**Accurate prediction of allele-specific gene expression (ASE) in hybrid poplar facilitated by high-quality genome assembly**

**Shi *et al.***

**Supplementary Note 1: 46 features used in the XGBoost** **machine-learning modeling of allele-specific gene expression (ASE).**

|  |  |
| --- | --- |
| **Sequence divergence** | |
| Gene\_length | Difference in gene length between a pair of alleles. |
| Exon\_length | Difference in exon length between a pair of alleles. |
| Intron\_length | Difference in intron length between a pair of alleles. |
| *Ka* | Nonsynonymous substitution rate of a pair of alleles. |
| *Ks* | Synonymous substitution rate of a pair of alleles. |
| *Ka/Ks* | *Ka/Ks* of a pair of alleles. |
| TFBS\_shared | The number of shared TFBS types within 2 kb upstream of a pair of alleles. |
| TFBS\_unique | The number of unique (specific to one allele) TFBS types within 2 kb upstream of a pair of alleles. |
| **Structural divergence** | |
| Exon\_number | Difference in exon number of a pair of alleles. |
| Intron\_number | Difference in intron number of a pair of alleles. |
| TE \_length\_inside\_exon | Difference in the total length of TE inserted into exon of a pair of alleles. |
| TE \_number\_inside\_exon | Difference in the total number of TE inserted into exon of a pair of alleles. |
| TE \_length\_inside\_intron | Difference in the total length of TE inserted into intron of a pair of alleles. |
| TE \_number\_inside\_intron | Difference in the total number of TE inserted into intron of a pair of alleles. |
| **Methylation difference** | |
| mCG\_gene | Difference of CG methylation frequency in the gene body of a pair of alleles. |
| mCHG\_gene | Difference of CHG methylation frequency in the gene body of a pair of alleles. |
| mCHH\_gene | Difference of CHH methylation frequency in the gene body of a pair of alleles. |
| mCG\_exon | Difference of CG methylation frequency in the gene body of a pair of alleles. |
| mCHG\_exon | Difference of CHG methylation frequency in exon of a pair of alleles. |
| mCHH\_exon | Difference of CHH methylation frequency in exon of a pair of alleles. |
| mCG\_intron | Difference of intron CG methylation frequency in intron of a pair of alleles. |
| mCHG\_intron | Difference of CHG methylation frequency in intron of a pair of alleles. |
| mCHH\_intron | Difference of CHH methylation frequency in intron of a pair of alleles. |
| mCG\_upstream | Difference of CG methylation frequency in the upstream 2kb of a pair of alleles. |
| mCHG\_upstream | Difference in CHG methylation frequency in the upstream 2kb of a pair of alleles. |
| mCHH\_upstream | Difference in CHH methylation frequency in the upstream 2kb of a pair of alleles. |
| mCG\_downstream | Difference in CG methylation frequency in the downstream 2kb of a pair of alleles. |
| mCHG\_downstream | Difference in CHG methylation frequency in the downstream 2kb of a pair of alleles. |
| mCHH\_downstream | Difference in CHH methylation frequency in the downstream 2kb of a pair of alleles. |
| mCG\_exon\_1st | Difference in CG methylation frequency at the first exon of a pair of alleles. |
| mCHG\_exon\_1st | Difference in CHG methylation frequency at the first exon of a pair of alleles. |
| mCHH\_exon\_1st | Difference in CHH methylation frequency at the first exon of a pair of alleles. |
| mCG\_intron\_1st | Difference in CG methylation frequency at the first intron of a pair of alleles. |
| mCHG\_intron\_1st | Difference in CHG methylation frequency at the first intron of a pair of alleles. |
| mCHH\_intron\_1st | Difference in CHG methylation frequency at the first intron of a pair of alleles. |
| **TE occupancy and affinity** | |
| TE\_gene\_distance | Difference between the distance to the nearest TE of a pair of alleles. |
| TE\_shared\_number | The number of same TE types of a pair of alleles. |
| TE\_unique\_number | The number of unique TE types of a pair of alleles. |
| TE\_occupation\_upstream2kb | The difference in the length of TE occupation within 2kb upstream of a pair of alleles. |
| TE\_occupation\_upstream5kb | The difference in the length of TE occupation within 5kb upstream of a pair of alleles. |
| TE\_occupation\_upstream10kb | The difference in the length of TE occupation within 10kb upstream of a pair of alleles. |
| TE\_occupation\_downstream2kb | The difference in the length of TE occupation within 2kb downstream of a pair of alleles. |
| TE\_occupation\_downstream5kb | The difference in the length of TE occupation within 5kb downstream of a pair of alleles. |
| TE\_occupation\_downstream10kb | The difference in the length of TE occupation within 10kb downstream of a pair of alleles. |
| **Tissue** | |
| Tissue | The type of tissue to which the sample belongs to. |
| **Treatment** | |
| Treatment | Experimental treatment. |

