nature genetics

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41588-018-0237-2

In the format provided by the authors and unedited.

Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L.

Jisen Zhang[1,20]*, Xingtan Zhang[1,20], Haibao Tang[1,20], Qing Zhang[1,20], Xiuting Hua[1], Xiaokai Ma[1], Fan Zhu[2], Tyler Jones[3], Xinguang Zhu[4], John Bowers[5], Ching Man Wai[6], Chunfang Zheng[7], Yan Shi[1], Shuai Chen[1], Xiuming Xu[1], Jingjing Yue[1], David R. Nelson[8], Lixian Huang[1], Zhen Li[1], Huimin Xu[1], Dong Zhou[1], Yongjun Wang[1], Weichang Hu[1], Jishan Lin[1], Youjin Deng[1], Neha Pandey[2], Melina Mancini[2], Dessireé Zerpa[2], Julie K. Nguyen[2], Liming Wang[1], Liang Yu[2], Yinghui Xin[2], Liangfa Ge[2], Jie Arro[2], Jennifer O. Han[2], Setu Chakrabarty[2], Marija Pushko[2], Wenping Zhang[1], Yanhong Ma[1], Panpan Ma[1], Mingju Lv[4], Faming Chen[9], Guangyong Zheng[9], Jingsheng Xu[1], Zhenhui Yang[1], Fang Deng[1], Xuequn Chen[1], Zhenyang Liao[1], Xunxiao Zhang[1], Zhicong Lin[1], Hai Lin[1], Hansong Yan[1], Zheng Kuang[1], Weimin Zhong[1], Pingping Liang[1], Guofeng Wang[1], Yuan Yuan[1], Jiaxian Shi[1], Jinxiang Hou[1], Jingxian Lin[1], Jingjing Jin[10], Peijian Cao[10], Qiaochu Shen[1], Qing Jiang[1], Ping Zhou[1], Yaying Ma[1], Xiaodan Zhang[1], Rongrong Xu[1], Juan Liu[1], Yongmei Zhou[1], Haifeng Jia[1], Qing Ma[1], Rui Qi[1], Zhiliang Zhang[1], Jingping Fang[1], Hongkun Fang[1], Jinjin Song[1], Mengjuan Wang[1], Guangrui Dong[1], Gang Wang[1], Zheng Chen[1], Teng Ma[1], Hong Liu[1], Singha R. Dhungana[11], Sarah E. Huss[2], Xiping Yang[12], Anupma Sharma[13], Jhon H. Trujillo[14], Maria C. Martinez[14], Matthew Hudson[15], John J. Riascos[14], Mary Schuler[2], Li-Qing Chen[2], David M. Braun[11], Lei Li[1], Qingyi Yu[13], Jianping Wang[1,12], Kai Wang[1], Michael C. Schatz[16], David Heckerman[17], Marie-Anne Van Sluys[18], Glaucia Mendes Souza[19], Paul H. Moore[3], David Sankoff[7], Robert VanBuren[6], Andrew H. Paterson[5], Chifumi Nagai[3]* and Ray Ming[1,2]*

[1]Fujian Agriculture and Forestry University and University of Illinois at Urbana-Champaign School of Integrative Biology Joint Center for Genomics and Biotechnology, National Sugarcane Engineering Technology Research Center, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Key Laboratory of Genetics, Breeding and Multiple Utilization of Corps, Ministry of Education, Fujian Agriculture and Forestry University, Fuzhou, China. [2]Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [3]Hawaii Agriculture Research Center, Kunia, HI, USA. [4]Institute for Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai, China. [5]Department of Plant Biology, University of Georgia, Athens, GA, USA. [6]Department of Horticulture, Michigan State University, East Lansing, MI, USA. [7]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario, Canada. [8]Department of Microbiology, Immunology and Biochemistry, University of Tennessee HSC, Memphis, TN, USA. [9]Chinese Academy of Sciences–Max-Planck-Gesellschaft Partner Institute for Computational Biology, Chinese Academy of Sciences, Shanghai, China. [10]China Tobacco Gene Research Center, Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou, China. [11]Division of Biological Sciences, University of Missouri, Columbia, MO, USA. [12]Department of Agronomy, University of Florida, Gainesville, FL, USA. [13]Texas A&M AgriLife Research, Texas A&M University System, Dallas, TX, USA. [14]Centro de Investigación de la Caña de Azúcar de Colombia (Cenicaña), Cali, Colombia. [15]Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [16]Departments of Computer Science and Biology, Johns Hopkins University, Baltimore, MD, USA. [17]Microsoft Research, Redmond, WA, USA. [18]Departamento de Botânica, Instituto de Biociências, Universidade de São Paulo, São Paulo, Brazil. [19]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, Brazil. [20]These authors contributed equally: Jisen Zhang, Xingtan Zhang, Haibao Tang, Qing Zhang. *e-mail: zjisen@fafu.edu.cn; cnagai@harc-hspa.com; rayming@illinois.edu

Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L.

Supplementary Notes

Hi-C scaffolding and chromosomal assembly

The major problem of scaffolding polyploid genome is that Hi-C signals are frequently detected between allelic haplotypes and any existing stat of art Hi-C scaffolding program, such as LACHESIS [1] and SALSA [2], links the allelic haplotypes together. To solve the problem, we developed a new Hi-C scaffolding pipeline, called ALLHIC, specifically tailored to the highly heterozygous diploids or polyploid genomes. ALLHIC pipeline contains a total of 4 steps: *prune*, *partition*, *optimize* and *build*.

Prune. Pruning function will firstly allow us to detect allelic contigs based on a well-assembled close related species, for instance Sorghum in our case. Signals (normalized Hi-C reads) between allelic contigs are removed from the input BAM files. In polyploid genome assembly, haplotypes that share high similarity are likely to be collapsed. Signals between the collapsed regions and nearby haplotypes result in chimeric scaffolds and are also removed in the *prune* step (Supplementary Figure 2).

Partition. ALLHIC partition applied the clustering algorithm implemented in LACHESIS package[1]. Briefly, partition works by iteratively merging contigs that contain significant number of links between them, based on a hierarchical linkage algorithm. In LACHESIS [1], this step requires a given number of partitions k. For each set of homeologous group for each of the 8 chromosome of AP85, our target is to construct $k = 4$ partitions for each of the 8 sets, for a total of 32 partitions. For AP85 datasets, we have experimented with $k = 4, 8, 16, 32, \ldots$ for each chromosome set, stopping when the homeolog alleles are separated in different partitions, which we then consider is the optimal k.

Optimize. The goal of the ALLHIC *optimize* step is to optimize the ordering and orientation of all the contigs in a given partition. This is traditionally called 'scaffolding' in the genome assembly field. We previously developed a method ALLMAPS [3] that performs scaffolding that seek to maximize the collinearity between the genome assembly and multiple maps by using Genetic Algorithm (GA). We chose to use Genetic Algorithm (GA) instead of some other heuristics such as local search, hill climbing, and greedy strategy to avoid getting stuck in local optima [3].

In ALLHIC, we used an objective function to maximize the total score based on the HiC data. The objective function, for each partition, is defined as the following:

$$S = \sum_{i,j \in contigs} \frac{L(i,j)}{D(i,j)}$$

where, $L(i,j)$ is the number of HiC links between two contigs i and j; and $D(i,j)$ is the distance in base pairs between the mid-points of two contigs i and j. Note that this definition of score S only influences the relative ordering of the contigs and not affected by their orientations. During the GA steps, the score S is repeatedly improved by mutating the current population of solutions, where the mutations operators are: 'inversion', which randomly selects two points in each solution and reverses the order of the scaffolds in between; 'insertion' which randomly translocates a scaffold and inserts it next to another randomly selected scaffold. These two mutation operators represent both large-scale and small-scale changes, and is set to be equally likely at 50%. For crossover operator, we use the 'Partially Mapped Crossover' (PMX) function that was shown to speed up convergence. The overall GA scheme is configured with mutation and crossover probability of 0.2 and 0.7, respectively, which were selected to offer a relatively rapid convergence. The population size is set at 100, and is allowed to evolve until there is no change of best solution in the last 5,000 generations as convergence criteria[3].

The orientations of the contigs are optimized separate from the ordering. The orientation problem uses a different objective function:

$$P = \sum_{i,j \in contigs} \sum_{\substack{k \in all\ links \\ between\ i\ and\ j}} \frac{1}{d_k}$$

where d_k is the distance of the two ends of a Hi-C link between contig i and j. Note that this distance is different depending on the relatively orientations between contig i and j, up to 4 distinct values $i+|j+,\ i+|j-,\ i-|j+,\ i-|j-$. The scoring function P is then simply the

sum of the reciprocal of the link distance. Compared to scoring function S, the calculation of P takes into account possible orientations, and is much more expensive to compute. The orientations of the contigs are optimized using a greedy method, which we find that works well empirically. We have the following operations: FLIPWHOLE, where we reverse the orientations for all contigs; and FLIPONE, where we reverse the orientations for a single contig. During each operation, we keep the changes if the score P improves, otherwise we reject the flipping proposal. An illustration of the ALLHIC *optimize* step can be seen in Supplementary Figure 3. We can see that over the course of the GA evolution, the HiC contacts are increasingly becoming diagonalized, while the synteny to related genomes are incrementally improved Supplementary Figure 3.

Build. Building the Hi-C based chromosome-level assembly requires a user-customized text file, which demonstrates the allelic relationship among the super-scaffolds generated in the last step. Similar to *prune* step, our *build* method firstly removes Hi-C signals between alleles and then searches for the best linkage signal for each super-scaffold. Super-scaffolds are clustered together once they have mutually best Hi-C signals under careful manual inspection. Position of unanchored contigs are further determined if they have best Hi-C signal with corresponding super-scaffolds. Contigs within each cluster are used for the second round of *optimize* step and a final Hi-C based chromosome-level assembly is released.

The Hi-C assembly generated 32 chromosomal level scaffolds, with length ranging from 54 Mb to 127 Mb. A total of 76,131 contigs, accounting for 92.3 % of assembled genome, were anchored in the 32 pseudo-chromosomes (Supplementary Table 7).

Validation. The key steps in ALLHIC have been validated using a variety of real datasets. To validate the *prune* and *partition* step, we constructed a 'synthetic' genome by mixing the Hi-C data of two rice subspecies (*Oryza sativa spp. japonica* and *Oryza sativa indica*). Since the true genome sequences of the two subspecies are available, this has allowed us to compare the final reconstruction of contigs. More details will be explained elsewhere. We have also tested ALLHIC on a variety of genomes including the reconstruction of Arabidopsis ecotype Ler0 and alfalfa genome, where we could obtain the Hi-C data as well as the draft genome assembly.

Assessing the quality of the contig orientation based on Hi-C link distribution

In order to evaluate the confidence of the contig orientation based on Hi-C scaffolding, Hi-C reads were re-mapped to the chromosomal level assembly using BWA[4] program and only uniquely mapped reads were retained. We have computed the posterior probability of the orientation for each contig. We start by modeling the distribution of Hi-C links by extracting the sizes of all intra-contig Hi-C links. The resulting distribution $f(x)$ is shown in Supplementary **Figure 20**. This empirical distribution has a similar power decay to the distribution obtained in previous studies in human [1,5].

Intuitively, we expect to find more proximal Hi-C links than distal Hi-C links, which serves as the basis for the probabilistic framework to quantitatively estimate the confidence of each contig orientation. If the contig should instead be placed in a different orientation, then the collection of links from this contig to the other contigs on the same chromosome would show more proximal links, with the total likelihood given by the link distance distribution. On the other hand, if either orientation provides the similar likelihood scores, then we would call the contig orientation "low confidence", or "random oriented". Formally, if we define D_{i^+j} to be the set of inter-contig link distances when contig i assumes the forward orientation and D_{i^-j} to be the set of inter-contig link distances when contig i assumes the reverse orientation, we have:

$$\boldsymbol{P}(\boldsymbol{D_{i^+j}}|\boldsymbol{o_i}=+)\sim\prod_{\boldsymbol{j\neq i}}\boldsymbol{f}(\boldsymbol{D_{i^+j}})$$

$$\boldsymbol{P}(\boldsymbol{D_{i^-j}}|\boldsymbol{o_i}=-)\sim\prod_{\boldsymbol{j\neq i}}\boldsymbol{f}(\boldsymbol{D_{i^-j}})$$

Assuming we have equal prior probability of taking either forward or reverse orientation, then we have the posterior probability:

$$\boldsymbol{P}\left(\mathbf{o_i}=+\middle|\mathbf{D_{ij}}\right)=\frac{\mathbf{P}(\mathbf{D_{i^+j}}|\mathbf{o_i}=+)}{\mathbf{P}\left(\mathbf{D_{i^+j}}\middle|\mathbf{o_i}=+\right)+\mathbf{P}(\mathbf{D_{i^-j}}|\mathbf{o_i}=-)}$$

This posterior probability allows us to infer the relative confidence of the inferred orientations per contig. Naturally, longer contigs tend to have higher high confidence in their orientation, since there are more inter-contig links that could allow more accurate inference of the orientation. Contigs with fewer links to other contigs (in the extreme case, a single link) could show similar likelihood of assuming either orientation, which could be considered "randomly oriented". We have assessed the orientation of each placed contig probabilistically. The

percentage of contigs with highly confident orientations increases with contigs of longer size. Specifically, over contigs of size of 100kb, we have 85.3% that have a highly confident orientation (Supplementary **Table 25**). Conversely, contigs with more confident orientations also appear to be longer. The set of contigs over 99% posterior probability has an average length of 42.9kb, compared with an average of 38.0kb for all contigs (Supplementary **Table 26**). At 99% posterior probability cutoff, we have 64.4% of the assembled genome that are considered to be highly confident (Supplementary **Table 26**).

Genetic maps and validation of assembly

The haploid clone AP85-441 (1n = 4x = 32) is weak and sterile, not suitable for making a mapping population. The ultra-high density genetic map was constructed from a mapping population between double haploid AP83-108 (2n = 4x = 64) and its progenitor octoploid SES208 (2n = 8x = 64), the same anther parent for the haploid AP85-441. We sequenced 54 F1 individuals from a backcross between octoploid SES208 and a doubled haploid AP83-108 5× genome equivalents each. We then identified SNPs using the contigs from the genome assembly as a reference to detect 2,105,205 segregating SNPs[6]. A locus for each assembly contig was generated for by determining the consensus supported by ⩾75% of the individual SNP loci. A map was constructed based on 7,262 high confidence contigs for which the consensus call was unambiguous for all 54 individuals, and supported by at least 10 individual SNPs. The contigs for which a high confidence mapping loci could be generated represented 451.3Mbp and 998,370 individual SNPs. The genetic map assembled into 44 linkage groups using MapDisto[7]. Comparison of contig orders from the genetic map and the Hi-C based assembly revealed that the genetic map only covered approximately 50% of the sugarcane genome, with several whole chromosomes or chromosome arms in the whole genome assembly not represented in the genetic map. Because AP83-108 and AP85-441 are from two different gametes, two different sets of 32 chromosomes were randomly sorted into the sequenced haploid and mapped double haploid genomes, hence the missing chromosomes from random assortment and chromosome arms from recombination through two meiosis events, one in 1983 when the double haploid AP83-108 was generated and the other in 1985 when the haploid AP85-441 was generated. The contig orders and chromosomal assignments mostly agreed between the genetic map assignments and the Hi-C

assembly. For 89% of the contigs, the genetic map and Hi-C assembly were in agreement in chromosomal assignment and order. For 7% of the contigs, the genetic map disagreed with the Hi-C assembly, but assigned the contig to a different homologue of the chromosome, and 4% of the contigs that did not match were assigned to an unrelated chromosome.

37 BACs of *S. spontaneum* AP85-441 (NCBI accession numbers: MH182499-MH182581 and KU685404-KU685417) were used to assess the quality of genome assembly. These BACs were assembled into a single contig and were blasted against genome assembly, which showed that 100.00 % of sequences were mapped and 99.33 % of bases were recovered in our genome (Supplementary Table 8). Genome completeness was assessed based on 248 ultra-conserved core eukaryotic genes (CEGs) in CEGMA and 1,440 conserved plant genes in BUSCO with default parameters. The two programs reported 88.31 % and 95.4 % of completeness, respectively (Supplementary Tables 9 and 10). Moreover, we mapped illumina sequencing reads (~ 80 x) from short-insert size libraries back to AP85-441 genome using BWA (version 0.7.8). Results revealed that nearly 98.3 % of them have a good alignment with our genome assembly and 97.35 % of our genome assembly was covered by illumina reads (Supplementary Table 11). We also tested the potential cross-mapping problem due to closely related homoeologous regions based on illumina reads. A collection of 64 million reads were tested based on BWA mem alignments with default parameters and only 0.76 % (0.485 million) of them were cross-mapped different homologous chromosomes.

General pattern of collinearity between *S. bicolor* and *S. spontaneum*

We produced a genomic dot-plot of syntenically conserved orthologous gene pairs in *Sorghum bicolor* and *Saccharum spontaneum* by CoGe SynMap[8] (Supplementary **Figure 8A**). The dot-plot showed a general pattern of conservation of chromosome-level synteny in all four *Saccharum* homeologs corresponding to each of the Sorghum chromosomes 1,2,3,4,6,7,9 and 10, while *Sorghum* chromosomes 5 and 8 have both been fragmented into two parts that were translocated to other chromosomes. These translocations were reflected in the comparisons of all four homeologous chromosomes in both cases, indicating that these events predated the two *Saccharum* WGDs.

We also observed genomic inversion reflected in all of *Saccharum* 4A,4B,4C and 4D when compared to Sorghum. In addition, there were inversions affecting only two homeologous chromosomes, i.e., that occurred after the first WGD, but before the second WGD. This happened independently in *Saccharum* chromosomes 2 and 7 (both aligned to *Sorghum* chromosome 8). In this case, the homeology between 7A and 7B and between 7C and 7D appeared to be derived in the most recent WGD. Similarly for the homeology between 2A and 2B and between 2C and 2D, an inversion had well separated 2AB and 2CD. We also detected an inversion in one chromosome that occurred after the second WGD, namely, in *Saccharum* chromosome 6C aligning to *Sorghum* chromosome 5.

These genomic rearrangements that clearly distinguished subsets of the homeolog chromosomes such as chromosome 2 and 7, provided strong evidence of three disjoint time periods in the evolution of *Saccharum* – between speciation from *Sorghum* and the first WGD, between this WGD and the second WGD, and after both events until the present. In particular, there might be a significant time period between the two WGDs. We call this hypothesis the 'time-gap' model. The 'time-gap' model has some support based on the genomic rearrangement pattern between *S. bicolor* and *S. spontaneum*, specifically the inversions that involved only subsets of the homeologous chromosomes.

Testing the 'time-gap' model based on sequence divergence between homeologs

We further tested the 'time-gap' hypothesis that the two WGDs were separated by a significant period of time that elapsed between the two WGDs. Under the 'time-gap' model, there would be pairs of homeologs more similar to one another that were derived from the more recent WGD events. In addition to defining the similarities based on rearrangement patterns, we could also exploit the sequence similarities between the homeologous genes. As detailed below, a two-tiered sequence divergence pattern, if identified, could serve as evidence of significant time lapse between the WGDs.

We examined the similarities in the set of all paralogous gene pairs found in syntenic blocks. No apparent partition reflecting two events can be directly inferred from the distribution of the

similarities between pairs of paralogs, since only one prominent peak could be identified in the distribution (Supplementary **Figure 8B**). The wide spread of this distribution had led to the difficulties in separating the two events. Global distribution of homeolog divergence therefore did not have the required resolution for distinguishing evolutionarily recent events such as the two very recent WGDs in the *S. spontaneum* lineage.

We developed the following framework for a more sensitive analysis of gene pair similarities, exploiting the fact that subsets of paralogs are often located on all four of a set of homeologous chromosomes. Suppose U and W are two homeologous chromosomes after the first WGD, as depicted in Supplementary **Figure 8C**. Under the 'time-gap' model, these two chromosomes then diverged for some significant period of time. Following the subsequent WGD, suppose U gave rise to new homologous chromosomes V and X while W gave rise to Y and Z. (N.B: V,X,Y,Z is some permutation of A,B,C,D). The two pairs (V,X) and (Y,Z) should each be more similar than (V,Y), (V,Z),(X,Y) and (X,Z). This two-tiered partition of sequence divergence is then expected under a 'time-gap' model of WGD. The divergence can be measured by average sequence similarity between the gene pairs along the chromosomes.

For simplicity, we call this a 'perfect' pattern, as expected under the time-gap model, that partitions the 6 pairs into 2 highly similar disjoint pairs versus 4 less similar pairs. Such 'perfect' partition not only applies to whole chromosomes, but also applies to individual genes or chromosomal regions. The two other perfect configurations are (V,Y),(X,Z) vs. (V,X),(V,Z),(X,Y),(Y,Z) and (V,Z),(X,Y) vs. (V,Y),(V,X),(Y,Z),(X,Z), for a total of three configurations. Mathematically, among all 6 possible pairs, there are a total of $\binom{6}{2} = 15$ configurations to choose the top two pairs – and when chosen at random, 3 of these configurations are expected to turn out as 'perfect'. Therefore, under a null, random hypothesis, we can expect 20% 'perfect' configurations. In what follows, we attempted to prove or disprove the 'time-gap' model below based on analyses at various scales ranging from individual genes, regions to whole chromosomes.

First of all, we examined genes in sets of four (forming six pairs) in syntenically conserved positions on four homeologous chromosomes. We have identified 295 such sets of genes in the *Saccharum* genome using the CoGe tool. Only 46 (16%) show a perfect pattern, not better than random. Quadruples of genes considered individually, then, did not help us establish a time interval between the two WGDs.

Examining next all the gene pairs in sets of entire homeologous chromosomes, only one such set (3A,3B,3C,3D) satisfies the definition of a perfect pattern. For all other chromosome sets excluding homologs of Sb5 and Sb8, if (V,X) is the most similar pair, the second most similar pair is not (Y,Z), but one of the other four possibilities (V,Y), (V,Z),(X,Y) or (X,Z). Therefore, the perfect pattern occurs just in one out of eight sets of homeologs (13%), which is again less than the 20% expected from random chromosomal similarities. As an example, we illustrate all pairwise comparisons within (3A,3B,3C,3D) set in Supplementary **Figure 8D.** Although there are increased and decreased similarities in local regions in several pairwise comparisons in parallel, we could identify very few areas of perfect patterns that would be expected under the 'time-gap' model. Therefore, quadruples of homeologous chromosomes did not help us establish a time interval between the two WGDs, either.

Could this lack of partition of sequence divergence be explicable in terms of recombination events between chromosomes that are not sister homeologs under the most recent WGD? Suppose V and X accidentally recombined or were homogenized at some point in time resulting in two new chromosomes V' and X' – each of which is more similar to W on some segment, and more similar to Y on the rest of the chromosome, so that these "more similar" regions are disjoint. These local regions should still show perfect configurations, at least within those regions. If there are many more accidental recombination events so that regions undisturbed by recombination are shorter and more numerous, these undisturbed local tracts of regions should each be perfect.

By searching for 'perfect' partitions in smaller regions rather than whole chromosomes, we should greatly enrich the possibilities of finding fragments that are undisturbed by accidental recombination, if this was indeed prevalent. Thus, as a final, most sensitive approach, we aligned

each set of four homeologous chromsomes according to the 295 quadruples of syntenically conserved paralogs mentioned above, as illustrated in Supplementary **Figure 8E**. Small regions thus defined were then merged, resulting in a total of 32 regions across the genome. Within each of these regions, we took the average of all gene pairs in the six two-way comparisons, not only those defined by conserved quadruples of genes. We again found only marginally more perfect fragments than random (7/32 = 22%).

In conclusion, relying on the inversions we find on *Saccharum* chromosomes 2C and 2D versus 2A and B, and chromosome 7C and D versus 7A and 7B, we had postulated that some time must have elapsed between the two WGDs since these homeologous chromosomes were clearly separated by inversions that occurred in the time between the two WGDs. However, the amount of time must have been short on the evolutionary scale, or there has been extensive level of homogenization between the homeologs [9], since we failed to detect the perfect pattern that could predict gene pair divergence at any level, from quadruples of whole homeologous chromosomes, to quadruples of small syntenic fragments, to quadruples of paralogous genes in the *S. spontaneum* AP85 genome.

Differentiation of genomic diversity among four homologous haploid sets

The reads mapped to each of four homologous haploid sets (A, B, C and D) of AP85-441 genome were retrieved for each of 64 accessions using the SAMtools[10] and Bedtools[11]. The four sets of retrieved reads for each of 64 accessions were mapped to each of eight chromosomes in a consensus monoploid genome separately using Bowtie2[12] with default parameters. The variants were called from cohort of 256 BAM files generated from previous step for each of eight chromosomes. The HaplotypeCaller of GATK[13] was used to estimate the SNPs and Indels for putative diploids using the default parameters. The HaplotypeCaller outputted 17,531,765 unfiltered variants (SNPs and Indels). The distribution of calling depths (DP) of each raw variant were estimated as a criterion for variants filtering. Low depths and repetitive variants were removed from the raw VCF file if they had DP < 1 or DP > 5, minQ < 20. We allowed the variants sites with max-missing rate as 50%. These filtering strategies reduced the raw unfiltered set of variants (SNPs and Indels) to the working set of 68,911 variants.

Genomic diversity among different polyploidy accessions

To test the effects of polyploidization on the genetic diversity, we therefore compare the population nucleotide diversity (π) among accessions with different ploidy levels. We used 1,000-kb sliding window and 500-kb step to calculate the values of each statistic. In addition, we divided the 64 accessions into four groups (ploidy 6, 8, 10 and 13-16) depend on their ploidy level. The four groups are used to calculate the pairwise Weir and Cockerham's *Fst* between two of them using VCFtools (v0.1.12b) [14] with 1,000-k sliding window and 500-k step.

References

1 Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* **31**, 1119-1125, doi:10.1038/nbt.2727 (2013).

2 Ghurye, J., Pop, M., Koren, S., Bickhart, D. & Chin, C. S. Scaffolding of long read assemblies using long range contact information. *BMC Genomics* **18**, 527, doi:10.1186/s12864-017-3879-z (2017).

3 Tang, H. *et al.* ALLMAPS: robust scaffold ordering based on multiple maps. *Genome biology* **16**, 3, doi:10.1186/s13059-014-0573-1 (2015).

4 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

5 Dudchenko, O. *et al.* De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92-95, doi:10.1126/science.aal3327 (2017).

6 Bowers, J. E., Pearl, S. A. & Burke, J. M. Genetic Mapping of Millions of SNPs in Safflower (Carthamus tinctorius L.) via Whole-Genome Resequencing. *G3* **6**, 2203-2211, doi:10.1534/g3.115.026690 (2016).

7 Heffelfinger, C., Fragoso, C. A. & Lorieux, M. Constructing linkage maps in the genomics era with MapDisto 2.0. *Bioinformatics* **33**, 2224-2225, doi:10.1093/bioinformatics/btx177 (2017).

