DNA Research, 2019, 26(5), 423–431
doi: 10.1093/dnares/dsz020
Advance Access Publication Date: 3 October 2019
Full Paper

OXFORD

Full Paper

The genome of *Populus alba* x *Populus tremula* var. *glandulosa* clone 84K

Deyou Qiu[1]*,†, Shenglong Bai[2†], Jianchao Ma[2†], Lisha Zhang[1], Fenjuan Shao[1], Kaikai Zhang[1,3], Yanfang Yang[1], Ting Sun[2], Jinling Huang[2], Yun Zhou[2], David W. Galbraith[2,4], Zhaoshan Wang[1], and Guiling Sun[2]*

[1]State Key Laboratory of Tree Genetics and Breeding, Key Laboratory of Tree Breeding and Cultivation of National Forestry and Grassland Administration, The Research Institute of Forestry, Chinese Academy of Forestry, Beijing 100091, China, [2]Key Laboratory of Plant Stress Biology, State Key Laboratory of Cotton Biology, School of Life Sciences, Henan University, Kaifeng 475004, China, [3]College of Horticulture, Agricultural University of Hebei, Baoding 071001, China, and [4]School of Plant Sciences and Bio5 Institute, The University of Arizona, Tucson, AZ 85721, USA

*To whom correspondence should be addressed: Tel. +86 010 62889641. Fax. +81 010 22872015. Email: qiudy@caf.ac.cn (D.Q.); Tel. +86 0371 23881387. Fax. +81 0371 23881387.
Email: sung@vip.henu.edu.cn (G.S.)

†These authors contributed equally to this work.
Edited by Dr Satoshi Tabata

Received 28 April 2019; Editorial decision 23 August 2019; Accepted 9 September 2019

Abstract

Poplar 84K (*Populus alba* x *P. tremula* var. *glandulosa*) is a fast-growing poplar hybrid. Originated in South Korea, this hybrid has been extensively cultivated in northern China. Due to the economic and ecological importance of this hybrid and high transformability, we now report the *de novo* sequencing and assembly of a male individual of poplar 84K using PacBio and Hi-C technologies. The final reference nuclear genome (747.5 Mb) has a contig N50 size of 1.99 Mb and a scaffold N50 size of 19.6 Mb. Complete chloroplast and mitochondrial genomes were also assembled from the sequencing data. Based on similarities to the genomes of *P. alba* var. *pyramidalis* and *P. tremula*, we were able to identify two subgenomes, representing 356 Mb from *P. alba* (subgenome A) and 354 Mb from *P. tremula* var. *glandulosa* (subgenome G). The phased assembly allowed us to detect the transcriptional bias between the two subgenomes, and we found that the subgenome from *P. tremula* displayed dominant expression in both 84K and another widely used hybrid, *P. tremula* x *P. alba*. This high-quality poplar 84K genome will be a valuable resource for poplar breeding and for molecular biology studies.

Key words: poplar 84K, genome sequencing, subgenome assignment, *P. alba*, *P. tremula*

1. Introduction

The genus *Populus* comprises around 30 economically and ecologically important species.[1] Besides wood products, these species provide a range of services including bioenergy production, carbon sequestration, bioremediation, nutrient cycling, biofiltration, and habitat diversification.[2] Due to its modest genome size, rapid growth rate, simple vegetative propagation, and short rotation cycle, coupled to high levels of genetic diversity as well as facile genetic

manipulation, this genus has become a model for the study of tree molecular biology.[3] Up to now, several high-quality poplar genomes have become available, including *P. trichocarpa* (Torr. and Gray),[4] *P. euphratica* Oliv,[5] *P. pruinosa*,[6] and *P. alba* var. *pyramidalis*.[7] Unfortunately, these species are not easily transformed by *Agrobacterium tumefaciens*.[8] Interestingly, it seems that only hybrid poplar species, which have much higher transformation rates, have been widely used for genetic transformation, examples including *P. alba* x *P. tremula var. glandulosa* clone 84K,[9] *P. tremula* x *P. alba*,[10] *P. alba* x *P. grandidentata*,[11] and *P. alba* x *P. tremula*.[12]

White poplar *P. alba*, commonly called silver poplar, is widely distributed in central Europe and central Asia. It requires abundant light and ample moisture, and stands up well to flooding. Due to its attractive green-and-white leaves, it is often planted as an ornamental tree. Its extensive root system and good tolerance of salt make it an effective tree on windy coasts. This species has also been widely used in short rotation plantation for timber and pulp production. *P. tremula*, commonly called aspen, is widely distributed in cool temperate regions of Europe and Asia. It is usually found in mountains at high altitudes. *P. tremula* is a very hardy species and can tolerate long, cold winters. It has been widely planted for timber, firewood, and veneer production. Natural populations of *P. alba* and *P. tremula* or its varieties hybridize frequently and have been selected as the parental species for artificial hybrid breeding. Two main hybrids, *P. tremula* x *P. alba* and *P. alba* x *P. tremula var. glandulosa*, have been frequently used in molecular biological studies due to their high rates of transformation. In 2016, Mader et al. reported a 900 Mb draft genome of a female interspecific hybrid *P. tremula* x *P. alba* clone INRA 717-1B4, with a N50 contig length of 3,850 bp.[13] However, the genome sequence of this hybrid has not been assembled to the level of chromosomes. Evidently, further development of high-quality genome sequences in hybrid poplar species having high transformation efficiencies will be critical to advances in molecular biology and genetic engineering of this woody genus.

Here we focus on the clone 84K, a male *P. alba* x *P. tremula var. glandulosa* interspecific hybrid. This hybrid resulted from a breeding program led by Professor SinKyu Hyun (Seoul National University, Korea), and was first introduced into China in 1984 by Professor Qiwen Zhang (The Research Institute of Forestry, Chinese Academy of Forestry, Beijing). Attractive characteristics of this interspecific hybrid include a high growth rate and excellent adaptation to diverse environments. This superior clone has been commonly used in short rotation plantation for timber, firewood, and pulp production. More importantly, it is easily accessible for genetic transformation using *A. tumefaciens*, and this clone is widely used by scientists in transgenic experiments as a model for woody species.[14] Many transgenic poplar lines based on this clone have been created for commercial applications in China. Regrettably, no genome sequence is currently available for this clone. In this study, we describe a *de novo* assembly of the genome sequence of the hybrid poplar (*P. alba* x *P. tremula* var. *glandulosa*) clone 84K (poplar 84K, hereafter), and identify two subgenomes via comparison to the genomes of *P. alba* var. *pyramidalis* and *P. tremula*. The genome resources of this hybrid will facilitate further gene functional analyses, optimization of genetic transformation experiments, and poplar breeding practices, as well as comparative genomic analyses across different poplars.