**Note:**

The weighted methylation levels are used to represent the regional methylation levels. Weighted methylation level: total number of reads for all methylated C sites in the region/total coverage in the region. And the illustration on the origin of genetic and epigenetic features can be seen in Figure 4 **(Figure 4)**.

**Supplementary Note 2: Library construction and sequencing**

For HiFi sequencing, single-molecule real-time circular consensus sequencing (CCS: https://github.com/PacificBiosciences/ccs) library preparation was performed according to the manufacturer protocols. Briefly, high-quality genomic DNA was extracted from the “84K” leaves using the DNeasy Plant Mini Kit (QIAGEN) and checked for integrity using the Agilent 4200 Bioanalyzer. Subsequently, 8 μg of genomic DNA was sheared using g-TUBEs (Covaris) and purified using AMPure PB magnetic beads. Sequencing libraries were prepared using Pacific Biosciences SMRTbell Template Prep Kit 2.0 according to manufacturer’s guidelines and subjected to fragment size selection (fragment molecules ≥ 11 kb) prior to HiFi sequencing on the PacBio Sequel II platform. The generated libraries were sequenced on three SMRT cells.

For ONT sequencing, leaves of the same “84K” individual were collected to extract high-quality genomic DNA using the cetyltrimethylammonium bromide (CTAB) method (Doyle and Doyle, 1987). After purification and quantification, adaptors were ligated using the Genomic DNA Ligation Reaction Kit (SQK-LSK109, Oxford Nanopore Technologies, UK). Then, sequencing libraries with ~20 kb DNA inserts were prepared and sequenced on two flow cells on the Nanopore PromethION platform.

For Hi-C library preparation, we followed a standard procedure (Xie et al., 2015). After the leaf tissues were fixed with formaldehyde and lysed, the cross-linked DNA was digested with MboI restriction enzyme and the digested fragments were biotinylated at the 5' overhangs. The blunt-end fragments were ligated to generate chimeric junctions, which were further purified, physically sheared, and enriched for biotin-containing fragments. Subsequently, the resulting biotinylated DNA fragments were subjected to end-repair, adaptor ligation and polymerase chain reaction, and paired-end sequencing libraries were constructed. These libraries were then sequenced using an Illumina Novaseq 6000 machine with 350 bp fragment.

**Supplementary Note 3: Genome assembly and quality assessment**

*De novo* genome assembly involved the following: primary assembly, Hi-C scaffolding, gap-filling, and optimization (**Supplementary Fig. 2**). First, HiFi reads obtained by sequencing in PacBio CCS mode were directly assembled with hifiasm v0.13-r308 (-o hifiasm -t32 -1 mat.yak -2 pat.yak -l 1 ccs.fa) (Cheng et al., 2021). We used a haplotype-switching resolution method, trio-binning which used short reads generated from parents to bin long reads generated from offspring prior to assembly (Koren et al., 2018). The method has been recommended for development of telomere-to-telomere (T2T) or gapless genome assembly efforts (Nurk et al., 2022).

