Hybrid origin of *Populus tomentosa* Carr. identified through genome sequencing and phylogenomic analysis

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], Bing Guo[2,3,4], Nada Siddig Mustafa[2,3], Shanwen Li[5], Quanzheng Yun[6], Stephen R. Keller[7], Jianfeng Mao[1,2,3,✉], Rengang Zhang[6,✉], Steven H. Strauss[8,✉]

†: Contributed equally to this work.

✉: To whom correspondence may be addressed: X.A. (email: anxinmin@bjfu.edu.cn), J. M. (email: jianfeng.mao@bjfu.edu.cn), R. Z.(email: zhangrengang@ori-gene.cn) or S.H.S. (email : Steve.Strauss@Oregonstate.Edu)

1. Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, Beijing Forestry University, Beijing 100083, China.
2. National Engineering Laboratory for Tree Breeding, College of Biological Sciences and Technology, Beijing Forestry University, Beijing 100083, 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 100083, China.
4. Shanxi Academy of Forestry, Taiyuan 030012, China.
5. Shandong Academy of Forestry, Jinan 250014, China.
6. Ori-Gene Technology Co.,Ltd. Beijing 102206, China.
7. Department of Plant Biology, University of Vermont, 111 Jeffords Hall, Burlington, VT 05405, USA
8. Department of Forest Ecosystems and Society, Oregon State University, Corvallis, OR 97331 USA.

Abstract

Populus tomentosa is widely distributed and cultivated in the Northern and Central China, where it is of great economic and ecological importance. However, the origin of *P. tomentosa* remains controversial. Here, we used a PacBio+Hi-C+Illumina strategy to sequence and assemble its 740.2 Mb (2n) genome. The assembly accounts for greater than 92.1% of the 800-megabase genome, comprises 38 chromosomes, and contains 59,124 annotated protein-coding genes. Phylogenomic analyses elucidated dynamic genome evolution events among its closely related white poplars, and revealed that *tomentosa* is comprised of two subgenomes, which we deomonstrate is likely to have resulted from hybridization between *Populus adenopoda* as the female, and *Populus alba* var. *pyramidalis* as the male, around 3.93 Mya. We also detected structural variations and allele-indels across genome. Our study presents a high quality and well assembled genome, unveils the origin of the widely distributed and planted *P. tomentosa*, and provides a powerful resource for comparative plant biology, breeding, and biotechnology.

Key words: *Populus tomentosa*; PacBio long-read sequencing; Genome assembly; Hybridization; Forest biotechnology.

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 (Ballouz et al., 2019; Sherman and 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.

Long-lived perennial forest trees present unique challenges and opportunities for evolutionary genomics, due to abundant structural and nucleotide diversity within their genomes, even among close congeners. 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 *Populus euphratica*, *Populus pruinosa*, *Populus tremula* and *Populus tremuloides*, and hybrid 84K (*P. alba* x *P. tremula var. glandulosa*) have also been published (Lin et al., 2018; Ma et al., 2013; Qiu et al., 2019; Yang et al., 2017). However, high genetic heterozygosity and short read lengths limit the quality of these genome assemblies, which 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; Mardis, 2013).

Populus tomentosa, also known as Chinese white poplar, is indigenous and widely distributed across large areas of China (Gao et al., 2019). Moreover, it is also the first tree species planted in large-scale artificial plantations in China. Like other white poplars, *P. tomentosa* has become a useful model for genetic research on trees (An et al., 2011; Chen et al., 2018; Wang et al., 2014), 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 as another species in the *Leuce* section, but is now considered a hybrid of *Populus alba* and *Populus adenopoda* or a tri-hybrid of the previous two taxa with *Populus tremula* (Dickmann and Isebrands, 2001). In recent years, similar conclusions were reached using limited molecular markers (He, 2005; Wang et al., 2014). However, there is as yet a lack of strong genomic evidence to support this hypothesis.

Here, we present *de novo* assembles for *P. tomentosa* (clone GM15) by the combined application of PacBio, Illumina and Hi-C sequencing platforms. We herein provide two high quality assemblies for all chromosomes whose phylogenetic affinities demonstrate the hybrid origin of this species. From previous phylogenetic analysis on chloroplast genomes of seven *Leuce* poplars (Gao et al., 2019), we deduced that the ancestors of *P. tomentosa* are *P. adenopoda* (female parent) and *P.alba* var. pyramidalis (male parent). This contrasts with previous suggestions that the male parent is *P.alba* (Dickmann and Isebrands, 2001; Wang et al., 2014). Furthermore, we uncovered chromosome structural variations and allele-indels across *P. tomentosa* genome. These findings will help to elucidate the mechanisms of speciation in *Populus*, as well as the role of artificial and natural selection in poplar genome evolution. This study provides the first highly contiguous assembly of a *Populus* genome since publication of the original *P. trichocarpa* reference (Tuskan et al. 2006), expands our understanding of the unique biology of large woody perennials and provides a powerful tool for comparative biology, breeding and biotechnology.

Results

Genome Assembly, Anchoring and Annotation. To obtain a high-quality reference genome for *P. tomentosa* "LM50 (male clone)," we cultivated anthers (Li et al., 2013) and regenerated a plantlet (GM15) from which DNA was extracted for sequencing. We sequenced and assembled the genome through a combination of PacBio, Hi-C and Illumina methods. Using the PacBio platform (Chin et al., 2013), we constructed 20-Kb libraries to sequence the genome, generating 11 cells There were 6.24 million PacBio single-molecule reads (~54 Gb), corresponding to ~70 ×coverage of the diploid genome, whose size was estimated to be ~800 Mb by K-mer analysis (Supplementary Fig. S1 and Table S1). We then performed PacBio assembly using an overlap-layout-consensus method implemented in CANU (Koren et al., 2017; Wang et al., 2018a). The primary assembly was further refined using chromosome contact maps. Hi-C (*in vivo* fixation of chromosomes) technology (Dudchenko et al., 2017; Durand et al., 2016a; Putnam et al., 2016) was employed to construct chromosomes, then the assembly was corrected by genome data of *P. tomentosa* generated from an Illumina (HiSeq X-10) platform with Pilon over five iterations (Walker et al., 2014) (Table S2). A total of 38 chromosome-scale pseudomolecules were successfully constructed (Fig.S2, Table S3).

Simultaneously, transcripts derived from several white poplar species, including *P. alba*, *P. adenopoda*, *P. davidiana* (NCBI SRA), and *P. grandidentata* (NCBI SRA)—together with genomes derived from *P. alba* var. pyramidalis (Ma et al., 2019), *P. tremula* and *P. tremuloides* (Lin et al., 2018), and *P. trichocarpa* (Tuskan et al., 2006)--were used to assess the chromosome-scale pseudomolecules resulting from Hi-C analysis (based on co-phylogenetic analysis). The *P. tomentosa* genome was successfully separated into two subgenomes (2×19 choromosomes). Mapping of syntenic regions within the assembly showed clear chromosome-to-chromosome correspondence and also extensive synteny among different chromosomes, as expected for the highly duplicated *Populus* genome (Fig.1a).

Subsequently, we analyzed genome redundancy using the random best model of blasr (Chaisson and Tesler, 2012). The coverage depth distribution of BUSCO (Benchmarking

Universal Single-Copy Orthologs) (Simao et al., 2015) for duplicated and single-copy core genes was identical, showing an expected Poisson distribution (Fig.S3). This indicates that these duplicated genes were not derived from assembling redundancy. More than 96% of complete genes identified by BUSCO could be detected in the genome. Meanwhile, the proportion of transcriptome data mapped to the genome reached 97.8%, suggesting that the genome was of high quality and nearly complete.

After integration of transcriptome, Illumina and Hi-C with the PacBio assemblies, a *de novo* assembly resulted in a draft diploid genome of 740.2 Mb for *P. tomentosa*. It was comprised of 38 chromosome-scale pseudomolecules that covered 92.1% of the 800-megabase genome of *P. tomentosa* (Table 1, Fig.1). Compared with previous poplar genome assemblies (Ma et al., 2019; Ma et al., 2013; Tuskan et al., 2006; Yang et al., 2017), the *P. tomentosa* assembly quality in the present study was improved significantly; the sizes of contig N50 and scaffold N50 reached 0.96 Mb and 18.91 Mb, with the longest contig and scaffold being 5.5 Mb, and 46.7 Mb, respectively (Supplementary Table S4).

Using a combination of RepeatModeler (http://www.repeatmasker.org/RepeatModeler/) and RepeatMasker (http://www.repeatmasker.org/), 1,001,718 repeats were identified and masked (*de novo* identification, classification, and masking were run under default parameters, respectively). Collectively, these repeats were 307.6 Mb in size and comprised ~41.5% of the genome (Fig.1a). 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 was unknown elements, making up 9.8% of the genome, followed by 5.6% Helitron repeats and 5.4% DNA elements (Fig.1b) (details in Supplementary Table S5) .

To annotate genes, we trained the AUGUSTUS parameter model (Stanke et al., 2008) using single-copy core genes identified by BUSCO (Simao et al., 2015), followed by five rounds of optimization. We annotated the remaining unmasked *P. tomentosa* genome using a comprehensive strategy combining evidence-based techniques (RNA-Seq data and homologous protein) and *ab initio* gene prediction (Fig.1c). Using the MAKER2 gene annotation pipeline (Cantarel et al., 2008), we incorporated 73,919 protein sequences from two plant species and 137,918 transcripts assembled from *P. tomentosa* RNA-seq data. A total

of 59,124 high quality gene models were identified, with an average coding-sequence length of 1.31 kb, 6.04 exons per gene, 430 amino acids (aa) per protein. There was 28.5% genome coverage with an average length of 210.8 Mb (Table 1, Supplementary Table S6, Table S7).

The annotated genes were associated with the three onotological classes: biological process, cellular components, and molecular functions (Fig. 1c). Using tRNAScan-SE (Lowe and Eddy, 1997) and RNAMMER (Lagesen et al., 2007), 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 ncRNAs with a length of 218,117 bp using RfamScan (Kalvari et al., 2018). Finally, we performed alignments with protein databases using BLAT (Kent, 2002), with a maximum annotation ratio of 98.6% (Supplementary Table S8) .

Comparative Genomics and Evolution. We compared 19,594 gene family (59,124 genes) in the *P. tomentosa* genome with those of other three sequenced poplar genomes including *P. trichocarpa*, *P. euphratica*, and *P. pruinosa* using OrthoMCL (Zhang Z et al., 2003). A total of 22,386 gene families (142,738 genes) were identified by homolog clustering. 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. pruinose*, respectively (Fig. 2a, Supplementary Table S9).

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 (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 poplar and Salix species both underwent a common whole genome duplication (WGD) event ($Ks \approx 0.25$), which was similar to a paleopolyploidization event occurred in the majority of flowering plants (Myburg et al., 2014; Otto, 2007). This result is consistent with a

previous study on *Salix suchowensis* (Dai et al., 2014). Another peak represented the divergence between *Populus* and *Salix* occurred (*Ks* ≈ 0.12) (Fig.2b). Further analysis showed that section *Leuce* and *P. trichocarpa* have a divergence at *Ks* ≈ 0.035, and *P. adenopoda* and *P. alba* at *Ks* ≈ 0.025. Subsequently, *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) (Fig. 2c).

To study the parental origin of *P. tomentosa*, we derived 1,052 single-copy orthologous genes from nine poplars, including *P. trichocarpa, P. alba, P. alba var. pyramidalis, P. adenopoda, P. tremula, P. tremuloides*, and *P. davidiana* for phylogenetic analysis. We constructed phylogenetic trees using *Salix suchowensis* as an outgroup using RaxML (Stamatakis, 2014), based on the GTR+GAMMA model (Allman et al., 2014) and maximum likelihood analysis (Guindon et al., 2010). Finally, referencing the fossil-based divergence time of *Populus* and *Salix* at 48 Mya (Boucher et al., 2003; Dai et al., 2014; Manchester et al., 1986; Manchester et al., 2006), we estimated dates for taxonomic divergence.