2. Materials and methods

2.1. Plant materials

Poplar 84K (*P. alba* x *P. tremula var. glandulosa*) was grown in the greenhouse at 20 °C, under a 12-h light/12-h dark illumination cycle. Fresh young leaves were collected from 1-month-old plants and immediately frozen in liquid nitrogen. Genomic DNA was isolated for library construction following the steps of the CTAB method described by Doyle et al.[15] Total RNA of the leaves and shoots from 1-month-old plants was extracted with TRI Reagent (Sigma, St. Louis, MO) for transcriptome sequencing (RNA-seq).

2.2. Genome sequencing and RNA-seq

Two DNA libraries with 270 and 500 bp insertions, and one DNA library with 20 kb insertions were constructed and separately sequenced on Illumina NovaSeq 6000 and PacBio Sequel platforms. Detailed methods for DNA library construction can be found in Supplementary Materials. For transcriptome sequencing (RNA-seq), four RNA libraries for leaves and one library for shoots were constructed according to the TrueSeq® RNA Sample Preparation protocol, and were sequenced on an Illumina NovaSeq sequencing system.

2.3. Estimation of genome size

The 1C value of poplar 84K was measured using flow cytometry with propidium iodide (PI) as the DNA stain and *Arabidopsis thaliana* (col-0) as the standard plant as described previously.[16] The genome size of poplar 84K was calculated using the equation provided by the regression analysis and the known *A. thaliana* 2C DNA content (see Supplementary Materials for details). We also used 28.44 Gb of Illumina NovaSeq short reads to estimate the genome size and other features using the GCE software based on *k-mer* depth-frequency distribution.[17] The versions and main parameters of GCE and other software packages used in this study are provided in Supplementary Table S1.

2.4. Assembly of the poplar 84K genome sequences

SMRT subreads were corrected, trimmed, and assembled using CANU[18] (see Supplementary Materials for details), and then polished the draft assembly using Arrow.[19] Lastly, Pilon[20] was used to perform two rounds of error correction using Illumina NovaSeq reads from the 270 and 500 bp insert libraries.

2.5. Identification of potential contamination and organelle genomes

For contigs less than 1 Mb, we performed a BLASTN[21] search against the non-redundant nucleotide (NT, downloaded on 4 March 2018) database in GenBank with an *E*-value of 1E-5. Contigs having the most matches to non-plant species were designated as environmental sequence contamination. For the identification of the chloroplast genome, the complete chloroplast genome sequence of *P. tremula* x *P. alba* (accession number: NC_028504.1)[22] was used as a query to search against all the remaining PacBio contigs with an *E*-value of 1E-5. The mitochondrial genome of *P. tremula* x *P. alba* (accession number: NC_028329.1)[22] was chosen to search against all remaining PacBio contigs with an *E*-value of 1E-5 for mitochondrial fragments. Gene structure annotations and figures of the organelle genomes were produced using GESEQ.[23]

2.6. Pseudomolecule construction

Hi-C mapping was employed to facilitate pseudomolecule construction and the details were described in Supplementary Materials. The PacBio contigs were divided into fragments having a length of 300 kb, and were then error corrected, clustered, ordered, and orientated by the LACHESIS software[24] operating on the valid interaction read pair dataset. Finally, contact maps were plotted using the HiCPlotter software.[25]

2.7. Evaluation of genome completeness, continuity, and accuracy

To evaluate the completeness of genome assembly, we checked the mapping rates by aligning RNA-seq reads from five libraries and DNA short reads from four libraries to the final assembly using HISAT2[26] and BWA-MEM,[27] respectively, and performed BUSCO analysis.[28] We assessed the continuity of the final assembly of poplar 84K along with two other poplar genomes, *P. trichocarpa*[4] and *P. deltoides* (assembly version 2.1, produced by the US Department of Energy Joint Genome Institute in collaboration with the user community), using LTR_retriever[29] by calculating the LTR Assembly Index (LAI) score that evaluates the *de novo* assembly quality of intergenic and repetitive regions. Raw LAI is defined as the ratio of intact LTR retrotransposon length to total LTR sequence length.[29] After standardization, the corrected LAI can be used as a reference-free genome metric regardless of genome size, LTR retrotransposon content, and gene content completeness.[29] The accuracy of the final assembly was estimated by aligning Illumina short reads using BWA-MEM[27] and GATK[30] to call variants.

2.8. Gene prediction and annotation

Tandem repeats within the poplar 84K genome were identified with Tandem Repeat Finder.[31] To identify known transposable elements (TEs), we used RepeatMasker,[32] loaded with the Repbase[33] and the Dfam[34] library databases, Homology-based ncRNA annotation was performed by mapping plant miRNA and snRNA genes from the Rfam database (release 14.0)[35] to the poplar 84K genome using infernal.[36] tRNAscan-SE[37] was used for tRNA annotation. RNAmmer[38] was used to predict rRNAs and their subunits. Protein-coding genes were predicted based on transcriptomic, homologous, and *de novo* methods, the details of which along with functional annotation of the predicted gene models were described in Supplementary Materials.

2.9. Subgenome assignment

The chromosome pairs were firstly identified with MCScanX,[39] based on protein collinearity between all the 38 Hi-C linkage groups in poplar 84K and all the 19 chromosomes of *P. trichocarpa*. The pairwise relationship of the chromosomes was further confirmed by their DNA-level collinearity identified using nucmer.[40] For each chromosome pair, the chromosomes derived from the parental species were distinguished by similarity to the genomes of related species *P. alba* var. *pyramidalis*[7] and *P. tremula*,[41] and were assigned to subgenome A and G, respectively.