We then used purge\_dups v1.2.5 (default parameters) (Guan et al., 2020) and performed manual controls to ensure that only reads with very shallow depth, very short/long length, high similarity, and a depth of approximately 1/2 the normal depth were removed. The assembly described above yielded the v0.1 version of the “84K” genome (**Supplementary Table 3**). The Hi-C reads were mapped to the v0.1 “84K” genome assembly using Juicer v1.7.6 (https://github.com/aidenlab/juicer) (Durand et al., 2016). We used an automated process to correct, order, and orientation errors by using the 3D-DNA scaffolding pipeline (Dudchenko et al., 2017). Juicebox v11.08 (Durand et al., 2016)was used to fine-tune the assembled scaffolds in a graphical and inter-active fashion through manual adjustments. To further improve the accuracy of the assembly, each chromosome was individually re-scaffolded with 3D-DNA and manually adjusted using Juicebox, including boundary adjustment, removal of incorrect insertions, and alignment adjustment, etc. Subsequently, 38 chromosomes and 10 contigs were created. Gaps were identified and their length was set to 100 bp. Then, the gaps were filled with ONT contigs or ONT ultra-long reads. ONT reads were assembled as contigs using NextDenovo v2.3.0 (genome\_size = 450m) (<https://github.com/Nextomics/NextDenovo>). The obtained contigs were compared with the assembled sequences using unimap (<https://github.com/lh3/unimap>) to fill the gaps. This step reduced the number of gaps from 55 to 11. The ONT ultra-long reads were further used for the gap filling using tgs-gapcloser v1.1.1 (--minmap\_arg ‘-x map-ont’) (Xu et al., 2020b), and the 38 chromosomes with only two gaps were obtained. The two previous “84K” genome assemblies did not cover these gaps (Liu et al., 2019; Qiu et al., 2019). Finally, we further optimized the genome assembly in two ways. First, we polished the assembly a second time using NextPolish v1.1.0 (Hu et al., 2020) with HiFi reads or Illumina short reads. Second, we compared the unplaced contig sequences with chromosome and organelle genome sequences using Redundans v0.14a (Pryszcz and Gabaldon, 2016) to identify redundancies in the unplaced sequences.

In addition, the chloroplast genome (Pt) was assembled based on Illumina short reads using GetOrganelle v1.6.0 (Jin et al., 2020). The complete mitochondrial genome (Mt) was obtained in three steps. First, mitochondria-derived contigs were obtained from Illumina data using GetOrganelle. Then, HiFi reads were compared with Minimap2 v2.17 (Li, 2018) to obtain mitochondria HiFi reads. Finally, the complete cyclic molecule was assembled with SMARTdenovo (Liu et al., 2021a).

The accuracy and structural completeness of the genome assembly was assessed by the ratio of genome collinearity with other *Populus* species using minimap2. BUSCO (Benchmarking Universal Single-Copy Orthologs) v2.0.1 (lineage dataset: embryophyta\_odb9) (Simao et al., 2015) was used to assess genome completeness. In addition, PacBio long reads and Illumina reads were mapped to the genome assembly using Minimap2 v2.17 (Li, 2018) and bwa v0.7.17 (https://github.com/lh3/bwa) (Li, 2013), respectively. Additionally, the transcriptome assembled in the current study was also mapped to the genome assembly using HiSat2 v2.1.0 (https://github.com/infphilo/hisat2) (Kim et al., 2015).

**Supplementary Note 4:** **Gene prediction and functional annotation**

First, protein sequences extracted from *Arabidopsis thaliana* and and 17 Salicaceae species (*Salix brachista*, *S. dunnii*, *S. purpurea*, *S. suchowensis*, *S. viminalis*, *Populus alba*, *P. alba*\_var\_*pyramidalis*, *P. alba*\_x\_*P. glandulosa*, *P. davidiana*\_x\_*P. alba*\_var\_*pyramidalis*, *P. deltoides*, *P. euphratica*, *P. ilicifolia*, *P. pruinosa*, *P. simonii*, *P. tremula*, *P. tremuloides*, and *P. trichocarpa*) were merged, followed by redundancy removal using CD-HIT v4.6. Subsequently, the assembled transcriptome assembly and the protein sequences were aligned to the repeat-masked reference genome assembly using BLASTn and tBLASTx from BLAST v2.2.28+ (Boratyn et al., 2012), and we further optimized the alignment using Exonerate v2.4.0 (Slater and Birney, 2005). The complete genes identified by BUSCO were used for *ab initio* gene prediction using AUGUSTUS v3.2.3 (Stanke et al., 2008). The gene model prediction using the MAKER v2.31.9 (Cantarel et al., 2008) was finalized using AUGUSTUS v3.2.3 (Stanke et al., 2008). The quality of gene prediction was assessed using the annotation edit distance (AED) for each of the predicted genes as part of MAKER. Non-coding RNA (ncRNA) prediction was performed using specific databases and packages, i.e. barrnap v0.9 (<https://github.com/tseemann/barrnap>), tRNAscan-SE v2.0 (Lowe and Eddy, 1997), and Rfam database (version 9.1) (<http://eggnogdb.embl.de/>). In addition, the genomes of the two parent genomes (subgenomes A and G) and 8 Salicaceae species (*Salix purpurea*, *S. suchowensis*, *S. brachista*, *S. dunnii*, *Populus ilicifolia*, *P. simonii*, *P. deltoides* and *P. trichocarpa*) were used for rDNA annotation with barrnap (--kingdom euk --threads 8). The distribution of rDNAs was mapped using TBtools v1.098769 (Chen et al., 2020a).

