A chromosome-level reference genome of the hornbeam, *Carpinus fangiana*

Xiaoyue Yang1,4, Zefu Wang2,4, Lei Zhang2, Guoqian Hao3, Jianquan Liu1,2 & Yongzhi Yang1*

**Background & Summary**

Betulaceae, also known as the birch family, includes over 160 species of trees or shrubs1. It is divided into two subfamilies, Coryloideae and Betuloideae: Betuloideae comprises the genera *Alnus* and *Betula*, while Coryloideae comprises *Corylus*, *Ostryopsis*, *Carpinus* and *Ostrya*. These subfamilies and their genera are readily distinguished based on their different morphological characteristics, such as the samara of Coryloideae, the nuts of Betuloideae, and their different types of pollen2. In addition, cell biological investigations have revealed that Betulaceae species have very different chromosome numbers: the basic chromosome number is eight for *Carpinus*, *Ostrya*, *Ostryopsis* species, eleven for *Corylus* species, and fourteen for *Alnus* and *Betula* species3,4.

Several Betulaceae species, notably those belonging to the genera *Betula*, *Alnus*, and *Carpinus*, are important components of forests in temperate regions, mountains, and subtropical areas, as well as important sources of timber and materials for traditional Chinese medicine. Some species of *Betula* and *Carpinus* are used as ornamental trees and widely planted in large parks and gardens. *Alnus* species can form symbioses with nitrogen-fixing bacteria of the genus *Frankia*, helping to enhance soil fertility5. The fruits of *Corylus*, known as hazelnuts, are economically important. The birch family thus has remarkable ecological, economic, medicinal, and ornamental value. Additionally, Betulaceae is a relict family, and there are many reliable fossils of this family that have provided important paleobotanic insights6. However, only a few species of the family have been studied extensively in ways that could support their further development and utilization.

A few genomes of Betulaceae species have been published in recent years. The genomes of two Betuloideae members, *Betula pendula* (scaffold N50: 0.53 Mb)7 and *Alnus glutinosa* (scaffold N50: 0.10 Mb)8, were presented in 2017 and 2018, and the *B. pendula* genome was further anchored to fourteen chromosomes. The only published Coryloideae genomes are those of two ironwood trees from the genus *Ostrya*: *O. rehderiana* (scaffold N50: 2.31 Mb) and *O. chinensis* (scaffold N50: 0.81 Mb), which were reported in 20189. However, no genomes representing any of the other three genera in Coryloideae have been disclosed and there are no published chromosome-level genomes for this subfamily.

---

1Stat Key Laboratory of Grassland Agro-Ecosystem, Institute of Innovation Ecology, Lanzhou University, 730000, Lanzhou, China. 2Stat Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, 610000, Sichuan, China. 3Sichuan Tea College, Yibin University, Yibin, 644007, Sichuan, China. 4These authors contributed equally: Xiaoyue Yang and Zefu Wang. *email: yangyongzhi2008@gmail.com*
To enrich the available genomic resources for Betulaceae, we sequenced the whole genome of *Carpinus fangiana* (Fig. 1), a member of the most species-rich genus in Coryloideae\(^\text{10}\). A total of 77.85 Gb (~200x) next-generation data and 52.19 Gb (~130x) Hi-C data were used to assemble the genome. The assembly produced a genome having a total length of 386.19 Mb, with 357.84 Mb being anchored to eight chromosomes. To our knowledge, this is the first reported chromosome-level Coryloideae genome assembly. The contig N50 and scaffold N50 were 35.32 kb and 1.91 Mb, respectively. Structural annotation of the genome revealed a total of 27,381 protein-coding genes, of which 94.36% were functionally annotated. The genome was also predicted to contain 4,440 non-coding genes based on a comprehensive annotation. This chromosome-level genome of *C. fangiana* will greatly facilitate further biological studies on Betulaceae as well as the development and commercial exploitation of the genus.

### Methods

**Sampling, library construction and sequencing.** Fresh leaves were collected from a wild *C. fangiana* tree in Ebian, Sichuan, China (N: 29° 1’ 44’’; S: 102° 59’ 30’’; Fig. 1) and immediately dried over silica gel. Genomic DNA was then extracted from the dried leaves using the modified Cetyltrimethylammonium Ammonium Bromide (CTAB)\(^\text{11}\) method. Sequencing libraries with different insert sizes were constructed using a library construction kit (Illumina). Short paired-end libraries were constructed with insert sizes of 230, 500, and 800 bp, while the insert sizes used to construct mate pair libraries were 2, 5, 10, and 20 kb. The Illumina HiSeq 2000 platform was used to sequence 150 bp paired-end reads for all these libraries in accordance with the manufacturer’s instructions. These procedures generated a total of 115.12 Gb (~200x) raw data for *C. fangiana* genome assembly (Table 1).

