Received: 18 May 2021 | Revised: 19 August 2021 | Accepted: 9 September 2021

DOI: 10.1111/1755-0998.13507

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

# High quality haplotype-resolved genome assemblies of *Populus tomentosa* Carr., a stabilized interspecific hybrid species widespread in Asia

Xinmin An[1,2,3] | Kai Gao[2,3] | Zhong Chen[1,2,3] | Juan Li[2,3] | Xiong Yang[2,3] | Xiaoyu Yang[2,3] | Jing Zhou[2,3] | Ting Guo[2,3] | Tianyun Zhao[2,3] | Sai Huang[2,3] | Deyu Miao[2,3] | Wasif Ullah Khan[2,3] | Pian Rao[2,3] | Meixia Ye[2,3] | Bingqi Lei[2,3] | Weihua Liao[2,3] | Jia Wang[2,3] | Lexiang Ji[2,3] | Ying Li[2,3] | Bin Guo[2,3,4] | Nada Siddig Mustafa[2,3] | Shanwen Li[5] | Quanzheng Yun[6] | Stephen R. Keller[7] | Jian-Feng Mao[1,2,3] | Ren-Gang Zhang[6] | Steven H. Strauss[8]

[1]Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, Beijing Forestry University, Beijing, China

[2]National Engineering Laboratory for Tree Breeding, College of Biological Sciences and Technology, Beijing Forestry University, Beijing, China

[3]Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants, MOE, College of Biological Sciences and Technology, Beijing Forestry University, Beijing, China

[4]Shanxi Academy of Forestry, Taiyuan, China

[5]Shandong Academy of Forestry, Jinan, China

[6]Ori-Gene Technology Co., Ltd., Beijing, China

[7]Department of Plant Biology, University of Vermont, Burlington, Vermont, USA

[8]Department of Forest Ecosystems and Society, Oregon State University, Corvallis, Oregon, USA

**Correspondence**

Xinmin An and Jian-Feng Mao, Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, Beijing Forestry University, Beijing, China.
Email: anxinmin@bjfu.edu.cn; jianfeng.mao@bjfu.edu.cn

Ren-Gang Zhang, Ori-Gene Technology Co., Ltd. Beijing, China.
Email: zhangrengang@ori-gene.cn

Steven H. Strauss, Department of Forest Ecosystems and Society, Oregon State University, Corvallis, Oregon, USA.
Email: steve.strauss@oregonstate.edu

**Funding information**

The China National Key R&D Program during the 14th Five-year Plan Period "Forest Tree Genome Editing Technology"; National Science and Technology Major Project of China, Grant/Award Number: 2018ZX08021002-002-004; National Natrual Science Foundation of China, Grant/Award Number: 31570661 and 31170631

## Abstract

*Populus* has a wide ecogeographical range spanning the Northern Hemisphere, and interspecific hybrids are common. *Populus tomentosa* Carr. is widely distributed and cultivated in the eastern region of Asia, where it plays multiple important roles in forestry, agriculture, conservation, and urban horticulture. Reference genomes are available for several *Populus* species, however, our goals were to produce a very high quality de novo chromosome-level genome assembly in *P. tomentosa* genome that could serve as a reference for evolutionary and ecological studies of hybrid speciation throughout the genus. Here, combining long-read sequencing and Hi-C scaffolding, we present a high-quality, haplotype-resolved genome assembly. The genome size was 740.2 Mb, with a contig N50 size of 5.47 Mb and a scaffold N50 size of 46.68 Mb, consisting of 38 chromosomes, as expected with the known diploid chromosome number ($2n = 2x = 38$). A total of 59,124 protein-coding genes were identified. Phylogenomic analyses revealed that *P. tomentosa* is comprised of two distinct subgenomes, which we deomonstrate is likely to have resulted from hybridization between *Populus adenopoda* as the female parent and *Populus alba* var. *pyramidalis* as the male parent, with an origin of approximately 3.93 Ma. Although highly colinear,

Xinmin An, Kai Gao and Zhong Chen contributed equally to this work.

significant structural variation was found between the two subgenomes. Our study provides a valuable resource for ecological genetics and forest biotechnology.

KEYWORDS
forest biotechnology, haplotype-resolved genome assembly, hybridization, PacBio long-read sequencing, *Populus tomentosa*

## 1 | INTRODUCTION

The genomics revolution has spurred unprecedented growth in the sequencing and assembly of whole genomes in a wide variety of model and non-model organisms (Ellegren, 2014). While this has fueled the development of large genomic diversity panels for studies into the genetic basis of adaptive traits, reliance on a single well-assembled reference genome within a species or across a set of closely related congeners poses significant limitations on genetic and evolutionary inferences (Sherman & Salzberg, 2020). The challenge is particularly acute when working with large, structurally diverse, hybrid or heterozygous genomes, for which low coverage and biases in variant calling may result when mapping short read sequences against a divergent reference genome. Such complexity challenges plant genome assembly, for which a homozygous line is desirable (Daccord et al., 2017; Wu et al., 2018). However many plants, including *Populus*, are outcrossing in nature and thus heterozygous genomic regions are ubiquitous and major contributors to phenotypic variation (Schnable & Springer, 2013). Moreover, direct sequencing of naturally heterozygous lines or artificially domesticated hybrid lines can provide deep views on their genetic complexity and evolution history (Minio et al., 2019).

The genus *Populus* (poplars, cottonwoods, and aspens) has emerged as the leading model in tree ecological genomics and biotechnology, including development of the reference genome assembly for *Populus trichocarpa*–the first tree to undergo whole genome sequencing (Tuskan et al., 2006). In recent years, the whole genomes of *P. euphratica*, *P. tremula* and *tremuloides*, *P. alba* var. *pyramidalis* and *P. alba* have also been published (Lin et al., 2018; Liu et al., 2019; Ma et al., 2019; 2013a; 2013b). However, high genetic heterozygosity and limited application of the third generation sequencing technology has limited the quality of many of these genome assemblies, which often remain highly fragmented into thousands of scaffolds (Ambardar et al., 2016).

The availability of multiple highly contiguous, well-assembled *Populus* reference genomes would greatly facilitate accurate inferences of synteny, recombination, and chromosomal origins (Lin et al., 2018). Diverse well-assembled reference genomes would also provide a fundamental tool for functional genomics, genetic engineering, and molecular breeding in this economically important genus (Zhang et al., 2019). It would also improve phylogenomic analyses of the *Populus* pan-genome (Pinosio et al., 2016; Zhang et al., 2019), without the need for reliance on reference-guided mapping and variant calling based solely on the *P. trichocarpa* reference. Recent advances in approaches to whole genome sequencing, including chromosome conformation capture (Hi-C; van Berkum et al., 2010) and long-read sequencing offer a means to go beyond fragmented draft genomes and generate nearly comprehensive de novo assemblies (El-Metwally et al., 2014).

*Populus tomentosa*, also known as Chinese white poplar, is indigenous and widely distributed across large areas of China (An et al., 2011). Moreover, it is also the first tree species planted in large-scale artificial plantations in China. Its characteristics include rapid growth, a thick and straight trunk, high environmental stress tolerance, and a long lifespan (typically 100–200 years, but sometimes over 500 years). These traits make *P. tomentosa* valuable from economic, ecological and evolutionary perspectives, with applications that include timber, pulp and paper, veneer, plywood, bioremediation, wind breaks, carbon capture, and prevention of soil erosion. Like other white poplars, *P. tomentosa* has become an important model for genetic research on trees (An et al., 2011), but at present no genome sequence is available and the origin, evolution and genetic architecture of the *P. tomentosa* genome are unclear. It has been proposed that *P. tomentosa* is a distinct species in the *Populus* section (Dickmann & Isebrands, 2001). However, the origin of *P. tomentosa* has remained controversial. Although *P. tomentosa* was proposed to contain two genetic types with different maternal parents (Wang et al., 2019), suggestions of a hybrid origin were based on a limited set of molecular markers and an incomplete collection of provenance materials. Thus, its ancestry and genome structure remains unclear. Our study adds to knowledge of the species by providing a much greater understanding of genomic architecture and structural composition following inferred interspecies hybridization.

Here, we present de novo assembles for *P. tomentosa* (clone GM15) by the combined application of PacBio, Illumina and Hi-C sequencing technologies. We herein provide two high-quality haplotype-resolved assemblies for all chromosomes whose phylogenetic affinities demonstrate the hybrid origin of this species. Combining phylogenetic analyses of chloroplast and mitochondrial genomes in this study, we deduced that the progenitors of *P. tomentosa* are *P. adenopoda* (female parent) and *P. alba* var. *pyramidalis* (male parent). Furthermore, we uncovered extensive structural variations across the genome. These findings help to elucidate the mechanisms of speciation in *Populus*, and expand our understanding of the genomic biology of *Populus*. Meanwhile, *P. tomentosa*, as a stabilized interspecific hybrid species with many good traits, it is widespread in Asia, unveiled parental origination and high quality haplotype-resolved genome will be of interest to the wider community.

## 2 | MATERIALS AND METHODS

### 2.1 | *In vitro* regeneration and validation

We collected the branches with floral buds from an elite male *P. tomentosa* clone (LM50), and water-cultured in a greenhouse. Subsequently, we performed anther-induced regeneration, and measured the ploidy of regenerated anther plantlets (referencing a previous study: [Li et al., 2013]). The nucleus DNA amount (C-value) was detected using a Cell Laboratory Quanta SC (Beckman Coulter). Chromosome counts were carried out on root-tip cells, and the micrographs were taken with an FSX-100 microscope camera system (Olympus). The leaves of *P. tomentosa* (Clone LM50) treated in the same way was used as a diploid control. Furthermore, The genotype of anther plantlets were identified using 19 allele-specific primer pairs located on each of the 19 chromosomes (Table S1).

### 2.2 | Genomic DNA library construction and sequencing

The plantlet GM15 generated by in vitro anther culture and regeneration system, was selected for genome sequencing. Genomic DNA was extracted using the Qiagen DNeasy Plant Mini Kit. DNA quality was evaluated by agarose gel electrophoresis and its quantity determined using a NanoDrop spectrophotometer (Thermo Fisher Scientific). The short-insert PCR-free genomic DNA library (300–500 bp) was constructed following the manufacturer's protocol (Illumina Inc.) for paired-end sequencing on the Illumina HiSeq X Ten sequencer. For SMRT sequencing, the 20-kb genomic DNA library was costructed following the manufacturer's protocol (Pacific Biosciences). Finally, 50 μg of high-quality genomic DNA was sequenced on the PacBio Sequel platform (Pacific Biosciences) to generate 11 SMRT cells. To compare the kmer-based genomic characteristics between GM15 and LM50, LM50 was also sequenced on the Illumina platform with the same procedures, including DNA extraction, library construction and sequencing.

### 2.3 | RNA-seq library construction and sequencing

Total RNA of roots, stems and leaves of the plantlet GM15 were extracted using the Qiagen RNeasy Plant Mini Kit (Qiagen). RNA quality was evaluated by agarose gel electrophoresis and its quantity determined using a NanoDrop spectrophotometer (Thermo Fisher Scientific). To assist gene annotation, 2 μg of total RNA from extracted tissues was used to construct the RNA-seq library and sequenced on the Illumina HiSeq X Ten platform following the protocol of manufacturer (New England Biolabs).

### 2.4 | Genome size estimation and assembly

Using the Genome Characteristic Estimation (GCE) program v1.0 (Liu et al., 2013), the genome sizes of GM15 and LM50 were estimated by 17-mer analysis based on PCR-free Illumina short reads. Based on the post-filtered PacBio reads (after filtering out reads shorter than 1000 bp), de novo assembly was conducted using an overlap-layout-consensus method in CANU v1.5 (Koren et al., 2017). Subsequently, the primary draft assembly was polished using Arrow (https://github.com/PacificBiosciences/GenomicConsensus) to improve accuracies.