2.10. Evolutionary analysis of chloroplast and mitochondrial genome

Phylogenetic tree reconstruction of chloroplast and mitochondrial protein coding regions was described in Supplementary Materials. The organelle-DNA similar fragments in nuclear genome were detected using BLASTN with the chloroplast and mitochondrial contigs as query to search against poplar 84K genome. The parameters of BLASTN and the filtration setting were as described previously.[42]

2.11. Analysis of genome collinearity and variation between subgenome A and G

Collinear blocks between subgenome A and subgenome G were determined using MCscanX[39] and plotted with Circos,[43] treating at least eight genes as a collinear block. We further aligned the DNA sequences of the two subgenomes using nucmer,[40] and only retained uniquely anchored sequences larger than 10 kb for variance calling by the Assemblytics software.[44]

2.12. Transcriptional profiles of the allelic genes in subgenomes

Being that the RNA-seq experiments performed in this study did not contain replicates and were from few tissues, we downloaded the RNA-seq datasets of four tissues in poplar 84K and of four tissues in *P. tremula* x *P. alba* from GenBank SRA database (Supplementary Table S2) and mapped them to the genome of poplar 84K by HISAT2.[26] Gene expression values were calculated with FPKM and TMP values using the Stringtie software.[45] Allelic genes determined by MCSCANX, and those showing transcriptional bias between subgenome A and G were identified using edgeR[46] with a 4-fold change and a *P*-value of <0.05. The qRT-PCR method was described in Supplementary Materials.

3. Results and discussion

3.1. Genome assembly and identification of potential contamination and organelle genomes

Two DNA libraries carrying 270 and 500 bp insertions, and one DNA library carrying 20 kb insertions were constructed and separately sequenced on Illumina NovaSeq 6000 and PacBio Sequel platforms, producing 56.3 Gb (Supplementary Table S3a) and 48.4 Gb (Supplementary Table S3b) data. The mean and N50 of PacBio subread lengths were 8.1 kb and 13.7 kb, respectively (Supplementary Table S3b, Fig. S1). The haploid genome size of poplar 84K was estimated to be 470.155 ± 5.94 Mb by flow cytometry (Supplementary Fig. S2, Table S4) and 427.2 Mb based on *k-mer* depth distribution (Supplementary Fig. S3). Heterozygosity was estimated to be 2.16% using GCE software.[17] A 753.8-Mb genome assembly of poplar 84K was constructed (Table 1), comprising 1,384 contigs and a contig N50 of 2.24 Mb. The difference between assembly size and estimated genome size may be due to the heterozygous character of poplar 84K and was investigated further in the following analysis. To remove fungal and bacterial DNA contamination, for 1,142 contigs less than 1 Mb (total 157,986,688 bp), we performed a BLASTN[21] search against the non-redundant nucleotide (NT) database in GenBank with an *E*-value of 1E-5. A total of 5,264,546 bp from 114 contigs were identified as bacterial or fungal contamination, and was excluded from further analysis. Three contigs with length of 40,693, 115,255, and 64,703 bp were found with high identities to the chloroplast sequence of *P. tremula* x *P. alba* (NC_028504.1).[22] The chloroplast genome (156,462 bp) was then assembled from these contigs based on a collinear relationship with the chloroplast sequence of *P. tremula* x *P. alba* (NC_028504.1).[22] One contig of 874,696 bp was found to encode the mitochondrial

Table 1. Statistics of genome and subgenome assembly of poplar 84K using different sequencing data

Length (bp) /number	PacBio assembly	Hi-C assembly	Pseudomolecules	Subgenome A	Subgenome G
N90	429,473	12,722,015	13,367,995	13,017,457	13,367,995
N80	959,980	13,867,023	14,133,090	13,907,236	14,133,090
N70	1,381,280	14,269,014	16,373,422	16,345,784	16,373,422
N60	1,818,021	17,170,365	17,428,093	17,170,365	17,629,401
N50	2,239,559	19,641,611	19,971,399	18,769,030	20,657,227
N40	2,832,264	20,657,227	20,657,227	20,018,725	21,475,114
N30	3,174,752	21,475,114	21,827,254	21,194,248	22,685,518
N20	4,522,243	23,093,407	24,417,576	24,417,576	25,197,140
N10	6,295,541	49,053,667	49,053,667	49,053,667	49,243,677
Maximum length	12,124,769	49,243,677	49,243,677	49,053,667	49,243,677
Minimum length	1,245	1,245	6,540,224	11,888,367	6,540,224
Total length	753,822,854[a]	747,538,837[b]	710,053,368	356,027,211	354,026,157
Total sequences	1,384	841	38	19	19
Gap numbers	—	505	505	227	278

[a] The original assembly obtained using CANU.
[b] Popar 84K nuclear genome assembly after removing organellar sequences and microorganism contamination.