A High-through chromosome conformation capture (Hi-C) library for the *C. fangiana* genome was also constructed. To this end, fresh leaves were fixed with formaldehyde to induce DNA cross-linking, after which

### Table 1. DNA sequencing metrics of *C. fangiana*, before and after quality control. Note: The data contains Next-generation and Hi-C sequencing data. The estimated genome size is 396.74 Mb.

| Sequesting technique | Library type | Insert size (bp) | Read length (bp) | Amount of sequence | Depth (x-times) |
|----------------------|--------------|------------------|------------------|--------------------|-----------------|
|                      |              |                  | Raw data (Gb)    | clean data (Gb)    | Raw data        | clean data      |
| Next-generation      | paired-end   | 230              | 150              | 11.32              | 10.92           | 28.54           | 27.52           |
|                      | paired-end   | 500              | 150              | 10.28              | 10.21           | 25.91           | 25.73           |
|                      | paired-end   | 800              | 150              | 15.82              | 15.64           | 39.88           | 39.42           |
|                      | mate pair    | 2,000            | 150              | 16.49              | 6.55            | 41.56           | 16.51           |
|                      | mate pair    | 5,000            | 150              | 13.25              | 9.71            | 33.39           | 24.47           |
|                      | mate pair    | 10,000           | 150              | 17.97              | 10.71           | 43.50           | 27.00           |
|                      | mate pair    | 20,000           | 150              | 29.99              | 14.12           | 75.59           | 35.59           |
|                      | Total        |                  |                  | 115.12             | 77.85           | 290.17          | 196.23          |
| Hi-C                 | Hi-C         | 300-700          | 150              | 52.54              | 52.19           | 132.43          | 131.55          |

---

Fig. 1 Photograph and location of the *C. fangiana* tree sampled for genome sequencing. (a) A photograph of a *C. fangiana* individual on Emei Mountain, Leshan, Sichuan, China. (b) Location of the *C. fangiana* sample used for genome sequencing.

---

### Table 1. DNA sequencing metrics of *C. fangiana*, before and after quality control. Note: The data contains Next-generation and Hi-C sequencing data. The estimated genome size is 396.74 Mb.
the DNA was digested with HindIII. The resulting sticky ends were biotinylated and proximity-ligated to form chimeric junctions that were enriched for, and physically sheared into 300–700 bp fragments. These chimeric fragments were sequenced on the Illumina HiSeq platform, generating 52.54 Gb (~130x) of Hi-C data (Table 1).

We also harvested six tissues (bark, branch, bract, flower, fruit, leaf) for total RNA sequencing. These samples were flash frozen in liquid nitrogen, and total RNA was extracted using the modified CTAB method. cDNA libraries were then constructed using the NEBNext Ultra RNA Library Prep Kit for Illumina (NEB). The Illumina HiSeq 2500 platform was used to sequence these libraries with a read length of 2 \times 150 bp, generating over 5.50 Gb raw data for each tissue (Table 2).

**Preprocessing and genome size estimation.** Quality control checks on the raw genome data were performed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Potential adapters in reads were removed using Scythe (http://github.com/vsbuffalo/scythe) and low-quality reads were discarded by Sickle (http://github.com/vsbuffalo/scythe). The program Lighter was then used to correct sequence errors in the remaining reads. For mate pair reads, we also used FastUniq to remove duplicates. In total, 77.85 Gb, ~200x high-quality next-generation sequencing data and 52.19 Gb, ~130x high-quality Hi-C data were generated for de novo assembly of the *C. fangiana* genome (Table 1).

Quality control of transcriptome data was performed using a custom Perl script. Reads were discarded if (1) the proportion of unidentified nucleotides in one read exceeded 5%, or (2) over 65% of the read’s bases had a phred quality below 8. After eliminating low-quality reads, the quantity of retained data for each tissue was above 5.50 Gb (Table 2). The RNA-seq reads were then assembled using Trinity. CD-Hit was used to eliminate redundant transcript sequences, and candidate coding regions in the transcript sequences were identified by TransDecoder.

