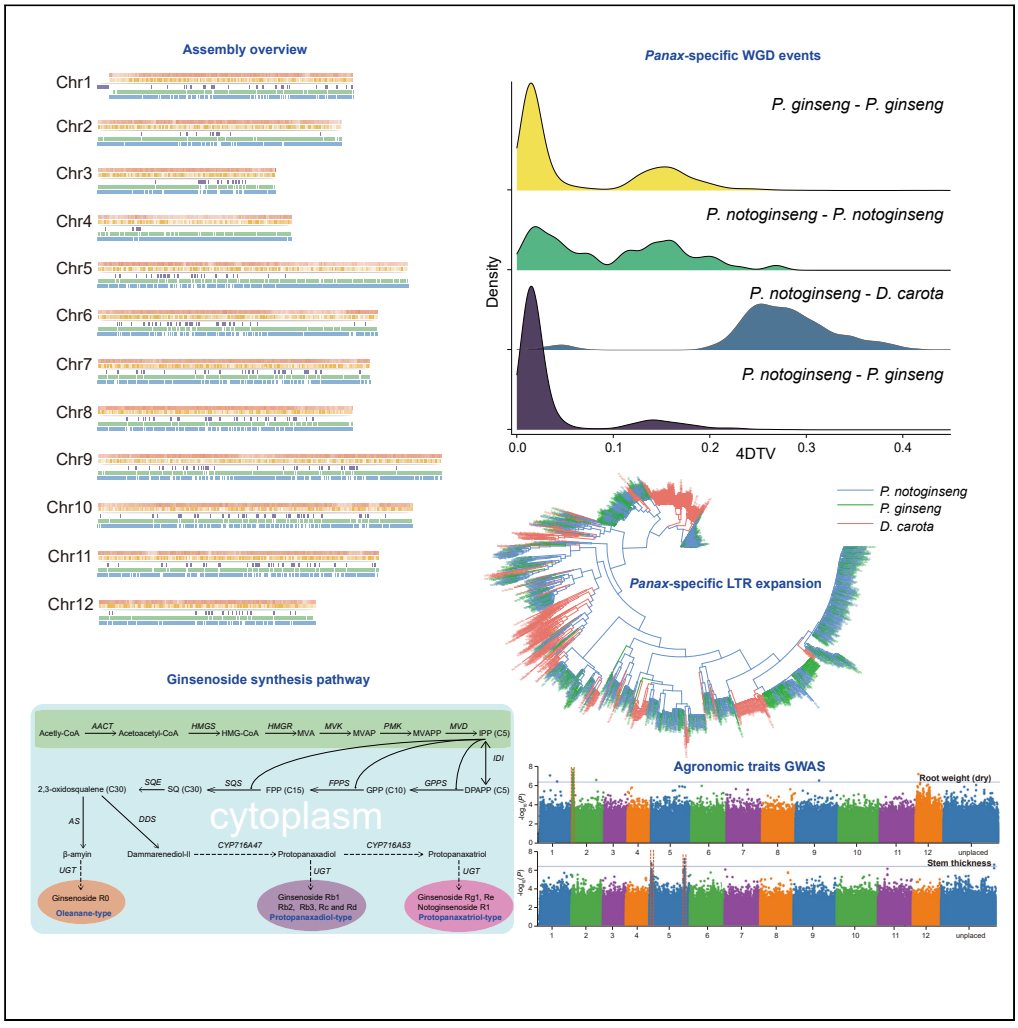
iScience

CellPress
OPEN ACCESS

Article

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



Guangyi Fan, Xiaochuan Liu, Shuai Sun, ..., Stephen Kwok-Wing Tsui, Xin Liu, Simon Ming-Yue Lee

kwtsui@cuhk.edu.hk (S.K.-W.T.)
liuxin@genomics.cn (X.L.)
simonlee@umac.mo (S.M.-Y.L.)

HIGHLIGHTS

A chromosome-level *P. notoginseng* genome assembly is provided

Panax-specific WGD events and TE expansion contribute to the larger genome size

Candidate genes involved in the ginsenoside synthesis pathway are revealed

The associated genes with agronomic traits were identified using GWAS data

Fan et al., iScience *23*, 101538
September 25, 2020
https://doi.org/10.1016/j.isci.2020.101538

iScience

CellPress
OPEN ACCESS

Article

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

Guangyi Fan,[1,2,4,9,14] Xiaochuan Liu,[2,14] Shuai Sun,[2,8,14] Chengcheng Shi,[2,14] Xiao Du,[2,14] Kai Han,[2,14] Binrui Yang,[1] Yuanyuan Fu,[2] Minghua Liu,[6] Inge Seim,[10,11] He Zhang,[2] Qiwu Xu,[2] Jiahao Wang,[2] Xiaoshan Su,[2] Libin Shao,[2] Yuanfang Zhu,[2] Yunchang Shao,[3] Yunpeng Zhao,[5] Andrew KC. Wong,[8] Dennis Zhuang,[8] Wenbin Chen,[3] Gengyun Zhang,[4] Huanming Yang,[3,12] Xun Xu,[3,13] Stephen Kwok-Wing Tsui,[6,*] Xin Liu,[2,3,4,7,15,*] and Simon Ming-Yue Lee[1,*]

SUMMARY

The Chinese ginseng *Panax notoginseng* is a domesticated herb with significant medicinal and economic value. Here we report a chromosome-level *P. notoginseng* genome assembly with a high (~79%) repetitive sequence content. The juxtaposition with the widely distributed, closely related Korean ginseng (*Panax ginseng*) genome revealed contraction of plant defense genes (in particular *R*-genes) in the *P. notoginseng* genome. We also investigated the reasons for the larger genome size of *Panax* species, revealing contributions from two *Panax*-specific whole-genome duplication events and transposable element expansion. Transcriptome data and comparative genome analysis revealed the candidate genes involved in the ginsenoside synthesis pathway. We also performed a genome-wide association study on 240 cultivated *P. notoginseng* individuals and identified the associated genes with dry root weight (63 genes) and stem thickness (168 genes). The *P. notoginseng* genome represents a critical step toward harnessing the full potential of an economically important and enigmatic plant.

INTRODUCTION

The genus *Panax* in the flowering plant family Araliaceae contains several medicinally and economically important ginseng species, including *Panax ginseng* (Korean ginseng), *Panax quinquefolius* (American ginseng), and *Panax notoginseng* (Chinese ginseng; sānqī in Chinese) (Briskin, 2000). Unlike *P. ginseng* and *P. quinquefolius*, which are widely distributed in several countries in the northern hemisphere (including the United States, Canada, China, and the Koreas), *P. notoginseng* is restricted to several mountain areas in Southern China (e.g., Wenshan Prefecture in Yunnan Province, Guo et al., 2010). *P. notoginseng* is susceptible to a wide range of pathogens, and identification of the genes conferring disease resistance is a major focus of research (Ou et al., 2011). The cultivation of *P. notoginseng* faces several challenges, including diseases that decrease production quality and yield (Baeg and So, 2013), and a potentially sparse repertoire of disease resistance genes. The major active components in ginseng genus *Panax* (Leung and Wong, 2010; Yang et al., 2014) are ginsenosides. Ginsenosides are steroid-like compounds with diverse pharmacological properties in addition to a role in plant defense. These include hepatoprotection, renoprotection, estrogen-like activities, and protection against cerebrocardiovascular ischemia and dyslipidemia (Li et al., 2009; Ng, 2006; Son et al., 2009; Xiang et al., 2011; Yang et al., 2010). For example, the Danshen dripping pill (which comprises *P. notoginseng*, *Salvia miltiorrhiza* [the Chinese herbal plant dānshēn], and synthetic borneol) is undergoing phase III clinical trials in the United States as a potential treatment for cardiovascular disease (Luo et al., 2015).

Two *P. notoginseng* genomes were published in 2017. Chen and colleagues' genome assembly (Chen et al., 2017a) is ~2.39 Gb (contig N50 of 16 kb and scaffold N50 of 96 kb), and Zhang and colleagues' assembly (Zhang et al., 2017) is ~1.85 Gb (scaffold N50 of 158 kb and contig N50 of 13.2 kb) (Figure 1C). Surprisingly, there is a notable difference in assembly size (~540 Mb) between the two versions (the estimated genome size using flow cytometry analysis is about 2.31 Gb). The Chen et al. assembly has ~75.94%

[1]State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macao 999078, China

[2]BGI-QingDao, BGI-Shenzhen, Qingdao 266555, China

[3]BGI-Shenzhen, Shenzhen 518083, China

[4]State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China

[5]The Key Laboratory of Conservation Biology for Endangered Wildlife of the Ministry of Education, College of Life Sciences, Zhejiang University, Hangzhou 310058, China

[6]School of Biomedical Sciences, The Chinese University of Hong Kong, Hong Kong, China

[7]BGI-Fuyang, BGI-Shenzhen, Fuyang 236009, China

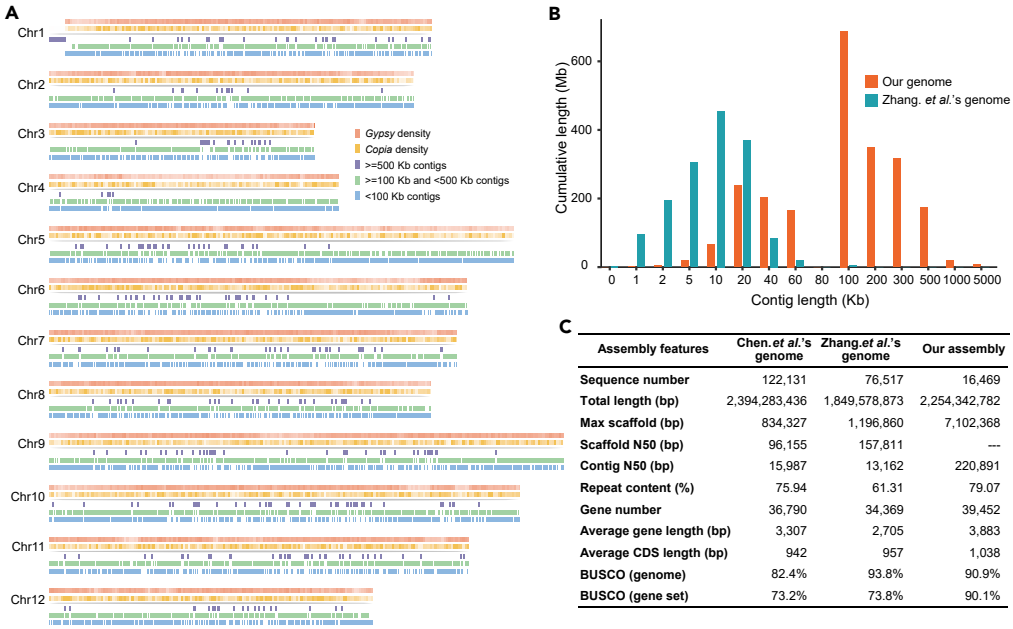
[8]System Design Engineering, University of Waterloo, Ontario, N2L 3G1 Canada

[9]Qingdao-Europe Advanced Institute for Life Sciences, BGI-Shenzhen, Qingdao 266555, China

[10]Integrative Biology Laboratory, College of Life Sciences, Nanjing Normal University, Nanjing 210046, China

[11]Comparative and Endocrine Biology Laboratory, Translational Research Institute-Institute of Health and Biomedical Innovation, School of Biomedical Sciences, Queensland University of

Continued



C

Assembly features	Chen.*et al.*'s genome	Zhang.*et al.*'s genome	Our assembly
Sequence number	122,131	76,517	16,469
Total length (bp)	2,394,283,436	1,849,578,873	2,254,342,782
Max scaffold (bp)	834,327	1,196,860	7,102,368
Scaffold N50 (bp)	96,155	157,811	---
Contig N50 (bp)	15,987	13,162	220,891
Repeat content (%)	75.94	61.31	79.07
Gene number	36,790	34,369	39,452
Average gene length (bp)	3,307	2,705	3,883
Average CDS length (bp)	942	957	1,038
BUSCO (genome)	82.4%	93.8%	90.9%
BUSCO (gene set)	73.2%	73.8%	90.1%

Figure 1. Comparison of the *P. notoginseng* Genome Assembly with Previous Assemblies

(A) The characters of each chromosome of *P. notoginseng*.

(B) Comparison of the contig length of our assembly with the published assembly.

(C) Comparison of the assembly assessment values among the three assembly versions.

repetitive sequences (~1.71 Gb), whereas the Zhang et al. assembly has 61.31% repetitive sequences (~1.13 Gb), suggesting that *P. notoginseng* has a large number of repetitive sequences. Repetitive elements are abundant in many plant species and pose a significant challenge to genome sequencing and assembly (Jiao et al., 2017). Here, we report on a third, more continuous and chromosome-level genome assembly of *P. notoginseng*.

RESULTS

Sequencing and Genome Assembly

We generated 178.2-Gb-long Nanopore reads (74.25×, with an average length of 11.49 kb) and 13.0-Gb-long PacBio reads (6×, with an average read length of 9 kb) (Table S1 and Figures S1 and S2). Using these long reads and 75.86-fold massively parallel sequencing data (Table S2), we obtained a genome assembly spanning 2.24 Gb—with a contig N50 of 220.89 kb and a 90.90% BUSCO (Seppey et al., 2019) completeness score (Tables S3 and S4 and Figures 1A and S3). This assembly represents a 10-fold increase in N50 compared with the two previous assemblies (Chen et al., 2017a; Zhang et al., 2017) (Figures 1B and 1C). Moreover, to meet the requirement of a chromosome-level reference genome for genome-wide association study (GWAS) analysis, we constructed a Hi-C library and sequenced 295.32-Gb data (123.05×) on the DNBSEQ sequencing platform (Table S5 and Figures S4 and S5). We anchored 2.0 Gb of scaffold sequences (88.89% of the total assembly) into 12 *P. notoginseng* chromosomes (Table S6), generating a final integrated *P. notoginseng* assembly. We annotated 1.78-Gb repetitive sequences spanning 79.07% of the genome (Table S7 and Figure 1C). We predicted 39,452 protein-coding genes with a BUSCO completeness score of 90.1% (Table S8 and Figure S6), an increase from the two previous assemblies (Chen assembly: 36,790 genes and 73.2% completeness; Zhang assembly: 34,369 genes and 73.8% completeness). It is also worth mentioning that the genome size and gene number of *P. notoginseng* are both notably lower than those of *P. ginseng* (genome assembly: 2.98 Gb; gene: 59,352) (Kim et al., 2018).

Genome Evolution of Ginseng Species

The phylogenetic tree showed that *P. notoginseng* and *P. ginseng* diverged approximately 4.7 mya and *Panax* species diverged with carrot (*Daucus carota*) ~64.4 mya (Figure 2A). We noticed that the genome size of *P. notoginseng* and *P. ginseng* were considerably larger than that of carrot (421.5 Mb) (Table S9).