8 Lyons, E., Pedersen, B., Kane, J. & Freeling, M. The Value of Nonmodel Genomes and an Example Using SynMap Within CoGe to Dissect the Hexaploidy that Predates the Rosids. *Tropical Plant Biology* **1**, 181-190, doi:10.1007/s12042-008-9017-y (2008).

9 Chalhoub, B. *et al.* Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science* **345**, 950-953, doi:10.1126/science.1253435 (2014).

10 Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

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

12 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012).

13 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303 (2010).

14 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).

Supplementary Tables

Supplementary Table 1. Summary of AP85-441 BAC pools for Nextera DNA libraries construction

Number. of sample	No. of BAC clone pooled	Pooling strategy	Total BAC clone
603	48	Column pool	28944
96	64	Column pool	6144
1	35	Column pool	35
1	33	Column pool	33

Supplementary Table 2. The assembled results of BAC sequencing by ALLPATH-LG, SOAPdenovo and SPAdes

	N50(Kb)	Average contig size(Kb)	Contigs number	Total contigs size(Mb)
AllpathLG	7.38	4.53	566,004	2,564
SOAPdenovo	5.30	2.43	1,553,498	3,775
SPAdes	6.70	1.27	3,724,763	4,723

Supplementary Table 4. Statistics of Pacbio sequencing and correction

Items	Raw reads	FALCON corrected reads	Canu corrected reads
Total Number of reads	41,995,530	8,405,944	19,842,259
Total Number of sequenced Bases (Gb)	249	60.73	96.61
Mean reads length (bp)	5,937	7225	4,869
N50 (bp)	9,132	8697	7,759
Coverage (X)*	77.81	18.98	30.19

*Coverage (x) = (read count * read length) / estimated genome size.

Supplementary Table 5. Statistics of contig-level assembly

Items	BAC contigs	FALCON assembly	Canu assembly
Assembly size (Mbp)	3,468	2,041	3,132
No. of contigs	2,611,145	33,665	91,867
Maximum length (bp)	210,478	772,193	400,016
N90 (bp)	358	32,113	17,099
N80 (bp)	768	52,933	23,216
N70 (bp)	1,699	70,240	30,024
N60 (bp)	3,654	88,254	37,193
N50 (bp)	6,190	108,148	45,023
Average length (bp)	1,328	60,648	34,095

Supplementary Table 6. Assessment of AP85-441 genome assembly using 37 published BAC contigs

Dataset	Number	Total Length (bp)	Accuracy (%)	Bases covered by assembly (%)	Sequences covered by assembly (%)	With >90% sequence in same chromosome		With >50% sequence in same chromosome	
						Number	Percent (%)	Number	Percent (%)
BAC contigs	37	4,082,009	99.72	99.33	100	34	91.89	37	100.00

Supplementary Table 7. Statistics of Hi-C sequencing and mapping

Statistics of mapping	
Clean Paired-end Reads	1,001,283,445
Unmapped Paired-end Reads	43,139,596
Unmapped Paired-end Reads Rate (%)	4.987
Paired-end Reads with Singleton	637,840,383
Paired-end Reads with Singleton Rate(%)	62.858
Multi Mapped Paired-end Reads	201,902,772
Multi Mapped Ratio (%)	20.25
Unique Mapped Paired-end Reads	118,400,694
Unique Mapped Ratio (%)	11.904
Statistics of valid reads	
Unique Mapped Paired-end Reads	118,400,694
Dangling End Paired-end Reads	13,468,211
Dangling End Rate (%)	11.375
Self Circle Paired-end Reads	1,725,886
Self Circle Rate (%)	1.458
Dumped Paired-end Reads	22,993,334
Dumped Rate (%)	19.42
Interaction Paired-end Reads	78,664,945
Interaction Rate (%)	64.44
Lib Valid Paired-end Reads	69,851,750
Lib Valid Rate (%)	88.797
Lib Dup (%)	11.203

Supplementary Table 8. Overview of chromosome level assembly based on Hi-C data

	Haplotype A		Haplotype B		Haplotype C		Haplotype D	
	No. of contigs	**Length (Mb)**	**No. of contigs**	**Length (Mb)**	**No. of contigs**	**Length (Mb)**	**No. of contigs**	**Length (Mb)**
Chr1	3,203	114	3,582	123	2,771	99	3,168	117
Chr2	3,139	122	2,933	114	3,133	127	2,787	109
Chr3	2,112	78	2,551	101	2,580	96	1,597	62
Chr4	1,962	75	2,127	77	2,196	81	2,218	83
Chr5	2,391	91	2,387	93	2,319	91	2,201	85
Chr6	2,670	106	2,254	90	2,342	92	2,305	91
Chr7	2,021	79	2,052	81	2,258	86	2,218	85
Chr8	1,786	68	1,668	65	1,668	65	1,429	54

	Number of sequences	**Length of sequences (Mb)**
Anchored contigs	76,131	2,892
Unanchored contigs	15,704	240
Gaps	76,009	7.6
Total	91,835	3,132

Supplementary Table 9. Completeness of the genome based on CEGMA

Description	Fully mapped CEGs	Fully+partially mapped CEGs
Number of CEGs present in the assembly	219	233
Completeness of the genome (%)	88.31	93.95
Average number of orthologs per CEG	4.11	4.46
CEGs with more than one ortholog (%)	94.52	97.42

Supplementary Table 10. BUSCO analysis of Genome assembly

Description	Number	Percentage (%)
Complete BUSCOs (C)	1373	95.4
Complete and single-copy BUSCOs (S)	207	14.4
Complete and duplicated BUSCOs (D)	1166	81
Fragmented BUSCOs (F)	9	0.6
Missing BUSCOs (M)	58	4
Total BUSCO groups searched	1440	100

Supplementary Table 11. Assessment of genome consistency

Iterms	Statistics
Number of reads	1,649,618,792
Data size (Gb)	249
Mapped bases (Gb)	245
Map rate (%)	98.3
Genome Length (Mbp)	3,141
Mean Depth	77.8
Coverage Rate (%)	97.35

Supplementary Table 12. Identification of centromeres in AP85-441 genome

	Haplotype A		Haplotype B		Haplotype C		Haplotype D	
	Position	**Length (Mb)**	**Position**	**Length (Mb)**	**Position**	**Length (Mb)**	**Position**	**Length (Mb)**
Chr1	49.35-59.55	10.20	93.30-93.80	0.50	52.40-60.30	7.90	54.50-55.10	0.60
Chr2	52.70-57.30	4.60	52.65-53.00	0.35	57.40-59.05	1.65	52.65-53.00	0.35
Chr3	41.55-42.30	0.75	58.75-59.85	1.10	57.70-58.10	0.40	38.35-38.70	0.35
Chr4	NA	NA	36.10-36.35	0.25	38.05-49.90	11.85	45.55-50.35	4.80
Chr5	43.75-44.10	0.35	38.90-44.55	5.65	55.50-55.85	0.35	44.75-45.10	0.35
Chr6	35.65-35.95	0.30	28.00-28.25	0.25	28.10-32.05	3.95	NA	NA
Chr7	NA	NA	30.55-32.80	2.25	32.75-36.50	3.75	16.75-17.00	0.25
Chr8	34.05-37.00	2.95	31.80-32.95	1.15	34.70-34.95	0.25	NA	NA

Supplementary Table 13. Go enrichment of *S. sponntaneum* specific genes

GO term	Ontology	Description	Number in input list	p-value	FDR
GO:0009611	P	response to wounding	21	8.80E-15	7.30E-12
GO:0009605	P	response to external stimulus	22	3.50E-12	1.40E-09
GO:0004867	F	serine-type endopeptidase inhibitor activity	13	2.50E-08	1.40E-05
GO:0030414	F	peptidase inhibitor activity	13	1.20E-06	0.00022
GO:0004866	F	endopeptidase inhibitor activity	13	1.20E-06	0.00022
GO:0015935	C	small ribosomal subunit	6	4.60E-05	0.0075
GO:0033279	C	ribosomal subunit	8	7.70E-05	0.0075

Supplementary Table 14. Statistics of TEs in AP85-441 genome

	Number	Length (Mb)	% of repeats	% of genome
Total repeat fraction	3,182,244	1,842.06	100	58.65
Class I: Retroelement	1,436,832	1,432.72	77.78	45.62
LTR Retrotransposon	1,075,834	1,305.67	70.88	41.57
Ty1/Copia	281,132	445.64	24.19	14.19
Ty3/Gypsy	669,834	817.80	44.4	26.04
Other	124,868	42.24	2.29	1.34
Non-LTR Retrotransposon	263,418	103.50	5.62	3.3
LINE	204,645	94.77	5.14	3.02
SINE	58,773	8.74	0.47	0.28
Unclassified retroelement	97,580	23.54	1.28	0.75
Class II: DNA transposon	1,089,153	292.73	15.89	9.32
TIR				
CMC [DTC]	209,471	91.74	4.98	2.92
hAT	102,389	29.26	1.59	0.93
Mutator	141,536	51.45	2.79	1.64
Tc1/Mariner	192,332	31.20	1.69	0.99
PIF/Harbinger	284,093	60.32	3.27	1.92
Other	33,000	2.4	0.13	0.08
Helitron	86,446	16.26	0.88	0.52
Tandem Repeats	575,983	82.79	4.49	2.64
Unknown	17,632	33.82	1.84	1.08

Supplementary Table 15. Identification of genetic variation comparing to monoploid genome in AP85-441

Haplotype	Variation	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Chr8
A	SNPs	337031	320931	232602	215337	215401	256202	206367	178904
	No. of Indels(1-10bp)	54801	51033	36429	33178	35756	15581	32646	11674
	No. of large Indels(>10bp)	1966	1735	1393	1195	1176	141	1166	85
	Size of Indels(1-10bp)	88367	81843	59583	53562	56202	16488	52659	12349
	No. of Repeat expansion/contraction	135	110	97	79	67	26	57	12
	Size of Repeat expansion/contraction	50145	37593	32856	30599	21713	1909	22655	752
B	SNPs	356807	301635	251918	225795	230974	222584	211161	165709
	No. of Indels(1-10bp)	24755	49793	17839	34846	15025	34725	33294	26367
	No. of large Indels(>10bp)	249	1782	202	1241	149	1138	1138	919
	Size of Indels(1-10bp)	25991	80797	18905	56345	15916	55118	53229	41899
	No. of Repeat expansion/contraction	34	109	13	85	9	60	67	49
	Size of Repeat expansion/contraction	3767	35132	2735	28908	2911	22959	23445	22608
C	SNPs	291154	336697	262036	226124	213312	222447	231938	139934
	No. of Indels(1-10bp)	46632	22403	42622	37334	23906	35751	15312	22245
	No. of large Indels(>10bp)	1714	223	1779	1295	1157	1088	123	688
	Size of Indels(1-10bp)	75615	23652	70245	60389	53426	55700	15091	34735
	No. of Repeat expansion/contraction	147	18	120	79	64	61	19	38
	Size of Repeat expansion/contraction	59782	2576	48656	32512	22168	21457	4675	15821
D	SNPs	352292	286183	196513	243866	194158	231982	211172	179459
	No. of Indels(1-10bp)	56610	43969	30794	17226	31204	35816	35331	29216
	No. of large Indels(>10bp)	2264	1526	1198	135	1098	1174	1126	913
	Size of Indels(1-10bp)	94113	70851	50970	18220	60003	56398	55601	46536
	No. of Repeat expansion/contraction	164	103	92	15	75	54	66	53
	Size of Repeat expansion/contraction	59489	36012	36558	1064	27346	19062	21666	17410

Supplementary Table 16. Investigation of potential collapsed and deleted regions

Chromosome	Region	Number of reads	Read depth
Chr1A	95M-114M	11156370	86.12
Chr1B	97M-123M	14242900	81.33
Chr1D	100M-116M	10255168	91.30
Chr3A	33M-51M	9003293	75.03
Chr3B*	32M-75M	30554803	106.59
Chr3C	42M-71M	16076724	83.16
Chr8A*	50M-67M	13119744	110.79
Chr8B	44M-65M	10633789	75.89
Chr8D	48M-66M	10155852	80.29

*indicate collapsed region

Supplementary Table 19. The genomic rearrangement regions in AP85-441 genome.

AP85-41 Chromosomes	**Corresponding *Sorghum bicolor* Chromosomes**	**Rearranged genomic position (Mb)**
Chr2A	Chr8L	97.607~121.273
Chr2B	Chr8L	93.656~113.725
*Chr2C	Chr8L	98.526~125.935
Chr2D	Chr8L	85.874~108.847
Chr5A	Chr5S	63.652~90.065
Chr5B	Chr5S	59.856~92.060
*Chr5C	Chr5S	57.627~89.054
Chr5D	Chr5S	59.695~84.556
Chr6A	Chr5L	72.406~104.890
Chr6B	Chr5L	59.697~89.339
Chr6C	Chr5L	61.386~90.160
*Chr6D	Chr5L	54.550~90.550
Chr7A	Chr8S	59.392~77.337
Chr7B	Chr8S	63.900~80.424
Chr7C	Chr8S	65.584~85.723
*Chr7D	Chr8S	62.034~83.264

Note: * indicated the rearrangement regions were selected for Fisher's Exact test of R gene distributions.

Supplementary Table 20. Statistics of Fisher's Exact test of R genes in rearranged regions and non-rearranged regions

	In rearranged regions	Not in rearranged regions	Fisher's Exact test
R genes	171	190	P-value < 2.2e-16
Not R genes	3,725	31,800	
R gene alleles	293	335	P-value < 2.2e-16
Not R gene alleles	11,945	97,713	

Supplementary Table 21. Summary of statistical comparisons of genomic diversity (π, Tajima' D and SNPs density) between genomic rearranged regions (RAR) and non-rearranged regions (Non-RAR)

Regions comparisons		π (Ave±S.E.)	Tajima's D (Ave±S.E.)	SNPs density (Ave±S.E.)
Chr2A	RAR	0.00032±0.00003*#	-0.559±0.063*#	414.28±41.64*#
	Non-RAR	0.00023±0.00001	-0.713±0.03	322.64±21.26
Chr2B	RAR	0.00029±0.00004*	-0.702±0.063	422.59±58.38*
	Non-RAR	0.00021±0.00001	-0.721±0.03	297.14±15.73
Chr2C	RAR	0.00029±0.00003*	-0.652±0.048	427.99±58.11*
	Non-RAR	0.00024±0.00001	-0.671±0.029	337.51±19.56
Chr2D	RAR	0.00026±0.00002*#	-0.723±0.055	390.46±34.67*#
	Non-RAR	0.0002±0.00001	-0.74±0.031	274.59±16.76
Chr5A	RAR	0.00025±0.00002	-0.757±0.048	367.08±35.24*#
	Non-RAR	0.00021±0.00002	-0.649±0.038	269.36±22.92
Chr5B	RAR	0.00022±0.00002*	-0.585±0.054*#	290.9±32.43*
	Non-RAR	0.00028±0.00002	-0.809±0.037	455.7±51.61
Chr5C	RAR	0.00021±0.00002*#	-0.634±0.05*#	279.43±26.31*#
	Non-RAR	0.0002±0.00001	-0.749±0.036	294.49±25.51
Chr5D	RAR	0.00027±0.00003*#	-0.672±0.058	392.29±48.16*#
	Non-RAR	0.00018±0.00001	-0.66±0.037	258.44±21.79
Chr6A	RAR	0.00028±0.00003*#	-0.68±0.052	375.47±39.84
	Non-RAR	0.00023±0.00002	-0.688±0.033	310.16±23.92
Chr6B	RAR	0.00026±0.00002*#	-0.644±0.065	389.21±37.95*#
	Non-RAR	0.0002±0.00002	-0.764±0.042	294.3±34.67
Chr6C	RAR	0.00026±0.00003	-0.662±0.057	363.44±39.32
	Non-RAR	0.00024±0.00002	-0.707±0.036	329.92±25.39
Chr6D	RAR	0.00025±0.00002*#	-0.593±0.049*#	334.77±29.11*#
	Non-RAR	0.00017±0.00001	-0.78±0.037	236.24±16.85
Chr7A	RAR	0.00021±0.00002*#	-0.686±0.078	335.69±44.34#
	Non-RAR	0.00019±0.00001	-0.722±0.035	268.07±22.25
Chr7B	RAR	0.0002±0.00002	-0.625±0.086	276.57±33.42
	Non-RAR	0.00019±0.00001	-0.717±0.035	269.34±19.55
Chr7C	RAR	0.00023±0.00003	-0.669±0.062	330.12±43.24*
	Non-RAR	0.0002±0.00001	-0.674±0.043	260.24±18.81
Chr7D	RAR	0.00025±0.00003*#	-0.695±0.067	374.07±59.98*
	Non-RAR	0.0002±0.00002	-0.758±0.035	281.23±23.53
Average	**RAR**	**0.00025±0.00003*#**	**-0.659±0.052*#**	**360.27±48.41*#**
	Non-RAR	**0.00021±0.00001**	**-0.72±0.011**	**297.46±12.65**

Significant *P* values showing in the table by * indicates $p< 0.05$ using T-test; # indicates $p< 0.05$ using Mann–Whitney U test

Supplementary Table 22. Go enrichment of gene models in genomic non-rearranged region (non-RAR)

GO ID	GO Name	GO Category	*FDR*	*P*-Value	Tested number
GO:0009507	chloroplast	C	1.05E-07	1.87E-11	236
GO:0008137	NADH dehydrogenase (ubiquinone) activity	F	8.92E-07	4.77E-10	3
GO:0050136	NADH dehydrogenase (quinone) activity	F	8.92E-07	4.77E-10	3
GO:0016655	oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor	F	1.23E-06	8.74E-10	5
GO:0003954	NADH dehydrogenase activity	F	2.41E-06	2.15E-09	4
GO:0015074	DNA integration	P	5.48E-05	5.86E-08	439
GO:0019843	rRNA binding	F	2.09E-04	2.61E-07	31
GO:0009534	chloroplast thylakoid	C	9.12E-04	1.46E-06	55
GO:0031976	plastid thylakoid	C	9.12E-04	1.46E-06	55
GO:0016020	membrane	C	0.00132785	2.37E-06	2881
GO:0009535	chloroplast thylakoid membrane	C	0.00136061	2.91E-06	49
GO:0055035	plastid thylakoid membrane	C	0.00136061	2.91E-06	49
GO:0031224	intrinsic component of membrane	C	0.005334457	1.24E-05	1788
GO:0048038	quinone binding	F	0.006142299	1.53E-05	5
GO:0044436	thylakoid part	C	0.006219596	1.66E-05	68
GO:0016021	integral component of membrane	C	0.007102412	2.03E-05	1737
GO:0022904	respiratory electron transport chain	P	0.008392109	2.69E-05	6
GO:0042651	thylakoid membrane	C	0.008392109	2.60E-05	59
GO:0009772	photosynthetic electron transport in photosystem II	P	0.009418009	3.36E-05	1
GO:0016651	oxidoreductase activity, acting on NAD(P)H	F	0.009418009	3.29E-05	32
GO:0044425	membrane part	C	0.010548712	3.95E-05	1905
GO:0009579	thylakoid	C	0.012805514	5.25E-05	87
GO:0006259	DNA metabolic process	P	0.012805514	5.05E-05	554
GO:0031984	organelle subcompartment	C	0.014153984	6.06E-05	89
GO:0034357	photosynthetic membrane	C	0.0176916	7.88E-05	66
GO:0015986	ATP synthesis coupled proton transport	P	0.02090508	1.01E-04	4
GO:0015985	energy coupled proton transport, down electrochemical gradient	P	2.09E-02	1.01E-04	4
GO:0022414	reproductive process	P	2.46E-02	1.27E-04	135
GO:0000003	reproduction	P	0.024555639	1.27E-04	135
GO:0044702	single organism reproductive process	P	0.025944473	1.39E-04	122
GO:0043228	non-membrane-bounded organelle	C	0.026166439	1.49E-04	394
GO:0043232	intracellular non-membrane-bounded organelle	C	0.026166439	1.49E-04	394

Supplementary Table 23. *Saccharum spontaneum* accessions (with ploidy and sampling locations information) used for resequencing and population genomics analysis

Accessions	Ploidy	Location
AP85-68	8	
s15-95	8	
FJ-89-1-1	11	Fujian, China
FU-89-1-16	11	Fujian, China
Gugu	8	Kenya
GZ-78-1-11	9	Guizhou, China
HN-2	10	Hainan, China
Holes1	10	Coimbatore, India
IK76-067	10	Panajam, Borneo, Indonesia
IN76-086	10	
IN84-021	10	Raha, Muna Regency, South East Sulawesi, indonesia
IN84-089	6	Salodik, Banggai Regency, Central Sulawesi, Indonesia
IND81-03	8	
IND81-05	8	
IND81-08	10	
IND81-13	10	
IND81-14	8	
IND81-15	8	
IND81-17	10	
IND81-18	10	
Iranspon	14	Iran
PCANOR84	8	Philippines
PPGN84-0	6	
PTAR84-0	9	
S-spont-jing	8	
S-spontI	10	
SaudiAra	12	Coimbatore, India
SC-79-2-11	11	Sichuan, China
SES004A	10	India
SES014	12	India
SES072	6	India
SES113A	6	
SES184B	15	India
SES186	7	

SES196	8	India
SES197A	8	India
SES208	8	India
SES234	8	Malaysia
SES239/43	9	
SES264	8	India
SES275	6	India
SES289	10	
SES294	8	India
SES297B	10	India
SES341	6	India
SES365	8	India
SES517	8	India
SES519	10	India
SES561	16	
SES602	10	
Shoaguan	8	Sichuan, China
SLC92-81	8	
SLC92-94	8	
SM7916	6	
Taiwansp	12	Taiwan, China
Tongza	8	China
US48-61	13	
US56-14-4	10	
US60-004	6	
US78-500	8	Khyber-Pakhtunkhwa, Pakistan
Yacheng-jing	10	China
YN-76-1-20	14	Yunnan,China
YNMZ	10	Yunnan,China
YNXD	11	Yunnan,China

Supplementary Table 24. Go enrichment of gene models in genomic rearranged region (RAR)

GO ID	GO Name	GO Category	*FDR*	*P*-Value	Tested number
GO:0009987	cellular process	P	6.14E-14	1.09E-17	393
GO:0016043	cellular component organization	P	4.99E-12	2.52E-15	22
GO:0071840	cellular component organization or biogenesis	P	4.99E-12	2.67E-15	30
GO:0044464	cell part	C	2.26E-11	1.61E-14	412
GO:0044237	cellular metabolic process	P	2.27E-11	2.31E-14	322
GO:0005623	cell	C	2.27E-11	2.42E-14	414
GO:0044260	cellular macromolecule metabolic process	P	9.69E-11	1.21E-13	203
GO:0005622	intracellular	C	4.63E-10	6.60E-13	362
GO:0043170	macromolecule metabolic process	P	8.74E-10	1.40E-12	233
GO:0044238	primary metabolic process	P	1.39E-09	2.48E-12	314
GO:0032991	macromolecular complex	C	1.54E-09	3.02E-12	50
GO:0044424	intracellular part	C	2.41E-09	5.15E-12	356
GO:0044422	organelle part	C	2.41E-09	5.58E-12	60
GO:0044446	intracellular organelle part	C	2.88E-09	7.20E-12	60
GO:0008152	metabolic process	P	1.42E-08	3.81E-11	452
GO:0003676	nucleic acid binding	F	2.16E-08	6.16E-11	92
GO:0097159	organic cyclic compound binding	F	2.40E-08	7.71E-11	237
GO:1901363	heterocyclic compound binding	F	2.40E-08	7.68E-11	237
GO:0071704	organic substance metabolic process	P	5.34E-08	1.81E-10	356
GO:0043226	organelle	C	4.12E-07	1.47E-09	314
GO:0043229	intracellular organelle	C	5.23E-07	1.96E-09	314
GO:0010467	gene expression	P	6.61E-07	2.59E-09	64
GO:0034641	cellular nitrogen compound metabolic process	P	7.34E-07	3.01E-09	146
GO:0005488	binding	F	7.86E-07	3.36E-09	397
GO:0006807	nitrogen compound metabolic process	P	9.13E-07	4.07E-09	162
GO:0043228	non-membrane-bounded organelle	C	9.10E-06	4.38E-08	25
GO:0043232	intracellular non-membrane-bounded organelle	C	9.10E-06	4.38E-08	25
GO:0009059	macromolecule biosynthetic process	P	1.05E-05	5.25E-08	65
GO:0043227	membrane-bounded organelle	C	1.23E-05	6.52E-08	300
GO:0043043	peptide biosynthetic process	P	1.23E-05	6.59E-08	14
GO:0034645	cellular macromolecule biosynthetic process	P	1.40E-05	7.75E-08	63
GO:0006412	translation	P	1.59E-05	9.06E-08	14
GO:0043231	intracellular membrane-bounded organelle	C	1.97E-05	1.19E-07	298
GO:0006996	organelle organization	P	1.97E-05	1.20E-07	14
GO:0044271	cellular nitrogen compound biosynthetic process	P	2.32E-05	1.45E-07	68
GO:0006518	peptide metabolic process	P	2.44E-05	1.56E-07	18
GO:1905039	carboxylic acid transmembrane transport	P	5.13E-05	3.38E-07	23
GO:0044085	cellular component biogenesis	P	5.24E-05	3.73E-07	13
GO:0005840	ribosome	C	5.24E-05	3.74E-07	8
GO:0044249	cellular biosynthetic process	P	5.24E-05	3.72E-07	105
GO:0043604	amide biosynthetic process	P	6.07E-05	4.65E-07	17
GO:1990904	ribonucleoprotein complex	C	6.07E-05	4.64E-07	18
GO:0030529	intracellular ribonucleoprotein complex	C	6.07E-05	4.64E-07	18
GO:1903825	organic acid transmembrane transport	P	7.16E-05	5.61E-07	23
GO:0005634	nucleus	C	8.16E-05	6.54E-07	65
GO:0005737	cytoplasm	C	9.94E-05	8.32E-07	294
GO:0046942	carboxylic acid transport	P	9.94E-05	8.30E-07	23