### 2.5 | Hi-C library sequencing and chromosome anchoring

The Hi-C library was prepared using standard procedures (NEBNext Uitra$^{T}$II DNA Library Prep Kit for Illumina) and sequenced on the Illumina HiSeq X Ten platform. The Hi-C reads were first mapped to the above draft genome using Juicer v1.5.6 (Durand, Shamim, et al., 2016). To account for the high duplication level in *Populus*, only the aligned reads with mapping quality >40 were used to conduct the Hi-C association chromosome assembly with the 3D-DNA (v170123) pipeline with parameters "-m haploid -t 5000 -s 2" (Dudchenko et al., 2017). Visualization was carried out using Juicebox v1.6 (Durand, Robinson, et al., 2016), This yielded a chromosomes-scale draft assemby containing 38 chromosomes. Subsequently, the assembly was polished with Arrow over three iterations using PacBio reads and finally polished using Illumina short reads with Pilon v1.22 over five iterations. Finally, the assembly qualiy was comprehensively assessed by mapping Illumina, PacBio and RNA-seq data using bowtie2 v2.3.4, BLASR v5.1 (Chaisson & Tesler, 2012) and Hisat2 v2.1.0 (Kim et al., 2015), respectively. Simultaneously, the completeness of genome was also assessed using Benchmarking Universal Single-Copy Orthologues (BUSCO, v2.0.1; Simao et al., 2015).

### 2.6 | Genome annotation

Repeat families were de novo identified and classified using the RepeatModeler v1.0.8, subsequently genome was masked using RepeatMasker v4.0.7, and protein-coding genes were annotated using the MAKER2 (v2.31.9) annotation pipeline (Cantarel et al., 2008). The single copy core genes identified by BUSCO (Simao et al., 2015) were used to train the AUGUSTUS model v3.3.1, and five rounds of optimization were carried out. MAKER2 pipeline was carried out with combining ab initio prediction, EST sequence alignments and protein sequence alignments, and finally integrated these data with annotation edit distance (AED) score calculated for quality control. The completeness of the annotated proteome was assessed using BUSCO. Protein sequences were aligned with transcripts using BLAT v36 (Kent, 2002). Functional annotation was performed by aligning protein sequences with the protein databases using BLAT v36 (Kent, 2002; identity > 30%, and the E < 1e−5). The tRNA and rRNA were predicted using tRNAScan-SE v1.3.1 and RNAmmer v1.2, respectively. Other non-coding RNAs were annotated using RfamScan v1.0. Counts of transposable elements (TEs) per family were tallied from the above RepeatModeler output, and subsequently tested for differences in frequencies between the two subgenomes. We first calculated a Chi-square test of independence for the overall null hypothesis that the frequencies across the TE families were proportional between two subgenomes. We then followed up with individual tests for each TE category, testing the null hypothesis that the frequences were equal between two subgenomes.

### 2.7 | Chromosome grouping and subgenome recombination test

We first collected genome data of 4 poplars and *Salix suchowensis* (Dai et al., 2014), and de novo transcriptomes assembly of other white poplars (Table S2). Then we performed gene family clustering using OrthoMCL on protein sequences, and conducted further collinearity analysis using MCScanX (Wang et al., 2012). We identified 1052 single copy and collinear orthologous genes to construct gene trees. Subsequently, the total 38 chromosomes of *P. tomentosa* were partitioned into two subgenomes (2 × 19 chromosomes) based on phylogenic distance. As there is currently no chromosome-scale assembly of *P. alba* var. *pyramidalis* and *P. adenopoda*, we instead used a draft assembly of *P. alba* var. *pyramidalis* and transcriptome of *P. adenopoda*. Further, we slected 5345 single copy orthologous genes, which are collinear allele pairs between two subgenomes of *P. tomentosa* and are homologus to those of *P. alba* var. *pyramidalis* (PA) and *P. adenopoda* (PD), to measure their distances through *Ks*, and reconfirm that *P. tomentosa* genome was composed of subgenome A (PtA) and subgenome D (PtD).

To investigate potential recombination between homologous gene pairs of the subgenomes, we compared the synonymous substitution rates of parent and progeny alleles. Our hypothesis tests is as follows: (1) if *Ks* (PD-PtD) is smaller than *Ks* (PD-PtA), and *Ks* (PA-PtA) is smaller than *Ks* (PA-PtD), then it supports the expectation of no recombination events between the orthologous genes. (2) if *Ks* (PD-PtD) is larger than *Ks* (PD-PtA), and *Ks* (PA-PtA) is larger than *Ks* (PA-PtD), it supports the expectation of recombination events between orthologous genes. If none of the above conditions are true, it may be due to false positive homology, gene loss, imbalance of evolution rate, or other causes and was excluded from recombinant analysis.

### 2.8 | Molecular phylogenetic tree, whole-genome duplication and divergence events

Based on collinear homologous gene pairs, including interspecific orthologues and intraspecific paralogues with tandem repeats excluded, we aligned protein sequences using MUSCLE v3.8 (Edgar, 2004), then used PAL2NAL v14 to carry out codon alignment (Suyama et al., 2006). The YN model-based *Ka* and *Ks* calculation was performed using KaKs_Calculator v2.0 (Zhang et al., 2006). Finally, we constructed a molecular phylogenetic tree using RAxML (Stamatakis, 2014) based on the GAMMA+GTR model. Assuming the divergence time of *Populus* and *Salix*–48 Ma (Manchester et al., 1986) as fossil calibration, we estimated dates for divergent events of poplar species using r8s (Sanderson, 2003). We also contructed two phylogenetic trees of the chloroplast and mitochondrial genomes derived from 15 and eight white poplars, respectively.

### 2.9 | Chromosomal structure variations and GO enrichment analysis

We conducted genome-wide synteny analysis between *P. tomentosa* and *P. trichocarpa*, and subgenome synteny analysis between subgenomes A and D in *P. tomentosa* using MCScanX (Wang et al., 2012). Genome-wide structural variations (insertion, INS; deletion, DEL; inversion, INV; translocation, TRANS; copy number variation, CNV) between homologous chromosome pairs were identified using MUMmer v3.1 and SVMU (Structural Variants from MUMmer) v0.3 (Chakraborty et al., 2019). We extracted GO annotation data of genes mapping to the SV regions, and performed futher GO enrichment analysis. Annotation results were summarized through the mapping to the Plant GOSlim.

## 3 | RESULTS

### 3.1 | Ploidy determination, genotype identification and genome size estimation

To create a plant that was suitable for genome sequencing, and also was more juvenile to promote transformability and regenerability when making transgenic plants, we regenerated plantlets from anther callus of *P. tomentosa* (using a male elite clone LM50, that otherwise shows low transformation efficiency). Although from anther culture, the conservation of ploidy level of the regenerated plantlet (GM15; Figure 1a) was determined by a number of approaches as follows. Flow cytometry showed that both GM15 and LM50 were diploids (Figure 1b). It was further confirmed by chromosome counts (Figure 1c), and their genotypes appeared to be identical and heterozygous based on 19 allele-specific primers (Table S1; Figure 1d). In addtion, the k-mer frequency distributions of both GM15 and LM50 are almost the same, suggesting a genome size of approximately 800 Mb, as expected for a diploid poplar (Figure 1e). We therefore conclude that the anther regenerated clone GM15 used for seqeuencing developed from somatic cells in the anther, not from gametes; it is thus a legitimate representative of its parent *P. tomentosa* genotype, LM50.

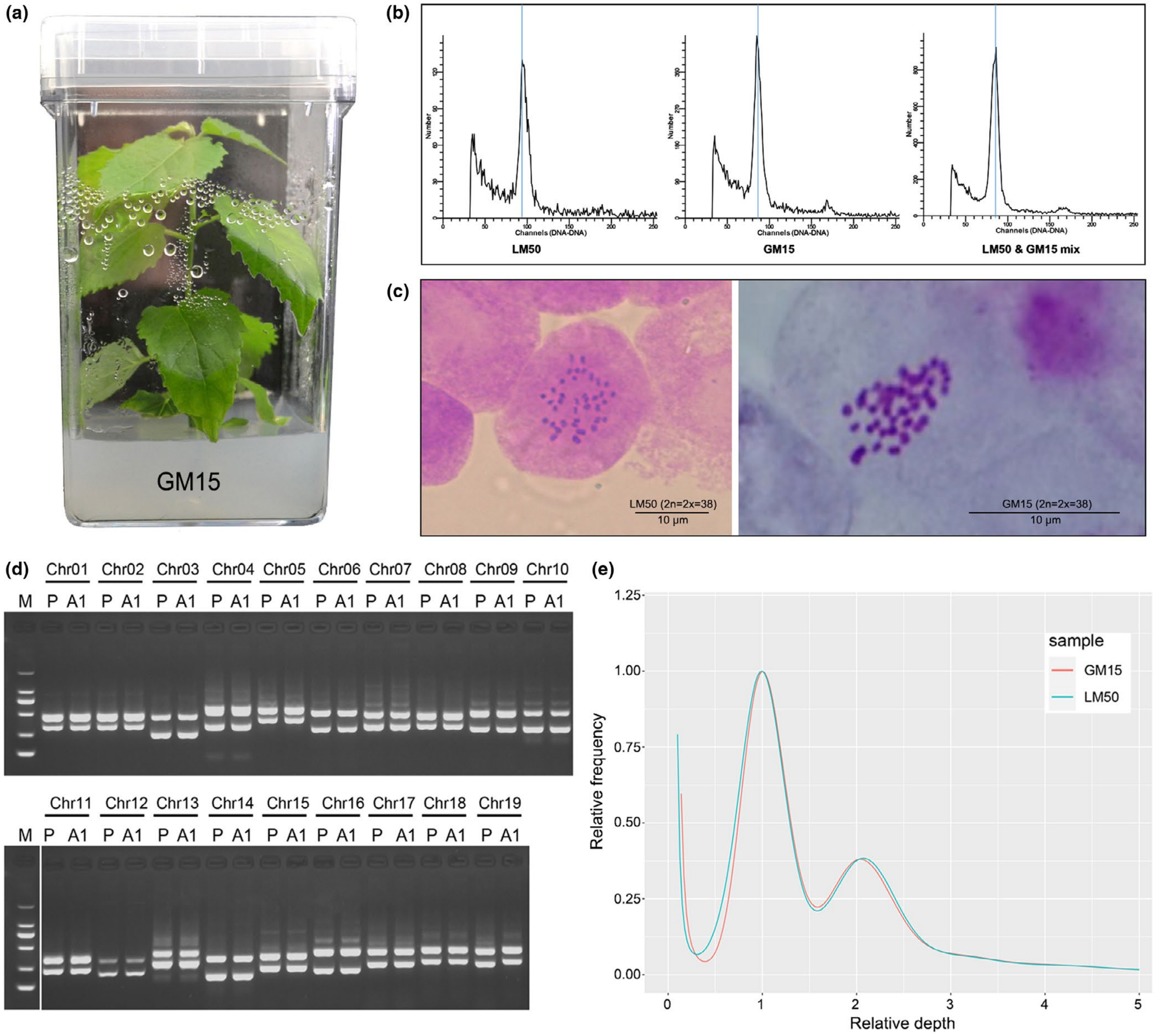

FIGURE 1 Ploidy and genotype identification of parent LM50 and its anther plant GM15. (a) The regenerated individual (GM15) from anther of *P. tomentosa* male clone LM50. (b) Ploidy detection of LM50 and GM15 by flow cytometry. (c) Chromosome counting of LM50 and GM15. (d) Genotype identification of LM50 and GM15 by PCR using allele-specific primers derived from 19 chromosomes. (M) marker DL2000, (P) Parent (male clone LM50), (A1) Anther plant GM15. (e) Normalized k-mer frequency distributions of LM50 and GM15

## 3.2 | Genome assembly and chromosome anchoring

To obtain a high-quality reference genome for *P. tomentosa*, we sequenced and assembled the genome of GM15 employing a combination of PacBio, Hi-C and Illumina methods. Its size was estimated to be ~800 Mb by K-mer analysis (Figure 1e, Table S3). A total of post-filtered ~54 Gb (6.24 million subreads, ~70× coverage) PacBio data was assembled to generate a primary draft assembly. To obtain a chromosome-scale assembly, Hi-C reads (430 million reads, 65 Gb, – 80× coverage) were sequenced and finally a total of 38 chromosome-scale pseudomolecules were successfully anchored (Figure 2, Table S4), generating a diploid genome size of 740.2 Mb. The 38 chromosome-scale pseudomolecules covered 92.1% of the estimated 800 Mb genome (Table 1). The sizes of contig N50 and scaffold N50 reached 0.96 and 17.13 Mb, with the longest contig and scaffold being 5.47, and 46.68 Mb, respectively (Table 1).