Phylogenetic analysis indicated that the divergence event between *Leuce* poplars and section *Tacamahaca* (*P. trichocarpa*) occurred at approximately 13.4 Mya. *P. adenopoda*, an ancestor of *P. tomentosa*, was the first to separate from the *Leuce* family as an independent clade approximately 9.3 Mya. Subsequently, the aspen tribe and white poplars tribe underwent a divergence event (approximately 8.4 Mya). Another ancestor of *P. tomentosa*, *P. alba var.* pyramidalis, gave rise to an independent variant of *P. alba* at approximately 4.8 Mya. Approximately 3.9 Mya, *P. tomentosa*, a new white poplar hybrid, was created by hybridization between *P. adenopoda* and *P. alba var.* pyramidalis (Fig. 2d).

We re-constructed phylogenetic trees of subgenome A, subgenome D of *P. tomentosa* and other poplars (Fig.S4), as well as for each of the corresponding 19 pairs of chromosomes (Fig.S5). All of these analyses supported the hypothesis that *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 tree, combining with our previous phylogenetic analyses of chloroplast genomes from *Populus* section *Leuce* (Gao et al., 2019), which indicated that P.tomentosa is much closer to *P. adenopoda* than other *Leuce* poplars, and following the

law of maternal inheritance in chloroplast, we deduce that *P. alba* var. *pyramidalis* and *P. adenopoda* respectively played male and female parent roles in the hybrid formation of *P. tomentosa* during evolution.

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 (Fig. 2e) 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 collinear genes in *P. tomentosa* and *P. trihcocarpa*, and other dispersed distribution-blocks in the dot plot, suggested the collinearity of paralogous genes on non-homologous chromosomes between the two poplars (Fig. 2e). These findings show that both of the *P. tomentosa* sub-genomes are highly syntenic with each other and *P. trichocarpa*.

To investigate potential recombination events between the two sub-genomes, the distance of 5,345 single copy orthologus genes from *P. tomentosa*, *P. alba* var. *pyramidalis* and *P. adenopoda* were estimated by *Ks*. We conducted tests whether recombination occurs between two subgenomes of *P. tomentosa* based on *Ks* comparisons; results indicated that no recombination events occurred within 4,309 gene loci (80.62%), recombination events did occur in 38 gene loci (0.87%), and 998 gene loci did not meet either of above two hypotheses (Table S10, Fig.S6). This suggests that the two parental genomes may be largely still intact in *P. tomentosa*, at least with respect to genic composition. Based on the speculation that *P. tomentosa* is F1 hybrid, therefore, it is reasonable to maintain basic independence and stability of the two sub-genomes in *P. tomentosa* genome.

Chromosome structural variation in the *P. tomentosa* genome. 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 sub-genome level, with dispersed collinear blocks among homologous and nonhomologous chromosomes (Fig. 3, center). We infer that these may have arisen from hybridization and duplication events occurred in *Populus* prior to their divergence.

To study genome-wide structural variation (SV), including copy number variation (CNV), deletions (DEL), insertions (INS), inversions (INV), and translocations (TRANS) among chromosome pairs (Fig. 3, rings 1-5), we conduted alignments using MUMmer, and subsequently called them out using SVMU (Structural Variants fromMUMmer) 0.3 (https://github.com/mahulchak/svmu). The results 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 (6,654) and DEL (6,231) accounted for the majority (83%). The other variant numbers were 1,602 and 694, and 299 for INV, TRANS and CNV, respectively, which together accounted for 27% of the total number of SVs observed (Table S11). The vast majority of INS, DEL, and CNV variations occurred between homologous chromosome pairs, whereas TRANS were generally seen between non-homologous pairs (Table S1, Fig.S7).

By plotting the distribution of five SV types along 38 *P. tomentosa* chromosomes, we noticed that total 299 CNVs presented irregular and sporadic distribution across the whole genome (Fig.3①). Relatively, there are more CNVs distributed on Chr01A and Chr01D, wherea fewer CNVs were distributed on Chr07A, Chr07D, Chr13A, Chr13D, Chr15A and Chr15D. 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 (Fig.3②). Similarly, INSs were present at high-density and showed a slight preference for telomere regions of Chr07A, Chr07D, Chr15A, Chr15D, Chr18A and Chr18D (Fig.3③). In contrast, INVs had a more uneven distribution across the genome (Fig.3 ④). 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 (Fig.3 ⑤).

Genome-wide scanning of indels (insertion/deletion) among allele paires was conducted using a web analytics tool (http://qb.cshl.edu/assemblytics/) (Nattestad and Schatz, 2016). A total 188,575 indels across 15,052 alleles were identified. We noticed that most of indels were evenly distributed on 38 chromosomes in *P. tomentosa*, and a high-density of indels were seen in regions close to telomere of Chr01A, Chr01D, Chr11A, Chr11D Chr12A, Chr12D

298 Chr16A, Chr16D, Chr17A and Chr17D. Exceptions were the high-density indels seen on
299 the middle and end of Chr18A and Chr18D (Fig.3 ⑥).

300 We performed GO enrichment analysis for total 15,480 SV variants using Plant GoSlim
301 database, and detecd 23 GO categories significantly over-represented with respect to the
302 whole set of *P. trichocarpa* genes (Fig.4). Ten of them (“motor activity”, “transporter
303 activity”, “DNA binding”, “transport”, “metabolic process”, “lysosome”, “nuclear envelope”,
304 “peroxisome”, “cell wall” and “extracellular region”) were over-represented in genes affected
305 by INS, three (“chromatin binding”, “translation” and “ribosome”) were over- represented in
306 genes affected by CNV, three (“hydrolase activity”, “response to biotic stimulus” and “lipid
307 metabolic process”) were over- represented also in genes affected by both INS and TRANS,
308 two (“cell differentiation” and “growth”) were over-represented also in genes affected by INV,
309 two (“vacuole” and “circadian rhythm”) were over-represented also in genes affected by
310 TRANS, one (“endosome”) was over-represented also in genes affected by both EDL and
311 CNV, one (“carbohydrate binding”) was over-represented also in genes affected by EDL,
312 CNV and TRANS, and one (“plasma membrane”) was over-represented also in genes
313 affected by both CNV and TRANS. Overall, functional annotation showed enrichments
314 associated with all of the major GO categories (“Molecular Function”, “Biological Process”,
315 and “Cellular Component”).

316

317

318 **Discussion**

319 *P. tomentosa*, also known as Chinese white poplar, is indigenous and widely distribueted and
320 cultivated in a large area of China (Gao et al., 2019). Moreover, it was also the first tree
321 species planted in large-scale artificial plantations in China. Here, we integrated advanced
322 SMRT sequencing technology (PacBio), Illumina correction and chromosome conformation
323 capture (Hi-C) to assemble a high quality genome. It contained 6.24 million PacBio
324 single-molecule reads (~54 Gb, ~70 ×coverage of whole genome) with a mean contig N50 of
325 0.99 Mb and a mean scaffold N50 of 18.91 Mb. The longest contig N50 was 5.47 Mb, and the
326 longest scaffold N50 was 46.68 Mb. In comparion to several published poplar genomes, the
327 assembly quality of *P. tomentosa* was better than those of *P. trichocarpa* (Tuskan et al., 2006),

P. euphratica (Ma et al., 2013), *P. pruinosa* (Yang et al., 2017), *P. alba* var. pyramidalis (Ma et al., 2019). Even though the *P. alba* genome has a little bit longer contig N50 (1.18 Mb) than *P. tomentosa*, it has not been associated with specific chromosomes yet (Liu et al., 2019) (Supplementary Table 4). 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 they were resolved into a single haploid genome rather than two diploid subgenomes (Liu et al., 2019; Ma et al., 2019; Ma et al., 2013; Tuskan et al., 2006; Yang et al., 2017). However, this case is very similar to the genome of a hybrid poplar (84K) 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 S3).

We presented evidence for divergence and duplication events in *Populus*, as well as within the *P. tomentosa* lineage. Like other many flowering plants (e.g., Myburg et al., 2014; Otto, 2007), *Salicaceae* species underwent a common palaeohexaploidy event, followed by a palaeotetraploidy event before the divergence of *Salix* and *Populus*. Subsequently, poplar speciation occurred gradually. *Leuce* poplars and *P. trichocarpa* differentiated from each other approximately 13.44 Mya ($Ks \approx 0.035$). The ancestors of *P. tomentosa*, *P. adenopoda* and *P. alba var.* pyramidalis successively diverged from *Leuce* family approximately 9.3 Mya and 4.8 Mya. *Populus tomentosa* emerged from a hybridization event approximately 3.9 Mya. This finding differs from previous proposals on the origin of *P. tomentosa* (Dickmann and Isebrands, 2001; Wang et al., 2014).

Unlike other sequened poplars (Ma et al., 2013; Tuskan et al., 2006; Yang et al., 2017), the *P. tomentosa* genome consists of two parts, subgenome A (*P. alba* var. *pramidalis*) and subgenome D (*P. adenopoda*) (Fig. 1 and Table 1). Hi-C, as a chromosome conformation capture-based method, has become a mainstream techniqus for the study of the 3D organization of genomes (Belaghzal et al., 2017; Ma et al., 2018). Based on both Hi-C analysis (Fig. S2) and mapping of transcriptome (derived from *P. adenopoda* and genome data of *P. alba* var. *pramidalis)*, we were able to partion the *P. tomentosa* genome into two

358 subgenomes. Phylogenetic analysis reveals the relationships among three white poplars (Fig.
359 2d, Fig. S4). Futher, 19 chromosome-phylogenetic trees provide solid evidence of the hybrid
360 origin of the entire genome (Fig. S5). In addtion, the previous phylogenetic analysis of the
361 chloroplast genome of *P. tomentosa* showed that it originated from *P. adenopoda* (Gao et al.,
362 2019).

363 Our analysis of recombination events within genes showed that the *P. tomentosa*
364 subgenomes have largely been maintained despite sharing the same nucleus for at least 3.93
365 million years. Comparision of 5,345 single copy orthologs from *P. tomentosa*, *P. alba* var.
366 *pyramidalis* and *P. adenopoda* showed an absence of recombination among 99.13% of them
367 (Fig. S6, Supplementary Table S10). This suggests that two subgenomes of *P. tomentosa* have
368 been maintained relatively intact. Similarly, such karyotype stability has been observed in the
369 paleo-allotetraploid *Cucurbita* genomes (Sun et al., 2017) and in newly synthesized
370 allotetraploid wheat containing genome combinations analogous to natural tetraploids
371 (Huakun et al., 2013; Zhang et al., 2013). However, this contrasts with the frequent
372 homoeologous exchanges among subgenomes of allotetraploid cotton during its long
373 evolutionary history after polyploidization (Li et al., 2015), as well as the considerable
374 rearrangement events in many other polyploid plants (Simon and Wendel, 2014).

375 As in *Cucurbita* subgenomes (Sun et al., 2017), the karyotypic stability in *P. tomentosa*
376 genome could be due to the rapid divergence between the two diploid genome donors in their
377 repetitive DNA composition, which may have prevented meiotic pairing of homologous
378 chromosomes and subsequent exchanges. This hypothesis may be supported by the
379 observtion that dual-spindle formation in zygotes keeps parental genomes separate in some
380 taxa but not others, such as for early mammalian embryos (Johansson et al., 2013;
381 Reichmann et al., 2017). We also found that transposon abundance and distribution varied
382 significantly. For example, LTR (Copia, Gypsy), LINE, SINE(tRNA), DNA(CMC-EnSpm,
383 hAT-Ac, hAT-Tag1, PIF-Harbinger) and RC (Helitron).The few CNVs found in *P. tomentosa*
384 may also be a reason for maintaining relative independence, stability and specifity of its
385 subgenomes.

386 Transposable elements are also important factors that can cause CNVs, INSs and DELs
387 due to their capacity to mobilize gene sequences within the genome (Kidwell and Lisch, 1997;

Morgante et al., 2007; Pinosio et al., 2016), both in the wild and in breeding processes (Lisch, 2013; Olsen and Wendel, 2013; Sanseverino et al., 2015; Sun et al., 2018). Often these movements are associated with agronomically important traits such as skin or flesh color of the orange, grape, and peach (Butelli et al., 2012; Falchi et al., 2013; Kobayashi et al., 2004)—and their large variation among subgenomes may provide a degree of "fixed heterosis" that contributes to the high productivity and wide distribution of *P. tomentosa*.

