iScience, Volume *23*

Supplemental Information

The Chromosome Level Genome and Genome-wide Association Study for the Agronomic Traits of *Panax Notoginseng*

Guangyi Fan, Xiaochuan Liu, Shuai Sun, Chengcheng Shi, Xiao Du, Kai Han, Binrui Yang, Yuanyuan Fu, Minghua Liu, Inge Seim, He Zhang, Qiwu Xu, Jiahao Wang, Xiaoshan Su, Libin Shao, Yuanfang Zhu, Yunchang Shao, Yunpeng Zhao, Andrew KC. Wong, Dennis Zhuang, Wenbin Chen, Gengyun Zhang, Huanming Yang, Xun Xu, Stephen Kwok-Wing Tsui, Xin Liu, and Simon Ming-Yue Lee

Supplementary Information

Supplementary Tables

Table S1. The summary of sequencing data generated by Nanopore. Related to Figure 1, Figure S1 and Figure S2.

Category	Raw data	Corrected data
Base (bp)	178,190,734,045	90,087,661,205
Reads (#)	27,101,176	7,843,544
>5K Reads (#)	13,727,370 (50.65%)	7,822,498 (99.73%)
>5K Base (bp)	149,030,234,572 (83.64%)	90,004,843,293 (99.91%)
>7K Reads (#)	10,074,833 (37.17%)	6,634,864 (84.59%)
>7K Base (bp)	127,085,125,807 (71.32%)	82,276,443,072 (91.33%)
>10K Reads (#)	5,436,814 (20.06%)	3,499,559 (44.62%)
>10K Base (bp)	88,184,506,908 (49.49%)	56,000,524,443 (62.16%)
>13K Reads (#)	2,941,838 (10.86%)	1,882,065 (24.00%)
>13K Base (bp)	59,958,457,599 (33.65%)	37,710,163,159 (41.86%)
>15K Reads (#)	2,076,236 (7.66%)	1,327,213 (16.92%)
>15K Base (bp)	47,912,047,660 (26.89%)	29,988,350,872 (33.29%)
Mean Length (kb)	6.56	11.49
N50 (kb)	9.92	11.61
Median Length (kb)	5.10	9.48

Table S2. **The summary of MPS sequencing data. Related to Figure 1.**

Library	Raw read (M)	Raw base (Mb)	Clean read (M)	Clean base (Mb)	Depth (×)
250	404.57	60686.16	322.06	48309.66	19.64
500	775.64	77564.16	652.08	65207.79	26.51
800	888.85	88885.46	730.91	73090.98	29.71

Note: Genome depth is calculated from the genome size estimated by *k*-mer analysis (here: 2.46 Gb).

Table S3. **17-mer statistics information based on short insert-size reads. Related to Figure S3.**

***k*-mer No.**	**Peak**	**Genome Size**	**Used Bases**	**Used Reads**	**×**
101,730,529,320	41	2,463,818,076	121,107,773,000	1,211,077,730	48.81

Table S4. Statistics of the assembly using Smart*denovo*. Related to Figure 1.

Type	Smartdenovo	Pilon
Total number	16,469	16,469
Total length of (bp)	2,242,091,458	2,254,342,782
Gap number (bp)	-	5
Average length (bp)	136,140.11	136,884.00
Contig N50 (bp)	219,818	220,891
Contig N90 (bp)	59,598	59,761
Maximum length (bp)	7,102,366	7,102,368
Minimum length (bp)	7,760	7,763
GC content is (%)	33.82	34.02
BUSCO score	C:57.3%, F:6.6%, M:36.1%	C:90.9%, F:2.2%, M:6.9%

Table S5. The summary of sequencing data generated by Hi-C library using BGISEQ-500. Related to Figure 1, Figure S4 and Figure S5.

Type	R1	R2
Total	2,953,199,263	2,953,199,263
Mapped	2,510,887,432	2,441,423,245
Global	2,484,156,564	2,410,785,590
Local	26,730,868	30,637,655
Mapping ratio	84.97%	82.63%

Table S6. Statistics of the final chromosome assembly using Hi-C data. Related to Figure 1.

Chromosome ID	Length (bp)
chr1	219,051,668
chr2	200,043,122
chr3	197,455,767
chr4	178,740,292
chr5	177,835,634
chr6	173,391,873
chr7	162,606,337
chr8	162,228,736
chr9	155,173,254
chr10	137,708,257
chr11	123,133,882
chr12	113,002,069

Table S7. The statistics of transposable elements of updating genome assembly. Related to Figure 1.

	RepBase TEs		**TE Proteins**		***De novo***		**Combined TEs**	
	Length (bp)	**%**	**Length**	**%**	**Length (bp)**	**%**	**Length (bp)**	**%**
DNA	27,715,502	1.23	7,997,523	0.35	89,117,447	3.95	114,716,567	5.09
LINE	5,631,253	0.25	1,766,072	0.08	6,560,244	0.29	13,350,999	0.59
LTR	10,134	0.00	-	0.00	15,481	0.00	25,615	0.00
SINE	341,429,901	15.15	363,629,823	16.13	1,674,265,611	74.27	1,697,987,823	75.32
Other	5,335	0.00	240	0.00	47,504	0.00	53,079	0.00
Unknown	-	0.00	-	0.00	1,228,103	0.05	1,228,103	0.05
Total	369,109,612	16.37	373,385,492	16.56	1,750,333,052	77.64	1,782,496,423	79.07

Table S8. Summary of the gene prediction of *P. notoginseng*. Related to Figure 1 and Figure S6.

Gene set	**Gene number**	**BUSCO assessment**
Original version	41,917	C:91.0%[S:82.2%,D:8.8%],F:3.4%,M:5.6%,n:1440
Filtered the genes overlapping with TEs (>0.8)	39,452	C:90.1%[S:81.5%,D:8.6%],F:3.3%,M:6.6%,n:1440
Filtered the genes overlapping with TEs (>0.5)	38,242	C:88.4%[S:79.9%,D:8.5%],F:3.3%,M:8.3%,n:1440

Table S9. Comparison of the repetitive sequences of six species. Related to Figure 2.

Species	DNA (%)	LINE (%)	SINE (%)	LTR (%)	Unknown (%)	Total TEs length (bp)	Total TEs (%)
D. carota	13.49	2.19	0.22	31.72	1.41	195,464,165	46.37
C. annuum	5.24	2.39	0.16	62.93	0.13	2,018,820,950	68.76
S. tuberosum	7.42	3.13	0.29	43.69	0.77	407,295,270	52.69
S. lycopersicum	5.09	1.88	0.16	47.73	0.96	445,626,787	53.81
P. ginseng	4.01	0.57	0.01	66.25	0.11	2,082,049,069	69.75
P. notoginseng	5.09	0.59	0.00	75.32	0.05	1,782,496,423	79.07

Table S10. **Statistics information of transcriptomes sequencing of eight samples. Related to Figure 3.**

Sample	Type	reads number	percentage
R1	Total Reads	68,570,216	--
	Total BasePairs	6,171,319,440	--
	Total Mapped Reads	60,706,857	88.53%
R2	Total Reads	66,892,688	--
	Total BasePairs	6,020,341,920	--
	Total Mapped Reads	60,796,034	90.89%
R3	Total Reads	65,258,974	--
	Total BasePairs	5,873,307,660	--
	Total Mapped Reads	60,457,069	92.64%
L1	Total Reads	69,236,606	--
	Total BasePairs	6,231,294,540	--
	Total Mapped Reads	63,832,103	92.19%
L2	Total Reads	68,605,032	--
	Total BasePairs	6,174,452,880	--
	Total Mapped Reads	62,249,641	90.74%
L3	Total Reads	65,041,040	--
	Total BasePairs	5,853,693,600	--
	Total Mapped Reads	58,782,315	90.38%
F2	Total Reads	69,007,174	--
	Total BasePairs	6,210,645,660	--
	Total Mapped Reads	62,707,160	90.87%
F3	Total Reads	68,125,310	--
	Total BasePairs	6,131,277,900	--
	Total Mapped Reads	63,348,369	92.99%

Supplementary Figures

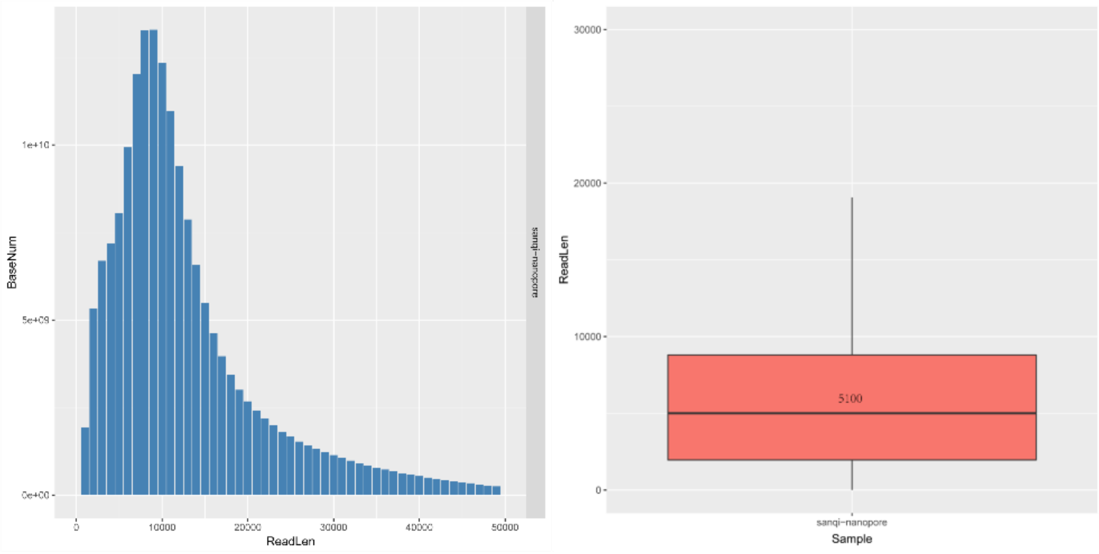

Figure S1. Summary of raw long read length. Related to Table S1.