## 3.3 | Genome quality assessment, assortment and annotation

A number of other indices showed that the genome was of high quality. The Illumina and PacBio data covered 99.45 and 99.76% of the whole genomes, repectively. The mapping rate of RNA-seq data was 97.8%. In total, 96.5% of BUSCO genes were represented as

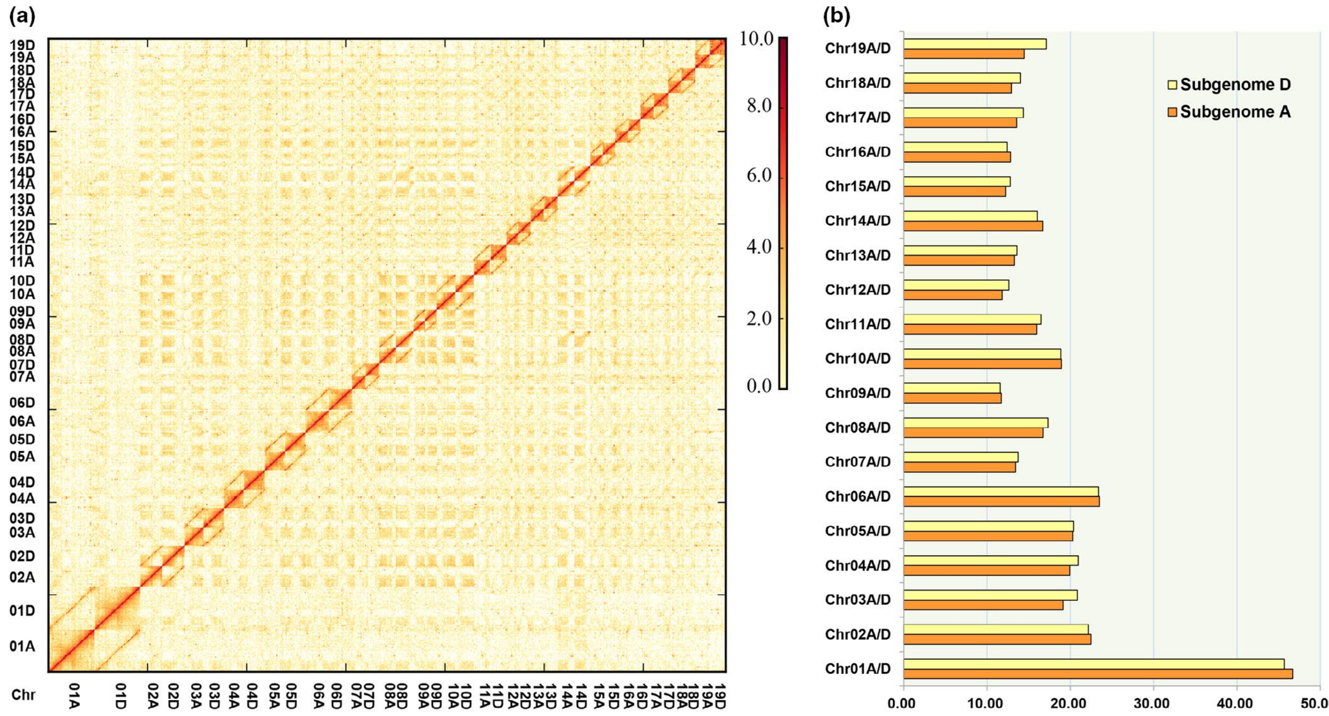

FIGURE 2 Hi-C interaction heatmap based on the chromosome-scale assembly. (a) The map represents the contact matrices generated by aligning the Hi-C data to the chromosome-scale assembly. (b) The length statistics of each chromosome for the two subgenomes resulting from the 3D-DNA pipelines

complete. The coverage depth of duplicated and single-copy BUSCO core genes was identical, showing an expected Poisson distribution (Figure S1), suggesting that the duplicated genes were not derived from assembling redundancy. In addtion, no large-scale switch errors were observed in the Hi-C heatmap, suggesting it is a well haplotype-resolved assembly.

Based on the phylogenies with other poplars, the *P. tomentosa* genome was successfully partitioned into two subgenomes (2 × 19 choromosomes), with sizes of 336.7 and 344.4 Mb, respectively (Tables 1 and S5a,b). Mapping of syntenic regions within the assembly showed clear synteny between homologous chromosome pairs and also extensive synteny among different chromosomes, as expected for the highly duplicated *Populus* genome (Figure 3). Furthermore, the coverage depth of the two subgenomes by both the PacBio and Illumina reads was also uniform, suggesting an accurate haplotype-resolved assembly (Figure S2). Compared with previous poplar genome assemblies (Ma et al., 2013a; 2013b; Tuskan et al., 2006; Yang et al., 2017), the result of comprehensive assessments showed that the *P. tomentosa* assembly quality in the present study was substantially improved (Table S6).

In total, 1,001,718 repeats were identified, and these repeats were 307.6 Mb in size and comprised ~41.6% of the genome (Figure 3a). Long-terminal repeats (LTR) were the most abundant, making up 17.5% of the genome. 13.3% of these were LTR/Gypsy elements, and 4.0% were LTR/Copia repeats. Second to LTR were unknown elements, making up 9.8% of the genome. This was followed by 5.6% of Helitron repeats and 5.4% of DNA elements (Figure 3b; details in Table S7). There were only slight differences in repeat size between two subgenomes; total repeats sizes were 133.7 and 137.9 Mb in subgenomes A and D (discussed further below), respectively, and the sizes of LTRs were 54.9 and 58.3 Mb, respectively (Table 1).

Transposable element (TE) abundance and distribution varied significantly between the subgenomes (Figure S3). All categories of the TEs were widely distributed on the two subgenomes (Figure S3 ①–⑦). For example, a total 109,542 and 114,615 TEs of Class I were found in subgenome A and subgenome D, respectively (4.6% increase), and a total 101,190 and 97,478 TEs of Class II were found in the two subgenomes (3.6% decrease), respectively (Figure S4). Further Pearson's Chi-square test showed that all TE categories but the LINE's were significant differences after Bonferonni correction ($\alpha = 0.0071$; Table S8).

Based on the repeat-masked *P. tomentosa* genome, we first annotated the protein-coding genes by incorporating evidences of 73,919 homologous protein sequences and 137,918 transcripts assembled from *P. tomentosa* RNA-seq data. A total of 59,124 protein-coding gene models were annotated, with an average coding-sequence length of 1.31 kb, 6.04 exons per gene, and 430 amino acids (aa) per protein. The gene regions covered 28.5% of the genome with a total length of 210.8 Mb (Tables 1, S9, S10). A BUSCO assessment to the proteome showed 95.8% completeness, and about 90% of the protein sequences were full-length supported by known proteins and transcripts, suggesting the high quality of the gene annotations. The annotated genes were then associated with the three onotological classes: biological process, cellular components, and molecular functions, all leading GO terms at level 2 can

**TABLE 1** Statistics for the *P. tomentosa* draft genome

| Assembly feature | Subgenome A (*P. alba* var. *pyramidalis*) | Subgenome D (*P. adenopoda*) | Genome of *P. tomentosa* |
|---|---|---|---|
| Estimated genome size by K-mer | -- | -- | 800 Mb |
| Number of contigs | 802 | 845 | 4,025 |
| Contig N50 (bp) | 994,455 | 968,830 | 964,137 |
| Longest contig (bp) | 3,787,650 | 5,467,932 | 5,467,932 bp |
| Contig N90 (bp) | 251,218 | 233,161 | 82,943 |
| Number of scaffolds | 19 | 19 | 2407 |
| Scaffold N50 (bp) | 18,914,766 | 18,843,764 | 17,128,596 |
| Longest scaffold (bp) | 46,677,810 | 45,691,089 | 46,677,810 |
| Scaffold N90 (bp) | 12,249,758 | 12,631,484 | 11,723,923 |
| Assembly length (bp) | 336,656,027 | 344,390,102 | 740,184,868 |
| GC content (% of genome) | 33.42 | 33.17 | 33.60 |
| Gap number | 783 | 826 | 1618 |
| Assembly (% of genome) | -- | -- | 92.11 |
| Repeat annotation (bp/% of assembly) | | | |
| LTR | 54,889,361/16.30 | 58,264,760/16.92 | 129,608,743/17.51 |
| Caulimovirus | 300,287/0.09 | 469786/0.14 | 849,811/0.11 |
| Copia | 13,528,294/4.02 | 13,431,562/3.90 | 29,658,574/4.00 |
| Gypsy | 40,866,300/12.14 | 44,086,909/12.80 | 98,553,170/13.31 |
| LINE | 3,312,508/0.98 | 2,828,020/0.82 | 7,766,907/1.05 |
| SINE | 1,979,481/0.59 | 1,814,936/0.53 | 3,925,279/0.53 |
| DNA | 18,854,707/5.60 | 19,016,245 /5.52 | 40,009,905/5.41 |
| RC/Helitron | 18,559,627/5.51 | 19,083,511/5.54 | 41,759,263/5.64 |
| Unknown | 30,157,396/8.96 | 30,468,761/8.85 | 72,521,254/9.80 |
| Satellite | 319,555/0.09 | 481,846/0.14 | 1,632,425/0.22 |
| Simple repeat | 4,442,224/1.32 | 4,738,265/1.38 | 9,640,201/1.30 |
| Low complexity | 1,092,089/0.32 | 1,136,945/0.33 | 2,331,509/0.31 |
| Total repeats | 133,663,317/39.70 | 137,952,318/40.06 | 310,333,451/41.93 |
| Gene annotation(counts) | | | |
| Coding gene | | | |
| Coding gene number | 28,512 | 28,605 | 59,124 |
| Coding gene number (AED < 0.5) | 27,532 | 27,604 | 57,015 |
| Average gene region length (bp) | 3,429.49 | 3,417.22 | 3,398.78 |
| Average transcript length (bp) | 1,609.1 | 1,602.21 | 1,596.97 |
| Average CDS length (bp) | 1,322.74 | 1,313.56 | 1,313.07 |
| Average exons per transcript | 5.83 | 5.84 | 5.79 |
| Average exon length (bp) | 276.11 | 274.54 | 275.86 |
| Average intron length (bp) | 75.90 | 78.32 | 83.89 |
| Non-coding gene | 1345 | 1331 | 3170 |
| tRNA number | 308 | 308 | 662 |
| rRNA number | 64 | 61 | 436 |
| Other non-coding gene number | 973 | 962 | 2072 |
| Total gene number | 29,857 | 29,936 | 62,294 |

be categorized into 24 groups (Figure 3c). The maximum function annotation ratio with protein databases is 98.6% (Table S11). We predicted 662 tRNAs with a total length of 49,659 bp (average length per tRNA: 75 bp), and 436 rRNAs (106 28S rRNAs, 106 18S rRNAs, and 224 5S rRNAs) with a length of 610,293 bp. We also annotated 2,072 other non-coding RNAs with a total length of 218,117 bp.