Before genome assembly, we estimated the *C. fangiana* genome's size by performing a combined analysis using Jellyfish and GenomeScope. Reads from the short-insert libraries were first processed by Jellyfish to assess their k-mer distribution, using a k value of 17. Then, GenomeScope was used to estimate the genome size based on the k-mer distribution (Fig. 2). The genome was thereby estimated to be around 396.74 Mb long.

**Genome assembly.** Preliminary de novo assembly of the *C. fangiana* genome was performed with Platanus, which can effectively manage high-throughput data from heterozygous samples. Assembly using Platanus proceeded via three steps: (1) contig-assembly, in which de Bruijn graphs were constructed using the clean reads from short paired-end libraries and the sequences of contigs were then displayed in the graphs; (2) scaffolding,
in which reads from all next-generation libraries (short paired-end and mate pair) were mapped to contigs, after which contigs considered to be linked were combined into scaffolds; (3) gap closing, in which reads that mapped onto scaffolds were collected to cover the gaps between them. GapCloser20 was used to further close the gaps based on reads from all the paired-end libraries, after which the automated HaploMerger2 pipeline21 was used to rebuild the above assembly and implement flexible and sensitive error detection. After discarding scaffolds smaller than 1 kb, a high-quality de novo assembled C. fangiana genome was obtained. The size of this genome (386.19 Mb) was 97.34% of the estimated value (396.74 Mb) and its GC content was 37.59%. The scaffold N50 and N90 values were 1.91 Mb and 0.43 Mb, while the contig N50 and N90 were 35.32 kb and 8.54 kb (Table 3).

The HiC-Pro22 program was used for quality assessment of the Hi-C data. Valid interaction pairs were mapped to and used for error correction of the contigs and scaffolds assembled based on the next-generation sequencing data. Next, the contigs and scaffolds were anchored to chromosomes using LACHESIS23. In total, 357.84 Mb of scaffolds were assembled into eight chromosomes (Table 4). Finally, we obtained a high-quality chromosome-level genome with a total size of 386.25 Mb. The contig N50 and scaffold N50 values of this chromosome-level assembly were 34.85 kb and 37.11 Mb, respectively (Table 3).

Heterozygosity assessment and repeat annotation. To assess the heterozygosity of the C. fangiana genome, we first mapped reads from the 500 bp library to the assembled genome using the BWA-MEM algorithm from the Burrows-Wheeler Aligner (BWA) package24. SAMtools25 was used to convert the mapping results to BAM format, sort them, and remove duplicates. The Picard package (http://broadinstitute.github.io/picard/) was used to replace read groups in the bam file. Two programs (RealignerTargetCreator and IndelRealigner) from the Genome Analysis ToolKit (GATK)26 package were used to avoid misalignments and account for the effects of indels. The SAMtools command ‘mpileup’ was used to generate a VCF format file, and the program bcftools from the SAMtools package was used to detect single nucleotide polymorphisms (SNPs). Finally, based on the SNPs, the heterozygosity was calculated to be 0.38% using a custom Perl script.

Repetitive sequences and transposable elements (TEs) in the C. fangiana genome were identified using a combined procedure incorporating de novo and homology-based approaches at the DNA and protein levels. Tandem repeats were annotated using Tandem Repeat Finder (TRF)27. A repeat library for the C. fangiana genome was generated using RepeatModeler (http://www.repeatmasker.org) to facilitate de novo annotation. RepeatMasker28 (http://www.repeatmasker.org) was used to identify and classify the TEs at the DNA level. We also used RepeatProteinMasker to perform a WU-BLASTX search against the TE protein database in order to identify and

| Type                        | De novo assembly | Hi-C assembly |
|-----------------------------|-----------------|---------------|
| Scaffold length (bp)        | 386,190,506     | 386,249,499   |
| Gap length (bp)             | 30,727,985      | 30,804,875    |
| Scaffold number              | 4,789           | 4,602         |
| Longest scaffold (bp)       | 8,871,445       | 60,187,804    |
| Scaffold N50 (bp)           | 1,908,393       | 37,105,143    |
| Scaffold N90 (bp)           | 425,779         | 595,656       |
| Contig length (bp)          | 355,461,404     | 355,441,862   |
| Contig number               | 21,775          | 22,086        |
| Longest contig (bp)         | 1,041,408       | 912,918       |
| Contig N50 (bp)             | 35,323          | 34,845        |
| Contig N90 (bp)             | 8,542           | 8,427         |
| GC content                  | 37.59%          | 37.55%        |

Table 3. Summary of C. fangiana genome assembly. Note: The estimated genome size is 396.74 Mb. GC content of the genome without N.