Methods

Study system. The genus *Populus*, collectively known as the poplars, comprises six taxonomically distinct sections (*Leuce, Aigeiros, Tacamahaca, Leucoides, Turanga* and *Abaso*) consisting of nearly 30 tree species found in the Northern Hemisphere. Poplars often segregate into geographically and morphologically distinct subspecies and varieties due to their large natural range, and show considerable genetic variation in their adapative characteristics (Evans et al., 2014). As predominantly dioecious, wind-pollinated species with low barriers to crossability among consectional taxa, numerous natural and induced hybrids have been described (Dickmann and Isebrands, 2001). Poplars are pioneer species that have among the highest potential rate of biomass growth of temperate tree species (Lin et al., 2018). These traits are often enhanced in interspecific hybrids, promoting their commercial and ecological value, with multiple uses including biofuel, fiber, lumber, pulp, veneer, bioremediation, and windbreaks (Stettler et al., 1996). *Populus* generally exhibits rapid growth, easy propagation, a modest genome size (<500 Mb), and is amendable to genetic transformation (Tuskan et al., 2006). Moreover, *Populus* displays abundant genetic variation at different taxonomic levels: among sections within the genus, as well as among species, provenances, populations, individuals, and genes (Stettler et al., 1996). This diversity underlies local adaptation commonly found in the genus (Evans et al., 2014; Holliday et al., 2016; Wang et al., 2018b), and facilitates the study of species origins, evolution and divergence (Wang et al., 2016). These characteristics render *Populus* an ideal system for genetics and functional genomics studies of trees and other perennial plants (Ma et al., 2013).

Populus tomentosa Carr., an indigenous white poplar that is widespread in parts of China, is classified in section *Leuce* along with the aspens. It is a dominant species in many of the

418 ecosystem it occurs in, and is widely distributed within a 1,000,000 km^2 area in the Yellow
419 River, Huaihe and Haihe regions (Gao et al., 2019). Its characteristics include rapid growth, a
420 thick and straight trunk, environmental stress tolerance, and a long lifespan (typically
421 100-200 years, but sometimes over 500 years). These traits make *P. tomentosa* valuable from
422 economic, ecological and evolutionary perspectives, with applications that include timber,
423 pulp and paper, veneer, plywood, bioremediation, wind break, carbon capture, and prevention
424 of soil erosion.
425
426 ***In vitro* culture of anthers and regeneration**. Branches with floral buds were collected
427 from an approximately 35-year-old *P. tomentosa* clone (LM50), which is a male elite
428 individual in *P. tomentosa* tribe, on January 5, 2015, and cultured in containers with clean
429 water at 24 ± 1 ℃ under 16/8 h light/dark conditions in a greenhouse. We referenced
430 previous anther induced regenerated system (Li et al., 2013) and adjusted the culture medium.
431 After determination of microspore developmental stages and pretreatment, anthers were
432 cultured on callus induction medium (H medium, with 1.0 mg/L 6-BA, 1.0mg/L NAA, 5.5g/L
433 agarose, and 30 g/L sucrose, pH 5.8) in the dark at 24 ± 1 ℃ and 60–65 % relative humidity
434 for 6 months during which anthers were transferred monthly onto fresh media. Callus was
435 transferred to shoot induction medium (MS medium, with 0.5 mg/L 6-BA, 0.05 mg/L NAA,
436 5.5g/L agarose, and 30 g/L sucrose, pH 5.8) to induce adventitious buds under conditions in
437 similar to those described above. After 5 weeks of culture, shoots regenerated and 1.5-2.0 cm
438 shoot sections were cut off, and transferred to the shoot rooting medium (1/2 MS medium
439 with 0.3 mg/L IBA 0.3mg/L, 5.5g/L agarose, and 20g/L sucrose, pH 5.8). Plantlets were
440 generated 30-40 days later.
441
442 **Plant material and DNA/RNA extraction.** The plantlet GM15 generated by *in vitro*
443 culture and regeneration system of anther from LM50, was selected for genome sequencing.
444 Tissue cultured plantlets of GM15 were grown in the tissue culture room under natural light
445 supplemented with artificial light (16 h light/8 h dark). Fresh leaves, stems and roots were
446 harvested, immediately frozen in liquid nitrogen and stored at - 80 ˚C for genome and
447 transcriptome sequencing. Young leaves collected from *P. alba*, *P. alba var.* primidalis, *P.*

448 *davidiana*, and *P.a denopoda* were also forzen in liquid nitrogen and stored at - 80 ˚C for
449 transcriptome sequencing.
450 Genomic DNA and total RNA were extracted using the Qiagen DNeasy Plant Mini Kit
451 and the Qiagen RNeasy Plant Mini Kit, respectively, following the manufacturer's
452 instructions (Qiagen, Valencia, CA, USA). DNA and RNA quality were evaluated by agarose
453 gel electrophoresis and their quantities determined using a NanoDrop spectrophotometer
454 (Thermo Fisher Scientific, Waltham, MA, USA).
455
456 **Genomic and RNA-seq library construction and sequencing.** A total of 2 μg of intact
457 genomic DNA was fragmented and used to construct short-insert PCR-free libraries (300 and
458 500 bp) following the manufacturer's protocol (Illumina Inc., San Diego, CA, USA). The
459 DNA libraries were used in paired-end sequencing (~265 million paired-end reads, ~40 G
460 bases, ~50× coverage) on the Illumina HiSeq X-10 sequencer. For SMRT sequencing,
461 genomic DNA was fragmented using g-Tubes (Covaris, Inc., Woburn, MA, USA) and 18–22
462 kb DNA fragments were further purified using AMPure magnetic beads (Beckman Coulter,
463 Fullerton, CA, USA). DNA Template Prep Kit 4.0 V2 (Pacific Biosciences, USA) was used
464 to constructed 20-kb library following the manufacturer's protocol (Pacific Biosciences,
465 Menlo Park, CA, USA). Finally, 50 μg of high-quality genomic DNA was used to generate 11
466 SMRT cells and then sequenced on the PacBio Sequel platform (Pacific Biosciences). To
467 assist prediction and annotation of genes from the *P. tomentosa* genome assembly, 2 μg of
468 total RNA from various tissues was used to construct RNA-seq library and sequence on the
469 Illumina HiSeq X-10 platform following the protocol of NEBNext Ultra RNA Library Prep
470 Kit for Illumina (New England Biolabs Ipswich, MA, USA).
471
472 ***De novo* genome assembly and estimation of genome size**. A total of 6.24 M PacBio
473 post-filtered reads were generated from the 11 SMRT cells, producing a total of ~6.24 million
474 reads, and ~54 G bases (~ 70× coverage) of single-molecule sequencing data. *De novo*
475 assembly was conducted using an overlap-layout-consensus method in CANU (Koren et al.,
476 2017). The draft assembly was polished with Arrow
477 (https://github.com/PacificBiosciences/GenomicConsensus) to improve accuracies The size

of contig N50 reached 1.8 Mb, yielding a final assembly with a total length of 740.18 Mb. We estimated genome size by k-mer distribution analysis with the program Jellyfish (k = 17) using Illumina short reads, and obtained an estimate of 803.6 Mb by using the Genome Characteristic Estimation (GCE) program (Liang et al., 2013). For additional details about k-mer distribution, see Supplementary Table S2.

Hi-C library sequencing and scaffold anchoring. The Hi-C library was prepared using standard procedures described as follows. A total of 700 ng of high molecular-weight genomic DNA was cross-linked *in situ*, extracted, and then digested with a restriction enzyme. The sticky ends of the digested fragments were biotinylated, diluted, and randomly ligated to each other. After ligation, cross-links were reversed and the DNA was purified from protein. Purified DNA was treated to remove biotin that was not internal to the ligated fragments. Biotinylated DNA fragments were enriched and sheared to a fragment size of 300–500 bp again before preparing the sequencing library, which was sequenced on a HiSeq X-10 platform (Illumina). This yielded a total of 430 M reads, and 65 G bases (~81 × coverage).

The Hi-C reads were first compared to the above draft genome using Juicer (Durand et al., 2016b). To account for high duplication level, only the aligned sequences with map-quality score >40 were used to conduct Hi-C association chromosome assembly (Dudchenko et al., 2017). Visualization was carried out using Juicebox (Durand et al., 2016a). In Hi-C assembling, parameters were set to -m haploid -t 5,000 -s 2 -c 19. In another words, sequences above 5,000 bp in length were assembled, followed by two rounds of correction, and finally split into 19 chromosomes. Misjoined sequences were split in the process of correction, causing contigs to increase and N50 falls. For some obvious chromosomal splitting errors, we performed local optimization using the assembly method described above. Subsequently, the assembly was polished with Arrow over three iterations using PacBio reads and finally corrected using Illumina short reads with Pilon over five iterations.

Transcriptome assembly and genome annotation. RNA-seq reads were preprocessed using Cutadapt to remove contaminating sequences from adaptors and sequences with low base quality. We employed a combination strategy of *de novo* and genome-guilded, consisting

of 1) HiSat2 (Pryszcz and Gabaldon, 2016) + StringTie (Pertea et al., 2015), 2) HiSat2 (Pryszcz and Gabaldon, 2016) + Trinity (Grabherr et al., 2011) genome-guild mode, and 3) Trinity (Grabherr et al., 2011) *de novo* mode, to assemble the transcripts. Combining all of the transcripts and removing redundant sequences by CD-HIT (Fu et al., 2012) with 95 % identity and 95 % coverage, a total of 137,918 transcriptional sequences (Table 5) were yileded. These transcripts were used as EST evidence for subsequent gene annotation.

We annotated *de novo* repeats using the RepeatModeler (running in default parameter environment), subsequently masked repeat library using RepeatMasker pipeline (running in default parameter environment), and predicted genes using the MAKER2 annotation pipeline. Homologous protein evidence for the MAKER2 pipeline were provided in the form of 73,919 non-redundant protein sequences from *Arabidopsis* (TAIR10) and *P. trichocarpa* (JGI3.0) . The single copy core genes identified by BUSCO (Simao et al., 2015) were used to train the AUGUSTUS (Stanke et al., 2008) parameter model, and five rounds of optimization were carried out. MAKER2 pipeline was carried out with combining *ab initio* prediction, EST sequence alignmenst and protein sequence alignmetns, and finally integrated these data with AED score calculated for quality control (Hoff et al., 2016). tRNA and rRNA were predicted using tRNAScan-SE (Lowe and Eddy, 1997) and RNAmmer (Lagesen et al., 2007), respectively. Other non-coding RNAs were annotated using RfamScan. Functional annotation was performed by aligning protein sequences with the protein database using BLAT (Kent, 2002) (identity >30 %, and the E < 1e -5).

Molecular phylogenetic tree, whole-genome duplication and divergence events, subgenome recombination test. We firstly collected genome data of *P. tremula* (Lin et al., 2018), *P. tremuloides* (Lin et al., 2018), *P. trichocarpa* (Tuskan et al., 2006), and *Salix_suchowensis* (Dai et al., 2014), and *de novo* transcriptomes assembly of *P. davidiana*, and *P. grandidentata, P. adenopoda, P. alba* and *P. tomentosa* (Table S12). Then we performed gene family clustering using OrthoMCL (default parameters) on protein sequences, and conducted further collinearity analysis of homologous gene pairs derived from genome data of poplars and *Salix suchowensis* using MCScanX (Wang et al., 2012)

Based on collinear homologous gene pairs, including interspecific orthologs and

intraspecific paralogs without tandem repeats, we aligned protein sequences using MUSCLE (Edgar, 2004), then used PAL2NAL to carry out codon alignment (Suyama et al., 2006). The YN model-based *Ka* and *Ks* calculation was performed using KaKs_Calculator (Zhang et al., 2006). The *Ks* distributions of the collinear homologous pairs (inparalogs and orthologs) were used to infer whole genome doubling (WGD) events and divergence events in species genome.