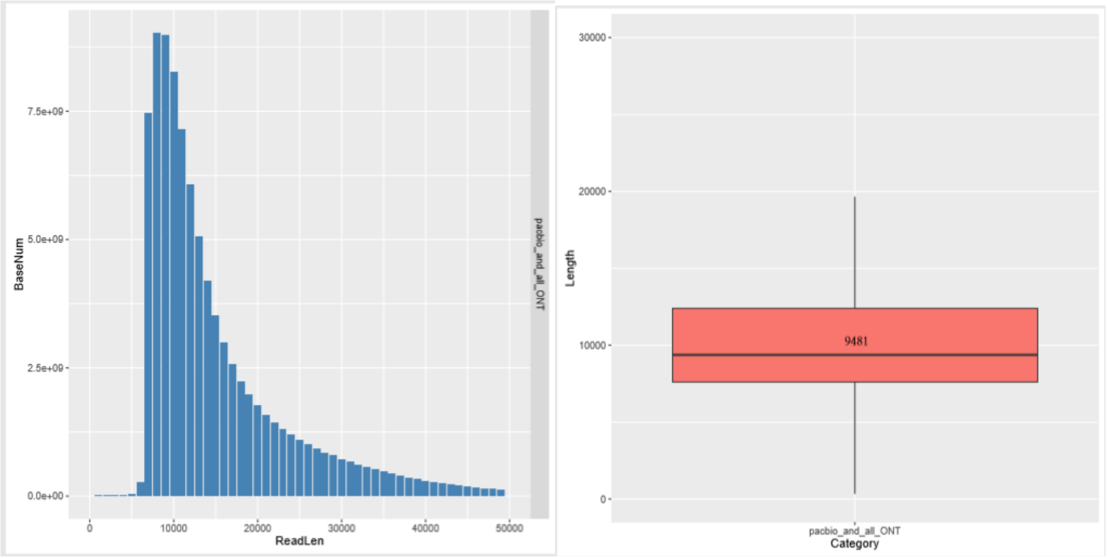

Figure S2. Summary of the length of the corrected long reads. Related to Table S1.

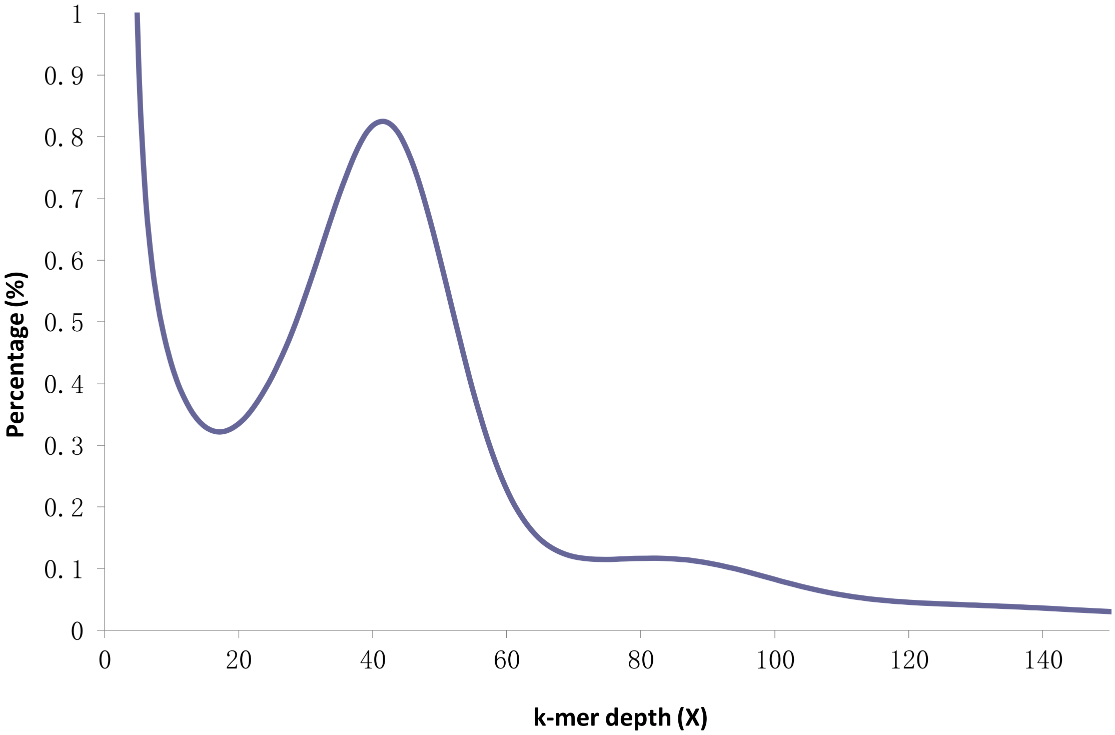

Figure S3. **The 17-mer depth distribution of *P. notoginseng*. Related to Table S3**

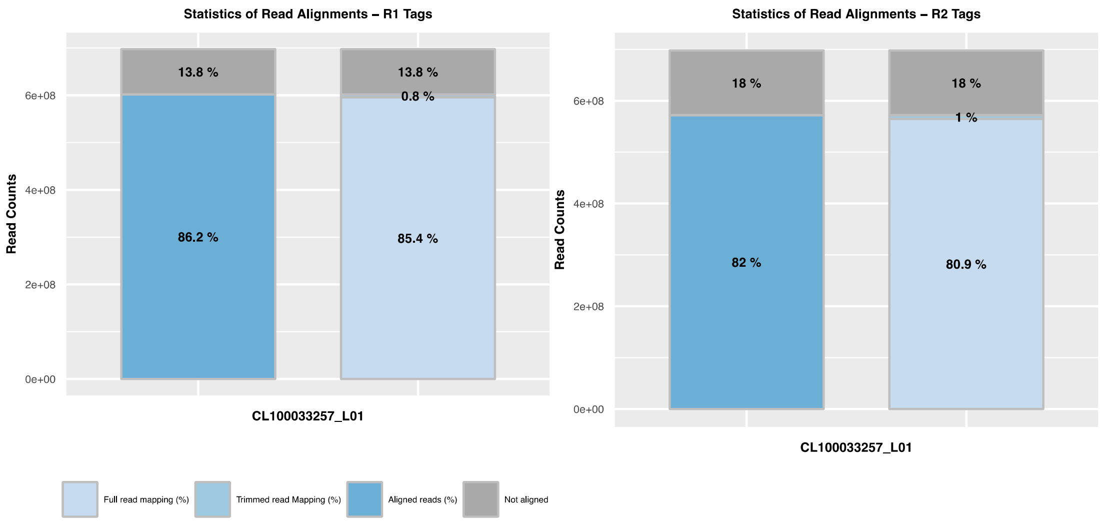

Figure S4. Quality control of Hi-C read. Statistics for the type of separated pair-end read alignment. The aligned read ratio shown in the left bar including full-read and trimmed read mapping. Related to Table S5.

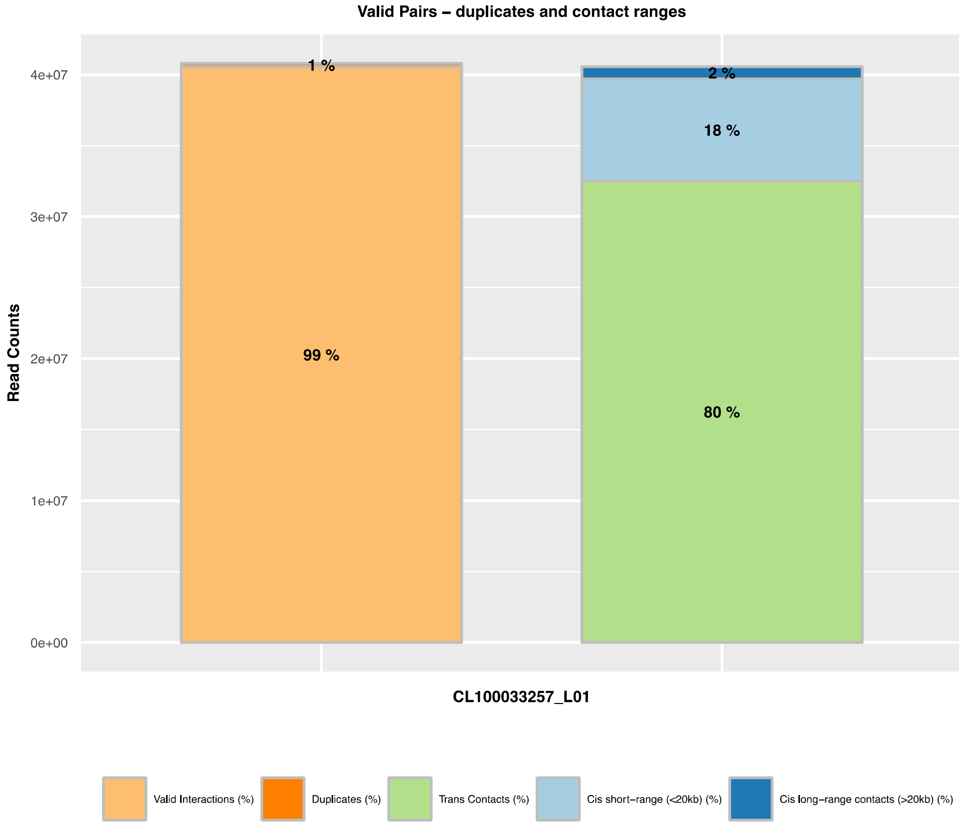

Figure S5. Quality control of Hi-C read. The left bar shows the ratio of duplication for the valid read pairs. For all the non-duplicated reads, the percentage of cis and trans contacts are shown (right bar). Related to Table S5