Technology, Brisbane 4102, Australia

[12]Guangdong Provincial Academician Workstation of BGI Synthetic Genomics, BGI-Shenzhen, Shenzhen 518120, China

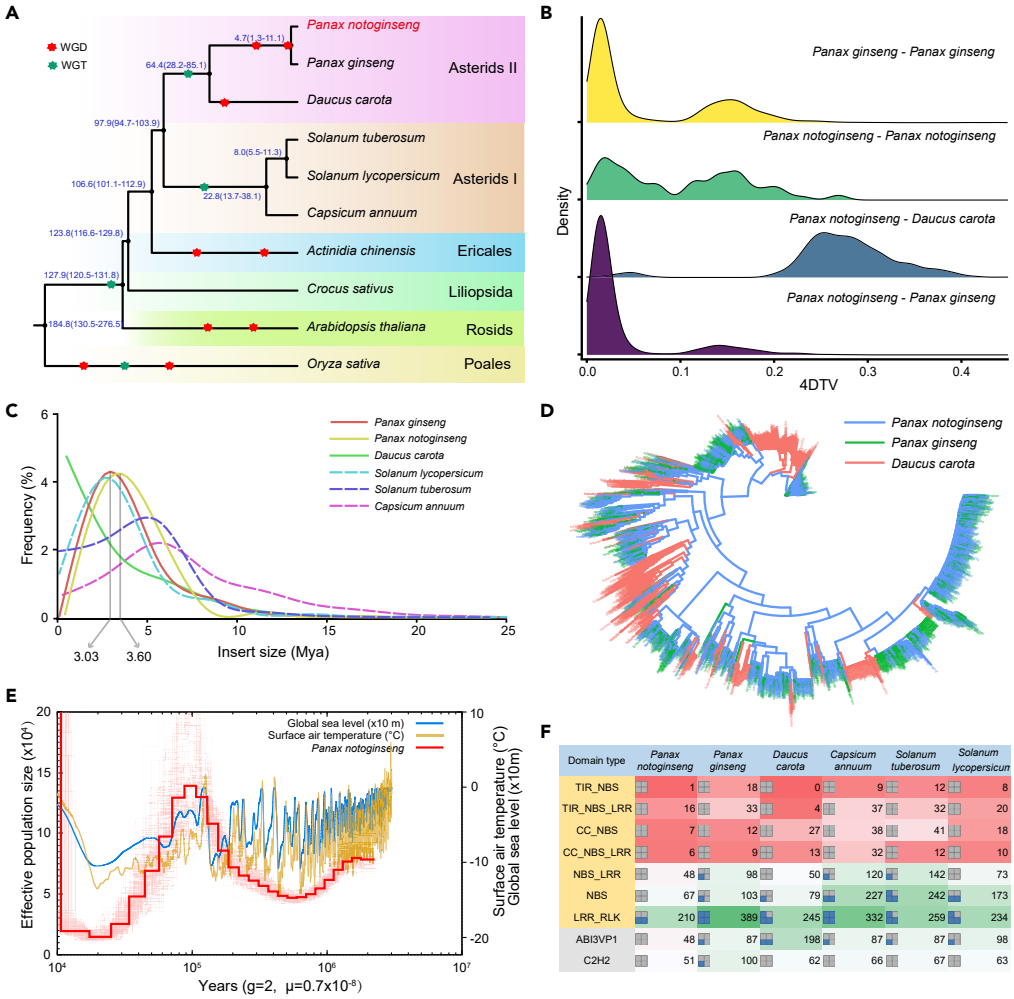
[13]Guangdong Provincial Key Laboratory of Genome Read and Write, BGI-Shenzhen, Shenzhen 518120, China

[14]These authors contributed equally

[15]Lead Contact

*Correspondence: kwtsui@cuhk.edu.hk (S.K.-W.T.), liuxin@genomics.cn (X.L.), simonlee@umac.mo (S.M.-Y.L.)

https://doi.org/10.1016/j.isci.2020.101538



Domain type	*Panax notoginseng*	*Panax ginseng*	*Daucus carota*	*Capsicum annuum*	*Solanum tuberosum*	*Solanum lycopersicum*
TIR_NBS	1	18	0	9	12	8
TIR_NBS_LRR	16	33	4	37	32	20
CC_NBS	7	12	27	38	41	18
CC_NBS_LRR	6	9	13	32	12	10
NBS_LRR	48	98	50	120	142	73
NBS	67	103	79	227	242	173
LRR_RLK	210	389	245	332	259	234
ABI3VP1	48	87	198	87	87	98
C2H2	51	100	62	66	67	63

Figure 2. Genome Evolution and Disease Resistance

(A) Phylogenetic tree and divergence time of two *Panax* species. The events of WGD and WGT were represented by red and green asterisks, respectively.

(B) Detection of whole-genome duplication events of two *Panax* species by 4-fold degenerate synonymous sites (4DTv) comparisons.

(C) Calculated LTR insertion time of two *Panax* species compared with other related species.

(D) The maximum likelihood tree constructed using LTR copia. For simplicity, 1,000 LTR sequences were randomly selected for each species. Blue, green, and red colors represent *P. notoginseng*, *P. ginseng*, and *D. carota*, respectively.

(E) Pairwise sequential Markovian coalescent (PSMC) analysis of the historical effective population size of *P. notoginseng*. Global sea level and surface air temperatures are shown in blue and yellow lines, respectively.

(F) Comparison of number of R-genes and two transcription factor genes among six species.

We next investigated the possible drivers for the genome size variation. Fourfold degenerate synonymous site (4DTv) comparisons on gene synteny blocks (Wang et al., 2012) revealed two recent *Panax*-specific whole-genome duplication (WGD) events (Figure 2B). The time of the most recent WGD event closely matches the divergence time of these two species, suggesting that it occurred before their speciation (Figure 2B). We also observed a higher proportion of transposable elements (TEs) in *P. notoginseng* (~79.07%) and *P. ginseng* (~79.52%) compared with carrot (~46.37%) (Table S9). A similar pattern was observed in their sister clades: pepper (81% of 3.48 Gb genome) versus the potato (~52.69% of 840 Mb genome) and tomato (53.81% of 950 Mb genome). We calculated the insert time of long terminal repeats (LTRs) (the major TEs of these species), revealing that *P. notoginseng* and *P. ginseng* experienced an LTR expansion ~3.03–3.6 mya (Figure 2C). In agreement, a phylogenetic tree of representative LTRs showed that *P. ginseng* and *P. notoginseng* share all LTR subtypes, indicating the *Panax*-specific LTR expansion occurred before the split of *P. ginseng* and *P. notoginseng* (Figure 2D). The LTR expansion is not observed in the carrot, suggesting that it contributes to the relatively larger genome sizes of genus *Panax*.

Reduced Plant Defense Gene Repertoire in *P. notoginseng*

Pairwise sequentially Markovian coalescent (PSMC) analysis (Li and Durbin, 2011) revealed a sharp decline in the effective population size of the mountain-dwelling *P. notoginseng* from 100,000 to 10,000 years ago consistent with the reduction of the atmospheric surface air temperature during this time (Figure 2E). Despite the relatively recent divergence time of extant ginseng species, *P. notoginseng* and *P. ginseng* have profoundly different disease resistance capabilities (Chen et al., 2017b; Lee et al., 2015; Mao et al., 2013; Ryu et al., 2014) and contemporary effective population sizes (Jang et al., 2020; Li et al., 2011b; Pan et al., 2016). Disease resistance genes (*R*-genes) serve to detect and respond to plant pathogens (Gururani et al., 2012). Most *R*-genes encode proteins with nucleotide-binding site and leucine-rich-repeat (NBS-LRR) domains. We identified 356 *R*-genes in *P. notoginseng*. This is notably fewer than that in *P. ginseng* (662), and also less than those of five other related species (418 in carrot, 796 in pepper, 741 in potato, and 537 in tomato) (Figure 2F). Moreover, we compared the *R*-gene subfamilies of the two *Panax* species and found that the NBS-LRR and NBS subfamilies have contracted in *P. notoginseng* (Figures S7–S10). In addition, transcription factor gene families involved in stress responses (Kielbowicz-Matuk, 2012; Zhuang et al., 2011) and correlated with the disease-resistant phenotype (Li et al., 2011a) have contracted in *P. notoginseng*. These include genes of the ABI3/VP1 family (48 in *P. notoginseng*, 87 in *P. ginseng*, 198 in carrot, 87 in pepper, 87 in potato, and 98 in tomato) and the C2H2-type zinc finger transcript factor family (51 in *P. notoginseng*, 100 in *P. ginseng*, 62L in carrot, 66 in pepper, 67 in potato, and 63 in tomato) (Figure 2F).

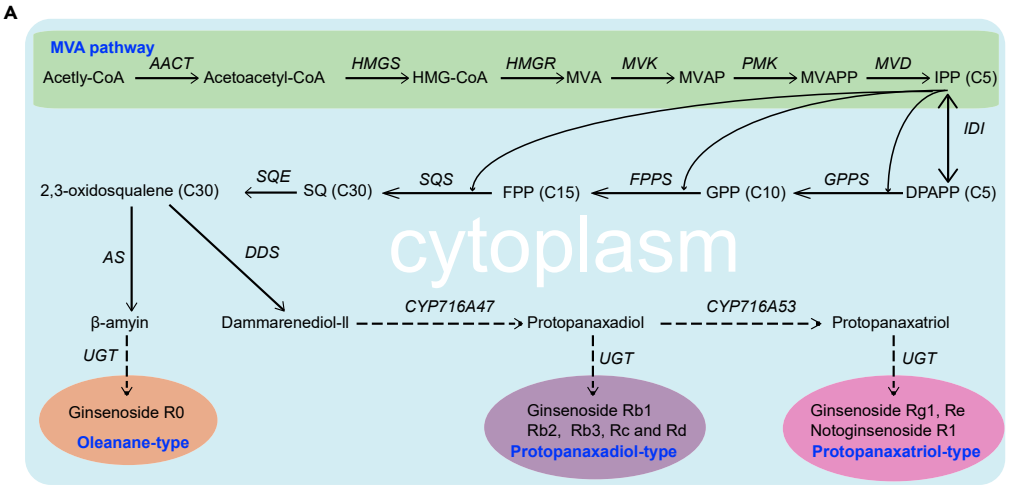
Identification of Genes in Ginsenoside Biosynthesis Pathway

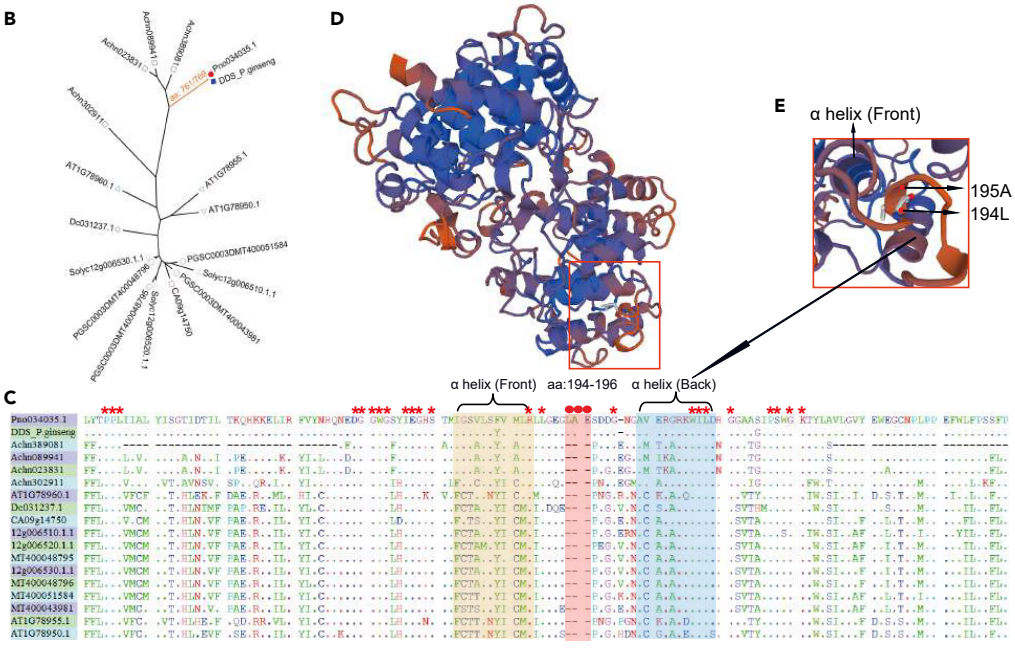
The major active ingredient of ginseng is ginsenosides (tetracyclic triterpenoid saponins). Ginsenosides are synthesized from dammarenediol-II after hydroxylation via cytochrome P450 (*CYP450*) (Shibuya et al., 2006) and glycosylation by UDP-glycosyltransferases (UGTs) (Choi et al., 2005). There are several gene families involved in ginseng biosynthesis, including dammarenediol synthase (DDS), *CYP716*, and UGTs (Figure 3A). We identified 320 and 532 CYP450 genes in 31 subfamilies in *P. notoginseng* (8 *CYP716*) and *P. ginseng* (15 *CYP716*), respectively (Figure S11). We also identified 185 and 308 UGT genes in 12 subfamilies in *P. notoginseng* (12 *UGT71* and 17 *UGT74*) and *P. ginseng* (29 *UGT71* and 37 *UGT74*), respectively (Figure S12). As expected, the gene numbers of ginsenoside biosynthesis gene families is smaller in *P. notoginseng* compared with *P. ginseng*. A notable exception is the *DDS* family, where a single copy is present in both species. DDS is highly conserved in ginseng species (98% similarity), differing by only eight amino acids (Figure 3B). Interestingly, we identified three *Panax*-specific amino acid residue insertions (L194, A195, and E196) in *P. notoginseng* and *P. ginseng* DDS (Han et al., 2006) (Figure 3C). This insertion is located in the cyclase-N domain of the protein (Figures 3D and 3E). We speculate that this variant is critical for the synthesis of ginsenosides.

P. notoginseng produces several different ginsenosides. These include protopanaxadiol-type (e.g., Rb1, Rb2, Rc, and Rd) and protopanaxatriol-type (e.g., Re, Rf, and Rg1) ginsenosides found in different tissues and developmental stages throughout the plant lifespan. Protopanaxatriol-type ginsenosides are primarily found in roots, whereas protopanaxadiol-type ginsenosides are predominant in aerial parts (leaves and flowers). To identify the candidate genes involved in the ginsenoside biosynthesis pathway, we interrogated transcriptome data from different tissues and life stages (Table S10). Four *CYP716* genes (*PN006374*, *PN008424*, *PN011429*, and *PN029913*), three *UGT71* genes (*PN006700*, *PN006701*, and *PN013865*), and two *UGT74* genes (*PN000315* and *PN008014*) were expressed at a higher level in the roots and tissues of older plants (Figure 3F) ($p < 0.01$). These genes may be involved in protopanaxatriol-type ginsenoside biosynthesis. Three *UGT71* genes (*PN000453*, *PN025151*, and *PN025152*) and three *UGT74* genes (*PN000316*, *PN024572*, and *PN033259*) showed higher expression in aerial tissues and may thus be associated with protopanaxadiol-type ginsenoside biosynthesis (Figure 3F).

