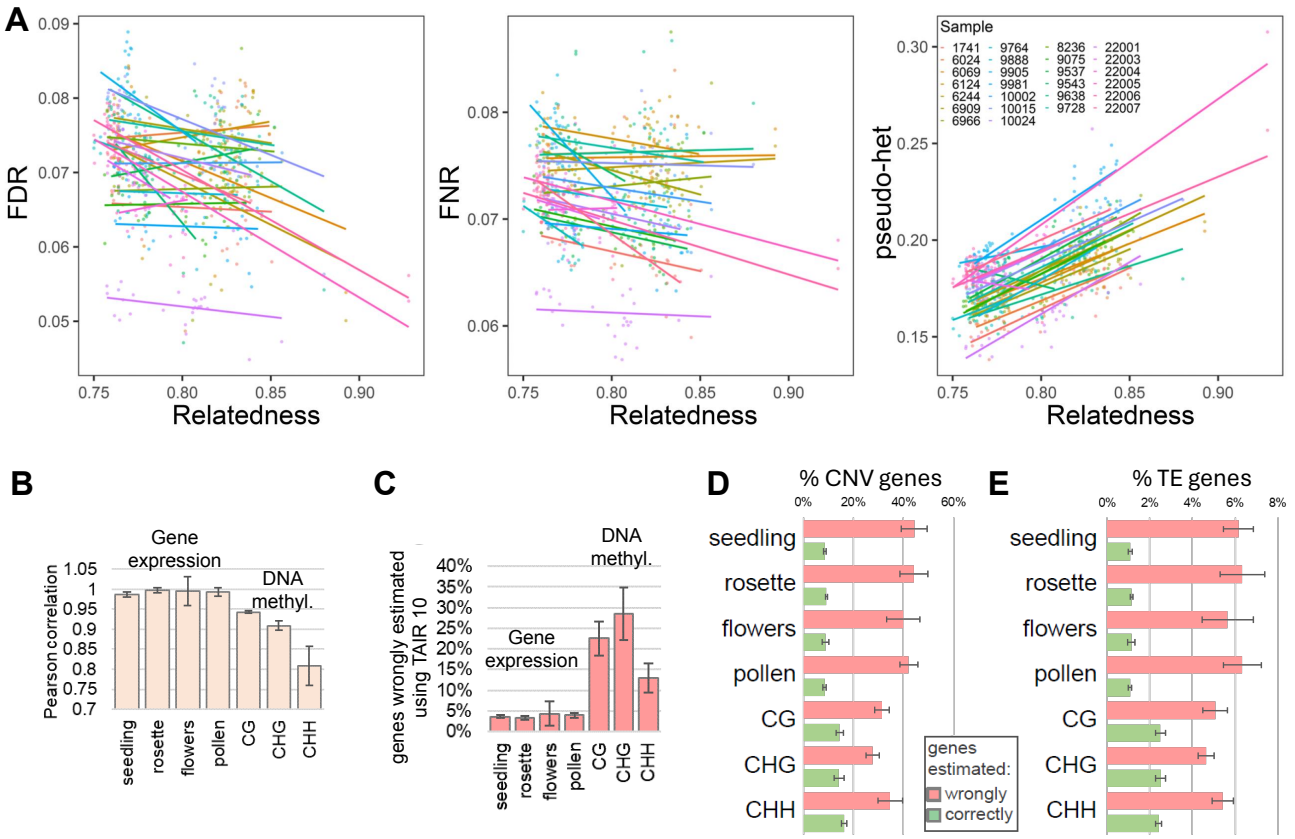


Figure 8 – Read-mapping and reference bias. A. SNP-calling error rates systematically depend on the choice of reference genome and the relationship between the reference and the sampled genome. Each line is the regression of SNP-calling errors for each sample mapping to all the others with varying degrees of relatedness. Samples are identified by colors. See Supplementary Note 5 for details. **B.** Correlation between expression- and methylation-level estimates derived from mapping reads to the TAIR10 reference genome and the genome from the sampled individual. **C.** The percentage of genes for which expression or methylation levels differ by more than 30% or 50%, respectively, when mapping reads to TAIR10 instead of to their "own" genome. Example expression scatter plots in **D-E.** Enrichment of copy-number variable genes and TE genes among those with discordant expression- or methylation-level estimates.

SNPs from short reads to be that such sites lie in regions where no calls are made, because the regions are not reliably covered by reads due to mapping problems. The extent of such regions depends on the mapping parameters used, but less conservative read-mapping will generally come at a cost of higher error rates. In our test SNP-calling with PCR-free data, we were able to reduce the fraction of missing SNPs to below 20%, but our FDR was then ∼7%. Consistent with a trade-off between conservative and aggressive read-mapping, *bona fide* SNP-calling errors, be they false positives or negatives, were overwhelmingly due to read mismapping (Supplementary Fig. 12). Rampant pseudo-heterozygosity caused by segregating duplications that are absent from the reference genome is particularly worrisome, in agreement with previous observations[49]. Finally, differences in local sequence alignment between algorithms also contributed to discrepancies, whereas traditional base-calling errors played an insignificant role.

Reference bias

It is obvious that read-mapping to a reference genome can cause reference bias, *i.e.*, results depend on which genome is used as reference. Because we now have multiple genomes to choose the reference from, our data allow for systematic investigation of this problem. Starting with SNP-calling, we found that all SNP error rates depend both on the reference genome and on the relatedness between the reference and the sampled genome, often in unpredictable ways (Fig. 8A; Supplementary Note 5). Since many population genetic analyses rely on SNPs to estimate the relatedness between samples, this is troubling, and the consequences of non-randomly varying levels of bias in samples for downstream analyses such as GWAS or demographic inference merit further investigation.

Reference bias may also affect standard -omics techniques that quantify molecular phenotypes by mapping short reads to a reference genome rather than to the genome of the individual being studied. We illustrated this problem for tran-

scriptome and methylome profiling by comparing the result of mapping RNA-seq and BS-seq reads to the TAIR10 reference as well as to the actual genome of the accession in question. Gene expression estimates between the two approaches were strongly correlated on average (Fig. 8B), but a subset of genes diverged markedly (Fig. 8C, Extended Data Fig. 26), and these were strongly enriched for copy-number variable genes and TE genes (Fig. 8D-E). Methylome profiling was even more sensitive to the choice of genome for mapping, not surprising given that methylation in *A. thaliana* predominantly targets TEs, which vary greatly in copy number between accessions and are prime targets for cryptic mismapping. A scan for Differentially Methylated Regions (DMRs) in 100-bp windows across the genome revealed that profiling methylation by mapping reads to the TAIR10 reference genome rather than to each sample's own genome produced a large number of spurious DMRs (Supplementary Fig. 17 and Supplementary Note 5). How serious this problem is will depend on the application, but it seems clear that it must be considered.

Discussion

Over the last several years, population samples of more-or-less complete eukaryotic genomes have been appearing at an increasing rate, and there has been much excitement over the (variously defined) "pan-genome"[55]. At the same time, it has become abundantly clear that characterizing all the differences between individual genomes is extremely difficult. The problem is manifestly not a technical one, because even with perfect chromosomal sequences, we have to decide how to align them and how to interpret the differences. For complex structural variation, especially in highly diverged regions[45], this is not trivial, and there is no obvious "gold standard" by which to evaluate algorithms. Furthermore, the ultimate reason for comparing genomes matters. If we are interested in using the pattern of polymorphism to answer questions about evolutionary history and mutational origins, we must interpret the pattern using models that reflect actual historical events. However, if the goal is simply to develop genetic markers, it may be irrelevant whether there is any correspondence between designated variants and the molecular processes that generated them.

With these caveats in mind, we tried two different approaches. First, we aligned the genomes using Pannagram, a whole-genome multiple-alignment pipeline, and used it to create a pan-genome coordinate system anchored to the TAIR10 reference genome, making it possible to visualize the alignments using familiar genome browsers (see Data availability). Second, we created a pan-genome graph using PGGB, which was develop for human data[14,24]. Comparison of the results revealed systematic discrepancies that reflected biological differences between humans and *A. thaliana*, which has higher levels of polymorphism, stronger population structure, and plenty of recently active TEs—all of which complicate graph-building, and often lead to misleading or uninterpretable results. For these reasons, as well as general ease of interpretability, we based most of our analysis on the output from Pannagram.

SVs make a substantial contribution to sequence polymorphism in *A. thaliana*. In addition to massive variation in tandem repeat regions (Fig. 1), the readily alignable chromosome arms (Fig. 2) are highly polymorphic, with two accessions differing at over ∼190k SVs covering over ∼11 Mb—in addition to perhaps 700k SNPs (the precise number depends heavily on how aggressively SNPs are called). The allele-frequency distribution of SVs suggests purifying selection, and also a mutational process that involves insertion of longer segments coupled with deletion of shorter segments. This is consistent with TE activity, and indeed we show that the overwhelming majority of presence-absence variants over 100 bp involves annotated TE sequences. As expected from an active mobile-ome[36,56], we found thousands of examples of what appear to be recent insertions of presumably complete (∼5 kb) TEs. Importantly, although many of these correspond perfectly to annotated TEs, many do not, demonstrating that our understanding of the mobile-ome remains highly incomplete—as is also becoming evident from long-read transcriptome analysis[50]. To explore this further, we developed a graph-based analysis that identified several putative new TE families, providing a glimpse into just how much is left to discover about the mobile-ome.

Turning to the gene-ome, we note that the term "pan-genome" was originally invented to describe the rather amorphous genomes of prokaryotes[53], which appear to be characterized by a relatively small number of essential (or "core") genes, and a large cloud of presumably dispensable (or "accessory") genes, which are shared even between distantly related taxa thanks to rampant horizontal gene transfer[57]. It has been argued that this paradigm should apply to eukaryotes as well, for example in the context of rapidly evolving plant-pathogen interactions[58], but the overall picture in *A. thaliana* is clearly very different: the gene-ome is highly conserved, with 87% of genes detected in all 27 genomes, and the number of segregating genes growing considerably more slowly with sample size than other types of variation (Fig. 7). Furthermore, we distinguish between two types of segregating genes: a minority that has homologs in the closely related *A. lyrata*, and a majority that does not (Fig. 6). The former, which correspond either to gene duplications or segregating deletions of ancestral genes, tend to be expressed at significantly higher levels than the latter, which often were characterized by TE-like epigenetic silencing. The extent to which these are really part of the gene-ome will require further investigation, ideally including functional evidence.

Finally, we demonstrate that using well-established algorithms for SNP-calling, transcriptome profiling, and methylation profiling using short-read data against a single reference genome can be problematic in species with active TEs and

many gene duplications. These methods may work well in humans, but in other species, cryptic multi-mapping may be a substantial problem. Even in the relatively simple *A. thaliana* genomes, cryptic duplications result in a substantial fraction of heterozygous SNP calls[49]. All these issues become even more relevant in an outcrosser, where spurious heterozygosity is not obvious. While a complete abandonment of single reference genomes is still impractical, we strongly suggest to use a sample of diverse high-quality genomes for short read mapping to obtain an estimate of the magnitude of the problem.

In conclusion, we recall Aravinda Chakravarti's prediction at the dawn of the SNP era that models would be needed to "make sense out of sequence"—and that this would lead to "a rejuvenation of population genetics"[59]. We think that the advent of complete whole-genome polymorphism data will have a similar effect. Most obviously, whole-genome polymorphism data will revolutionize our understanding of TE dynamics by finally allowing us to observe the actual mobileome, *i.e.*, which sequences have recently been moving or are still moving, and to distinguish them from the accumulated layers of dead and decaying TE sequences that litter most genomes. Through Pannagram we provide a software package for doing this (for a related approach, see Sierra & Durbin[60]).

More subtly, in order to make sense of complex polymorphisms, we need to understand the history of mutations that gave rise to them. The problem is analogous to phylogenetic analysis, where the estimated relationship between species is used to make sense of complex traits—what evolved first, and what evolved multiple times? In the present context, we need to estimate local coalescent trees (the so-called Ancestral Recombination Graph[61]) and use them to infer the sequence of mutational and recombination events that gave rise to the sampled sequences. This analysis must be informed by a better understanding of the molecular mechanisms that cause structural variation. Although a considerable literature on phylogeny-guided statistical alignment exists[62-66], methods for doing this on a whole-genome scale, using an appropriate population genetic framework[54] and incorporating knowledge about molecular mechanisms are missing.

In this context, it is also important to remember that alignment-based methods developed for humans often do not work well in other species for a variety of reasons, including many species having much higher levels of polymorphism and TE activity than humans[67]. *Arabidopsis thaliana* genomes are not unusual in these respects—on the contrary, other plants are considerably worse. In species like maize, where intergenic space is essentially unalignable even between relatively closely related agricultural varieties[68], the idea of representing a whole-genome alignment as a graph that captures all variation may be neither practicable nor useful.

REFERENCES

1. Gallone, B. *et al.* Domestication and divergence of saccharomyces cerevisiae beer yeasts. *Cell* **166**, 1397–1410.e16 (2016).
2. Istace, B. *et al.* de novo assembly and population genomic survey of natural yeast isolates with the oxford nanopore MinION sequencer. *Gigascience* **6**, 1–13 (2017).
3. Peter, J. *et al.* Genome evolution across 1,011 saccharomyces cerevisiae isolates. *Nature* (2018).
4. Walkowiak, S. *et al.* Multiple wheat genomes reveal global variation in modern breeding. *Nature* 1–7 (2020).
5. Alonge, M. *et al.* Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145–161.e23 (2020).
6. Jayakodi, M. *et al.* The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature* **588**, 284–289 (2020).
7. Hufford, M. B. *et al.* De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, 655–662 (2021).
8. Liu, Y. *et al.* Pan-Genome of wild and cultivated soybeans. *Cell* **182**, 162–176.e13 (2020).
9. Qin, P. *et al.* Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542–3558.e16 (2021).
10. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372** (2021).
11. Zhou, Y. *et al.* Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* **606**, 527–534 (2022).
12. Li, R. *et al.* A sheep pangenome reveals the spectrum of structural variations and their effects on tail phenotypes. *Genome Res.* **33**, 463–477 (2023).
13. Ferguson, S. *et al.* Plant genome evolution in the genus eucalyptus driven by structural rearrangements that promote sequence divergence. *bioRxiv* 2023.04.19.537464 (2023).
14. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
15. Wlodzimierz, P. *et al.* Cycles of satellite and transposon evolution in arabidopsis centromeres. *Nature* (2023).
16. He, Q. *et al.* A graph-based genome and pan-genome variation of the model plant setaria. *Nat. Genet.* (2023).
17. Shi, T. *et al.* The super-pangenome of populus unveil genomic facets for adaptation and diversification in widespread forest trees. *bioRxiv* 2023.07.18.549473 (2023).
18. Kang, M. *et al.* The pan-genome and local adaptation of *Arabidopsis thaliana*. *Nat. Commun.* **14**, 6259 (2023).

19. Cochetel, N. *et al.* A super-pangenome of the north american wild grape species. *Genome Biol.* **24**, 290 (2023).

20. Beaulieu, C. *et al.* The *Marchantia* pangenome reveals ancient mechanisms of plant adaptation to the environment. *bioRxiv* 2023.10.27.564390 (2023).

21. Li, X. *et al.* Large-scale gene expression alterations introduced by structural variation drive morphotype diversification in brassica oleracea. *Nat. Genet.* (2024).

22. Lian, Q. *et al.* A pan-genome of 69 Arabidopsis thaliana accessions reveals a conserved genome structure throughout the global species range. *Nat. Genet.* 1–10 (2024).

23. Eizenga, J. M. *et al.* Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020).

24. Garrison, E. *et al.* Building pangenome graphs. *bioRxiv* (2023).

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

26. Clark, R. M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317**, 338–342 (2007).

27. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627–631 (2010).

28. 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).

29. 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**, 481–491 (2016).

30. Zou, Y.-P. *et al.* Adaptation of *Arabidopsis thaliana* to the Yangtze River basin. *Genome Biol.* **18**, 239 (2017).

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

32. Fulgione, A., Koornneef, M., Roux, F., Hermisson, J. & Hancock, A. M. Madeiran *Arabidopsis thaliana* reveals ancient long-range colonization and clarifies demography in Eurasia. *Mol. Biol. Evol.* (2017).

33. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).