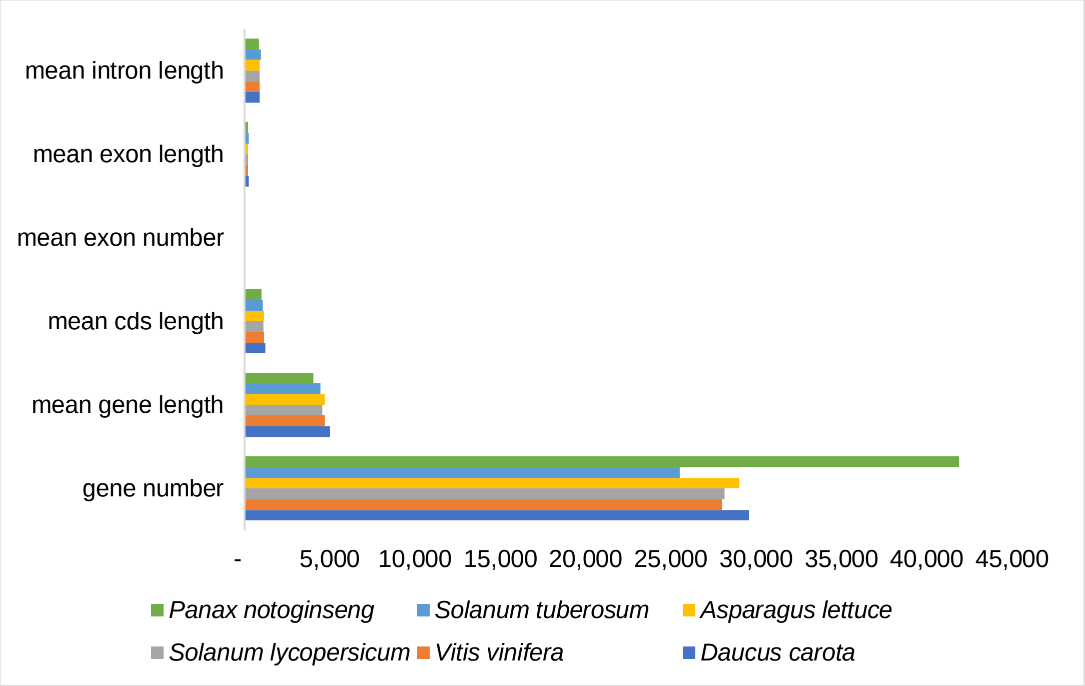

Figure S6. Comparison of the gene structures among *P. notogiseng* and other five species. Related to Table S8.

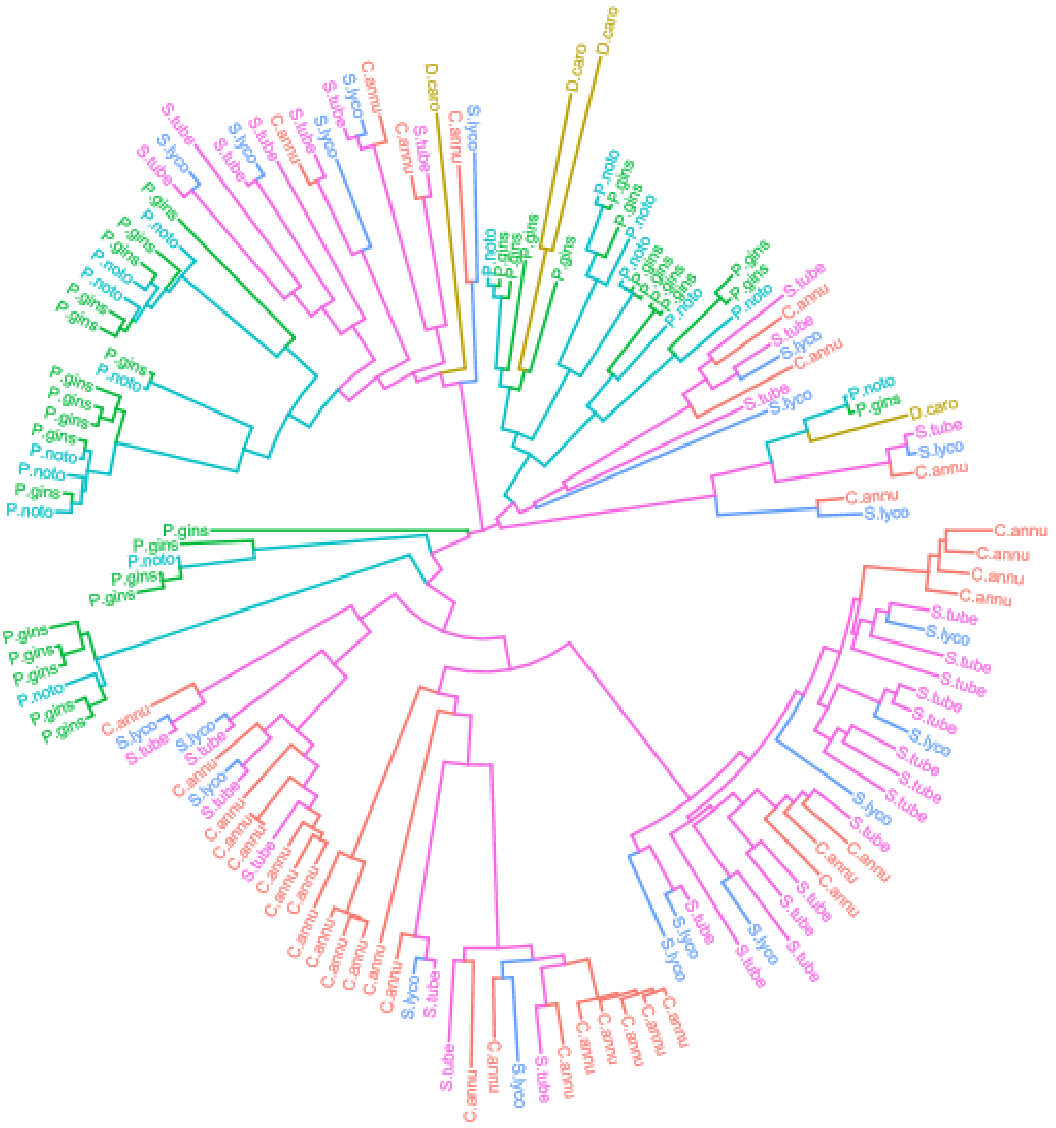

Figure S7. **Comparison of the TIR_NBS_LRR R-genes,** P.noto is the genes of *P. notoginseng*, P.gins is the genes of *P. ginseng*, S.tube is the genes of *Solanum tuberosum*, C.annu is the genes of *Capsicum annuum*, S.lyco is the genes of *Solanum lycopersicum*, D.caro is the genes of *Daucus carota*. Related to Figure 2.

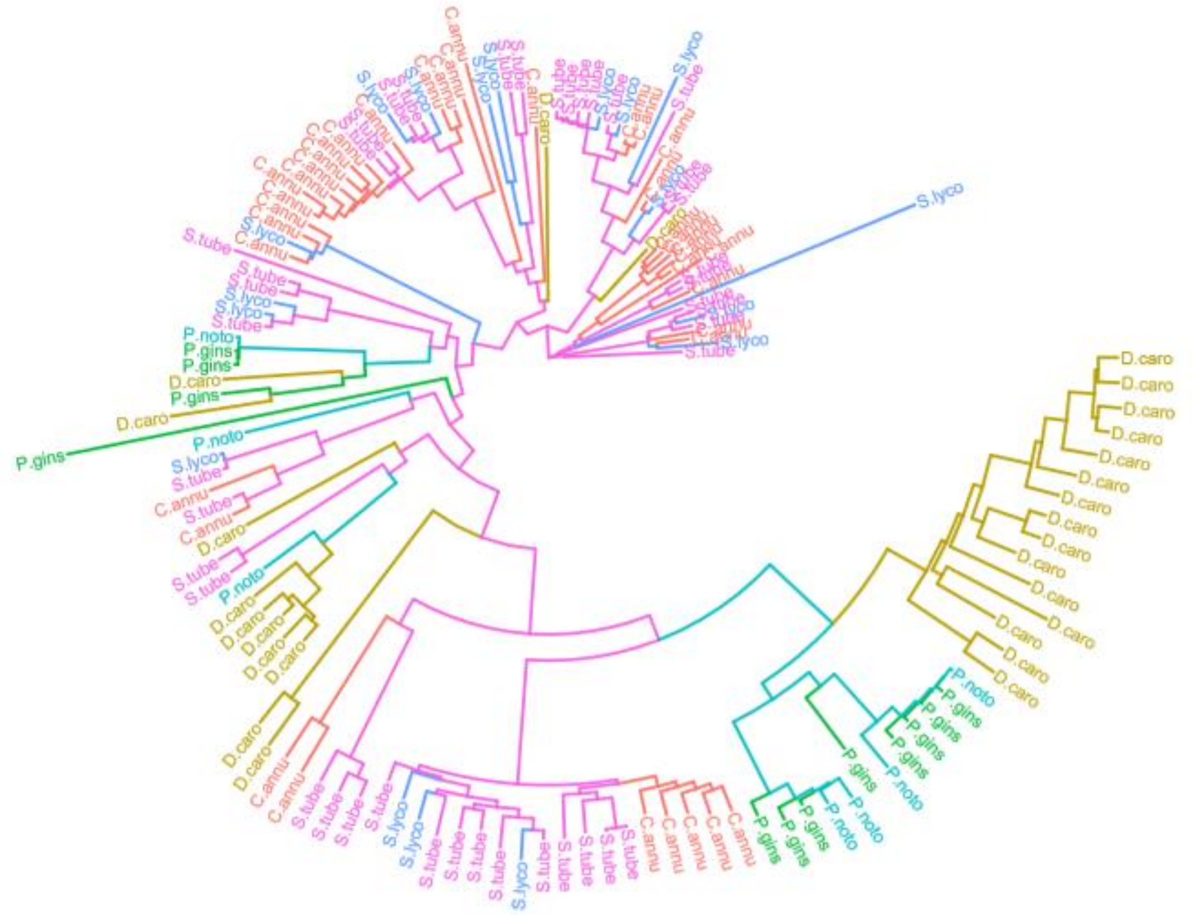

Figure S8. **Comparison of the CC_NBS R-genes**, P.noto is the genes of *P. notoginseng*, P.gins is the genes of *P. ginseng*, S.tube is the genes of *Solanum tuberosum*, C.annu is the genes of *Capsicum annuum*, S.lyco is the genes of *Solanum lycopersicum*, D.caro is the genes of *Daucus carota*. Related to Figure 2.