| Type | Sequence Number | Sequence Length (bp) | GenBank accession |
|------|----------------|----------------------|-------------------|
| Cfa01 | 128            | 62,383,991           | CM017321          |
| Cfa02 | 97             | 51,103,020           | CM017322          |
| Cfa03 | 107            | 42,654,226           | CM017323          |
| Cfa04 | 135            | 44,816,785           | CM017324          |
| Cfa05 | 88             | 39,651,540           | CM017325          |
| Cfa06 | 104            | 40,118,261           | CM017326          |
| Cfa07 | 92             | 39,687,453           | CM017327          |
| Cfa08 | 109            | 37,421,582           | CM017328          |
| Total Sequences Clustered (Ratio %) | 860 (16.32)   | 357,836,858 (92.66) |
| Total Sequences Ordered and Oriented (Ratio %) | 677 (78.72)   | 319,127,541 (89.18) |

Table 4. Summary of the assembled chromosomes in the C. fangiana genome.
classify TEs at the protein level. Finally, long terminal repeats (LTR) were identified using LTR-FINDER. In total, the C. fangiana genome was found to contain 158.69 Mb repetitive sequences, accounting for 41.08% of its length (Table 5). As shown in Table 5, the most common classifications assigned to these repetitive elements were Unknown (15.97% of the assembled genome) and LTRs (14.57% of the assembled genome).

**Gene annotation.** Structural annotation of gene models was performed by applying a combination of de novo, homology-based, and transcriptome-based methods to the repeat-masked genome. The de novo approach was implemented using Augustus, Geneid, GeneMark, glimmerHMM, and SNAP. For homology-based prediction, TBLASTN was used to align predicted protein sequences from Arabidopsis thaliana, Vitis vinifera, Prunus persica, Ostrya chinensis, Ostrya rehderiana, and Juglans regia to the C. fangiana genome with an E-value threshold of 1E-05. Then, GeneWise was used to obtain accurate spliced alignments by aligning homologous sequences to matched proteins. Transcriptome-based prediction was performed with the Program to Assemble Spliced Alignments (PASA), which was used to predict protein-coding regions based on the assembled transcripts of the six different C. fangiana tissues. The gene models obtained from the de novo, homology-based, and transcriptome-based annotations were combined to form a consensus gene set using EVidenceModeler (EVM). After strict filtering, a total of 27,381 non-redundant protein-coding genes were annotated in the C. fangiana genome (Table 6).

### Table 5. Repeat element metrics for the C. fangiana genome.

| Type               | Length (bp) | Percent (%) |
|--------------------|-------------|-------------|
| DNA                | 14,244,548  | 3.69        |
| LINE               | 15,452,667  | 4.00        |
| Low_complexity     | 1,653,498   | 0.43        |
| LTR                | 56,262,090  | 14.57       |
| Other              | 660         | 1.71E-04    |
| RC                 | 1,272,200   | 0.33        |
| tRNA               | 5,881       | 1.52E-03    |
| Satellite          | 232,066     | 0.06        |
| Simple_repeat      | 7,594,441   | 1.97        |
| SINE               | 281,915     | 0.07        |
| Unknown            | 61,686,663  | 15.97       |
| All                | 158,686,629 | 41.08       |