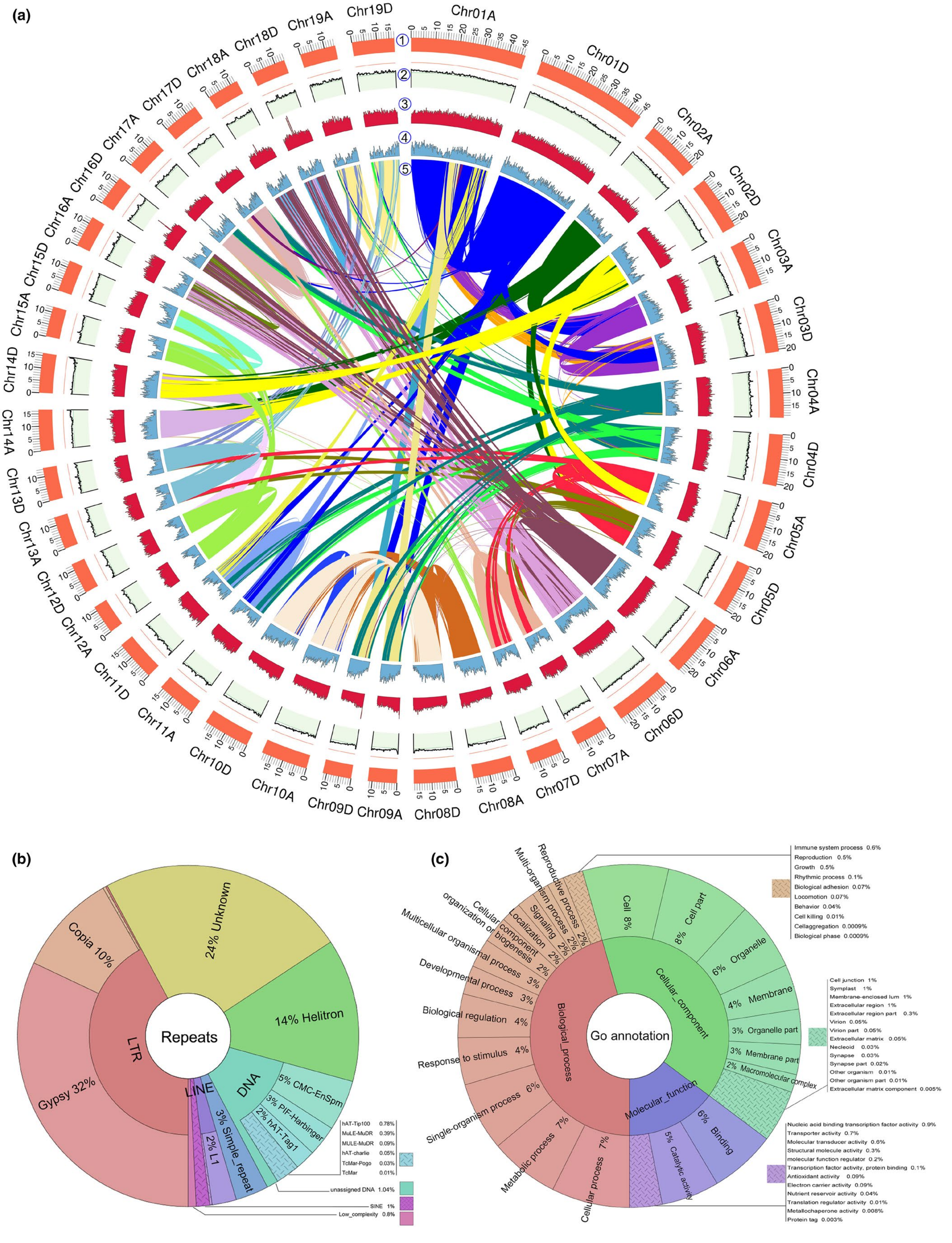
(a)
Chr01A
Chr01D
Chr02A
Chr02D
Chr03A
Chr03D
Chr04A
Chr04D
Chr05A
Chr05D
Chr06A
Chr06D
Chr07A
Chr07D
Chr08A
Chr08D
Chr09A
Chr09D
Chr10A
Chr10D
Chr11A
Chr11D
Chr12A
Chr12D
Chr13A
Chr13D
Chr14A
Chr14D
Chr15A
Chr15D
Chr16A
Chr16D
Chr17A
Chr17D
Chr18A
Chr18D
Chr19A
Chr19D
①
②
③
④
⑤
(b)
Repeats
LTR
Gypsy 32%
Copia 10%
24% Unknown
14% Helitron
DNA
5% CMC-EnSpm
3% PIF-Harbinger
2% hAT-Tag1
LINE
3% Simple_repeat
2% L1
hAT-Tip100 0.78%
MuLE-MuDR 0.39%
MULE-MuDR 0.09%
hAT-charlie 0.05%
TcMar-Pogo 0.03%
TcMar 0.01%
unassigned DNA 1.04%
SINE 1%
Low_complexity 0.8%
(c)
Go annotation
Biological_process
Cellular_component
Molecular_function
Cell 8%
8% Cell part
6% Organelle
4% Membrane
3% Organelle part
3% Membrane part
2% Macromolecular complex
6% Binding
5% Catalytic activity
Cellular process 7%
Metabolic process 7%
Single-organism process 6%
Response to stimulus 4%
Biological regulation 4%
Developmental process 3%
Multicellular organismal process 3%
Cellular component organization or biogenesis 2%
Localization 2%
Signaling 2%
Multi-organism process 2%
Reproductive process 2%
Immune system process 0.6%
Reproduction 0.5%
Growth 0.5%
Rhythmic process 0.1%
Biological adhesion 0.07%
Locomotion 0.07%
Behavior 0.04%
Cell killing 0.01%
Cellaggregation 0.0009%
Biological phase 0.0009%
Cell junction 1%
Symplast 1%
Membrane-enclosed lum 1%
Extracellular region 1%
Extracellular region part 0.3%
Virion 0.05%
Virion part 0.05%
Extracellular matrix 0.05%
Necleoid 0.03%
Synapse 0.03%
Synapse part 0.02%
Other organism 0.01%
Other organism part 0.01%
Extracellular matrix component 0.005%
Nucleic acid binding transcription factor activity 0.9%
Transporter activity 0.7%
Molecular transducer activity 0.6%
Structural molecule activity 0.3%
molecular function regulator 0.2%
Transcription factor activity, protein binding 0.1%
Antioxidant activity 0.09%
Electron carrier activity 0.09%
Nutrient reservoir activity 0.04%
Translation regulator activity 0.01%
Metallochaperone activity 0.008%
Protein tag 0.003%

**FIGURE 3** Characterization of the *Populus tomentosa* genome. (a) *Populus tomentosa* genome overview. Genome features in 200 kb intervals across the 38 chromosomes. Units on the circumference show megabase values and chromosomes. ① Choromosome karyotype. ② GC content (33.6%. red line 50%, green line 30%). ③ Repeat coverage (45–1,937 repeats). ④ Gene density (3–164 genes). ⑤ Homologous syntenic blocks. (b) Distribution of repeat classes in the *P. tomentosa* genome. (c) Distribution of predicted genes among different high-level gene ontology (GO) biological process terms

## 3.4 | Comparative genomics and evolution

We compared 19,594 gene families, cointaining 59,124 genes, in the *P. tomentosa* genome with those of other three sequenced poplar genomes including *P. trichocarpa*, *P. euphratica*, and *P. pruinosa*. A total of 22,386 gene families (142,738 genes) were identified. In addition, 14,738 gene families (119,375 genes) were shared by all four poplar species, and 1,154 gene families consisting of 2,038 genes were found to be unique to *P. tomentosa* based on OrthoMCL "mutual optimization." Similarly, 646/1,349, 179/261, and 399/1,041 gene families/genes were found to be unique to *P. trichocarpa*, *P. euphratica* and *P. pruinosa*, respectively (Figure 4a, Table S12).

To phase the chromosome pairs and study the parental origin of *P. tomentosa*, we selected 1,052 orthologous genes that appear to be allelic between each *P. tomentosa* chromosme pair and are single-copy genes in poplars of other sections, and then constructed gene trees to assess phylogenetic distances. The allelic gene-pairs of each *P. tomentosa* chromsosme pair were observed to be clearly closest to either *P. alba* var. *pyramidalis* (PA) or *P. adenopoda* (PD), respectively, on most gene trees. Thus, it sucessfully divided a total of 38 chromosomes of *P. tomentosa* into two subgenomes (2 × 19 chromosomes) based on phylogenetic distances. To confirm the results, we measured *Ks* distances among two subgenomes of *P. tomentosa*, *P. alba* var. *pyramidalis* (PA) and *P. adenopoda* (PD) using 5345 single-copy orthologous genes. The results were also consistent; we refer to the genome of *P. tomentosa* as comprised of subgenome A (putatively derived from *P. alba* var. *pyramidalis*) and subgenome D (putatively derived from *P. adenopoda*).

To investigate potential recombination events between the two subgenomes, the synonymous (*Ks*) distance between 5345 single copy orthologus genes of *P. tomentosa*, *P. alba* var. *pyramidalis* and *P. adenopoda* was estimated. We found that there was limited apparent recombination events within the large majority of gene loci (4309: 80.62%), though a low level of recombination appeared to occur (38 loci, 0.87%); and 998 loci (18.7%) did not meet either of above two hypotheses and thus were uninformative (Table S13, Figure S5). This suggests that the two parental subgenomes may be largely still intact in *P. tomentosa*, at least with respect to genic composition.

We reconstructed phylogenetic trees of subgenome A, subgenome D and of other poplars (Figure S6), as well as for each of the corresponding 19 pairs of chromosomes (Figure S7). All of these analyses supported the hypothesis that the *P. tomentosa* genome originated from hybridization between *P. adenopoda* and *P. alba* var. *pyramidalis*. Based on the fact that the *P. alba* var. *pyramidalis* is a male clone, and no female clone are found, together with phylogenetic analyses of both chloroplast and mitochondrial genomes from section *Populus* (Figures S8–S9), which indicated that *P. adenopoda* is the maternal parent of *P. tomentosa*, we deduce that *P. alba* var. *pyramidalis* and *P. adenopoda* were the male and female parents, respectively, in the hybrid formation of *P. tomentosa*.

To address dates of divergence and duplication events in poplars, we conducted collinearity analysis of homologous gene pairs derived from *Populus* species vs. *Salix suchowensis* using MCScanX (Wang et al., 2012). From the *Ks* (synonymous substitution rate) distribution, we inferred a whole genome duplication event (WGD; based on paralogous pairs) and a species divergence event (based on orthologous pairs). The *Ks* distribution among syntenic genes of the four poplar species and *S. suchowensis* contained two peaks. One peak indicated that both poplar and *Salix* species underwent a common WGD event (*Ks* ≈ 0.25). Such WGD events are known to have occurred frequently in the evolution of angiosperms (Jiao et al., 2011; Myburg et al., 2014; Otto, 2007; Van de Peer et al., 2017). This result is also consistent with a previous study on *Salix suchowensis* (Dai et al., 2014). Another peak that represents divergence between *Populus* and *Salix* is also visible (*Ks* ≈ 0.12; Figure 4b). Further analysis showed that section *Populus* and *P. trichocarpa* have a divergence at *Ks* ≈ 0.035, and *P. adenopoda* and *P. alba* at *Ks* ≈ 0.025. Subsequently, as a variant, *P. alba* var. *pyramidalis* is separated from *P. alba* at *Ks* ≈ 0.008. The hybridization event between *P. adenopoda* and *P. alba* var. *pyramidalis* subsequently occurred, followed by the emergence of *P. tomentosa* (*Ks* ≈ 0.005; Figure 4c).