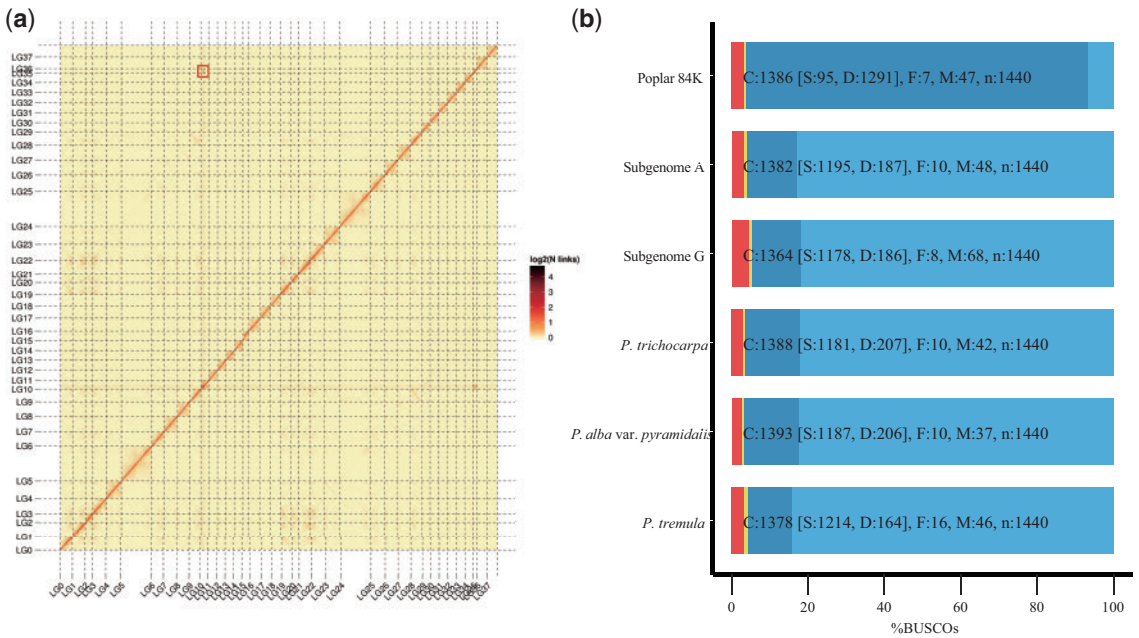

Figure 1. Assessment of the assembled 84K genome. (a) Interaction frequency distributions of Hi-C linkage groups. The log2 of the valid interaction link number of Hi-C data between any pair of 500 kb non-overlapping bins were calculated and is displayed as a heatmap by HiCPlotter.[25] The black/white bar of the heatmap indicates the interaction frequency of the Hi-C links. The square indicates the abnormal high frequency of interaction between linkage group 35 (chromosome 11 of subgenome G) and linkage group 10 (chromosome 12 in subgenome A). (b) BUSCO analysis of the genomes from poplar 84K, P. trichocarpa, P. alba var. pyramidalis, and P. tremula. (C) complete; (S) single-copy; (D) duplicated; (F) Fragmented; (M) Missing.

genome. Gene structure annotations and figures of the organelle genomes were produced using GESEQ (Supplementary Figs S4 and S5).[23] After assignment of microbial contamination and organelles, 747.5 Mb from poplar 84K nuclear genome were obtained. With the 95.2 Gb clean reads generated by Hi-C sequencing (Supplementary Table S3d), a total of 1,042 contigs were clustered into 38 groups for pseudomolecule construction. Of these, 544 contigs, with total length of 710 Mb, were ordered and oriented in all linkage groups (Fig. 1a, Table 1).

3.2.Evaluation of completeness, continuity, and accuracy of final assembly

We found 95.64–97.88% of the RNA-seq reads and 95.74–97.74% of the DNA short reads could be aligned to the final assembly (Supplementary Table S5). BUSCO analysis[27] showed that 1,386 (96.3%) of 1,440 plant single-copy orthologues were complete, but 1,291 (93.1%) of them presented as duplicated copies. Similar percentages of the 1,440 plant single-copy orthologues were detected in *P. trichocarpa*,[4] *P. alba* var. *Pyramidalis*,[7] and *P. tremula*,[41] and at

least 85.1% of them were single copy in these species (Fig. 1b). This suggested that the heterozygous regions were obtained, based on which these heterozygous regions could be separated using their parental genome information.

We assessed the continuity of the final assembly of poplar 84K in comparison to two poplar reference genomes, *P. trichocarpa*[4] and *P. deltoides* (assembly version 2.1, produced by the US Department of Energy Joint Genome Institute in collaboration with the user community). The standardized LAI scores are 9.34 in *P. trichocarpa*, 7.95 in *P. deltoides*, and 14.79 in poplar 84K. This implies the poplar 84K assembly has achieved reference genome quality.

The accuracy of the final assembly was estimated by aligning Illumina short reads using BWA-MEM[27] and GATK[30] to call variants. A total of 318 homozygous SNPs and 6,398 homozygous INDELs were considered as errors, and 35,735 heterozygous SNPs and 24,860 heterozygous INDELs were found.

3.3. Genome annotation and subgenome assignment

We identified 184.0 Mb of TE sequence (24.4% of the assembly) (Supplementary Table S6). The largest class of TEs comprised retrotransposons, accounting for 18.2% of the assembly, and consisted mostly of Gypsy and Copia retrotransposon families. DNA transposons accounted for 3.8% of the assembly. These analyses also identified 1,983 miRNAs, 1,312 tRNAs, 1,140 rRNAs, and 1,126 snRNAs (Supplementary Table S7).

Four RNA libraries for leaves and one library for shoots were constructed, and sequenced using the Illumina NovaSeq platform, generating 145.8 million pair reads for genome annotation (Supplementary Table S3c). Finally, 85,755 consensus protein-coding genes were predicted, the average gene length, average transcript length, average CDS length, and exon number per gene being 2,948 bp, 2,937 bp, 1,075 bp and 4.48, respectively (Supplementary Table S8). Functional annotation of the predicted protein-coding genes revealed that 72,574 of the total of 85,755 genes (84.6%) could be assigned putative functions (Supplementary Table S9).

Chromosome pairs were first identified based on protein collinearity between all the 38 Hi-C linkage groups in poplar 84K and all the 19 chromosomes of *P. trichocarpa* (Supplementary Fig. S6). The pairwise relationship of the chromosomes was further confirmed by their DNA-level collinearity (Supplementary Fig. S7). For each chromosome pair, the chromosomes derived from the parental species were distinguished by similarity to the genomes of related species *P. alba* var. *pyramidalis* and *P. tremula* (Supplementary Fig. S8).

The chromosome numbering and orientation were determined by comparison to *P. trichocarpa* (Supplementary Fig. S6 and Table S10). Ultimately, 356 Mb from female parental species *P. alba* and 354 Mb from male parental species *P. tremula* var. *glandulosa* were obtained, and were designated as subgenome A and subgenome G. BUSCO analysis indicated each subgenome has characteristics similar to those of *P. trichocarpa*, *P. alba* var. *pyramidalis* and *P. tremula* (Fig. 1b), which implies high completeness of the two subgenomes. Specifically, the numbers of duplicated genes dropped from 1,291 (89.7%) to 187 in subgenome A and to 186 in subgenome G. This supports our previous speculation that combination of two haploid genomes causes the assembly size to increase by 30%.

3.4. Characteristics of the poplar 84K genome

The characteristics of the two subgenomes, along with sequencing depths, are shown in Fig. 2, including the chromosome length, gene density, TE content, GC content, and location of collinear regions. Chromosome 11 of subgenome G (6.54 Mb) was much shorter than that of subgenome A (17.42 Mb) (Fig. 2 and Supplementary Table S10). This may be due to large fragment loss during interspecies hybridization and/or subsequent chromosome stabilization, or defects of the current assembly algorithm in dealing with the regions with high identities. Another anomalous region is a 6.92 Mb fragment at the 5′ end of chromosome 12 of subgenome A, which showed a doubled sequencing depth as compared with other regions (Fig. 2). This 6.92 Mb region also showed a high frequency of interaction with the short chromosome 11 of subgenome G (Fig. 1a). We speculate that one additional 6.92 Mb region derived from recent large DNA fragment duplication may be responsible for the connection to chromosome 11 of subgenome G.