Finally, we referencing 1052 single copy orthologous genes derived from 10 plant species, constructed a molecular phylogenetic tree using RAxML (Stamatakis, 2014) based on the GAMMA+GTR model. Assuming the divergence time of *Populus* and *Salix*–48 Mya (Boucher et al., 2003; Dai et al., 2014; Manchester et al., 1986; Manchester et al., 2006) as fossil calibration, we estimated dates for WGD and divergent events of poplar species using r8s (Sanderson, 2003).

In additon to, we slected 5,345 single copy orthologous genes, which are collinear allele pairs between two subgenomes of *P. tomentosa* (PtA, PtD), and are homologus to those of *P. alba* var. pyramidalis (PA) and *P. adenopoda* (PD), to measure their distances through *Ks*, to investigate poptential recombinantion between homologus gene pairs based two hypothese: (1) if meet Ks (PD-PtD) < Ks (PD-PtA) and Ks (PA-PtA) < Ks(PA-PtD), then support the expectation of no recombination events between orthologous genes. (2) if meet Ks (PD-PtD) > Ks (PD-PtA) and Ks (PA-PtA) > Ks (PA-PtD), then support the expectation of recombination events between orthologous genes. If none of the above conditions are true, it may be false positive homology, gene loss, imbalance of evolution rate, etc., which shall not be taken as the judgment condition.

Whole-genome synteny, chromosomal structure variations and indels in alleles. We conducted genome-wide synteny analysis between *P.tomentosa* and *P.trichocarpa*, and subgenome synteny analysis between subgenome A and subgenome 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 corresponding chromosome pairs in subgenomes were detected using MUMmer, and chromosome variation was identified using SVMU (Structural Variants from MUMmer) 0.3

(https://github.com/mahulchak/svmu). Besed on previous synety anlysis of homologous, alleles on homologous chromosomes were aligned using muscle, and indels were called out using a web tool (http://qb.cshl.edu/assemblytics/) (Nattestad and Schatz, 2016).

Reference

Allman, E., Ane, C., Rhodes, J., 2014. Identifiability of the GTR+Gamma substitution model of DNA evolution. Astronomy & Astrophysics 567.

Ambardar, S., Gupta, R., Trakroo, D., Lal, R., Vakhlu, J., 2016. High Throughput Sequencing: An Overview of Sequencing Chemistry. Indian Journal of Microbiology 56, 1-11.

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, 89-100.

Arabidopsis Genome, I., 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796-815.

Ballouz, S., Dobin, A., Gillis, J.A., 2019. Is it time to change the reference genome? Genome Biol 20, 159.

Belaghzal, H., Dekker, J., Gibcus, J.H., 2017. Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. Methods 123, 56-65.

Boucher, L.D., Manchester, S.R., Judd, W.S., 2003. An extinct genus of Salicaceae based on twigs with attached flowers, fruits, and foliage from the Eocene Green River Formation of Utah and Colorado, USA. American journal of botany 90, 1389-1399.

Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackay, S., Bailey, P., Reforgiato-Recupero, G., Martin, C., 2012. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell 24, 1242-1255.

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A.S., Yandell, M., 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18, 188-196.

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.

Chen, Z., Rao, P., Yang, X., Su, X., Zhao, T., Gao, K., Yang, X., An, X., 2018. A Global View of Transcriptome Dynamics During Male Floral Bud Development in Populus tomentosa. Scientific Reports 8, 722.

Chia, J.M., Song, C., Bradbury, P.J., Costich, D., de Leon, N., Doebley, J., Elshire, R.J., Gaut, B., Geller, L., Glaubitz, J.C., Gore, M., Guill, K.E., Holland, J., Hufford, M.B., Lai, J., Li, M., Liu, X., Lu, Y., McCombie, R., Nelson, R., Poland, J., Prasanna, B.M., Pyhajarvi, T., Rong, T., Sekhon, R.S., Sun, Q., Tenaillon, M.I., Tian, F., Wang, J., Xu, X., Zhang, Z., Kaeppler, S.M., Ross-Ibarra, J., McMullen, M.D., Buckler, E.S., Zhang, G., Xu, Y., Ware, D., 2012. Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet 44, 803-807.

Chin, C.S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., Turner, S.W., Korlach, J., 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10, 563-+.

610 Choulet, F., Alberti, A., Theil, S., Glover, N., Barbe, V., Daron, J., Pingault, L., Sourdille, P., Couloux,
611 A., Paux, E., Leroy, P., Mangenot, S., Guilhot, N., Le Gouis, J., Balfourier, F., Alaux, M.,
612 Jamilloux, V., Poulain, J., Durand, C., Bellec, A., Gaspin, C., Safar, J., Dolezel, J., Rogers, J.,
613 Vandepoele, K., Aury, J.M., Mayer, K., Berges, H., Quesneville, H., Wincker, P., Feuillet, C.,
614 2014. Structural and functional partitioning of bread wheat chromosome 3B. Science 345,
615 1249721.
616 Dai, X.G., Hu, Q.J., Cai, Q.L., Feng, K., Ye, N., Tuskan, G.A., Milne, R., Chen, Y.N., Wan, Z.B.,
617 Wang, Z.F., Luo, W.C., Wang, K., Wan, D.S., Wang, M.X., Wang, J., Liu, J.Q., Yin, T.M., 2014.
618 The willow genome and divergent evolution from poplar after the common genome duplication.
619 Cell Res 24, 1274-1277.
620 Dickmann, D.I., Isebrands, J.G., 2001. Poplar Culture in North America. NRC Research Press.
621 Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S.,
622 Machol, I., Lander, E.S., Aiden, A.P., Aiden, E.L., 2017. De novo assembly of the Aedes aegypti
623 genome using Hi-C yields chromosome-length scaffolds. Science 356, 92-95.
624 Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., Aiden, E.L.,
625 2016a. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom.
626 Cell Syst 3, 99-101.
627 Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S., Huntley, M.H., Lander, E.S., Aiden, E.L., 2016b.
628 Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst
629 3, 95-98.
630 Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput.
631 Nucleic acids research 32, 1792-1797.
632 El-Metwally, S., Ouda, O.M., Helmy, M., 2014. Next-Generation Sequencing Platforms. Annual
633 Review of Analytical Chemistry.
634 Ellegren, H., 2014. Genome sequencing and population genomics in non-model organisms. Trends
635 Ecol Evol 29, 51-63.
636 Evans, L.M., Slavov, G.T., Eli, R.M., Joel, M., Priya, R., Wellington, M., Brunner, A.M., Wendy, S.,
637 Lee, G., Jin-Gui, C., 2014. Population genomics of Populus trichocarpa identifies signatures of
638 selection and adaptive trait associations. Nature Genetics 46, 1089-1096.
639 Falchi, R., Vendramin, E., Zanon, L., Scalabrin, S., Cipriani, G., Verde, I., Vizzotto, G., Morgante, M.,
640 2013. Three distinct mutational mechanisms acting on a single gene underpin the origin of yellow
641 flesh in peach. Plant J 76, 175-187.
642 Fu, L.M., Niu, B.F., Zhu, Z.W., Wu, S.T., Li, W.Z., 2012. CD-HIT: accelerated for clustering the
643 next-generation sequencing data. Bioinformatics 28, 3150-3152.
644 Gao, K., Li, J., Khan, W.U., Zhao, T., Yang, X., Yang, X., Guo, B., An, X., 2019. Comparative
645 genomic and phylogenetic analyses of Populus section Leuce using complete chloroplast genome
646 sequences. Tree Genetics & Genomes 15.
647 Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L.,
648 Raychowdhury, R., Zeng, Q.D., Chen, Z.H., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di
649 Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011.
650 Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature
651 Biotechnology 29, 644-U130.
652 Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New
653 Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the

654 Performance of PhyML 3.0. Syst Biol 59, 307-321.
655 He, C., 2005. Study on Genetic diversity and origin of Populus tomentosa Carr. . Dissertation, Beijing
656 Forestry University.
657 Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G., Vaillancourt, B., Penagaricano,
658 F., Lindquist, E., Pedraza, M.A., Barry, K., de Leon, N., Kaeppler, S.M., Buell, C.R., 2014.
659 Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26, 121-135.
660 Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M., Stanke, M., 2016. BRAKER1: Unsupervised
661 RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32,
662 767-769.
663 Holliday, J.A., Zhou, L., Bawa, R., Zhang, M., Oubida, R.W., 2016. Evidence for extensive
664 parallelism but divergent genomic architecture of adaptation along altitudinal and latitudinal
665 gradients in Populus trichocarpa. New Phytologist 209, 1240-1251.
666 Huakun, Z., Yao, B., Xiaowan, G., Yuzhu, D., Sachin, R., Bangjiao, Z., Chunming, X., Ning, L., Bao,
667 Q., Fangpu, H., 2013. Intrinsic karyotype stability and gene copy number variations may have
668 laid the foundation for tetraploid wheat formation. Proceedings of the National Academy of
669 Sciences of the United States of America 110, 19466-19471.
670 International Rice Genome Sequencing, P., 2005. The map-based sequence of the rice genome. Nature
671 436, 793-800.
672 Johansson, J., Bolmgren, K., Jonzen, N., 2013. Climate change and the optimal flowering time of
673 annual plants in seasonal environments. Global change biology 19, 197-207.
674 Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A.,
675 Finn, R.D., Petrov, A.I., 2018. Rfam 13.0: shifting to a genome-centric resource for non-coding
676 RNA families. Nucleic acids research 46, D335-D342.
677 Kent, W.J., 2002. BLAT--the BLAST-like alignment tool. Genome Res 12, 656-664.
678 Kidwell, M.G., Lisch, D., 1997. Transposable elements as sources of variation in animals and plants.
679 Proceedings of the National Academy of Sciences of the United States of America 94, 7704-7711.
680 Kobayashi, S., Goto-Yamamoto, N., Hirochika, H., 2004. Retrotransposon-induced mutations in grape
681 skin color. Science 304, 982.
682 Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., Phillippy, A.M., 2017. Canu: scalable
683 and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome
684 Res 27, 722-736.
685 Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., Ussery, D.W., 2007. RNAmmer:
686 consistent and rapid annotation of ribosomal RNA genes. Nucleic acids research 35, 3100-3108.
687 Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R.J., Ma, Z., Shang, H., Ma, X., Wu, J., 2015.
688 Genome sequence of cultivated Upland cotton (GossypiumÂ hirsutum TM-1) provides insights
689 into genome evolution. Nature Biotechnology 33, 524-530.
690 Li, Y., Li, H., Chen, Z., Ji, L.X., Ye, M.X., Wang, J., Wang, L.N., An, X.M., 2013. Haploid plants
691 from anther cultures of poplar (Populus x beijingensis). Plant Cell Tiss Org 114, 39-48.
692 Lin, Y.C., Wang, J., Delhomme, N., Schiffthaler, B., Sundstrom, G., Zuccolo, A., Nystedt, B.,
693 Hvidsten, T.R., de la Torre, A., Cossu, R.M., Hoeppner, M.P., Lantz, H., Scofield, D.G., Zamani,
694 N., Johansson, A., Mannapperuma, C., Robinson, K.M., Mahler, N., Leitch, I.J., Pellicer, J., Park,
695 E.J., Van Montagu, M., Van de Peer, Y., Grabherr, M., Jansson, S., Ingvarsson, P.K., Street, N.R.,
696 2018. Functional and evolutionary genomic inferences in Populus through genome and
697 population sequencing of American and European aspen. Proceedings of the National Academy

of Sciences of the United States of America 115, E10970-E10978.

Lisch, D., 2013. How important are transposons for plant evolution? Nat Rev Genet 14, 49-61.

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. Sci China Life Sci 62, 609-618.

Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic acids research 25, 955-964.

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, 451-460.