34. Alonge, M. *et al.* Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).

35. Lamesch, P. *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–10 (2012).

36. Quadrana, L. *et al.* The *Arabidopsis thaliana* mobilome and its impact at the species level. *Elife* **5** (2016).

37. 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).

38. Fultz, D., McKinlay, A., Enganti, R. & Pikaard, C. S. Sequence and epigenetic landscapes of active and silent nucleolus organizer regions in Arabidopsis. *Sci Adv* **9**, eadj4509 (2023).

39. Schmuths, H., Meister, A., Horres, R. & Bachmann, K. Genome size variation among accessions of *Arabidopsis thaliana*. *Ann. Bot.* **93**, 317–321 (2004).

40. Long, Q. *et al.* Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from sweden. *Nat. Genet.* **45**, 884–890 (2013).

41. Kidwell, M. G. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**, 49–63 (2002).

42. Zapata, L. *et al.* Chromosome-level assembly of *Arabidopsis thaliana* ler reveals the extent of translocation and inversion polymorphisms. *Proc. Natl. Acad. Sci. U. S. A.* (2016).

43. Fransz, P. *et al.* Molecular, genetic and evolutionary analysis of a paracentric inversion in *Arabidopsis thaliana*. *Plant J.* **88**, 159–178 (2016).

44. 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).

45. Cork, J. M. & Purugganan, M. D. High-diversity genes in the Arabidopsis genome. *Genetics* **170**, 1897–1911 (2005).

46. Stupar, R. M. *et al.* Complex mtDNA constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2: implication of potential sequencing errors caused by large-unit repeats. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 5099–5103 (2001).

47. Fields, P. D. *et al.* Complete sequence of a 641-kb insertion of mitochondrial DNA in the *Arabidopsis thaliana* nuclear genome. *Genome Biol. Evol.* **14**, evac059 (2022).

48. Puchta, H. The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. *J. Exp. Bot.* **56**, 1–14 (2005).

49. Jaegle, B. *et al.* Extensive sequence duplication in arabidopsis revealed by pseudo-heterozygosity. *Genome Biol.* **24**, 1–19 (2023).

50. Borredá, C., Leduque, B., Colot, V. & Quadrana, L. Transposable element products, functions, and regulatory networks in arabidopsis. *bioRxiv* 2024.04.02.587720 (2024).

51. Kornienko, A. E., Nizhynska, V., Molla Morales, A., Pisupati, R. & Nordborg, M. Population-level annotation of lncRNAs in Arabidopsis reveals extensive expression variation associated with transposable element-like silencing. *Plant Cell* **36**, 85–111 (2023).

52. Cheng, C.-Y. *et al.* Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* **89**, 789–804 (2017).

53. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U. S. A.* **102**, 13950–13955 (2005).

54. Nordborg, M. Coalescent theory. In Balding, D., Moltke, I. & Marioni, J. (eds.) *Handbook of Statistical Genomics*, 145–175 (Wiley, 2019), 1 edn.

55. Schreiber, M., Jayakodi, M., Stein, N. & Mascher, M. Plant pangenomes for crop improvement, biodiversity and evolution. *Nat. Rev. Genet.* (2024).

56. Stuart, T. *et al.* Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *Elife* **5** (2016).

57. McInerney, J. O., McNally, A. & O'Connell, M. J. Why prokaryotes have pangenomes. *Nat Microbiol* **2**, 17040 (2017).

58. Dong, S., Raffaele, S. & Kamoun, S. The two-speed genomes of filamentous pathogens: waltz with plants. *Curr. Opin. Genet. Dev.* **35**, 57–65 (2015).

59. Chakravarti, A. Population genetics–making sense out of sequence. *Nat. Genet.* **21**, 56–60 (1999).

60. Sierra, P. & Durbin, R. Identification of transposable element families from pangenome polymorphisms. *bioRxiv* 2024.04.05.588311 (2024).

61. Lewanski, A. L., Grundler, M. C. & Bradburd, G. S. The era of the ARG: An introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics. *PLoS Genet.* **20**, e1011110 (2024).

62. Novák, A., Miklós, I., Lyngsø, R. & Hein, J. StatAlign: an extendable software package for joint bayesian estimation of alignments and evolutionary trees. *Bioinformatics* **24**, 2403–2404 (2008).

63. Chatzou, M. *et al.* Multiple sequence alignment modeling: methods and applications. *Brief. Bioinform.* **17**, 1009–1023 (2016).

64. Löytynoja, A. Phylogeny-Aware alignment with PRANK and PAGAN. In Katoh, K. (ed.) *Multiple Sequence Alignment: Methods and Protocols*, 17–37 (Springer US, New York, NY, 2021).

65. Redelings, B. D. BAli-Phy version 3: model-based co-estimation of alignment and phylogeny. *Bioinformatics* **37**, 3032–3034 (2021).

66. Kille, B., Balaji, A., Sedlazeck, F. J., Nute, M. & Treangen, T. J. Multiple genome alignment in the telomere-to-telomere assembly era. *Genome Biol.* **23**, 182 (2022).

67. Song, B., Buckler, E. S. & Stitzer, M. C. New whole-genome alignment tools are needed for tapping into plant diversity. *Trends Plant Sci.* (2023).

68. Morgante, M. *et al.* Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat. Genet.* **37**, 997–1002 (2005).

69. Ou, S. *et al.* Effect of sequence depth and length in long-read assembly of the maize inbred NC358. *Nat. Commun.* **11**, 2288 (2020).

70. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).

71. Maheshwari, S., Ishii, T., Brown, C. T., Houben, A. & Comai, L. Centromere location in Arabidopsis is unaltered by extreme divergence in CENH3 protein sequence. *Genome Res.* **27**, 471–478 (2017).

72. Simon, L. *et al.* Genetic and epigenetic variation in 5S ribosomal RNA genes reveals genome dynamics in *Arabidopsis thaliana*. *Nucleic Acids Res.* **46**, 3019–3033 (2018).

73. Rabanal, F. A. *et al.* Unstable inheritance of 45S rRNA genes in *Arabidopsis thaliana*. *G3* **7**, 1201–1209 (2017).

74. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

75. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods Mol. Biol.* **1962**, 1–14 (2019).

76. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).

77. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).

78. Ou, S. & Jiang, N. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).

79. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–8 (2007).

80. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob. DNA* **10**, 48 (2019).

81. Su, W., Gu, X. & Peterson, T. TIR-Learner, a new ensemble method for TIR transposable element annotation, provides evidence for abundant new transposable elements in the maize genome. *Mol. Plant* **12**, 447–460 (2019).

82. Shi, J. & Liang, C. Generic repeat finder: A High-Sensitivity tool for Genome-Wide de novo repeat detection. *Plant Physiol.* **180**, 1803–1815 (2019).

83. Xiong, W., He, L., Lai, J., Dooner, H. K. & Du, C. HelitronScanner uncovers a large overlooked cache of helitron transposons in many plant genomes. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 10263–10268 (2014).

84. Zhang, R.-G. *et al.* TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic Res* (2022).

85. Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

86. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).

87. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **1962**, 227–245 (2019).

88. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).

89. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).

90. Campbell, M. S. *et al.* MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164**, 513–524 (2014).

91. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).

92. Pertea, G. & Pertea, M. GFF utilities: GffRead and GffCompare. *F1000Res.* **9** (2020).

93. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).

94. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

95. Liao, Y., Smyth, G. K. & Shi, W. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

96. Kawakatsu, T. *et al.* Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. *Cell* **166**, 492–505 (2016).

97. Ramírez, F. *et al.* deeptools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–5 (2016).

98. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

99. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

100. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* (2013). 1303.3997.

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

102. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).

103. Sun, H., Ding, J., Piednoël, M. & Schneeberger, K. findGSE: estimating genome size variation within human and arabidopsis using k-mer frequencies. *Bioinformatics* (2017).

104. 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).

105. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).

106. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).

107. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 (2018).

108. Song, B. *et al.* AnchorWave: Sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. *Proc. Natl. Acad. Sci. U. S. A.* **119** (2022).

Methods

DNA extraction and sequencing

For long-read sequencing, we began with 3-week-old plants grown in soil that had been transferred to darkness for 24-48 h before harvesting to reduce the starch content. 20-30 g of flash-frozen rosette tissue from pooled individuals was ground in liquid nitrogen with pestle and mortar. Nuclei were isolated as described for accession Ey15-2[37], and high-molecular-weight (HMW) DNA purified with Genomic-tips 100G (Qiagen; #10243) following manufacturer's instructions. 10 μg of HMW DNA were sheared with either Megaruptor 3 (Diagenode; #B06010003) or a needle (FINEJECT® 26Gx1" 0.45x25mm; #14-13651) to ca. 75 kb, and used as input for long-read library preparation with the SMRTbell Express Template Preparation Kit 2.0 (Pacific Biosciences; #101-693-800). These libraries were size-selected with the BluePippin system (SageScience) with a 30 kb cut-off in a 0.75% DF Marker U1 high-pass 30–40 kb vs3 gel cassette (Biozym; #BLF7510). Libraries for accessions 9981 (Angit-1; CS76366) and 10002 (TueWal-2; CS76405) were sequenced on a Sequel II system (Pacific Biosciences), and the others on a Sequel I system.

To prepare PCR-free libraries for short-read sequencing, the genomic DNA was fragmented to 250-350 bp using a Covaris S2 Focused Ultrasonicator (Covaris). The libraries were prepared according to manufacturer's instructions with either the TruSeq DNA PCR-Free (Illumina, #20015962) or the NxSeq® AmpFREE Low DNA Library Kit (Lucigen; # 14000-2). In total, libraries for 89 accessions (including the main 27 for which we assembled their genomes) were sequenced in paired-end mode on an HiSeq 3000 system (Illumina).

The ultra-HMW DNA extraction and sample preparation for optical maps was done as described[37,69] at Corteva Agriscience (Johnston, Indiana, USA) using the Direct Labeling and Stain (DLS) technology (Bionano Genomics).

Assembly

The CLR subreads were assembled with Canu v1.71[33]. Since accessions 9981 and 10002 had been sequenced at higher coverage on a Sequel II instrument, only about 200X genome coverage worth of reads were used for assembly. We performed two rounds of polishing on the resulting contigs of all assemblies: first with the CLR subreads and Arrow v2.3.2 (https://github.com/PacificBiosciences/GenomicConsensus), and then with PCR-free short reads and Pilon v1.22[70].

For scaffolding, we generated hybrid scaffolds with optical maps for eight accessions (Supplementary Table 1) using Bionano Access v1.5 and Bionano Solve v3.6. The assembly was performed in pre-assembly mode using parameters `non-haplotype` and `no-CMPR-cut`, without `extend-split`. Based on what we learned from these hybrid assemblies, we set the parameters for *in silico* scaffolding of the other genomes. We scaffolded contigs > 150 kb with RagTag v1.1.1[34] (`scaffold -q 60 -f 10000 -I 0.6 -remove-small`) using the TAIR10 reference with hard-masked centromeres, rDNAs, telomeres, and nuclear insertions of organelles to prevent misplacement of contigs due to reference bias[37]. All scaffolded assemblies were manually curated to specifically discard low-confidence centromere satellite-rich contigs or to invert contigs with satellite repeats at their edges indicative of their correct orientation. These edits were implemented in the agp files, which were converted to fasta format with the RagTag `agp2fa` function[34]. Agp files before and after manual curation are available in the GitHub repository of this project.

Repeat annotation

Repetitive elements were annotated as described[37]. We ran RepeatMasker v4.0.9 (`-cutoff 200 -nolow -gff -xsmall`) using a custom library that included various consensus sequences for the CEN178[71], 5S rDNA[72], 45S rDNA[73], and telomere repeats. We identified insertions of organellar DNA in the nuclear genomes by aligning the organellar genomes from the TAIR10 reference[35] with minimap2 v2.16[74] (`-cx asm5`). We annotated tRNAs with tRNAscan-SE v2.0.6[75], and TEs with Extensive de-novo TE Annotator (EDTA) v1.9.7[76] (`--step all --sensitive 1 --anno 1`), a pipeline that combines several TE annotation tools (LTRharvest, LTR_FINDER, LTR_retriever, TIR-Learner, HelitronScanner, TEsorter)[77-84]. Finally, to understand the causes of contig breaks, we determined what type of repetitive element was closest to each contig edge, considering the first 2 kb from each edge in contigs >10 kb.

Pannagram

Pannagram is a toolkit designed for reference-free **pan**-genome alignment, **ann**otation, and **ana**lysis, as well as for generating dia**grams**.

We represent the whole-genome alignment as a matrix of corresponding positions, where rows represent accessions, and columns represent homologous positions. The construction of the alignment is done in a reference-free manner (see below). However, to visualize the alignment in genome browsers, columns must be sorted in some manner, *e.g.*, to correspond to the TAIR10 sequence order. Then, columns of the pan-genome are used as positions in the pan-genome coordinate system.

To perform reference-free whole-genome alignment, we developed a 3-step pipeline. First, we use several accessions as reference and build draft pairwise alignments between each and all other accessions. This process results in several reference-based matrices of corresponding positions. Next, we intersect these matrices, selecting only those columns that are present in all reference-biased matrices, which produces

reliable and reference-independent correspondences. In the final step, we resolve unaligned sequences between blocks of corresponding positions using multiple sequence alignment tools. Once the reference-free alignment is complete, it can be sorted according to the desired order of accessions. In our case, we employ an alphabetical order, with the TAIR10 genome first.

For the pairwise alignments between a reference genome (not necessarily TAIR10) and another accession, the focal accession genome is divided into blocks of 5000 bp, and each block is then mapped to the corresponding chromosomes of the reference genome using BLAST, with exactly one best hit retained for each block through this process. Next, the BLAST hits that are not in close proximity to each other in both genomes are removed. An additional BLAST search is performed to align corresponding unaligned sequences between remaining hits.

To resolve any unaligned blocks after the reference-randomization procedure, the MAFFT[85] method is used. Blocks longer than 30 kb cannot be aligned within a reasonable time using MAFFT, so they are considered to be highly diverged. We found the final unaligned regions to be primarily associated with centromeric regions, rDNA clusters, telomeres and complex regions of multiple and long insertions and deletions, which are regions that are not of primary interest in this paper.

Given the whole-genome alignment, SNPs can simply be output as sequence differences. However, sequence differences can arise from ambiguities in local alignment and do not necessarily correspond to SNPs (see also Supplementary Note 5). If we take all sequence differences as SNPs, a pair of accessions differs at over 800k positions, on average, but if we restrict ourselves to isolated sequence differences, the number shrinks to 600k.

Pan-genome graph

Graph construction

We constructed genome graphs for each of the five chromosomes using the pan-genome Graph Builder (PGGB) pipeline. First, we prepared the assemblies by splitting them into chromosomes and removing all unplaced contigs. To enforce linearity for simpler analysis and comparison, we used a modified version of accession 22001 with the genome rearranged to a consensus pan-genomic order (suffix: "f"). We added the TAIR10 reference genome to the graph to enable anchoring and presentation of results in a reference frame work.

We executed the PGGB pipeline (downloaded 25 January 2024) with the parameters `-s 10000 -p 90 -n 27`. PGGB consists of three methods: An all-against-all alignment with wfmash (v0.12.4-5-g0b191bb), graph induction using seqwish (v0.7.9-2-gf44b402), and two rounds of pangenome ordering (odgi v0.8.3-26-gbc7742ed) followed by normalization with smoothxg (v0.7.2-11-g9970e0d). The graph was used for the analysis of the pan genome, synteny and variation detection using vg deconstruct.

Similarity

We exploited graph properties to classify different levels of similarity between genomes. Nodes traversed in all accessions are labeled as core, nodes traversed in only one accession are private, and all other nodes (`> 1 and < 28 traversals`) are shell (soft). Note that nodes can be traversed multiple times by the same genome, which affects the total number of core nodes. Since each node contains a specific sequence, node count can easily be converted to the actual amount of sequence and respective genomic location.

Synteny windows

Every node in the graph can be translated to its exact position for each path. This direct connection allows us to create sliding window approaches for each sample/path using graph-based statistics. Here, we used non-overlapping windows of 300 kb and calculated the average similarity (see above) of these regions. This was performed for each graph and path independently and the results represented in a heat map.