Functional annotation of protein-coding genes was performed using three strategies. First, annotation was performed using eggNOG-mapper v2.0.5 (Huerta-Cepas et al., 2017). Protein-coding genes were compared to the eggNOG homologous gene database to annotate gene functions (Jensen et al., 2008). Second, a sequence similarity search was performed for functional annotation. The predicted gene models were aligned with the protein archived in the SwissProt database (Bairoch and Apweiler, 2000), the Translated European Molecular Biology Laboratory (TrEMBL) database (Bairoch and Apweiler, 2000), the NCBI nonredundant protein database (NR), and the Arabidopsis databases using Diamond (Buchfink et al., 2015) to determine the best matching alignments based on E < 1e5 and identity > 30% criteria. Third, domain similarity searches were also performed. Using InterProScan v5.27-66.0 (Jones et al., 2014), motifs and functional domains were determined by searching protein databases, such as PRINTS, Pfam, SMART, PANTHER, and CDD.

**Supplementary Note 5: Phylogenetics and gene collinearity in the Salicaceae**

The protein-coding sequences of the two parental genomes (subgenomes A and G) and 12 Salicaceae species (*Salix purpurea*, *S. suchowensis*, *S. brachista*, *S. viminalis*, *S. dunnii*, *Populus euphratica*, *P. pruinosa*, *P. ilicifolia*, *P. simonii*, *P. deltoides*, *P. trichocarpa,* and *P. tremuloides*) were used for comparative genomics **(Supplementary Table 15)**. OrthoFinder2 v2.3.12 was used to construct the orthogroups (Emms and Kelly, 2019). Based on 2,106 single-copy orthologs, we used IQTREE v1.6.7 (Nguyen et al., 2015) to construct a maximum likelihood (ML) phylogenetic tree using the best-fit model (JTT+ F + R4). MAFFT v7.407 (Katoh and Standley, 2013) was used to align homeologs before converting the aligned protein sequences to codon alignment. Concatenated amino acid sequences were trimmed using trimAL v1.4 (Capella-Gutierrez et al., 2009) with -gt 0.8 -st 0.001 -cons 60. The MCMCTree program from PAML v4.9j (Yang, 2007) was used to estimate divergence times under independent substitution rate (clock = 1) setting, with the following details: K80 substitution model, 5e5 iterations, and discarded 1e5 iterations as burn-in, and a fossil date time of 48–52 MYA split time between *Populus* and *Salix* (Chen et al., 2019). CAFÉ v4.2.1 (Han et al., 2013) was then used to infer gene family expansion and contraction based on the chronogram of the 14 species analyzed.

Gene collinearity analysis on chromosome level in eight species, four willows, four poplars (*Salix purpurea*, *S. suchowensis*, *S. brachista*, *S. dunnii*, *Populus ilicifolia*, *P. simonii*, *P. deltoides,* and *P. trichocarpa*) and the parental subgenomes A and G was performed using MCScanX (Wang et al., 2012).