To study the parental origin of *P. tomentosa*, we constructed phylogenetic trees using *Salix suchowensis* as an outgroup. Phylogenetic analysis indicated that the divergence event between section *Populus* and section *Tacamahaca* (*P. trichocarpa*) occurred at approximately 13.4 Ma. *P. adenopoda*, an ancestor of *P. tomentosa*, was the first to separate from the *Populus* family as an independent clade at approximately 9.3 Ma. Subsequently, the aspen group and white poplars group underwent a divergence event (approximately 8.4 Ma). Another ancestor of *P. tomentosa*, *P. alba* var. *pyramidalis*, gave rise to an independent variant of *P. alba* at approximately 4.8 Ma. Approximately 3.9 Ma, *P. tomentosa* was created by hybridization between *P. adenopoda* (female) and *P. alba* var. *Pyramidalis* (male; Figure 4d). Further phylogenetic trees constructed using both chloroplast and mitochondria genomes of white poplar species and *P. trichocapa* supported that the most probable female parent of *P. tomentosa* is *P. adenopoda* (Figure S8 and S9).

Whole-genome synteny analysis revealed pairs of *P. trichocarpa* -homologous regions shared between chromosomes corresponding to the two subgenomes of *P. tomentosa*. A dot plot (Figure 4e) indicated that most of the common linear segments of homologous chromosomes were shared between *P. trichocarpa*, subgenome A and subgenome D. The diagonal distribution ("/") indicated orthologous

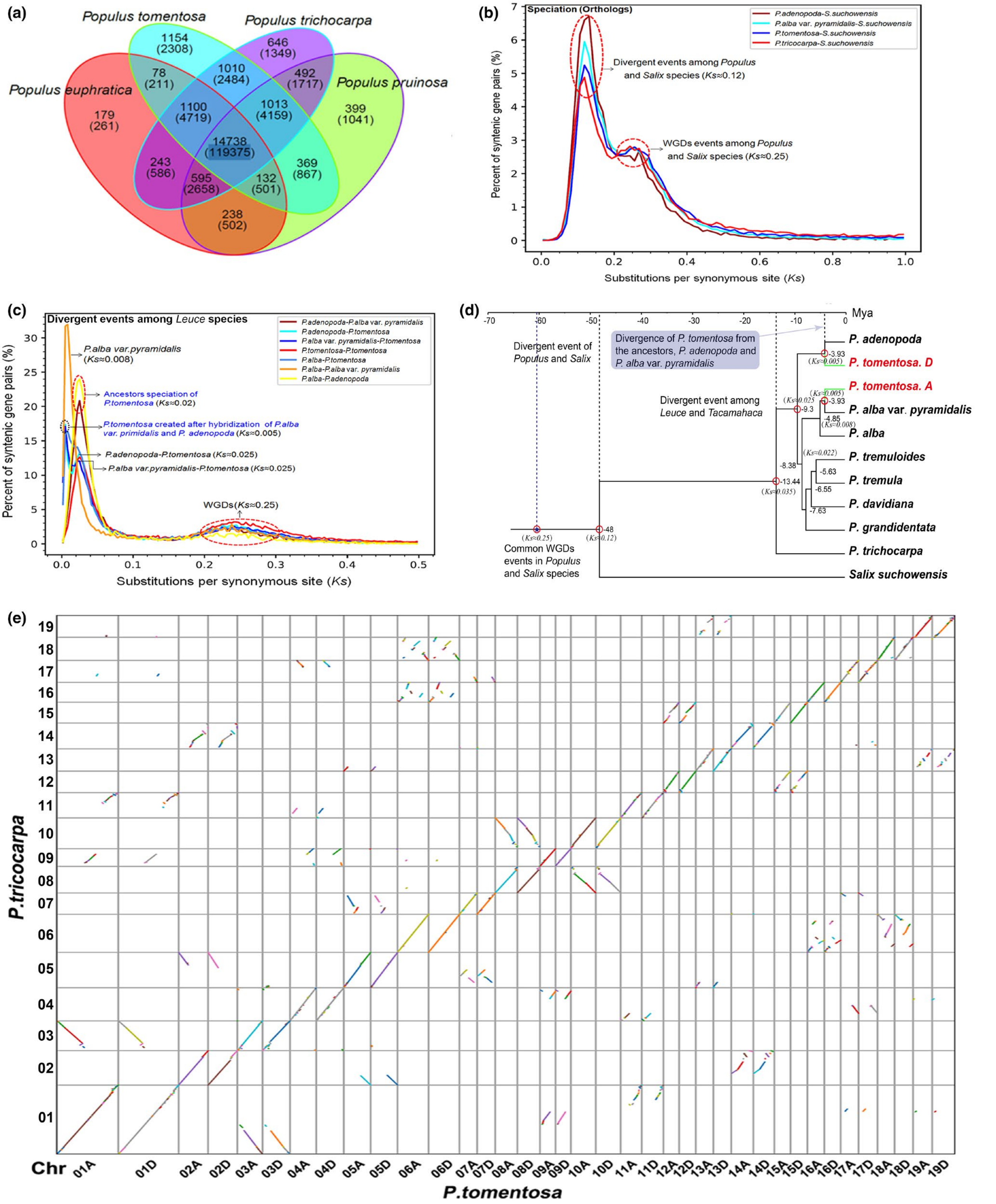

**FIGURE 4** s (a) Shared gene families among *P. tomentosa* and three other poplars. The numbers indicate the number of families and genes (within each category). (b) Interspecific divergence in *Salicaceae* species, and intraspecific divergence in *Populus* species inferred by synonymous substitution rates (*Ks*) between collinear orthologous and paralogous pairs respectively. (c) Common genome duplication events (*Ks* = 0.25, ~60 Ma) in *Salix* and *Populus* species (Dai et al., 2014; Ma et al., 2019), *P. tomentosa* speciation (*Ks* = 0.005, ~3.93 Ma) and divergence events of other poplars as revealed through *Ks* analysis. (d) Inferred phylogenetic tree across 10 plant species using r8s (Sanderson, 2003). WGD events of *Salicaceae* species are placed. (e) Synteny between the *P. tomentosa* genome (the horizontal axis) and *P. trichocarpa* genome (the vertical axis). The *P. tomentosa* chromosomes were inferred to be syntenous with *P. trichocarpa* chromosomes based on orthologous genes from OrthoMCL analysis

collinear genes in *P. tomentosa* and *P. trichocarpa*, and other dispersed distribution-blocks in the dot plot, suggested the collinearity of paralogous genes on nonhomologous chromosomes between the two poplars (Figure 4e). These findings show that both of the *P. tomentosa* subgenomes are highly syntenic with *P. trichocarpa*.

### 3.5 | Chromosome structural variation and GO analysis

To investigate the differences between subgenome A and subgenome D, we performed synteny analysis between paralogs in the *P. tomentosa* genome. This revealed collinear in-paralogous gene pairs, and suggested general collinearity at the subgenome level, with dispersed collinear blocks among homologous and nonhomologous chromosomes (Figure 5, center). We found 65,864 paralogous gene-pairs, 1,434 collinear blocks, and 65,444 collinear gene-pairs between the two subgenomes (Table S14). We infer that these may have arisen from duplication events that occurred in *Populus* prior to its divergence as a section of *Populus*.

Genome-wide structural variation (SV), including copy number variation (CNV), deletions (DEL), insertions (INS), inversions (INV), and translocations (TRANS) among chromosome pairs (Figure 5, rings 1–5 referred to as circled numbers such as "①" hereafter), indicated that there were abundant chromosome structural variations in the *P. tomentosa* genome. Across the whole genome we detected 15,480 structural variations in total, of which INS (6654) and DEL (6231) accounted for the majority (83%). The other variant numbers were 1602 and 694, and 299 for INV, TRANS and CNV, respectively, which together accounted for 27% of the total number of SVs observed (Table S15). The vast majority of INS, DEL, and CNV variations occurred between homologous chromosome pairs, whereas TRANS were generally seen between nonhomologous pairs (Table S15, Figure S10). We also found that SVs impacted gene structure and amino acid sequences. For example, allele pair Potom02G0005700 and Potom02G0411400 with indel caused changes of nucleic acid and amino acid sequences (Figure S11).

We observed that a total of 299 CNVs had an irregular and sporadic distribution across the whole genome (Figure 5). Relatively, high-density CNVs were seen on Chr17A and Chr17D (0.54/Mb), Chr09A and Chr09D (0.47/Mb), whereas comparably low-density CNVs distributed on Chr06A and Chr06D (0.13/Mb), Chr13A and Chr13D (0.15/Mb), Chr07A and Chr07D (0.18/Mb; Figure 5 ②). We also noticed that most of DELs were almost evenly distributed through the whole genome, showing a slight preference for the telomere regions of Chr12A, Chr12D, Chr17A, Chr17D, Chr18A and Chr18D (Figure 5 ③). Similarly, INSs were present at high-density and showed a slight preference for telomere regions of Chr07A, Chr07D, Chr15A, Chr15D, Chr18A and Chr18D (Figure 5 ④). In contrast, INVs had a more uneven distribution across the genome (Figure 5 ⑤). INVs were more abundant on Chr01A and Chr01D, whereas their distribution was limited on other chromosomes. TRANS were very sparsely distributed on chromosomes, with only a few detected on Chr02D, Chr07D, Chr08D, Chr13D and Chr14D (Figure 5 ⑥).

We performed GO enrichment analysis for the genes located in the total 15,480 SVs region and detected 23 GO categories significantly overrepresented with respect to the whole set of genes (Figure 6). Ten of them ("motor activity," "transporter activity," "DNA binding," "transport," "metabolic process," "lysosome," "nuclear envelope," "peroxisome," "cell wall" and "extracellular region") were overrepresented in genes affected by INS, three ("chromatin binding," "translation" and "ribosome") were overrepresented in genes affected by CNV, three ("hydrolase activity," "response to biotic stimulus" and "lipid metabolic process") were overrepresented also in genes affected by both INS and TRANS, two ("cell differentiation" and "growth") were overrepresented also in genes affected by INV, two ("vacuole" and "circadian rhythm") were overrepresented also in genes affected by TRANS, one ("endosome") was overrepresented also in genes affected by both DEL and CNV, one ("carbohydrate binding") was overrepresented also in genes affected by DEL, CNV and TRANS, and one ("plasma membrane") was overrepresented also in genes affected by both CNV and TRANS. Overall, functional annotation showed enrichments associated with all of the major GO categories (Figure 6a).

To explore the biological importance of the SVs, we further annotated genes which were highly enriched in above GO categories. We found that many genes with CNV, INS and DEL regions are involved in disease-resistance and sugar metabolism pathways (Figure 6b). For examples, Potom05G0191000 and Potom05G0207500 with CNV, Potom06G0303900 and Potom01G0355800 genes with DEL, all of which encode LRR receptor-like serine/threonine-protein kinase FLS2, which may be important for disease resistance. The disease-resistant genes in INS region are mainly annotated as nitro oxide synthase, enhanced disease susceptibility 1 protein and pathogenesis related protein 1, which are involved in plant hormone signal transduction and plant-pathogen interaction. More interestingly, we found 3 copies of both Potom05G0191000 and Potom05G0207500 in subgenome *P. adenopoda*, and 11 copies of both Potom05G0191000 and Potom05G0207500 in subgenome *P. alba* var. *pyramidalis*. Previous studies in, *Glycine max* (McHale et al., 2012) also indicated that structural variations such as CNV are common in genes related to disease resistance and biological stress. More copy numbers of both Potom05G0191000 and Potom05G0207500 may help explain why the elite individual LM50 shows strong disease resistance—a trait that is known for among forest growers. Of course, this hypothesis needs functional validation.