Saturation analysis

A saturation analysis was performed using a bootstrapping approach. In each iteration we removed a specific number of paths from our graph and performed the same pan-genome categorization as above (s. 6.4.2 Similarity). In addition, we added the total pan-genome, which describes the total amount of sequence (core+shell+private sequence). We performed 20 different (unique) combinations for each size (number of genomes).

Deconstructing the graph

To achieve full insights into graph variation and cover all bubbles in the graph, `VG deconstruct` was run multiple times with each accession reference path once (`vg deconstruct -a -e`. After, the reported VCF (v1.54.0 "Parafada") files were converted to a BED file with all important information provided. In addition, each chromosome was merged and genotype information was concluded and added. Bubbles were identified by start and end position, also reporting all traversals within these bubbles. Scripts can be found the in the repository.

The mobile-ome

The mobile-ome refers to the collection of insertions and deletions that are likely to have occurred recently and are therefore not fixed in our sample. We hypothesize that each mobile event results in a structural variant (SV), specifically a

presence-absence polymorphism at the location of the insertion or deletion. Consequently, our initial approach involves extracting all presence-absence SVs and systematically decomposing them step by step. To distinguish between simple bi-allelic presence-absence polymorphisms (indels) and complex multi-allelic polymorphisms, we analyzed the lengths of alleles within the SVs. We distinguish two types based on the similarity threshold s, with $s = 0.9$ in our case. We consider a simple indel as one that contains alleles of two length types: those that are shorter than $(1 - s)$ of the SV length (absence allele) and those that are longer than s of the SV length (presence allele). The distinction between simple and complex presence-absence polymorphisms is partially a computational construct to filter SVs and simplify further analysis. Simple indels and complex presence-absence polymorphisms form a continuum, and by relaxing the similarity threshold ($s < 0.9$, in our case), some complex SVs become classified as simple indels. Additionally, there is an natural bias towards complex presence-absence polymorphisms. Consider a scenario with a simple presence-absence polymorphism where an indel occurs within the presence allele. If the presence allele was initially observed in only one accession, then the new event does not reclassify the initial region as not belonging to the simple presence-absence polymorphisms category. However, if the presence allele was observed in multiple accessions, the new event is likely to be reclassified as complex presence-absence polymorphisms. To simplify and clarify the analysis, we considered only the simple polymorphisms. In order to determine the known portion of the mobile-ome within indels, we conducted BLAST searches using pan-genome consensus sequences of indels against known *A. thaliana* TEs, as well as against themselves. The indels that exhibited some similarity to known TEs were divided into the following groups:

is complete Significant similarity to known TEs and can be classified as TEs themselves.

contains complete Contained regions with similarity to known TEs but also additional sequences.

is fragment Contained only partially sequences with similarity to known TEs.

contains fragment Partial coverage by BLAST hits of TE segments but also additional sequences unrelated to known TEs.

We consider all these indels as parts of the mobile-ome. Indels without similarity to known TEs but showing nested similarities within the indel data set (where one sequence is a subsequence of another) were considered as potential candidates for new mobile-ome elements. In order to investigate their potential function, we obtained all six open reading frames (ORFs) within each of these indels. From each translated sequence, we selected either all continuous stretches without stop codons that were longer than 100 codons, or the longest stretch that exceeded 30 codons without a stop. Subsequently, we performed a BLAST search using the obtained amino acid sequences against the NCBI protein database and classified the potential proteins for each indel based on the ordered keywords found in the BLAST hits: “transpos”, “reverse”, “transcriptase”, “repeat”, “zinc”, “receptor”, and “disease”. Indels that only had blast hits with descriptions such as “hypothetical protein”, “unnamed protein product”, “uncharacterized protein”, “predicted protein”, “PREDICTED:”, “putative protein”, and “unknown” were categorized as “undefined proteins”. Indels without any BLAST hits were classified as “no protein”.

Gene annotation

Preliminary annotation

Gene annotation was mainly based on Augustus (v3.3.3)[86]. Augustus-predicted gene models were trained using parameters obtained from “hints” from three different sources. First, we ran BUSCO (v4.0.1)[87] with `-m` genome option. Second, the *A. thaliana* reference gene annotation was projected onto each genome using Liftoff[88] with the `-exclude_partial` and `-copies` options. Third, the RNA-seq data for each accession were used: wiggle hints were generated using `bam2wig` and `wig2hints` and EST hints were generated using `bam2hints` (all three tools provided by Augustus). Augustus was run with the following non-default parameters:

```
--softmasking 1
--species=BUSCO_retraining
--gff3=on
--extrinsicCfgFile=Custom_Config
--hintsfile=Liftoff_hints
```

For every accession, the GFF3 output of Augustus was run through the Augustus-provided tool `getAnno.pl` to translate gene annotations into protein sequences. Finally, for each annotation the Augustus output was combined and evaluated using `augustus_GFF3_to_EVM_GFF3.pl` (provided by `EVidenceModeler`[89]).

Apart from the Augustus-generated annotations, we used two types of additional independent evidence for gene models: the SNAP *de novo* annotation tool[90] and Cufflinks transcriptome assembly[91]. Annotations produced by Augustus, SNAP, and Cufflinks were combined and then subdivided into 1 Mb windows with 1 kb overlap using `partition_EVM_input.pl` (provided by `EVidenceModeler`). We ran `EVidenceModeler` with annotation gff files, the assembly fasta file, the partitions and a weight matrix. We chose weights for each input based on their ability to recreate the Araport11 gene annotation. Running `EVidenceModeler` produced the final annotation compilation for each accession. We retained only one longest isoform for each gene using `gffread`[92].

Reconciling annotations

To enable comparison between the independent annotation, we utilized the pan-genome coordinate system, reconciling discrepancies using majority voting. Our approach generated a total of 33,223 putative gene: 4,133 of these were the result of splitting preliminary annotations; 1,306 were the result of merges. For details, see Supplementary Note 4.1.

Ancestry analysis

All protein-coding sequences from all accessions were compared using DIAMOND's `blastp` module[93] (version 2.0.11) against the *A. lyrata* MN47 proteome (version 2, GenBank: GCA_944990045.1) and the best hit was further considered as the *A. lyrata* homolog. Next, each annotation group from *A. thaliana* was assigned to the *A. lyrata* homolog that was common to at least 50% of its members. *A. thaliana* annotation groups were defined to be ancestral relative to the *A. lyrata* gene, if they were part of a co-linear segment (of at least two) genes. To that end, all *A. thaliana* genes were ordered according to their relative position in the pan-genome coordinate system. Each pair of consecutive genes in *A. thaliana* were assigned to the same co-linear segment as their homologs in *A. lyrata*, if the homologs were separated by fewer than six genes. The ancestral state was defined as 'similar' for cases where the genes from *A. lyrata* and *A. thaliana* were not part of the same co-linear segment but shared at least 80% sequence identity over at least 80% of the length *A. thaliana* gene. Please see further details in Supplementary Note 4.2 for TE analysis and Supplementary Note 4.3 for origin of new genes.

Expression analysis

RNA-seq read mapping and gene expression calculation

Raw RNA-seq reads from 7-day-old seedlings, 9-leaf rosettes, flowers, and pollen[51] were aligned either to the TAIR10 reference genome or the corresponding accession genome using STAR v2.7.1[94] with custom options:

```
--alignIntronMax 6000
--alignMatesGapMax 6000
--outFilterIntronMotifs RemoveNoncanonical
--outFilterMismatchNoverReadLmax 0.1
--outFilterMismatchNmax 999
--outFilterMismatchNoverLmax 0.3
--outFilterMultimapNmax 1
--alignSJoverhangMin 8
--outSAMattributes NH HI AS nM NM MD jM jI XS
```

(for read statistics, see Supplementary Table 4). Expression levels were assessed using `featurecounts` from Subread v2.0.1[95] on each RNA-seq sample with either the TAIR10 gene annotation or the accession-specific annotations produced in this study. The expression levels were normalized by calculating TPMs, the number of read counts was divided by the gene length in kb and the total number of counts covering the annotation in million reads.

Expression of duplicated genes

The expression of gene copies was calculated only on the whole-locus level (disregarding exon-intron structure, as many copies were not part of our annotation and we did not have exon information) using the same method, but normalizing expression to the total number of counts covering the full annotation for a given sample and not the total number of reads covering the gene copies.

Mapping to TAIR10 vs. own genome

To determine whether the gene expression calculation was consistent between RNA-seq mapping in TAIR10 versus accession-specific genomes, we focused on the annotation groups with a one-to-one correspondence with an Araport11 gene. For each RNA-seq sample, we obtained the Pearson's correlation coefficient between the exonic counts obtained from TAIR10 mapping and accession-genome mapping. We also estimated the number of genes that were correctly or wrongly estimated using TAIR10 mapping. We called a gene "wrong" if the number of counts in TAIR10 and in the own genome differed by more than 30% (`Ncounts_min/Ncounts_max` $\leq$ `0.7`). Only genes with at least 6 counts in either calculation were analyzed.

ChIP-seq analyses

We used ChIP-seq data from 6 accessions and sRNA-seq data from 14 accessions[51] as well as BS-seq data from 12 accessions[96].

We used STAR[94] to map ChIP-seq reads with these non-default options:

```
--alignIntronMax 5
--outFilterMismatchNmax 10
--outFilterMultimapNmax 1
--alignEndsType EndToEnd
--twopassMode Basic
```

The ChIP-seq data were log2-normalized to input using bamCompare (deeptools package[97]) using

```
--operation log2
--effectiveGenomeSize 119481543
--ignoreDuplicates
--outFileFormat bedgraph
```

The ChIP-seq coverage was estimated using bedtools map -mean[98]. The ChIP-seq coverage was further normalized to obtain value range similarity across accessions. For this, we applied quantile-normalization using an R function:

```
function(x) {
 (x-quantile(x,.20)) / (quantile(x,.80) - quantile(x,.20))
}
```

which equalized the 20% and 80% quantile values of each ChIP-seq sample. After quantile-normalization, the replicated samples were averaged.

To process the sRNA-seq data we trimmed the reads using cutadapt[99]: `cutadapt -a AACTGTAGGCACCATCAAT --minimum-length 18`. We then used STAR[94] with the following non-default options to map sRNA-seq reads to the corresponding genome:

```
--runRNGseed 12345
--alignEndsType Extend5pOfRead1
--alignIntronMax 5000 --alignSJDBoverhangMin 1
--outReadsUnmapped Fastx --outSAMmultNmax 100
--outSAMprimaryFlag AllBestScore
--outSAMattributes NH HI AS nM NM MD jM jI XS
--outFilterMultimapNmax 10
--outFilterMatchNmin 16
--outFilterMatchNminOverLread 0.66
--outFilterMismatchNmax 2
--outFilterMismatchNoverReadLmax 0.05
--outFilterIntronMotifs RemoveNoncanonicalUnannotated
--twopassMode None
```

We finally extracted 24-nt reads, calculated read coverage for each position of the genome using `genomeCoverageBed` (bedtools v.2.27.1), normalized it by the total number of uniquely-mapped reads in each sample and calculated 24-nt small RNA coverage for each locus of interest using `bedtools map -mean` function.

Reference bias in determining DNA methylation

We focused on the annotation groups in our accessions with a one-to-one correspondence with an Araport11 gene (the current annotation of the TAIR10 genome). For each accession and each gene, we calculated the CG, CHG and CHH methylation levels as the ratio of the total number of methylated and unmethylated reads in the specific context within the gene. We next obtained the Pearson's correlation coefficient between the methylation level estimates obtained from TAIR10 mapping and accession-genome mapping. We also estimated the number of genes that were correctly or wrongly estimated using TAIR10 mapping. For each methylation context, we called a gene "wrong" if the methylation level in TAIR10 and the own genome differed by more than 50% ($\texttt{methlevel_min}/\texttt{methlevel_max} \leq 0.5$). For further information, please see Supplementary Note 6.

Data availability

The genome assemblies and raw data (PacBio CLR and Illumina PCR-free short reads) for the 27 *A. thaliana* genomes analyzed in this study are deposited in the European Nucleotide Archive (ENA) under project accession number PRJEB73474 (ERP158243). Illumina PCR-free data for 61 additional accessions used to investigate the contribution of satellite repeats can be accessed under project accession number PRJEB73476 (ERP158245). In addition, assemblies, annotations, TEs, RNA-Seq, BS-seq, and ChIP-Seq data can be downloaded from 1001genomes.org, where a collection of accession-based JBrowse2 genome viewers and a pangenome JBrowse2 are also available.

Code availability

All scripts for this papers are located on GitHub in an *ad hoc* 1001 Genomes+ repository. The Pannagram toolkit is found in its own GitHub repository as are the Pangenome-graph methods. The genome assembly pipeline is likewise found in its own GitHub repository.

Acknowledgments

We thank Zhigui Bao, Alejandra Duque and Gal Ofir for discussion and comments on the manuscript. Elizaveta Grigoreva supplied the polymorphism data for Fig. 2C. This work was funded by the Deutsche Forschungsgemeinschaft (DFG), the Austrian Science Fund (FWF) and the Biotechnology and Biological Sciences Research Council (BBSRC) through ERA-CAPS Project 1001GenomesPlus (BBSRC BB/S004661/1; P.K., D.W., M.N.), the Austrian Academy of Sciences (M.N.), the Max Planck Society (D.W.), and the European Union's Framework Programme for Research and Innovation Horizon 2020 (2014-2020) under the Marie Curie Skłodowska Grant Agreement Nr. 847548 (H-J.L.).

Author contributions

P.K., D.W. and M.N. conceived and supervised the project. P.F., S.N, A.H. and Y.G. contributed materials. F.R., A.K., M.C., A.M., V.L., V.N., I.R., C.L, and F.B generated data. F.R., H.A., B.J. and T.W. assembled the genomes. A.I., H.A., M.C. and C.K. annotated the genomes. H.L. and A.K analyzed expression and epigenetics. A.I. and S.V. performed the SV analysis. A.I. analyzed the mobile-ome. A.I., H.A., A.K., and V.V. analyzed the gene-ome. H.L. and A.K. analyzed short-read mapping bias. J.F. developed the genome browser. A.I, S.V., F.R., H.L, H.A., A.K., D.W. and M.N. wrote the paper.

Competing interests

D.W. holds equity in Computomics, which advises plant breeders. D.W. also consults for KWS SE, a plant breeder and seed producer with activities throughout the world. J.F. is an employee of Tropic TI, Lda. All other authors declare no competing interests.