GO:0046943	carboxylic acid transmembrane transporter activity	F	1.07E-04	9.13E-07	23
GO:0044267	cellular protein metabolic process	P	1.12E-04	9.82E-07	113
GO:0009579	thylakoid	C	1.38E-04	1.24E-06	1
GO:0003723	RNA binding	F	1.38E-04	1.25E-06	14
GO:0015849	organic acid transport	P	1.43E-04	1.33E-06	23
GO:0005342	organic acid transmembrane transporter activity	F	1.54E-04	1.46E-06	23
GO:0019538	protein metabolic process	P	1.64E-04	1.58E-06	133
GO:0009098	leucine biosynthetic process	P	1.69E-04	1.69E-06	5
GO:0003852	2-isopropylmalate synthase activity	F	1.69E-04	1.69E-06	5
GO:0043603	cellular amide metabolic process	P	1.74E-04	1.80E-06	23
GO:0003333	amino acid transmembrane transport	P	1.74E-04	1.82E-06	21
GO:0044391	ribosomal subunit	C	1.74E-04	1.83E-06	4
GO:0044428	nuclear part	C	2.90E-04	3.10E-06	10
GO:0022607	cellular component assembly	P	2.92E-04	3.17E-06	6
GO:0008514	organic anion transmembrane transporter activity	F	2.92E-04	3.23E-06	23
GO:0006865	amino acid transport	P	3.17E-04	3.56E-06	21
GO:0009534	chloroplast thylakoid	C	3.72E-04	4.38E-06	0
GO:0015171	amino acid transmembrane transporter activity	F	3.72E-04	4.29E-06	21
GO:0031976	plastid thylakoid	C	3.72E-04	4.38E-06	0
GO:0016746	transferase activity, transferring acyl groups	F	5.72E-04	6.86E-06	46
GO:0043234	protein complex	C	5.72E-04	6.94E-06	21
GO:0090304	nucleic acid metabolic process	P	6.57E-04	8.08E-06	102
GO:0006551	leucine metabolic process	P	7.60E-04	9.48E-06	5
GO:0071555	cell wall organization	P	7.87E-04	1.01E-05	3
GO:0003735	structural constituent of ribosome	F	7.87E-04	1.00E-05	10
GO:0045229	external encapsulating structure organization	P	8.21E-04	1.07E-05	3
GO:0005198	structural molecule activity	F	8.59E-04	1.13E-05	12
GO:0046483	heterocycle metabolic process	P	8.74E-04	1.20E-05	130
GO:0044763	single-organism cellular process	P	8.74E-04	1.19E-05	124
GO:0043933	macromolecular complex subunit organization	P	8.74E-04	1.19E-05	6
GO:0044444	cytoplasmic part	C	8.84E-04	1.28E-05	252
GO:0070013	intracellular organelle lumen	C	8.84E-04	1.27E-05	9
GO:0043233	organelle lumen	C	8.84E-04	1.27E-05	9
GO:0031974	membrane-enclosed lumen	C	8.84E-04	1.27E-05	9
GO:0009535	chloroplast thylakoid membrane	C	9.39E-04	1.39E-05	0
GO:0055035	plastid thylakoid membrane	C	9.39E-04	1.39E-05	0
GO:0044436	thylakoid part	C	0.001094846	1.64E-05	1
GO:0098656	anion transmembrane transport	P	0.00178723	2.71E-05	24
GO:0004751	ribose-5-phosphate isomerase activity	F	0.002018125	3.13E-05	4
GO:0006139	nucleobase-containing compound metabolic process	P	0.002018125	3.10E-05	125
GO:0034357	photosynthetic membrane	C	0.00218854	3.43E-05	1
GO:0009058	biosynthetic process	P	0.002314685	3.67E-05	144
GO:0015711	organic anion transport	P	0.00301356	4.83E-05	23
GO:0006725	cellular aromatic compound metabolic process	P	0.003182796	5.22E-05	139
GO:1901566	organonitrogen compound biosynthetic process	P	0.003182796	5.20E-05	38
GO:0065003	macromolecular complex assembly	P	0.003358563	5.57E-05	6
GO:0016210	naringenin-chalcone synthase activity	F	0.003365754	5.64E-05	8
GO:0042651	thylakoid membrane	C	0.004115238	6.97E-05	1
GO:0009052	pentose-phosphate shunt, non-oxidative branch	P	0.004136114	7.08E-05	4
GO:0008509	anion transmembrane transporter activity	F	0.004971615	8.60E-05	23
GO:0055114	oxidation-reduction process	P	0.005764028	1.01E-04	51
GO:0048037	cofactor binding	F	0.007149533	1.26E-04	8

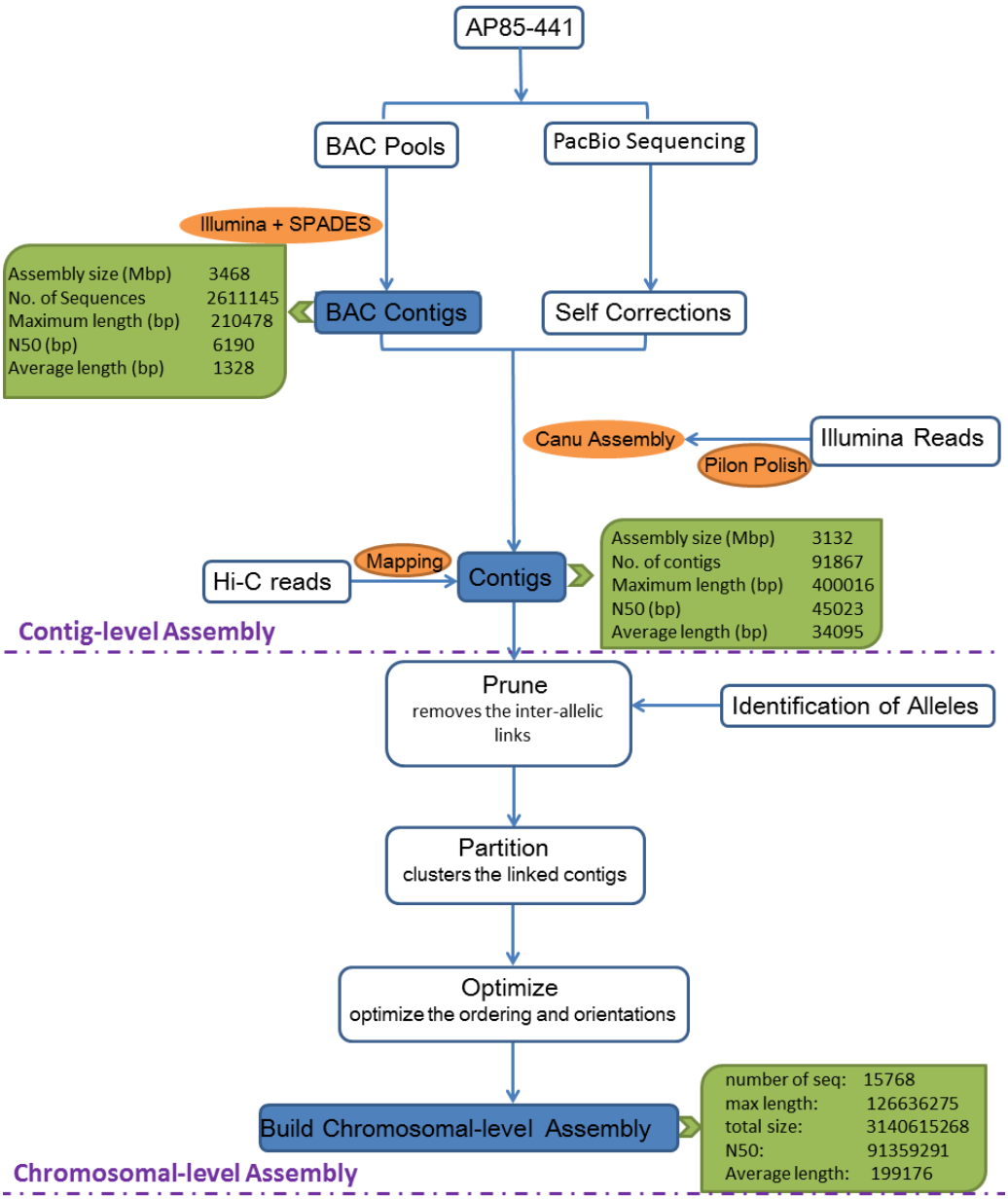
GO:0044425	membrane part	C	0.007384461	1.32E-04	186
GO:0034622	cellular macromolecular complex assembly	P	0.007832184	1.41E-04	6
GO:0006366	transcription from RNA polymerase II promoter	P	0.007858617	1.43E-04	0
GO:0009082	branched-chain amino acid biosynthetic process	P	0.007870507	1.45E-04	6
GO:0031981	nuclear lumen	C	0.008398219	1.56E-04	7
GO:0048519	negative regulation of biological process	P	0.008462689	1.59E-04	3
GO:0015934	large ribosomal subunit	C	0.008462689	1.60E-04	2
GO:0044710	single-organism metabolic process	P	0.008580497	1.64E-04	147
GO:0030312	external encapsulating structure	C	0.009208174	1.79E-04	5
GO:0005618	cell wall	C	0.009208174	1.79E-04	5
GO:1901360	organic cyclic compound metabolic process	P	0.00955625	1.87E-04	147
GO:1901564	organonitrogen compound metabolic process	P	0.00981607	1.94E-04	63
GO:0043140	ATP-dependent 3'-5' DNA helicase activity	F	0.010720076	2.24E-04	3
GO:0035596	methylthiotransferase activity	F	0.010720076	2.24E-04	3
GO:0050497	transferase activity, transferring alkylthio groups	F	0.010720076	2.24E-04	3
GO:0035600	tRNA methylthiolation	P	0.010720076	2.24E-04	3
GO:0009378	four-way junction helicase activity	F	0.010720076	2.24E-04	3
GO:0006091	generation of precursor metabolites and energy	P	0.010720076	2.16E-04	7
GO:0009081	branched-chain amino acid metabolic process	P	0.011596831	2.44E-04	6
GO:0016070	RNA metabolic process	P	0.014023564	2.97E-04	57
GO:0071669	plant-type cell wall organization or biogenesis	P	0.014378943	3.11E-04	0
GO:0003677	DNA binding	F	0.014378943	3.11E-04	41
GO:0031224	intrinsic component of membrane	C	0.014378943	3.13E-04	174
GO:0009067	aspartate family amino acid biosynthetic process	P	0.016593332	3.64E-04	9
GO:0042623	ATPase activity, coupled	F	0.017370429	3.84E-04	4
GO:0044434	chloroplast part	C	0.017497603	3.93E-04	7
GO:0046912	transferase activity, transferring acyl groups, acyl groups converted into alkyl on transfer	F	0.017497603	3.90E-04	6
GO:0071554	cell wall organization or biogenesis	P	0.018870628	4.31E-04	10
GO:0071944	cell periphery	C	0.018870628	4.28E-04	55
GO:0016021	integral component of membrane	C	0.01951423	4.49E-04	170
GO:1901576	organic substance biosynthetic process	P	0.02153316	4.99E-04	138
GO:0051276	chromosome organization	P	0.022260585	5.20E-04	4
GO:0036094	small molecule binding	F	0.024651721	5.80E-04	124
GO:0044427	chromosomal part	C	0.028266368	6.70E-04	1
GO:0044699	single-organism process	P	0.033825903	8.08E-04	270
GO:0044445	cytosolic part	C	0.03405238	8.19E-04	5
GO:0016020	membrane	C	0.034283305	8.35E-04	308
GO:0009507	chloroplast	C	0.034283305	8.37E-04	25
GO:0043167	ion binding	F	0.036816585	9.06E-04	222
GO:0016829	lyase activity	F	0.039861565	9.88E-04	8
GO:0003824	catalytic activity	F	4.79E-02	1.19E-03	446

Supplementary Table 25. Assessment of contigs with high confidence orientation in AP85 Hi-C scaffolding at different contig length cutoff.

Length cutoff (kb)	# of contigs	# of contig orientation with at least 95% of posterior probability	% of contig orientation with high confidence
0	76,131	48,333	63.5%
50	17,254	13,609	78.8%
100	2,979	2,542	85.3%

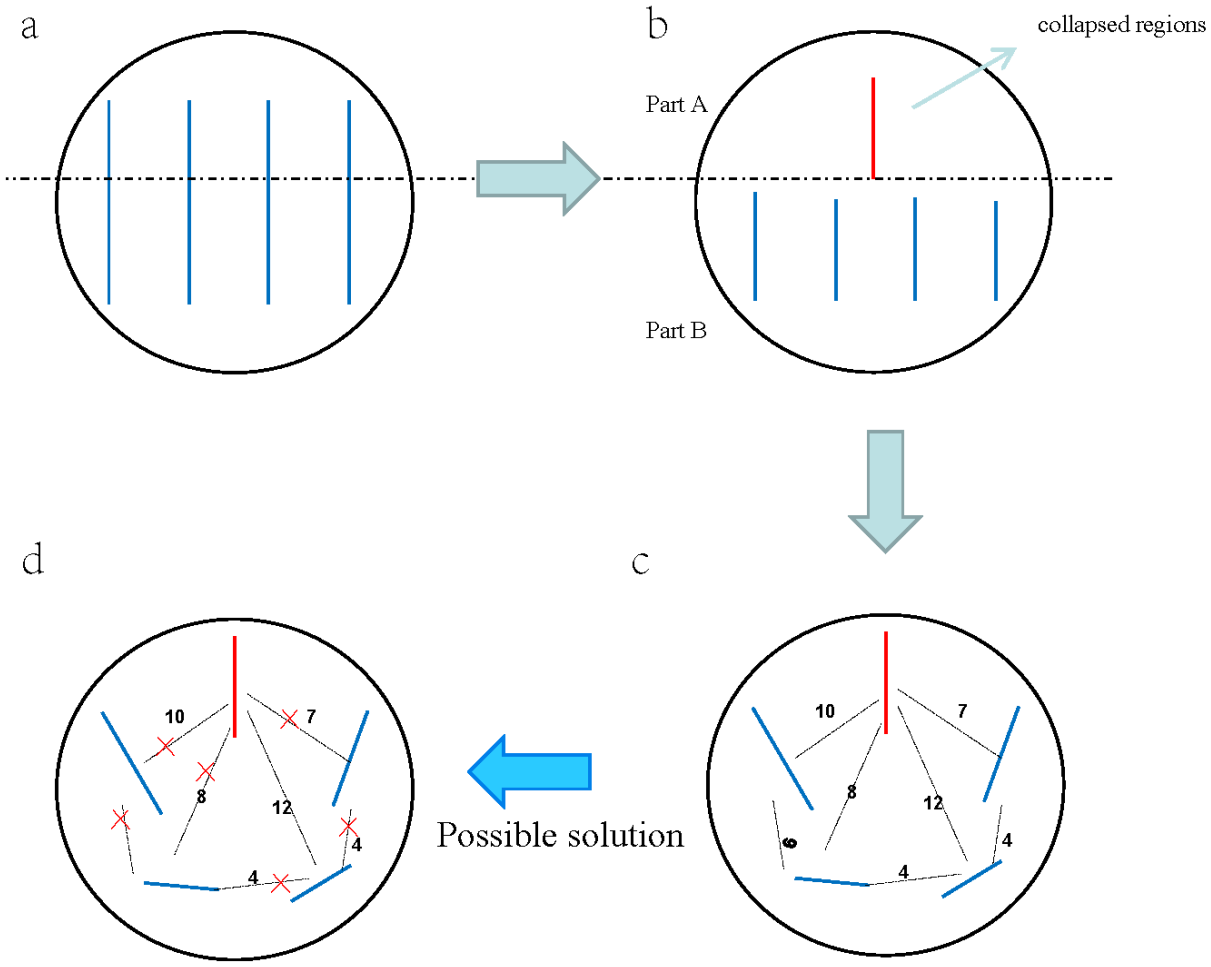
Supplementary Table 26. Coverage of assembled AP85 draft genome at different posterior probability cutoff.

Posterior Probability cutoff	# of contigs	Total length of contigs (bp)	% of genome	Average length (bp)
≥99%	46,970	2,018,831,091	64.4%	42,981
≥95%	48,333	2,051,790,506	65.5%	42,451
≥90%	49,099	2,069,452,786	66.1%	42,149
All	76,131	2,892,630,957	92.3%	37,995



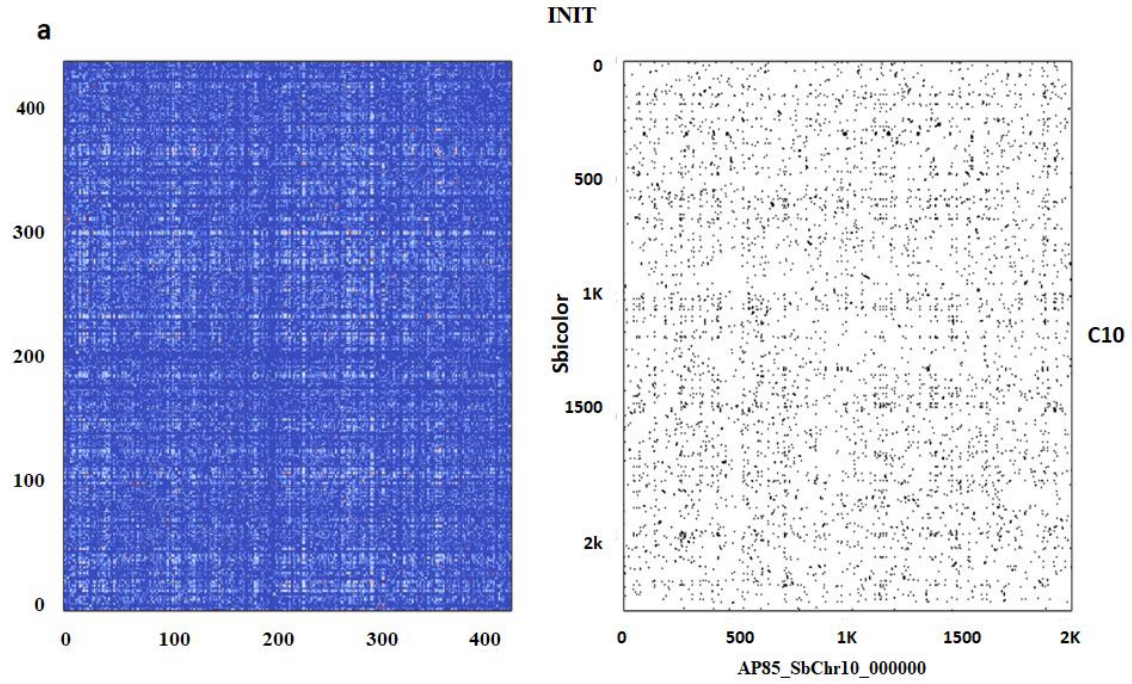
Supplementary Figure 1. Genome assembly strategy in AP85-441.

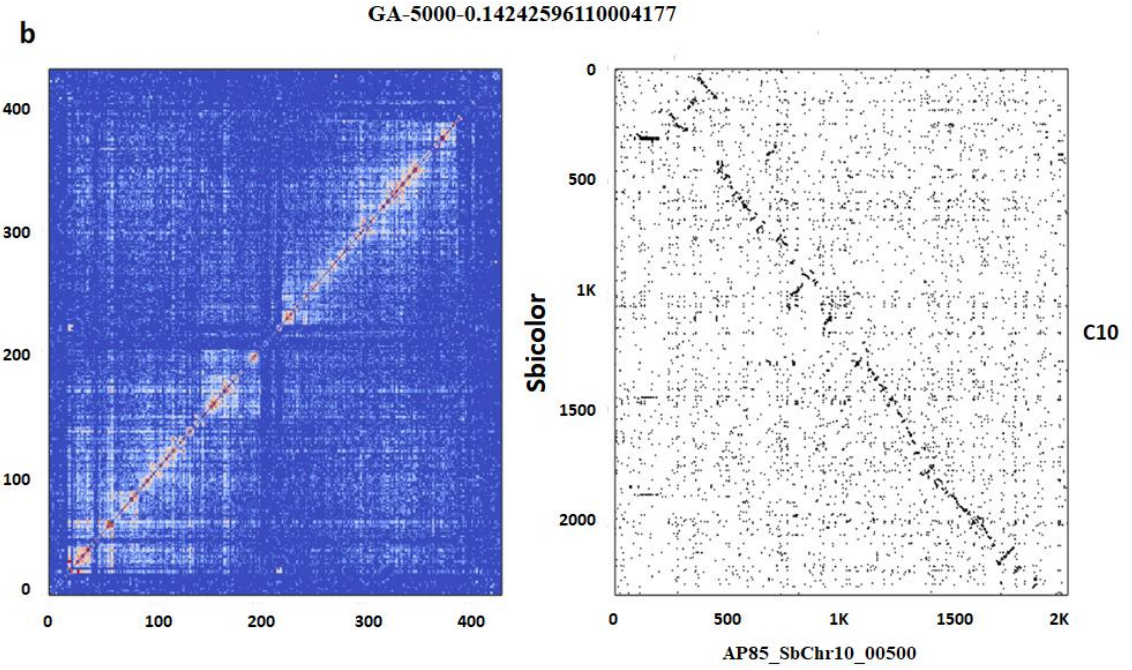
Contigs are assembled by integrating assembled BAC contigs and PacBio RSII long reads. Hi-C-based scaffolding is generated by our newly developed program, ALLHiC, including prune, partition, optimize and build steps (See methods for details).

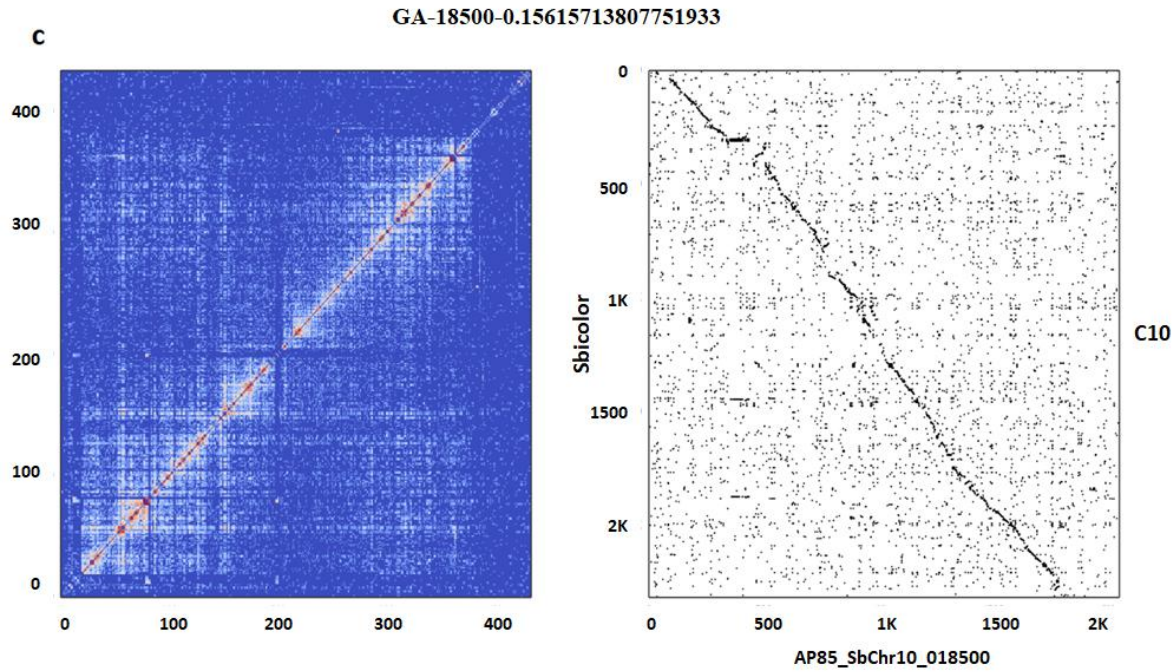


Supplementary Figure 2. Description of Hi-C scaffolding problem in polyploidy genome and application of prune approach to remove inter-haplotype links.

(a) Each line repreasents one haplotype. Totally four haplotypes in AP85 cell. (b) Some of the regions (Part A) are collapsed in assembly due to their high similarity. While, some other regions (Part B) will not be collapsed in assembly as they have high-level variations. Part A and B are two extreme cases. Red line indicates collapsed region and blue lines are separated contigs in assembly. (c) In HiC cluster step, the collapsed region will be detect to have signals with all four haplotypes and then cluster the allelic contigs together in to one group. Dash lines mean linkage signals and the numbers in dash lines are signal density. Larger number represent stronger linkage relationship. (d) Solution: Remove the signals which should not have in the linkage step. 1) remove signals between allelic regions. 2) Only retain one haplotype who has the strongest signal with collapsed contigs.