Ma, T., Wang, J.Y., Zhou, G.K., Yue, Z., Hu, Q.J., Chen, Y., Liu, B.B., Qiu, Q., Wang, Z., Zhang, J., Wang, K., Jiang, D.C., Gou, C.Y., Yu, L.L., Zhan, D.L., Zhou, R., Luo, W.C., Ma, H., Yang, Y.Z., Pan, S.K., Fang, D.M., Luo, Y.D., Wang, X., Wang, G.N., Wang, J., Wang, Q., Lu, X., Chen, Z., Liu, J.C., Lu, Y., Yin, Y., Yang, H.M., Abbott, R.J., Wu, Y.X., Wan, D.S., Li, J., Yin, T.M., Lascoux, M., DiFazio, S.P., Tuskan, G.A., Wang, J., Liu, J.Q., 2013. Genomic insights into salt adaptation in a desert poplar. Nat Commun 4.

Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., Hesson, J., Cavanaugh, C., Ware, C.B., Krumm, A., Shendure, J., Blau, C.A., Disteche, C.M., Noble, W.S., Duan, Z., 2018. Using DNase Hi-C techniques to map global and local three-dimensional genome architecture at high resolution. Methods 142, 59-73.

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, 156-160.

Manchester, S.R., Judd, W.S., Handley, B., 2006. Foliage and fruits of early poplars (Salicaceae : Populus) from the eocene of Utah, Colorado, and Wyoming. Int J Plant Sci 167, 897-908.

Mardis, E., 2013. Next-generation sequencing platforms. . Annu Rev Anal Chem 6, 287–303.

Morgante, M., De Paoli, E., Radovic, S., 2007. Transposable elements and the plant pan-genomes. Curr Opin Plant Biol 10, 149-155.

Morin, S.J., Eccles, J., Iturriaga, A., Zimmerman, R.S., 2017. Translocations, inversions and other chromosome rearrangements. Fertil Steril 107, 19-26.

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., Van Jaarsveld, I., Silva, O.B., Togawa, R.C., Pappas, M.R., Faria, D.A., Sansaloni, C.P., Petroli, C.D., Yang, X.H., Ranjan, P., Tschaplinski, T.J., Ye, C.Y., Li, T., Sterck, L., Vanneste, K., Murat, F., Soler, M., Clemente, H.S., Saidi, N., Cassan-Wang, H., Dunand, C., Hefer, C.A., Bornberg-Bauer, E., Kersting, A.R., Vining, K., Amarasinghe, V., Ranik, M., Naithani, S., Elser, J., Boyd, A.E., Liston, A., Spatafora, J.W., Dharmwardhana, P., Raja, R., Sullivan, C., Romanel, E., Alves-Ferreira, M., Lheim, C.K., Foley, W., Carocha, V., Paiva, J., Kudrna, D., Brommonschenkel, S.H., Pasquali, G., Byrne, M., Rigault, P., Tibbits, J., Spokevicius, A., Jones, R.C., Steane, D.A., Vaillancourt, R.E., Potts, B.M., Joubert, F., Barry, K., Pappas, G.J., Strauss, S.H., Jaiswal, P., Grima-Pettenati, J., Salse, J., Van de Peer, Y., Rokhsar, D.S., Schmutz, J., 2014. The genome of Eucalyptus grandis. Nature 510, 356-+.

Nattestad, M., Schatz, M.C., 2016. Assemblytics: a web analytics tool for the detection of variants

742 from an assembly. Bioinformatics 32, 3021-3023.
743 Olsen, K.M., Wendel, J.F., 2013. A bountiful harvest: genomic insights into crop domestication
744 phenotypes. Annu Rev Plant Biol 64, 47-70.
745 Otto, S.P., 2007. The evolutionary consequences of polyploidy. Cell 131, 452-462.
746 Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G.,
747 Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., Wang, X., Wicker, T.,
748 Bharti, A.K., Chapman, J., Feltus, F.A., Gowik, U., Grigoriev, I.V., Lyons, E., Maher, C.A.,
749 Martis, M., Narechania, A., Otillar, R.P., Penning, B.W., Salamov, A.A., Wang, Y., Zhang, L.,
750 Carpita, N.C., Freeling, M., Gingle, A.R., Hash, C.T., Keller, B., Klein, P., Kresovich, S.,
751 McCann, M.C., Ming, R., Peterson, D.G., Mehboob ur, R., Ware, D., Westhoff, P., Mayer, K.F.,
752 Messing, J., Rokhsar, D.S., 2009. The Sorghum bicolor genome and the diversification of grasses.
753 Nature 457, 551-556.
754 Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., Salzberg, S.L., 2015. StringTie
755 enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology
756 33, 290-+.
757 Pinosio, S., Giacomello, S., Faivre-Rampant, P., Taylor, G., Jorge, V., Le Paslier, M.C., Zaina, G.,
758 Bastien, C., Cattonaro, F., Marroni, F., Morgante, M., 2016. Characterization of the Poplar
759 Pan-Genome by Genome-Wide Identification of Structural Variation. Mol Biol Evol 33,
760 2706-2719.
761 Pryszcz, L.P., Gabaldon, T., 2016. Redundans: an assembly pipeline for highly heterozygous genomes.
762 Nucleic acids research 44.
763 Putnam, N.H., O'Connell, B.L., Stites, J.C., Rice, B.J., Blanchette, M., Calef, R., Troll, C.J., Fields, A.,
764 Hartley, P.D., Sugnet, C.W., Haussler, D., Rokhsar, D.S., Green, R.E., 2016. Chromosome-scale
765 shotgun assembly using an in vitro method for long-range linkage. Genome Res 26, 342-350.
766 Qiu, D., Bai, S., Ma, J., Zhang, L., Shao, F., Zhang, K., Yang, Y., Sun, T., Huang, J., Zhou, Y.,
767 Galbraith, D.W., Wang, Z., Sun, G., 2019. The genome of Populus alba x Populus tremula var.
768 glandulosa clone 84K. DNA research : an international journal for rapid publication of reports on
769 genes and genomes 26, 423-431.
770 Reichmann, J., Nijmeijer, B., Hossain, M.J., Eguren, M., Schneider, I., Politi, A.Z., Roberti, M.J.,
771 Hufnagel, L., Hiiragi, T., Ellenberg, J., 2017. Dual spindle formation in zygotes keeps parental
772 genomes apart in early mammalian embryos. bioRxiv, 198275.
773 Sanderson, M.J., 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the
774 absence of a molecular clock. Bioinformatics 19, 301-302.
775 Sanseverino, W., Henaff, E., Vives, C., Pinosio, S., Burgos-Paz, W., Morgante, M., Ramos-Onsins,
776 S.E., Garcia-Mas, J., Casacuberta, J.M., 2015. Transposon Insertions, Structural Variations, and
777 SNPs Contribute to the Evolution of the Melon Genome. Mol Biol Evol 32, 2760-2774.
778 Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen,
779 J.J., Cheng, J., Xu, D., Hellsten, U., May, G.D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya,
780 M.K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D.,
781 Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault,
782 M., Sethuraman, A., Zhang, X.C., Shinozaki, K., Nguyen, H.T., Wing, R.A., Cregan, P., Specht, J.,
783 Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R.C., Jackson, S.A., 2010. Genome sequence
784 of the palaeopolyploid soybean. Nature 463, 178-183.
785 Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton,

L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M.J., McMahan, L., Van Buren, P., Vaughn, M.W., Ying, K., Yeh, C.T., Emrich, S.J., Jia, Y., Kalyanaraman, A., Hsia, A.P., Barbazuk, W.B., Baucom, R.S., Brutnell, T.P., Carpita, N.C., Chaparro, C., Chia, J.M., Deragon, J.M., Estill, J.C., Fu, Y., Jeddeloh, J.A., Han, Y., Lee, H., Li, P., Lisch, D.R., Liu, S., Liu, Z., Nagel, D.H., McCann, M.C., SanMiguel, P., Myers, A.M., Nettleton, D., Nguyen, J., Penning, B.W., Ponnala, L., Schneider, K.L., Schwartz, D.C., Sharma, A., Soderlund, C., Springer, N.M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T.K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J.L., Dawe, R.K., Jiang, J., Jiang, N., Presting, G.G., Wessler, S.R., Aluru, S., Martienssen, R.A., Clifton, S.W., McCombie, W.R., Wing, R.A., Wilson, R.K., 2009. The B73 maize genome: complexity, diversity, and dynamics. Science 326, 1112-1115.

Sherman, R.M., Salzberg, S.L., 2020. Pan-genomics in the human genome era. Nat Rev Genet.

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, 3210-3212.

Simon, R.B., Wendel, J.F., 2014. Doubling down on genomes: polyploidy and crop plants. American journal of botany 101, 1711.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312-1313.

Stanke, M., Diekhans, M., Baertsch, R., Haussler, D., 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24, 637-644.

Stettler, R.F., Bradshaw, H.D., Jr, Heilman, P.E., Hinckley, T.M., 1996. Biology of Populus and its implications for management and conservation. Biology of & Its Implications for Management & Conservation, 542.

Sun, H., Wu, S., Zhang, G., Jiao, C., Guo, S., Ren, Y., 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, 1293-1306.

Sun, S., Zhou, Y., Chen, J., Shi, J., Zhao, H., Zhao, H., Song, W., Zhang, M., Cui, Y., Dong, X., Liu, H., Ma, X., Jiao, Y., Wang, B., Wei, X., Stein, J.C., Glaubitz, J.C., Lu, F., Yu, G., Liang, C., Fengler, K., Li, B., Rafalski, A., Schnable, P.S., Ware, D.H., Buckler, E.S., Lai, J., 2018. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat Genet 50, 1289-1295.

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.

Tian, Z., Zhao, M., She, M., Du, J., Cannon, S.B., Liu, X., Xu, X., Qi, X., Li, M.W., Lam, H.M., Ma, J., 2012. Genome-wide characterization of nonreference transposons reveals evolutionary propensities of transposons in soybean. Plant Cell 24, 4422-4436.

Tomato Genome, C., 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485, 635-641.

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R.R., Bhalerao, R.P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.L., Cooper, D., Coutinho, P.M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Dejardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjarvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leple, J.C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D.R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouze, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y., Rokhsar, D., 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313, 1596-1604.

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. J Vis Exp.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q.D., Wortman, J., Young, S.K., Earl, A.M., 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. Plos One 9.

Wang, A.Q., Wang, Z.Y., Li, Z., Li, L.M., 2018a. BAUM: improving genome assembly by adaptive unique mapping and local overlap-layout-consensus approach. Bioinformatics 34, 2019-2028.

Wang, J., Ding, J., Tan, B., Robinson, K.M., Michelson, I.H., Johansson, A., Nystedt, B., Scofield, D.G., Nilsson, O., Jansson, S., 2018b. A major locus controls local adaptation and adaptive life history variation in a perennial plant. Genome Biology 19, 72.

Wang, J., Street, N.R., Scofield, D.G., Ingvarsson, P.K., 2016. Variation in Linked Selection and Recombination Drive Genomic Divergence during Allopatric Speciation of European and American Aspens. Molecular Biology & Evolution 33, 1754-1767.

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 Res 40, e49.

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, e103645.

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, 1-7.

874 Zhang, H., Bian, Y., Gou, X., Dong, Y., Rustgi, S., Zhang, B., Xu, C., Li, N., Qi, B., Han, F., 2013.
875 Intrinsic karyotype stability and gene copy number variations may have laid the foundation for
876 tetraploid wheat formation. Proceedings of the National Academy of Sciences of the United
877 States of America 110, 19466-19471.
878 Zhang, L., Hu, J., Han, X., Li, J., Gao, Y., Richards, C.M., Zhang, C., Tian, Y., Liu, G., Gul, H., Wang,
879 D., Tian, Y., Yang, C., Meng, M., Yuan, G., Kang, G., Wu, Y., Wang, K., Zhang, H., Wang, D.,
880 Cong, P., 2019. A high-quality apple genome assembly reveals the association of a
881 retrotransposon and red fruit colour. Nat Commun 10, 1494.
882 Zhang, Z., Li, J., Zhao, X.Q., Wang, J., Ka-Shu, W.G., Yu, J., 2006. KaKs_Calculator: Calculating Ka
883 and Ks Through Model Selection and Model Averaging. Genomics,Proteomics & Bioinformatics
884 4, 259-263.
885 Zhang Z, L.J., Zhao X Q, et al. KaKs_Calculator: calculating Ka and Ks through model selection and
886 model averaging.[J]. Genomics,Proteomics & Bioinformatics, 2006, 4(4):259-263.Li, L.,
887 Stoeckert, C.J., Jr., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic
888 genomes. Genome Res 13, 2178-2189.