Further phylogenetic analyses of chloroplast proteins were conducted. The tree topology indicated that *P. trichocarpa* is closely related to *P. balsamifera* and *P. fremontii* (Supplementary Fig. S9), consistent with previous studies.[19] *P. yunnanensis* formed a clade with *P. alba*, *P. tremula* and *P. tremula* x *P. alba*, within which the poplar 84K genome formed a sister branch with the genome of *P. alba* (Supplementary Fig. S9). The mitochondrial genome in poplar 84K also showed maternal inheritance in that they were mostly related to those from *P. alba* (Supplementary Fig. S10).

The lateral transfer of organelle fragments into the nuclear genome has been widely reported in plants. To get an insight into the scale of such transfer and to exclude the possibility of misidentification of organelle genome in poplar 84K, we analysed the insertion times, sequence identities, and maximum and total lengths of organelle-DNA similar fragments in the nuclear genome. About 2,234 and 2,093 insertions with total length of 533 kb and 530 kb of chloroplast sequences, and 3,916 and 3,844 insertions with total length of 597.2 kb and 625.6 kb of mitochondrial sequences were found in all the chromosomes of subgenome A and G, respectively (Supplementary Table S11). The mean insertion lengths were 236 bp for chloroplasts and 155 bp for mitochondria (Supplementary Fig. S11). The longest chloroplast and mitochondrial insertion lengths were 15.2 kb (98.9% identity) and 19.1 kb (98.1% identity), respectively. The statistics of organelle-DNA similar fragments in the nuclear genome of poplar 84K were similar to the previous report of *P. trichocarpa*.[42]

We performed orthologous assignment and gene family comparison in the two poplar 84K subgenomes and four other *Populus* species using OrthoFinder.[47] Total of 30,342 clusters were obtained in them, among which 17,690 (58.30%) were shared by the four genomes and two poplar 84K subgenomes (Supplementary Fig. S12).

3.5. Genome collinearity and variation between subgenome A and G

About 323 Mb of regions of collinearity, including 28,974 allelic gene pairs, were identified between the two subgenomes (90.7% of subgenome A and 91.2% of subgenome G). We then aligned the DNA sequences of the two subgenomes and only retained uniquely anchored sequences larger than 10 kb for variant calling. We found 5,398,437 SNPs and 1,108,357 indels (1 bp–10 kb) in these collinear regions. These variations within allelic genes would influence their gene structures, expression patterns, functions, and regulation, which together contribute the morphologic and physiologic characteristics of hybrid 84K.

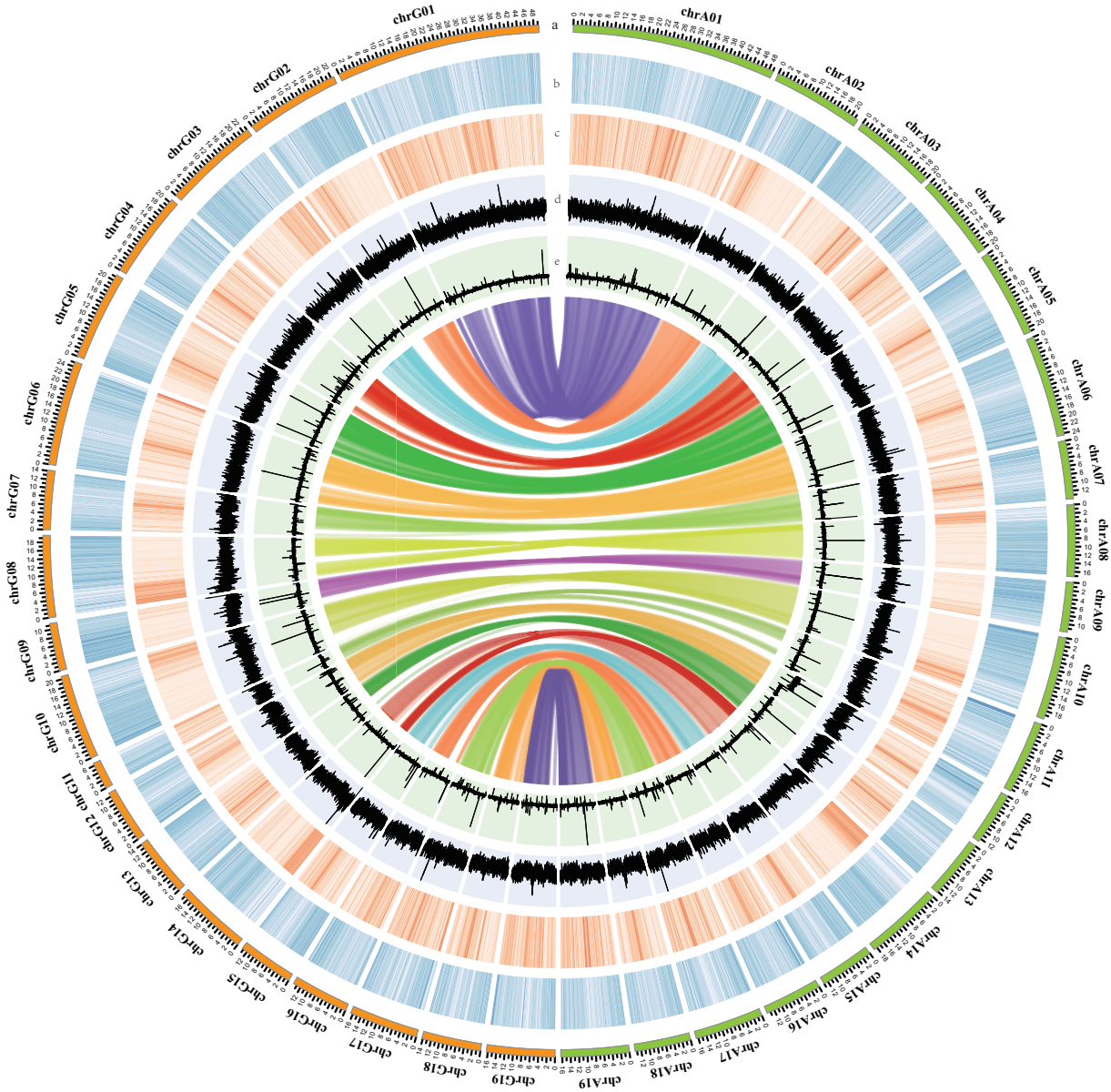

Figure 2. Characteristics of the poplar 84K genome. From the outer edge inward: (a) circles represent the subgenomes A (right) and G (left); (b) gene density on each chromosome; (c) repeat density at 10 kb intervals; (d) GC content at 10 kb intervals; (e) the sequencing depth of Illumina short reads at 10 kb intervals, and collinear blocks linked by grey lines.