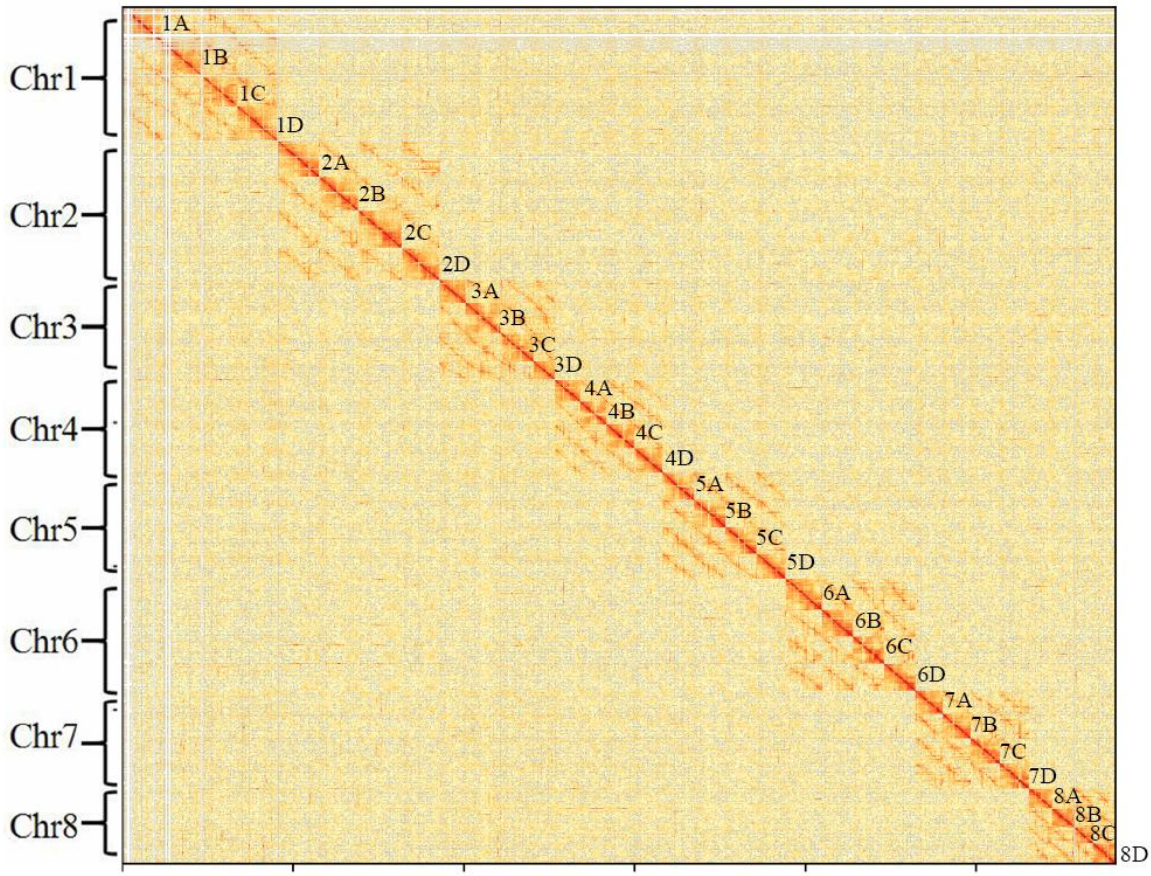

a
INIT
400
300
200
100
0
0
100
200
300
400
Sbicolor
0
500
1K
1500
2k
0
500
1K
1500
2K
C10
AP85_SbChr10_000000


GA-5000-0.14242596110004177
b
400
300
200
100
0
0
100
200
300
400
Sbicolor
0
500
1K
1500
2000
0
500
1K
1500
2K
C10
AP85_SbChr10_00500

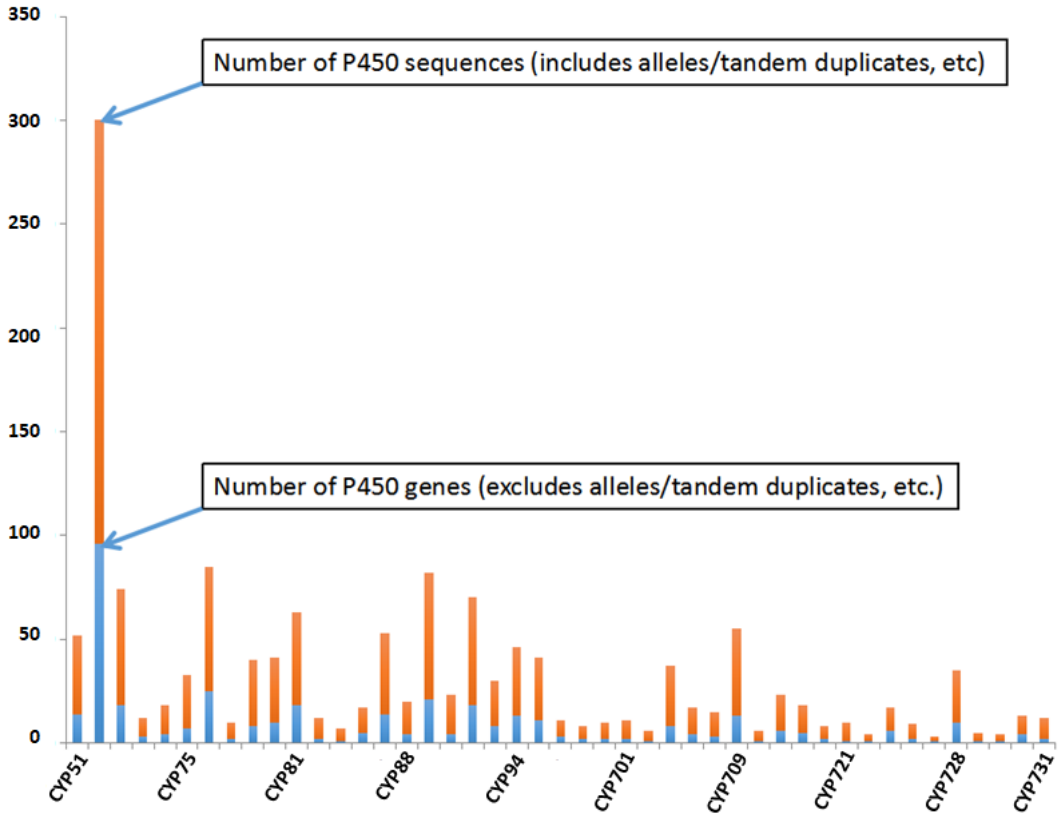


Supplementary Figure 3. Optimize step uses Genetic Algorithm (GA) to order and orient contigs within a partition.

To illustrate the GA process, we show the HiC contact heatmap (left) and genomic dot plot to sorghum Chromosome 10 to show synteny. (a) GA iteration 0 (b) GA iteration 5000 (c) GA iteration 18500. We can see that over the course of the GA evolution, the HiC contacts were increasingly becoming clustered around the 'diagonal', while the synteny to the related genome Sorghum was incrementally improved.

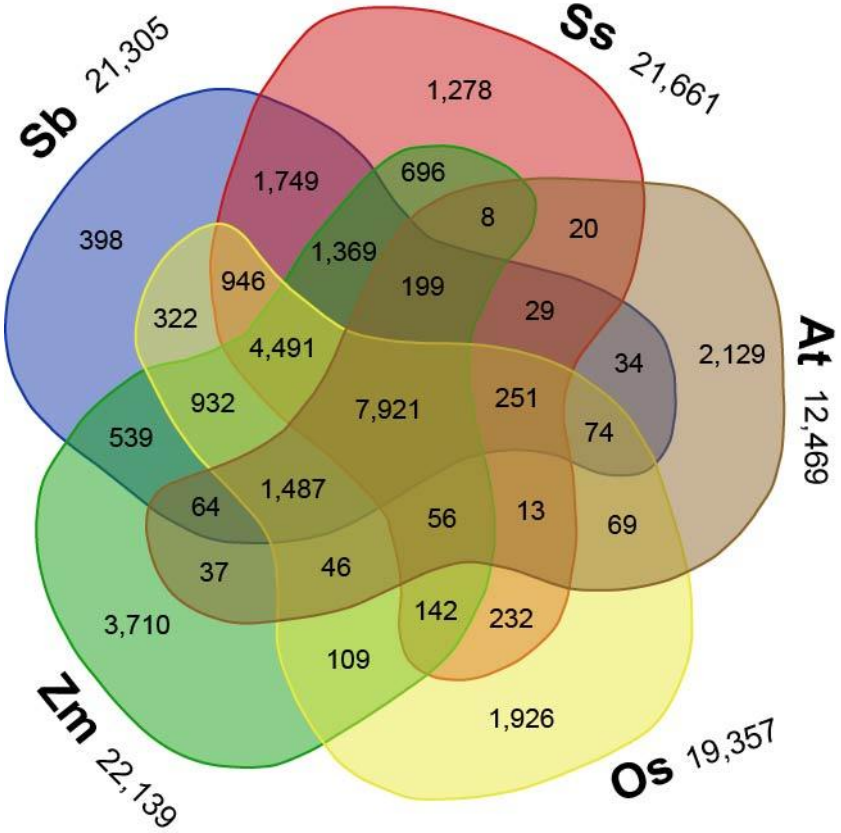


Supplementary Figure 4. Genome-wide analysis of chromatin interactions at 150-kb resolution in AP85-441 genome.



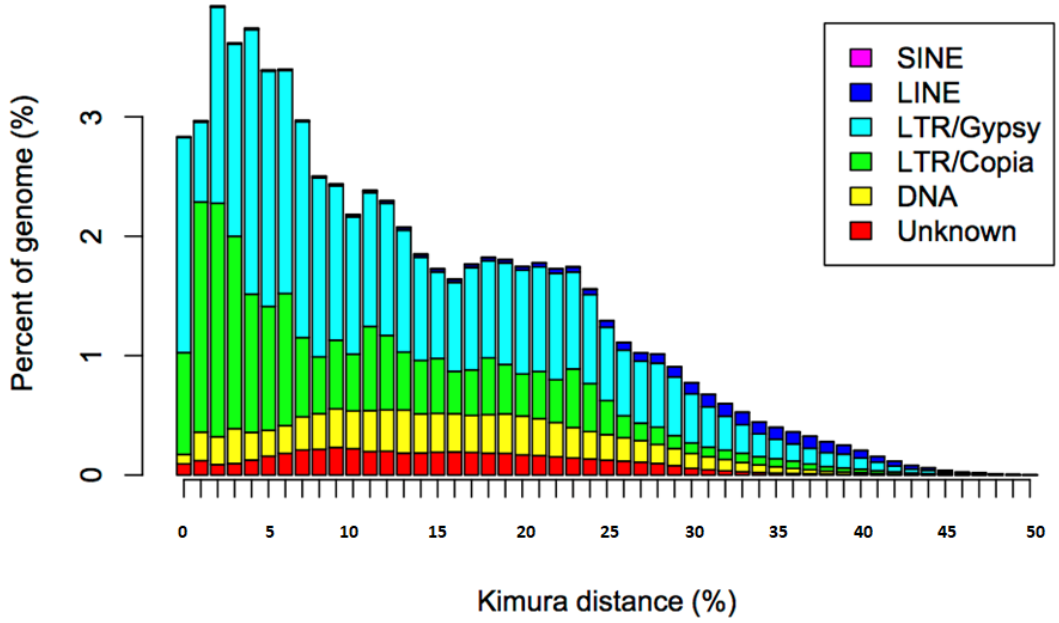
Supplementary Figure 5.The number of cytochrome P450s by family.

Red is the total number of P450s. Blue is the number of unique genes not counting alleles, tandem duplicates and other very close duplicates.



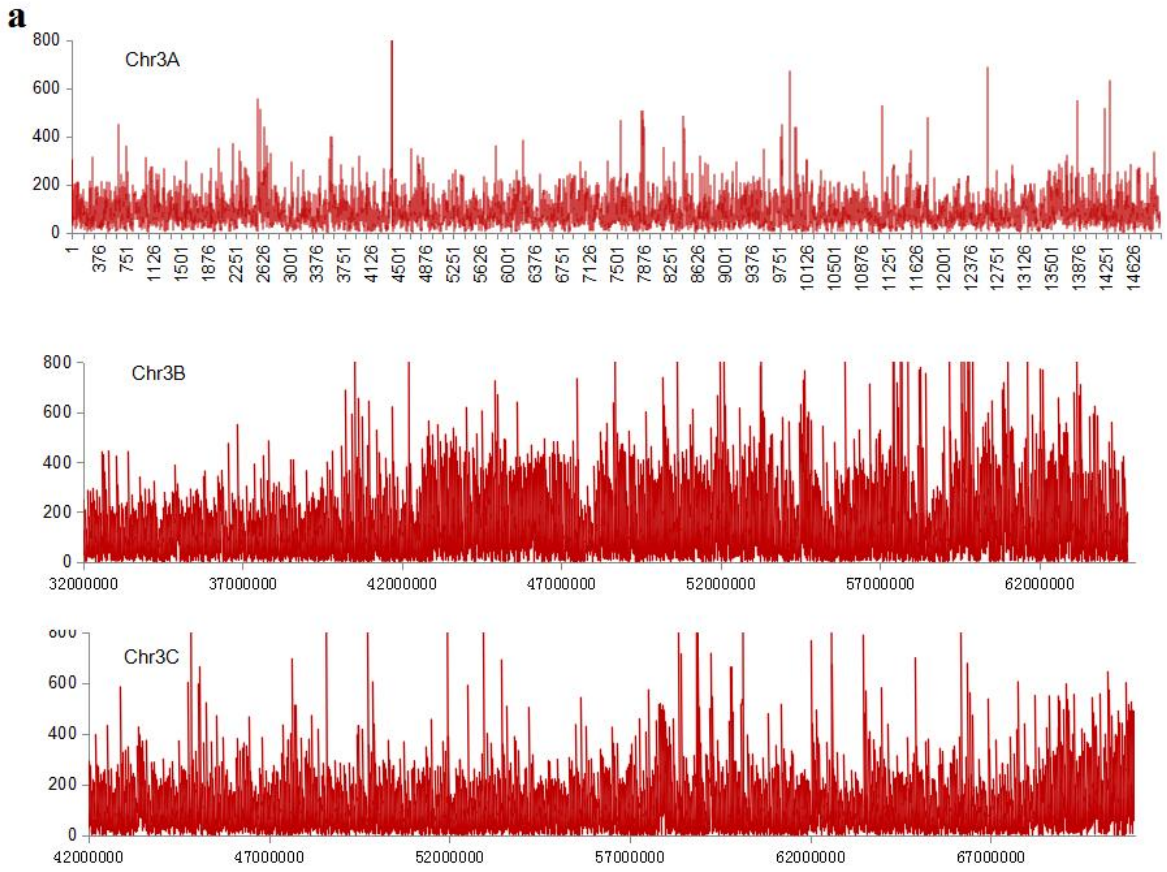
upplementary Figure 6. Orthologous gene families among *S. spontaneum*, *Sorghum*, *Arabidopsis*, rice and maize.

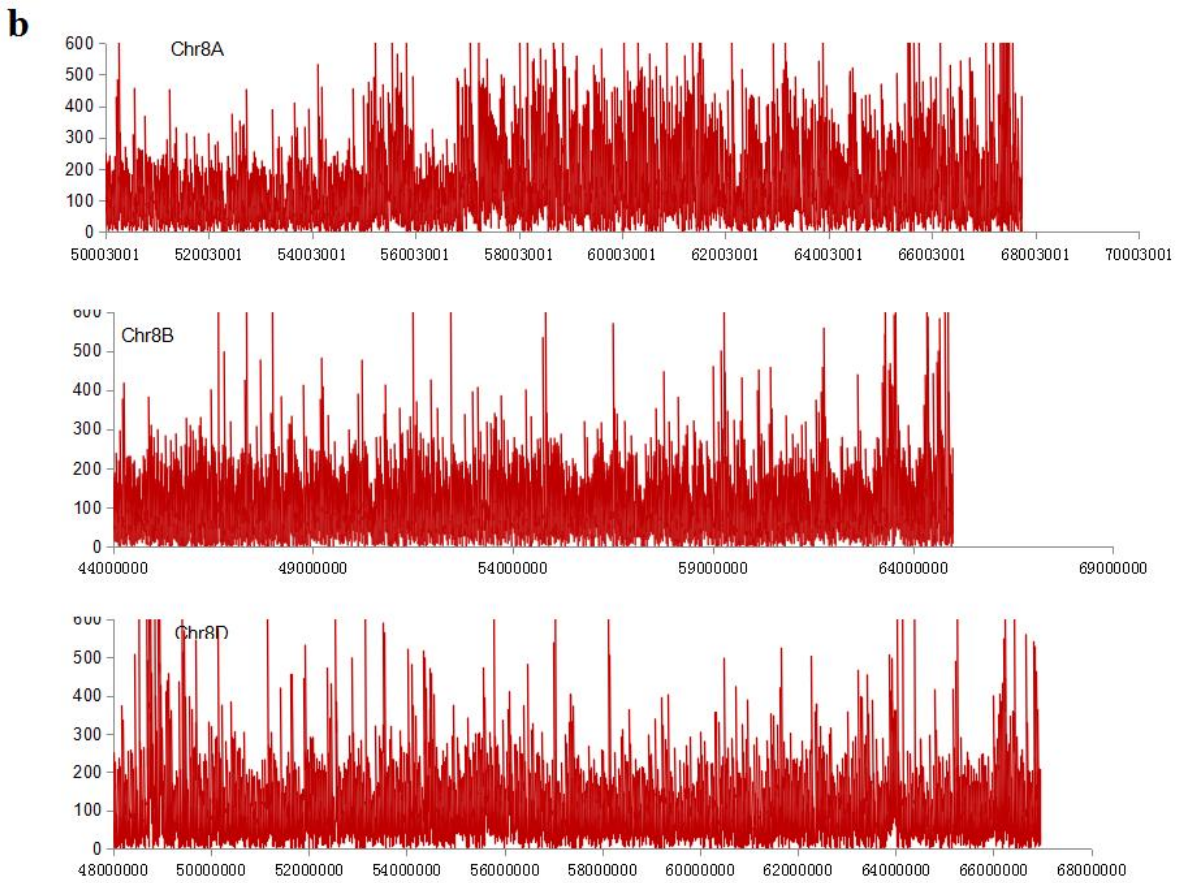
The numbers of gene families (clusters) are indicated for each species and species intersection.



Supplementary Figure 7. Kimura distance-based copy divergence analysis of transposable elements in AP85-441 genome.

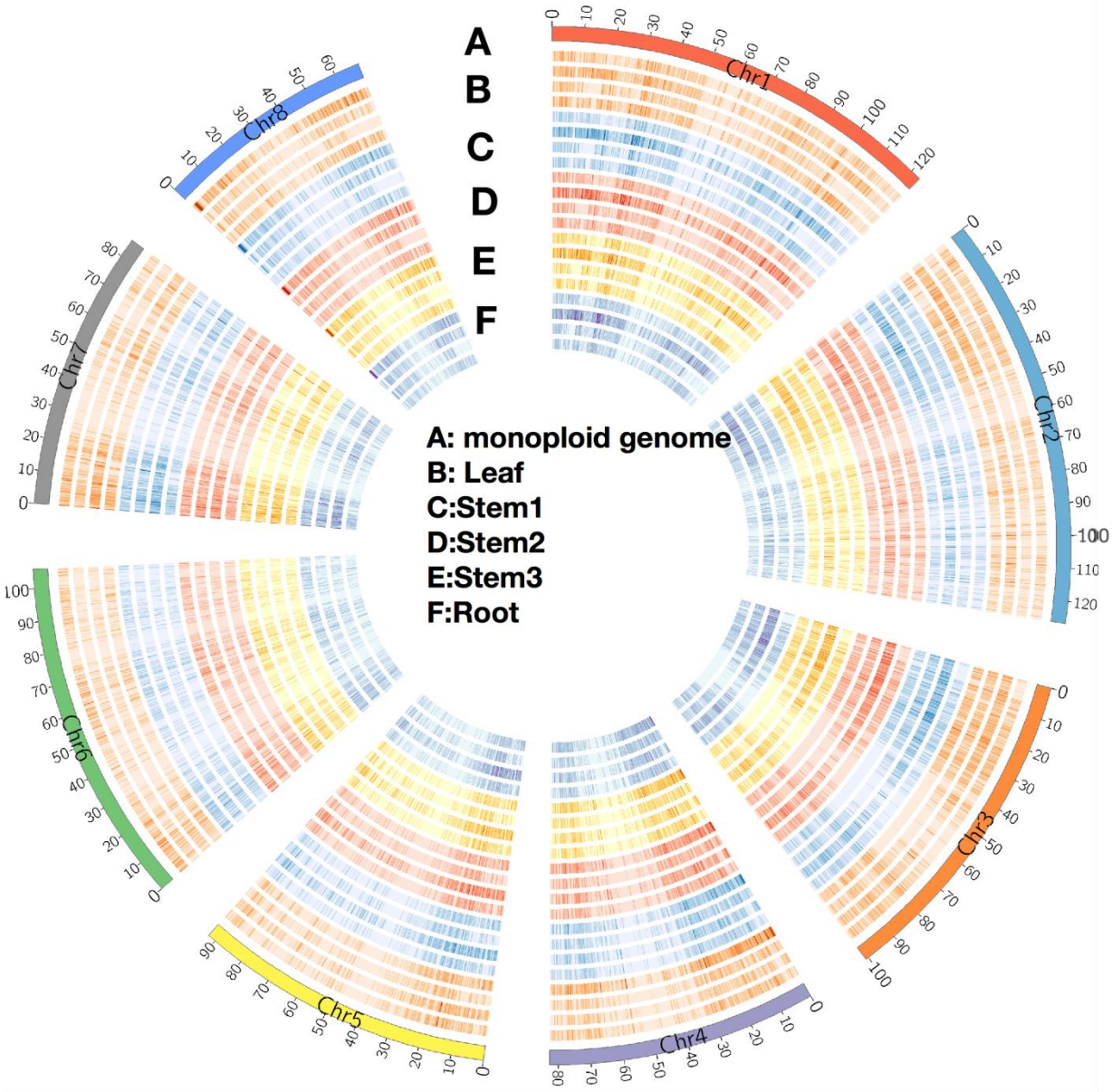
The graph represents percentage of genome (y-axis) for each type of TEs (SINE, LINE, LTR/Gypsy, LTR/Copia and DNAtransposons), clustered according to Kimura distances to their corresponding consensus sequences (x-axis, *K*-value from 0 to 50).


a
Chr3A
800
600
400
200
0
1
376
751
1126
1501
1876
2251
2626
3001
3376
3751
4126
4501
4876
5251
5626
6001
6376
6751
7126
7501
7876
8251
8626
9001
9376
9751
10126
10501
10876
11251
11626
12001
12376
12751
13126
13501
13876
14251
14626
Chr3B
800
600
400
200
0
32000000
37000000
42000000
47000000
52000000
57000000
62000000
Chr3C
600
400
200
0
42000000
47000000
52000000
57000000
62000000
67000000



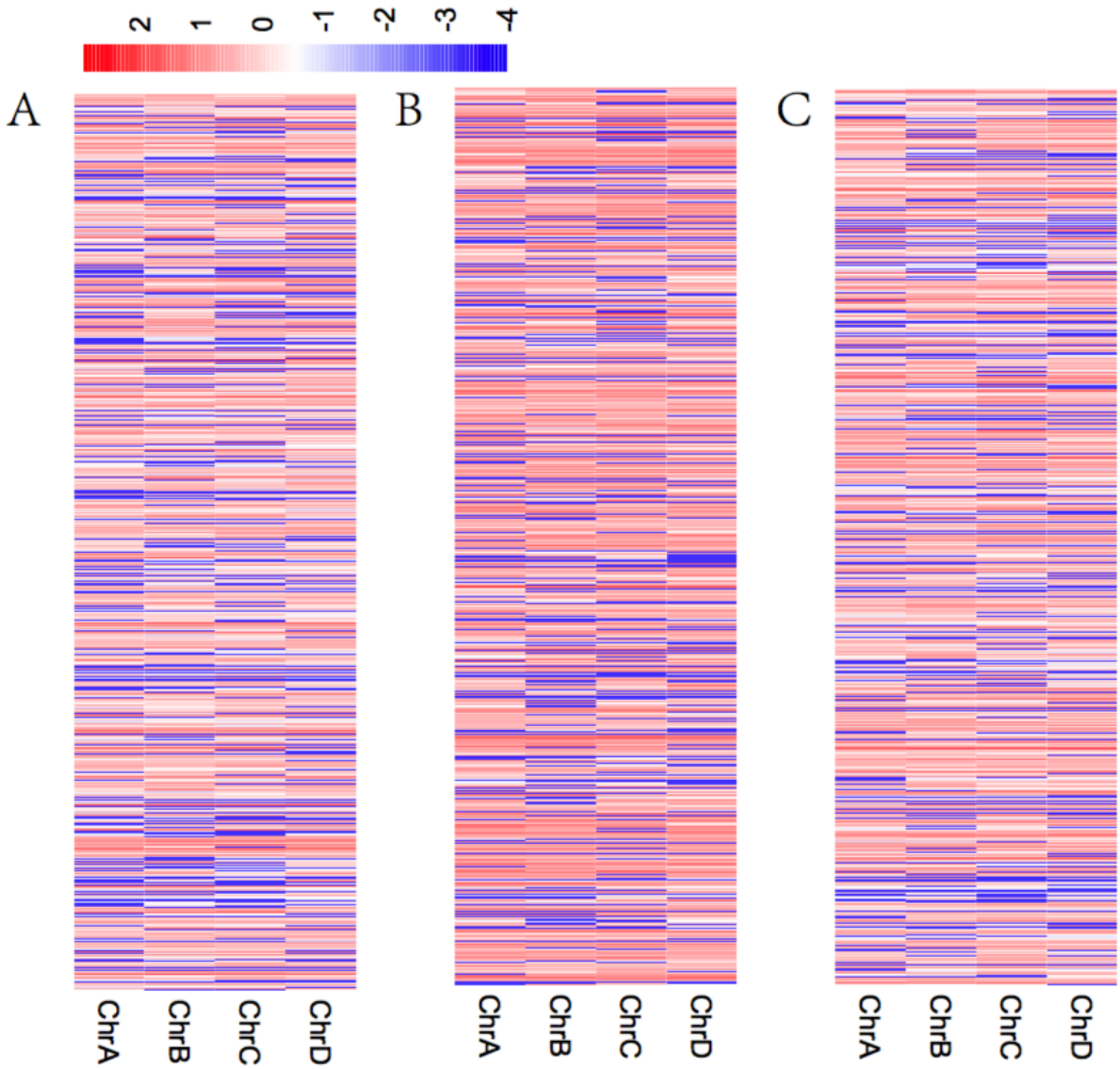
Supplementary Figure 8. Reads depth in collapsed region.

Chr3B (a) and Chr8A (b) have about greater depth of Illumina short reads, suggesting that they are the collapsed homologs.



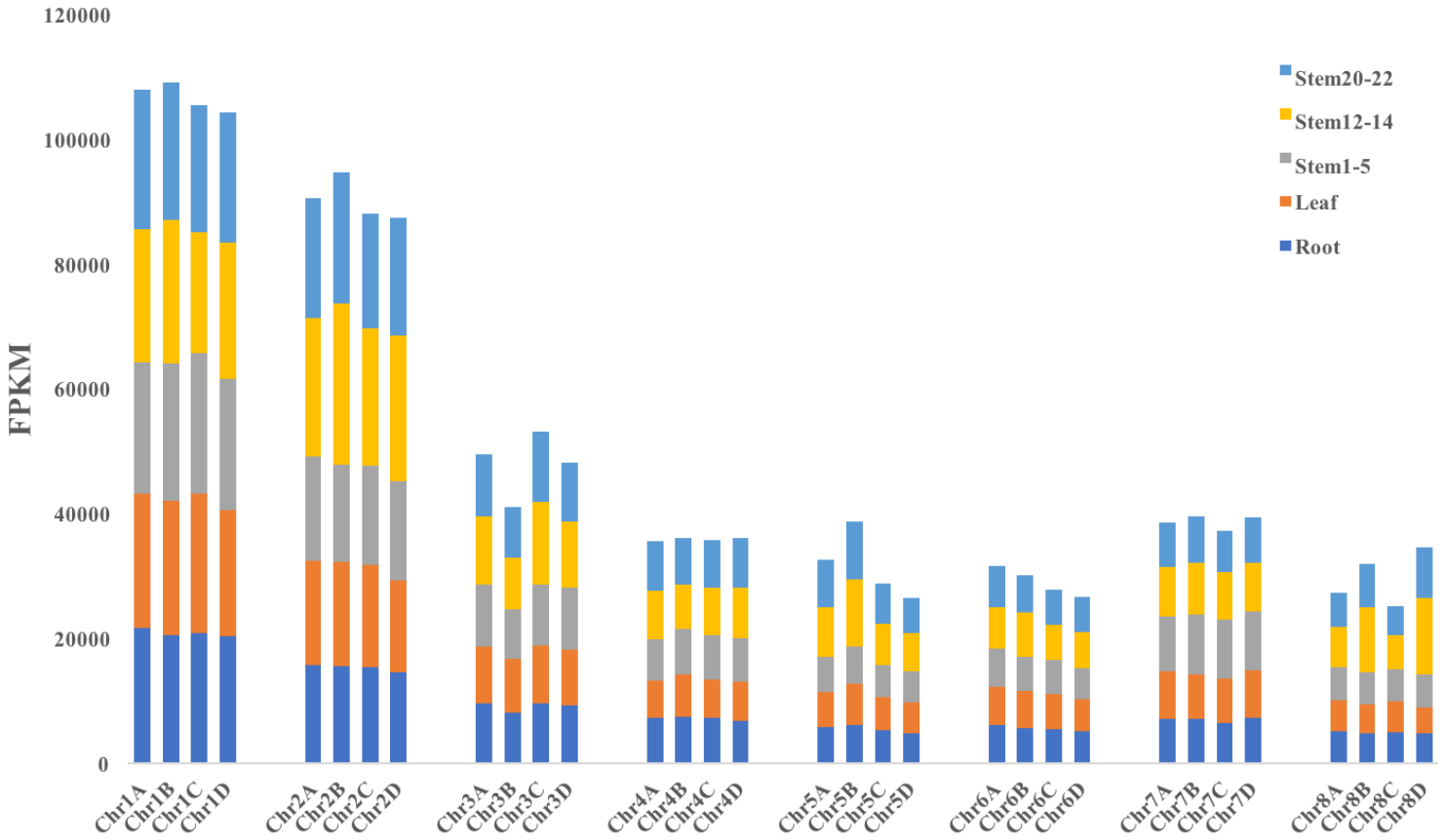
Supplementary Figure 9. The gene expression of five tissues in *S. spontaneum*.