889

890 **Acknowledgements**

891 Financial support was provided by the National Science and Technology Major Project of
892 China (No. 2018ZX08021002-002-004), the National Natrual Science Foundation of China
893 (No. 31870652, 31570661, 31170631), the National High Technology Research and
894 Development Program (No. 2013AA102703) and the Major State Basic Research
895 Development Program (No. 2012CB114505). We thank Professor Zhiyi Zhang for his
896 forward-looking suggestions, also thank Shuhua Mu and Zewu An for figure editing.

897 **Author contributions**

898 Xinmin An designed and managed the project. Rengang Zhang led the genome sequencing,
899 assembly and analyses. Jianfeng Mao designed and led evolutionary analyses. Steven H.
900 Strauss contributed to scientific analysis and interpretation, and edited manuscript. Stephen R.
901 Keller participed in writing manuscript. Kai Gao and Y.L. created anther plant. Kai Gao,
902 Zhong Chen, J.L., X.Y., X.Y.Y., J.Z., T.Y.Z., T.G., S.H., D.Y.M., W.K., B.G., S.W.L., and
903 Nada prepared and collected all plant materials. J.W., B.Q.L., W.H.L., and Q.Z.Y. prepared
904 RNA samples. P.R., M.X.Y., and L.X.J., preformed transcriptome assembly and anlysis.
905 Xinmin An, Kai Gao, and Zhong Chen wrote the manuscript with input from other authors.
906 All authors approved the manuscript before submission.

907

Data availability

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 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*).

Additional information

Supplementary information accompanies this paper

Competing interests: The authors declare no competing interests.

938 **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	2,407
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	1,618
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
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	1,345	1,331	3,170
tRNA number	308	308	662
rRNA number	64	61	436
other non-coding gene number	973	962	2,072
Total gene number	29,857	29,936	62,294

939

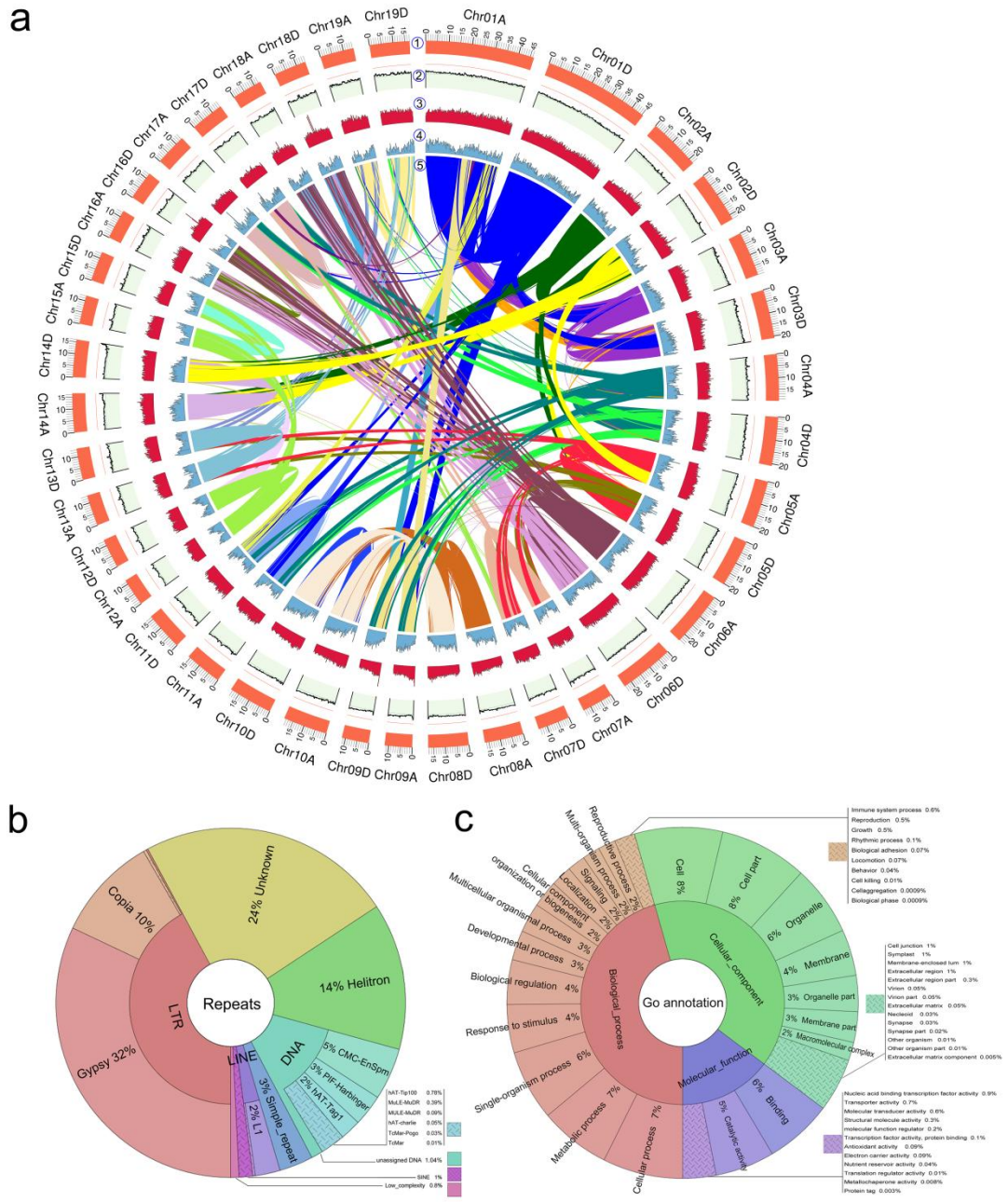


Figure 1 | Characterization of the *Populus tomentosa* genome. a. *Populus tomentosa* genome overview. Genome features in 1-Mb intervals across the 38 chromosomes. Units on the circumference show megabase values and chromosomes. ① Choromosome karyotype; ② GC content (33.6 %. red line 50 %, green line 30 %). ③ Repeat coverage (45–1,937 repeats per Mb). ④ Gene density (3–164 genes per Mb). ⑤ The innermost parts are homologous blocks (1,463 genes) from paralogous synteny analysis. **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.

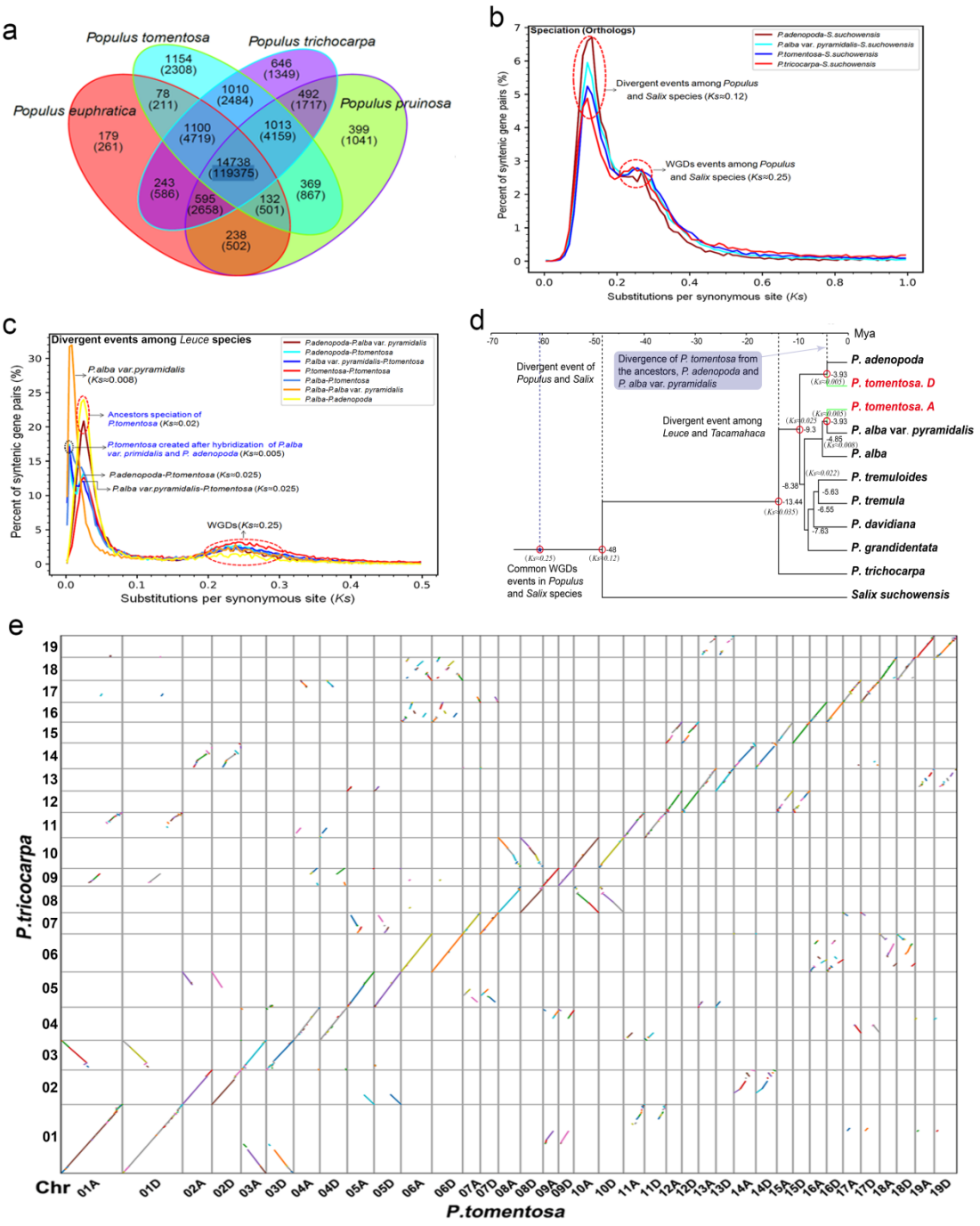


Figure 2 | Comparative genomic analysis of *P. tomentosa* with other plants. (**a**) Shared gene families among *P. tomentosa* and three other poplars. The numbers indicate the number of genes (within each category). **b** Interspecific divergence in *Salicaceae* species, and intraspecific divergence 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 Mya) in *Salix* and *Populus* species (Dai et al., 2014; Ma et al., 2019), *P. tomentosa* speciation (*Ks* = 0.05, ~3.93 Mya) and divergence events of other poplars as revealed through *Ks* analysis. (**d**)

957 Inferred phylogenetic tree across 11 plant species using r8s (Sanderson, 2003). WGD events of
958 *Salicaceae* species are placed. (**e**) Synteny between the *P. tomentosa* genome (the horizontal axis)
959 and *P. trichocarpa* genome (the vertical axis). The *P. tomentosa* chromosomes were inferred to be
960 syntenous with *P. trichocarpa* chromosomes based on orthologous genes from OrthoMCL analysis.