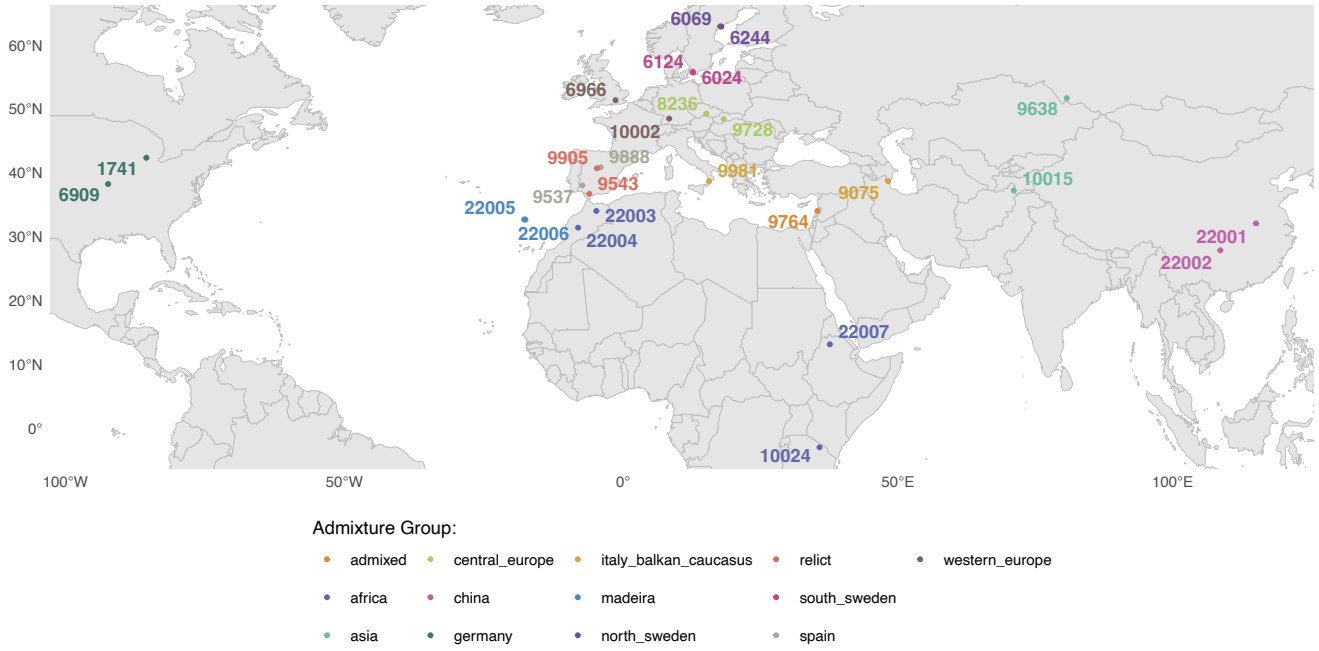
Extended Data Figures

1 Origin of the sequenced accessions. 7
2 The causes of contig breaks. 7
3 Scaffolded vs. unplaced contigs. 8
4 TE family size distribution. 8
5 Whole-genome alignments 9

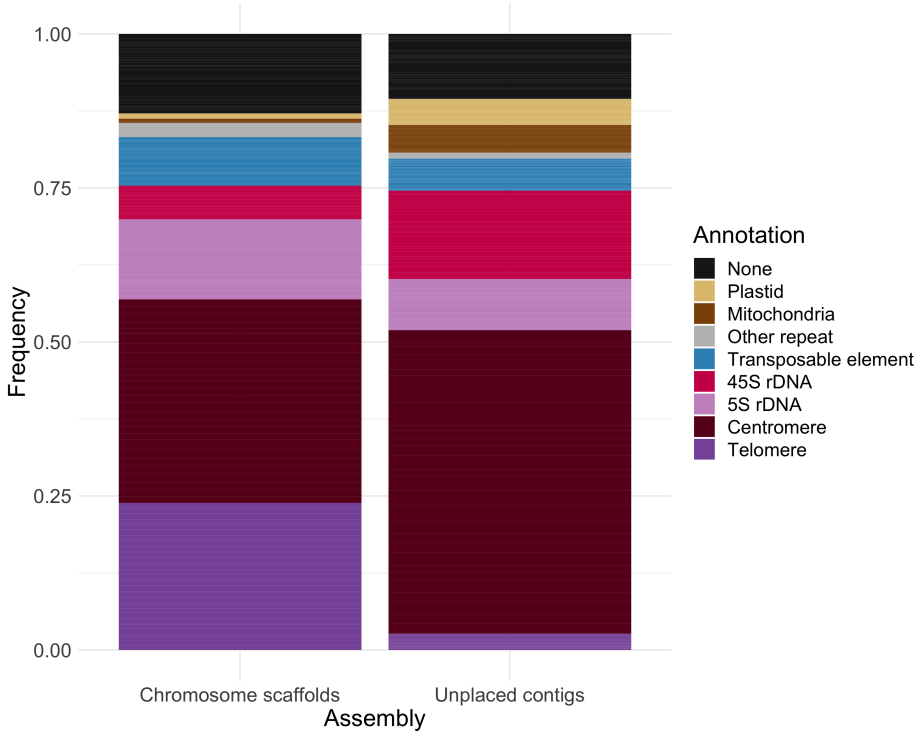
6 Simple and complex length variants 9
7 Comparing variation from Pannagram and PGGB 10
8 Comparing Pannagram and PGGB output on base-level 11
9 Location of indels with respect to genes. . . . 12
10 Frequency and size distribution of organellar inserts. 12
11 Size distribution of nodes in the graph base on TE-containing sSVs 13
12 The graph of the nestedness of existing *A. thaliana* TE annotations. 13
13 Very similar mobile elements have different names in *A. thaliana* TE annotation 14
14 A mobile element that contains annotated TEs. 15
15 Another mobile element that contains annotated TEs 16
16 Un-annotated TE with evidence of horizontal transfer 17
17 Putative novel mobile element family 18
18 Methylation of sSVs 19
19 Expression in sSVs 20
20 Details of gene-ome analysis 21
21 Chromosomal gene distribution 22
22 Gene silencing 23
23 Saturation curves for sSVs for different length ranges 24
24 Normalized saturation curves for different genome components 24
25 Pan-genome vs. reference genome coordinates for each chromosome 25
26 Reference bias: expression 26

Supplementary information

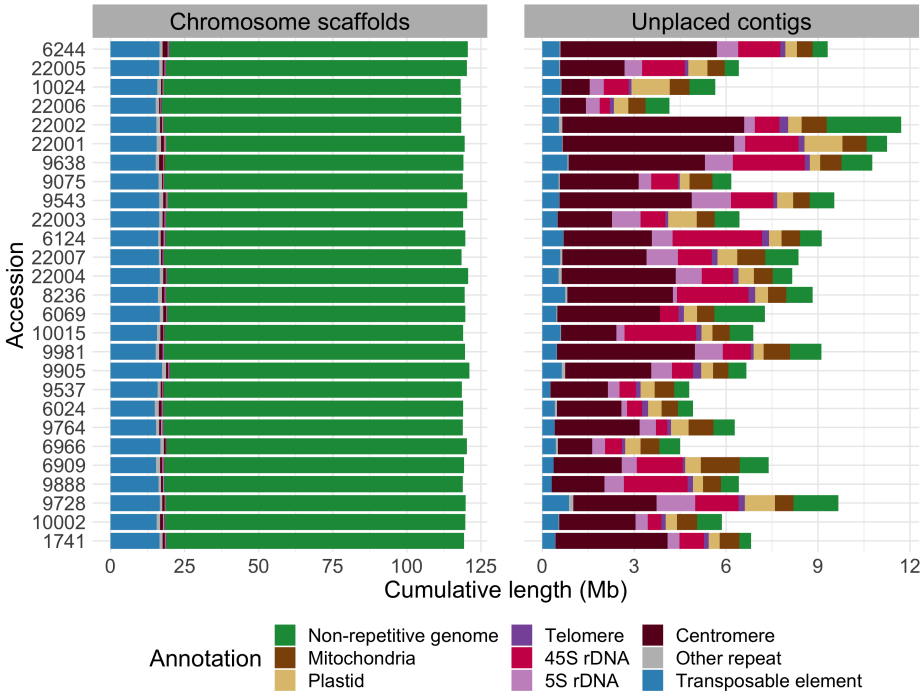
- Supplementary Note – Detailed information about specific analyses, with figures.
- Supplementary Table 1 – Origin and sequencing data for the 27 accessions.
- Supplementary Table 2 – Large inversions.
- Supplementary Table 3 – Organellar inserts.
- Supplementary Table 4 – RNA-Seq read-map data.



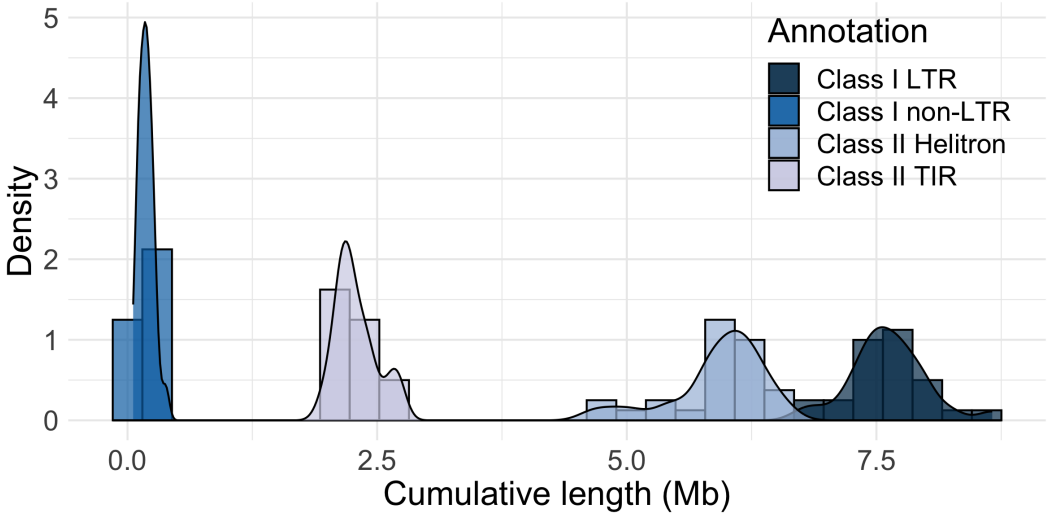
Extended Data Figure 1 – Origin of the sequenced accessions. Colors indicate "ADMIXTURE group"[29]. Note that 6909, which corresponds to the TAIR10 reference genome (Col-0) lacks reliable collection data (but hails from Germany based on genotype as well as historical records).



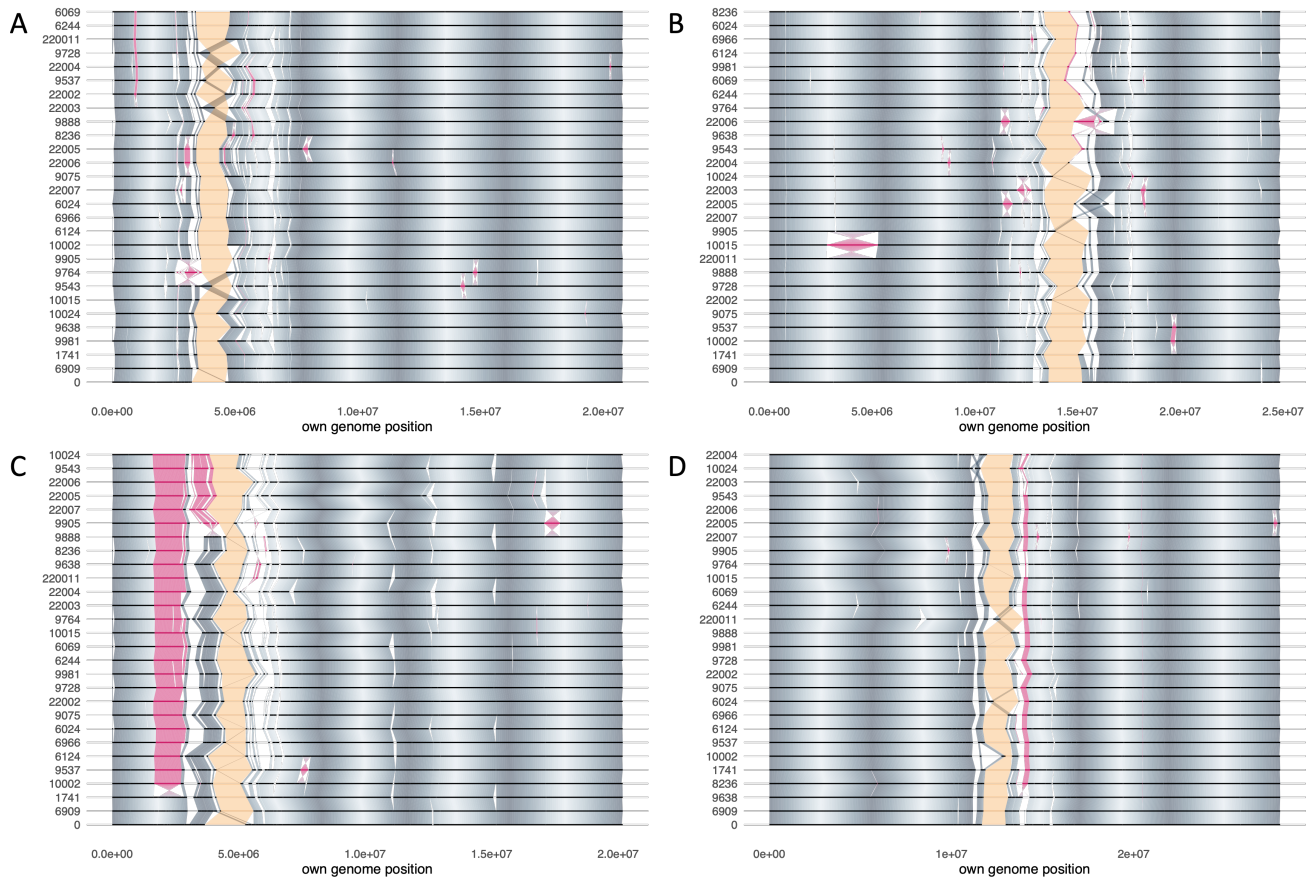
Extended Data Figure 2 – The causes of contig breaks. Stacked bar chart summarizing the type of repetitive element closest to each contig edge across the 27 assemblies, separately for scaffolded and unplaced contigs. 72% of the latter are terminated by centromeric or rDNA repeats.



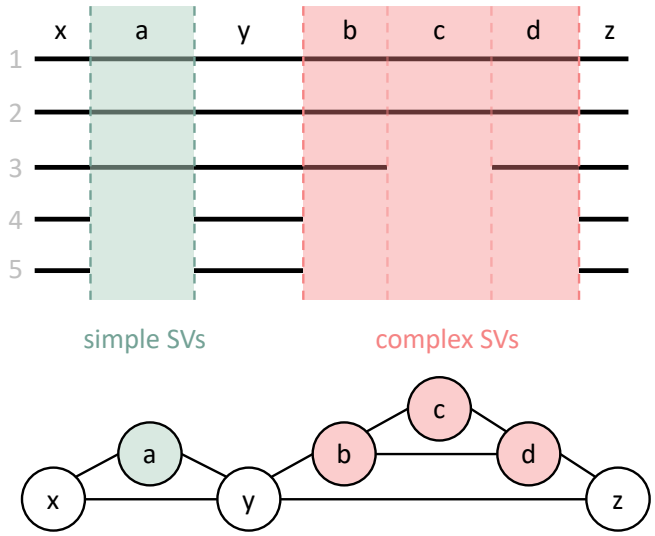
Extended Data Figure 3 – Scaffolded vs. unplaced contigs. The former correspond to the chromosome arms and contain mostly non-repetitive sequence and TEs, while the latter mostly contain centromeric and rDNA repeats, as well as organellar DNA sequence (*cf.* Fig. 1).



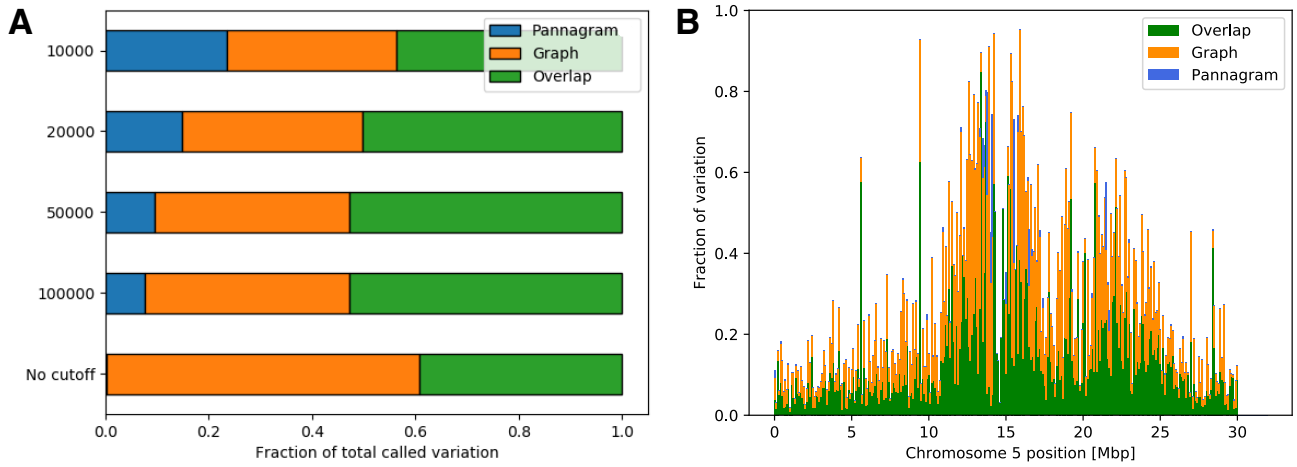
Extended Data Figure 4 – TE family size distribution. The cumulative length of all copies of a given TE family differs greatly across accessions, but the total TE content does not.