### Table 6. Summary of predicted protein-coding genes in the C. fangiana genome. Note: *UTR regions were contained.

| Gene set          | Number | Average gene length (bp) | Average CDS length (bp) | Average exons per gene | Average exon length (bp) | Average intron length (bp) |
|-------------------|--------|--------------------------|-------------------------|------------------------|-------------------------|---------------------------|
| **De novo prediction** |
| Augustus          | 36,499 | 3,740.33                 | 1,371.15                | 5.20                   | 342.17                  | 678.20                    |
| Geneid            | 43,054 | 4,539.67                 | 1,023.87                | 4.14                   | 247.27                  | 1,755.27                  |
| GeneMark          | 28,642 | 1,900.29                 | 892.05                  | 3.15                   | 283.15                  | 492.58                    |
| glimmerHMM        | 45,800 | 1,657.35                 | 867.05                  | 2.65                   | 327.78                  | 398.26                    |
| SNAP              | 63,982 | 1,087.42                 | 656.98                  | 2.62                   | 250.80                  | 220.80                    |
| **Homolog prediction** |
| Arabidopsis thaliana | 21,976   | 3,251.94               | 1,100.22               | 4.45                   | 247.27                  | 631.93                    |
| Vitis vinifera   | 23,733 | 3,293.62                 | 1,047.44                | 4.59                   | 228.23                  | 633.86                    |
| Prunus persica   | 24,493 | 3,204.43                 | 1,088.71                | 4.35                   | 250.14                  | 639.44                    |
| Juglans regia    | 25,252 | 3,200.15                 | 1,076.69                | 4.24                   | 253.84                  | 662.00                    |
| Ostrya rehderiana | 31,130   | 2,907.56               | 990.15                  | 4.00                   | 247.72                  | 647.70                    |
| Ostrya chinensis | 32,669 | 2,901.71                 | 958.97                  | 3.94                   | 243.51                  | 668.90                    |
| **RNA seq PASA**  | 33,115 | 5,076.06                 | 1,100.55                | 5.09                   | 414.69                  | 800.10                    |
| **EVM**           | 36,585 | 3,692.57                 | 1,283.06                | 4.67                   | 274.71                  | 1,197.00                  |
| **PASA update**   | 36,439 | 4,067.94                 | 1,384.96                | 5.27                   | 320.73                  | 1,253.00                  |
| **Final**         | 27,381 | 3,948.29                 | 1,415.09                | 5.16                   | 345.16                  | 1,165.54                  |
We also annotated non-coding RNAs in the *C. fangiana* genome. tRNAscan-SE\(^\text{44}\) was used to detect putative transfer RNAs (tRNAs) with eukaryotic parameters, resulting in the identification of 632 tRNAs. To identify other non-coding RNAs, INFERNAL\(^\text{45}\) was used to perform searches against the Rfam\(^\text{46}\) database, resulting in the identification of 936 ribosomal RNAs (rRNAs), 197 microRNAs (miRNAs), 117 small nuclear RNAs (snRNAs), and 232 small nucleolar RNAs (snoRNAs) (Table 8).

### Data Records

The sequencing data including the Illumina genome data (SRA accession: SRX6070999-SRX6071006), Hi-C data (SRA accession: SRX6071007), and Illumina transcriptome data (SRA accession: SRX6070994-SRX6070998, SRX6071008) were submitted to the NCBI Sequence Read Archive (SRA) database under BioProject accession number PRJNA54802747. The assembled genome was deposited at DDJB/ENA/GenBank under accession number VIBQ0000000048. Repeat annotations, gene model annotations and non-coding RNA annotations, the CDS sequences for the coding and non-coding genes, the protein sequences for the coding genes, as well as two custom Perl scripts were deposited at figshare\(^\text{49}\).

### Technical Validation

#### Assessment of the genome assembly.

We evaluated the completeness of the *C. fangiana* genome assembly in two ways. First, all the paired-end reads were mapped to the assembly genome with BWA. The aligned outputs were then analyzed using SAMtools. The mapping rate for each library was above 90% (Table 9). Furthermore, the coverage of the genome after gap elimination was 99.74%, with 95.05% having at least 100x coverage. Benchmarking Universal Single-Copy Orthologs (BUSCO)\(^\text{50}\) was also used to evaluate the completeness of the genome assembly. 95.30% of the “complete BUSCOs” were successfully identified in the assembly, and the proportion of “missing BUSCOs” was only 4.10% (Table 10). These results demonstrate the high reliability and completeness of the reported genome assembly.

Finally, we evaluated the assembly of the eight chromosomes. To this end, the anchored genome was split into ‘bins’ of 100 kb in length. The number of Hi-C read pairs covered by any two ‘bins’ was used to define the signal for the interaction between those ‘bins’, and these signal intensities were plotted in the form of a heat map. The signal intensities clearly divided the ‘bins’ into eight distinct groups, demonstrating the high quality of the chromosome assembly (Fig. 3).

#### Improvement of gene annotation quality.

To maximize the reliability of the gene annotation process, repeat regions in the assembled genome were masked before gene annotation. Mirroring the procedure used to filter gene annotation, EVM was initially used to merge the results obtained by *de novo*, homolog-based, and transcriptome-based predictions. Genes were then discarded if: (1) their CDS length was below 150 bp; (2) their putative coding regions could not be accurately translated into protein sequences; (3) they possessed early termination codons; or (4) they were only supported by *de novo* predictions. In addition, PASA was used to identify untranslated regions (UTRs).

