**Table S1 Illumina sequencing reads used for genome size estimation and the correction of the genome assembly**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Library | Data (Gb) | Depth (×) | Q20 (%) | Q30 (%) |
| 350 bp | 54.55 | 83.26 | 97.33 | 92.76 |
| Total | 54.55 | 83.26 | --- | --- |

**Table S2 Size distributions of the PacBio subreads**

|  |  |  |  |
| --- | --- | --- | --- |
| Length (bp) | Number (bp) | Total Length (bp) | Average Length (bp) |
| 500–2,000 | 638,299 | 769,627,686 | 1,205.75 |
| 2,000–4,000 | 613,551 | 1,806,954,490 | 2,945.08 |
| 4,000–6,000 | 469,478 | 2,330,983,492 | 4,965.05 |
| 6,000–8,000 | 397,190 | 2,771,110,809 | 6,976.79 |
| 8,000–10,000 | 350,922 | 3,151,254,832 | 8,979.93 |
| 10,000–12,000 | 315,116 | 3,459,904,616 | 10,979.78 |
| 12,000–14,000 | 279,637 | 3,629,020,526 | 12,977.61 |
| 14,000–16,000 | 244,232 | 3,657,754,402 | 14,976.56 |
| 16,000–18,000 | 215,024 | 3,651,178,915 | 16,980.33 |
| >18,000 | 1,472,248 | 41,754,242,787 | 28,360.88 |
| Total | 4,995,697 | 66,982,032,555 | 13,408 |

**Table S3 Statistical details of the PacBio subreads**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Type** | **Reads\_num** | **Total\_base (bp)** | **Read\_N50** | **Mean** | **Maximum** |
| Sequel | 4,995,697 | 66,982,032,555 | 22,633 | 13,408 | 90,903 |

**Table S4 The initial C. mollissima assembly, as generated by different tools**

|  |  |  |  |
| --- | --- | --- | --- |
| **Assembly tools** | **WTGB** | **Canu** | **WTGB+Canu (merge)** |
| Genome size (bp) | 782,890,668 | 994,461,872 | 793,805,471 |
| Contig number | 2,527 | 4,217 | 1,381 |
| Maximum contig length (bp) | 5,709,809 | 5,367,888 | 25,283,183 |
| Contig N50 (bp) | 1,046,151 | 579,746 | 3,382,036 |
| Contig N90 (bp) | 96,478 | 83,619 | 182,580 |
| GC content (%) | 35.13 | 35.10 | 35.09 |

**Table S5 Statistical details of the Hi-C library data**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **library-ID** | **Read Pairs Number** | **Base Number** | **GC Content (%)** | **≥Q30 (%)** | **Depth** |
| Hi-C01 | 370,041,098 | 110,833,450,402 | 37.66 | 91.85 | 170× |

**Table S6 Details of the alignment of the clean Hi-C data to the initial assembly**

|  |  |  |
| --- | --- | --- |
| **Mapping Type** | **Number** | **Ratio (%)** |
| Total read pairs | 370,041,098  | 100 |
| Mapped reads | 663,321,111  | 89.63 |
| Unique mapped read pairs | 210,652,422  | 56.93 |

**Table S7 Statistical details of all types of Hi-C sequencing pairs**

|  |  |  |
| --- | --- | --- |
| **Type** | **Number** | **Ratio (%)** |
| Unique paired alignments | 210,652,422  | 100 |
| Valid interaction pairs | 161,128,476  | 76.49 |
| Dangling end pairs | 10,555,700  | 5.01 |
| Re-ligation pairs | 2,518,240  | 1.2 |
| Self-cycle pairs | 25,336,948  | 12.03 |
| Dumped pairs  | 11,113,058  | 5.28 |

**Table S8 Evaluation using BUSCO**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Species** | **Complete BUSCOs** | **Complete and single-copy BUSCOs** | **Complete and duplicated BUSCOs** | **Fragmented BUSCOs** | **Missing BUSCOs** |
| chestnut | 1,492 (92.44%) | 1,433 (88.79%) | 59 (3.66%) | 24 (1.49%) | 98 (6.07%) |