Each color line from outer to inner representing haplotype A, B, C, D respectively. Notes: Stem-1: top internode (internode number 3), Stem-2: maturing internode (internode number 6) , Stem-3: mature internode (internode number 13).

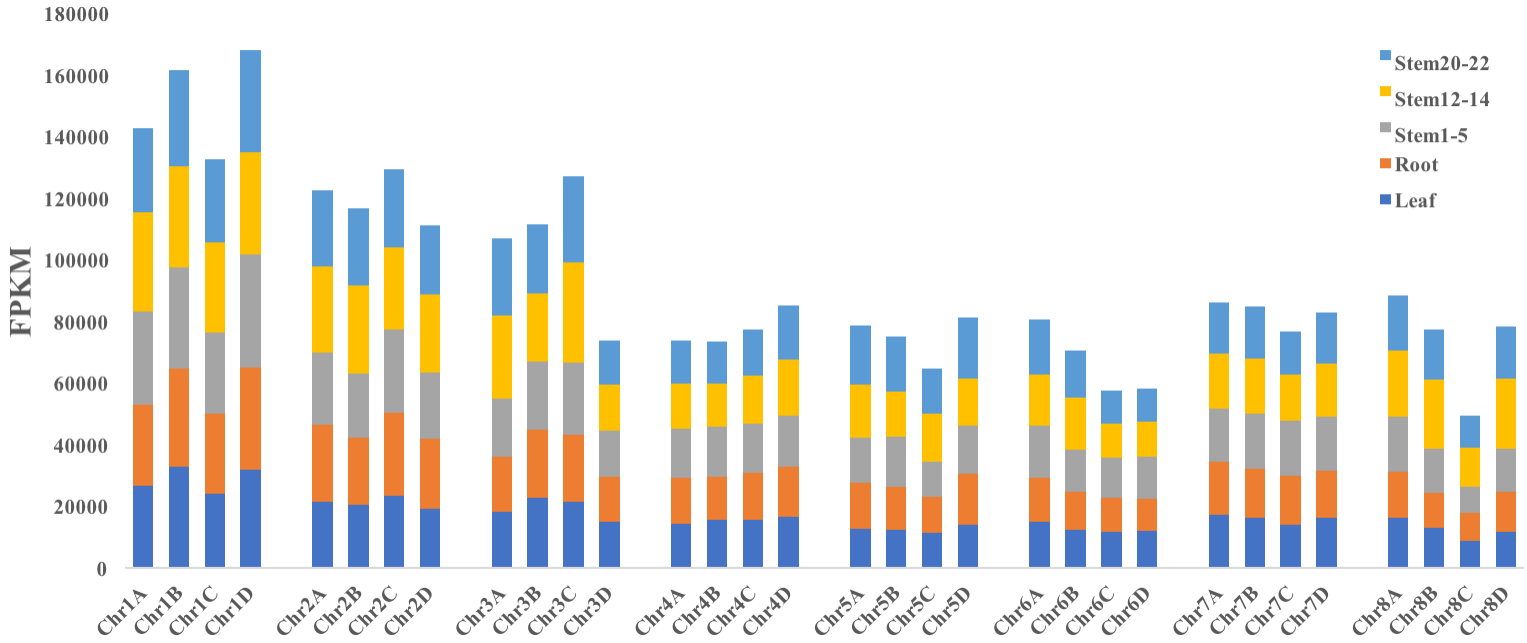


Supplementary Figure 10. Alleleic differential expression analysis of 4,289 genes with full of four alleles in AP85-441 genome.

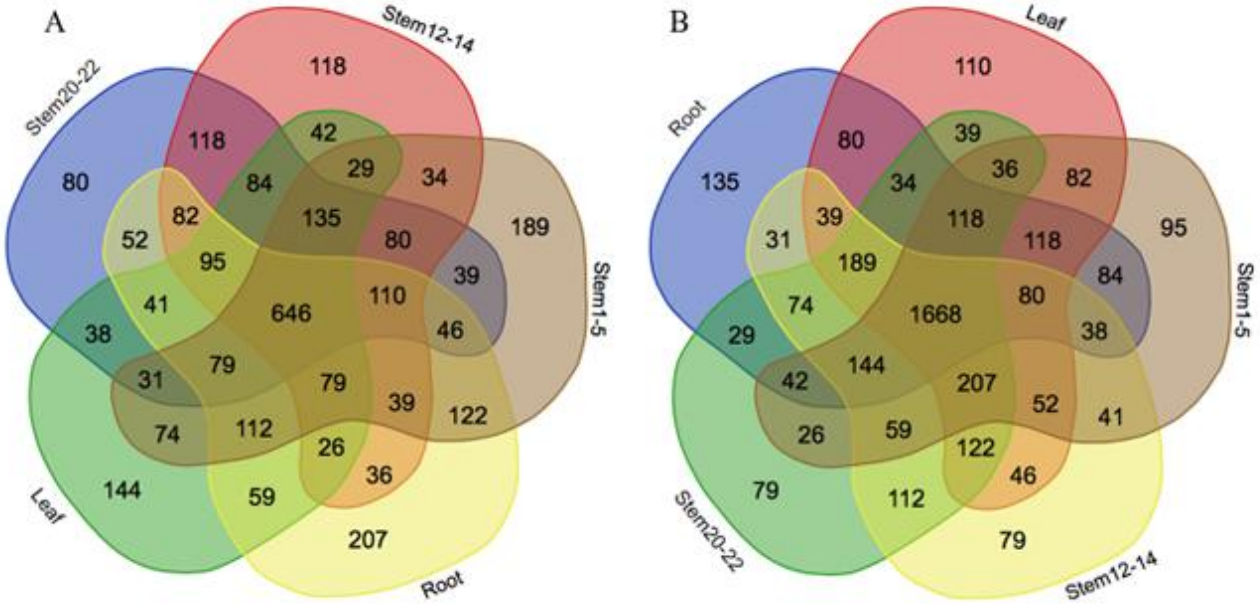
Gene expression in sample leaf (A), stem (B) and root (C). ChrA, ChrB, ChrC and ChrD represent four homologous groups.



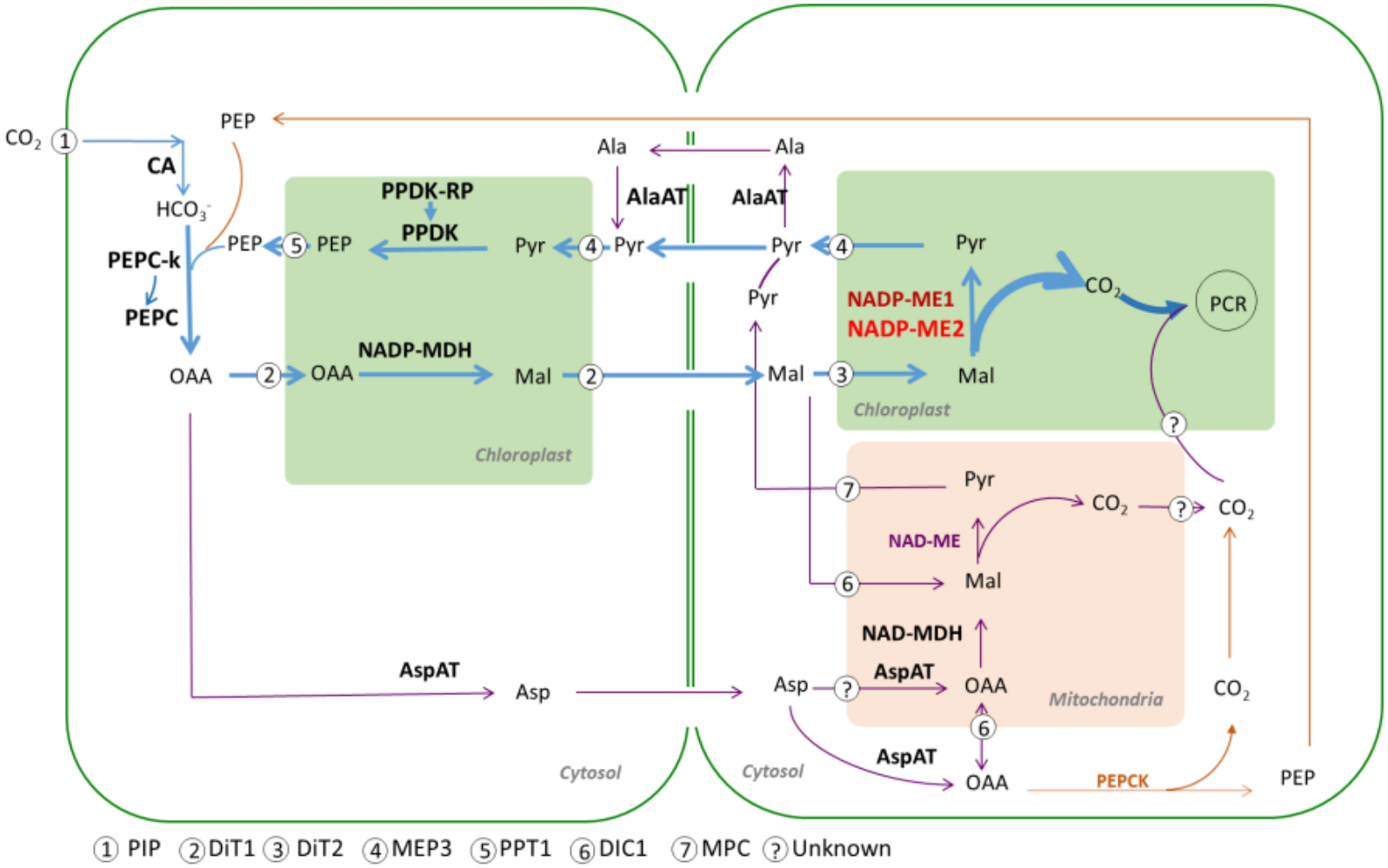
Supplementary Figure 11. The accumulation of gene expression for the genes with 4 alleles



Supplementary Figure 12. The accumulation of gene expression in all of genes.



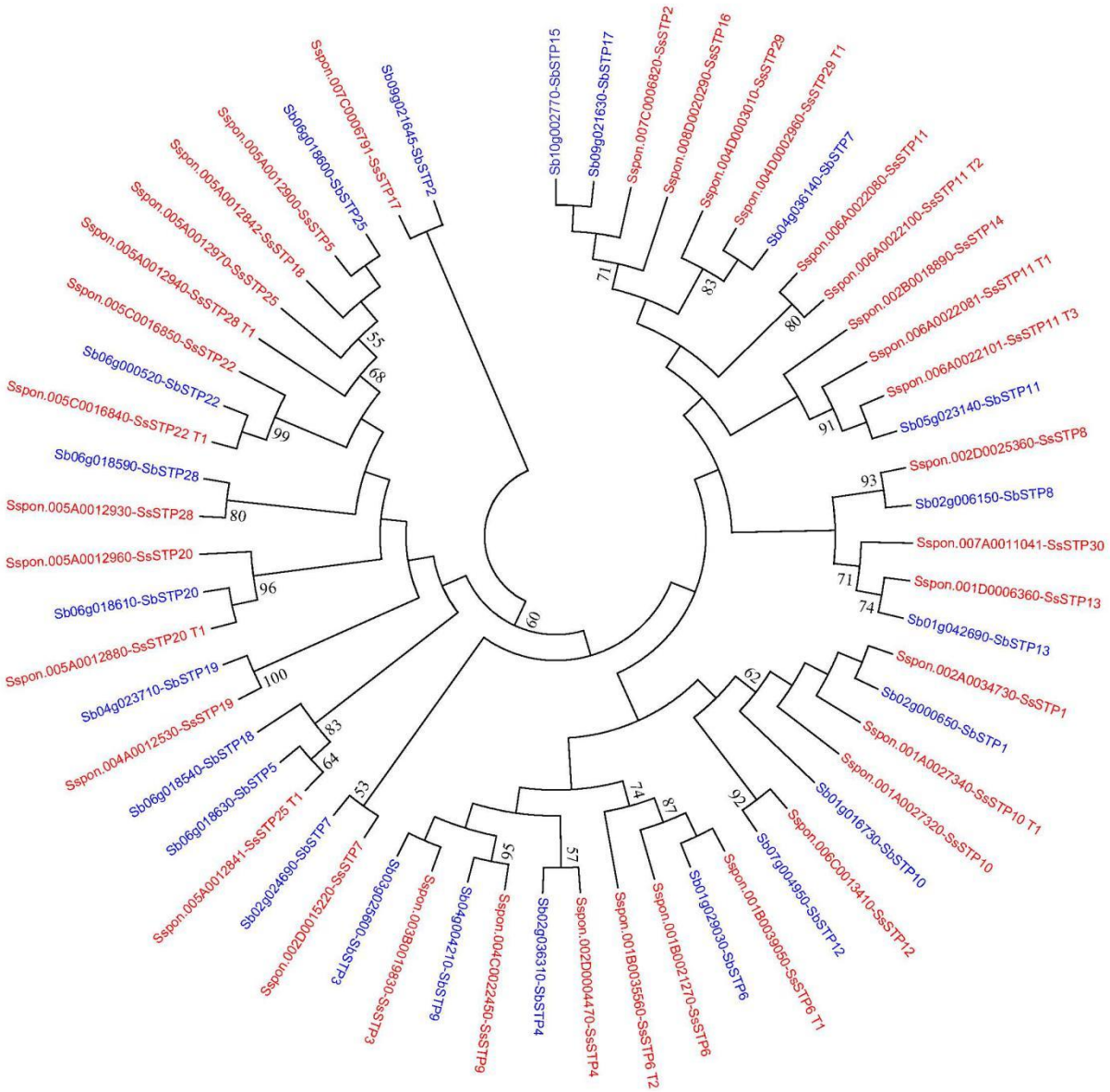
Supplementary Figure 13. Gene dominance common. Neutral (A) and non-Neutral (B) pattern.


CO2 ①
CA
HCO3-
PEPC-k
PEPC
PEP
OAA
②
PPDK-RP
PPDK
PEP
⑤ PEP
Pyr
④ Pyr
NADP-MDH
OAA
② OAA
Mal
②
Chloroplast
Ala
Ala
AlaAT
AlaAT
Pyr
④ Pyr
Pyr
NADP-ME1
NADP-ME2
CO2
PCR
Mal
③ Mal
Mal
Chloroplast
?
⑦ Pyr
CO2
? CO2
NAD-ME
⑥ Mal
NAD-MDH
AspAT
Asp
Asp
? AspAT
OAA
Mitochondria
CO2
⑥
AspAT
OAA
PEPCK
PEP
Cytosol
Cytosol
① PIP ② DiT1 ③ DiT2 ④ MEP3 ⑤ PPT1 ⑥ DIC1 ⑦ MPC ? Unknown

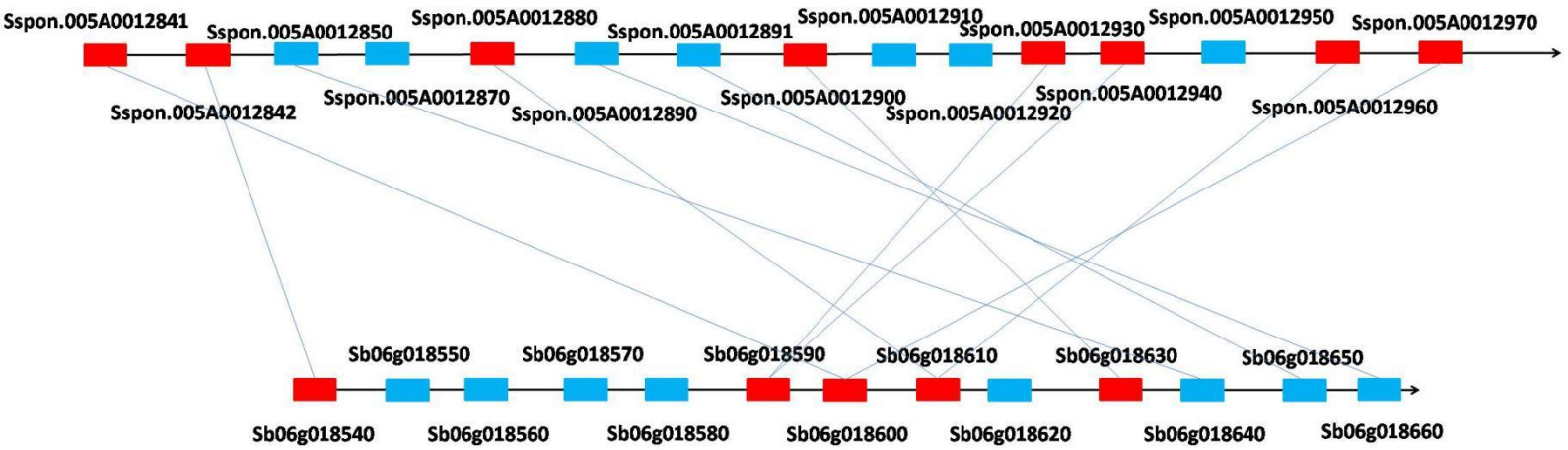
Supplementary Figure 14. An enhanced NADP-ME type C4 pathway in *S. spontaneum*.

The schematic of co-exists of the three decarboxylation pathways in sugarcane, NADP-ME (blue) is the dominant pathway, NAD-ME (purple) and PEPCK (orange) are subsidiary.

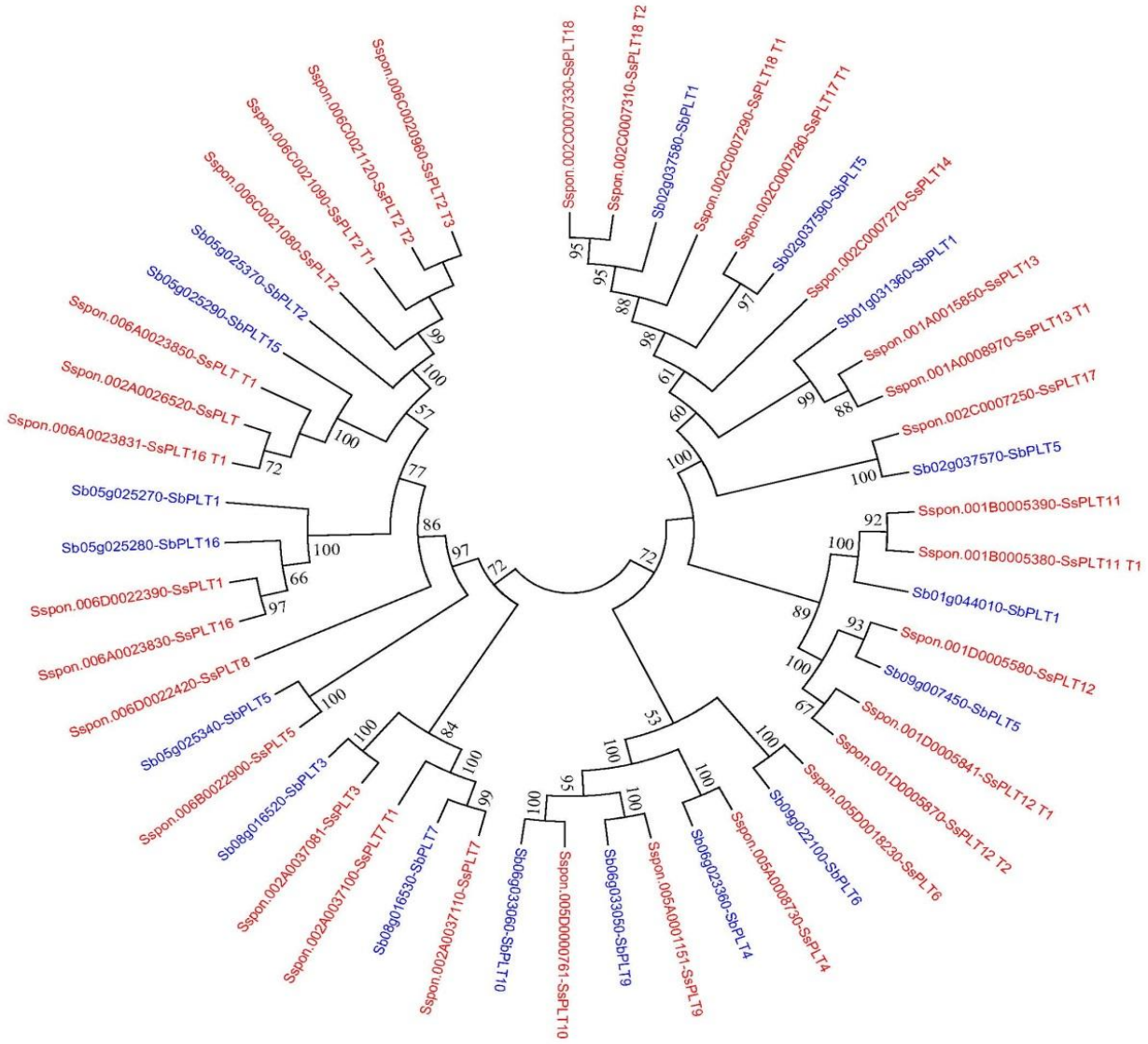
a


Sb10g002770-SbSTP15
Sb09g021630-SbSTP17
Sspon.007C0006820-SsSTP2
Sspon.008D0020290-SsSTP16
Sspon.004D0003010-SsSTP29
Sspon.004D0002960-SsSTP29 T1
Sb04g036140-SbSTP7
Sspon.006A0022080-SsSTP11
Sspon.006A0022100-SsSTP11 T2
Sspon.002B0018890-SsSTP14
Sspon.006A0022081-SsSTP11 T1
Sspon.006A0022101-SsSTP11 T3
Sb05g023140-SbSTP11
Sspon.002D0025360-SsSTP8
Sb02g006150-SbSTP8
Sspon.007A0011041-SsSTP30
Sspon.001D0006360-SsSTP13
Sb01g042690-SbSTP13
Sspon.002A0034730-SsSTP1
Sb02g000650-SbSTP1
Sspon.001A0027340-SsSTP10 T1
Sspon.001A0027320-SsSTP10
Sb01g016730-SbSTP10
Sspon.006C0013410-SsSTP12
Sb07g004950-SbSTP12
Sspon.001B0039060-SsSTP6 T1
Sb01g029030-SbSTP6
Sspon.001B0021270-SsSTP6
Sspon.001B0035660-SsSTP6 T2
Sspon.002D0004470-SsSTP4
Sb02g036310-SbSTP4
Sspon.004C0022450-SsSTP9
Sb04g004210-SbSTP9
Sspon.003B0019830-SsSTP3
Sb03g025600-SbSTP3
Sspon.002D0015220-SsSTP7
Sb02g024690-SbSTP7
Sspon.005A0012841-SsSTP25 T1
Sb06g018630-SbSTP5
Sb06g018540-SbSTP18
Sspon.004A0012530-SsSTP19
Sb04g023710-SbSTP19
Sspon.005A0012880-SsSTP20 T1
Sb06g018610-SbSTP20
Sspon.005A0012960-SsSTP20
Sspon.005A0012930-SsSTP28
Sb06g018590-SbSTP28
Sspon.005C0016840-SsSTP22 T1
Sb06g000520-SbSTP22
Sspon.005C0016850-SsSTP22
Sspon.005A0012940-SsSTP28 T1
Sspon.005A0012970-SsSTP25
Sspon.005A0012842-SsSTP18
Sspon.005A0012900-SsSTP5
Sb06g018600-SbSTP25
Sspon.007C0006791-SsSTP17
Sb09g021645-SbSTP2
71
83
80
91
93
71
74
62
92
74
87
57
95
53
64
83
100
96
80
99
68
55
60