3.6. Transcriptional bias of the allelic genes in the two subgenomes

We further explored the expressional bias of allelic genes in the two subgenomes by mapping RNA-seq data of four tissues from 84K and four tissues from backcross hybrid *P. tremula* x *P. alba* to the poplar 84 genome assembly (Supplementary Table S2). Of all the 28,974 allelic gene pairs, 6,780 gene pairs were found with transcriptional bias in poplar 84K, and 5,349 in *P. tremula* x *P. alba* (Fig. 3a and b, Supplementary Table S12). Among these allelic genes with transcriptional bias, 2,121 gene pairs showed bias in all four tissue in poplar 84K, the corresponding number in *P. tremula* x *P. alba* being 1,166 (Fig. 3a and b). We further calculated the number of genes that showed transcriptional bias to subgenomes A and G in two hybrids. Interestingly, we found more than 54.4% of allelic genes showed transcriptional bias to subgenome G, and less than 45.6% of allelic genes to subgenome A in all the tissues that were used (Table 2), which implies subgenome G plays a more important role during the growth process in both hybrids. We selected four allelic gene pairs and performed qRT-PCR analysis in shoots and leaves, which showed expression patterns that were consistent with the RNA-seq data (Supplementary Fig. S13, Table S13). The similarities of promoters were additionally analysed between the allelic pairs showing expression bias. The result indicated that a higher divergence in promoter regions was present in the allelic pairs having expression bias than those without detectable expression bias (Fig. 3c).

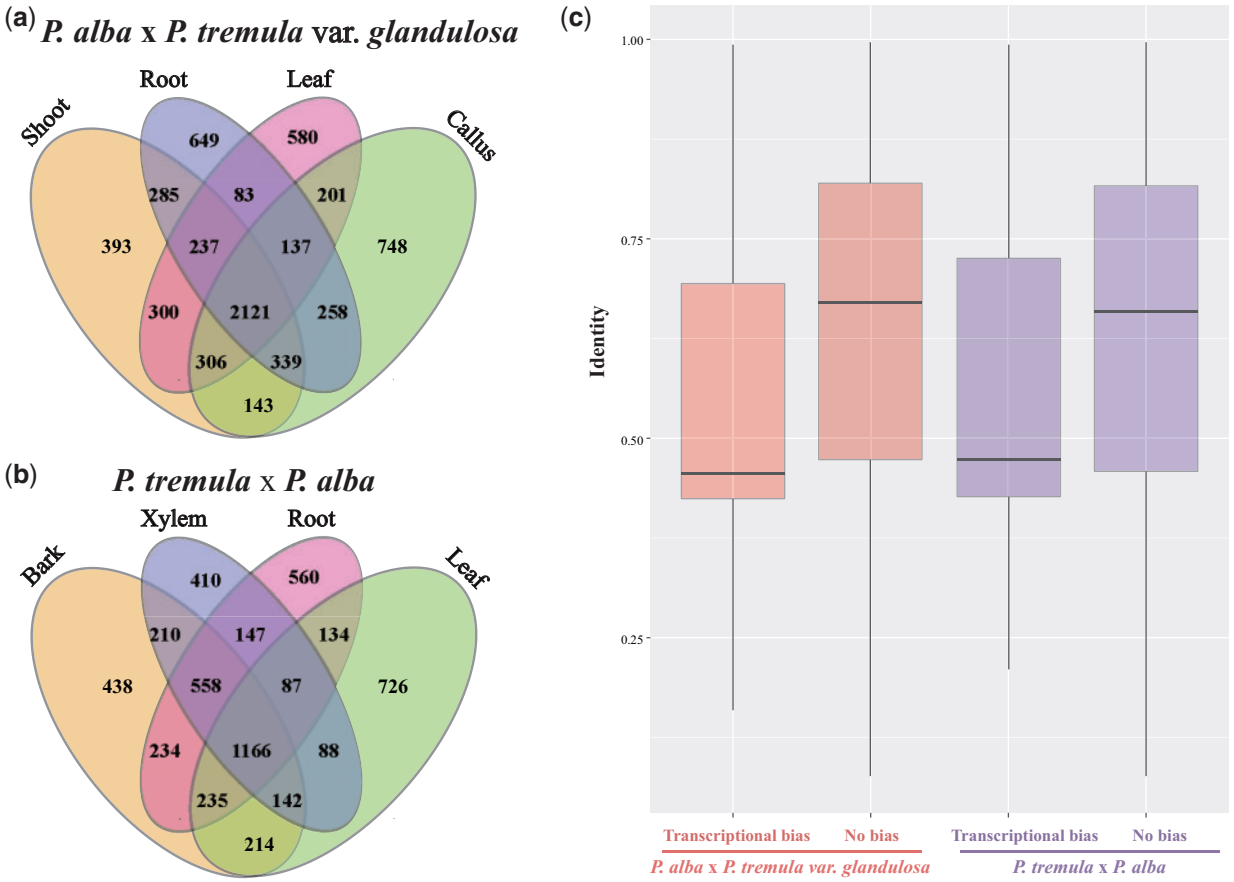

Figure 3. Allelic genes with transcriptional bias and their promoter region identity in poplar 84K and *P. tremula* x *P. alba*. (a) Identified allelic genes with transcriptional bias in poplar 84K. (b) Identified allelic genes with transcriptional bias in *P. tremula* x *P. alba*. (c) Boxplot of the identity in allelic gene promoter region. Significant differences of identity between allelic genes with transcriptional bias and no-bias were supported by the Kolmogorov–Smirnov test (*P*-value < 0.001) in poplar 84K and *P. tremula* x *P. alba*.