**Supplementary Note 6: Variation between the two parental genomes**

The nucmer alignment tool from the MUMmer toolbox v4.0.0 (Marcais et al., 2018; Zhao et al., 2018) was used to perform whole-genome alignment with the parameter settings “–maxmatch -c 100 -b 500 -l 50 -g 90”. Alignment results were filtered by identity (>90) and alignment length (>100). Finally, syntenic regions, structural rearrangements (inversions, translocations, and duplications), and sequence differences (SNPs, indels, etc.) between the two parental genomes were identified using SyRI v1.3 (Goel et al., 2019). The genome of the *P*. *alba* parent (subgenome A) was used as the reference and subgenome G (from the other parent, *P. tremula* var. *glandulosa*) as the query.

The distance between the inversion region on the chromosome and the nearest TE was also calculated with BEDtools v2.29.2 (Quinlan, 2014). Here, TEs include both randomly generated and observed TEs.Random data were generated using the random function in BEDtools v2.29.2 (Quinlan, 2014).To analyze the effects of structural variants (SVs) on gene expression levels, we calculated the average expression (in TPM) for each gene in the syntenic and the different SV regions across all samples to represent the gene expression levels. The two-sided Wilcoxon test was used to determine the significant differences in gene expression levels between the syntenic and SV regions.

**Supplementary Note 7: RNA*-*seq data and allelic gene expression**

We collected a total of 156 RNA-seq samples of the “84K” clone from different tissues and treatments, each with three independent biological replicates. Low-quality reads and adapters were removed and the remaining reads were quantified for gene expression estimation using Salmon v1.6.0 (--validateMappings --numBootstraps 100) (Patro et al., 2017), through which we obtained count values and normalized TPM values for each sample (**Supplementary Table 8**). If the TPM expression value of a gene/allele in TPM exceeded 0.5 in any sample, we considered it to be expressed. Differentially expressed alleles were identified using the DESeq2 package (Love et al., 2014). The following ranges of fold change (FC) were used as criteria to determine differential expression: 1) no-expression (both alleles were not expressed) (NE); 2) no-significant difference between a pair of alleles with *p*-adjust > 0.05 (Diff00), and 3) significant difference between a pair of alleles with *p*-adjust ≤ 0.05, ASE. The ASE group was further divided into different classes based on the FC in expression: (1) Diff0, when FC ≤ |2| (Diff0); (2) Diff2, when |2| < FC < |8| (Diff2); and (3) Diff8, when FC ≥ |8| (Diff8).

**Supplementary Note 8:** **DNA methylation quantification from ONT long reads**

To quantify DNA methylation (CG, CHG, and CHH) of the “84K” clone, we used DeepSignal-plant v.0.1.4 (Ni et al., 2021), which detects DNA 5mC methylation using a deep learning approach with bidirectional recurrent neural networks (BRNN) and long short-term memory (LSTM) units. In total, we used two replicates of 20× raw ONT read data. First, the raw nanopore were preprocessed by conversion to base sequences using Guppy v5.0.16 (guppy\_basecaller, --config dna\_r9.4.1\_450bps\_hac\_prom.cfg). The signal data (fast5 format) can be successfully converted into base sequences (fastq format). Then, tombo v1.5.1 (Stoiber et al., 2016) was used to manipulate re-squiggle (Raw Signal Genomic Alignment). Briefly, the re-squiggle algorithm aligned the raw signal (electric current nanopore measurements) to the “84K” genome assembly (tombo resquiggle). Once the data were processed, methylations for the CG, CHG, and CHH contexts were called using DeepSignal-plant under the default reference models (model.dp2.CNN.arabnrice2-1\_120m\_R9.4plus\_tem.bn13\_sn16.both\_bilstm.epoch6.ckpt). Then, the methylation frequencies of CG, CHG, and CHH sites were generated separately using scripts in the DeepSignal-plant pipeline (<https://github.com/PengNi/deepsignal-plant>) (**Supplementary Table 19**).

We used the methylation information for each genomic position obtained from DeepSignal-plant to calculate and visualize the average methylation levels of the different regions. At least five reads covering each cytosine methylation site were retained. To map the distribution of methylation levels along chromosomes, a 500 kb sliding window with a step size of 100 kb was defined using the makewindows function in BEDtools v2.29.2 (Quinlan, 2014), and the average methylation level within the window was calculated using methyGff in BatMeth2 (Lim et al., 2012). In addition, the gene body and 2 kb upstream and downstream regions were divided into 100 bins each.