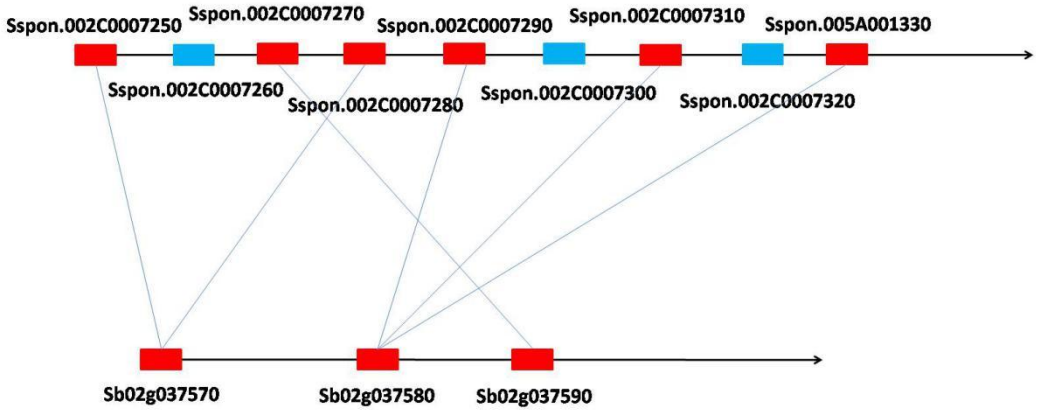
b


Sspon.005A0012841
Sspon.005A0012850
Sspon.005A0012880
Sspon.005A0012891
Sspon.005A0012910
Sspon.005A0012930
Sspon.005A0012950
Sspon.005A0012970
Sspon.005A0012842
Sspon.005A0012870
Sspon.005A0012890
Sspon.005A0012900
Sspon.005A0012920
Sspon.005A0012940
Sspon.005A0012960
Sb06g018550
Sb06g018570
Sb06g018590
Sb06g018610
Sb06g018630
Sb06g018650
Sb06g018540
Sb06g018560
Sb06g018580
Sb06g018600
Sb06g018620
Sb06g018640
Sb06g018660

c

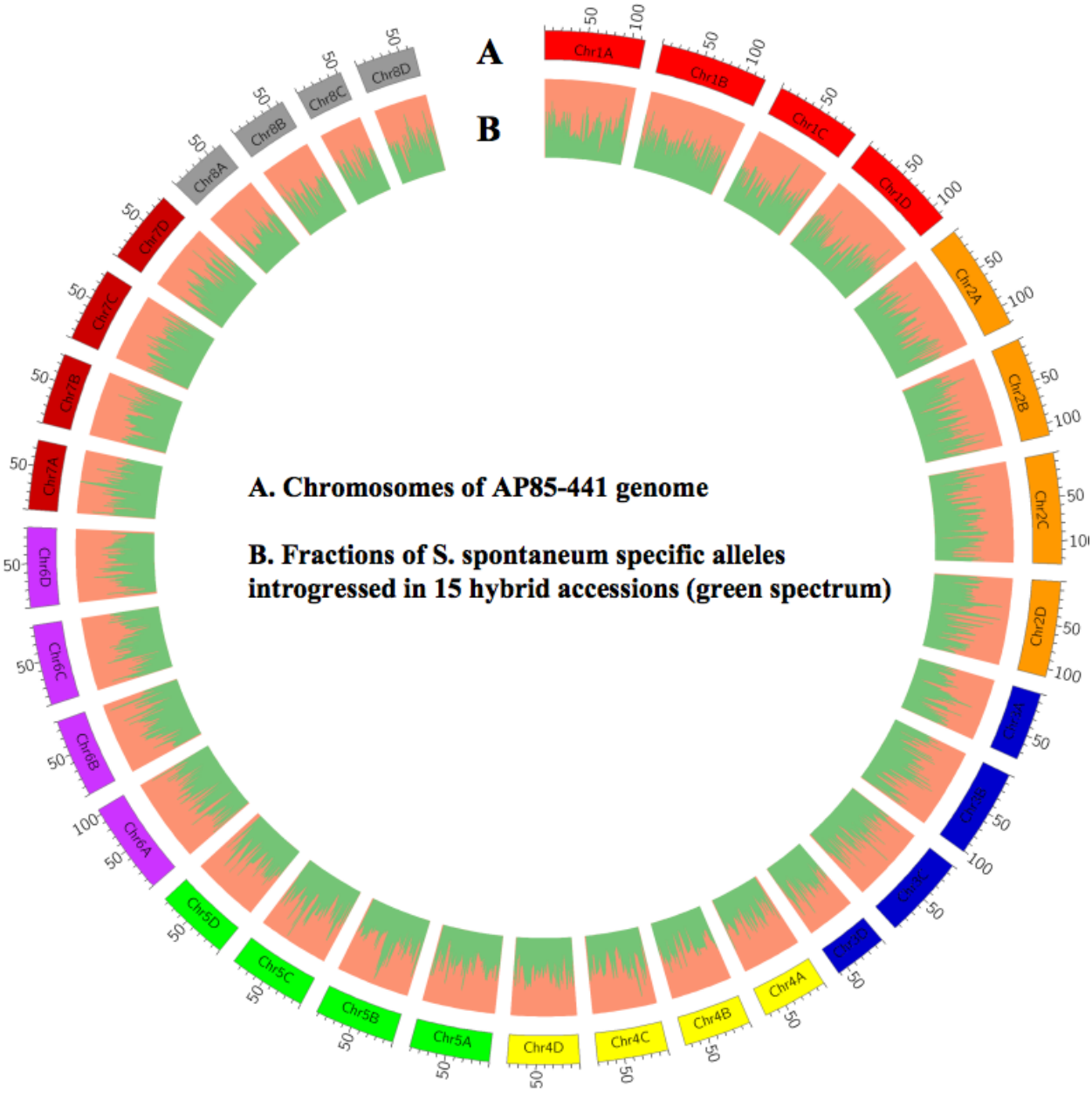

Sspon.002C0007330-SsPLT18
Sspon.002C0007310-SsPLT18 T2
Sb02g037580-SbPLT1
Sspon.002C0007280-SsPLT18 T1
Sspon.002C0007280-SsPLT17 T1
Sb02g037590-SbPLT5
Sspon.002C0007270-SsPLT14
Sb01g031360-SbPLT1
Sspon.001A0015850-SsPLT13
Sspon.001A0008970-SsPLT13 T1
Sspon.002C0007250-SsPLT17
Sb02g037570-SbPLT5
Sspon.001B0005390-SsPLT11
Sspon.001B0005380-SsPLT11 T1
Sb01g044010-SbPLT1
Sspon.001D0005580-SsPLT12
Sb09g007450-SbPLT5
Sspon.001D0005841-SsPLT12 T1
Sspon.001D0005870-SsPLT12 T2
Sspon.005D0018230-SsPLT6
Sb09g022100-SbPLT6
Sspon.005A0008730-SsPLT4
Sb06g023960-SbPLT4
Sspon.005A0001151-SsPLT9
Sb06g033050-SbPLT9
Sspon.005D0000761-SsPLT10
Sb06g033060-SbPLT10
Sspon.002A0037110-SsPLT7
Sb08g016530-SbPLT7
Sspon.002A0037100-SsPLT7 T1
Sspon.002A0037081-SsPLT7
Sb08g016520-SbPLT3
Sspon.006B0022900-SsPLT5
Sb05g025340-SbPLT5
Sspon.006D0022420-SsPLT8
Sspon.006A0023830-SsPLT16
Sspon.006D0022390-SsPLT1
Sb05g025280-SbPLT16
Sb05g025270-SbPLT1
Sspon.006A0023831-SsPLT16 T1
Sspon.002A0026520-SsPLT
Sspon.006A0023850-SsPLT T1
Sb05g025290-SbPLT15
Sb05g025370-SbPLT2
Sspon.006C0021080-SsPLT2
Sspon.006C0021090-SsPLT2 T1
Sspon.006C0021120-SsPLT2 T2
Sspon.006C0020980-SsPLT2 T3

d

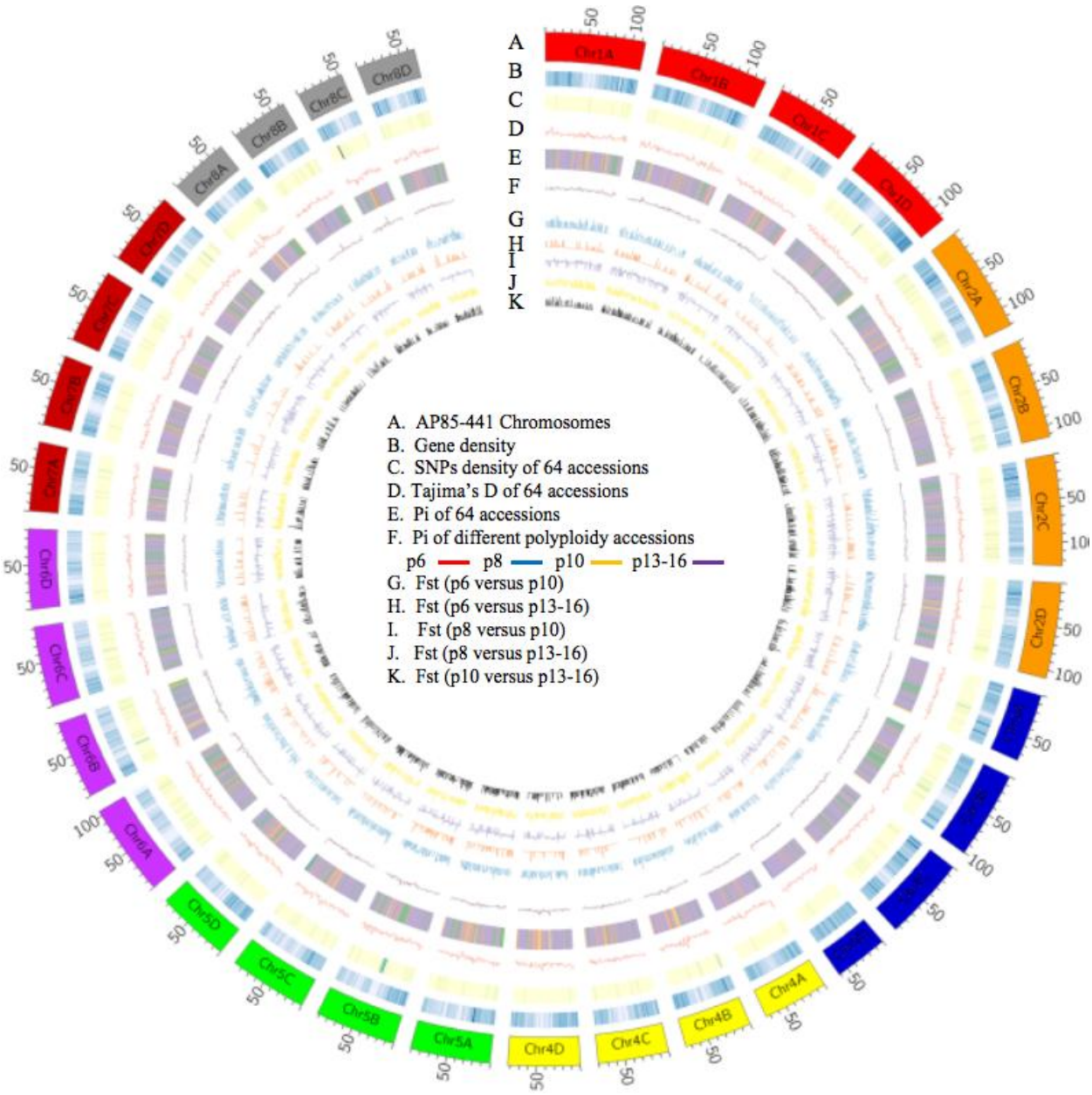


Supplementary Figure 15. Co-linear segment between *S. spontaneum* and Sorghum showing lots of tandem duplication events in sugar transporter gene families.

Genes in the STP family (a) and PLT family (c) were separately clustered using MEGA7 with the neighbor-joining method. Genes of *S. spontaneum* and Sorghum were marked with red and blue colors respectively (a,c). The co-linearity between the two species was identified by reciprocal blast analyses. The genes in the STP and PLT families were represented by red boxes and orthologous genes are linked by blue lines (b and d), while the other genes were represented with blue boxes.

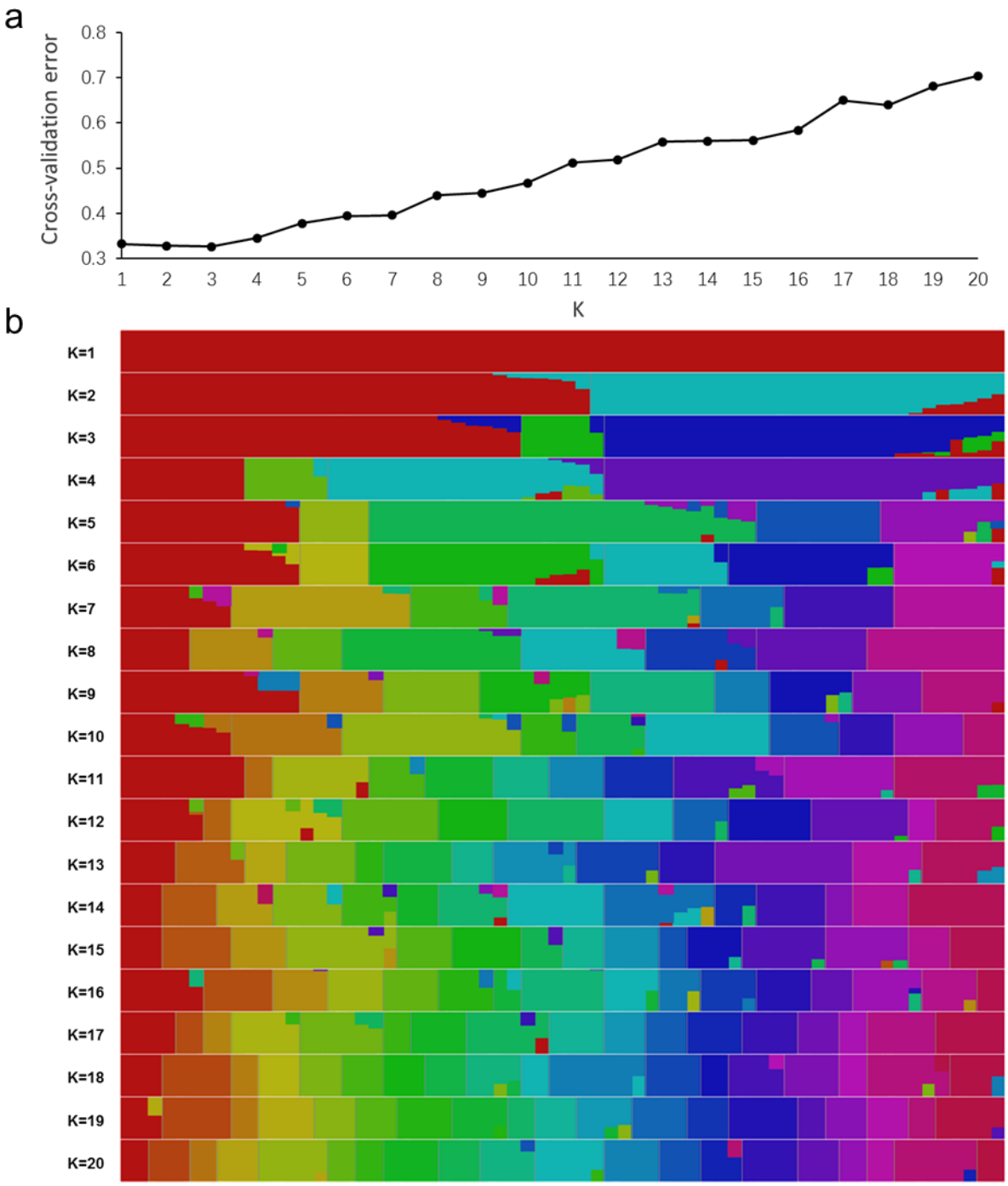


Supplementary Figure 16. Fractions of *S. spontaneum*specific alleles introgressed in 15 hybrid accessions (green spectrum) across AP85-441 genome.



Supplementary Figure 17. Genomic diversity map for *Saccharum spontaneum*.

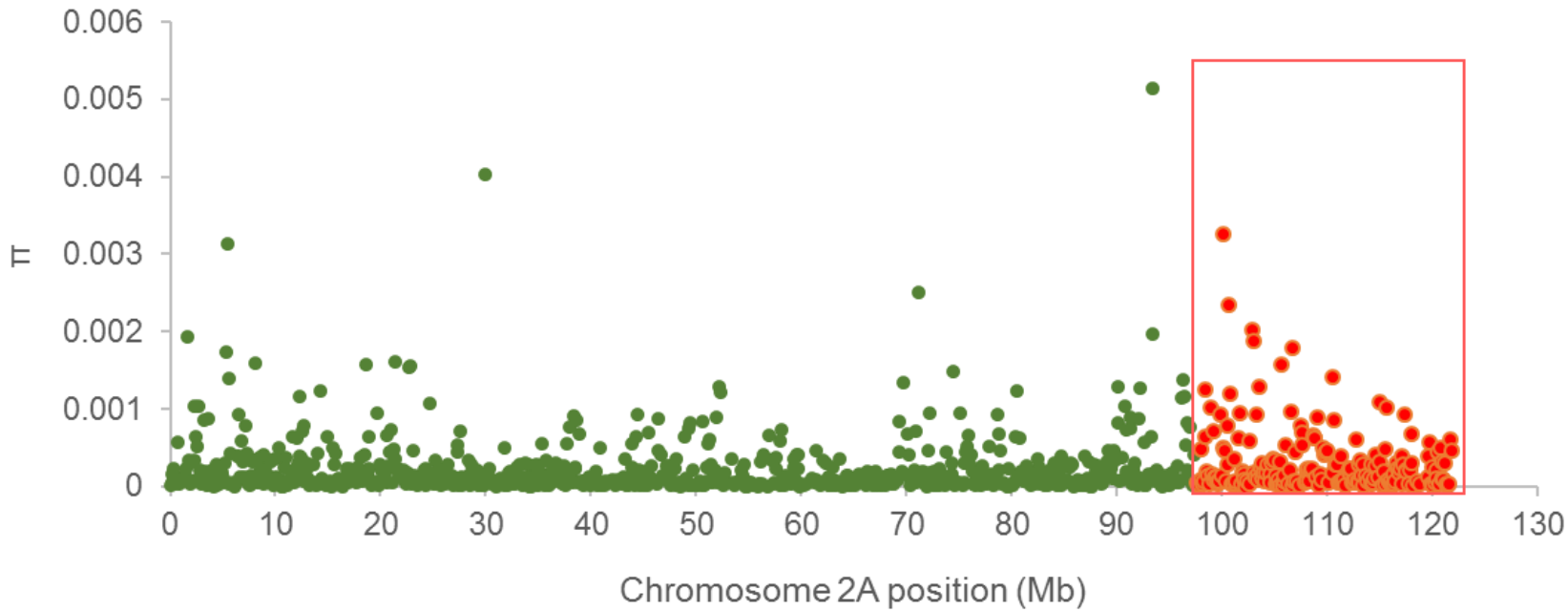
Circos plot showing variants diversity across the 8 chromosomes for 64 *S.spontaneum* accessions. From the outer to the center, represents: A. AP85-441 Chromosomes; B. Gene density; C. SNPs density of 64 accessions; D. Tajima's D of 64 accessions for testing the neutral evolution of variant sites at equilibrium between mutation and genetic drift; E. nucleotide diversity Pi (π) of 64 accessions; F. nucleotide diversity Pi (π) of different polyploidy accessions (p6 hexaploid, p8 octoploid, p10 decaploid, p13-16 tridecaploid to hexadecaploid); G. pairwise *Fst* between hexaploid and decaploid (p6 versus p10); H. pairwise *Fst* between hexaploid and hyperploid more thantridecaploid (p6 versus p13-16); I. pairwise *Fst* between octoploid and decaploid (p8 versus p10); J. pairwise *Fst* between octoploid and hyperploid more thantridecaploid (p8 versus p13-16); K. *Fst* between octoploid and hyperploid more thantridecaploid (p10 versus p13-16).



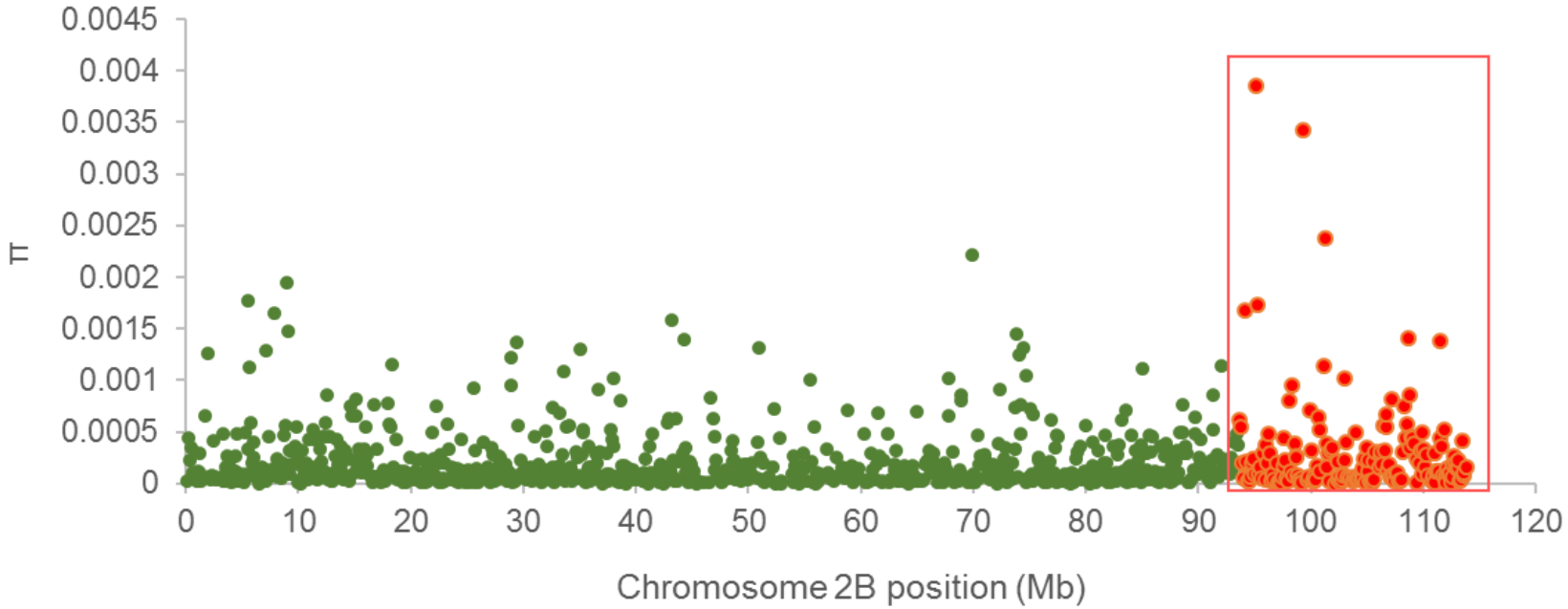
Supplementary Figure 18. Admixture results from K=1 to K=20.

(a) Cross-validate errors show K=3 is the optimal population cluster grouping; (b) population from K=1 to K=20 to show the population structure.

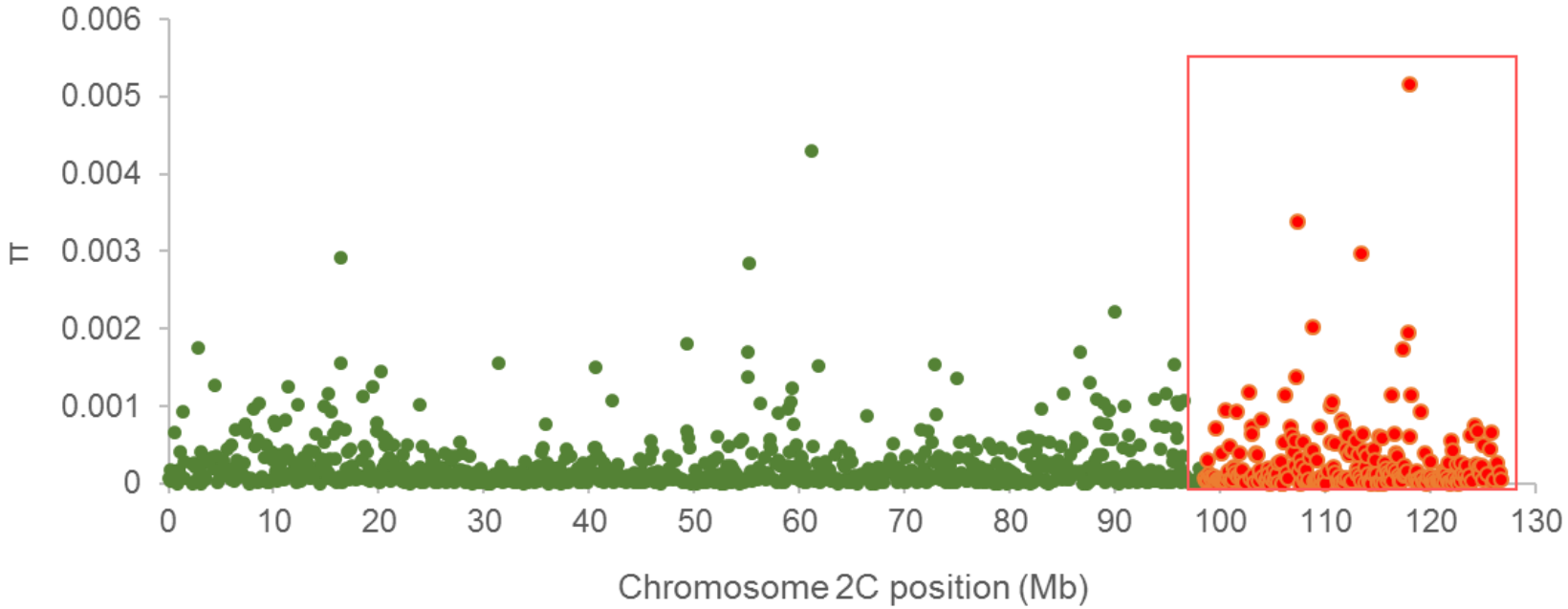
a1


0.006
0.005
0.004
0.003
0.002
0.001
0
π
0
10
20
30
40
50
60
70
80
90
100
110
120
130
Chromosome 2A position (Mb)

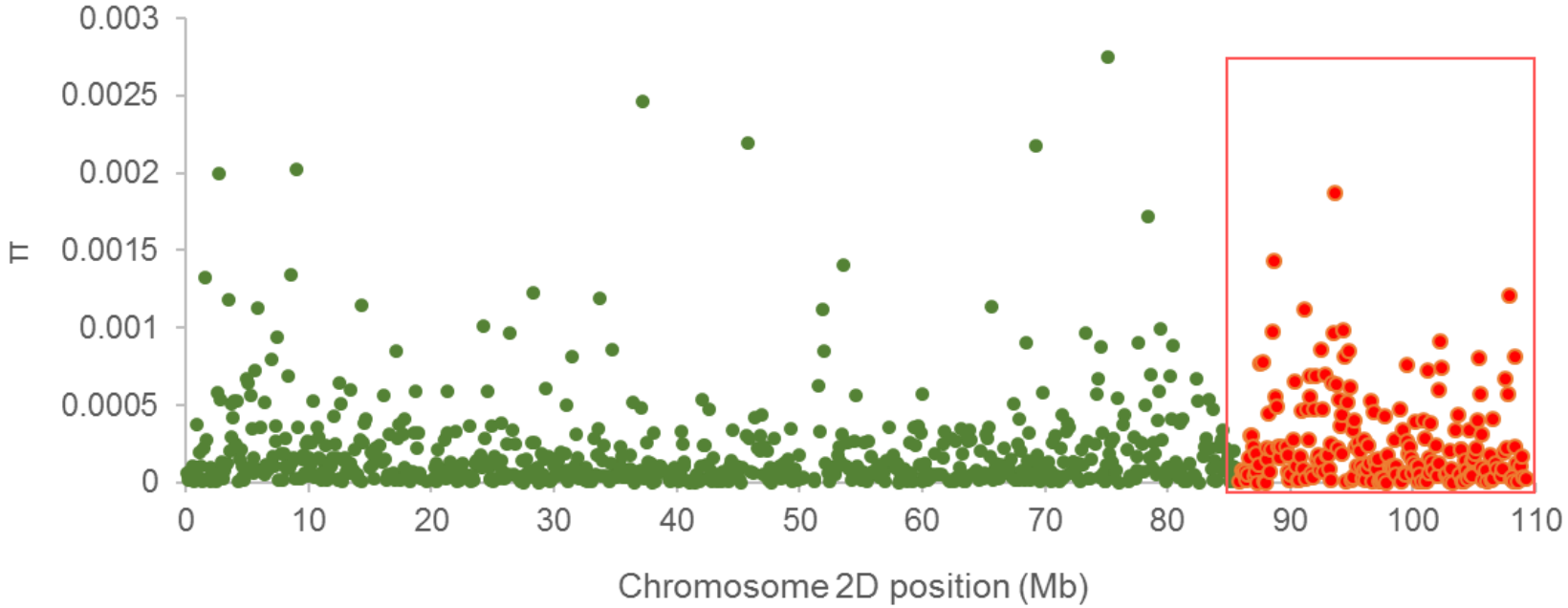
a2


0.0045
0.004
0.0035
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
π
0
10
20
30
40
50
60
70
80
90
100
110
120
Chromosome 2B position (Mb)