**Table S9 Alignment of the Illumina reads to the C. mollissima genome assembly**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Library** | **Total reads** | **Mapped reads** | **Mapped (%)** | **Properly mapped reads** | **Properly mapped (%)** |
| 350bp  | 422,116,203 | 415,232,639 | 98.37 | 387,854,968 | 93.08 |

**Table S10 Functional annotation of the predicted protein-coding genes**

|  |  |  |
| --- | --- | --- |
| **Database** | **Number of genes annotated**  | **Percentage (%)** |
| GO | 16,709 | 49.67% |
| KEGG | 9,388 | 27.91% |
| KOG | 19,248 | 57.22% |
| TrEMBL | 32,825 | 97.58% |
| Nr | 32,844 | 97.64% |
| All | 32,874 | 97.73% |

GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; KOG: eukaryotic orthologous groups of proteins; Nr：GenBank non-redundant protein database

**Table S11 Castanea mollissima gene prediction**

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Software** | **Species** | **Number of genes** |
| *Ab initio* | Genscan | - | 23,855 |
| Augustus | - | 35,934 |
| GlimmerHMM | - | 27,074 |
| GeneID | - | 42,283 |
| SNAP | - | 14,071 |
| Homology-based | GeMoMa | *Arabidopsis thaliana* | 24,250 |
| *Oryza sativa* | 25,274 |
| European ash | 25,855 |
| *Quercus suber* | 42,632 |
| RNA-seq | PASA | *-* | 29,117 |
| TransDecoder | *-* | 46,795 |
| GeneMarkS-T | *-* | 28,020 |
| Integration | EVM | - | 33,638 |

**Table S12 Statistical details of the identified genes**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Software** | **Gene Num** | **GeneLen****(bp)** | **AveGeneLen(bp)** | **ExonLen****(bp)** | **AveExonLen(bp)** | **IntronLen****(bp)** | **AveIntronLen(bp)** |
| EVM | 33,638 | 157,948,859 | 4,695.54 | 42,540,186 | 252.51 | 115,408,673 | 685.05 |

**Table S13 Statistical details of the identified repeat sequences**

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **Number** | **Length（bp）** | **Percentage (%)** |
| ClassI | 423,675  | 235,886,704  | 34.24 |
| ClassI/DIRS | 13,697  | 12,004,658  | 1.74 |
| ClassI/LARD | 12,837  | 3,961,957  | 0.58 |
| ClassI/LINE | 123,740  | 46,315,237  | 6.72 |
| ClassI/LTR/Copia | 111,135  | 59,762,237  | 8.67 |
| ClassI/LTR/Gypsy | 136,445  | 106,363,325  | 15.44 |
| ClassI/LTR/Unknown | 1,146  | 398,253  | 0.06 |
| ClassI/PLE | 526  | 133,012  | 0.02 |
| ClassI/SINE | 4,981  | 957,920  | 0.14 |
| ClassI/TRIM | 2,422  | 1,458,016  | 0.21 |
| ClassI/Unknown | 16,746  | 4,532,112  | 0.66 |
| ClassII | 101,191  | 38,048,432  | 5.52 |
| ClassII/Crypton | 9  | 637  | 0 |
| ClassII/Helitron | 40,400  | 12,826,319  | 1.86 |
| ClassII/MITE | 1,630  | 422,650  | 0.06 |
| ClassII/Maverick | 1,218  | 362,292  | 0.05 |
| ClassII/TIR | 54,312  | 23,681,234  | 3.44 |
| ClassII/Unknown | 3,622  | 755,300  | 0.11 |
| PotentialHostGene | 89,112  | 21,644,354  | 3.14 |
| SSR | 720  | 160,173  | 0.02 |
| Unknown | 278,155  | 71,100,097  | 10.32 |
| Total | 892,853  | 366,839,546  | 53.24 |



**Fig S1. Distribution of subread lengths**

The red dotted line represents the N50 length of subreads.



**Figure S2 Nineteen-k-mer estimation of genome size**

The genome size of *Castanea mollissima* was estimated based on Illumina genomic reads from insert size of ~350 bp library.



**Fig S3 Differences in the genes predicted by the *Ab initio*, homolog, and RNA-sequencing methods**