P. notoginseng Population Structure and Association Mapping of Agronomic Traits

To study the population structure and identify SNP markers possibly associated with agronomic traits, we collected and sequenced 240 individuals of *P. notoginseng*. An average of ~27.02 Gb of data was obtained (~11× sequencing depth) (Table S11), providing a good foundation for variation calling. Using the SNP data, we reconstructed the *P. notoginseng* population structure, revealing four distinct sub-populations (Figures S13 and S14). This structure was supported by phylogenetic tree and principal-component analysis (Figures 4A and 4B). The population structure revealed evidence of frequent gene flow between these sub-populations, probably due to their extensive horticulture. We filtered the SNP data using individual-level





F

Gene family	Gene ID	1-year root	2-year root	3-year root	2-year flower	3-year flower	1-year leaf	2-year leaf	3-year leaf
CYP716	PN006374	601.33	754.39	826.11	10.56	10.22	2.01	3.36	4.34
CYP716	PN008424	223.97	72.75	477	23.54	16.02	0.04	0	0
CYP716	PN011429	280.08	476.47	2824.53	58.55	273.89	92.9	45.45	108.51
CYP716	PN029913	20.94	25.72	173.39	6.05	3.18	0	0	0
UGT71	PN000453	0.29	2.17	0.46	25.6	94.49	150.68	89.25	188.95
UGT71	PN025151	0.33	2.28	0.33	8.56	1.57	33.93	22.05	34.04
UGT71	PN025152	0	2.68	0	12.24	0.9	37.14	16.11	30.87
UGT71	PN006700	28.3	42.23	327.37	3.16	7.76	3.84	1.91	6.38
UGT71	PN013865	217.39	253.5	1546.16	24.5	151.73	73.73	53.93	112.18
UGT71	PN006701	18.2	25.41	184.61	0.69	5.05	2.03	1.44	2.8
UGT74	PN008014	38.74	71.28	52.58	14.76	9.19	30.49	16.97	7.31
UGT74	PN000316	0.65	1.24	2.83	0.66	0.26	10.14	21.72	31.49
UGT74	PN024572	5.43	4.76	2.37	13.16	2.96	12.48	12.46	3.57
UGT74	PN033259	0.19	0	0.38	3.27	11.31	14.3	1.94	4.1
UGT74	PN000315	244.12	359.6	1269.61	51.87	182	92.41	51.39	128.49

Figure 3. The Pathway of Ginsenoside Biosynthesis and Several Key Gene Families

(A) The ginsenoside synthesis pathway from glycolysis involves terpenoid backbone biosynthesis.

(B) Gene tree of DDS genes of two *Panax* species and other species.

(C) Protein sequencing: multiple comparisons of DDS among *P. notoginseng*, *P. ginseng*, and other representative plants. Stars represent amino acids conserved in all protein sequences; red circles represent amino acids that are present in *P. notoginseng* and *P. ginseng*, but missing in all other plants.

Figure 3. *Continued*

(D) The protein 3D structure of DDS of *P. notoginseng* constructed by the SWISS-MODEL.

(E) The partial enlarged view of the protein 3D structure of DDS, and arrows indicate two amino acids specific to *P. notoginseng* and *P. ginseng*.

(F) Differentially expressed genes (CYP716, UGT71, and UGT74) among the different tissues and living years.

filter criteria, retaining ~11.8 M SNPs for GWAS analysis. The phenotypic (trait) data were normally distributed and not skewed (Figure S15). In the GWAS analysis, we considered the population structure (top 10 principal components) and the kinship (relatedness matrix) of seven phenotypic traits (Figures S16–S23). Two phenotypic traits (dry root weight and stem thickness) had a significant signal (Figure 4C). We detected 91 loci and 63 genes located on chromosome 2 associated with root weight (Figures 4D and S24). These included three genes encoding cysteine/histidine-rich C1 domain proteins (*PN005195*, *PN005196*, and *PN010902*) closest to the SNP peak of the Manhattan plot. Such genes mediate plant growth and root development and are required for plant cell death and pathogen defense in pepper (Hwang et al., 2014). A large number (128) of genes were associated with stem thickness (Figures 4D and S25). These included the nine genes closest to the SNP peak of the Manhattan plot. Among them, *APC6* (*PN038243*) gene was the most prominent. *APC6* controls the overall number of lateral roots and root elongation in the legume *Medicago truncatula* (Bangham and McMichael, 1990) and is also involved in the amount of vascular tissue in *Arabidopsis thaliana* (Marrocco et al., 2009). *WRKY71* (*PN005405*) accelerates flowering by regulating *FLOWERING LOCUS T*, and *LEAFY*, and also has pivotal roles in shoot branching by regulating the transcription of RAX genes and auxin pathways (Guo et al., 2015; Yu et al., 2016). We speculate that *WRKY71* might be of pivotal importance in the developmental plasticity and stem thickness of *P. notoginseng*. Similarly, Reduced Wall Acetylation 3 (*RWA3*; *P. notoginseng gene PN005404*) protein has a role in plant cell wall acetylation, revealing the importance of this process for plant growth and development (Manabe et al., 2013). Finally, plant ubox 17 (*PUB17*, *PN032338*) has E3 ubiquitin ligase activity. Ubiquitination regulates plant growth and development, including flowering and responses to abiotic and biotic stresses (Sharma et al., 2016). For the disease resistance trait, we identified 33 genes. Likely because disease resistance is a complex trait, the associated SNPs were separated into four chromosomes with indistinct signals (Figures S26 and S27). We assessed the association of these candidate genes using fastBAT (Bakshi et al., 2016), a gene set-based association test method (Figure S28). The disease resistance trait was significantly enriched for the RIG-I-like receptor signaling pathway, which functions to recognize different pathogens (Mayor et al., 2007) (Figure S29). Genes driving this enrichment included LRR receptor-like serine/threonine-protein kinases (e.g., *PN033297* and *PN010029*), members of large gene families with critical role in defense (Afzal et al., 2008).

DISCUSSION

We here report a chromosome-level genome assembly of *P. notoginseng*. This high-quality genome provides insights into ginseng (genus *Panax*) evolution, including the timing of WGD events and TE expansion in *Panax*. Moreover, we identify candidate genes associated with functional diversification in *Panax*. The reduced pathogen resistance in *P. notoginseng* may be attributed to its comparatively smaller disease resistance gene repertoire. A compensatory role for *P. notoginseng* ginsenoside in pathogen defense cannot be ruled out, however. Phenolics, alkaloids, and terpenoids are three major classes of chemicals involved in plant defenses (Freeman and Beattie, 2008). Ginsenosides have antimicrobial and antifungal actions, as shown in numerous laboratory studies (Bernards et al., 2006; Nicol et al., 2002). The addition of methyl jasmonate (a signaling molecule specifically expressed by plants in response to insect and pathogenic attacks) enhances overall ginsenoside production and conversion of PPD-type ginsenosides to PPT-type ginsenosides (Palazón et al., 2003). Similarly, pepper and tomato (two species phylogenetically close to *Panax*) produce alkaloids (capsaicinoids and tomatine, respectively), which function as deterrents against pathogens (Friedman, 2002; Kim et al., 2014). We also identify several key genes in the ginsenoside synthesis pathway and show a unique 3-amino acid insertion in *Panax* DDS, an enzyme in ginsenoside biosynthesis. Transcriptome data revealed ginsenoside biosynthesis pathway genes associated with distinct tissues and life stages. Population structure analysis of 240 individuals revealed four distinct populations likely marked by agriculture-enhanced gene flow.

We identified genes associated with target traits by GWAS analysis. Recently, there have been a number of studies related to the agricultural trait for the success of plant GWAS projects, for examples, GWAS for improving grain yield, stress resilience, and quality of bread wheat (Juliana et al., 2019); GWAS for

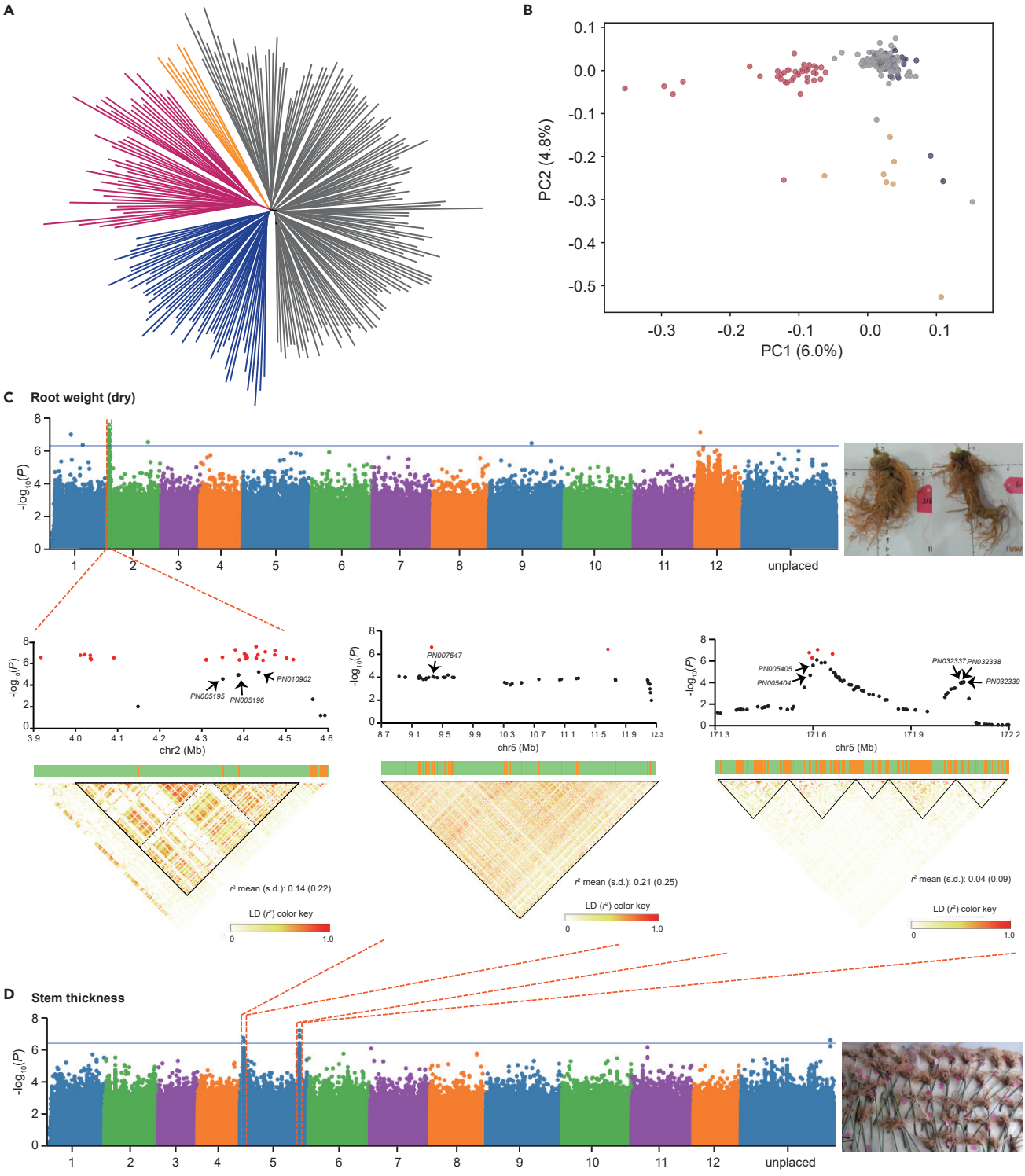


Figure 4. Population Structure and GWAS Analysis of 240 Individuals

(A) The SNP tree of 240 individuals.

(B) Principal-component analysis results of 240 individuals.

(C) The results of GWAS analysis and linkage disequilibrium blocks for the root weight based on SNP data.

(D) The results of GWAS analysis and linkage disequilibrium blocks for the stem thickness based on SNP data.

NAC42-activated nitrate transporter conferring high nitrogen use efficiency in rice (Tang et al., 2019); and GWAS of 12 agronomic traits in peach (Cao et al., 2016). Identifying molecular genetic marker and major effect QTL associated with important agricultural traits is of great interest to breeders. Root weight and stem thickness of *P. notoginseng* are two of many traits for the highly valued commercial markets. Some of these may be useful in ginseng breeding programs, for example, in concert with the development of transgenic ginseng lines using methods such as CRISPR (Schreiber et al., 2018), a method now possible with the availability of a high-quality ginseng genome. As a result, the findings in our GWAS study represent the valuable resources of *P. notoginseng*, providing new opportunities and foundation for geneticists and breeders to collectively explore the genetics underlying a wide array of agricultural traits.

Limitations of the Study

We sequenced the genome, transcriptome data, and population data and detected the candidate genes associated with different traits. The reference genome and the variations were used for the GWAS analysis mainly focused on these 12 chromosome sequences, which are about 88.98% of the whole genome sequences, suggesting that our work lacks part of GWAS results on the remaining 11.02% sequences. Besides, the candidate genes associated with agronomic traits were only detected by sequencing data with no experimental validation, and the function of these candidate SNP and gene markers should be further validated using experiment results or other technologies.

Resource Availability

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Xin Liu (liuxin@genomics.cn).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

The sequencing reads and genome assembly of *P. notoginseng* have been deposited in the CNGBdb under accession number CNP0001042, the link is http://ftp.cngb.org/pub/CNSA/data2/CNP0001042/CNS0223752/(for the whole-genome sequencing data, the accessions are from CNX0187192 to CNX0187217; for the RNA sequencing data, the accessions are from CNX0205085 to CNX0205093; for the genome assembly and annotations, the accession is CNA0013972). Meanwhile, all the sequencing data of *P. notoginseng* have been deposited in the NCBI under Bioproject number PRJNA656117 (for the whole-genome sequencing data, the accessions are from SRR11794023 to SRR11794041, for the RNA sequencing data, the accessions are from SRR12506286 to SRR12506294) and the genome assembly is available in the NCBI with the accession number JACBWS000000000.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101538.