We also found many genes involved in carbohydrate metabolism had structural variations including CNV, DEL and INS. They were, for example, as UDP-glucuronate 4-epimerase, alpha-1,4-galacturonosyltransferase, and beta-galactosidase (Figure 6b). In addition, Potom03G0262900 and Potom01G0217800 that showed INS variation were annotated as ADP sugar diphosphatase and pectinesterase, and involved ribose phosphorylation and pentose and glucuronate interconversions, respectively; they may be important for energy and growth. Finally, it well known that the existence of centromere and telomere plays an important role in

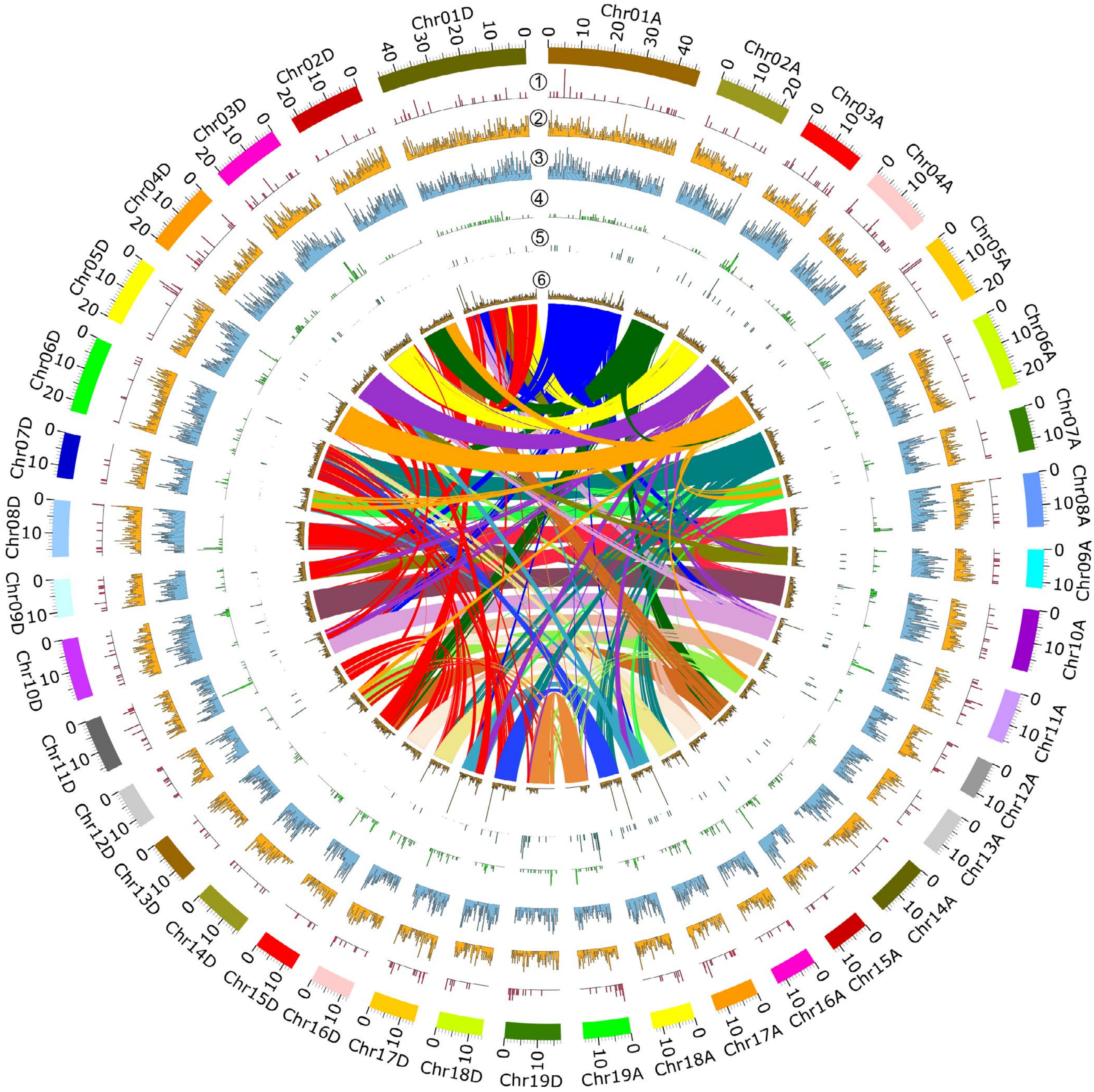

FIGURE 5 Synteny, structural variations and allele-indels analyses between subgenome A and subgenome D in *P. tomentosa*. ① Chromosome karyotype, ② Genomic distributions of copy number variations (CNV), ③ Genomic distributions of deletions (DEL), ④ Genomic distributions of insertions (INS), ⑤ Genomic distributions of inversions (INV), ⑥ Genomic distributions of translocations (TRANS), ⑦ Genomic distributions of indels between alleles of the two *P.tomentosa* subgenomes. ⑧ The inner part are synteny between subgenome A and subgenome D. The chromosomes of subgenome A were inferred to be syntenous with the chromosomes of subgenome D based on orthologous genes identified in OrthoMCL analysis

maintaining chromosome stability. Interestingly, we also found that the three genes Potom01G0282700, Potom12G0168500 and Potom12G0040500 showed INS variation, and are involved in meiotic DNA break processing and repairing, chromatin silencing at rDNA, and histone methylation. These genes may play a role in maintaining chromosome structure or reducing the rate of meiotic recombination that we observed.

## 4 | DISCUSSION

Although haploid induction was not successful, we obtained a more juvenile, easily regenerable and transformable individual GM15, which appears to be extremely similar to its parent tree LM50 based on ploidy, genotype and genome size evidences, and thus was considered suitable for sequencing. Here, we integrated advanced

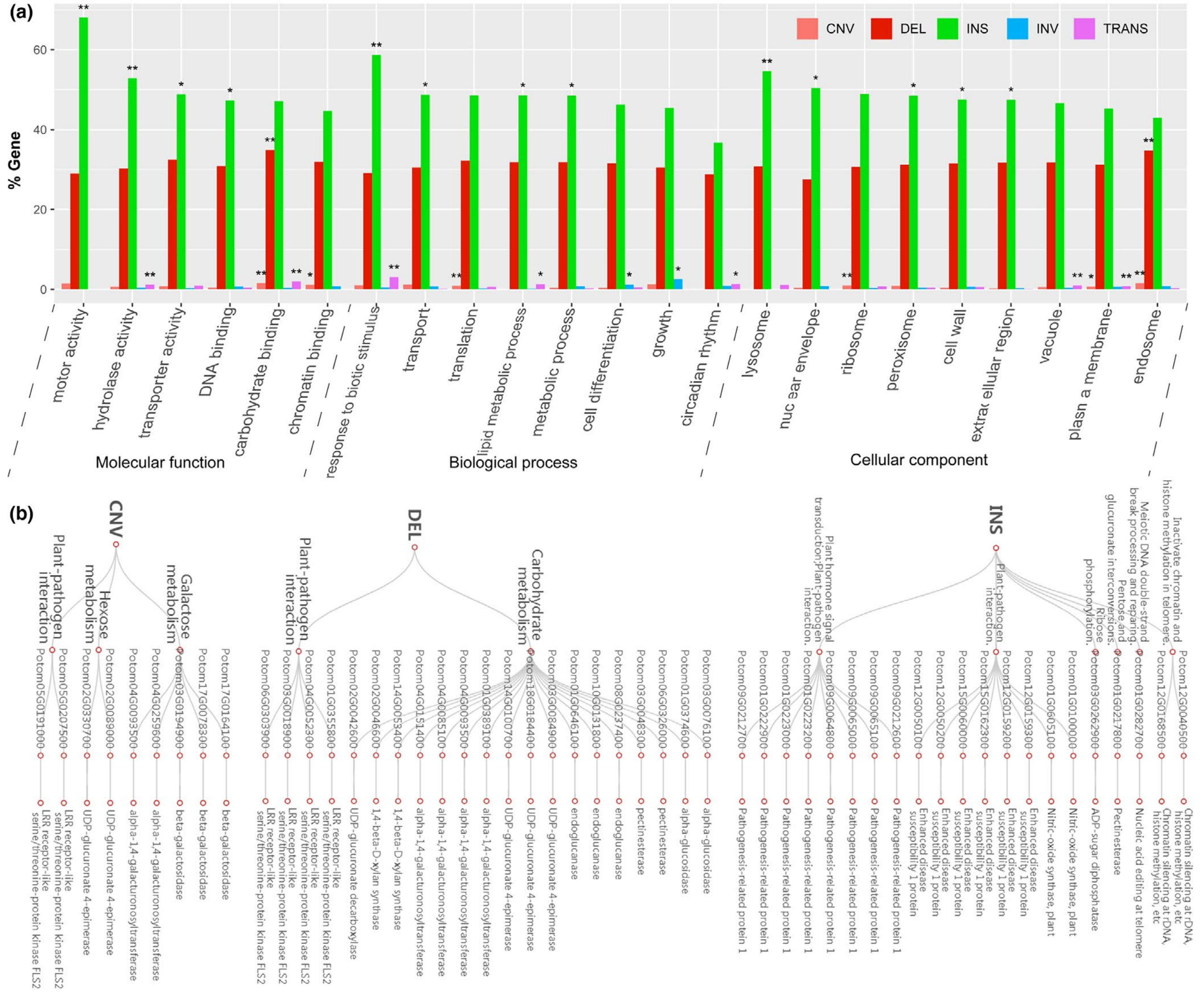

**FIGURE 6** Functional classification of GO annotations of genes associated with chromosome structural variations. (a) Frequencies of the GO terms for which an overrepresentation has been observed when comparing the subsets of genes included in copy number variations (CNV), deletions (DEL), insertions (INS), inversions (INV), and translocations (TRANS) with respect to the complete data set of *P. trichocarpa* annotated genes (ALL). *$p$-value <.05, **$p$-value <.01. (b) Structural variations, and involved pathways and important genes

SMRT sequencing technology (PacBio), Illumina correction and chromosome conformation capture (Hi-C) to assemble a high quality haplotype-resolved genome. In comparion to several published poplar genomes, including *P. trichocarpa* (Tuskan et al., 2006), *P. euphratica* (Ma et al., 2013a; 2013b), *P. pruinosa* (Yang et al., 2017), and *P. alba* var. *pyramidalis* (Ma et al., 2019), the assembly quality of *P. tomentosa* was of higher or comparable quality. Only for the *P. alba* genome was the contig N50 longer than for *P. tomentosa* (1.18 vs. 0.96 Mb); however, its contigs have not been associated with specific chromosomes yet (Liu et al., 2019; Table S7). The whole genome size of *P. tomentosa* is 740.2 Mb, which is comprised of the sum of subgenome A (*P. alba* var. *pyramidalis*) and subgenome D (*P. adenopoda*). It obviously differs with those of *P. trichocarpa* (422.9 Mb), *P. euphratica* (497.0 Mb), and *P. pruinosa* (479.3 Mb), *P. alba* var. *pyramidalis* (464.0 Mb) and *P. alba* (416.0 Mb), which respectively consist of 19 chromosomes as the allelic diversity in these diploids were subsumed into a single haploid genome rather than into two diploid subgenomes (Liu et al., 2019; Ma et al., 2013a; 2013b; Tuskan et al., 2006; Yang et al., 2017). However, this case is very similar to the genome of a hybrid poplar (84 K) recently published, which was subdivided into two subgenomes (*P. alba* and *P. tremula* var. *glandulosa*) with a total genome size of 747.5 Mb (Qiu et al., 2019; Table S7).