### Tables

**Table 7.** Summary of functional annotation in the *C. fangiana* genome.

| Type   | Gene number | % in genome |
|--------|-------------|-------------|
| Total  | 27,381      |             |
| GO     | 19,879      | 71.87       |
| KEGG   | 18,845      | 68.83       |
| InterProScan | 15,582      | 56.91       |
| Pfam   | 19,688      | 71.90       |
| Uniprot_sprot | 19,733      | 72.07       |
| Uniprot_trembl | 24,110      | 88.05       |
| All    | 25,836      | 94.36       |

**Table 8.** Summary of non-coding genes in the *C. fangiana* genome.

| Type  | Number | Average length (bp) | Total length (bp) | % of genome |
|-------|--------|---------------------|-------------------|-------------|
| tRNA  | 632    | 76.71               | 48,478            | 0.01255     |
| rRNA  | 936    | 122.70              | 114,844           | 0.03136     |
| miRNA | 197    | 124.27              | 24,481            | 0.00669     |
| snRNA | 117    | 141.58              | 16,565            | 0.00452     |
| snoRNA| 232    | 97.28               | 22,570            | 0.00616     |
| SRP   | 9      | 280.33              | 2,523             | 0.00069     |
| other ncRNA | 2,317 | 109.13             | 252,859           | 0.06905     |
| Total | 4,440  | 108.63              | 482,320           | 0.12490     |
This work relied on many software tools. The versions, settings and parameters of these tools are given below.

(1) **FastQC**: version 0.11.5, default parameters;
(2) **Scythe**: version 0.994 BETA, parameters: -q sanger --quiet;
(3) **Sickle**: version 1.33, parameters: pe -t sanger -q 20 -l 50 -n --quiet;
(4) **Lighter**: version 1.0.7, parameters: -K 21 360000000;
(5) **FastUniq**: version 1.1, default parameters;
(6) **Trinity**: trinityrnaseq-2.6.4, parameters: --seqType fq --JM 260G;
(7) **CD-Hit**: version 4.6, default parameters;
(8) **Jellyfish**: version 1.1.10, parameters: count command: -m 17 -s 4G -c 7, dump command: -c -t, histo command: default parameters;
(9) **GenomeScope**: version 2.0, parameters: 17 (k-mer length) 150 (read length);
(10) **Platanus**: version 1.2.1, default parameters for the all three steps,
(11) **GapCloser**: version 1.12, parameter: -l 150;
(12) **HaploMerger2**: version HaploMerger2_20151124, default parameters for the followed running processes: carrying out batchA to batchE with the recommended pipeline, among which batchA was repeated 3 times and batchD was repeated 2 times, respectively;
(13) **HiC-Pro**: version 2.10.0, default parameters;
(14) **LACHESIS**: released in 2017, parameters: CLUSTER_MIN_RE_SITES=36 CLUSTER_MAX_LINK_DENSITY=1 CLUSTER_NONINFORMATIVE_RATIO=8 ORDER_MIN_N_RES_IN_TRUN=22 ORDER_MIN_N_RES_IN_SHREDS 22;
(15) **BWA**: version 0.7.12-r1039, default parameters;
(16) **SAMtools**: version 1.5, parameters:

### Table 9. Mapping ratio of Illumina DNA reads for the *C. fangiana* genome.

| Reads | Mapping rate (%) | Coverage | Value (%) |
|-------|------------------|----------|-----------|
| Library (bp) | | | |
| 230 | 93.19 | at least 1x | 99.74 |
| 500 | 91.04 | at least 10x | 99.28 |
| 800 | 90.54 | at least 20x | 98.87 |
| 2 k | 99.07 | at least 30x | 98.87 |
| 5 k | 99.42 | at least 50x | 98.51 |
| 10 k | 98.93 | at least 80x | 97.84 |
| 20 k | 98.36 | at least 100x | 95.03 |

### Table 10. Assessment of BUSCOs in the *C. fangiana* genome.

| BUSCOs | Number | Percent |
|--------|--------|---------|
| Complete BUSCOs | 1,372 | 95.30% |
| Complete and single-copy BUSCOs | 1,329 | 92.30% |
| Complete and duplicated BUSCOs | 45 | 3.00% |
| Fragmented BUSCOs | 8 | 0.60% |
| Missing BUSCOs | 60 | 4.10% |
| Total BUSCO groups searched | 1,440 | |

**Fig. 3** Heat map of chromosomal interactions in the *C. fangiana* genome. Cfa01–Cfa08 represent the eight chromosomes in the *C. fangiana* genome. The horizontal and vertical coordinates represent the order of each ‘bin’ on the corresponding chromosome.

**Code availability**

This work relied on many software tools. The versions, settings and parameters of these tools are given below.