ACKNOWLEDGMENTS

This work was funded by The Science and Technology Development Fund, Macau SAR (File nos. 0058/2019/A1 and 0016/2019/AKP) Fund, Macau SAR and the Ministry of Science and Technology of China (MOST) joint funding scheme (File. no. FDCT 017/2015/AMJ), and University of Macau (MYRG2017-00112-ICMS, MYRG2019-00105-ICMS and MYRG2016-00056-FST) and the Guangdong Provincial Academician Workstation of BGI Synthetic Genomics (No. 2017B090904014) and the Guangdong Provincial Key Laboratory of Genome Read and Write (No. 2017B030301011). The data that support the findings of this study have been deposited in the CNSA (https://db.cngb.org/cnsa/) of CNGBdb.

AUTHOR'S CONTRIBUTIONS

S.M.-Y.L., X.X., H.Y., and X.L. designed the project. B.Y., X.S., and G.Z. prepared the samples and conducted the experiments. X.L., Y.F., and J.W. performed the genome assembly and annotation. C.S., Y.Z., A.K.C.W., D.Z., and H.Z. performed genome evolution analysis. S.S., X.D., K.H., and L.S. performed the population analysis. G.F., Y.Z., S.M.-Y.L., I.S., and S. K-W.T wrote and revised the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 20, 2020
Revised: August 11, 2020
Accepted: September 3, 2020
Published: September 25, 2020

REFERENCES

Afzal, A.J., Wood, A.J., and Lightfoot, D.A. (2008). Plant receptor-like serine threonine kinases: roles in signaling and plant defense. Mol. Plant Microbe Interact. *21*, 507–517.

Baeg, I.H., and So, S.H. (2013). The world ginseng market and the ginseng (Korea). J. Ginseng Res. *37*, 1–7.

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.

Bangham, C.R., and McMichael, A.J. (1990). HIV infection. Why the long latent period? Nature *348*, 388.

Bernards, M.A., Yoesef, L.F., and Nicol, R.W. (2006). The Allelopathic Potential of Ginsenosides (Netherlands: Springer).

Briskin, D.P. (2000). Medicinal plants and phytomedicines. Linking plant biochemistry and physiology to human health. Plant Physiol. *124*, 507–514.

Cao, K., Zhou, Z., Wang, Q., Guo, J., Zhao, P., Zhu, G., Fang, W., Chen, C., Wang, X., and Wang, X. (2016). Genome-wide association study of 12 agronomic traits in peach. Nat. Commun. *7*, 1–10.

Chen, W., Kui, L., Zhang, G., Zhu, S., Zhang, J., Wang, X., Yang, M., Huang, H., Liu, Y., Wang, Y., et al. (2017a). Whole-genome sequencing and analysis of the Chinese herbal plant Panax notoginseng. Mol. Plant *10*, 899–902.

Chen, Z.J., Ma, X.H., Dong, L.L., Zhang, L.J., Wei, G.F., Xiao, L.N., Wang, Y., Wei, F.G., Liu, W.L., Yu, Y.Q., et al. (2017b). DNA marker-assisted selection of medicinal plants (III)Evaluation of disease resistance of "Miaoxiang Kangqi 1" –a new cultivar of Panax notoginseng. Zhongguo Zhong Yao Za Zhi *42*, 2046–2051.

Choi, D.W., Jung, J., Ha, Y.I., Park, H.W., Dong, S.I., Chung, H.J., and Liu, J.R. (2005). Analysis of transcripts in methyl jasmonate-treated ginseng hairy roots to identify genes involved in the biosynthesis of ginsenosides and other secondary metabolites. Plant Cell Rep. *23*, 557–566.

Freeman, B.C., and Beattie, G.A. (2008). An Overview of Plant Defenses against Pathogens and Herbivores (The Plant Health Instructor).

Friedman, M. (2002). Tomato glycoalkaloids: role in the plant and in the diet. J. Agric. Food Chem. *50*, 5751–5780.

Guo, D., Zhang, J., Wang, X., Han, X., Wei, B., Wang, J., Li, B., Yu, H., Huang, Q., Gu, H., et al. (2015). The WRKY transcription factor WRKY71/EXB1 controls shoot branching by transcriptionally regulating RAX genes in Arabidopsis. Plant Cell *27*, 3112–3127.

Guo, H., Cui, X., An, N., and Cai, G. (2010). Sanchi ginseng (Panax notoginseng (Burkill) F. H. Chen) in China: distribution, cultivation and variations. Genet. Resour. Crop Evol. *57*, 453–460.

Gururani, M.A., Venkatesh, J., Upadhyaya, C.P., Nookaraju, A., Pandey, S.K., and Park, S.W. (2012). Plant disease resistance genes: current status and future directions. Physiol. Mol. Plant Pathol. *78*, 51–65.

Han, J.Y., Kwon, Y.S., Yang, D.C., Jung, Y.R., and Choi, Y.E. (2006). Expression and RNA interference-induced silencing of the dammarenediol synthase gene in Panax ginseng. Plant Cell Physiol *47*, 1653–1662.

Hwang, I.S., Choi, D.S., Kim, N.H., Kim, D.S., and Hwang, B.K. (2014). The pepper cysteine/histidine-rich DC1 domain protein CaDC1 binds both RNA and DNA and is required for plant cell death and defense response. New Phytol. *201*, 518–530.

Jang, W., Jang, Y., Kim, N.-H., Waminal, N.E., Kim, Y.C., Lee, J.W., and Yang, T.-J. (2020). Genetic diversity among cultivated and wild Panax ginseng populations revealed by high-resolution microsatellite markers. J. Ginseng Res. *44*, 637–643.

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X., Chin, C.S., et al. (2017). Improved maize reference genome with single-molecule technologies. Nature *546*, 524–527.

Juliana, P., Poland, J., Huerta-Espino, J., Shrestha, S., Crossa, J., Crespo-Herrera, L., Toledo, F.H., Govindan, V., Mondal, S., and Kumar, U. (2019). Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. Nat. Genet. *51*, 1530–1539.

Kielbowicz-Matuk, A. (2012). Involvement of plant C(2)H(2)-type zinc finger transcription factors in stress responses. Plant Sci. *185-186*, 78–85.

Kim, N.H., Jayakodi, M., Lee, S.C., Choi, B.S., Jang, W., Lee, J., Kim, H.H., Waminal, N.E., Lakshmanan, M., van Nguyen, B., et al. (2018). Genome and evolution of the shade-requiring medicinal herb Panax ginseng. Plant Biotechnol. J. *16*, 1904–1917.

Kim, S., Park, M., Yeom, S.I., Kim, Y.M., Lee, J.M., Lee, H.A., Seo, E., Choi, J., Cheong, K., Kim, K.T., et al. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. Nat. Genet. *46*, 270–278.

Lee, B.D., Dutta, S., Ryu, H., Yoo, S.J., Suh, D.S., and Park, K. (2015). Induction of systemic resistance in Panax ginseng against Phytophthora cactorum by native Bacillus amyloliquefaciens HK34. J. Ginseng Res. *39*, 213–220.

Leung, K.W., and Wong, A.S. (2010). Pharmacology of ginsenosides: a literature review. Chin. Med. *5*, 20.

Li, C.W., Su, R.C., Cheng, C.P., Sanjaya, You, S.J., Hsieh, T.H., Chao, T.C., and Chan, M.T. (2011a). Tomato RAV transcription factor is a pivotal modulator involved in the AP2/EREBP-mediated defense pathway. Plant Physiol. *156*, 213–227.

Li, H., Deng, C.Q., Chen, B.Y., Zhang, S.P., Liang, Y., and Luo, X.G. (2009). Total saponins of Panax notoginseng modulate the expression of caspases and attenuate apoptosis in rats following focal cerebral ischemia-reperfusion. J. ethnopharmacol. *121*, 412–418.

Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature *475*, 493–496.

Li, S., Li, J., Yang, X.-L., Cheng, Z., and Zhang, W.-J. (2011b). Genetic diversity and differentiation of cultivated ginseng (Panax ginseng CA Meyer) populations in North-east China revealed by inter-simple sequence repeat (ISSR) markers. Genet. Resour. Crop Evol. *58*, 815–824.

Luo, J., Song, W., Yang, G., Xu, H., and Chen, K. (2015). Compound Danshen (Salvia miltiorrhiza) dripping pill for coronary heart disease: an overview of systematic reviews. Am. J. Chin. Med. *43*, 25–43.

Manabe, Y., Verhertbruggen, Y., Gille, S., Harholt, J., Chong, S.L., Pawar, P.M., Mellerowicz, E.J., Tenkanen, M., Cheng, K., Pauly, M., et al. (2013). Reduced Wall Acetylation proteins play vital and distinct roles in cell wall O-acetylation in Arabidopsis. Plant Physiol. *163*, 1107–1117.

Mao, Z.S., Long, Y.J., Zhu, S.S., Chen, Z.J., Wei, F.G., and Zhu, Y.Y. (2013). Advances in root rot pathogen of Panax notoginseng research. J. Chin. Med. Mater. *36*, 2051–2054.

Marrocco, K., Thomann, A., Parmentier, Y., Genschik, P., and Criqui, M.C. (2009). The APC/C E3 ligase remains active in most post-mitotic Arabidopsis cells and is required for proper vasculature development and organization. Development *136*, 1475–1485.

Mayor, A., Martinon, F., De Smedt, T., Petrilli, V., and Tschopp, J. (2007). A crucial function of SGT1 and HSP90 in inflammasome activity links mammalian and plant innate immune responses. Nat. Immunol. *8*, 497–503.

Ng, T.B. (2006). Pharmacological activity of sanchi ginseng (Panax notoginseng). J. Pharm. Pharmacol. *58*, 1007–1019.

Nicol, R.W., Traquair, J.A., and MA, B. (2002). Ginsenosides as host resistance factors in American ginseng (Panax quinquefolius). Can. J. Bot. *80*, 557–562.

Ou, X., Jin, H., Guo, L., Yang, Y., Cui, X., Xiao, Y., and Liu, D. (2011). [Status and prospective on nutritional physiology and fertilization of Panax notoginseng]. Zhongguo Zhong yao Za Zhi *36*, 2620–2624.

Palazón, J., Cusidó, R.M., Bonfill, M., Mallol, A., Moyano, E., Morales, C., and Piñol, M.T. (2003). Elicitation of different Panax ginseng transformed root phenotypes for an improved ginsenoside production. Plant Physiol. Biochem. *41*, 1019–1025.

Pan, Y., Wang, X., Sun, G., Li, F., and Gong, X. (2016). Application of RAD sequencing for evaluating the genetic diversity of domesticated Panax notoginseng (Araliaceae). PLoS One *11*, e0166419.

Ryu, H., Park, H., Suh, D.S., Jung, G.H., Park, K., and Lee, B.D. (2014). Biological control of Colletotrichum panacicola on Panax ginseng by Bacillus subtilis HK-CSM-1. J. Ginseng Res. *38*, 215–219.

Schreiber, M., Stein, N., and Mascher, M. (2018). Genomic approaches for studying crop evolution. Genome Biol. *19*, 140.

Seppey, M., Manni, M., and Zdobnov, E.M. (2019). Busco: assessing genome assembly and annotation completeness. Methods Mol. Biol. *1962*, 227–245.

Sharma, B., Joshi, D., Yadav, P.K., Gupta, A.K., and Bhatt, T.K. (2016). Role of ubiquitin-mediated degradation system in plant biology. Front. Plant Sci. *7*, 806.

Shibuya, M., Hoshino, M., Katsube, Y., Hayashi, H., Kushiro, T., and Ebizuka, Y. (2006). Identification of beta-amyrin and sophoradiol 24-hydroxylase by expressed sequence tag mining and functional expression assay. FEBS J. *273*, 948–959.

Son, H.Y., Han, H.S., Jung, H.W., and Park, Y.K. (2009). Panax notoginseng attenuates the infarct volume in rat ischemic brain and the inflammatory response of microglia. J. Pharmacol. Sci. *109*, 368–379.

Tang, W., Ye, J., Yao, X., Zhao, P., Xuan, W., Tian, Y., Zhang, Y., Xu, S., An, H., and Chen, G. (2019). Genome-wide associated study identifies NAC42-activated nitrate transporter conferring high nitrogen use efficiency in rice. Nat. Commun. *10*, 1–11.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. *40*, e49.

Xiang, H., Liu, Y., Zhang, B., Huang, J., Li, Y., Yang, B., Huang, Z., Xiang, F., and Zhang, H. (2011). The antidepressant effects and mechanism of action of total saponins from the caudexes and leaves of Panax notoginseng in animal models of depression. Phytomedicine *18*, 731–738.

Yang, C.Y., Wang, J., Zhao, Y., Shen, L., Jiang, X., Xie, Z.G., Liang, N., Zhang, L., and Chen, Z.H. (2010). Anti-diabetic effects of Panax notoginseng saponins and its major anti-hyperglycemic components. J. ethnopharmacol. *130*, 231–236.

Yang, W.Z., Hu, Y., Wu, W.Y., Ye, M., and Guo, D.A. (2014). Saponins in the genus *Panax L.* (Araliaceae): a systematic review of their chemical diversity. Phytochemistry *106*, 7–24.

Yu, Y., Liu, Z., Wang, L., Kim, S.G., Seo, P.J., Qiao, M., Wang, N., Li, S., Cao, X., Park, C.M., et al. (2016). WRKY71 accelerates flowering via the direct activation of flowering locus t and leafy in Arabidopsis thaliana. Plant J. *85*, 96–106.

Zhang, D., Li, W., Xia, E.H., Zhang, Q.J., Liu, Y., Zhang, Y., Tong, Y., Zhao, Y., Niu, Y.C., Xu, J.H., et al. (2017). The medicinal herb Panax notoginseng genome provides insights into ginsenoside biosynthesis and genome evolution. Mol. Plant *10*, 903–907.

Zhuang, J., Sun, C.C., Zhou, X.R., Xiong, A.S., and Zhang, J. (2011). Isolation and characterization of an AP2/ERF-RAV transcription factor BnaRAV-1-HY15 in Brassica napus L. HuYou15. Mol. Biol. Rep. *38*, 3921–3928.