We presented evidence for divergence and duplication events in *Populus*, as well as within the *P. tomentosa* lineage. Like other many flowering plants (Otto, 2007), *Salicaceae* species underwent a common palaeohexaploidy event, followed by a palaeotetraploidy event before the divergence of *Salix* and *Populus* (Lin et al., 2018; Liu et al., 2019; Tuskan et al., 2006). Subsequently, poplar speciation occurred gradually. The progenitors of *P. tomentosa*, *P. adenopoda* and *P. alba* var. *pyramidalis*, successively diverged from section *Populus* approximately 9.3 Ma and 4.8 Ma. *Populus tomentosa* emerged from a hybridization event approximately 3.9 Ma. This finding differs from previous proposals on the origin of *P. tomentosa* (Wang et al., 2014). Unlike most other sequenced poplars (Ma et al., 2013a; 2013b; Tuskan et al., 2006; Yang et al., 2017), the *P. tomentosa* genome consists of subgenome A (*P. alba* var. *pramidalis*) and subgenome D (*P. adenopoda*; Figure 3 and Table 1). There also appears to be variation within *P. tomentosa* with respect to its hybrid origin. Based on a small number of marker genes, Wang et al. (2019) suggested that *P. alba* acted as the male parental species, but that the maternal parent could be either *P. adenopoda* or *P. davidiana* (for *P. tomentosa* types mb1 and mb2, respectively; Wang et al., 2019). However, *P. tomentosa* of Shandong provenance had not been collected in their experimental materials, quite coincidentally, the elite *P. tomentosa* clone LM50 in our study was from Shandong provenance, is different with *P. tomentosa* types mb1 and mb2. Thus, *P. tomentosa* may have a more complex evolutionary history than is fully understood, including possibly multiple independent origins.

Our analysis of recombination events within genes showed that the *P. tomentosa* subgenomes have largely remained independent, despite sharing the same nucleus for approximately 3.93 million years. To assess if this low rate of recombination would be expected given the time since the species' origin, we used recombination data from a recent study in the closely related European aspen (*P. tremula*; to generate an expected rate of recombination, assuming this non-hybrid species shows normal recombination rates for *Populus*). They estimated the recombination rate to be 15.6–16.1 cM/Mbp/generation (Apuli et al., 2020). In general, *P. tomentosa* has a long life cycle, the seedlings begin flowering after at least 7–8 years and thereafter annual flowering occurs during the reproductive phase (Zhu, 1992). Assuming a generation time of about 20 years, 31 recombinations per 1 kb gene would be expected—several orders of magnitude below our observation. This suggests that the two subgenomes of *P. tomentosa* have been maintained largely intact over many thousands of genertions, despite ample opportunity for recombination events to have occurred within the studied genes. The subgenome integrity of *P. tomentosa*, where there appears to be a low rate of normal meiotic products, is congruent with observations of very low fertility in the species. In a study of elite tree resourse of *P. tomentosa*, most of them showed weak fertility, a low rate of seed setting, germination and seedling surviving (Bai, 2015). Such characteristics and recent genetic analysis of *P. tomentosa* (Wang et al., 2019) suggest that *P. tomentosa* acts like the $F_1$ generation of a wide cross, with quite limited but not zero fertility.

SVs are increasingly being recognized as major factors underying phenotypic variation in eukaryotic organisms (Gabur et al., 2019). In plants, SVs have been proved to be closely related to many phenotypic variations such as of plant height (Zhou et al., 2015), and biotic stress resistance (Cook et al., 2012). In our study, we detected 15,480 SVs across the genome of GM15 of which 12,885 were INDELS and accounted for the majority of SVs (83%). GO analysis indicated INDELS are highly represented within genes with roles in plant-pathogen interaction and carbohydrate metabolism. They may therefore contribute to characteristics such as disease resistance and fast growth, for which *P. tomentosa* is well known. A few INDELS are also enriched in genes associated with meiotic DNA double-strand break processing and repair, as well as inactivation of chromation and histone methylation in telomeres. Perhaps such SVs contribute to retaining independence of the two subgenomes and maintaining karyotype stability in *P. tomentosa* —thus play a role in maintaining its putative "fixed heterosis," as discussed further below. We also found 299 CNVs, and GO analysis suggested an association with plant hormone signal transduction, plant-pathogen interaction, and sugar metablism. In sum, the many identified SVs in *P. tomentosa* provide logical focal points for study of their biological roles and phenotypic effects in relation to heterosis, evolution, breeding and biotechnology.

The mechanisms for the low recombination among subgenomes are unknown. *P. tomentosa* is well known for having low sexual fertility (Ma et al., 2013a; 2013b), probably a reflection of meiotic difficulties that give rise to abnormal gametes. As suggested for *Cucurbita* subgenomes (Sun et al., 2017), the low recombination rate in *P. tomentosa* genome could be due to the rapid divergence between the two parental species in their repetitive DNA composition, which may have inhibited meiotic pairing of homologous chromosomes and subsequent exchanges; as shown above, the transposon compositions of the two genomes differ significantly. In addition, TE activity can cause CNVs, INSs, TRANSs and DELs due to their capacity to mobilize and recombine gene sequences within and between chromosomes (Morgante et al., 2007), both in the wild and in breeding processes (Lisch, 2013). These SVs may further inhibit normal meiosis. Karyotype stability and rare recombination among subgenomes has been observed in paleo-allotetraploid *Cucurbita* genomes (Sun et al., 2017), and in newly synthesized allotetraploid wheat genome (Zhang et al., 2013). However, their functional connection to recombination rate suppression is unclear. The maintenance of subgenomes that we found in *P. tomentosa* may be advantageous in providing a degree of "fixed heterosis". This may help to explain *P. tomentosa's* high productivity and wide distribution in spite of its low sexual fertility.

## ACKNOWLEDGEMENTS

This research has been supported by the China National Key R&D Programme during the 14th Five-year Plan Period "Forest Tree Genome Editing Technology", the National Natrual Science Foundation of China (No. 31570661, 31170631), the National Science and Technology Major Project of China (No. 2018ZX08021002-002-004), the National High Technology Research and Development Programme (No. 2013AA102703) and the Major State Basic Research Development Programme (No. 2012CB114505). We thank Professor Zhiyi Zhang for his advice, also thank Shuhua Mu and Zewu An for figure editing.

## CONFLICT OF INTEREST

The authors have declared no conflict and competing interests for this article.

## AUTHOR CONTRIBUTIONS

Xinmin An designed and managed the project. Ren-Gang Zhang led the genome assembly and downstream analyses. Jianfeng Mao designed and led evolutionary analyses. Steven H. Strauss contributed to scientific analysis and interpretation, and edited the manuscript. Stephen R. Keller participated in scientific interpretation and writing the manuscript. Kai Gao and Ying Li created anther plant. Kai Gao, Zhong Chen, Juan Li, Xiong Yang, Xiaoyu Yang, Jing Zhou, Tianyun Zhao, Ting Guo, Sai Huang, Deyu Miao, Wasif Ullah Khan, Bin Guo, Shanwen Li and Nada Siddig Mustafa prepared and collected all plant materials. Jia Wang, Bingqi Lei, Weihua Liao, and Quanzheng Yun prepared RNA samples. Pian Rao, Meixia Ye, and Lexiang Ji performed transcriptome assembly and anlysis. Xinmin An, Kai Gao, and Zhong Chen wrote the manuscript with input from other authors. All authors approved the manuscript before submission.

## DATA AVAILABILITY STATEMENT

The raw reads generated in this study have been deposited in the NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) under the BioProject accession PRJNA613008. The genome assembly and annotation of *P. tomentosa* clone GM15 has been deposited at DDBJ/ENA/GenBank under the accession JAAWWB000000000. The transcriptome assemblies have been deposited at DDBJ/EMBL/GenBank under the accessions GIKW00000000 (*P. grandidentata*), GILB00000000 (*P. davidiana*), GIKX00000000 (*P. adenopoda*) and GILC00000000 (*P. alba*). The chloroplast genome assemblies also have been deposited at GenBank under accessions MW537051 (*P. alba* × *P. glandulosa*), MW537052 (*P. glandulosa*), MW537053 (*P. tremuloides*) and MK251149 (*P. tomentosa*). The GenBank accessions of mitogenomes generated in this study are MZ707540-MZ707543 (*P. adenopoda*), MZ707544-MZ707547 (*P. tomentosa*) and MZ675536 (*P. alba var. pyramidalis*).

## ORCID

*Xinmin An* https://orcid.org/0000-0001-9315-1753
*Stephen R. Keller* https://orcid.org/0000-0001-8887-9213
*Jian-Feng Mao* https://orcid.org/0000-0001-9735-8516

## REFERENCES

Ambardar, S., Gupta, R., Trakroo, D., Lal, R., & Vakhlu, J. (2016). High throughput sequencing: An overview of sequencing chemistry. *Indian Journal of Microbiology*, *56*(4), 1–11. https://doi.org/10.1007/s12088-016-0606-4

An, X. M., Wang, D. M., Wang, Z. L., Li, B., Bo, W. H., Cao, G. L., & Zhang, Z. Y. (2011). Isolation of a LEAFY homolog from *Populus tomentosa*: expression of *PtLFY* in *P. tomentosa* floral buds and *PtLFY*-IR-mediated gene silencing in tobacco (*Nicotiana tabacum*). *Plant Cell Reports*, *30*(1), 89–100. https://doi.org/10.1007/s00299-010-0947-0

Apuli, R. P., Bernhardsson, C., Schiffthaler, B., Robinson, K. M., Jansson, S., Street, N. R., & Ingvarsson, P. K. (2020). Inferring the Genomic Landscape of Recombination Rate Variation in European Aspen (*Populus tremula*). *G3 (Bethesda)*, *10*(1), 299–309. https://doi.org/10.1534/g3.119.400504

Bai, F. Y. (2015). *Evaluation of elite tree resourse and construction of parent population for breeding programme in Populus tomentosa carr.* (Master). Beijing Forestry University.

Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., & Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, *18*(1), 188–196. https://doi.org/10.1101/gr.6743907

Chaisson, M. J., & Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, *13*, 238. https://doi.org/10.1186/1471-2105-13-238

Chakraborty, M., Emerson, J. J., Macdonald, S. J., & Long, A. D. (2019). Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nature Communications*, *10*(1), 4872. https://doi.org/10.1038/s41467-019-12884-1

Cook, D. E., Lee, T. G., Guo, X., Melito, S., Wang, K., Bayless, A. M., Wang, J., Hughes, T. J., Willis, D. K., Clemente, T. E., Diers, B. W., Jiang, J., Hudson, M. E., & Bent, A. F. (2012). Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *Science*, *338*(6111), 1206–1209. https://doi.org/10.1126/science.1228746

Daccord, N., Celton, J.-M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., van de Geest, H., Bianco, L., Micheletti, D., Velasco, R., Di Pierro, E. A., Gouzy, J., Rees, D. J. G., Guérif, P., Muranty, H., Durel, C.-E., Laurens, F., Lespinasse, Y., Gaillard, S., … Bucher, E. (2017). High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nature Genetics*, *49*(7), 1099–1106. https://doi.org/10.1038/ng.3886

Dai, X., Hu, Q., Cai, Q., Feng, K., Ye, N., Tuskan, G. A., Milne, R., Chen, Y., Wan, Z., Wang, Z., Luo, W., Wang, K., Wan, D., Wang, M., Wang, J., Liu, J., & Yin, T. (2014). The willow genome and divergent evolution from poplar after the common genome duplication. *Cell Research*, *24*(10), 1274–1277. https://doi.org/10.1038/cr.2014.83

Dickmann, D. I., & Isebrands, J. G. (2001). *Poplar Culture in North America*. NRC Research Press.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., & Aiden, E. L. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*(6333), 92–95. https://doi.org/10.1126/science.aal3327

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E. L. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems*, *3*(1), 99–101. https://doi.org/10.1016/j.cels.2015.07.012

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems*, *3*(1), 95–98. https://doi.org/10.1016/j.cels.2016.07.002

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, *29*(1), 51–63. https://doi.org/10.1016/j.tree.2013.09.008

El-Metwally, S., Ouda, O. M., & Helmy, M. (2014). Next-generation sequencing platforms. *Annual Review of Analytical Chemistry*, *6*(4), 287.