961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983

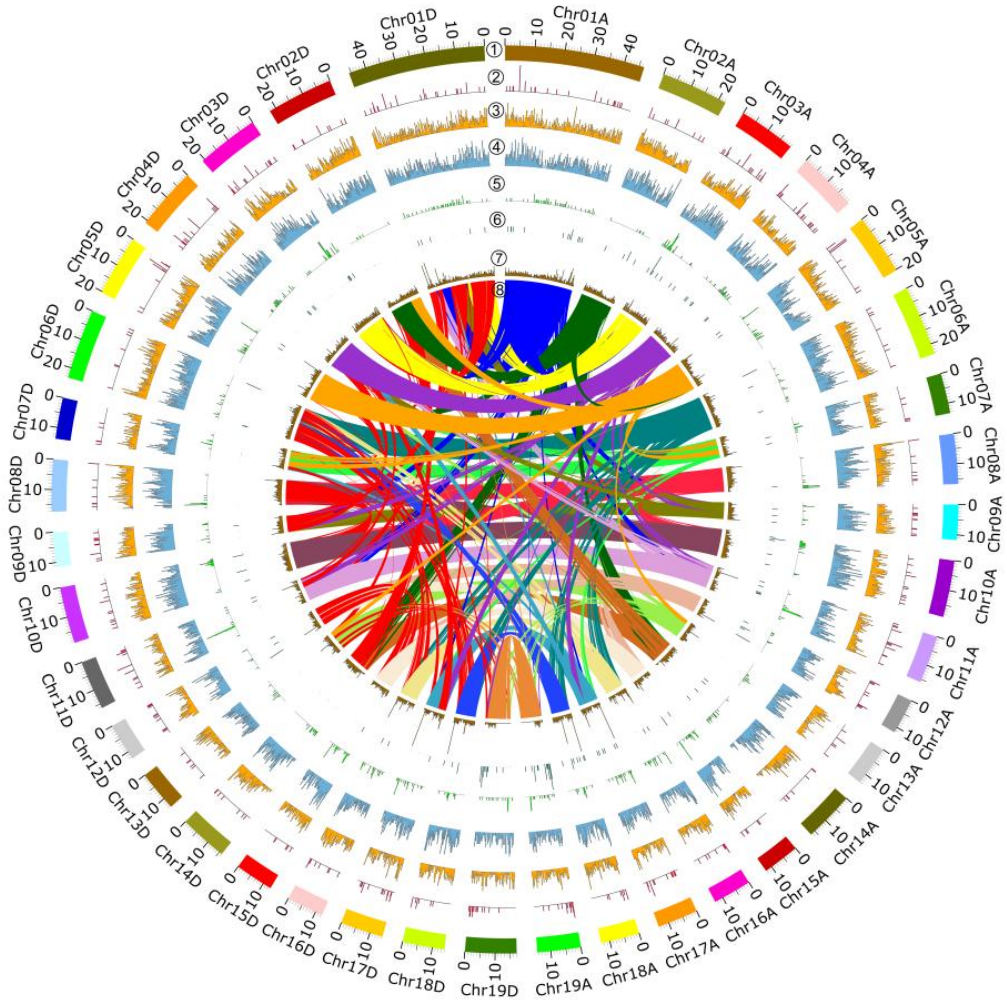


Figure 3 | Synteny, structural variations and allele-indels analyses between subgenome A and subgenome D in *P. tomentosa*. ① Chromosome karyotype, ②Genomic distributions of copy number variations (CNV), ③ Genomic distributions of deletions (DEL), ④ Genomic distributions of insertions (INS), ⑤ Genomic distributions of inversions (INV), ⑥ Genomic distributions of translocations (TRANS), ⑦Genomic 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.

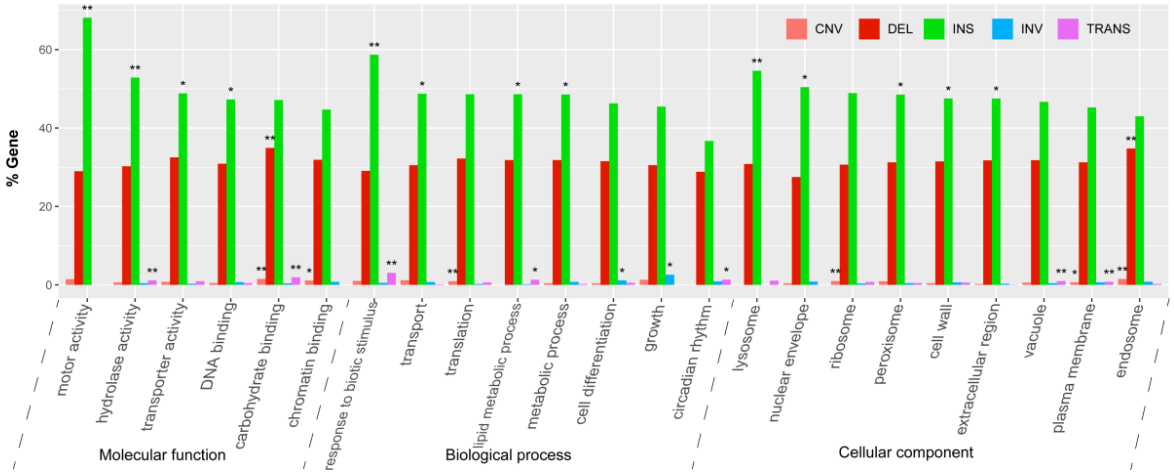


Figure 4 | Functional classification of GO annotations of genes associated with chromosome structural variations. Frequencies of the Gene Ontology terms for which an over-representation has been observed when comparing the subsets of genes included in copy number variations (CNV), deletions (DEL), insertions (INS), inversions (INV), and translocations (TRANS) with respect to the complete dataset of *P. trichocarpa* annotated genes (ALL). * *P*-value<0.05, ** *P*-value<0.01.

Fig.1

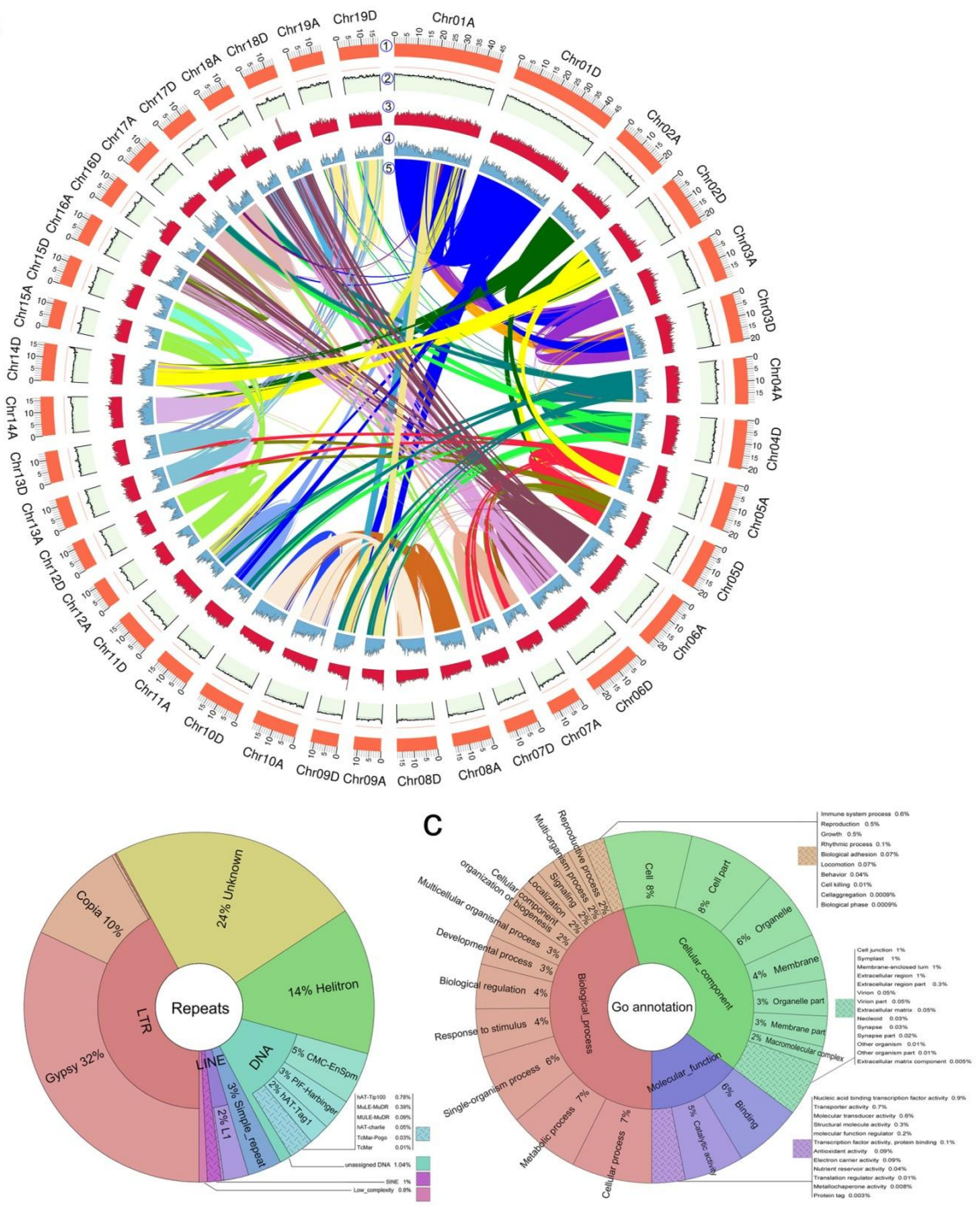

a
Chr01A
Chr01D
Chr02A
Chr02D
Chr03A
Chr03D
Chr04A
Chr04D
Chr05A
Chr05D
Chr06A
Chr06D
Chr07A
Chr07D
Chr08A
Chr08D
Chr09A
Chr09D
Chr10A
Chr10D
Chr11A
Chr11D
Chr12A
Chr12D
Chr13A
Chr13D
Chr14A
Chr14D
Chr15A
Chr15D
Chr16A
Chr16D
Chr17A
Chr17D
Chr18A
Chr18D
Chr19A
Chr19D
b
Repeats
LTR
Gypsy 32%
Copia 10%
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-charle 0.05%
TcMar-Pogo 0.03%
TcMar 0.01%
unassigned DNA 1.04%
SINE 1%
Low_complexity 0.6%
c
Go annotation
Biological_process
Cellular_component
Molecular_function
Reproductive process 2%
Multi-organism process 2%
Cellular component organization or biogenesis 2%
Localization 2%
Signaling 2%
Multicellular organismal process 2%
Developmental process 3%
Biological regulation 4%
Response to stimulus 4%
Single-organism process 6%
Metabolic process 7%
Cellular process 7%
Cell 8%
Cell part 8%
Organelle 6%
Membrane 4%
Organelle part 3%
Membrane part 3%
Macromolecular complex 2%
Binding 6%
Catalytic activity 5%
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%
Nucleoid 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%