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

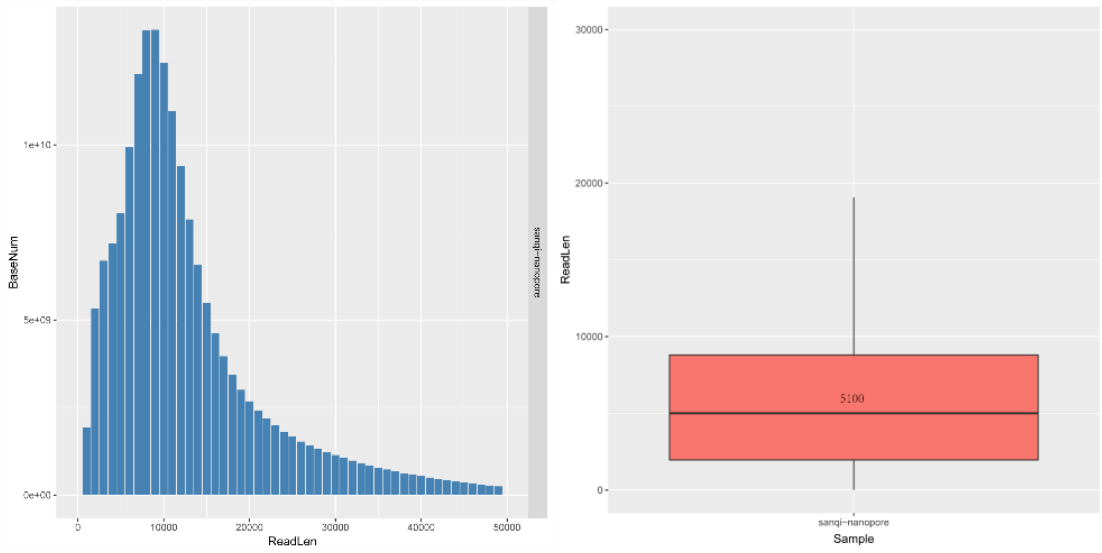


Figure S1. Summary of raw long read length. Related to Table S1.

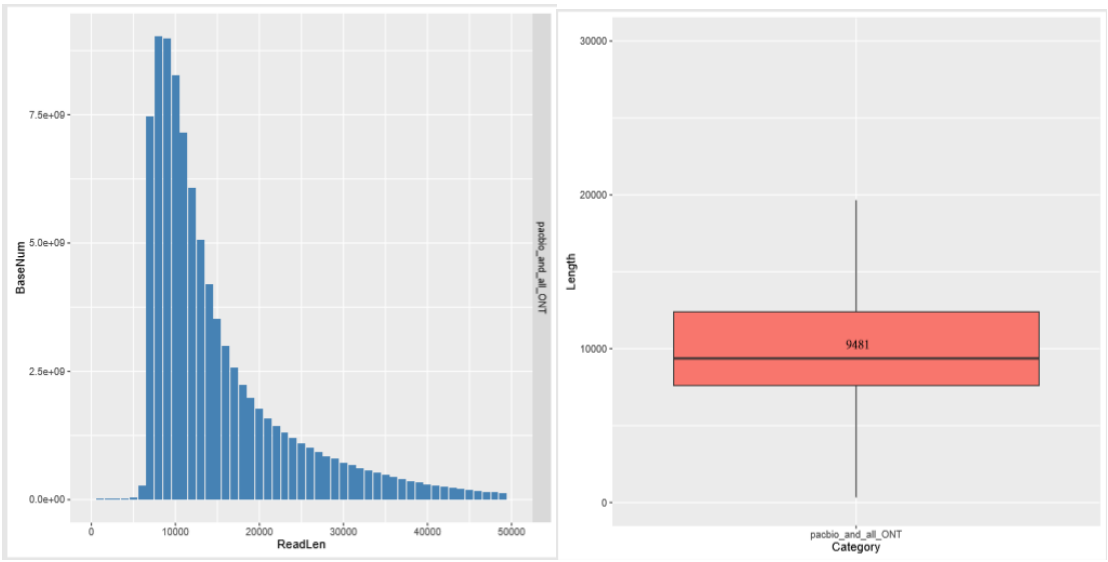


Figure S2. Summary of the length of the corrected long reads. Related to Table S1.

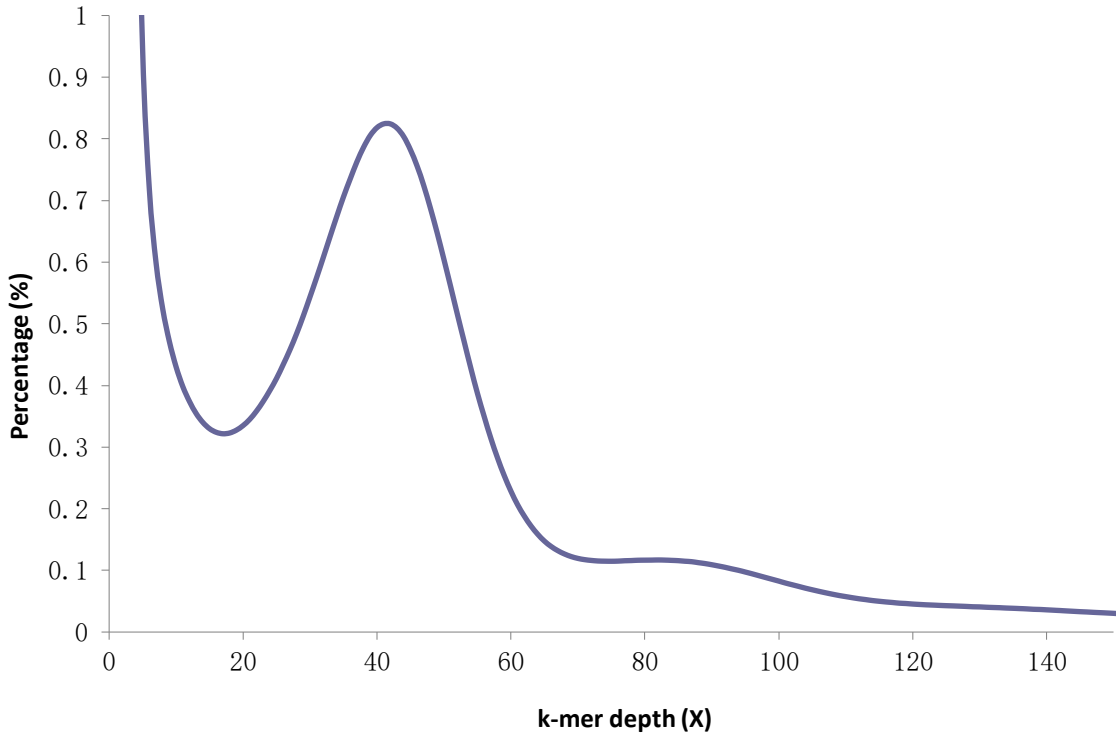


Figure S3. The 17-mer depth distribution of *P. notoginseng*. Related to Table S3

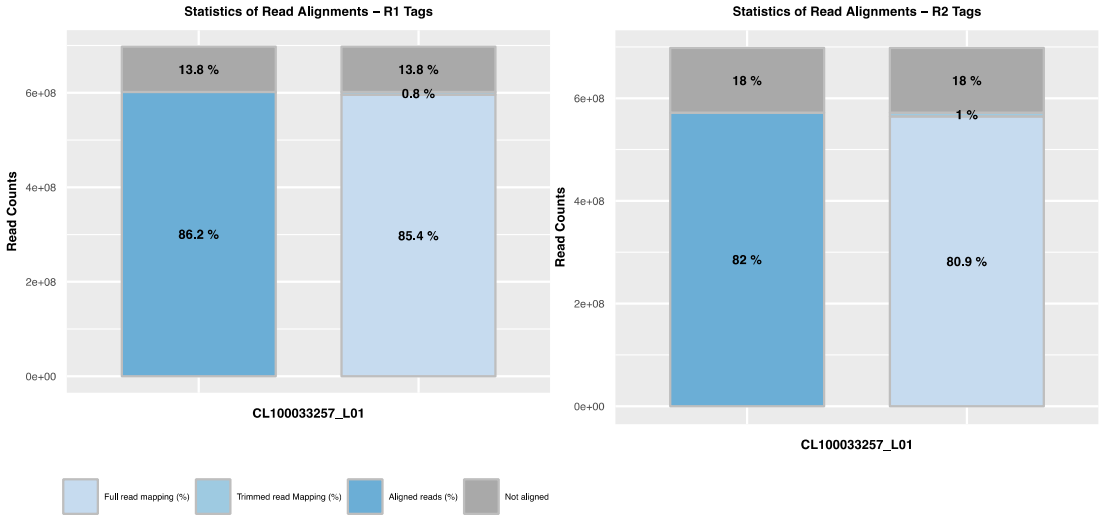


Figure S4. Quality control of Hi-C read. Statistics for the type of separated pair-end read alignment. The aligned read ratio shown in the left bar including full-read and trimmed read mapping. Related to Table S5.

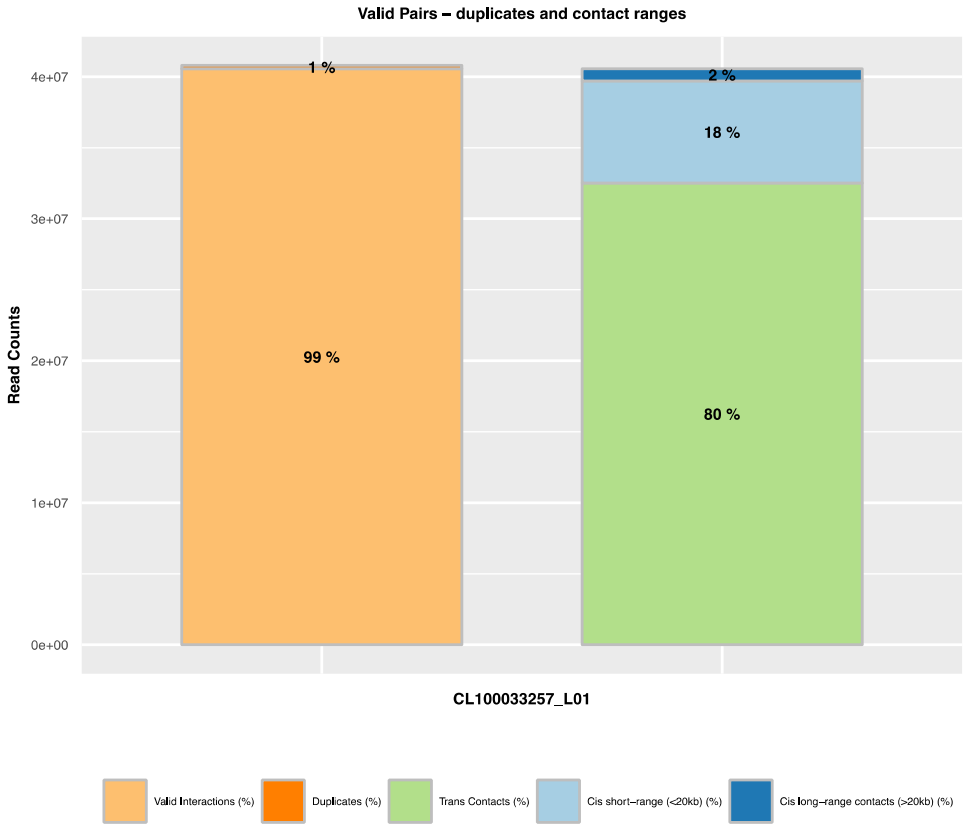


Figure S5. Quality control of Hi-C read. The left bar shows the ratio of duplication for the valid read pairs. For all the non-duplicated reads, the percentage of cis and trans contacts are shown (right bar). Related to Table S5

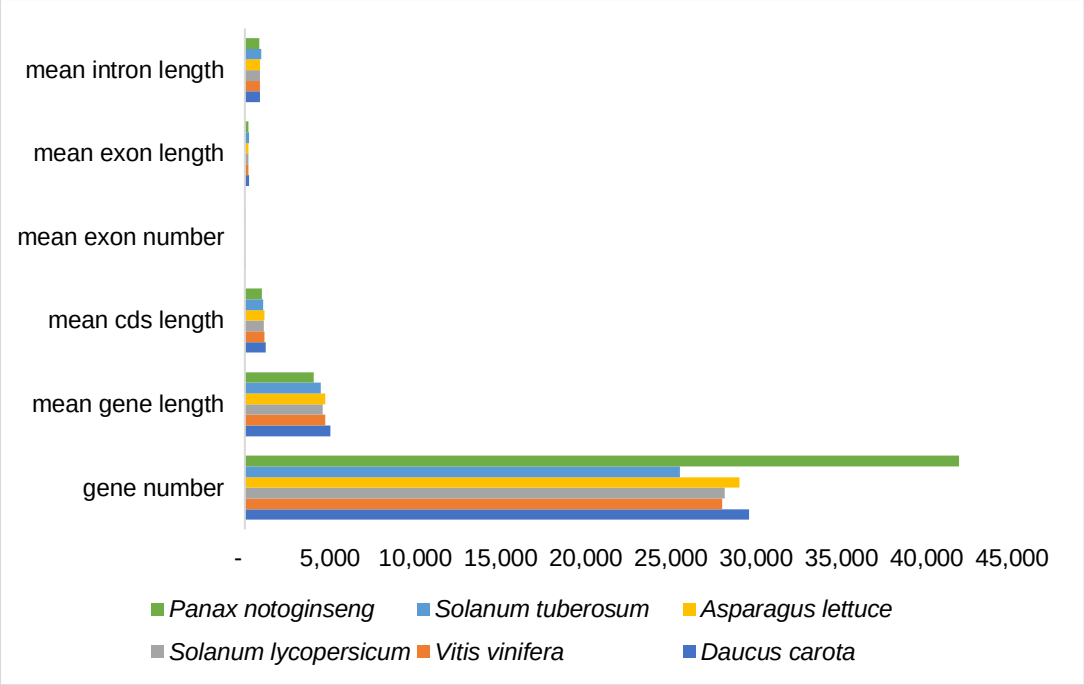


Figure S6. Comparison of the gene structures among *P. notogiseng* and other five species. Related to Table S8.