Table 2. Genes with transcriptional bias within different tissues in two poplar hybrids

Species	Tissues	DEG numbers	Dominant expression in subgenome A	Dominant expression in subgenome G
P. alba x *P. tremula* var. *glandulosa*	Shoot	4124	1,856 (45.00%)	2,268 (55.00%)
P. alba x *P. tremula* var. *glandulosa*	Rooting	4109	1,826 (44.44%)	2,283 (55.56%)
P. alba x *P. tremula* var. *glandulosa*	Callus	3965	1,802 (45.45%)	2,163 (54.55%)
P. alba x *P. tremula* var. *glandulosa*	Leaf	4253	1,936 (45.52%)	2,317 (54.48%)
P. tremula x *P. alba*	Leaf	2792	1,065 (38.14%)	1,727 (61.86%)
P. tremula x *P. alba*	Bark	3197	1,314 (41.10%)	1,883 (58.90%)
P. tremula x *P. alba*	Xylem	2808	1,150 (40.95%)	1,658 (59.05%)
P. tremula x *P. alba*	Root	3121	1,288 (41.27%)	1,833 (58.73%)

P. alba x *P. tremula* var. *glandulosa* = 84K.

Further examination uncovered 285 allelic gene pairs with bias to subgenome A (Supplementary Fig. S14a) and 339 with bias toward subgenome G, for all the tissues and in both hybrids (Supplementary Fig. S14b). These genes might be essential, playing important functions in both hybrids. Gene enrichment analysis of these two datasets revealed that protein metabolic process and nitrogen compound metabolic process were represented in both subgenomes. Subgenome A has more allelic gene pairs in the categories of monovalent inorganic cation transport and chromatin organization, whereas subgenome G has more in the categories of mRNA processing, ncRNA metabolic process, and sulphur compound metabolic process. As for their cellular component classifications, allelic gene pairs showing bias to subgenome A function in mitochondria and the nucleoplasm, and those in subgenome G play roles in plastids, endosomes, vesicles, and ribosomes (Supplementary Table S14). This implies these genes function in different cellular compartments and different biological processes.

More transcriptome data will be needed to obtain further insights into the transcriptional bias of each subgenome in different developmental stages under different stresses.

4. Summary and conclusions

We employed PacBio single-molecular real-time sequencing and Hi-C technology to generate a reference sequence of the poplar 84K (*Populus alba* x *Populus tremula* var. *glandulosa*) genome. The genome sequences were assembled into the chromosome levels, with a contig N50 size of 1.99 Mb and a scaffold N50 size of 19.6 Mb. About 356 Mb from the female parental species (*P. alba*) and 354 Mb from the male parental species (*P. tremula* var. *glandulosa*) were assigned. The two subgenomes showed high collinearity over 323 Mb. The allelic gene pairs with transcriptional bias toward subgenome A or G function in different cellular compartments and different biological processes. More allelic gene pairs showed transcriptional bias toward subgenome G in poplar 84K and another widely used hybrid *P. tremula* x *P. alba*. The dominant expression of subgenome G indicates that it plays a more important role than subgenome A during hybrid growth. The high-quality genome of poplar 84K provides an important gene resource for poplar breeding and molecular biology research.

Data availability

The sequencing reads from each sequencing library have been deposited at NCBI with the Project ID PRJNA556338 and CNGB Nucleotide Sequence Archive (CNSA) with the Project ID CNP0000339. Software versions and main parameters are provided in Supplementary Table S1 in Supplementary Materials.

Acknowledgements

This project was supported by the special fund from Key Laboratory of Tree Breeding and Cultivation of State Forestry Administration (ZDRIF201705), the National Natural Science Foundation of China (31771414), and the Program for Science and Technology Innovation Talents in Universities of Henan Province (18HASTIT041).

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at *DNARES* online.