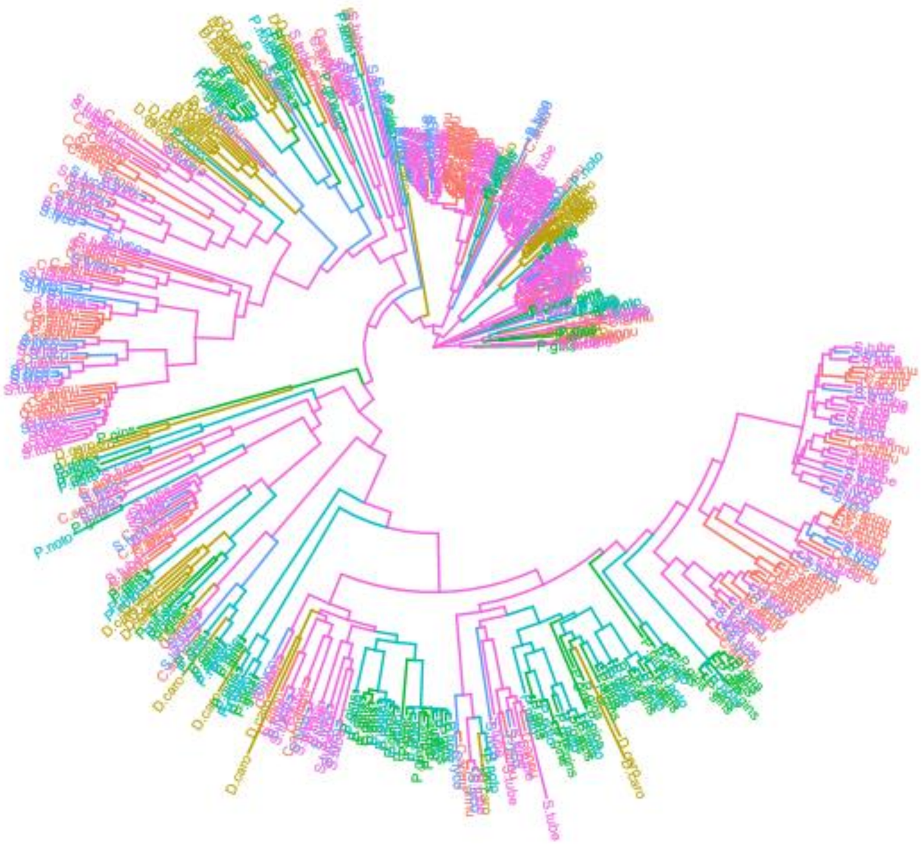

Figure S9. **Comparison of the NBS_LRR R-genes**, P.noto is the genes of *P. notoginseng*, P.gins is the genes of *P. ginseng*, S.tube is the genes of *Solanum tuberosum*, C.annu is the genes of *Capsicum annuum*, S.lyco is the genes of *Solanum lycopersicum*, D.caro is the genes of *Daucus carota*. Related to Figure 2.

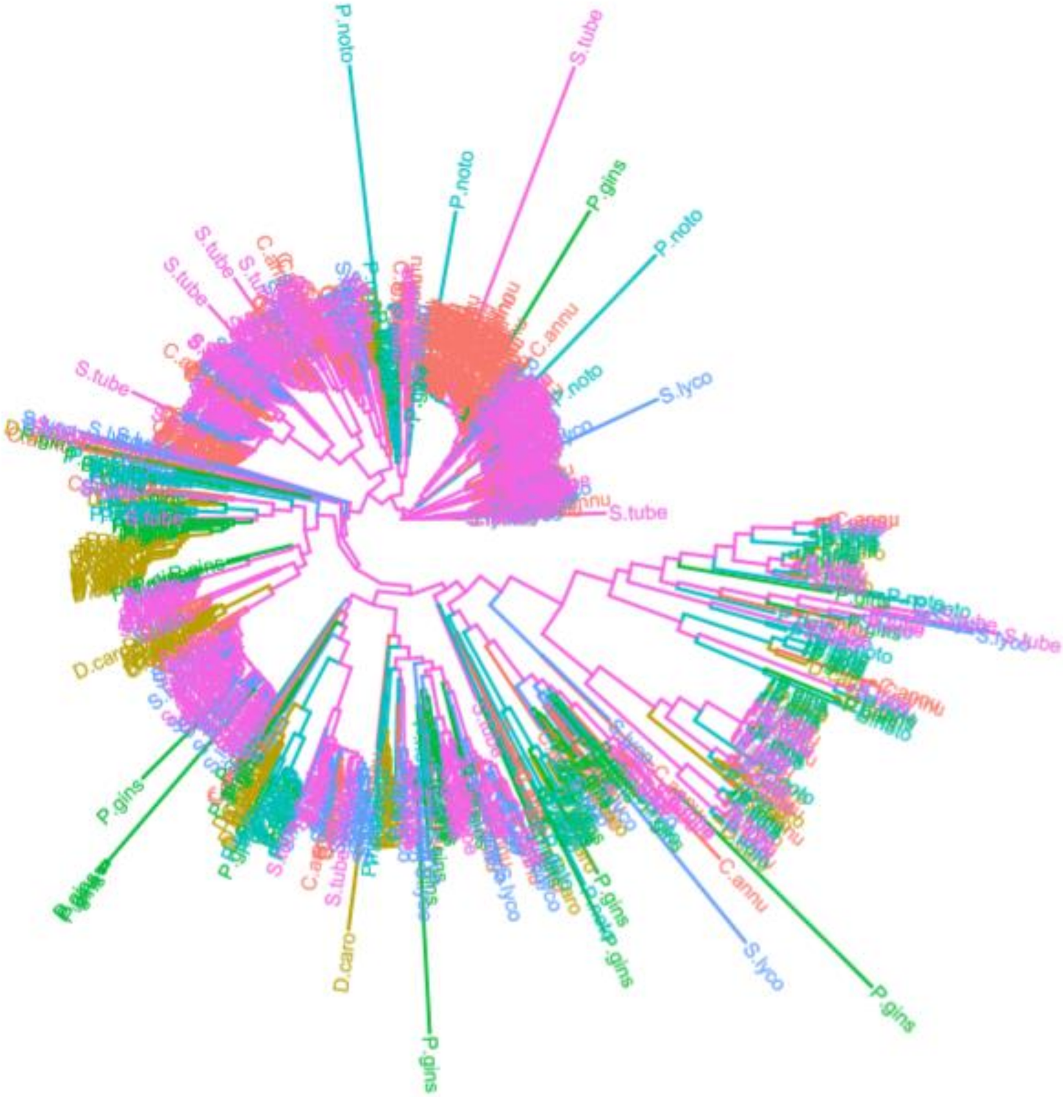

Figure S10. **Comparison of the NBS R-genes**, P.noto is the genes of *P. notoginseng*, P.gins is the genes of *P. ginseng*, S.tube is the genes of *Solanum tuberosum*, C.annu is the genes of *Capsicum annuum*, S.lyco is the genes of *Solanum lycopersicum*, D.caro is the genes of *Daucus carota*. Related to Figure 2.