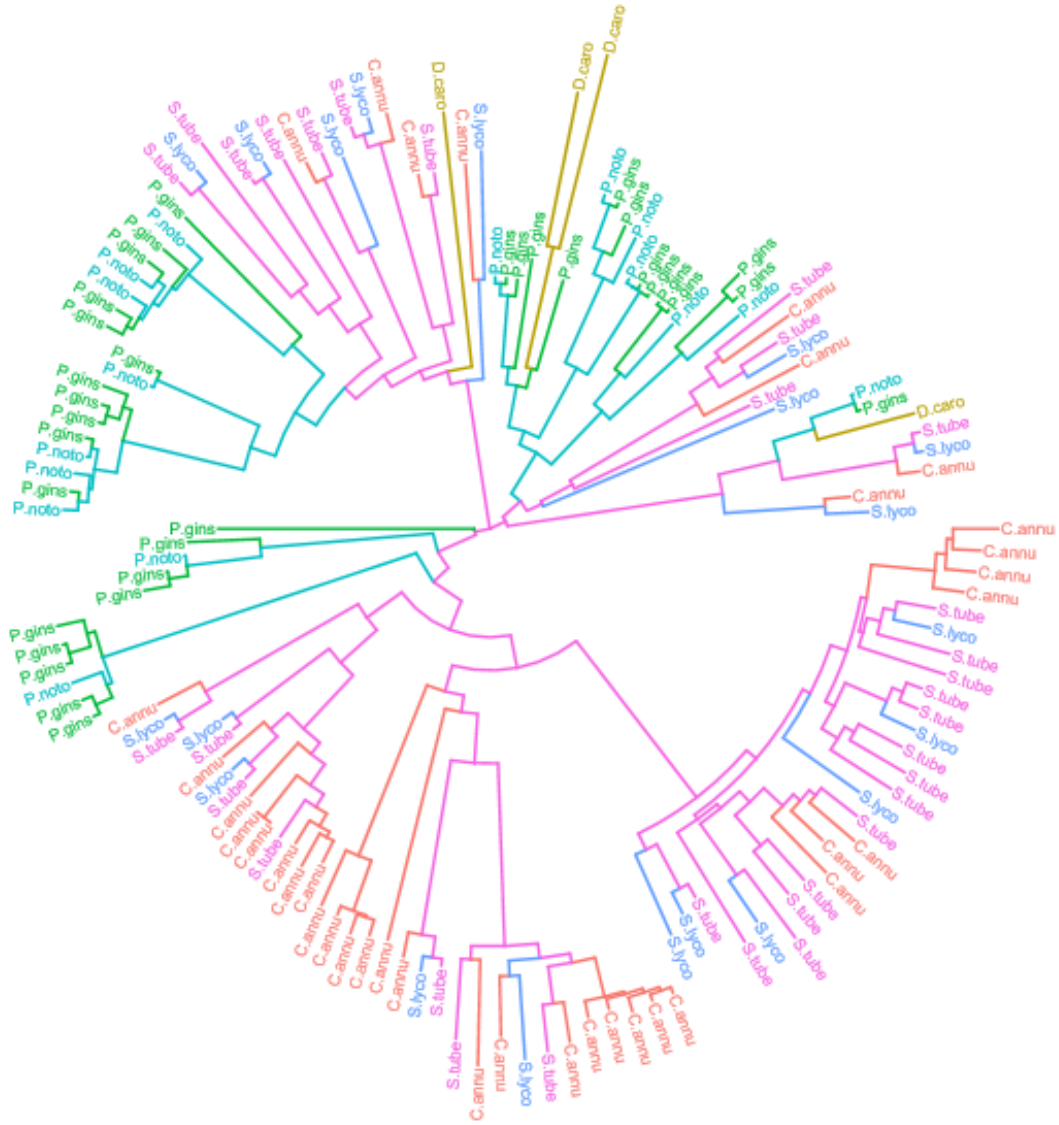


Figure S7. **Comparison of the TIR_NBS_LRR R-genes,** P.noto is the genes of *P. notoginseng*, P.gins is the genes of *P. ginseng*, S.tube is the genes of *Solanum tuberosum*, C.annu is the genes of *Capsicum annuum*, S.lyco is the genes of *Solanum lycopersicum*, D.caro is the genes of *Daucus carota*.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

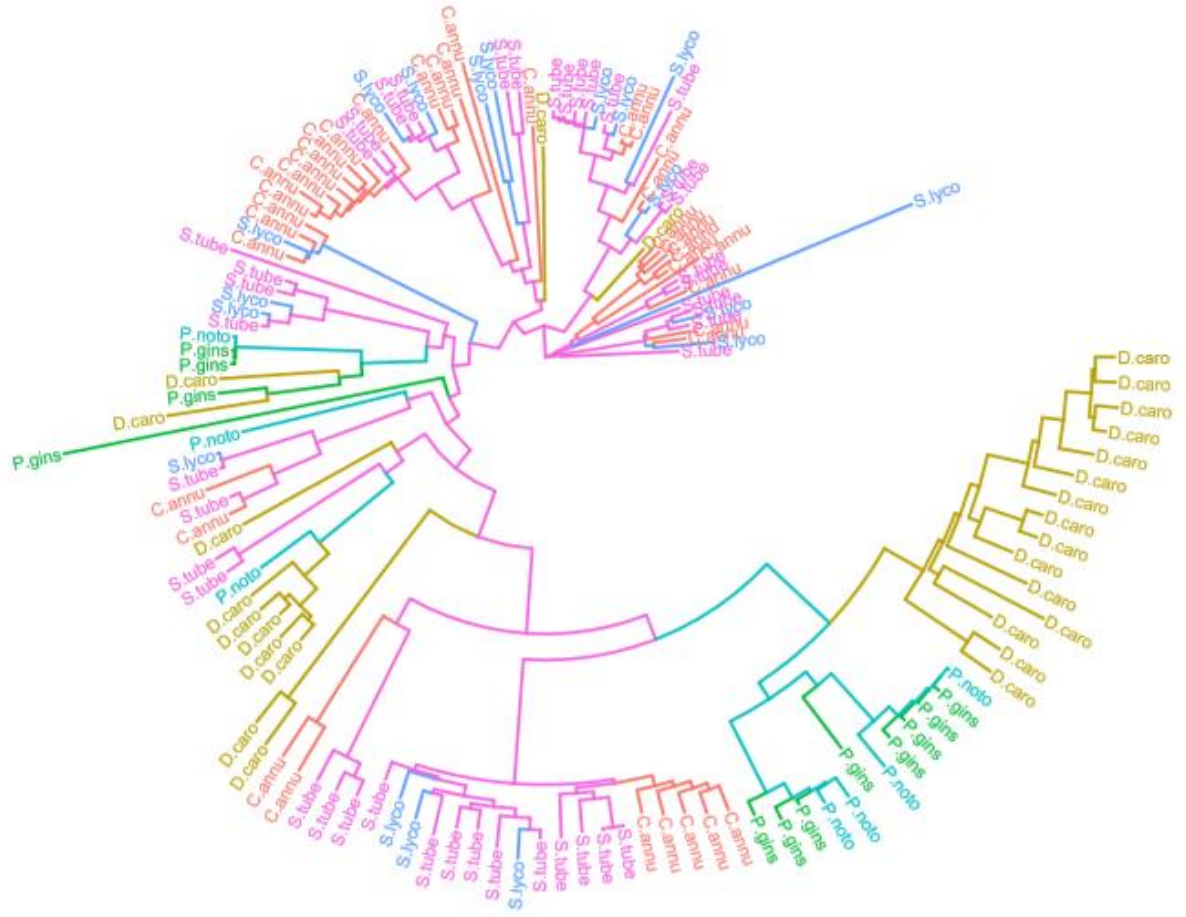


Figure S8. **Comparison of the CC_NBS R-genes**, P.noto is the genes of *P. notoginseng*, P.gins is the genes of *P. ginseng*, S.tube is the genes of *Solanum tuberosum*, C.annu is the genes of *Capsicum annuum*, S.lyco is the genes of *Solanum lycopersicum*, D.caro is the genes of *Daucus carota*. Related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

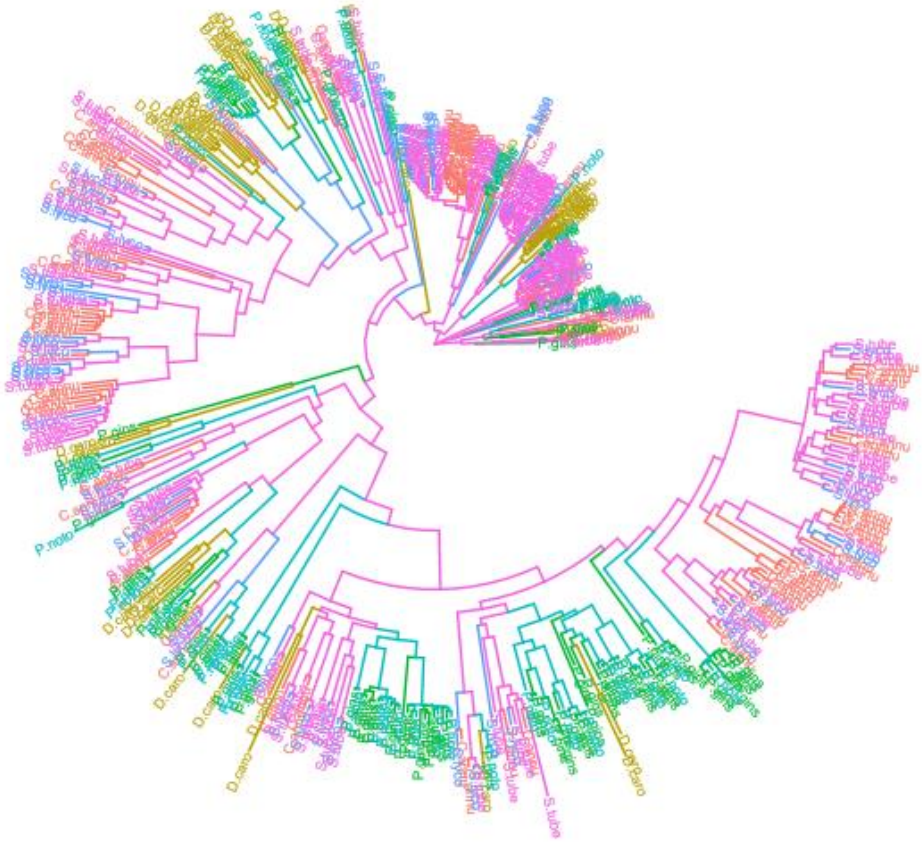


Figure S9. **Comparison of the NBS_LRR R-genes**, P.noto is the genes of *P. notoginseng*, P.gins is the genes of *P. ginseng*, S.tube is the genes of *Solanum tuberosum*, C.annu is the genes of *Capsicum annuum*, S.lyco is the genes of *Solanum lycopersicum*, D.caro is the genes of *Daucus carota*.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

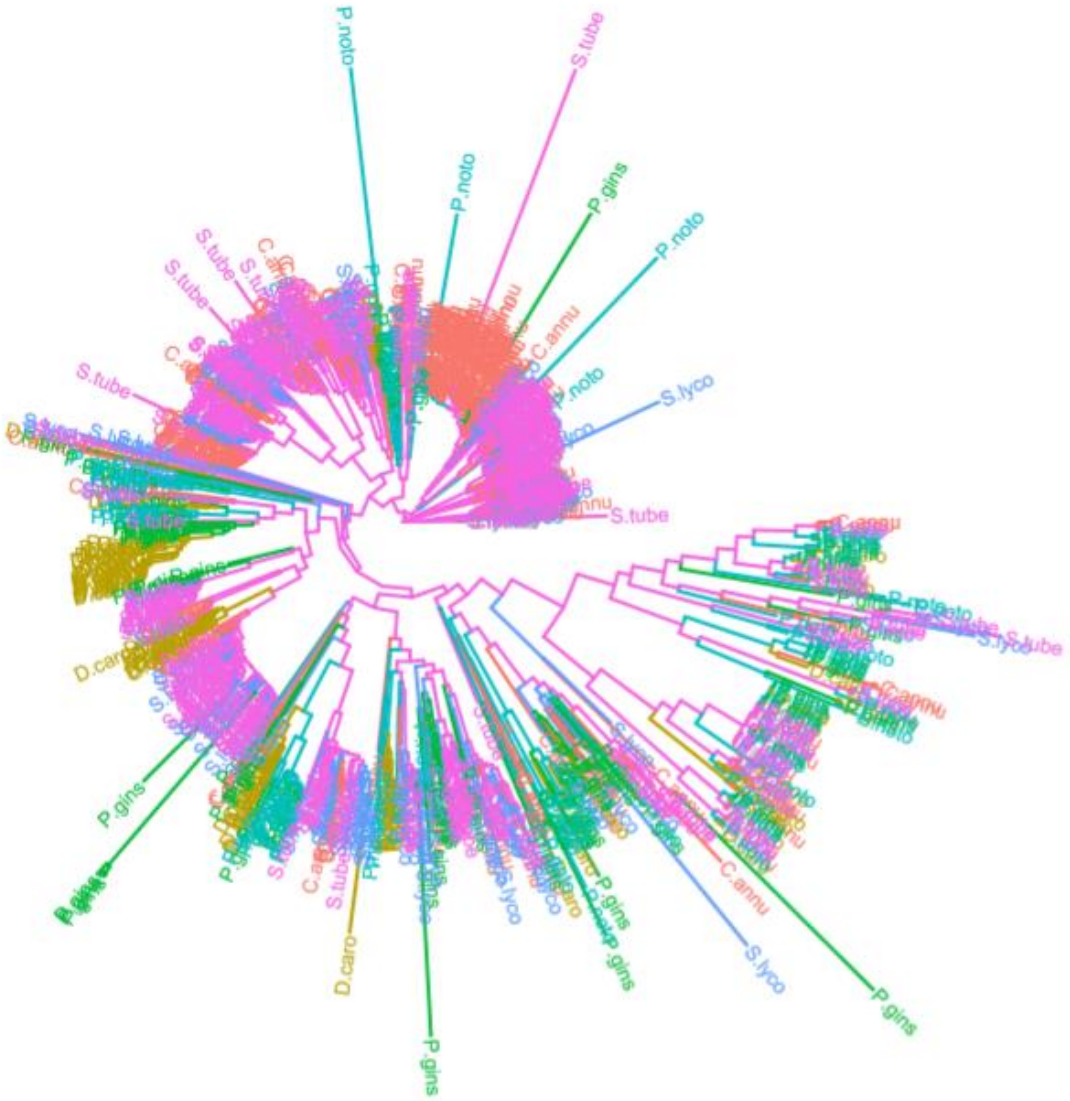


Figure S10. **Comparison of the NBS R-genes**, P.noto is the genes of *P. notoginseng*, P.gins is the genes of *P. ginseng*, S.tube is the genes of *Solanum tuberosum*, C.annu is the genes of *Capsicum annuum*, S.lyco is the genes of *Solanum lycopersicum*, D.caro is the genes of *Daucus carota*. Related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