Gabur, I., Chawla, H. S., Snowdon, R. J., & Parkin, I. A. P. (2019). Connecting genome structural variation with complex traits in crop plants. *TAG. Theoretical and Applied Genetics.*, *132*(3), 733–750. https://doi.org/10.1007/s00122-018-3233-0

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., Hu, Y. I., Liang, H., Soltis, P. S., Soltis, D. E., Clifton, S. W., Schlarbaum, S. E., Schuster, S. C., Ma, H., Leebens-Mack, J., & dePamphilis, C. W. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature*, *473*(7345), 97–100. https://doi.org/10.1038/nature09916

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/gr.229202

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116

Li, Y., Li, H., Chen, Z., Ji, L. X., Ye, M. X., Wang, J., & An, X. M. (2013). Haploid plants from anther cultures of poplar (*Populus* x *beijingensis*). *Plant Cell Tissue and Organ Culture*, *114*(1), 39–48. https://doi.org/10.1007/s11240-013-0303-5

Lin, Y. C., Wang, J., Delhomme, N., Schiffthaler, B., Sundstrom, G., Zuccolo, A., & Street, N. R. (2018). Functional and evolutionary genomic inferences in *Populus* through genome and population sequencing of American and European aspen. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(46), E10970–E10978. https://doi.org/10.1073/pnas.1801437115

Lisch, D. (2013). How important are transposons for plant evolution? *Nature Reviews Genetics*, *14*(1), 49–61. https://doi.org/10.1038/nrg3374

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., Li, Z., Chen, Y., Mu, D., & Fan, W. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv*. https://arxiv.org/vc/arxiv/papers/1308/1308.2012v1.pdf

Liu, Y. J., Wang, X. R., & Zeng, Q. Y. (2019). De novo assembly of white poplar genome and genetic diversity of white poplar population in Irtysh River basin in China. *Science China Life Sciences*, *62*(5), 609–618. https://doi.org/10.1007/s11427-018-9455-2

Ma, J., Wan, D., Duan, B., Bai, X., Bai, Q., Chen, N., & Ma, T. (2019). Genome sequence and genetic transformation of a widely distributed and cultivated poplar. *Plant Biotechnology Journal*, *17*(2), 451–460. https://doi.org/10.1111/pbi.12989

Ma, K., Song, Y., Huang, Z., Lin, L., Zhang, Z., & Zhang, D. (2013). The low fertility of Chinese white poplar: dynamic changes in anatomical structure, endogenous hormone concentrations, and key gene expression in the reproduction of a naturally occurring hybrid. *Plant Cell Reports*, *32*(3), 401–414. https://doi.org/10.1007/s00299-012-1373-2.

Ma, T., Wang, J., Zhou, G., Yue, Z., Hu, Q., Chen, Y., Liu, B., Qiu, Q., Wang, Z., Zhang, J., Wang, K., Jiang, D., Gou, C., Yu, L., Zhan, D., Zhou, R., Luo, W., Ma, H., Yang, Y., … Liu, J. (2013). Genomic insights into salt adaptation in a desert poplar. *Nature Communications*, *4*, 2797. https://doi.org/10.1038/Ncomms3797

Manchester, S. R., Dilcher, D. L., & Tidwell, W. D. (1986). Interconnected reproductive and vegetative remains of *Populus* (Salicaceae) FROM the Middle Eocene Green River formation, northeastern utah. *American Journal of Botany*, *73*(1), 156–160. https://doi.org/10.1002/j.1537-2197.1986.tb09691.x

McHale, L. K., Haun, W. J., Xu, W. W., Bhaskar, P. B., Anderson, J. E., Hyten, D. L., Gerhardt, D. J., Jeddeloh, J. A., & Stupar, R. M. (2012). Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiology*, *159*(4), 1295–1308. https://doi.org/10.1104/pp.112.194605

Minio, A., Massonnet, M., Figueroa-Balderas, R., Castro, A., & Cantu, D. (2019). Diploid genome assembly of the wine grape carmenere. *G3 (Bethesda)*, *9*(5), 1331–1337. https://doi.org/10.1534/g3.119.400030

Morgante, M., De Paoli, E., & Radovic, S. (2007). Transposable elements and the plant pan-genomes. *Current Opinion in Plant Biology*, *10*(2), 149–155. https://doi.org/10.1016/j.pbi.2007.02.001

Myburg, A. A., Grattapaglia, D., Tuskan, G. A., Hellsten, U., Hayes, R. D., Grimwood, J., Jenkins, J., Lindquist, E., Tice, H., Bauer, D., Goodstein, D. M., Dubchak, I., Poliakov, A., Mizrachi, E., Kullan, A. R. K., Hussey, S. G., Pinard, D., van der Merwe, K., Singh, P., … Schmutz, J. (2014). The genome of *Eucalyptus grandis*. *Nature*, *510*(7505), 356–362. https://doi.org/10.1038/nature13308

Otto, S. P. (2007). The evolutionary consequences of polyploidy. *Cell*, *131*(3), 452–462. https://doi.org/10.1016/j.cell.2007.10.022

Pinosio, S., Giacomello, S., Faivre-Rampant, P., Taylor, G., Jorge, V., Le Paslier, M. C., Zaina, G., Bastien, C., Cattonaro, F., Marroni, F., & Morgante, M. (2016). Characterization of the Poplar Pan-Genome by Genome-Wide Identification of Structural Variation. *Molecular Biology and Evolution*, *33*(10), 2706–2719. https://doi.org/10.1093/molbev/msw161

Qiu, D., Bai, S., Ma, J., Zhang, L., Shao, F., Zhang, K., Yang, Y., Sun, T., Huang, J., Zhou, Y., Galbraith, D. W., Wang, Z., & Sun, G. (2019). The genome of *Populus alba* x *Populus tremula* var. *glandulosa* clone 84K. *DNA Research*, *26*(5), 423–431. https://doi.org/10.1093/dnares/dsz020

Sanderson, M. J. (2003). r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, *19*(2), 301–302. https://doi.org/10.1093/bioinformatics/19.2.301

Schnable, P. S., & Springer, N. M. (2013). Progress toward understanding heterosis in crop plants. *Annual Review of Plant Biology*, *64*, 71–88. https://doi.org/10.1146/annurev-arplant-042110-103827

Sherman, R. M., & Salzberg, S. L. (2020). Pan-genomics in the human genome era. *Nature Reviews Genetics*, https://doi.org/10.1038/s41576-020-0210-7

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Sun, H., Wu, S., Zhang, G., Jiao, C., Guo, S., Ren, Y. I., Zhang, J., Zhang, H., Gong, G., Jia, Z., Zhang, F., Tian, J., Lucas, W. J., Doyle, J. J., Li, H., Fei, Z., & Xu, Y. (2017). Karyotype stability and unbiased fractionation in the paleo-allotetraploid *Cucurbita* genomes. *Molecular Plant*, *10*(10), 1293–1306. https://doi.org/10.1016/j.molp.2017.09.003

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*, W609–W612. https://doi.org/10.1093/nar/gkl315

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., & Rokhsar, D. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, *313*(5793), 1596–1604. https://doi.org/10.1126/science.1128691

van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., Dekker, J., & Lander, E. S. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. *Journal of Visualized Experiments*, (39), 1869. https://doi.org/10.3791/1869

Van de Peer, Y., Mizrachi, E., & Marchal, K. (2017). The evolutionary significance of polyploidy. *Nature Reviews Genetics*, *18*(7), 411–424. https://doi.org/10.1038/nrg.2017.26

Wang, D., Wang, Z., Kang, X., & Zhang, J. (2019). Genetic analysis of admixture and hybrid patterns of *Populus hopeiensis* and *P. tomentosa*. *Scientific Reports*, *9*(1), 4821. https://doi.org/10.1038/s41598-019-41320-z

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., Lee, T.-H., Jin, H., Marler, B., Guo, H., Kissinger, J. C., & Paterson, A. H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, *40*(7), e49. https://doi.org/10.1093/nar/gkr1293

Wang, Z., Du, S., Dayanandan, S., Wang, D., Zeng, Y., & Zhang, J. (2014). Phylogeny reconstruction and hybrid analysis of *Populus* (Salicaceae) based on nucleotide sequences of multiple single-copy nuclear genes and plastid fragments. *PLoS ONE*, *9*(8), e103645. https://doi.org/10.1371/journal.pone.0103645

Wu, S., Lau, K. H., Cao, Q., Hamilton, J. P., Sun, H., Zhou, C., Eserman, L., Gemenet, D. C., Olukolu, B. A., Wang, H., Crisovan, E., Godden, G. T., Jiao, C., Wang, X., Kitavi, M., Manrique-Carpintero, N., Vaillancourt, B., Wiegert-Rininger, K., Yang, X., … Fei, Z. (2018). Genome sequences of two diploid wild relatives of cultivated sweetpotato reveal targets for genetic improvement. *Nature Communications*, *9*(1), 4580. https://doi.org/10.1038/s41467-018-06983-8

Yang, W., Wang, K., Zhang, J., Ma, J., Liu, J., & Ma, T. (2017). The draft genome sequence of a desert tree *Populus pruinosa*. *Gigascience*, *6*(9), 1–7. https://doi.org/10.1093/gigascience/gix075

Zhang, H., Bian, Y., Gou, X., Dong, Y., Rustgi, S., Zhang, B., Xu, C., Li, N., Qi, B., Han, F., von Wettstein, D., & Liu, B. (2013). Intrinsic karyotype stability and gene copy number variations may have laid the foundation for tetraploid wheat formation. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(48), 19466–19471. https://doi.org/10.1073/pnas.1319598110

Zhang, L., Hu, J., Han, X., Li, J., Gao, Y., Richards, C. M., Zhang, C., Tian, Y. I., Liu, G., Gul, H., Wang, D., Tian, Y. U., Yang, C., Meng, M., Yuan, G., Kang, G., Wu, Y., Wang, K., Zhang, H., … Cong, P. (2019). A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. *Nature Communications*, *10*(1), 1494. https://doi.org/10.1038/s41467-019-09518-x

Zhang, Z., Li, J., Zhao, X. Q., Wang, J., Ka-Shu, W. G., & Yu, J. (2006). KaKs_Calculator: Calculating Ka and Ks Through Model Selection and Model Averaging. *Genomics*, *Proteomics & Bioinformatics*, *4*(4), 259–263. https://doi.org/10.1016/S1672-0229(07)60007-2

Zhou, Z., Jiang, Y. U., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., Fang, C., Shen, Y., Liu, T., Li, C., Li, Q., Wu, M., Wang, M., Wu, Y., Dong, Y., … Tian, Z. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology*, *33*(4), 408–414. https://doi.org/10.1038/nbt.3096

Zhu, Z. (1992). Collection, conservation and utilization of plus tree resource of *Populus tomentosa* in China. *Journal of Beijing Forestry University*, *14*(S3), 1–25.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** An, X., Gao, K., Chen, Z., Li, J., Yang, X., Yang, X., Zhou, J., Guo, T., Zhao, T., Huang, S., Miao, D., Ullah Khan, W., Rao, P., Ye, M., Lei, B., Liao, W., Wang, J., Ji, L., Li, Y., … Strauss, S. H. (2021). High quality haplotype-resolved genome assemblies of *Populus tomentosa* Carr., a stabilized interspecific hybrid species widespread in Asia. *Molecular Ecology Resources*, 00, 1–17. https://doi.org/10.1111/1755-0998.13507