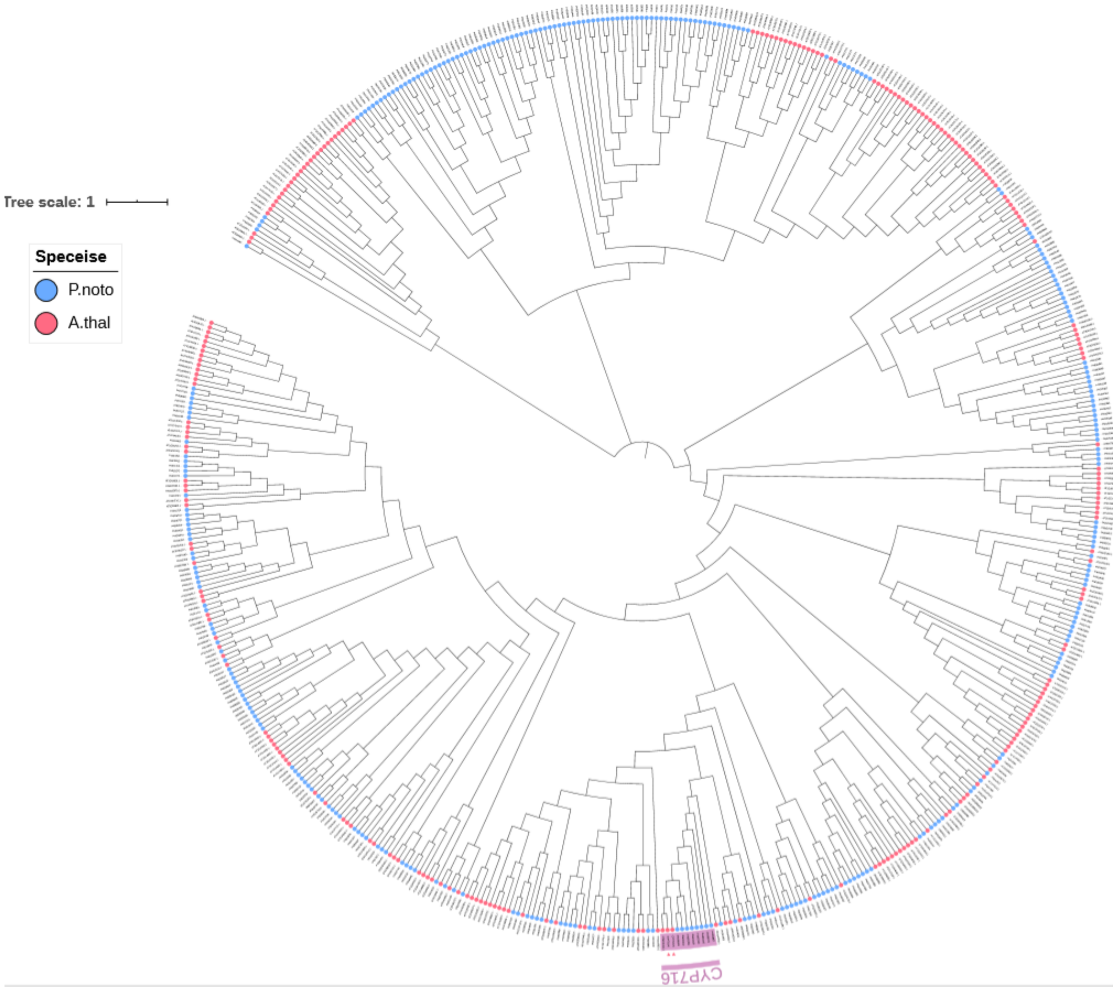

Figure S11. **The gene trees of CYP450 of *P. notoginseng* and *A. thaliana*. Related to Figure 3.**

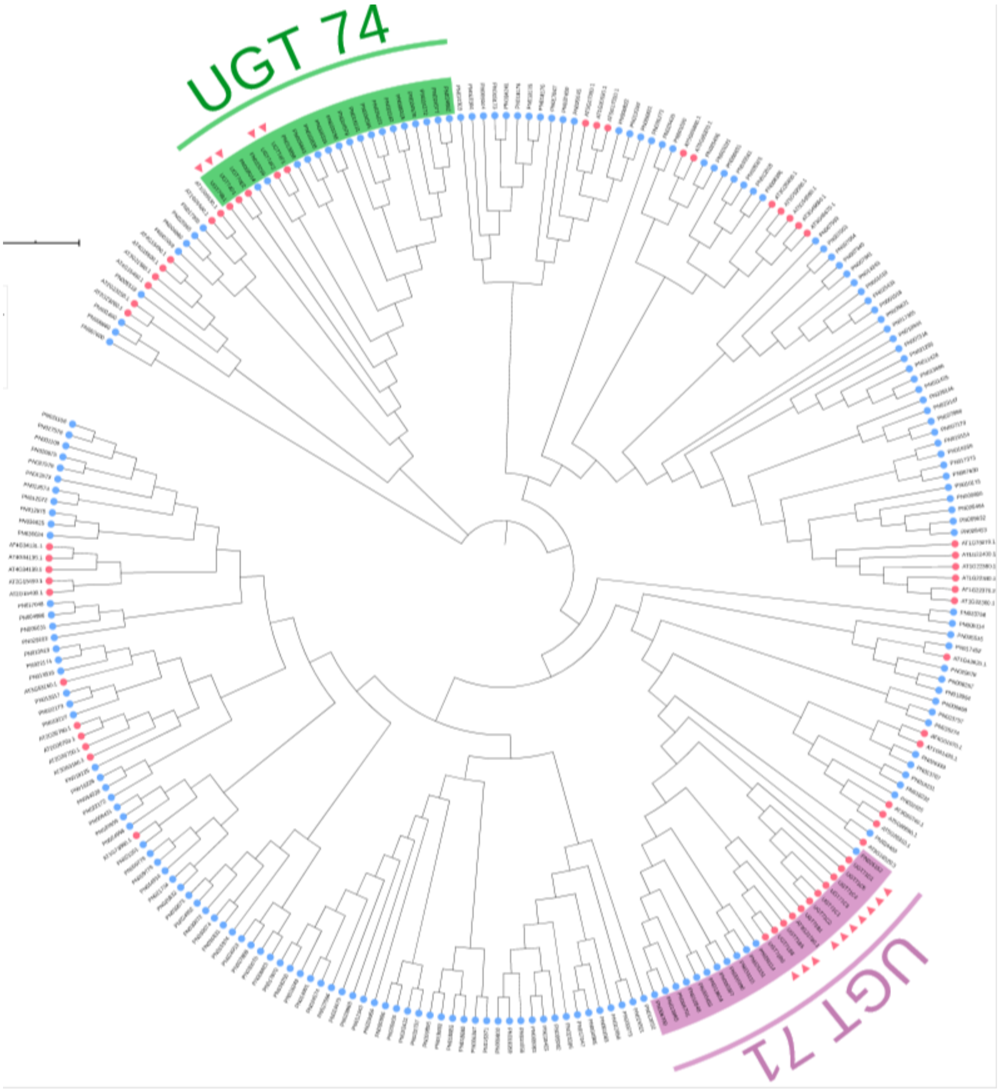

Figure S12. Phylogenetic analysis for classifying the UGT subfamilies of *P. notoginseng* based on the subfamily class of *P. ginseng* and *A. thaliana.* Related to Figure 3.

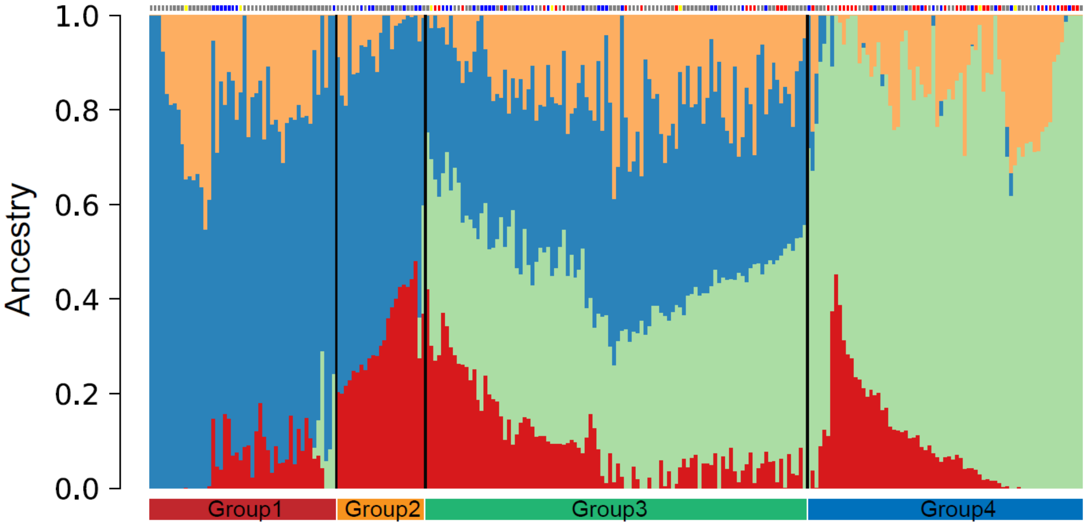

Figure S13. **The population structure of this resequencing population. Related to Figure 4.**

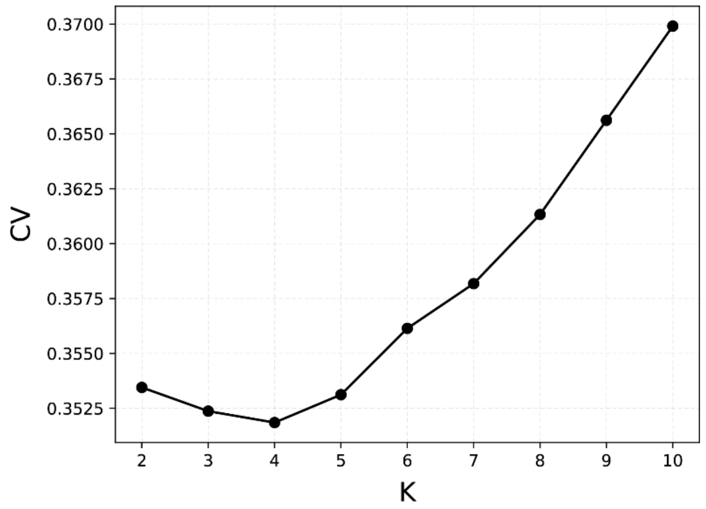

Figure S14. **The estimated best K if this population structure. Related to Figure 4.**