Fig.2

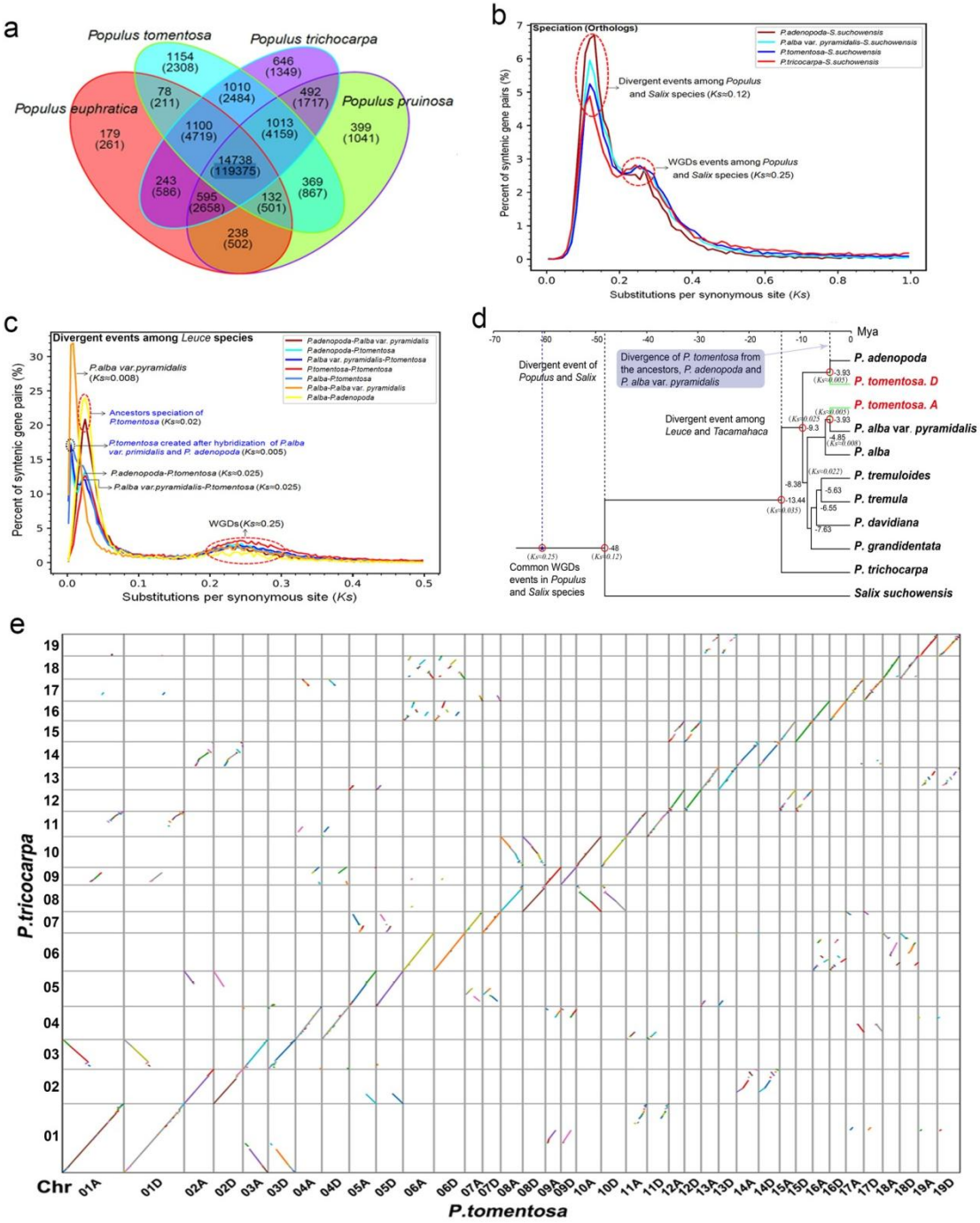

a
Populus tomentosa
Populus trichocarpa
Populus euphratica
Populus pruinosa
1154 (2308)
646 (1349)
78 (211)
1010 (2484)
492 (1717)
179 (261)
1100 (4719)
1013 (4159)
399 (1041)
14738 (119375)
243 (586)
595 (2658)
132 (501)
369 (867)
238 (502)
b
Speciation (Orthologs)
P.adenopoda-S.suchowensis
P.alba var. pyramidalis-S.suchowensis
P.tomentosa-S.suchowensis
P.tricocarpa-S.suchowensis
Divergent events among Populus and Salix species (Ks≈0.12)
WGDs events among Populus and Salix species (Ks≈0.25)
Percent of syntenic gene pairs (%)
Substitutions per synonymous site (Ks)
c
Divergent events among Leuce species
P.adenopoda-P.alba var. pyramidalis
P.adenopoda-P.tomentosa
P.alba var. pyramidalis-P.tomentosa
P.tomentosa-P.tomentosa
P.alba-P.tomentosa
P.alba-P.alba var. pyramidalis
P.alba-P.adenopoda
P.alba var.pyramidalis (Ks≈0.008)
Ancestors speciation of P.tomentosa (Ks≈0.02)
P.tomentosa created after hybridization of P.alba var. primidalis and P. adenopoda (Ks≈0.005)
P.adenopoda-P.tomentosa (Ks≈0.025)
P.alba var.pyramidalis-P.tomentosa (Ks≈0.025)
WGDs(Ks≈0.25)
Percent of syntenic gene pairs (%)
Substitutions per synonymous site (Ks)
d
Mya
Divergent event of Populus and Salix
Divergence of P. tomentosa from the ancestors, P. adenopoda and P. alba var. pyramidalis
Divergent event among Leuce and Tacamahaca
Common WGDs events in Populus and Salix species
P. adenopoda
P. tomentosa. D
P. tomentosa. A
P. alba var. pyramidalis
P. alba
P. tremuloides
P. tremula
P. davidiana
P. grandidentata
P. trichocarpa
Salix suchowensis
e
P.tricocarpa
P.tomentosa
Chr

Fig.3

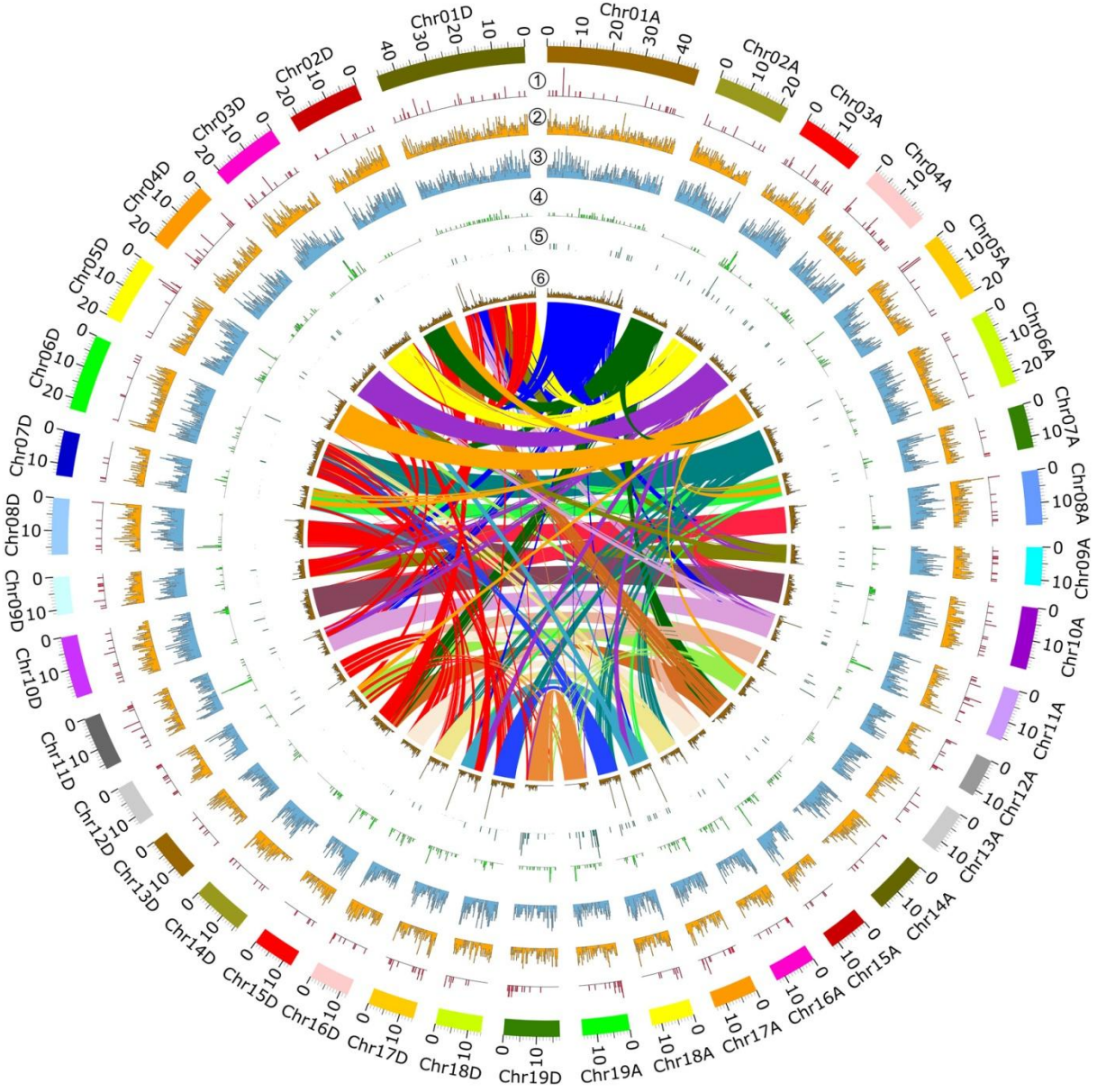
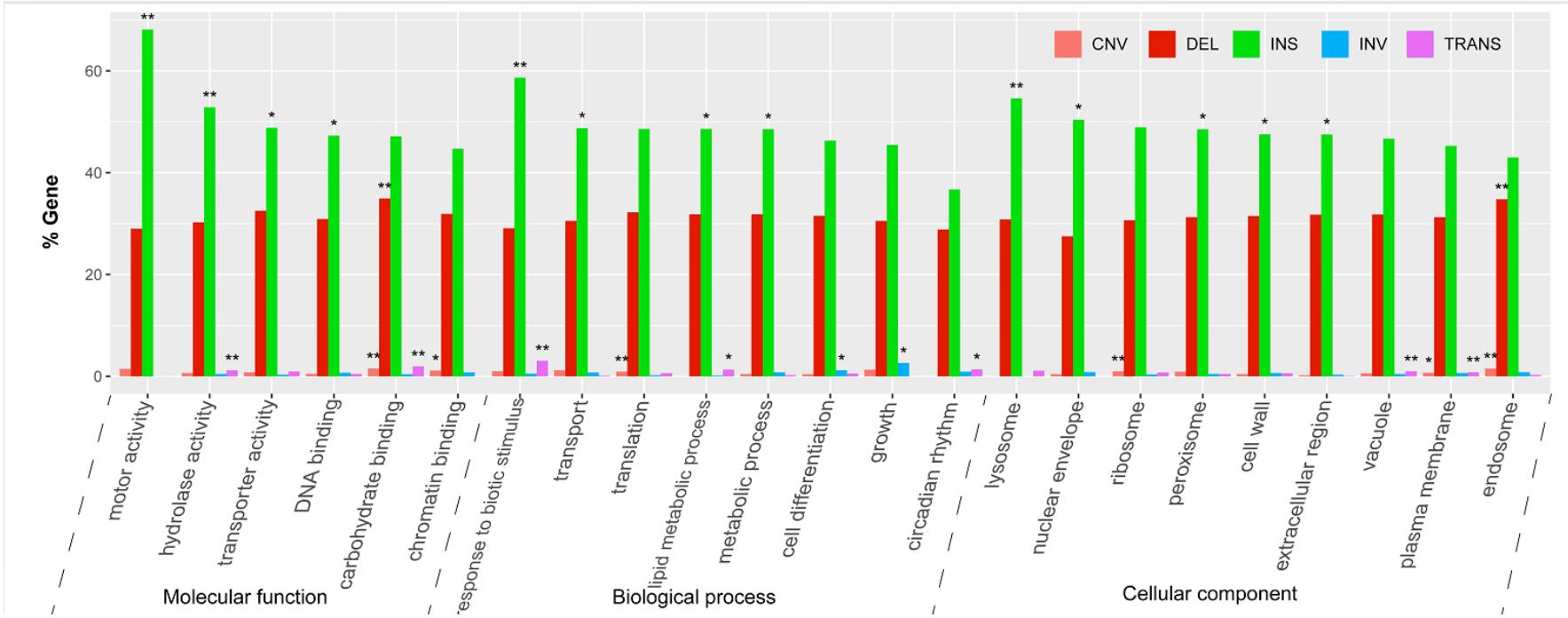

Chr01D
Chr01A
Chr02D
Chr02A
Chr03D
Chr03A
Chr04D
Chr04A
Chr05D
Chr05A
Chr06D
Chr06A
Chr07D
Chr07A
Chr08D
Chr08A
Chr09D
Chr09A
Chr10D
Chr10A
Chr11D
Chr11A
Chr12D
Chr12A
Chr13D
Chr13A
Chr14D
Chr14A
Chr15D
Chr15A
Chr16D
Chr16A
Chr17D
Chr17A
Chr18D
Chr18A
Chr19D
Chr19A
①
②
③
④
⑤
⑥

Fig.4


CNV
DEL
INS
INV
TRANS
% Gene
0
20
40
60
motor activity
hydrolase activity
transporter activity
DNA binding
carbohydrate binding
chromatin binding
response to biotic stimulus
transport
translation
lipid metabolic process
metabolic process
cell differentiation
growth
circadian rhythm
lysosome
nuclear envelope
ribosome
peroxisome
cell wall
extracellular region
vacuole
plasma membrane
endosome
Molecular function
Biological process
Cellular component