References

1. Stettler, R., Bradshaw, T., Heilman, P. and Hinckley, T. 1996, *Biology of Populus and Its Implications for Management and Conservation*. NRC Research Press, Ottawa.
2. Brunner, A.M., Busov, V.B. and Strauss, S.H. 2004, Poplar genome sequence: functional genomics in an ecologically dominant plant species, *Trends Plant Sci.*, **9**, 49–56.
3. Wullschleger, S.D., Jansson, S. and Taylor, G. 2002, Genomics and forest biology: populus emerges as the perennial favorite, *Plant Cell*, **14**, 2651–5.
4. Tuskan, G.A., Difazio, S., Jansson, S., et al. 2006, The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray), *Science*, **313**, 1596–604.
5. Ma, T., Wang, J.Y., Zhou, G.K., et al. 2013, Genomic insights into salt adaptation in a desert poplar, *Nat. Commun.*, **4**, 2797–805.
6. Yang, W., Wang, K., Zhang, J., Ma, J., Liu, J. and Ma, T. 2017, The draft genome sequence of a desert tree *Populus pruinosa*, *Gigascience*, **6**, 1–7.
7. Ma, J., Wan, D., Duan, B., et al. 2019, Genome sequence and genetic transformation of a widely distributed and cultivated poplar, *Plant Biotechnol. J.*, **17**, 451–60.
8. Song, J., Lu, S., Chen, Z.Z., Lourenco, R. and Chiang, V.L. 2006, Genetic transformation of *Populus trichocarpa* genotype Nisqually-1: a functional genomic tool for woody plants, *Plant Cell Physiol.*, **47**, 1582–9.
9. Wang, S.Y., Chen, Q.J., Wang, W.L., Wang, X.C. and Lu, M.Z. 2005, Salt tolerance conferred by over-expression of OsNHX1 gene in Poplar 84K, *Chinese Sci. Bull.*, **50**, 224–8.
10. Coleman, H.D., Canovas, F.M., Man, H., Kirby, E.G. and Mansfield, S.D. 2012, Enhanced expression of glutamine synthetase (GS1a) confers altered fibre and wood chemistry in field grown hybrid poplar (*Populus tremula* X *alba*) (717-1B4), *Plant Biotechnol J.*, **10**, 883–9.
11. Maloney, V.J. and Mansfield, S.D. 2010, Characterization and varied expression of a membrane-bound endo-beta-1,4-glucanase in hybrid poplar, *Plant Biotechnol J.*, **8**, 294–307.
12. Cho, J.S., Nguyen, V.P., Jeon, H.W., et al. 2016, Overexpression of PtrMYB119, a R2R3-MYB transcription factor from *Populus trichocarpa*, promotes anthocyanin production in hybrid poplar, *Tree Physiol.*, **36**, 1162–76.
13. Mader, M., Le Paslier, M.C., Bounon, R., et al. 2016, Whole-genome draft assembly of *Populus tremula* x *P. alba* clone INRA 717-1B4, *Silvae Genet.*, **65**, 74–9.
14. Shim, D., Kim, S., Choi, Y.I., et al. 2013, Transgenic poplar trees expressing yeast cadmium factor 1 exhibit the characteristics necessary for the phytoremediation of mine tailing soil, *Chemosphere*, **90**, 1478–86.
15. Doyle, J.J. and Doyle, J.L. 1990, Isolation of plant DNA from fresh tissue, *Focus*, **12**, 13–5.
16. Galbraith, D.W., Harkins, K.R., Maddox, J.M., Ayres, N.M., Sharma, D.P. and Firoozabady, E. 1983, Rapid flow cytometric analysis of the cell cycle in intact plant tissues, *Science*, **220**, 1049–51.
17. Liu, B., Shi, Y. and Yuan, J. 2013, Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv*, 1308.2012.
18. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M. 2017, Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation, *Genome Res.*, **27**, 722–36.
19. Chin, C.S., Alexander, D.H., Marks, P., et al. 2013, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data, *Nat. Methods*, **10**, 563–9.
20. Walker, B.J., Abeel, T., Shea, T., et al. 2014, Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement, *PLoS One*, **9**, e112963.
21. Camacho, C., Coulouris, G., Avagyan, V., et al. 2009, BLAST+: architecture and applications, *BMC Bioinformatics*, **10**, 421.
22. Kersten, B., Faivre Rampant, P., Mader, M., et al. 2016, Genome sequences of *Populus tremula* chloroplast and mitochondrion: implications for holistic poplar breeding, *PLoS One*, **11**, e0147209.
23. Tillich, M., Lehwark, P., Pellizzer, T., et al. 2017, GeSeq - versatile and accurate annotation of organelle genomes, *Nucleic Acids Res.*, **45**, W6–W11.
24. Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O. and Shendure, J. 2013, Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions, *Nat. Biotechnol.*, **31**, 1119–25.
25. Akdemir, K.C. and Chin, L. 2015, HiCPlotter integrates genomic data with interaction matrices, *Genome Biol.*, **16**, 198.
26. Kim, D., Langmead, B. and Salzberg, S.L. 2015, HISAT: a fast spliced aligner with low memory requirements, *Nat. Methods.*, **12**, 357–60.
27. Li, H. 2013, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv: 1303.3997*.
28. Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. 2015, BUSCO: assessing genome assembly and

annotation completeness with single-copy orthologs, *Bioinformatics*, **31**, 3210–2.
29. Ou, S., Chen, J. and Jiang, N. 2018, Assessing genome assembly quality using the LTR Assembly Index (LAI), *Nucleic Acids Res.*, **46**, e126.
30. McKenna, A., Hanna, M., Banks, E., et al. 2010, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Genome Res.*, **20**, 1297–303.
31. Benson, G. 1999, Tandem repeats finder: a program to analyze DNA sequences, *Nucleic Acids Res.*, **27**, 573–80.
32. Chen, N. 2004, Using RepeatMasker to identify repetitive elements in genomic sequences, *Curr. Protoc. Bioinformatics*, Chapter 4, Unit 4.10.
33. Bao, W., Kojima, K.K. and Kohany, O. 2015, Repbase update, a database of repetitive elements in eukaryotic genomes, *Mob. DNA*, **6**, 11.
34. Hubley, R., Finn, R.D., Clements, J., et al. 2016, The Dfam database of repetitive DNA families, *Nucleic Acids Res.*, **44**, D81–89.
35. Kalvari, I., Nawrocki, E.P., Argasinska, J., et al. 2018, Non-coding RNA analysis using the Rfam database, *Curr. Protoc. Bioinformatics.*, **62**, e51.
36. Nawrocki, E.P. and Eddy, S.R. 2013, Infernal 1.1: 100-fold faster RNA homology searches, *Bioinformatics*, **29**, 2933–5.
37. Lowe, T.M. and Chan, P.P. 2016, tRNAscan-SE on-line: integrating search and context for analysis of transfer RNA genes, *Nucleic Acids Res.*, **44**, W54–57.
38. Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T. and Ussery, D.W. 2007, RNAmmer: consistent and rapid annotation of ribosomal RNA genes, *Nucleic Acids Res.*, **35**, 3100–8.
39. Wang, Y., Tang, H., Debarry, J.D., et al. 2012, MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity, *Nucleic Acids Res.*, **40**, e49.
40. Kurtz, S., Phillippy, A., Delcher, A.L., et al. 2004, Versatile and open software for comparing large genomes, *Genome Biol.*, **5**, R12.
41. Lin, Y.C., Wang, J., Delhomme, N., et al. 2018, Functional and evolutionary genomic inferences in Populus through genome and population sequencing of American and European aspen, *Proc. Natl. Acad. Sci. U.S.A.*, **115**, E10970–E10978.
42. Smith, D.R., Crosby, K. and Lee, R.W. 2011, Correlation between nuclear plastid DNA abundance and plastid number supports the limited transfer window hypothesis, *Genome Biol. Evol.*, **3**, 365–71.
43. Krzywinski, M., Schein, J., Birol, I., et al. 2009, Circos: an information aesthetic for comparative genomics, *Genome Res.*, **19**, 1639–45.
44. Nattestad, M. and Schatz, M.C. 2016, Assemblytics: a web analytics tool for the detection of variants from an assembly, *Bioinformatics*, **32**, 3021–3.
45. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T. and Salzberg, S.L. 2015, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads, *Nat. Biotechnol.*, **33**, 290–5.
46. Robinson, M.D., McCarthy, D.J. and Smyth, G.K. 2010, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics*, **26**, 139–40.
47. Emms, D.M. and Kelly, S. 2015, OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy, *Genome Biol.*, **16**, 157.