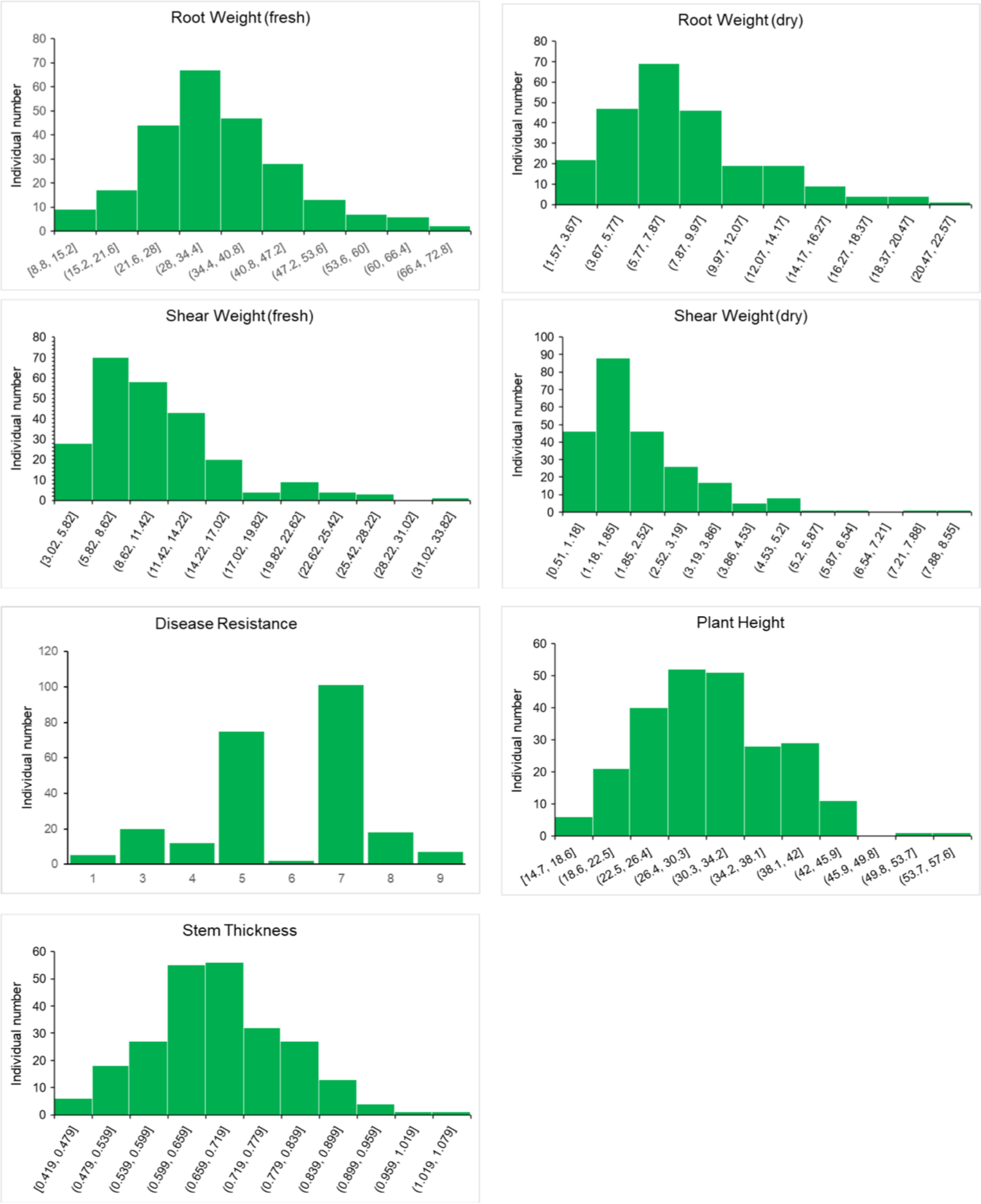

Figure S15. The statistics of sequencing data and seven phenotypic traits. Related to Figure 4.

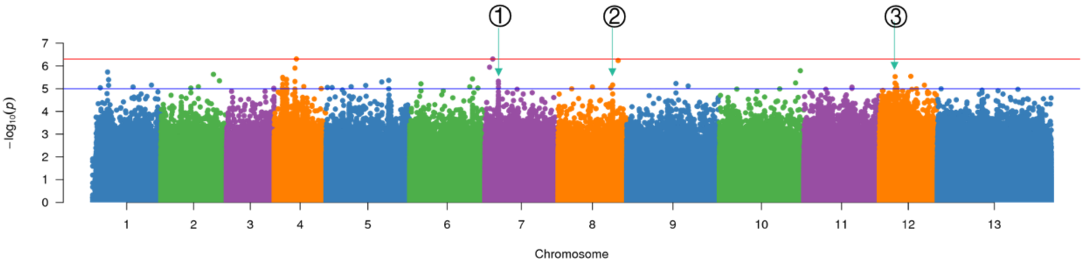

Figure S16. The Manhattan plot of the trait of plant height. Related to Figure 4.

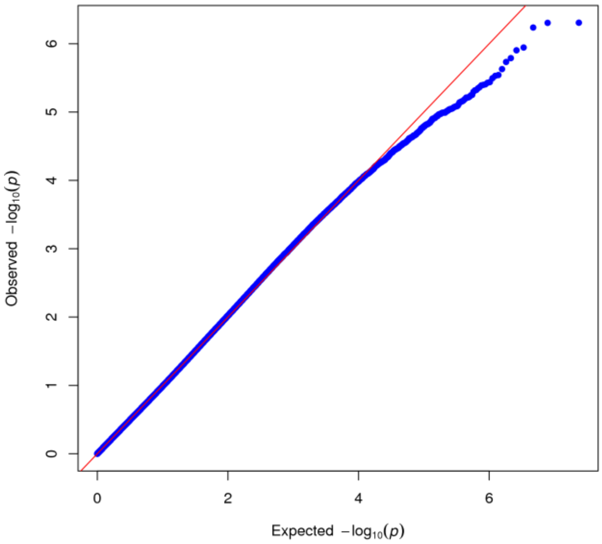

Figure S17. **The QQ plot of the trait of plant height. Related to Figure 4.**

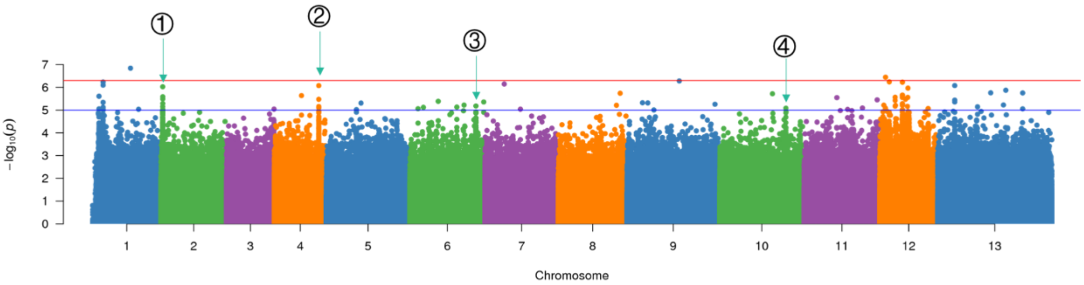

Figure S18. **The Manhattan plot of the trait of root weight (fresh). Related to Figure 4.**

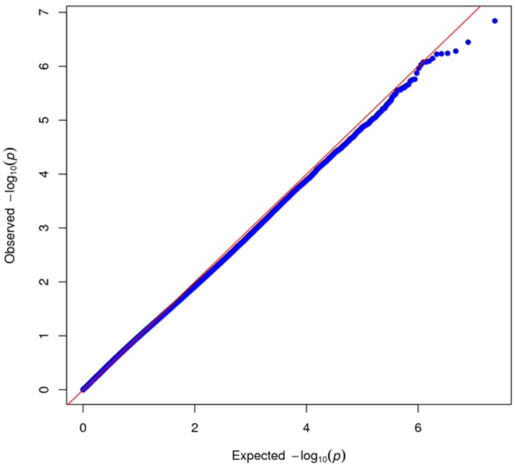

Figure S19. **The QQ plot of the trait of root weight (fresh). Related to Figure 4.**

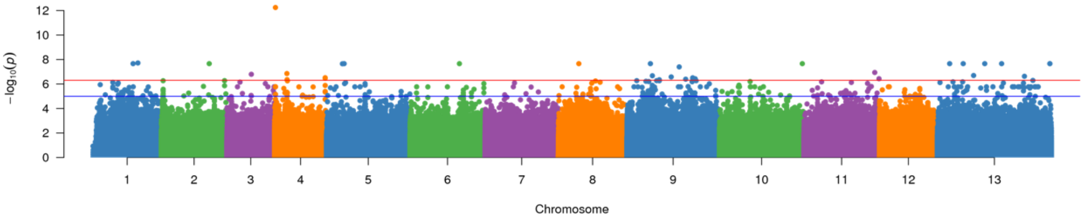

Figure S20. The Manhattan plot of the trait of shear weight (dry). Related to Figure 4.

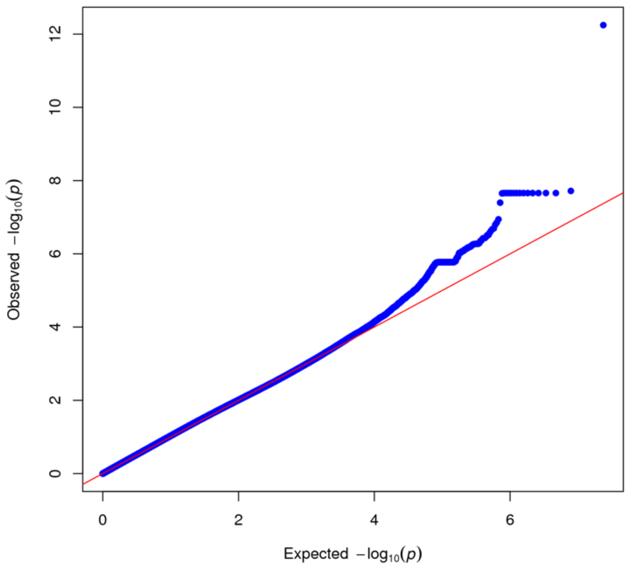

Figure S21. **The QQ plot of the trait of the shear weight (dry). Related to Figure 4.**