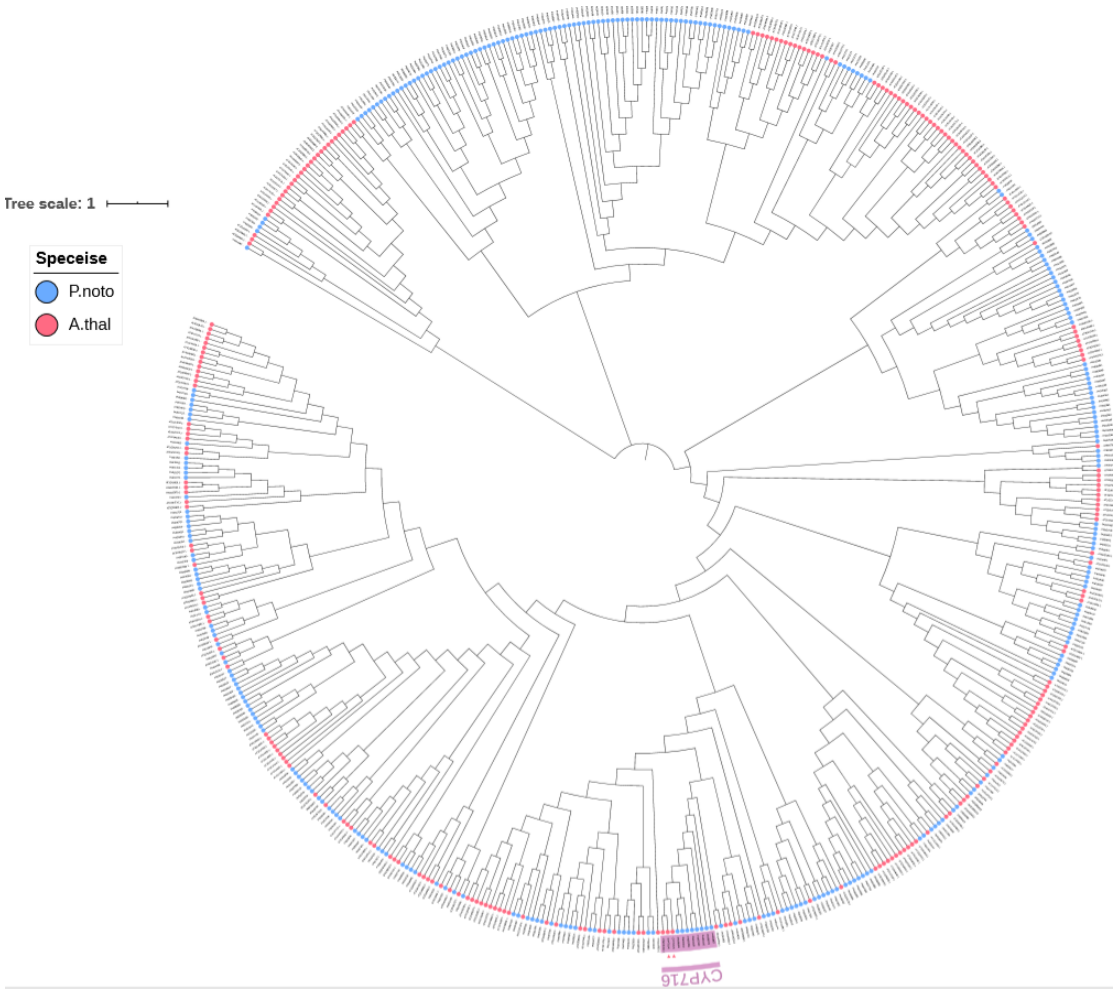


Figure S11. The gene trees of CYP450 of *P. notoginseng* and *A. thaliana*. Related to Figure 3.

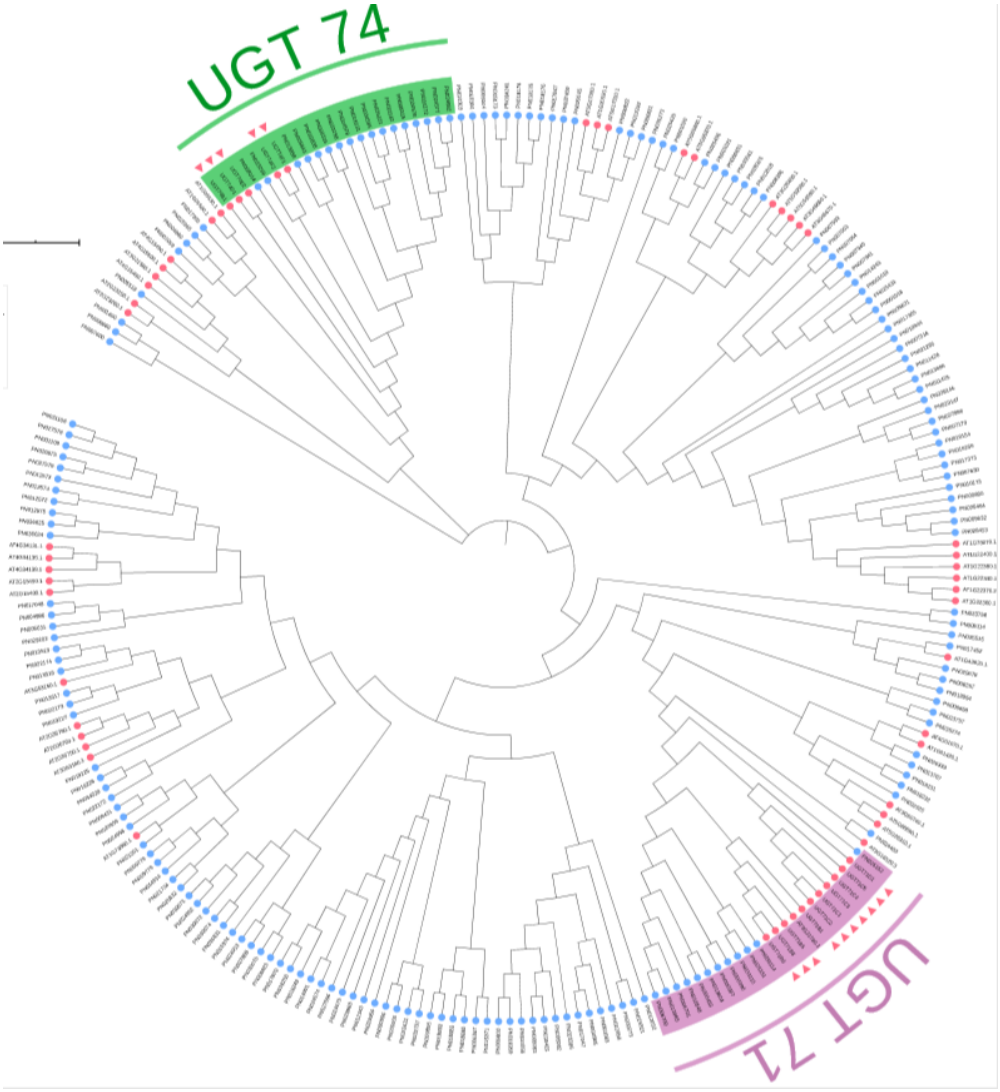


Figure S12. Phylogenetic analysis for classifying the UGT subfamilies of *P. notoginseng* based on the subfamily class of *P. ginseng* and *A. thaliana.* Related to Figure 3.

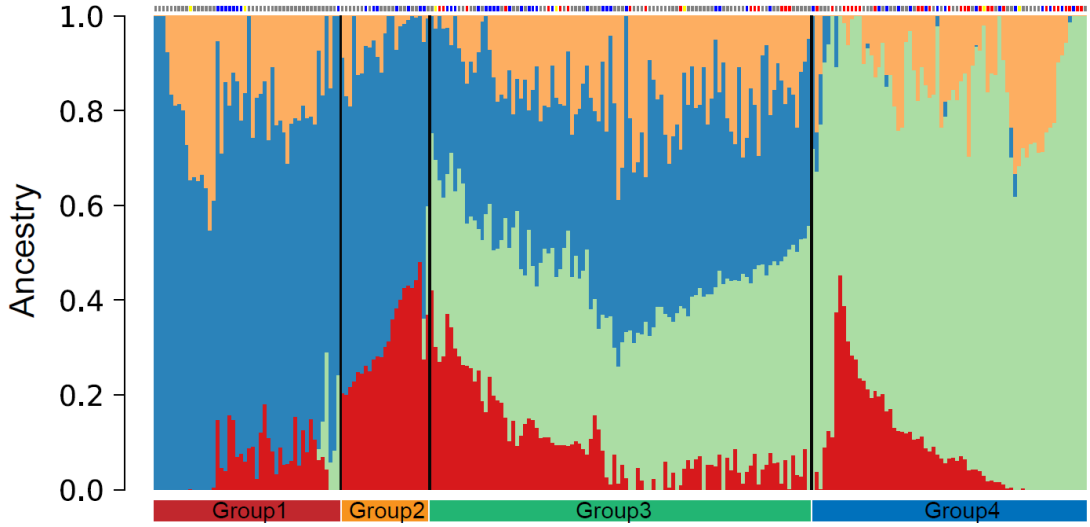


Figure S13. The population structure of this resequencing population. Related to Figure 4.

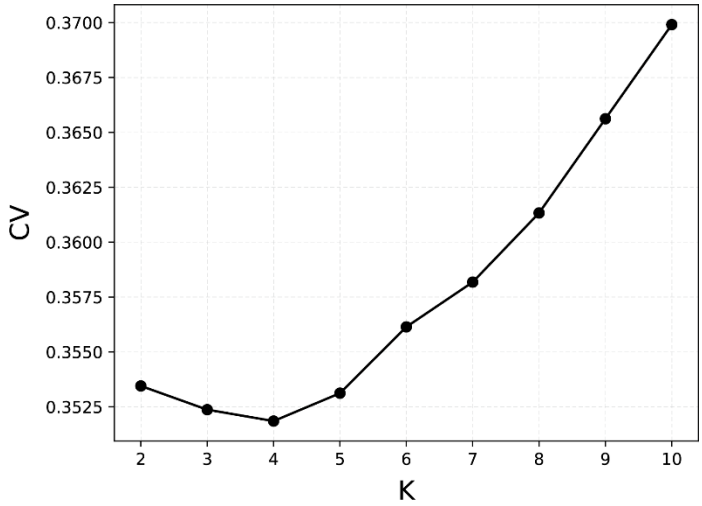


Figure S14. The estimated best K if this population structure. Related to Figure 4.

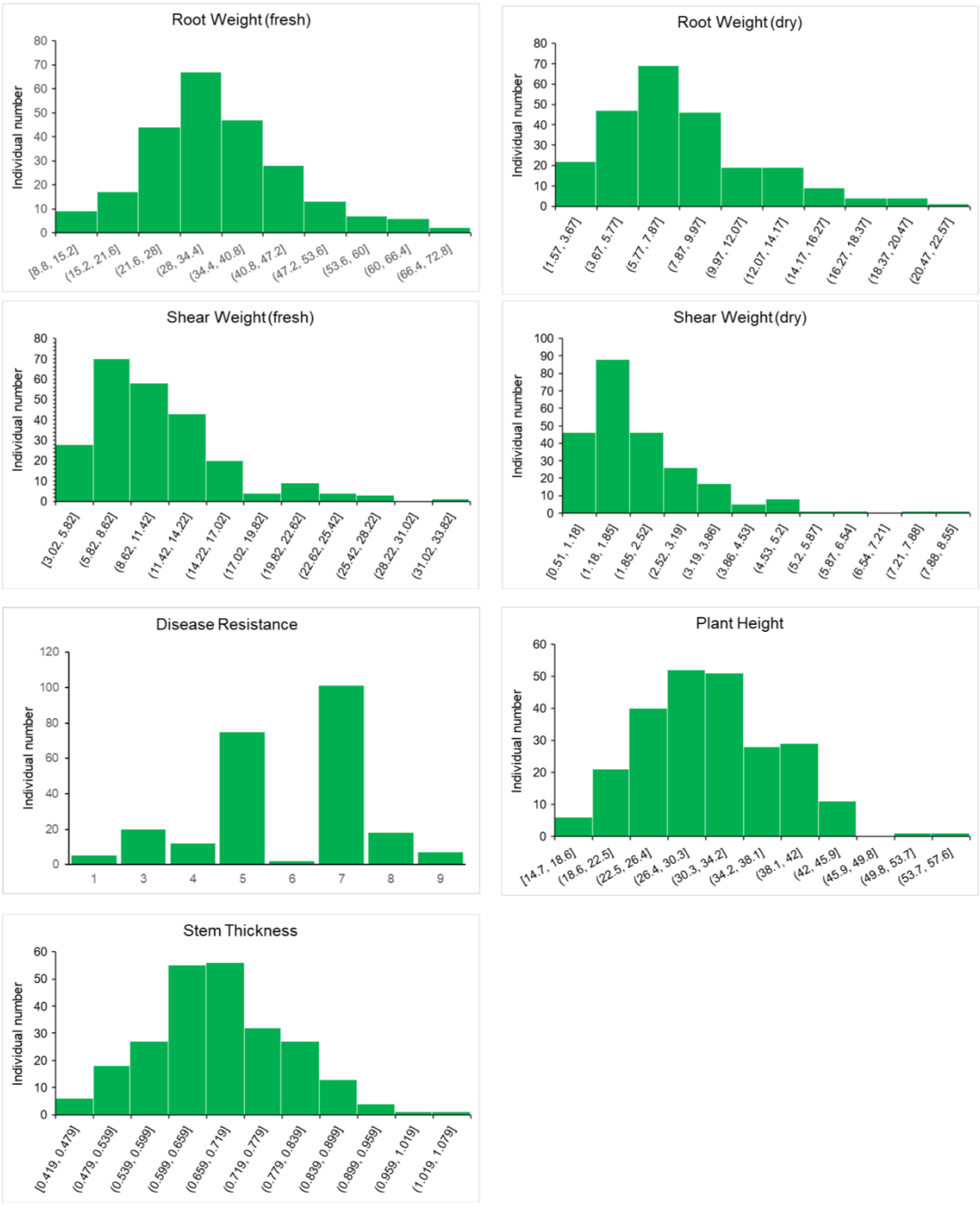


Figure S15. The statistics of sequencing data and seven phenotypic traits. Related to Figure 4.

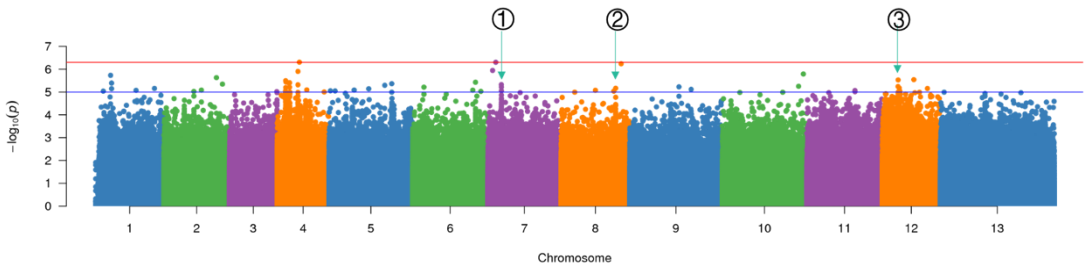


Figure S16. The Manhattan plot of the trait of plant height. Related to Figure 4.