**Supplementary Note 9: Feature extraction for machine-learning modeling**

To build the machine-learning (ML) model, a feature dataset was first created. Each column of this dataset represented a feature, and each row represented a pair of alleles in comparison. A total of 46 features were created in six categories (**Supplementary Table 21** and **Supplementary Notes 1**). The feature categories were described as follows: (1) Methylation features, including the difference of a pair of alleles in the average methylation frequency (three types: CG, CHG, CHH) of the gene body (mCG\_gene, mCHG\_gene, mCHH\_gene), exons (mCG\_exon, mCHG\_exon, mCHH\_exon), introns (mCG\_intron, mCHG\_intron, mCHH\_intron), sequences from upstream 2 kb (mCG\_upstream, mCHG\_upstream, mCHH\_upstream), sequences of downstream 2 kb (mCG\_downstream, mCHG\_downstream, mCHH\_downstream), first exon (mCG\_exon\_1st, mCHG\_exon\_1st, mCHH\_exon\_1st) and first intron (mCG\_intron\_1st, mCHG\_intron\_1st, mCHH\_intron\_1st). (2) TE occupancy and affinity, a factor of great interest in the study of gene expression. This category includes the distance difference of the closest TE insertion to an allele pair (TE\_gene\_distance), the number of TE insertion shared by an allele pair (TE\_shared\_number), the number of unique TE insertion in an allele pair (TE\_unique\_number), the difference of TE occupancy in upstream 2/5/10 kb of an allele pair (TE\_occupation\_upstream2kb, TE\_occupation\_upstream5kb, TE\_occupation\_upstream10kb), the difference of TE occupancy in downstream 2/5/10 kb of an allele pair (TE\_occupation\_downstream2kb, TE\_occupation\_downstream5kb, TE\_occupation\_ downstream10kb). (3) Sequence divergence in an allele pair, including that of gene length (Gene\_length), exon length (Exon\_length), intron length (Intron\_length), *Ka* (number of substitutions per non-synonymous site), *Ks* (number of substitutions per synonymous site), *Ka*/*Ks*, the number of transcription factor binding sites (TFBS) shared in the upstream 2 kb between alleles (TFBS\_shared), the number of TFBS unique in the upstream 2 kb between alleles (TFBS\_unique). (4) Structural divergence in a pair of alleles, including the number of exon (Exon\_number), the intron number (Exon\_number), the number of exons with TE insertion (TE\_number\_inside\_exon), the length of exon with TE insertion (TE\_length\_inside\_exon), the number of introns with TE insertion (TE\_number\_inside\_intron), the length of intron with TE insertion (TE\_length\_inside\_intron). (5) Tissue, the tissue from which the RNA-seq was done. (6) Treatment, the plant’s treatment from which the RNA-seq was done.

**Supplementary Note 10:** **Model construction**

Here, XGBoost modeling was used to understand whether allele-specific expression (ASE) was predictable from different genetic, epigenetic features, or experimental designs, and then which factors represented were the most important associations with the observed ASE. Our XGBoost modeling were implemented with R package, xgboost (Chen et al., 2015).

First, XGBoost modeling was performed on a dataset containing 46 features (predictor variables) with the following settings, eta =0.3, gamma = 0.001, max\_depth =2, nrounds = 100000, print\_every\_n = 100, early\_stopping\_rounds = 200 and default values for other parameters. Our dataset contained 1,220,274 cases in our dataset, 70% of which were used as training, and 30% as test set. Also, we used the same dataset for correlation analysis between features. We kept the features with significant correlation of less than 0.001 (p < 0.001). In addition, of the interrelated features, we retained the one that ranked highest in model importance (**Supplementary Fig. 19** and **Supplementary Table 21**).

As a primary step, XGBoost modeling (Model 0) was performed with all 46 features as predictors to predict four groups of ASEs (as defined above), which was used to rank the features. After feature selection, a XGBoost model (Model 1) was constructed to predict the four groups of ASEs with 15 selected features. In addition, another XGBoost classification model (Model 2) was created to predict two ASE groups (no ASE; ASE). Another XGBoost regression model (Model 3) was built to predict the difference in expression (in transcripts per million, TPM) of ASE.