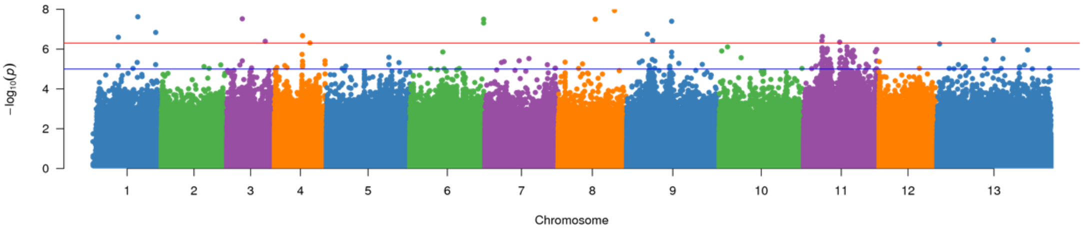

Figure S22. **The Manhattan plot of the trait of shear weight (fresh). Related to Figure 4.**

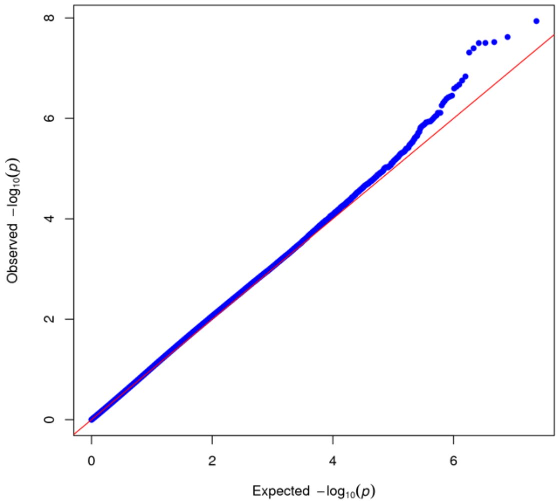

Figure S23. **The QQ plot of the trait of shear weight (fresh). Related to Figure 4.**

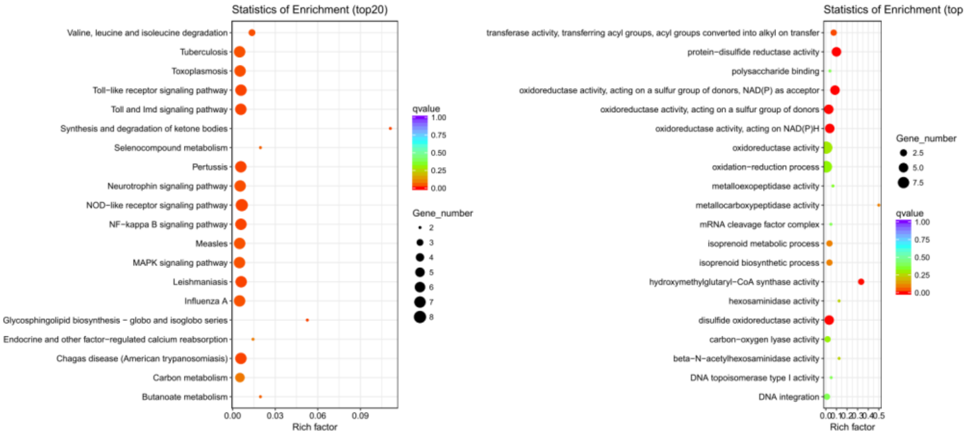

Figure S24. The KEGG and GO enrichment results of the root weight. Related to Figure 4.

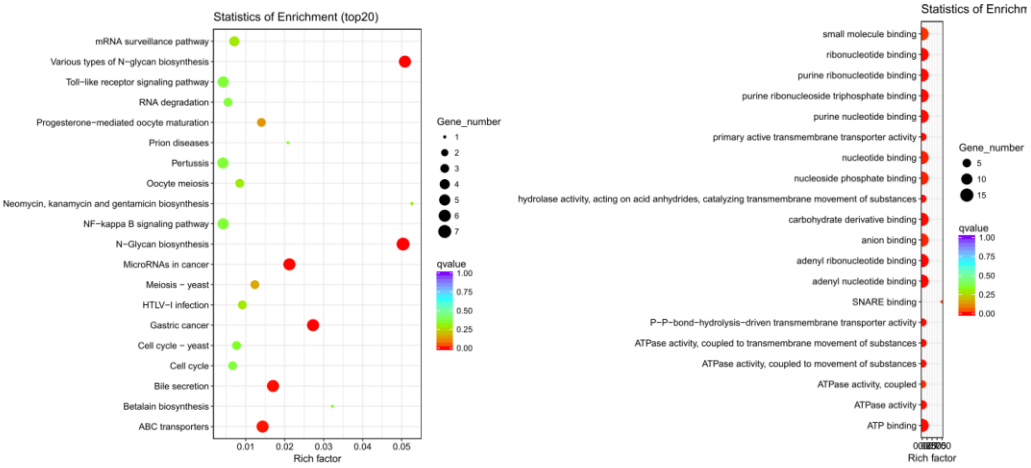

Figure S25. The KEGG and GO enrichment results of the stem thickness. Related to Figure 4.

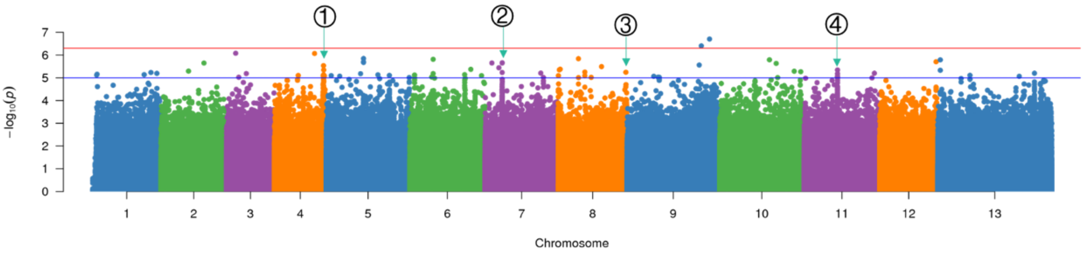

Figure S26. **The Manhattan plot of the trait of disease resistance. Related to Figure 4.**

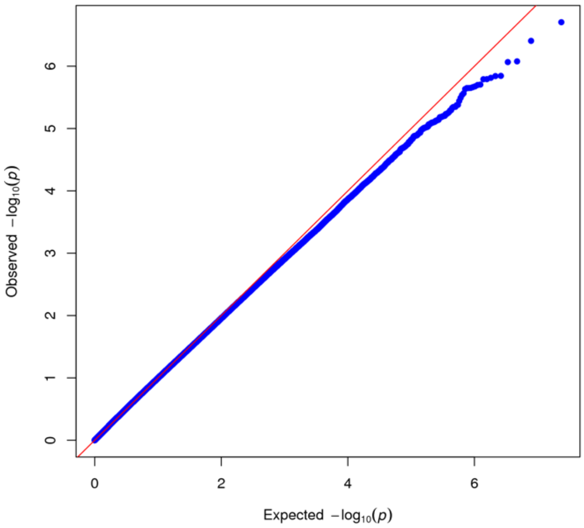

Figure S27. **The QQ plot of the trait of disease resistance. Related to Figure 4.**

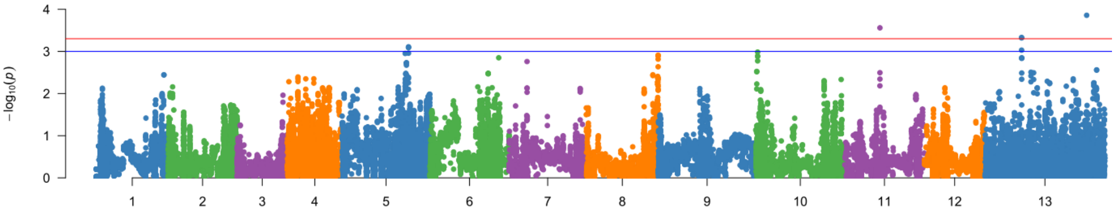

Figure S28. **The result of the gene set-based association test using fastBAT. Related to Figure 4.**

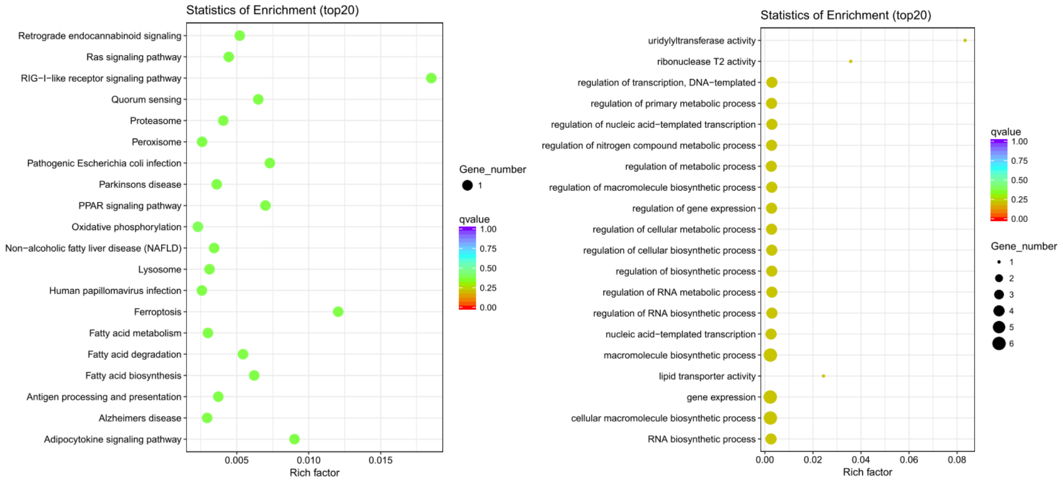

Figure S29. The KEGG and GO enrichment results of the disease resistance. Related to Figure 4.