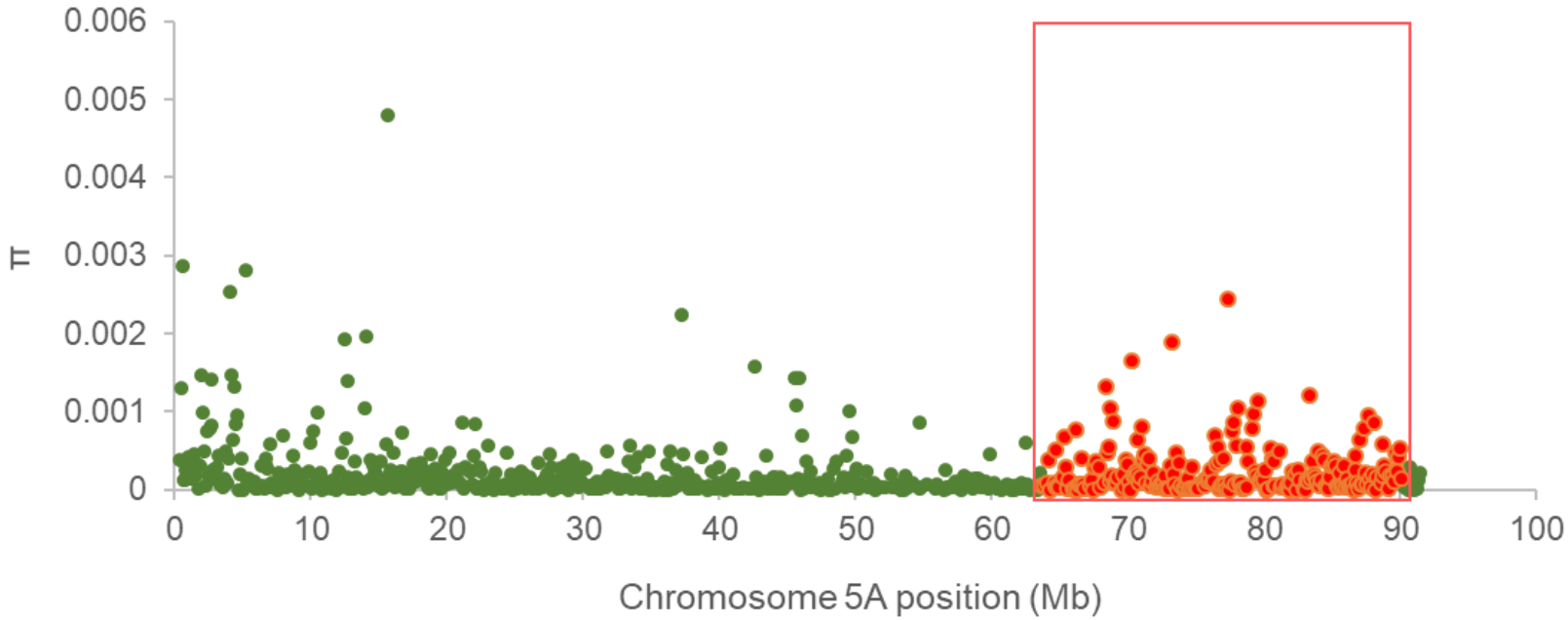
a3


0.006
0.005
0.004
0.003
0.002
0.001
0
π
0
10
20
30
40
50
60
70
80
90
100
110
120
130
Chromosome 2C position (Mb)

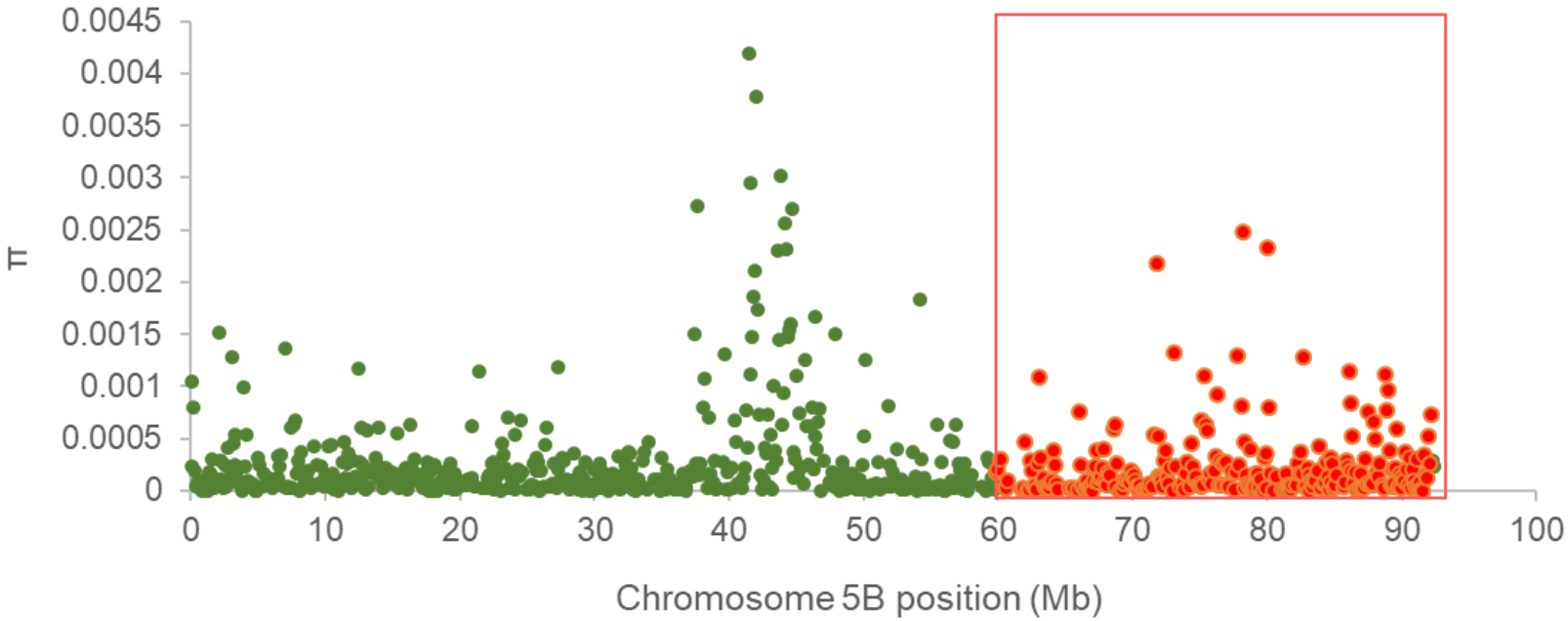
a4


0.003
0.0025
0.002
0.0015
0.001
0.0005
0
π
0
10
20
30
40
50
60
70
80
90
100
110
Chromosome 2D position (Mb)

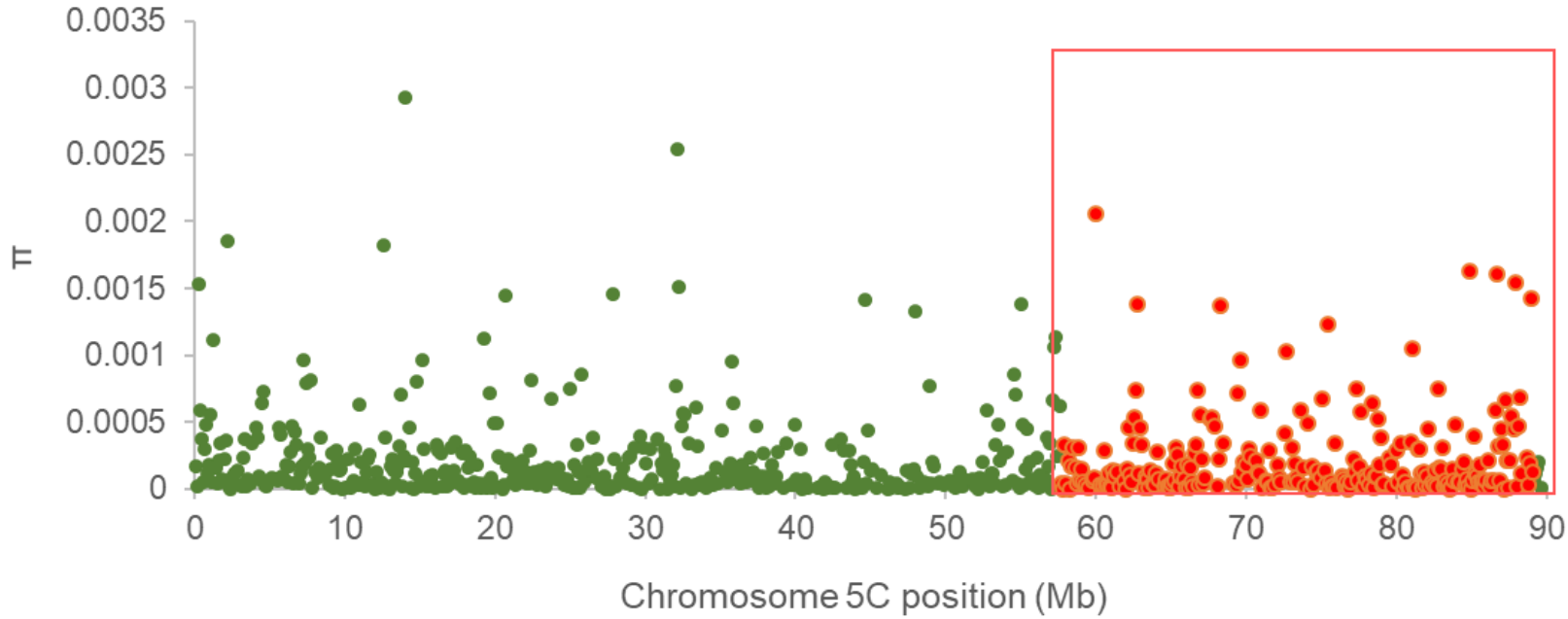
b1


0.006
0.005
0.004
0.003
0.002
0.001
0
π
0
10
20
30
40
50
60
70
80
90
100
Chromosome 5A position (Mb)

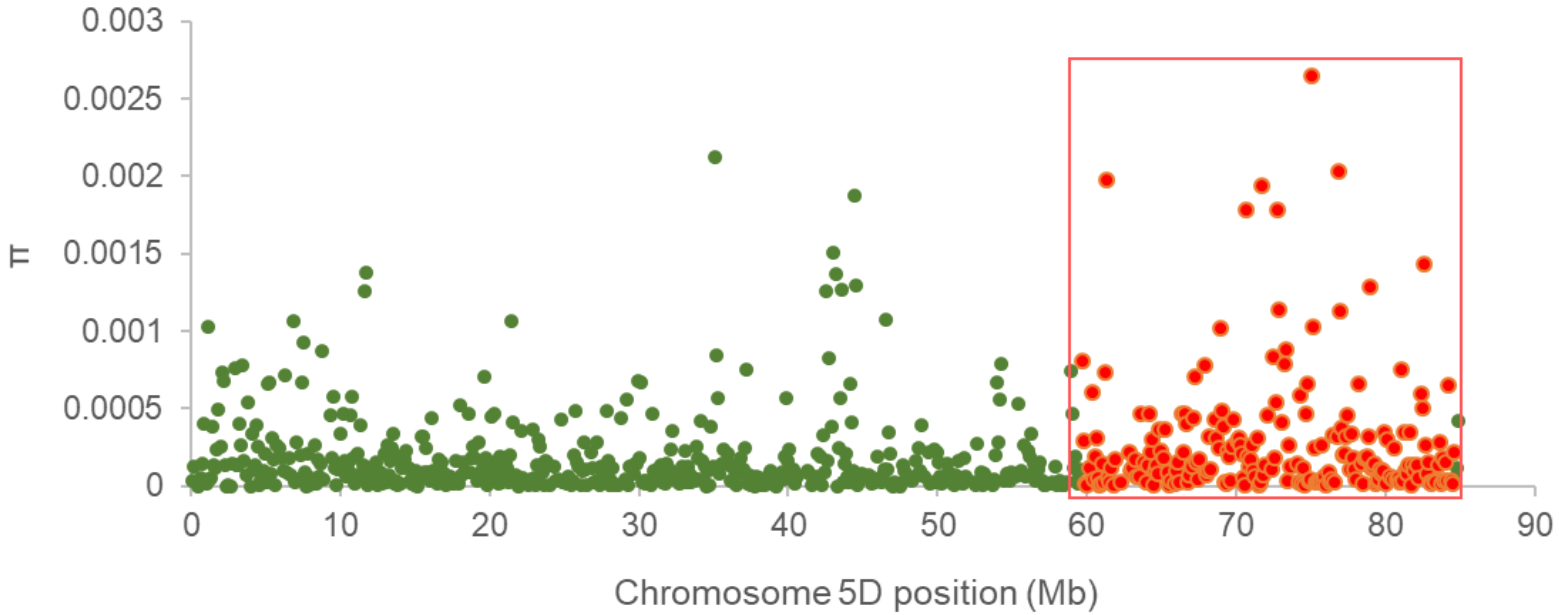
b2


0.0045
0.004
0.0035
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
π
0
10
20
30
40
50
60
70
80
90
100
Chromosome 5B position (Mb)

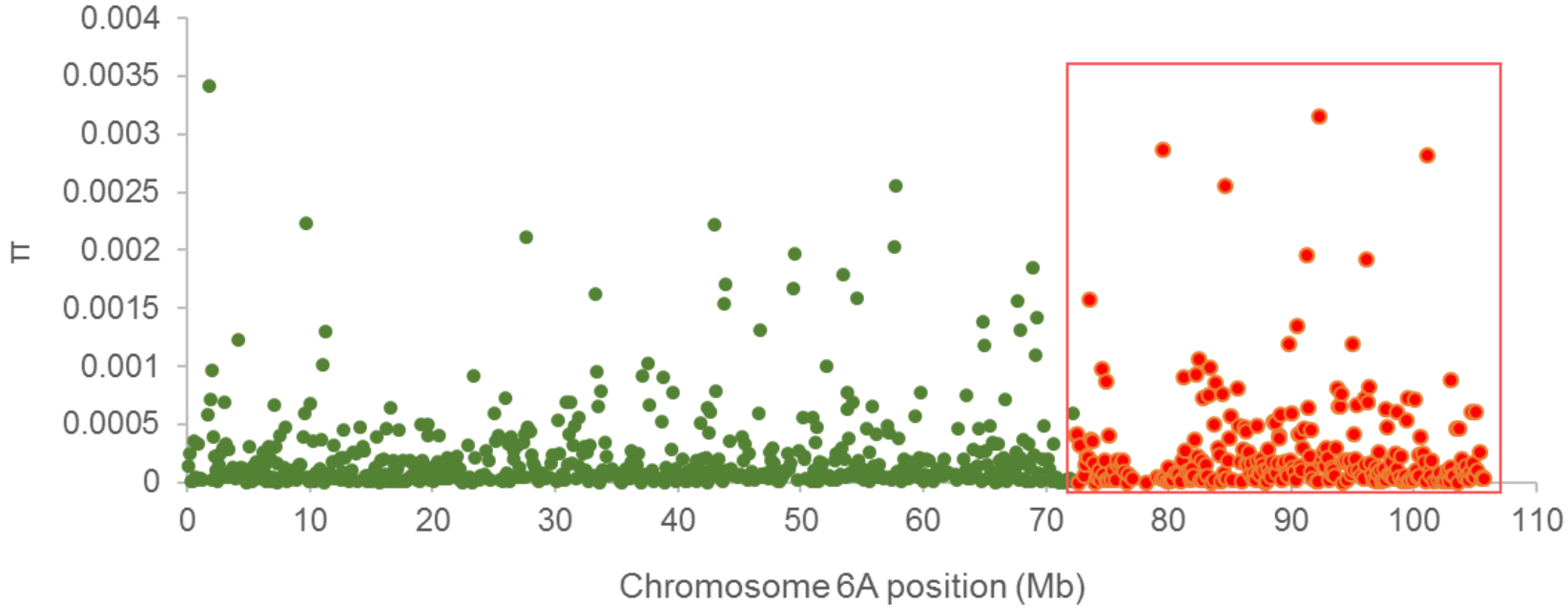
b3


0.0035
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
π
0
10
20
30
40
50
60
70
80
90
Chromosome 5C position (Mb)

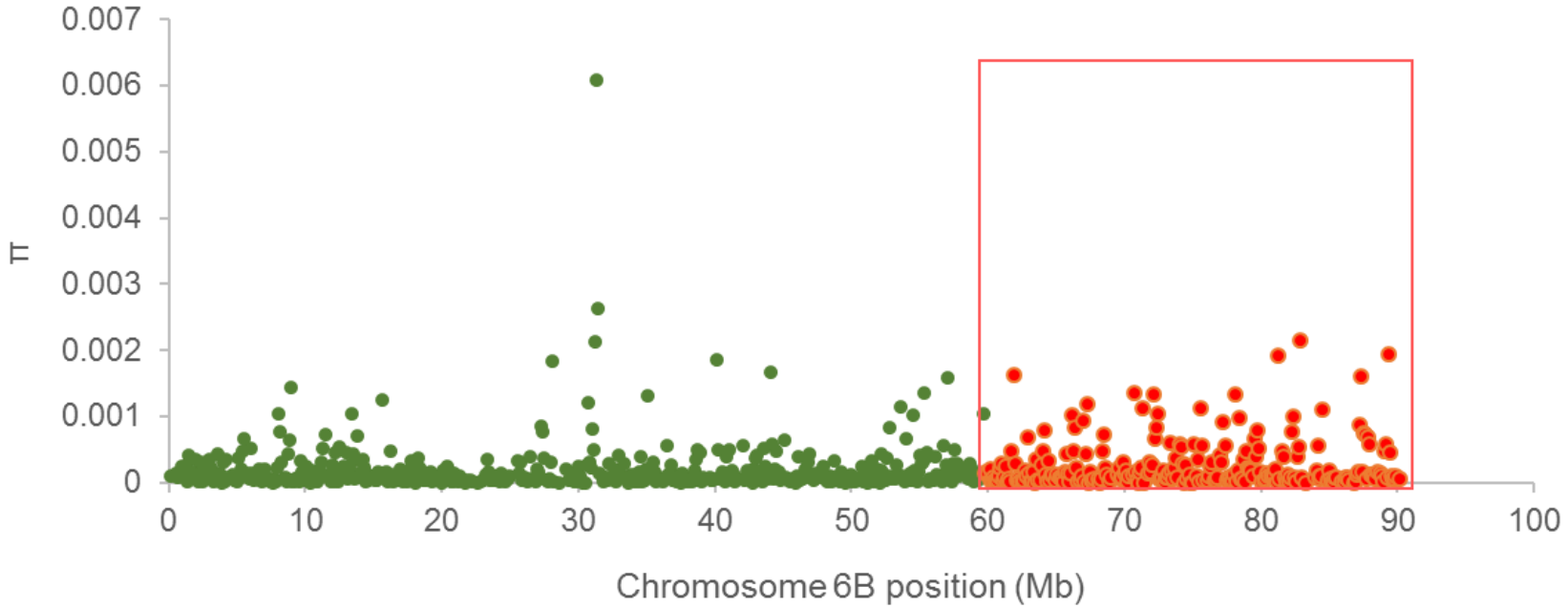
b4


0.003
0.0025
0.002
0.0015
0.001
0.0005
0
π
0
10
20
30
40
50
60
70
80
90
Chromosome 5D position (Mb)

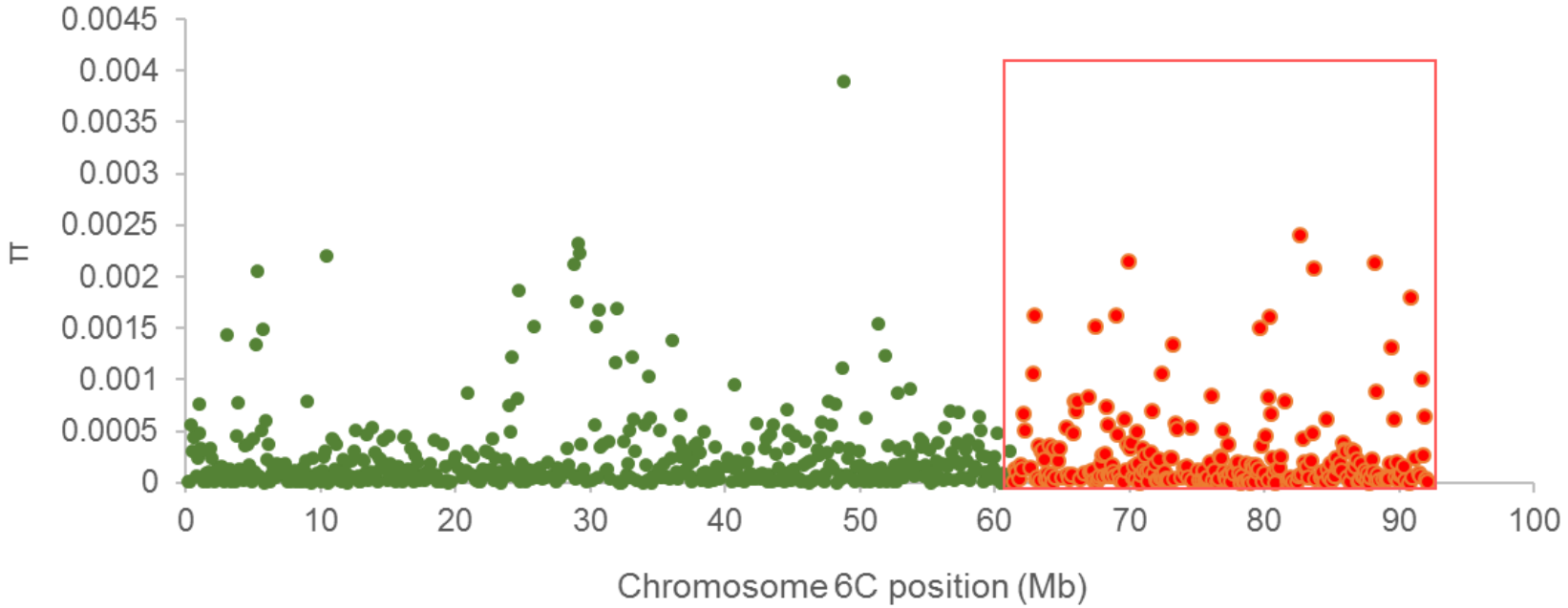
c1


0.004
0.0035
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
π
0
10
20
30
40
50
60
70
80
90
100
110
Chromosome 6A position (Mb)

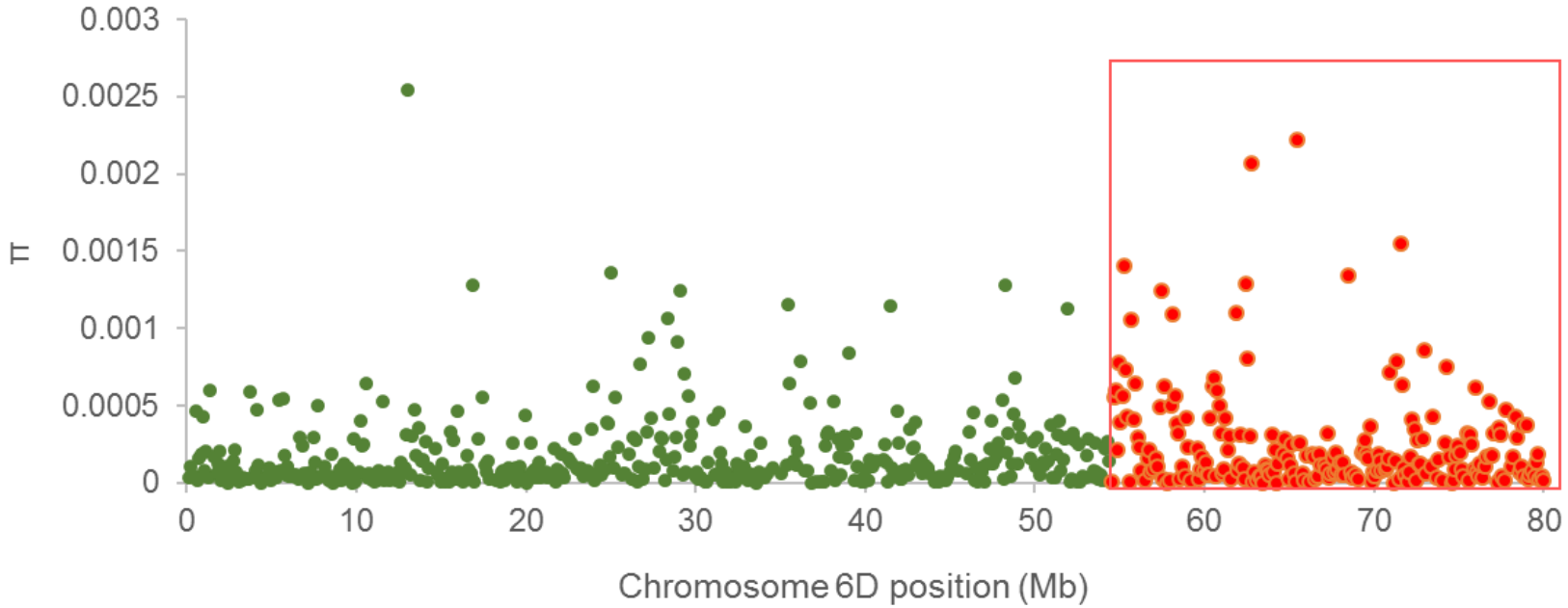
c2


0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
π
0
10
20
30
40
50
60
70
80
90
100
Chromosome 6B position (Mb)