(1) **FastQC**: version 0.11.5, default parameters;
(2) **Scythe**: version 0.994 BETA, parameters: -q sanger --quiet;
(3) **Sickle**: version 1.33, parameters: pe -t sanger -q 20 -l 50 -n --quiet;
(4) **Lighter**: version 1.0.7, parameters: -K 21 360000000;
(5) **FastUniq**: version 1.1, default parameters;
(6) **Trinity**: trinityrnaseq-2.6.4, parameters: --seqType fq --JM 260G;
(7) **CD-Hit**: version 4.6, default parameters;
(8) **Jellyfish**: version 1.1.10, parameters: count command: -m 17 -s 4G -c 7, dump command: -c -t, histo command: default parameters;
(9) **GenomeScope**: version 2.0, parameters: 17 (k-mer length) 150 (read length);
(10) **Platanus**: version 1.2.1, default parameters for the all three steps,
(11) **GapCloser**: version 1.12, parameter: -l 150;
(12) **HaploMerger2**: version HaploMerger2_20151124, default parameters for the followed running processes: carrying out batchA to batchE with the recommended pipeline, among which batchA was repeated 3 times and batchD was repeated 2 times, respectively;
(13) **HiC-Pro**: version 2.10.0, default parameters;
(14) **LACHESIS**: released in 2017, parameters: CLUSTER_MIN_RE_SITES=36 CLUSTER_MAX_LINK_DENSITY=1 CLUSTER_NONINFORMATIVE_RATIO=8 ORDER_MIN_N_RES_IN_TRUN=22 ORDER_MIN_N_RES_IN_SHREDS 22;
(15) **BWA**: version 0.7.12-r1039, default parameters;
(16) **SAMtools**: version 1.5, parameters:
view command: -bS, sort command: -O BAM, depth command: -Q 40; mpileup command: -DSug -C 50, default parameters for the rmdup, index and flagstat commands; (17) Picard: version 1.80, parameters: SORT_ORDER = coordinate RGPI = illumina RGPI = illumina; (18) GATK: version 3.3-0-g37228af, default parameters for the two programs RealignerTargetCreator and IndelRealigner; (19) bcf tools: version 0.1.19-44428 cd, parameters: view -Ncg; (20) TRF: version 4.07b, parameters: Match=2 Mismatch=7 Delta=7 PM=80 PI=10 Minscore=50 MaxPeriod=500 -d –h; (21) RepeatModeler: version 1.0.4, parameters: -pa 30 -database Fan; (22) RepeatMasker: version open-4.0.5, parameters: -pa 30 -species all -nolow -norma -no_is -gff; (23) RepeatProteinMasker: version 2.1, parameters: -engine abblast -nolowSimple -pvalue 1e-04; (24) LTR-FINDER: version 1.0.5, default parameters; (25) Augustus: version 2.5.5, parameters: --species=arabidopsis; (26) Geneid: version 1.4, parameters: -3 -P; (27) GeneMark: version 3.47, parameters: -gff3; (28) GlimmerHMM: version 3.0.4, default parameters; (29) SNAP: version 2006-07-28, default parameters; (29) GeneWise: version 2.4.1, parameters: -tfor/-trev -gff; (30) EVM: version 1.1.1, default parameters; (31) PASA: version 2.0.2, parameters: for Launch_PASA_pipeline.pl step: -C -R r Aligners blat gmap, default parameters for the below two steps: asa_asmbls_to_training_set.extract_reference_orfs.pl and pasa_asmbls_to_training_set.dbi; (32) BLASTP: version 2.2.30+, parameters: -eval 1e-5 -outfmt 7; (33) Interproscan: version 5.25-64.0, parameters: -dp -f tsv; (34) tRNAscan-SE: tRNAscan-SE-2.0, default parameters; (35) BUSCO: version 2.0, parameters: -m genome -c 20.