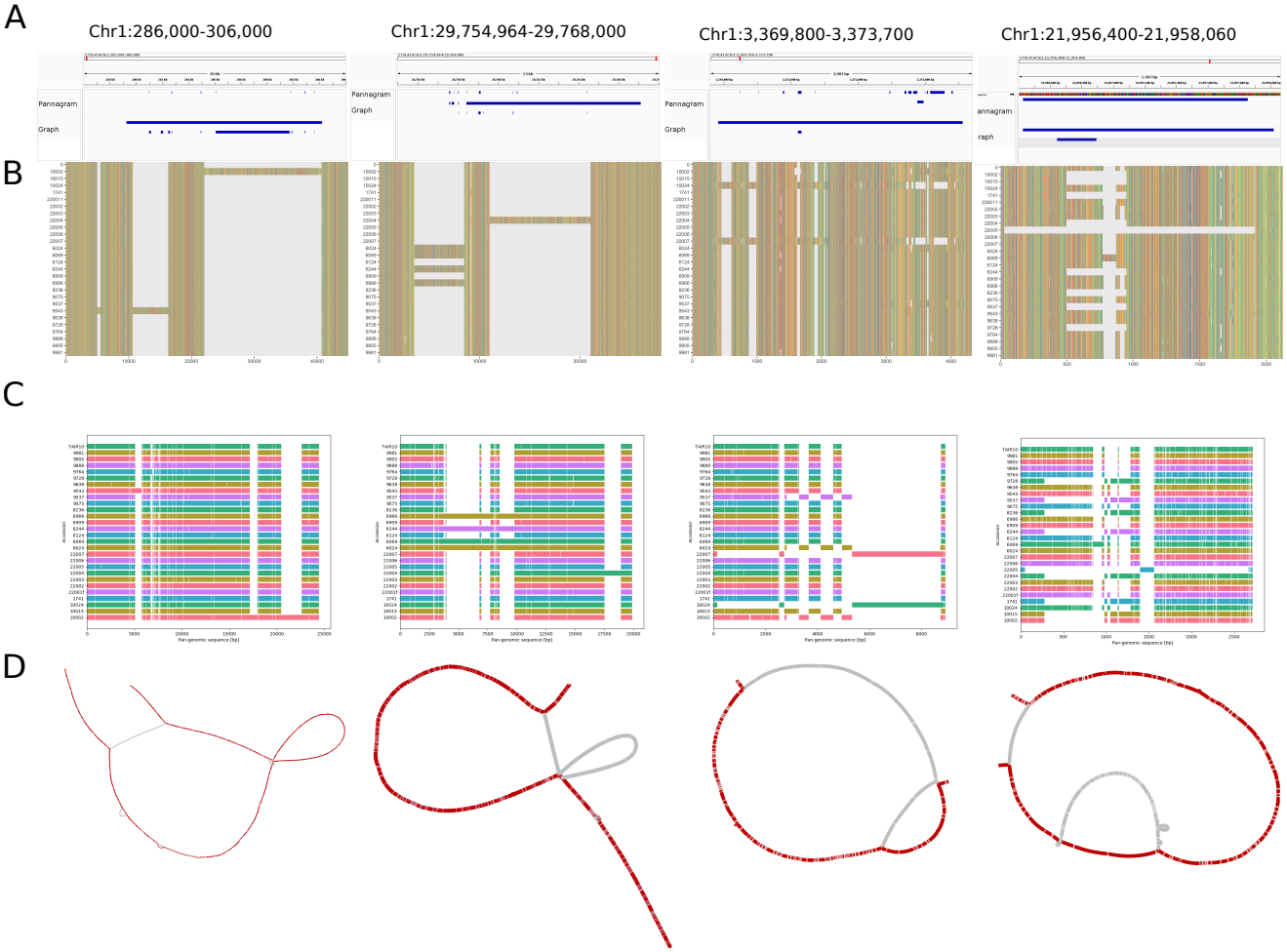
Extended Data Figure 5 – Whole-genome alignments. A. Chromosome 2. **B.** Chromosome 3. **C.** Chromosome 4. **D.** Chromosome 5. For chromosome 1 and further explanation, please see Fig 2A.



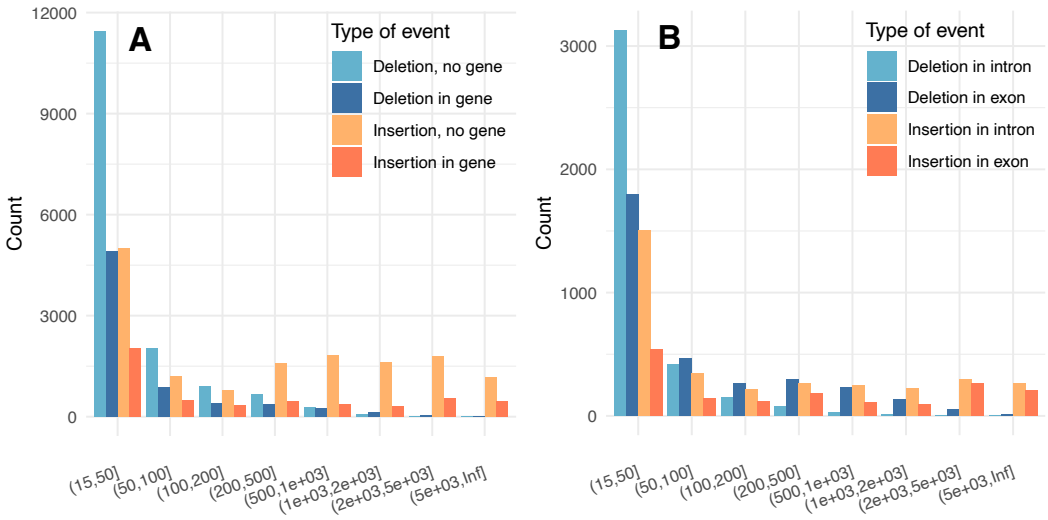
Extended Data Figure 6 – Simple and complex length variants. Cartoons illustrating our classification of length variants into simple and complex structural variants (SVs) in the whole-genome alignment and graph representations.



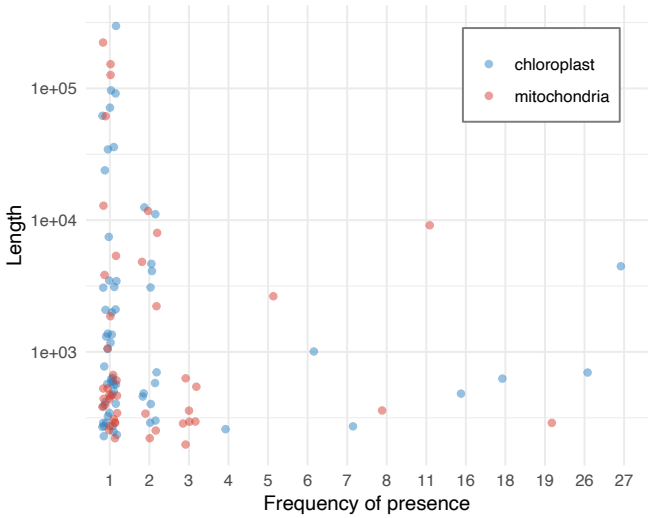
Extended Data Figure 7 – Comparing variation from Pannagram and PGGB. A. Bar plot displaying the overlap and method-specific detection of structural variation. Big bubbles are mostly due to technical artifacts and can be removed without losing biological relevant information. Removing big bubbles improves the overlap ratio. **B.** Positional representation of each (non)-overlapping variation represented on chromosome 1 of accession 6909 (Col-0). In regions close to the centromere we found a highly discrepancy between the two data sets. The data used in this plot is variation with a maximum size of 100 kb. Each bar covers 100 kb and the histogram is normalized by maximum count.



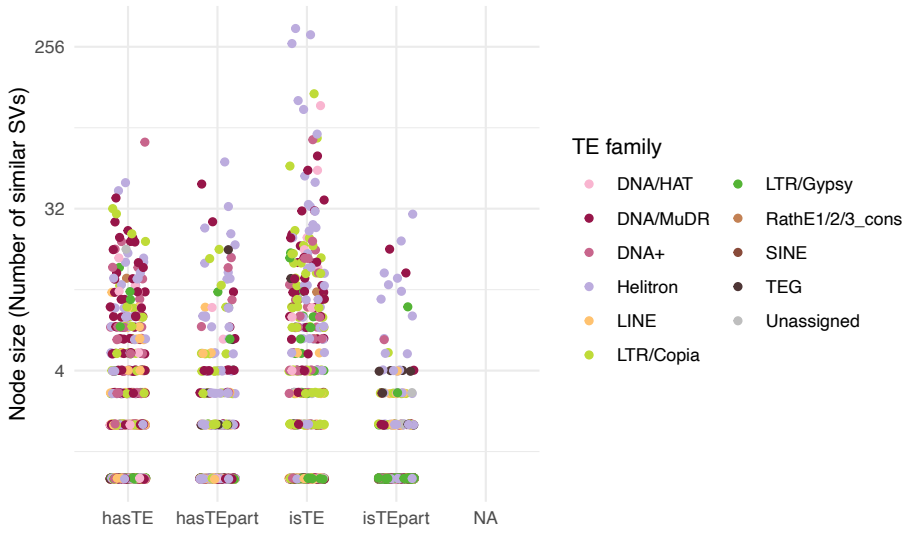
Extended Data Figure 8 – Comparing Pannagram and PGGB output on base-level. A. A selection of variation on shown on accession 1741. On top all SVs detected by Pannagram, on the bottom all variation called by the variation graph. **B.** Pannagram alignment on selected regions of chromosome 1 in accession 1741 (bottom). In top, we displayed the different variation length along the region (right: Pannagram, left: Genome graph). **C.** Graph alignment which was used to build genome graphs. **D.** 2D representations of the variation graphs at the given regions. Path of accession 1741 is highlighted with red.



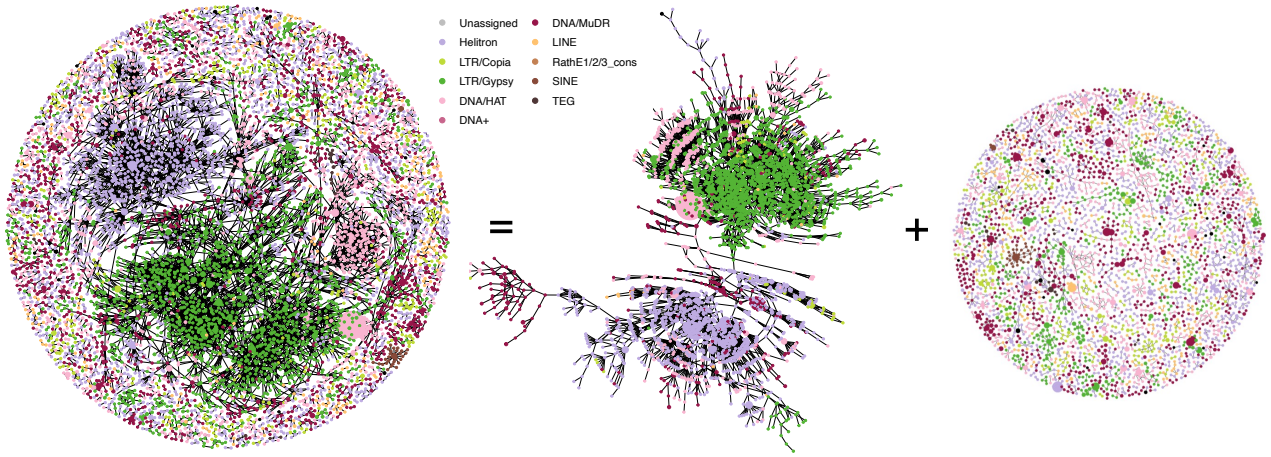
Extended Data Figure 9 – Location of indels with respect to genes. A. The number of putative insertion and deletions of varying lengths in gene and intergenic regions. **B.** Ditto for exons and introns.



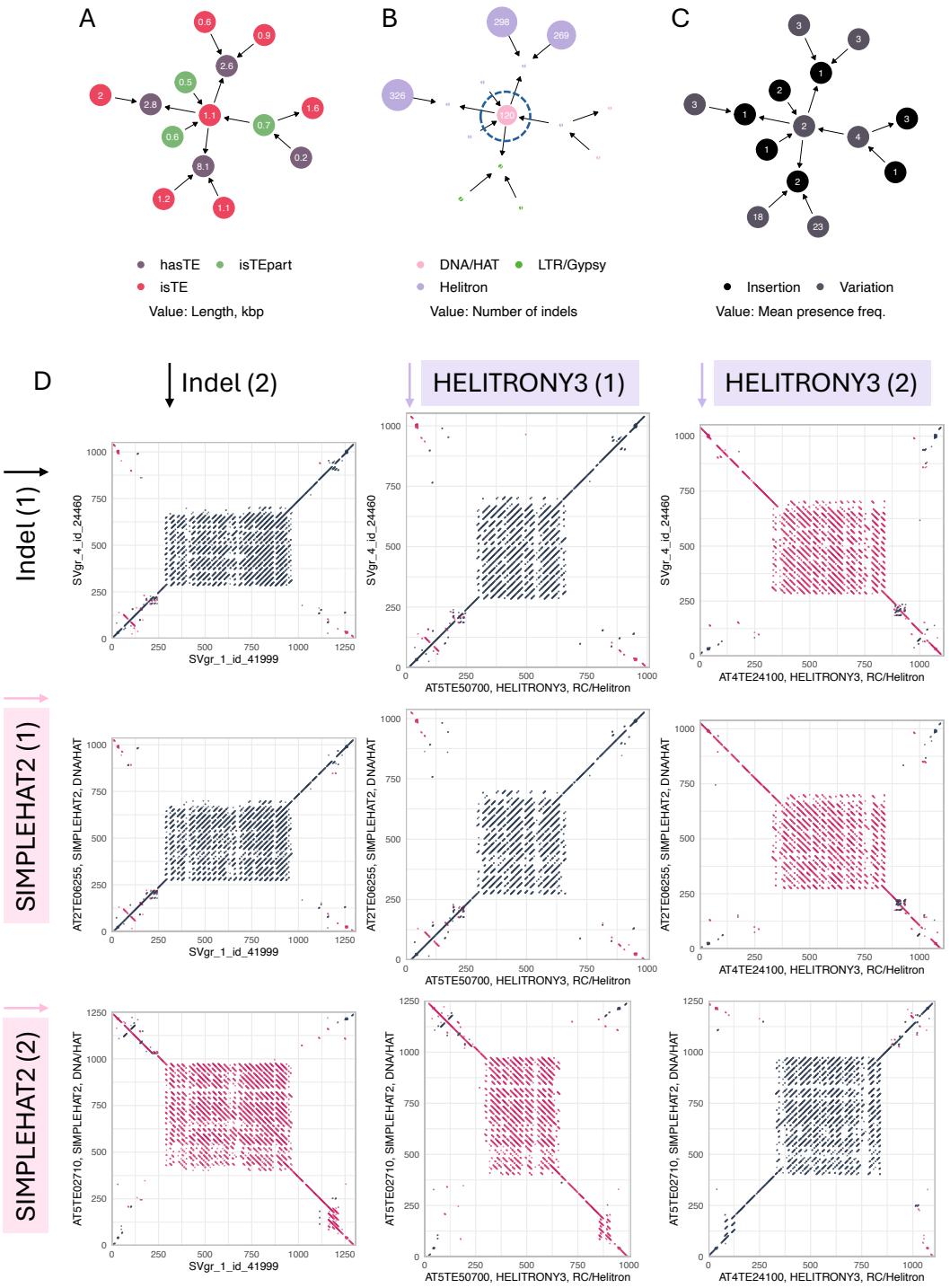
Extended Data Figure 10 – Frequency and size distribution of organellar inserts. Note that, since seven of these insertions were responsible for contig breaks in our chromosome scaffolds, we could not precisely determine their length (Extended Data Fig. 2).