To assess the predictability of each classification model, we calculated ROC curves (Receiver Operating characteristic Curves) and AUC (Area Under the Curve) values (Robin et al., 2021). In addition, the modeling results of this purely data-driven approach could be explained using SHapley Additive exPlanations (SHAP) to better interpret the model (default values for parameters) (Lundberg and Lee, 2017). Here, we used SHAP to explain the influence of the five highest-ranking features on the final prediction of ASE.

References:

**Bairoch A, Apweiler R** (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Research **28:** 45-48

**Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL** (2012) Domain enhanced lookup time accelerated BLAST. Biology Direct **7:** 1-14

**Buchfink B, Xie C, Huson DH** (2015) Fast and sensitive protein alignment using DIAMOND. Nature Methods **12:** 59-60

**Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M** (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research **18:** 188-196

**Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T** (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics **25:** 1972-1973

**Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R** (2020a) TBtools: an integrative toolkit developed for interactive analyses of big biological data. Mol Plant **13:** 1194-1202

**Chen JH, Huang Y, Brachi B, Yun QZ, Zhang W, Lu W, Li HN, Li WQ, Sun XD, Wang GY, He J, Zhou Z, Chen KY, Ji YH, Shi MM, Sun WG, Yang YP, Zhang RG, Abbott RJ, Sun H** (2019) Genome-wide analysis of Cushion willow provides insights into alpine plant divergence in a biodiversity hotspot. Nature Communications **10:** 5230

**Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou TJRpv-** (2015) Xgboost: extreme gradient boosting. R package version 0.4-2 **1:** 1-4

**Cheng H, Concepcion GT, Feng X, Zhang H, Li H** (2021) Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. Nature Methods **18:** 170-175

**Doyle J, Doyle JL** (1987) Genomic plant DNA preparation from fresh tissue-CTAB method. Phytochemical Bulletin **19:** 11-15

**Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL** (2017) *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science **356:** 92-95

**Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL** (2016) Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Systems **3:** 99-101

**Emms DM, Kelly S** (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology **20:** 238

**Goel M, Sun H, Jiao WB, Schneeberger K** (2019) SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol **20:** 277

**Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R** (2020) Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics **36:** 2896-2898

**Han MV, Thomas GW, Lugo-Martinez J, Hahn MW** (2013) Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Molecular Biology and Evolution **30:** 1987-1997

**Hu J, Fan J, Sun Z, Liu S** (2020) NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics **36:** 2253-2255

**Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P** (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. Molecular Biology And Evolution **34:** 2115-2122

**Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, Bork P** (2008) eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res **36:** D250-254

**Jin JJ, Yu WB, Yang JB, Song Y, dePamphilis CW, Yi TS, Li DZ** (2020) GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. Genome Biol **21:** 241

**Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S** (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics **30:** 1236-1240

**Katoh K, Standley DM** (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution **30:** 772-780

**Kim D, Langmead B, Salzberg SL** (2015) HISAT: a fast spliced aligner with low memory requirements. Nature Methods **12:** 357-360

**Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM** (2018) *De novo* assembly of haplotype-resolved genomes with trio binning. Nature Biotechnology **36:** 1174–1182

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

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

**Lim JQ, Tennakoon C, Li G, Wong E, Ruan Y, Wei CL, Sung WK** (2012) BatMeth: improved mapper for bisulfite sequencing reads on DNA methylation. Genome Biol **13:** R82

**Liu H, Wu S, Li A, Ruan J** (2021a) SMARTdenovo: a *de novo* assembler using long noisy reads. GigaByte **2021:** 1-9

**Liu YJ, Wang XR, Zeng QY** (2019) *De novo* assembly of white poplar genome and genetic diversity of white poplar population in Irtysh River basin in China. Science China Life Sciences **62:** 609-618

**Love MI, Huber W, Anders S** (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol **15:** 550

**Lowe TM, Eddy SR** (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25:** 955-964

**Lundberg SM, Lee SI** (2017) A unified approach to interpreting model predictions. In Proc. 31st Int. Conf. Neural Information Processing Systems **vol. 30:** 4768–4777

**Marcais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A** (2018) MUMmer4: A fast and versatile genome alignment system. PLoS Computational Biology **14:** e1005944

**Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ** (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution **32:** 268-274

**Ni P, Huang N, Nie F, Zhang J, Zhang Z, Wu B, Bai L, Liu W, Xiao CL, Luo F, Wang J** (2021) Genome-wide detection of cytosine methylations in plant from Nanopore data using deep learning. Nat Commun **12:** 5976

**Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, Aganezov S, Hoyt SJ, Diekhans M, Logsdon GA, Alonge M, Antonarakis SE, Borchers M, Bouffard GG, Brooks SY, Caldas GV, Chen NC, Cheng H, Chin CS, Chow W, de Lima LG, Dishuck PC, Durbin R, Dvorkina T, Fiddes IT, Formenti G, Fulton RS, Fungtammasan A, Garrison E, Grady PGS, Graves-Lindsay TA, Hall IM, Hansen NF, Hartley GA, Haukness M, Howe K, Hunkapiller MW, Jain C, Jain M, Jarvis ED, Kerpedjiev P, Kirsche M, Kolmogorov M, Korlach J, Kremitzki M, Li H, Maduro VV, Marschall T, McCartney AM, McDaniel J, Miller DE, Mullikin JC, Myers EW, Olson ND, Paten B, Peluso P, Pevzner PA, Porubsky D, Potapova T, Rogaev EI, Rosenfeld JA, Salzberg SL, Schneider VA, Sedlazeck FJ, Shafin K, Shew CJ, Shumate A, Sims Y, Smit AFA, Soto DC, Sović I, Storer JM, Streets A, Sullivan BA, Thibaud-Nissen F, Torrance J, Wagner J, Walenz BP, Wenger A, Wood JMD, Xiao C, Yan SM, Young AC, Zarate S, Surti U, McCoy RC, Dennis MY, Alexandrov IA, Gerton JL, O'Neill RJ, Timp W, Zook JM, Schatz MC, Eichler EE, Miga KH, Phillippy AM** (2022) The complete sequence of a human genome. Science **376:** 44-53

**Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C** (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods **14:** 417-419

**Pryszcz LP, Gabaldon T** (2016) Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res **44:** e113

**Qiu D, Bai S, Ma J, Zhang L, Shao F, Zhang K, Yang Y, Sun T, Huang J, Zhou Y, Galbraith DW, Wang Z, Sun G** (2019) The genome of *Populus alba* x *Populus tremula* var. *glandulosa* clone 84K. DNA Research **26:** 423-431

**Quinlan AR** (2014) BEDTools: the Swiss-army tool for genome feature analysis. Current Protocols in Bioinformatics **47:** 11.12.11–11.12.34

**Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M, Siegert S, Doering M, Billings Z** (2021) Package ‘pROC’. *In*. 2012-09-10 09: 34

**Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM** (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31:** 3210-3212

**Slater GS, Birney E** (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics **6:** 31

**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

**Stoiber M, Quick J, Egan R, Lee JE, Celniker S, Neely RK, Loman N, Pennacchio LA, Brown J** (2016) *De novo* identification of DNA modifications enabled by genome-guided nanopore signal processing. BioRxiv**:** 094672

**Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, Kissinger JC, Paterson AH** (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res **40:** e49

**Xie T, Zheng JF, Liu S, Peng C, Zhou YM, Yang QY, Zhang HY** (2015) *De novo* plant genome assembly based on chromatin interactions: a case study of *Arabidopsis thaliana*. Mol Plant **8:** 489-492

**Xu M, Guo L, Gu S, Wang O, Zhang R, Peters BA, Fan G, Liu X, Xu X, Deng L, Zhang Y** (2020b) TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. Gigascience **9:** giaa094

**Yang Z** (2007) PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution **24:** 1586-1591

**Zhao H, Wang S, Wang J, Chen C, Hao S, Chen L, Fei B, Han K, Li R, Shi C, Sun H, Wang S, Xu H, Yang K, Xu X, Shan X, Shi J, Feng A, Fan G, Liu X, Zhao S, Zhang C, Gao Q, Gao Z, Jiang Z** (2018) The chromosome-level genome assemblies of two rattans (*Calamus simplicifolius* and *Daemonorops jenkinsiana*). Gigascience **7:** 1-11