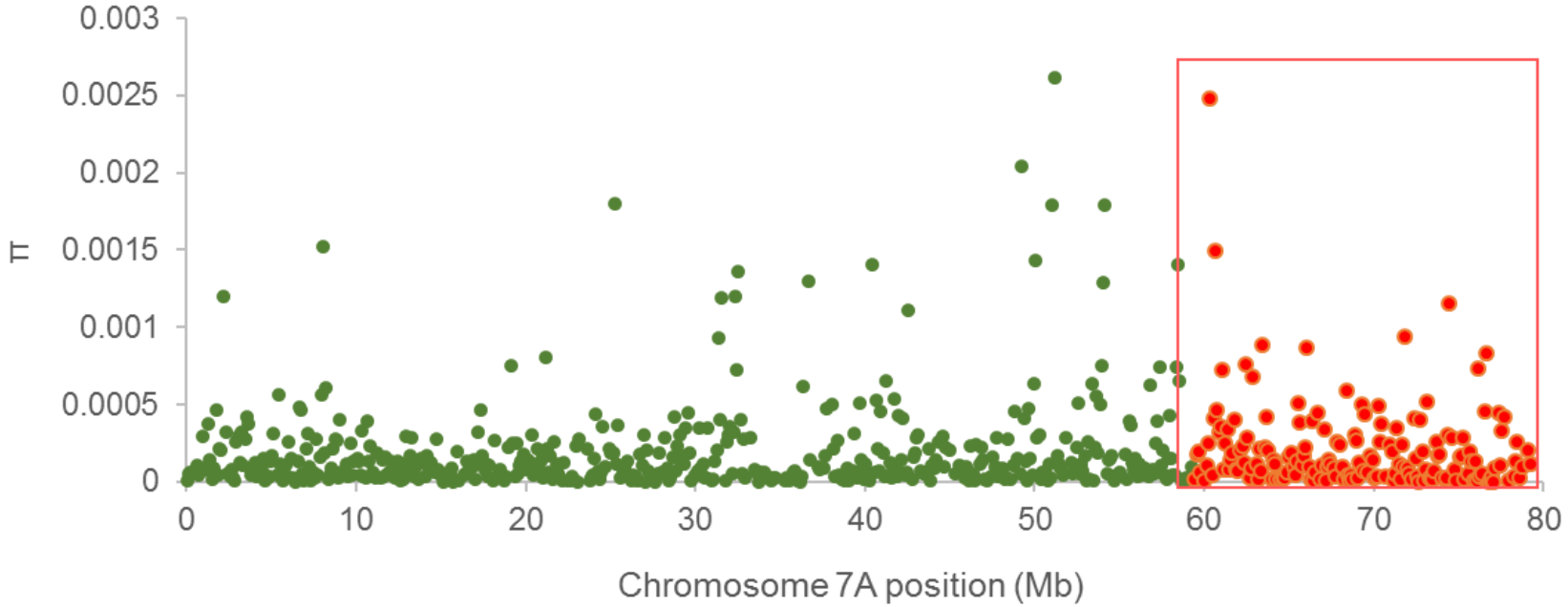
c3


0.0045
0.004
0.0035
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
π
0
10
20
30
40
50
60
70
80
90
100
Chromosome 6C position (Mb)

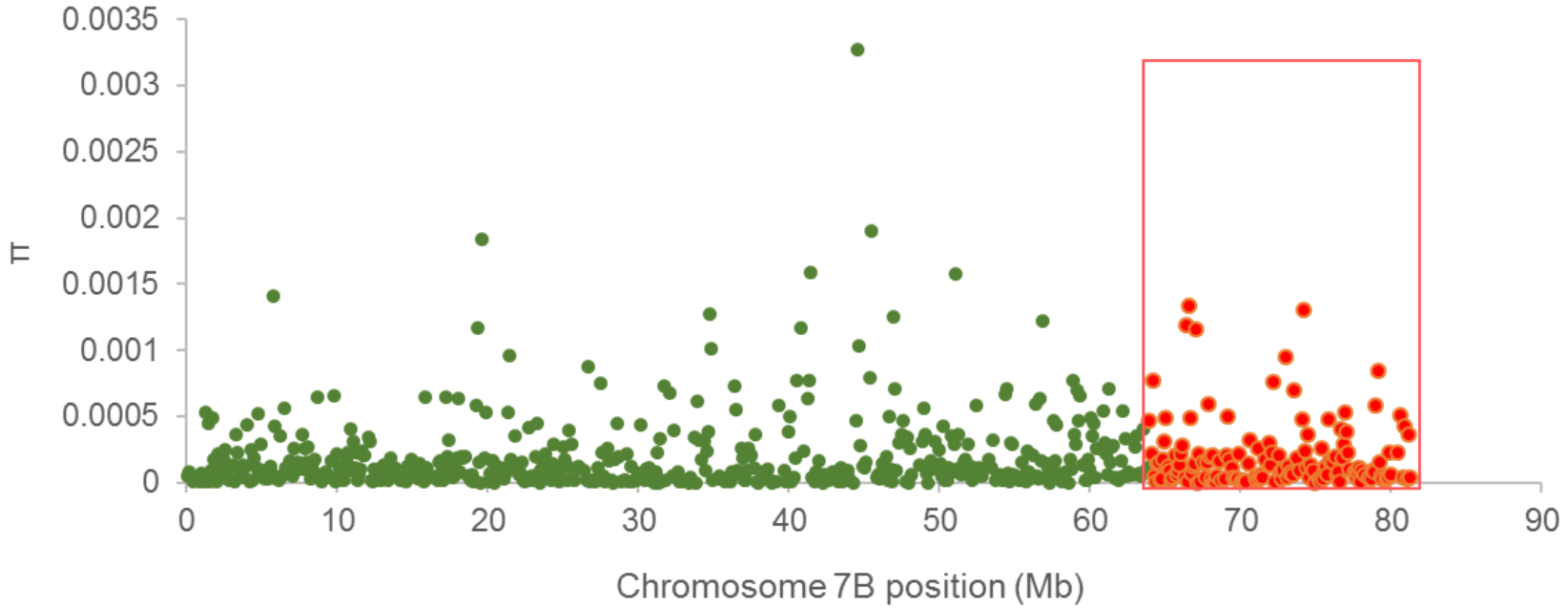
c4


0.003
0.0025
0.002
0.0015
0.001
0.0005
0
π
0
10
20
30
40
50
60
70
80
Chromosome 6D position (Mb)

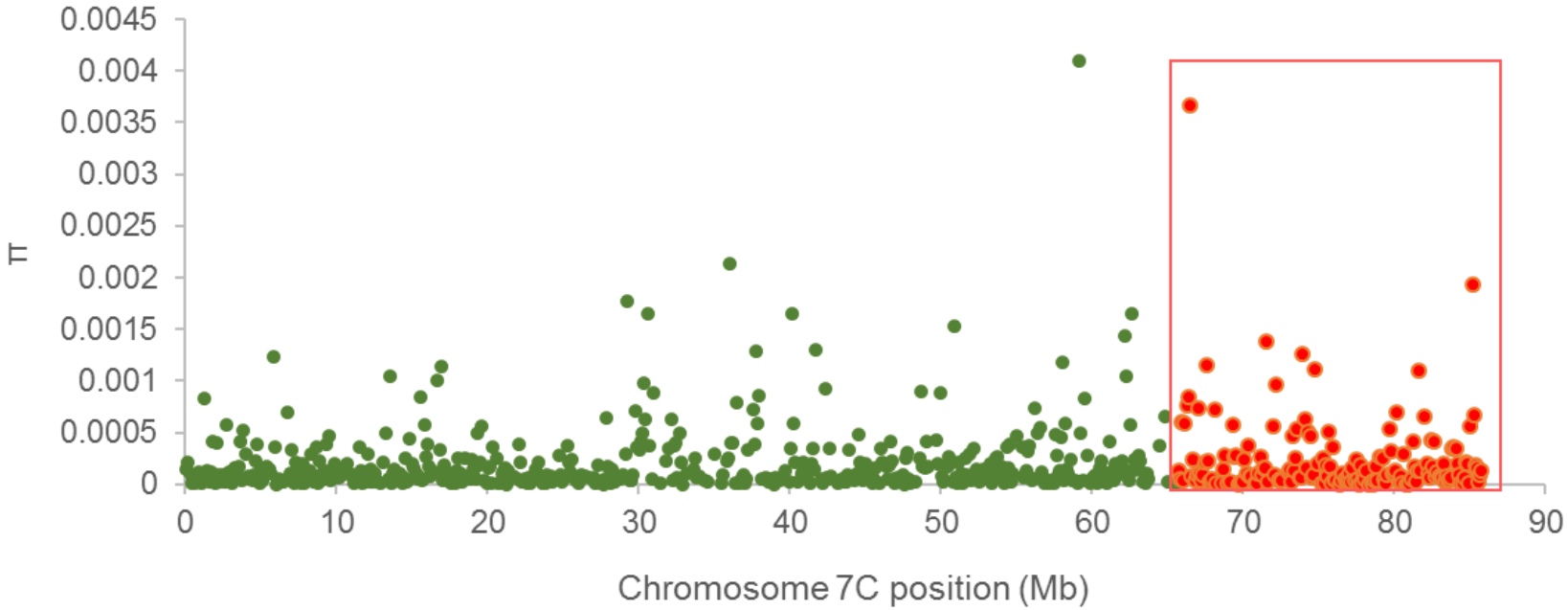
d1

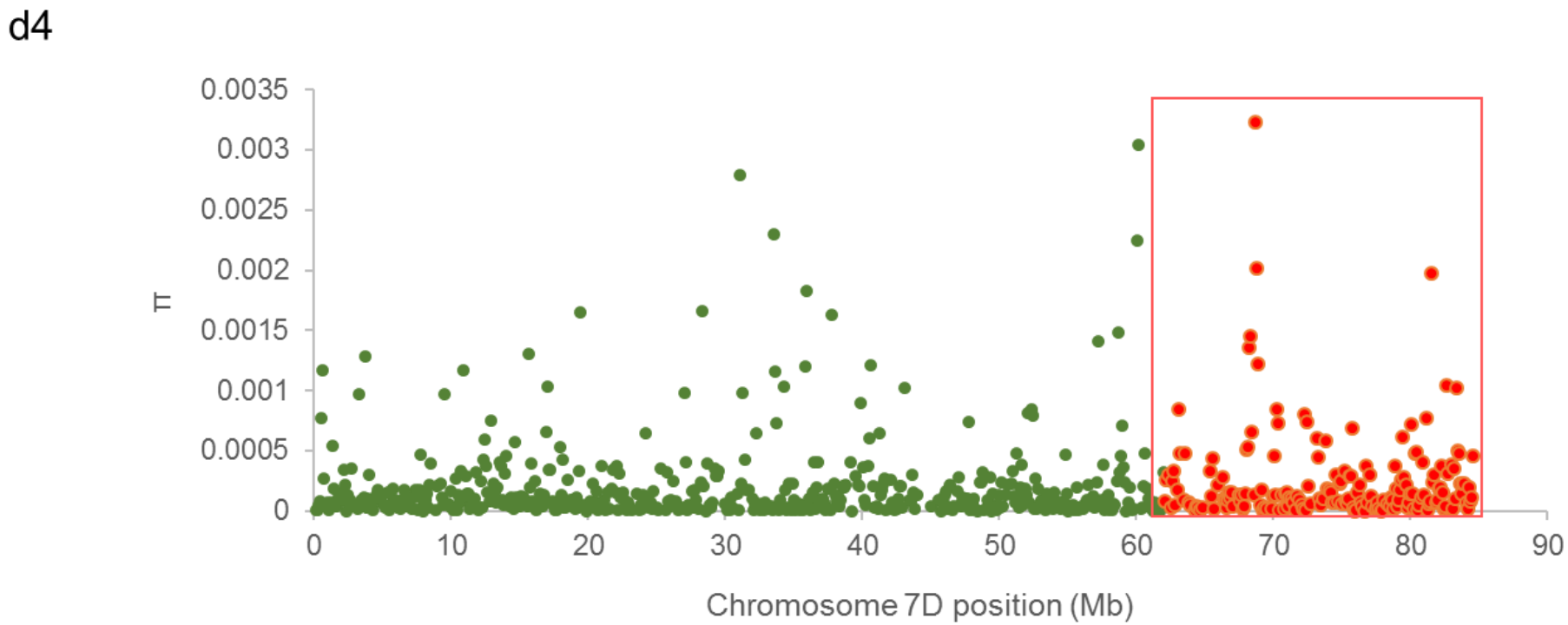

0.003
0.0025
0.002
0.0015
0.001
0.0005
0
π
0
10
20
30
40
50
60
70
80
Chromosome 7A position (Mb)

d2


0.0035
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
π
0
10
20
30
40
50
60
70
80
90
Chromosome 7B position (Mb)

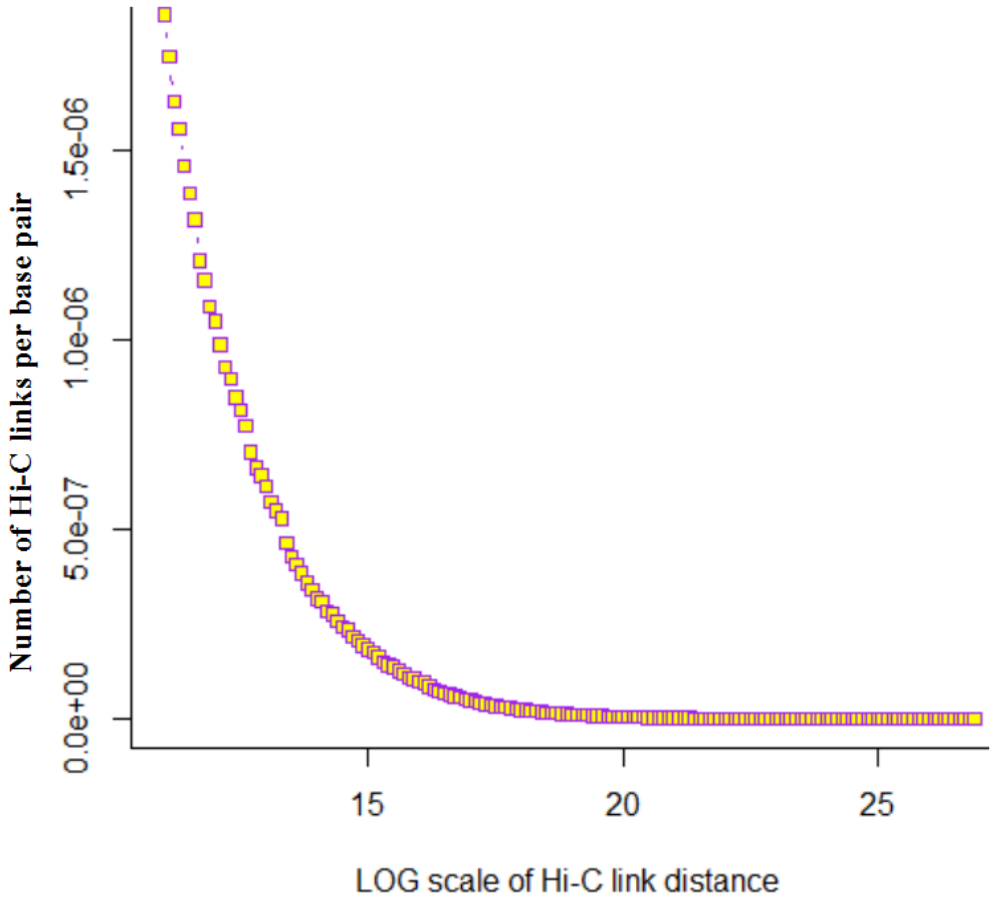
d3


0.0045
0.004
0.0035
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
π
0
10
20
30
40
50
60
70
80
90
Chromosome 7C position (Mb)



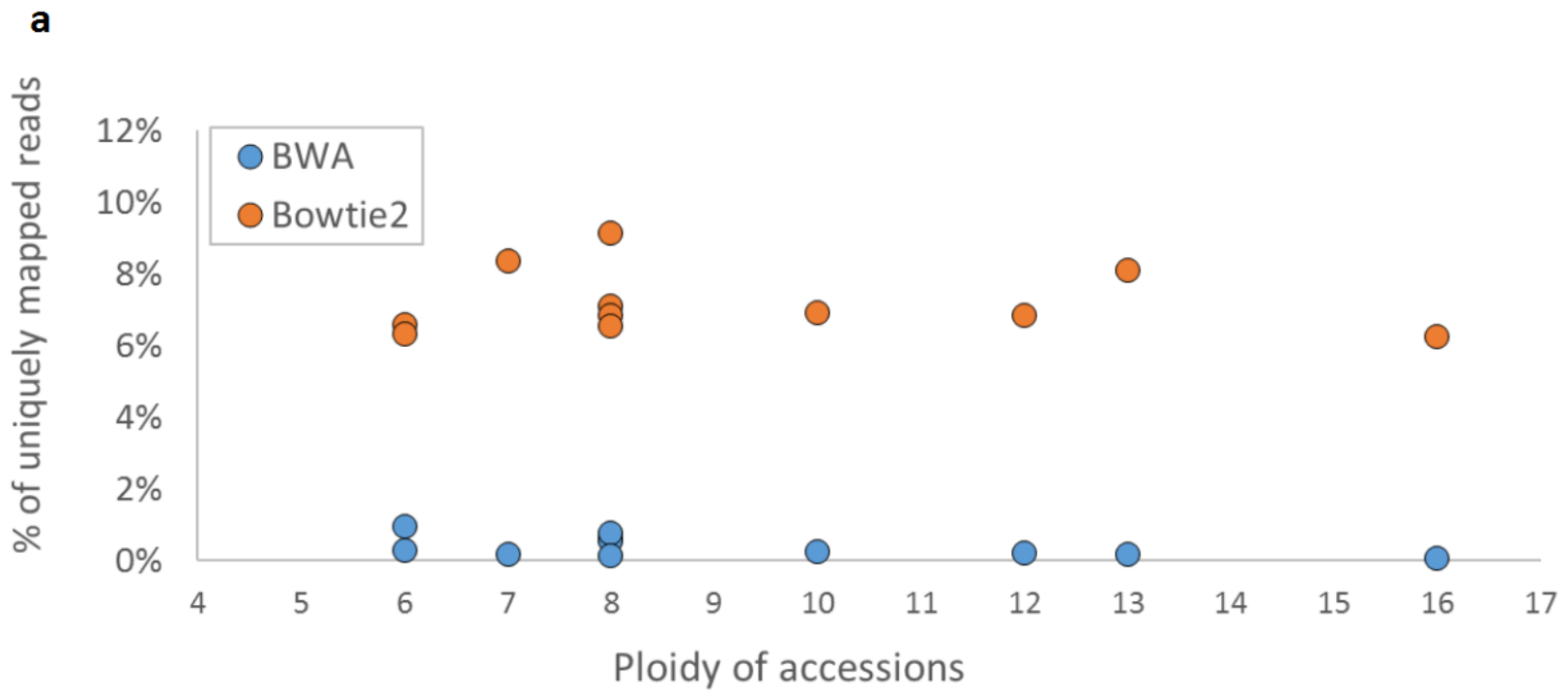
Supplementary Figure 19. Nucleotide diversity (pi) between genomic rearranged and non-rearranged regions.

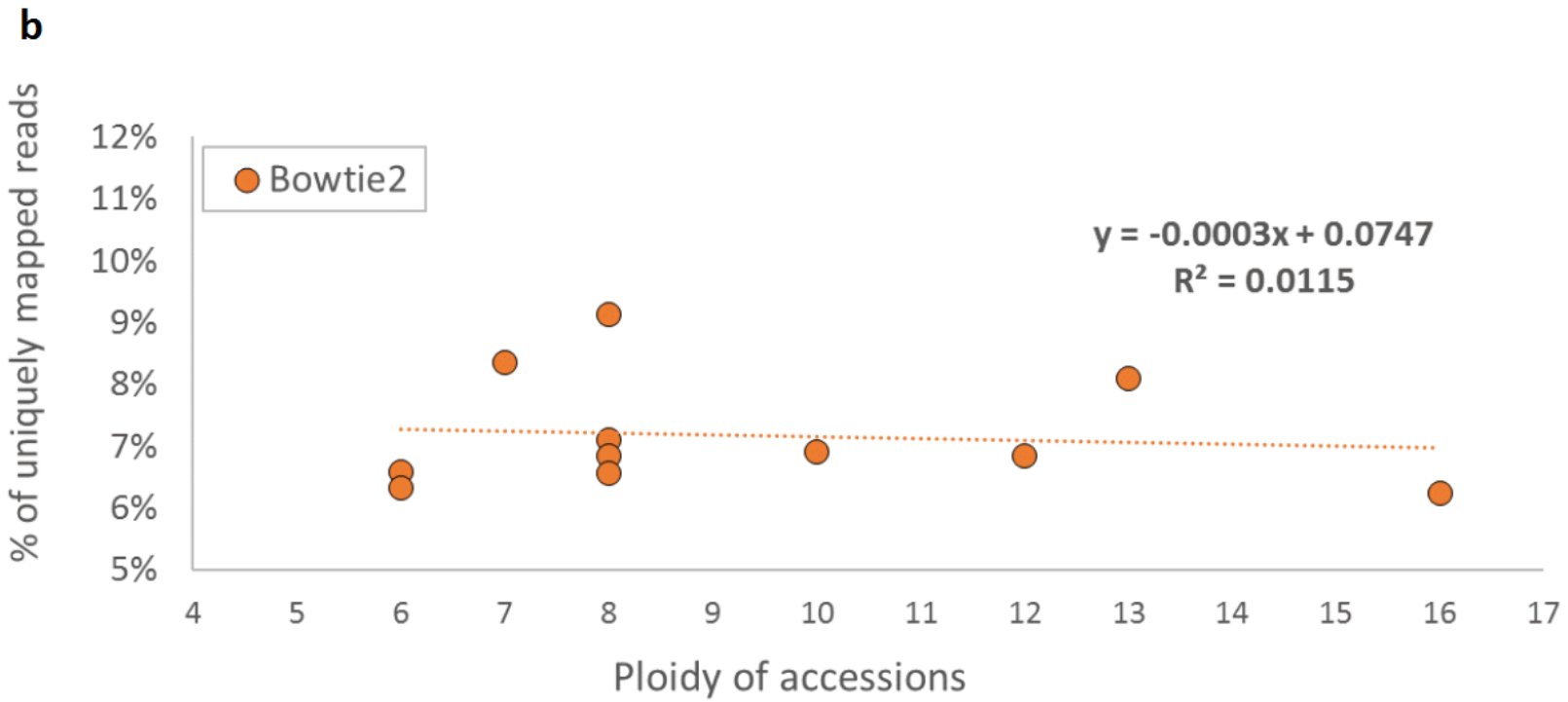
Chr2A-2D, a1-a4; Chr5A-5D, b1-b4; Chr6A-6D, c1-c4; Chr7A-7D, d1-d4. The red rectangles show the genomic rearranged regions on each homologous chromosome.

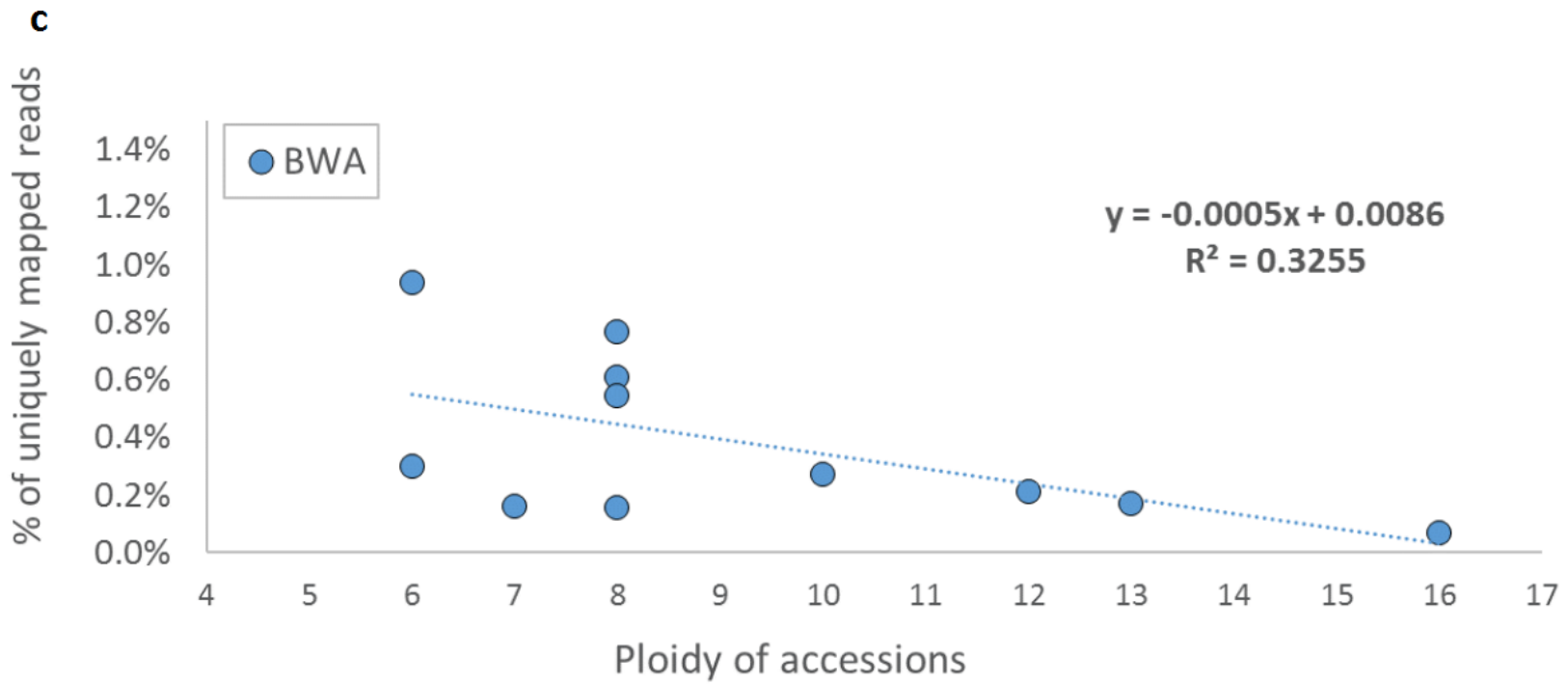


Supplementary Figure 20. Hi-C link size distribution based on all intra-contig links in the draft AP85 genome.

Given the size of the Hi-C link, we computed the density (number of links per base pair) of all Hi-C links of a given size. The link size distribution provides a basis for probabilistic inference of the relative confidence of contig ordering and orientations.


a
% of uniquely mapped reads
12%
10%
8%
6%
4%
2%
0%
BWA
Bowtie2
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Ploidy of accessions


b
% of uniquely mapped reads
12%
11%
10%
9%
8%
7%
6%
5%
Bowtie2
y = -0.0003x + 0.0747
R² = 0.0115
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Ploidy of accessions



Supplementary Figure 21. Comparison of uniquely mapped reads using BWA and Bowtie2 for ten *S. spontaneum* accessions with different ploidy.

(a) Percentages of uniquely mapped reads (UMR) using both BWA and Bowtie2; (b) Fitted curve shows the percentages of UMR decrease slowly with increasing of ploidy using Bowtie2 mapping; (c) Fitted curve shows the percentages of UMR decrease quickly with increasing of ploidy using BWA mapping.