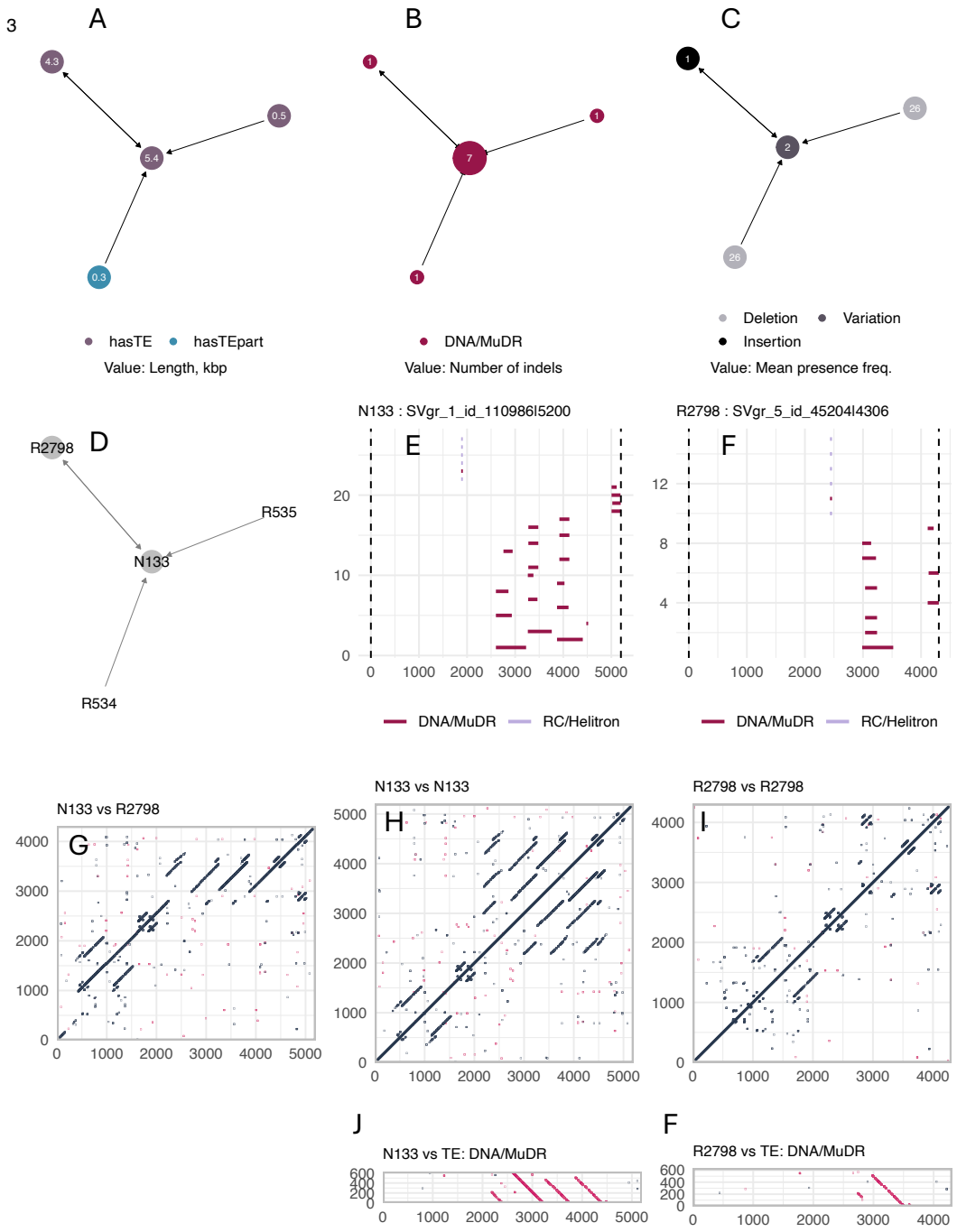
Extended Data Figure 11 – Size distribution of nodes in the graph base on TE-containing sSVs.



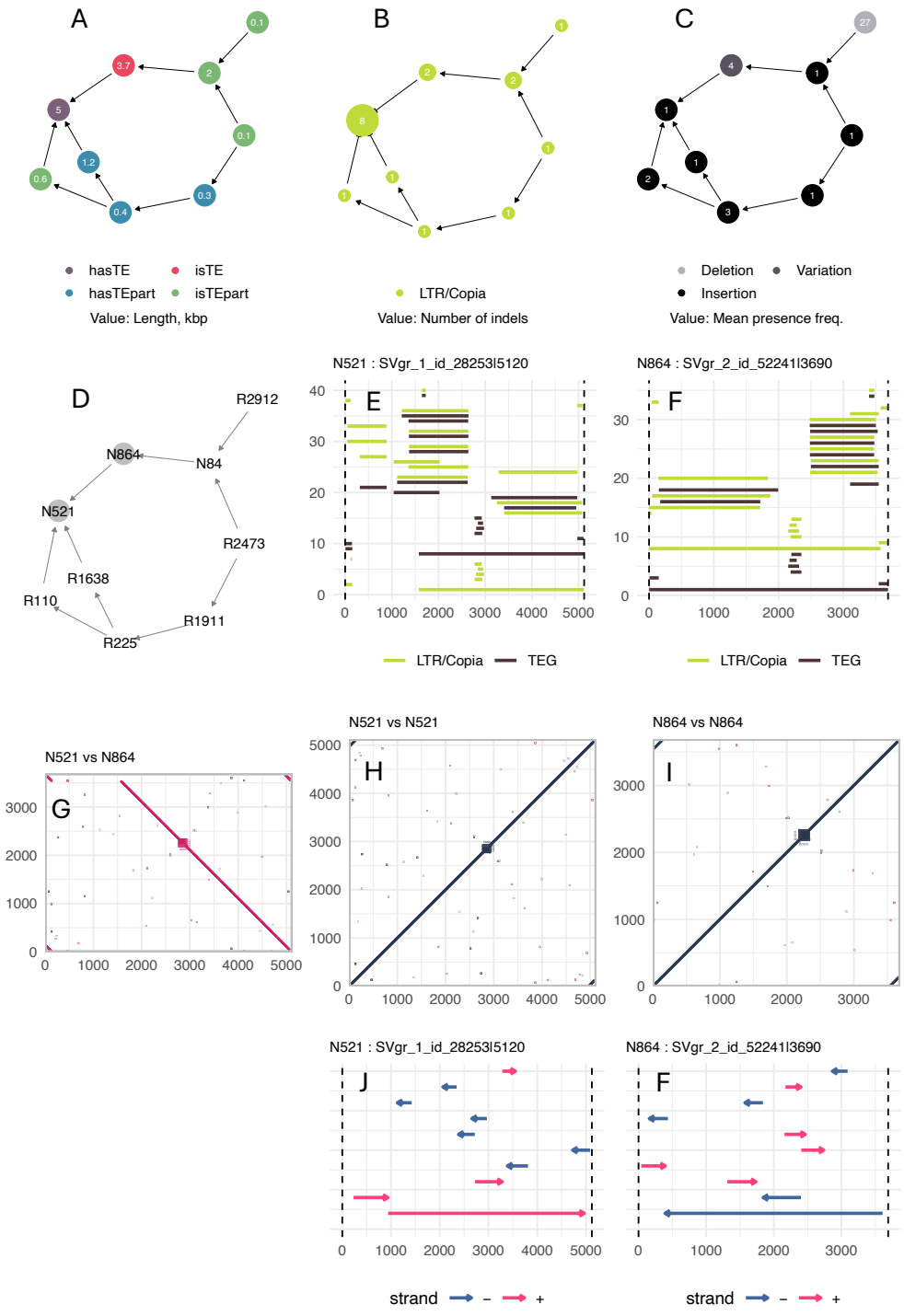
Extended Data Figure 12 – The graph of the nestedness of existing *A. thaliana* TE annotations. Each node is a cluster of similar sequences (with length and identity thresholds of 0.85), where the size of the node indicates its relative abundance. The graph is decomposed into one dominant connected component and a set of smaller ones.



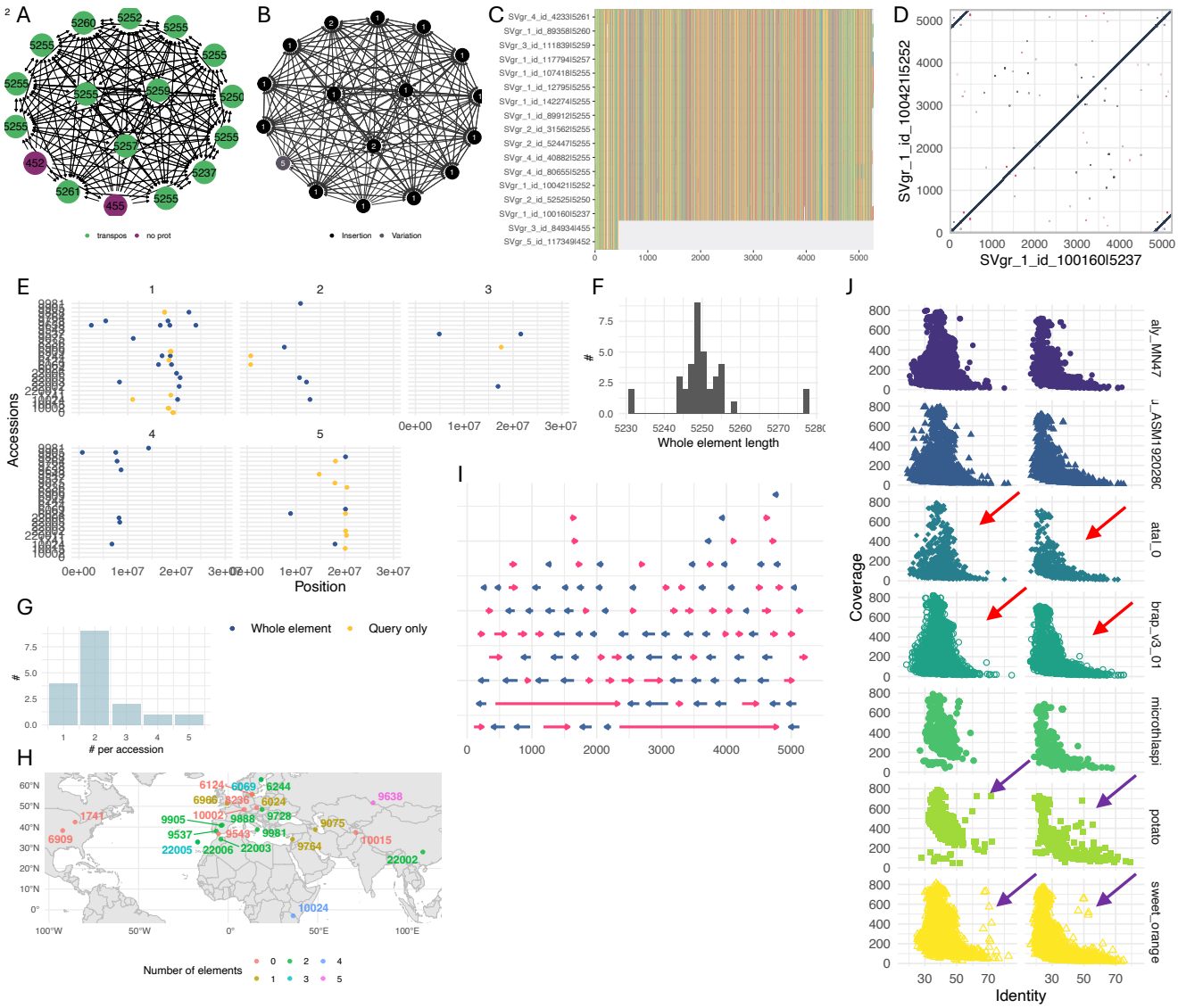
Extended Data Figure 13 – Very similar mobile elements have different names in *A. thaliana* TE annotation. A part of our Graph of Nestedness (Fig. 5) is: **(A)** colored based on TE content and shows length in Mb; **(B)** colored according to TE family and shows the number of cases in the pan-genome alignment; **(C)** colored to distinguish between insertions and deletions by the frequency of presence. **D.** Dot plots of two sSVs from the middle node from the above graph, two members of the annotated HELITRONY3 family, and three members of the SIMPLEHAT2 family. Dark color reflects the similarity on the forward strand, pink color on the reverse complement. Dot plots were constructed with a window parameter of 15 and the number of matches set to 12.



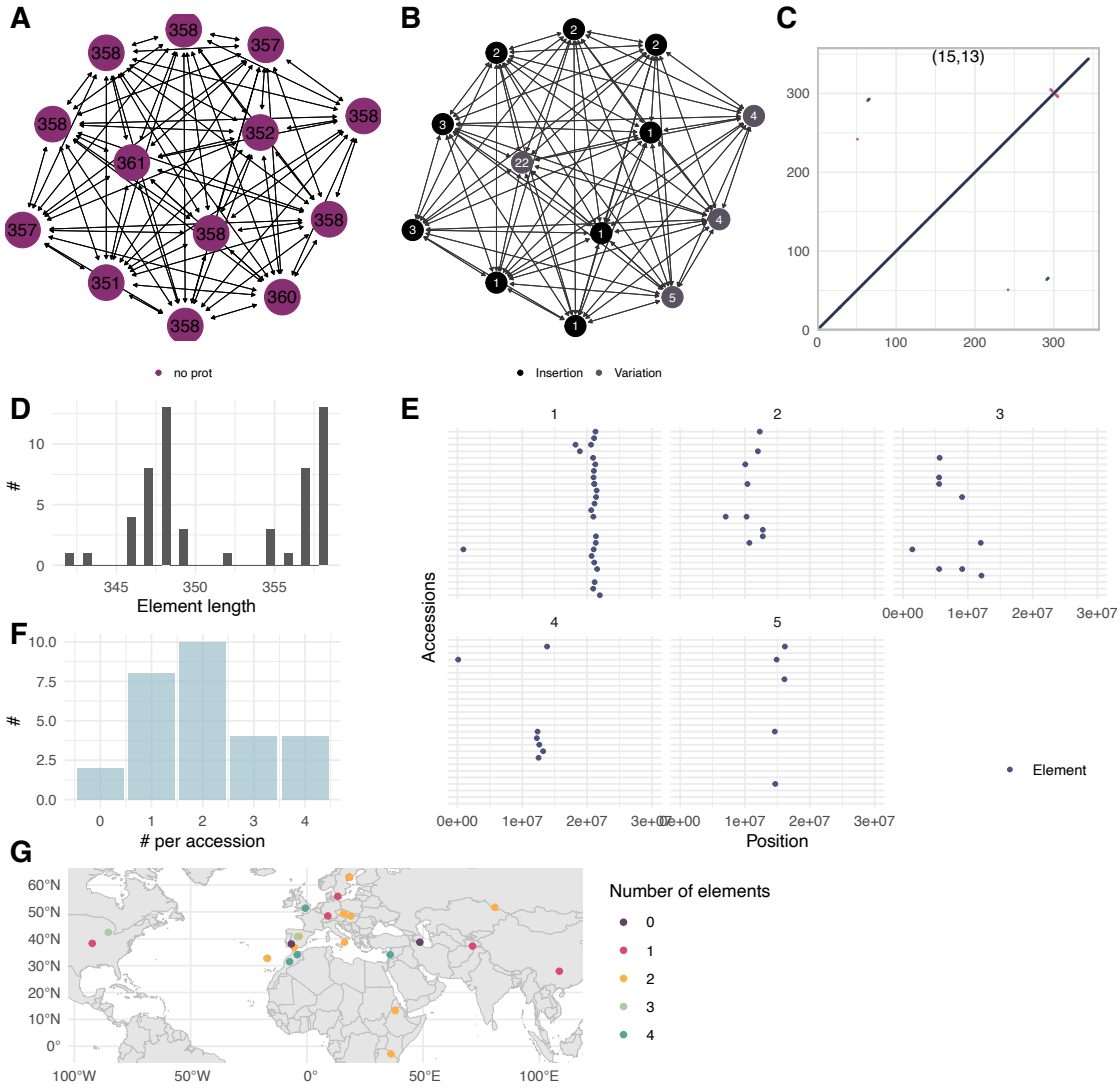
Extended Data Figure 14 – A mobile element that contains annotated TEs. A connected component from the graph of nestedness (Fig. 5) built on sSVs that overlap annotated TEs: **(A)** colored by TE content and labeled with lengths in Mb; **(B)** colored by TE family and labeled with the number of instances in the pangenome alignment; **(C** colored with phasing to insertions and deletions by the mean frequency of presence. **D.** The same connected component, but with node names. The gray color reflects two nodes of interest. **E.** BLAST results of annotated TEs vs one of the nodes of interest. **F.** BLAST results of annotated TEs vs another node of interest. **G.** Dot plot between sequences in the nodes of interest. **H-I.** Dot plot of sequences on itself. **J-F.** Dot plot of sequences on annotated full TEs.