References
1. Christenhusz, M. J. M. & Byng, J. W. The number of known plant species in the world and its annual increase. *Phytotaxa*. 261, 201–217 (2016).
2. Chen, Z. Phylogeny and phytogeography of the Betulaceae. *Acta Phytotaxon. Sin.*. 32, 1–31 (1994).
3. Botta, R., Emanuel, E., Me, G., Sacerdoti, E. & Vallania, R. Indagine cariologica in alcune specie del genere *Corylus* (In Italian). *Rivista di Ortoforparallticultura Italiana*. 70, 323–329 (1986).
4. Snow, N. & Goldblatt, P. Chromosome number in *Tydendron* (Fagales, Toidendraceae). *Ann. Mo. Bot. Gard.* 79, 906–907 (1992).
5. Jin, W. Y., Cai, X. F., Na, M. K., Lee, J. J. & Bae, K. H. Triterpenoids and Diarylheptainoids from *Alnus hirsuta* inhibit HIF-1 in AGS Cells. *Arch. Pharm. Res.*. 30, 412–418 (2007).
6. Crane, P. R. & Blackmore, S. Evolution, systematics, and fossil history of the Hamamelidaceae. (Published for the Systematics Association by the Clarendon Press, 1989).
7. Salpiùri, J. et al. Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nat. Genet.* 49, 904–912 (2017).
8. Griesmann, M. et al. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science*. 361, eaat1743 (2018).
9. Yang, Y. et al. Genomic effects of population collapse in a critically endangered ironwood tree *Ostrya rehderiana*. *Nat. Commun.*. 9, 5449 (2018).
10. Lancaster, R. & Rix, M. 705. *Carpinus fagifolia*. *Curtisi Botanical Magazine*. 28, 103–110 (2011).
11. Doyle, J. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull.*. 19, 11–15 (1987).
12. Chang, S. J., Puryear, J. & Cairney, J. A simple and efficient method for isolating RNA from pine trees. *Plant Mol. Biol. Rep.*. 11, 113–116 (1993).
13. Song, L., Song, L., Florea, L. & Langmead, B. Lighter: fast and memory-efficient sequencing error correction without counting. *Genome Biol.*. 15, 509 (2014).
14. Xu, H. et al. FastUniq: a fast de novo duplicates removal tool for paired short reads. *PloS One*. 7, e52249 (2012).
15. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*. 29, 644–652 (2011).
16. Li, W. Z., Jaroszewski, L. & Godzik, A. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*. 18, 77–82 (2002).
17. Marcias, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 27, 764–770 (2011).
18. Vurture, G. W. et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*. 33, 2202–2204 (2017).
19. Kajitani, R. et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.*. 24, 1384–1395 (2014).
20. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*. 1, 18 (2012).
21. Huang, S. et al. HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Res.*. 22, 1581–1588 (2012).
22. Servant, N. et al. Hi-C-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.*. 16, 1–11 (2015).
23. Burton, J. N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.*. 31, 1119–1125 (2013).
24. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*.*1303, 3997* (2013).
25. Li, H., Bob, H., ALEC, W., Tim, F. & Jue, R. The Sequence Alignment/Map format and SAMTools. *Bioinformatics*. 25, 2078–2079 (2009).
26. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*. 20, 1297–1303 (2010).
27. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*. 27, 573–580 (1999).
28. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Carr. Protoc. Bioinformatics*. 25, 4.10.1–4.10.14 (2009).
29. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.*. 35, W265–W268 (2007).
30. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*. 24, 637–644 (2008).
31. Parra, G., Blanco, E. & Guigó, R. Geneid in *Drosophila*. *Genome Res.*. 10, 511–515 (2000).
32. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Bioinformatics*. 26, 1107–1115 (1998).
33. Majors, W. H., Peta, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene finders. *Bioinformatics*. 20, 2878–2879 (2004).
34. Vop, H. Gene finding in novel genomes. *MC Bioinformatics*. 5, 59 (2004).
35. Gertz, E. M., Yu, Y. K., Agarwala, R., Schäffer, A. A. & Altschul, S. F. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. 4, 41 (2006).
Acknowledgements

This work was equally supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB31000000) and National Key Research and Development Program of China (2017YFC0505203), and further by National Natural Science Foundation of China (31590821 and 31900201), the Fundamental Research Funds for the Central Universities (Grant No. Izujbk-2019-77) and the Lanzhou University’s ”Double First-Class” Guided Project – Team Building Funding - Research Startup Fee for Jianquan Liu and Yongzhi Yang. We thank Dr. Mingcheng Wang and Cheng Zhang for their technical help.

Author contributions

Yongzhi Yang designed and conceived this work; Lei Zhang and Guoqian Hao collected the materials and prepared DNA and RNA for sequencing; Zefu Wang and Xiaoyue Yang analyzed the data. Xiaoyue Yang wrote the manuscript with other authors’ help; Jianquan Liu and Yongzhi Yang revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

The Creative Commons Public Domain Dedication waiver http://creativecommons.org/publicdomain/zero/1.0/ applies to the metadata files associated with this article.

© The Author(s) 2020