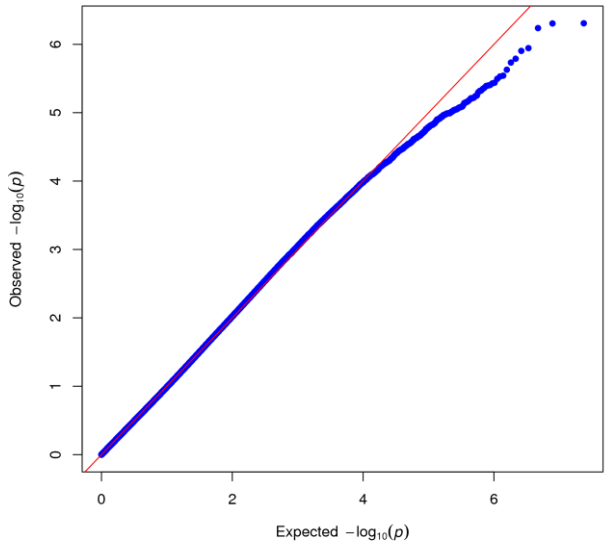


Figure S17. The QQ plot of the trait of plant height. Related to Figure 4.

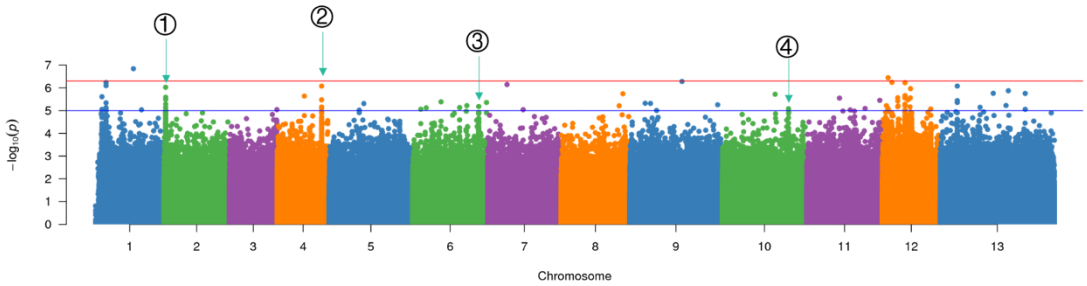


Figure S18. The Manhattan plot of the trait of root weight (fresh). Related to Figure 4.

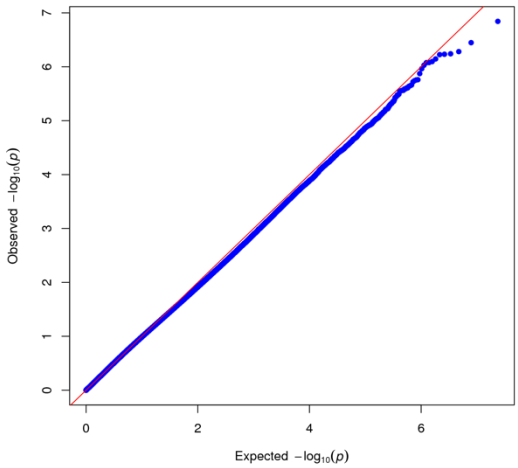


Figure S19. The QQ plot of the trait of root weight (fresh). Related to Figure 4.

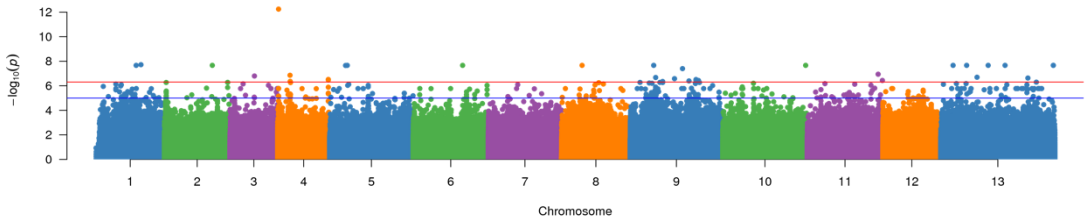


Figure S20. The Manhattan plot of the trait of shear weight (dry). Related to Figure 4.

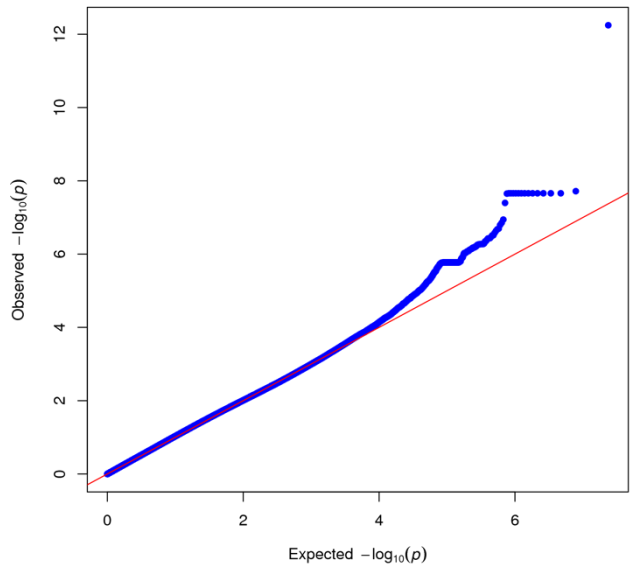


Figure S21. The QQ plot of the trait of the shear weight (dry). Related to Figure 4.

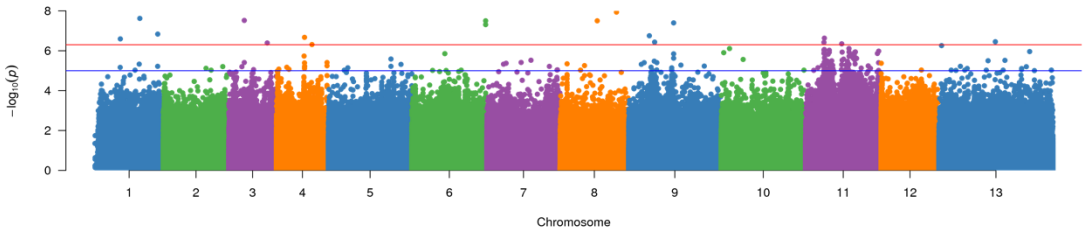


Figure S22. The Manhattan plot of the trait of shear weight (fresh). Related to Figure 4.

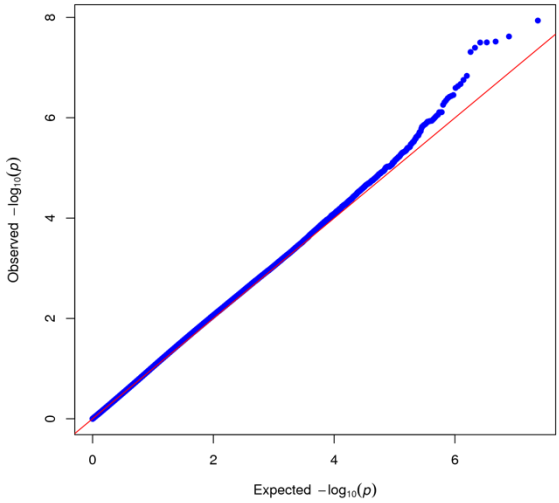


Figure S23. The QQ plot of the trait of shear weight (fresh). Related to Figure 4.

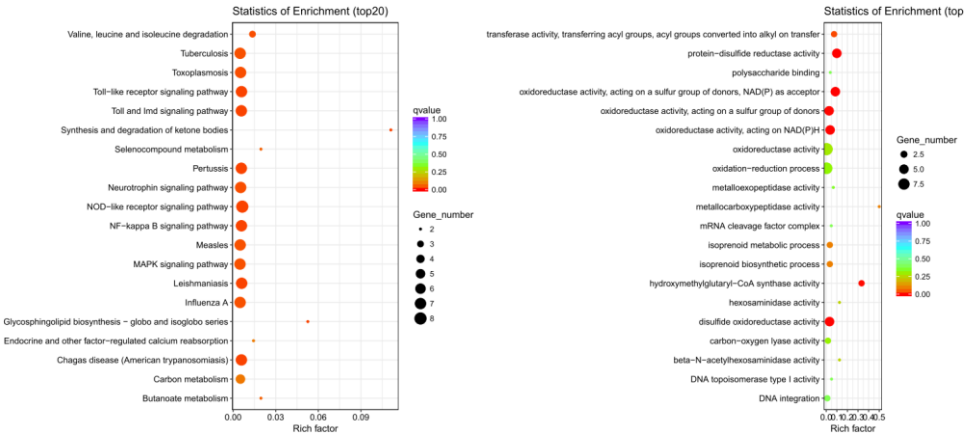


Figure S24. The KEGG and GO enrichment results of the root weight. Related to Figure 4.

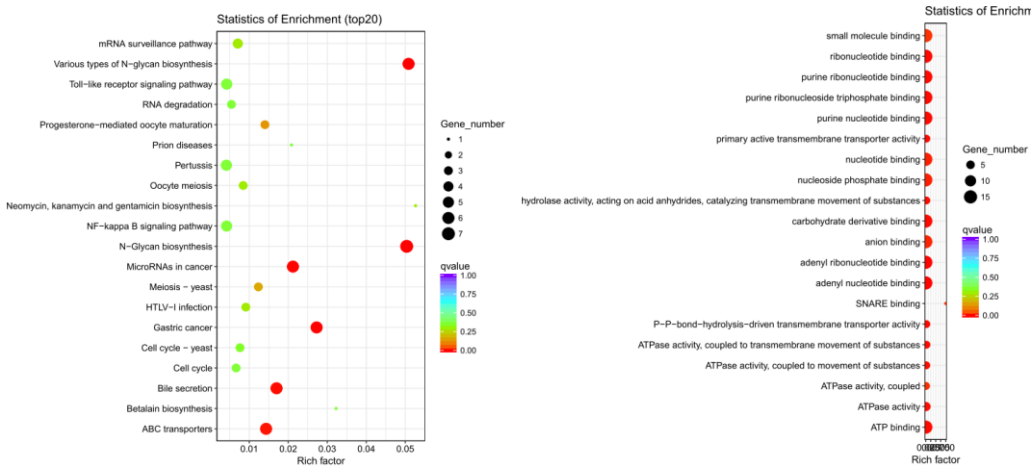


Figure S25. The KEGG and GO enrichment results of the stem thickness. Related to Figure 4.

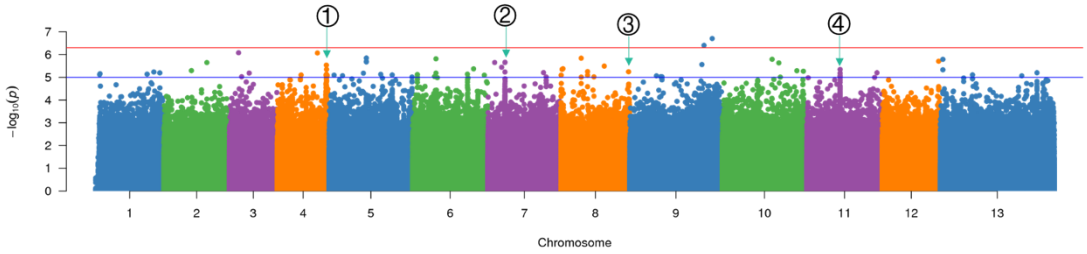


Figure S26. The Manhattan plot of the trait of disease resistance. Related to Figure 4.

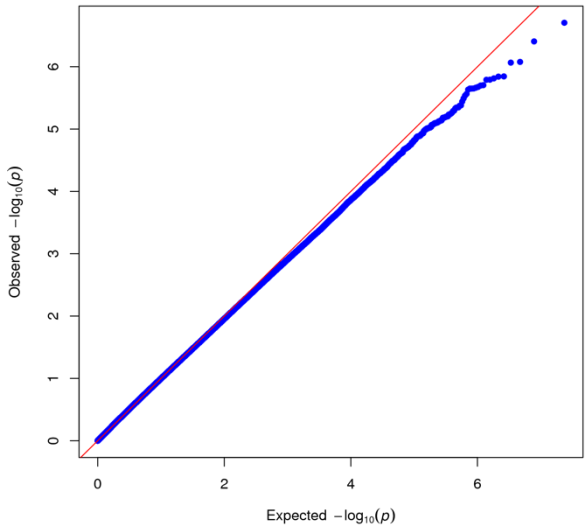


Figure S27. The QQ plot of the trait of disease resistance. Related to Figure 4.

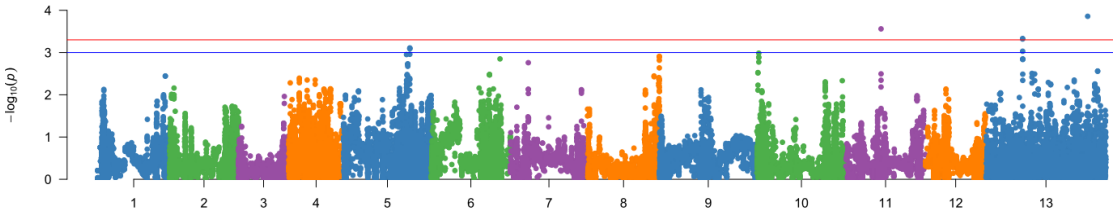


Figure S28. The result of the gene set-based association test using fastBAT. Related to Figure 4.

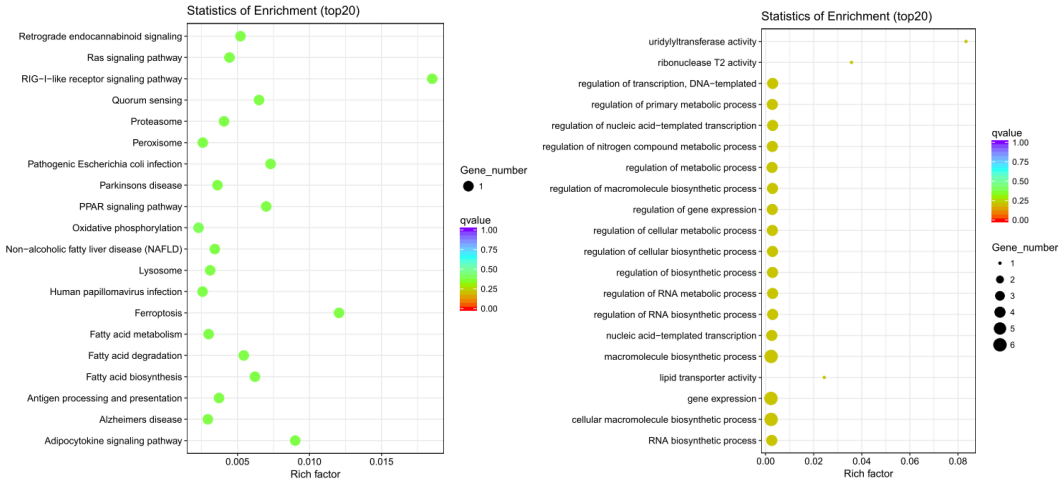


Figure S29. The KEGG and GO enrichment results of the disease resistance. Related to Figure 4.

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.