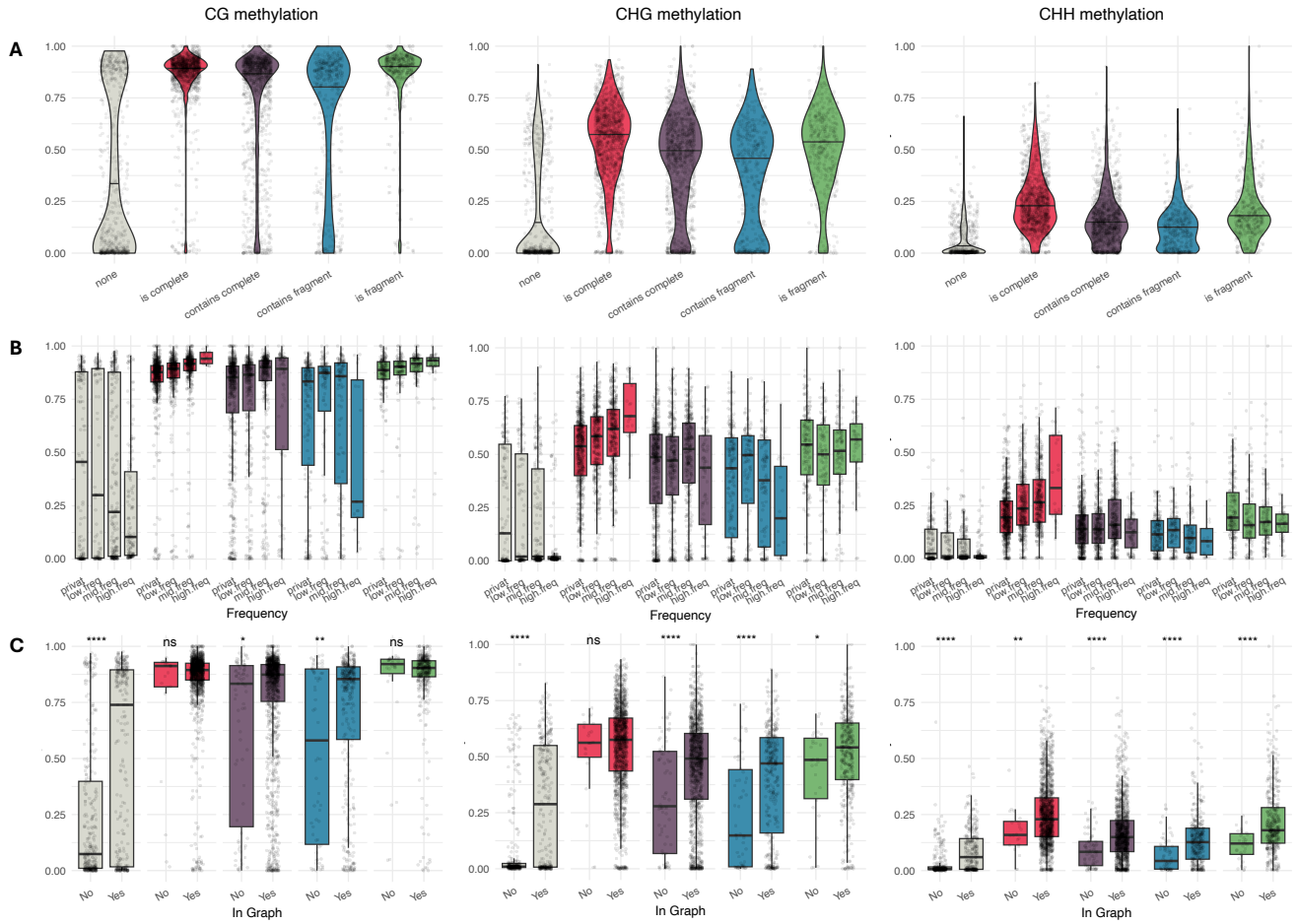
Extended Data Figure 15 – Another mobile element that contains annotated TEs. A connected component from the graph of nestedness (Fig. 5) built on sSVs that overlap annotated TEs: **(A)** colored by TE content and labeled with lengths in Mb; **(B)** colored by TE family and labeled with the number of instances in the pangenome alignment; **(C)** colored with phasing to insertions and deletions by the mean frequency of presence. **D.** The same connected component, but with node names. The gray color reflects two nodes of interest. **E.** BLAST results of annotated TEs vs one of the nodes of interest. **F.** BLAST results of annotated TEs vs another node of interest. **G.** Dot plot between sequences in the nodes of interest. **H-I.** Dot plot of sequences on itself. **J-F** Dot plot of sequences on annotated full TEs.



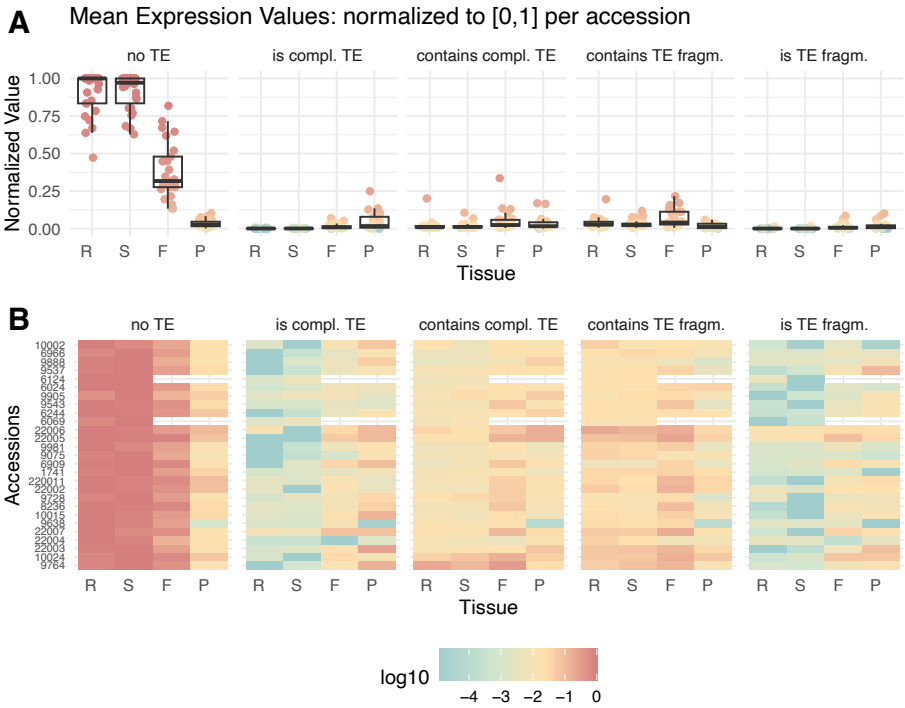
Extended Data Figure 16 – Un-annotated TE with evidence of horizontal transfer. A connected component from the graph of nestedness (Fig. 5) built on sSVs that have no overlap with known TEs for *A. thaliana*. **A.** The connected component is colored by keywords after the BLAST of ORFs and labeled with the length of sSVs. Each node represents one location of sSVs. **B.** The same but colored by the frequency of presence. **C.** Multiple alignment of all sequences from the graph. **D.** Dot plot of one of the long sequences to another copy. Lines at the angles reflect the presence of LTRs. **E.** Chromosomal distribution of the whole element (blue) and LTRs only (yellow). The figure contains all detected instances, not only those in sSVs. Note that the full element is absent from the *A. thaliana* reference genome. **F.** Length distribution of the whole elements (flanked by LTRs) in the population. **G.** Distribution of the number of occurrences of the whole element in accessions. **H.** Geographical distribution of the number of occurrences. **I.** Stretches of amino acids without stop codons in forward (pink) and reverse (blue) strands. **J.** BLAST results of coverage vs identity of the two longest amino acid stretches in all possible peptides of the selected plant species.



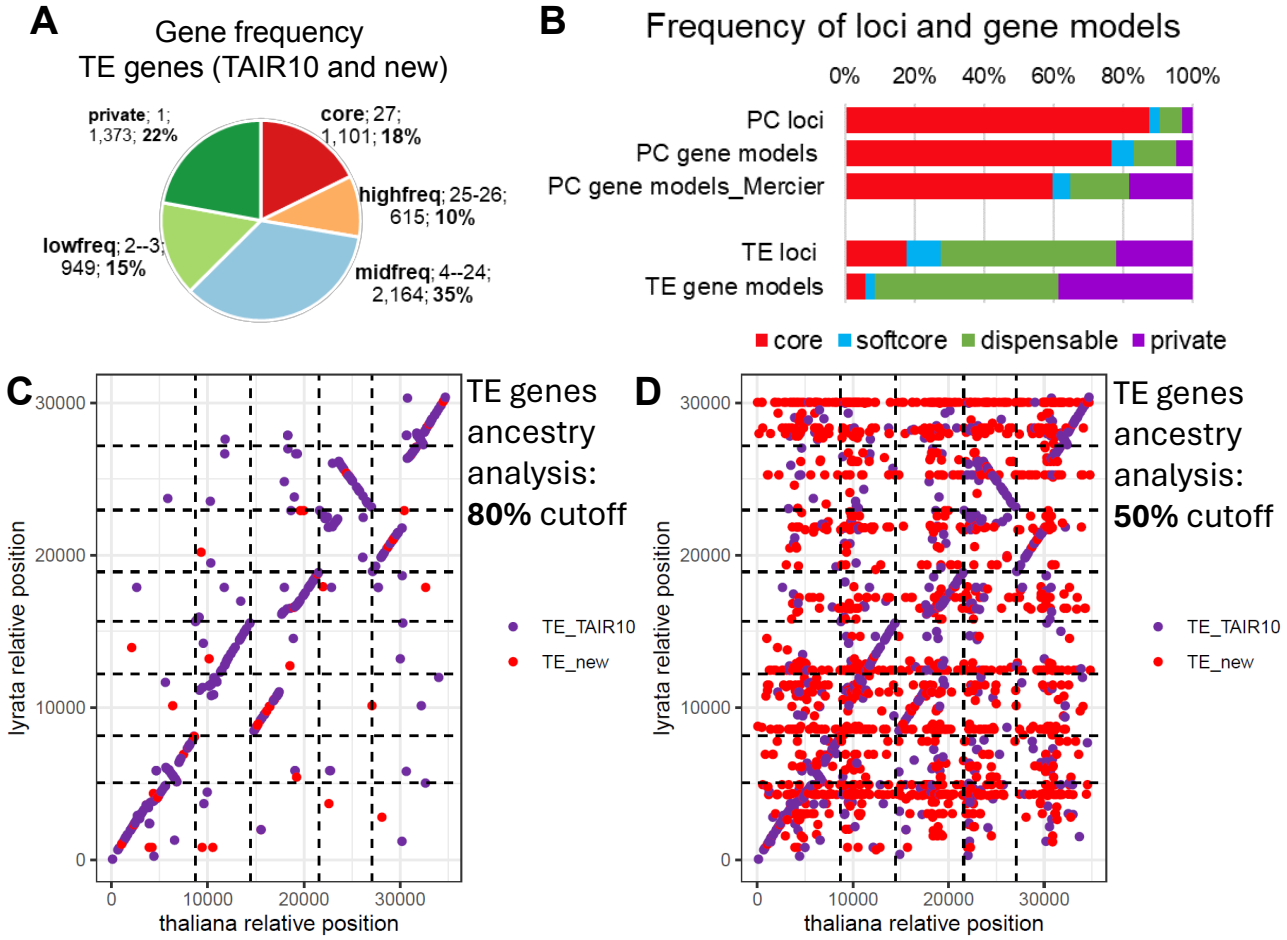
Extended Data Figure 17 – Putative novel mobile element family. A connected component from the graph of nestedness built on sSVs (Fig. 5) that have no overlap with known TEs in *A. thaliana* and no protein similarity to any known genes or TEs (based on an NCBI BLAST search). **A.** The connected component is uniformly colored as no sSV appear to encode a protein. Each node represents one sSV-location, and the numbers reflect the length of each sSVs. **B.** The same but colored by the frequency of presence. **C.** Length distribution of the elements. **D.** The dot-plot of one sSV shows that the element is not repetive. **E.** Chromosomal distribution of the whole element (blue). The figure contains all detected instances, not only those in sSVs. **F.** Distribution of the number of occurrences of the whole element in accessions. **G.** Geographical distribution of the number of occurrences. **J.** Results of the NCBI BLAST search with the default parameters on the database excluding *A. thaliana*.



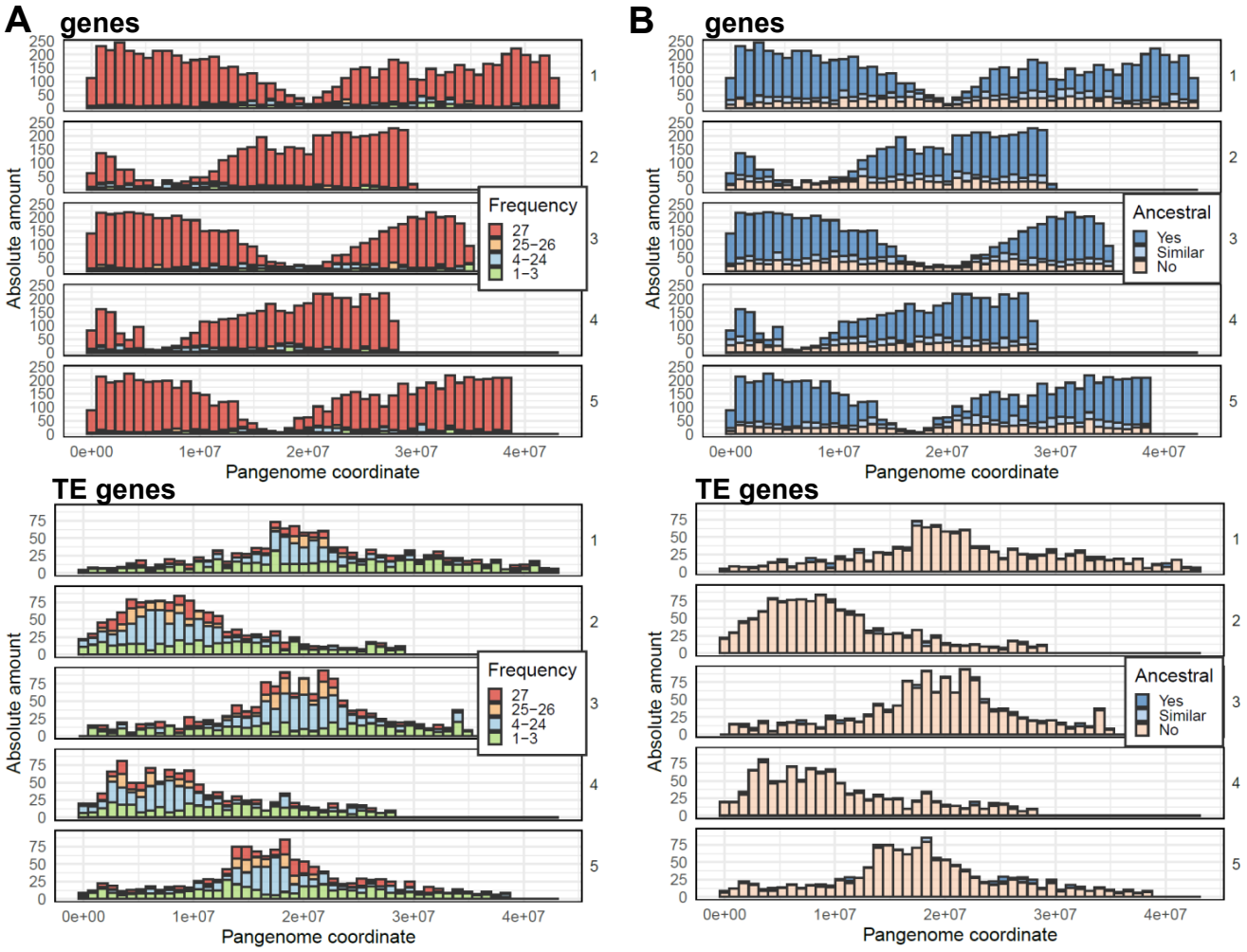
Extended Data Figure 18 – Methylation of sSVs. A. The distribution of methylation by TE content. **B.** The distribution of methylation by frequency of presence ("private" = 1, "low" = 2–3, "mid" = 4–24, "high" = 25–26). **C.** The distribution of methylation by membership in the graph of nestedness (Fig. 5).