Methods

Sequencing and genome assembly

Because of the high error rate of the long read data generated on the Nanopore and PacBio sequencing platforms, we used Canu (v1.7)(Koren et al., 2017) to correct the raw reads. The initial version of the *P. notoginseng* genome assembly was generated using the corrected raw reads and Smartdenovo (v1.0; available at https://github.com/ruanjue/smartdenovo) with the parameters '-c 1 -k 17'. We used Pilon (v1.22)(Walker et al., 2014) with the parameters '--chunksize 15000000 --diploid --changes' to refine the genome assembly using corrected long reads and MPS sequencing reads. To anchor the scaffolds of the assembly into chromosomes, we sequenced a Hi-C library(Belton et al., 2012) on the BGISEQ-500 sequencing platform. To construct the Hi-C library, leaves were cut into fragments and fixed in 1% formaldehyde (the reaction was stopped with glycine). Next, restriction enzyme *Mbo I* was added to digest the DNA, followed by 5′ overhang repair by 5U/ μl DNA Polymerase I. The Hi-C library was created by shearing 20 μg of DNA and capturing the biotin-containing fragments on streptavidin-coated beads. Following PCR, the standard circularization step required for BGISEQ-500 was carried out and DNA nanoball (DNB) prepared as previously described(Mak et al., 2017). The library was sequenced on a BGISEQ-500 sequencer with 50 bp paired-end reads. HiC-Pro(Servant et al., 2015) (v170123) was utilized for quality control (QC) of sequencing data with the partial parameter 'BOWTIE2_GLOBAL_OPTIONS = --very-sensitive -L 30 --score-min L,-0.6,-0.2 --end-to-end –reorder;BOWTIE2_LOCAL_OPTIONS = --very-sensitive -L 20 --score-min L,-0.6,-0.2 --end-to-end –reorder; IGATION_SITE = GATC; MIN_FRAG_SIZE = 100; MAX_FRAG_SIZE = 100000; MIN_INSERT_SIZE = 50; MAX_INSERT_SIZE = 1500'. We employed Juicer(Durand et al., 2016) (v1.5) and 3d-dna(Dudchenko et al., 2017) (version 170123) to obtain the contact matrices of chromatin and construct super-scaffolds (i.e., chromosomes) with the parameters '-m haploid -s 4 -c 5'.

Identification of repetitive sequences

We identified repetitive elements by integrating homology and de novo predictions. RepeatModeler (v1.0.8) (Sengupta et al., 2004) to obtain TEs predictions. Homology-based transposable elements (TEs) annotation were obtained by interrogating RepBase (v21.01) (Jurka et al., 2005) using RepeatMasker and RepeatProteinMask(Tarailo-Graovac and Chen, 2009). A non-redundant repeat annotation was obtained by combining the above data

Gene prediction and annotation

We predicted protein-encoding genes from homolog, de novo, and RNA-seq data. The results of the three methods were integrated using EVM(Haas et al., 2008) (v1.1.1), excluding genes without homolog and RNA-seq evidence. Protein sequences from closely related species (*Solanum tuberosum*, *Lactuca sativa*, *Solanum lycopersicum*, *Vitis vinifera*, and *Daucus carota*) were applied in homolog prediction by mapping them to the *P. notoginseng* genome assembly using tBLASTn(Mount, 2007) with a 1 × 10^{-5} *E*-value cut-off. For de novo prediction, BRAKER2(Hoff et al., 2016)(v2.1) was used with default parameters. RNA-data were aligned using HISAT2(Kim et al., 2015) (v2.1.0; a fast splice-aware aligner with low memory requirements), transcripts were predicted using StringTie(Pertea et al., 2015) (v1.3.4), and coding sequences (CDS) were identified using TransDecoder(Haas et al., 2013) (v 5.5.0). A final non-redundant reference gene set was generated by merging the three annotated gene sets using EVidenceModeler(Haas et al., 2008). The gene set was annotated by translating their coding sequences into proteins and interrogating the protein databases (Swiss-Prot (Bairoch and Apweiler, 2000), TrEMBL, KEGG(Kanehisa and Goto, 2000) and InterPro (Zdobnov and Apweiler, 2001)) using BLASTp (1 × 10^{-5} *E*-value cut-off) and InterProScan(Jones et al., 2014). BUSCO (Benchmarking Universal Single-Copy Orthologs) v3.0.1 (embryophyta_odb9 library) was used to evaluate the gene set and genome.

Identification of *R*-genes

Most *R*-genes in plants encode NBS-LRR proteins. According to the conservative structural characteristics of such domains, we used HMMER(Finn et al., 2011) (v3; http://hmmer.janelia.org/software) to screen the domains in the Pfam NBS (NB-ARC) family. We compared all NBS-encoding genes with the TIR HMM (PF01582) and LRR 1 HMM (PF00560) data sets using HMMER (V3). For the CC domains, we used paircoil2 (v2)(McDonnell et al., 2006) with a *P*-score cut-off of 0.025.

Gene cluster analysis

We used OrthoMCL (v1.4)(Li et al., 2003) to identify gene families. We constructed a phylogenetic tree based on the single-copy orthologous gene families using PhyML(Guindon et al., 2010). We used MCMCTREE (implemented in PAML v4.4)(Yang, 2007) to estimate the species divergence time. A ‘Correlated molecular clock’ and the ‘JC69’ model in the MCMCTREE program were used in our calculation.

Analysis of key gene families

Genes of interest in *A. thaliana* (such as *CYP450* and UGT genes) were found in the TAIR10 functional descriptions file. *P. notoginseng* were identified and classified using BLASTp with a 1 × 10^{-5} *E*-value cut-off. Gene trees were constructed using FastTree(Price et al., 2010) (v2.1.10) . The tree representation was constructed using iTOL(Letunic and Bork, 2016) (v5.5.1).

Gene expression analysis

Clean reads (see gene annotation section) were mapped to reference gene sequences using SOAP2 (Li et al., 2009), with no more than five mismatches allowed in the alignment. The gene expression level of each gene was calculated using the RPKM method (Mortazavi et al., 2008) (reads per kilobase transcriptome per million mapped reads) based on the unique alignment results. Referring to a previous study(Audic and Claverie, 1997), we used a stringent procedure to identify differentially expressed genes. The probability of a gene being expressed at equal levels in two groups was calculated based on a Poisson distribution.

Variation calling and population analysis

Low-coverage (11×) whole-genome sequencing of 240 *P. notoginseng* individuals was used to identify SNPs covering coding and regulatory regions. Sequencing reads were mapped to the reference genome using BWA (v0.7.12) (Li and Durbin, 2009). We used GATK (v4.0.6.0)(McKenna et al., 2010) to call SNPs and small indels. We re-constructed the population structure and determined the optimal number of sub-populations using Admixture (v1.3.0)(Alexander and Lange, 2011).

GWAS analysis of phenotypic traits

We recorded seven phenotypic traits of these samples subjected to whole-genome sequencing. The traits included disease resistance, the dry root weight, and the stem thickness. The phenotypic data showed an approximately normal distribution, so normalization transformation was not conducted. We filtered the SNP data using an individual-level filter: call rate ≥ 90% and site-level filter: call rate ≥ 90% and MAF ≥ 0.05. The filtered SNPs were subjected to GWAS analysis. We considered the population structure (the top 10 principal components were determined using PLINK (1.90b6.6) (Purcell et al., 2007)) and kinship (the relatedness matrix was calculated using EMMAX (beta-

07Mar2010)(Kang et al., 2010)). Genes associated with significant peaks in the Manhattan plot of these three phenotypic traits were considered genes of interest. When peaks were not obvious (e.g., associated SNPs were separated into several different chromosomes), we considered candidate genes using fastBAT(Bakshi et al., 2016), a gene set-based association test method (*P*-value cut-off of 0.05).

Supplementary References

Alexander, D.H., and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics *12*, 246.
Audic, S., and Claverie, J.M. (1997). The significance of digital gene expression profiles. Genome research *7*, 986-995.
Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic acids research *28*, 45-48.
Bakshi, A., Zhu, Z., Vinkhuyzen, A.A., Hill, W.D., McRae, A.F., Visscher, P.M., and Yang, J. (2016). Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. Sci Rep *6*, 32894.
Belton, J.M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. Methods *58*, 268-276.
Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., *et al.* (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science *356*, 92-95.
Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst *3*, 95-98.
Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res *39*, W29-37.
Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic biology *59*, 307-321.
Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., *et al.* (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc *8*, 1494-1512.
Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol *9*, R7.
Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics *32*, 767-769.
Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., *et al.* (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236-1240.
Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research *110*, 462-467.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research *28*, 27-30.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat Genet *42*, 348-354.

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

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*, 722-736.

Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res *44*, W242-245.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research *13*, 2178-2189.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics *25*, 1966-1967.

Mak, S.S.T., Gopalakrishnan, S., Caroe, C., Geng, C., Liu, S., Sinding, M.S., Kuderna, L.F.K., Zhang, W., Fu, S., Vieira, F.G., *et al.* (2017). Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. Gigascience *6*, 1-13.

McDonnell, A.V., Jiang, T., Keating, A.E., and Berger, B. (2006). Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics *22*, 356-358.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res *20*, 1297-1303.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods *5*, 621-628.

Mount, D.W. (2007). Using the basic local alignment search tool (BLAST). Cold Spring Harbor Protocols *2007*, pdb. top17.

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*, 290-295.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One *5*, e9490.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., *et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet *81*, 559-575.

Sengupta, S., Toh, S.A., Sellers, L.A., Skepper, J.N., Koolwijk, P., Leung, H.W., Yeung, H.W., Wong, R.N., Sasisekharan, R., and Fan, T.P. (2004). Modulating angiogenesis: the yin and the yang in ginseng. Circulation *110*, 1219-1225.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol *16*, 259.

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics *Chapter 4*, Unit 4 10.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., *et al.* (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One *9*, e112963.

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

Zdobnov, E.M., and Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847-848.