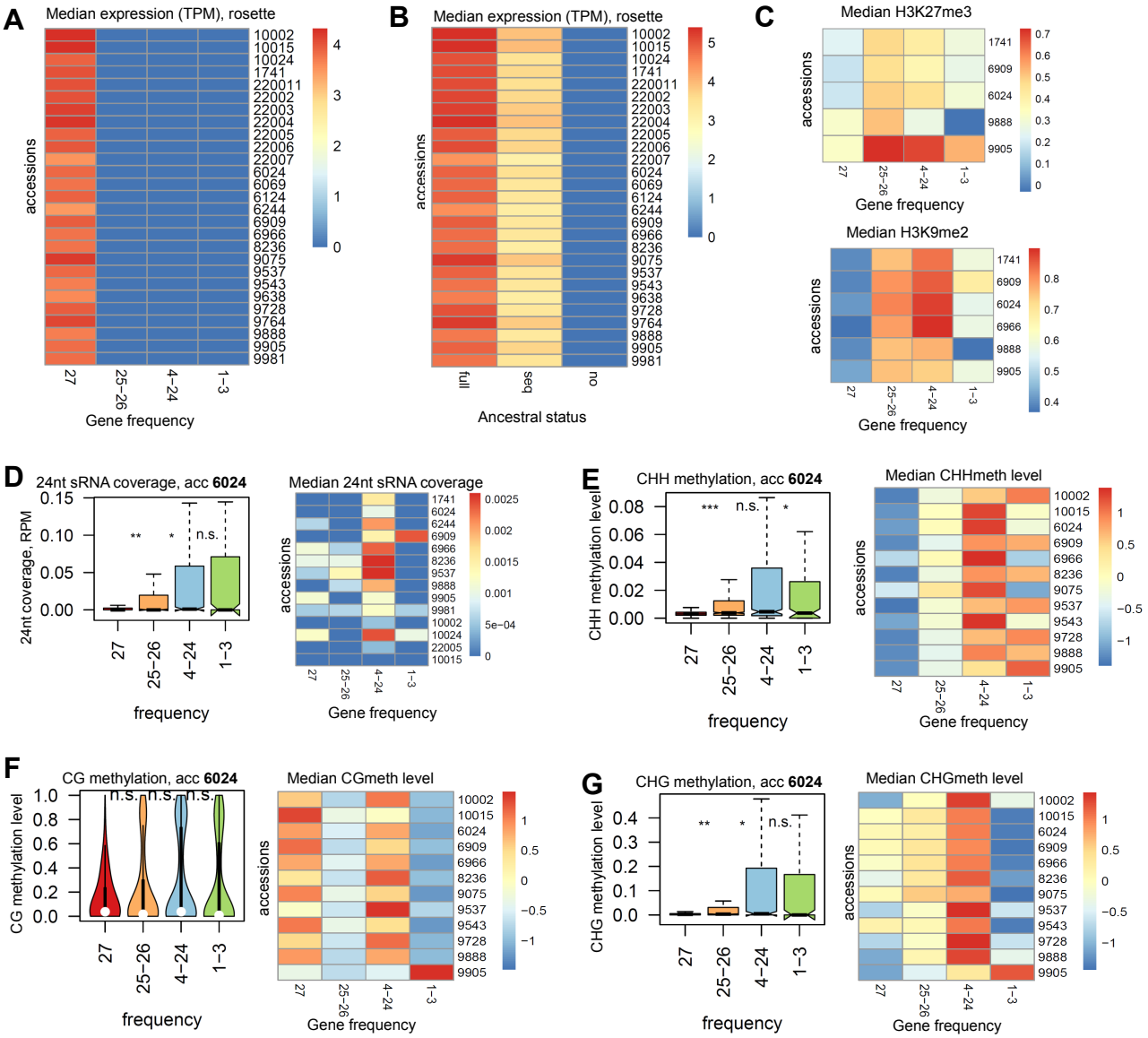
Extended Data Figure 19 – Expression in sSVs. Distribution of mean expression of sSVs in different TE-coverage categories and tissues: R stands for rosette; S for seedling; F for flower; and P for pollen. For each sSV, read counts were normalized by length. In each TE-coverage category in every tissue, expression values of sSVs were averaged. All average expression levels per accession were normalized to 1. The expression of "no TE" sSVs is the highest in all tissues except pollen, supporting the fact that these sSVs cover genes. In contrast, the "is TE" category shows increased expression in pollen. **A.** Each dot represents an expression value in one accession. **B.** Every row represents expression values in one accession.



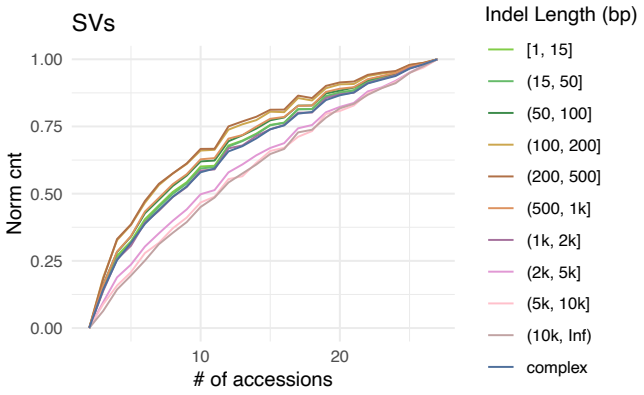
Extended Data Figure 20 – Details of gene-ome analysis. A. Gene frequency distribution for TE genes in our annotation. **B.** Comparison of our gene variation to the variation reported by the recently published study annotating genes from 69 accessions[22]. For both PC genes and TE genes gene model presence-absence variability is noticeably higher than gene locus presence-absence variability: this can be explained by the fact that the gene annotation pipeline might miss lowly expressed genes or genes not expressed in any of the four tissues we used. The gene model variability in the protein coding gene models of[22] is noticeably higher than in our data. This can be explained by the different approaches used for matching genes between accessions and estimating gene model variation: while we used pangenome coordinates, Lian *et al.*[22] used an orthogroup search approach. For example, if the gene locus is present in every accession and has a gene model in every accession, but in three accessions a shorter gene model was annotated, the orthogroup approach would produce two gene families for this locus and call both gene models dispensable, while our approach would treat this locus as one and would call such a gene (and gene models corresponding to this gene) core since it is represented by a gene model in every accession. **C.** Syntenic and sequence similarity analysis in *A. lyrata* for TE genes (TAIR10:TE and new:TE) with the same stringency as in Fig. 6C: 80% sequence similarity (see Methods). Genes are plotted not by their chromosomal position but by their consequential number along the chromosomes. The small number of data points outside of diagonals compared to PC genes (Fig. 6C) is caused by the fact that most TE genes are classified as non-ancestral and are not present on the plot as they were not found in the *A. lyrata* genome. **D.** Same syntenic and sequence similarity analysis in *A. lyrata* for TE genes as in (C) but using a less stringent sequence similarity cutoff of 50%.



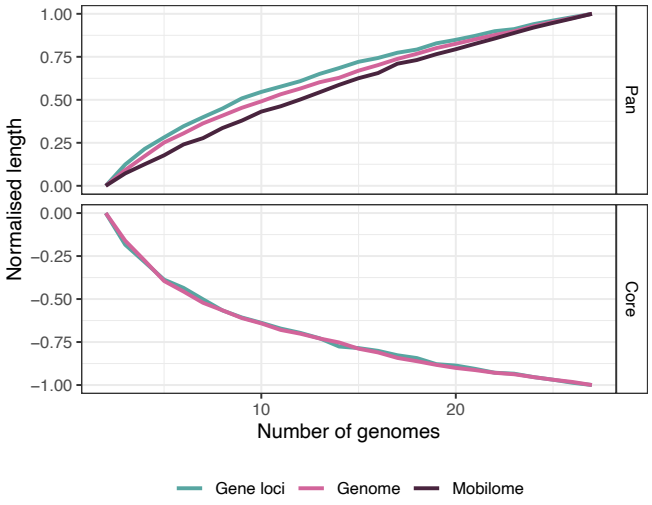
Extended Data Figure 21 – Chromosomal gene distribution for all 5 chromosomes. A. Chromosomal distribution of PC (top) and TE (bottom) genes of different frequencies. **B.** Chromosomal distribution of Ancestral:synteny+sequence similarity (Blue, "Yes"), Ancestral:only sequence similarity (Light blue. "Similar"), and non-ancestral (peach, "No") PC (top) and TE (bottom) genes. TAIR10:genes and new:genes are displayed under "genes" and new:TE are displayed under "TE genes".



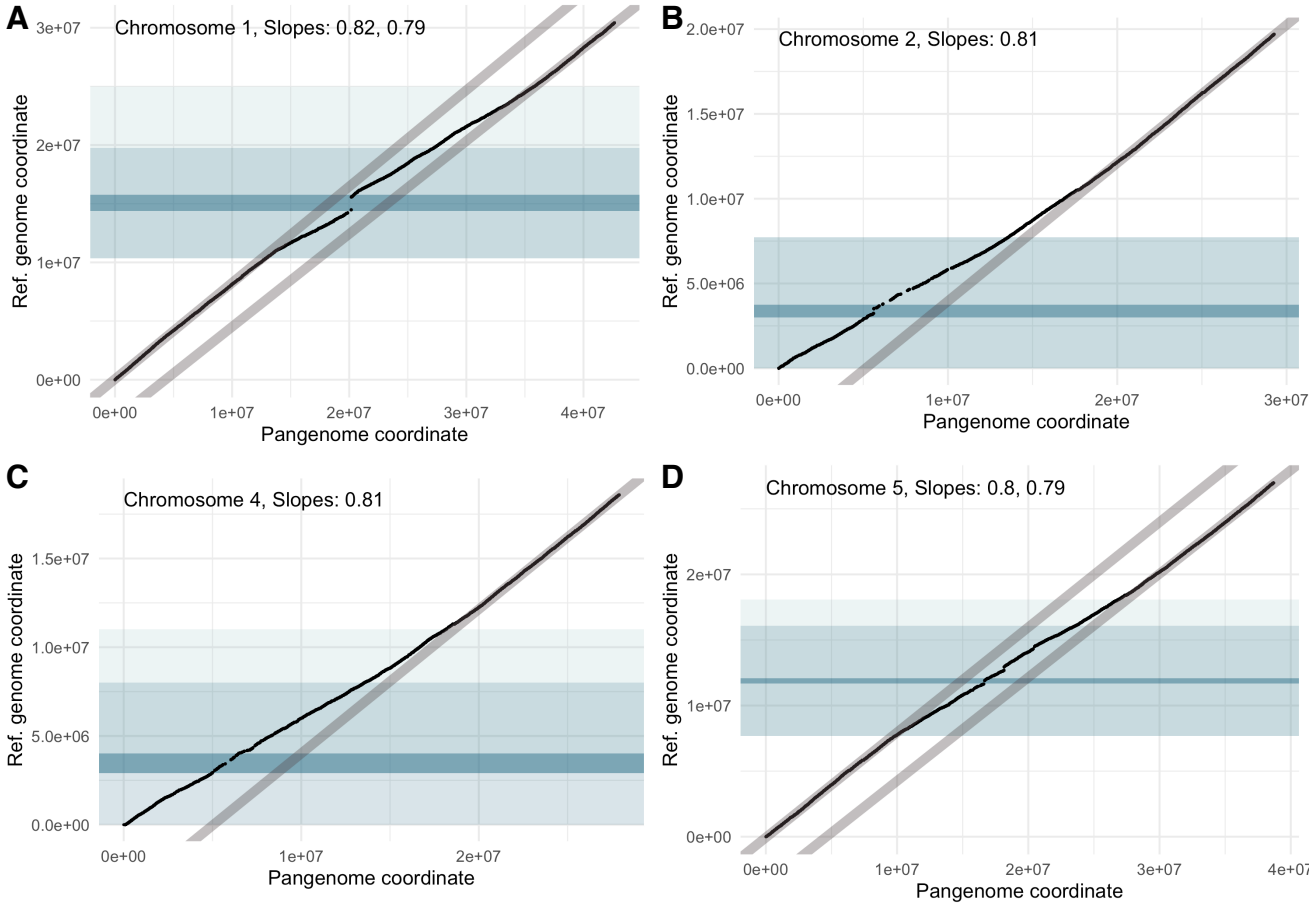
Extended Data Figure 22 – Gene silencing. A. Expression of genes vs. their presence frequency. The median TPM for each gene category is plotted for each accession. **B.** Expression of genes vs. their ancestral status. The median TPM for each gene category is plotted for each accession. **C.** H3K27me3 and H3K9me2 silencing[51] of genes vs. their presence frequency. The median of quantile and Input normalized ChIP-seq signal for each gene category is plotted for each accession. **D.** 24-nt small RNA targeting of genes of different frequency. Left: 24-nt locus coverage in flowers of accession 6024. Right: the median of 24-nt coverage is plotted for each gene category and accession. **E-G.** locus-wide CHH, CG and CHG methylation levels in leaves[96] for genes of different frequency. Left: accession 6024. Right: the median of CHH methylation level is plotted for each gene category and accession.



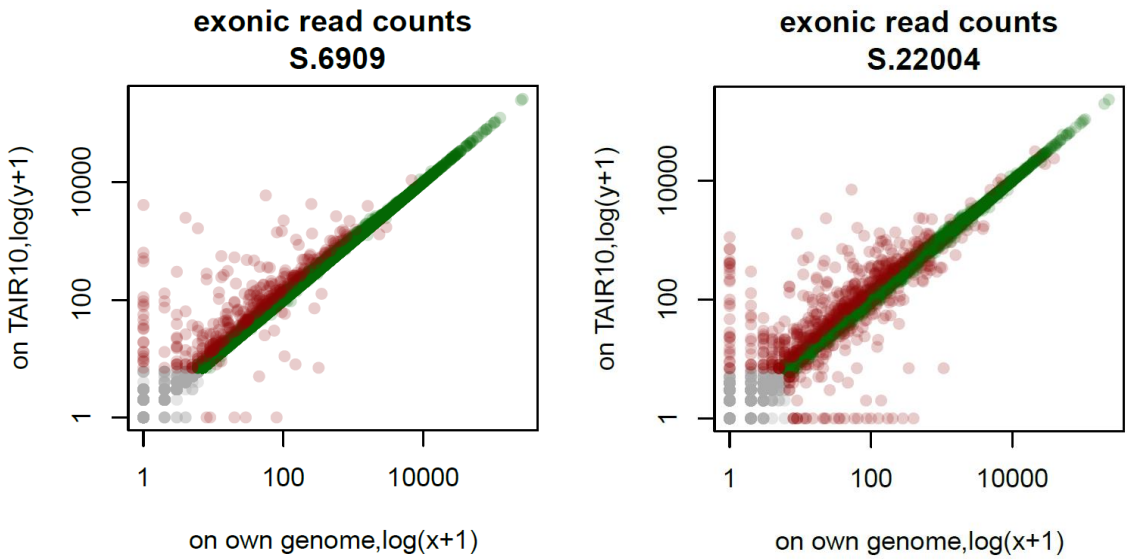
Extended Data Figure 23 – Saturation curves for sSVs for different length ranges. Each curve is drawn through points that are averaged values from 20 repetitions. The curves were normalized to start and end at the same points. Larger sSVs have a higher growth rate.



Extended Data Figure 24 – Normalized saturation curves for different genome components The dependence on sample size for the union ("pan") and intersection ("core") sequence length, separately for the full genome, the mobile-ome, and the gene-ome. Within the "pan" curves, the mobile-ome grows the fastest, the gene-one grows the slowest, and the whole genome grows at a intermediate rate. Gene-one and whole genome "core" curves show the same decay trend.



Extended Data Figure 25 – Pan-genome vs. reference genome coordinates for each chromosome. The pericentromeric region (light blue) shows a higher "dilution" of the spatial coordinates due to the increased number of SVs in this region. The centromeric region is dark blue.



Extended Data Figure 26 – Reference bias: gene expression. Raw counts in seedling RNA-seq samples for genes mapped to TAIR10 (using TAIR10 gene annotation) and to the own genome (using our annotation). Only genes matched one-to-one matched genes were analyzed.