A *Panax ginseng* pangenome reveals two independent alleles underlying yellow berry color

Woohyeon Cho
Seoul National University
Xuejiao Liao
China Academy of Chinese Medical Sciences
Ju-Young Ahn
Seoul National University
Hyeonah Shim
Seoul National University
Jung-Woo Lee
Rural Development Administration
Young Chang Kim
Rural Development Administration
Yong Chan Park
CNARES
Yurry Um
National Institute of Forest Science
Sang Hyun Moh
Plant Cell Research Institute of BIO-FD&C Co.
Haoyu Hu
China Academy of Chinese Medical Sciences
Shilin Chen
Chengdu University of Traditional Chinese Medicine
Deqiang Dou
Liaoning University of Traditional Chinese Medicine
Yidan Xi
Liaoning University of Traditional Chinese Medicine
Yun Sun Lee
Seoul National University
Yeon Jeong Kim
Seoul National University
Jee Young Park
Seoul National University
Hojin Ryu
Chungbuk National University
Jungyeon Kim
Seoul National University
Young-Ju Oh
Seoul National University
Hyun-Hee Kim
Sahmyook University

Nomar Espinosa Waminal
Institute of Plant Genetics and Crop Plant Research
Murukarthick Jayakodi
Texas A&M AgriLife Research and Extension Center
Jiang Xu
China Academy of Chinese Medical Sciences
Martin Mascher
Institute of Plant Genetics and Crop Plant Research
Tae-Jin Yang
tjyang@snu.ac.kr

Seoul National University

Research Article

Keywords: ginseng, pangenome, structural variation, berry colour, DFR

Posted Date: May 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9694379/v1

Additional Declarations: No competing interests reported.

Abstract

Background

Panax ginseng C. A. Meyer is a medicinally and economically important perennial herb valued for its diverse pharmacologically active compounds. However, the genetic basis of phenotypic variation in ginseng remains poorly understood because of its large, highly repetitive, and allotetraploid genome. Berry colour variation is a conspicuous trait in ginseng and may reflect differences in specialized metabolites such as flavonoids, yet its genetic basis has remained unclear. To investigate the genomic basis of this trait and capture broader genetic diversity, we constructed a chromosome-scale pangenome using sixteen diverse accessions, including cultivated varieties, wild-simulated ginseng, and approximately 50-year-old wild ginseng from Korea and Russia.

Results

We generated a chromosome-scale pangenome with multiple near telomere-to-telomere assemblies, providing highly contiguous and near-complete reference genomes for diverse ginseng accessions. Comparative analyses identified dihydroflavonol 4-reductase (*DFR*), a key enzyme in the anthocyanin biosynthesis pathway, as the major genetic determinant of berry colour. Two independent loss-of-function alleles were identified: a missense mutation (Gly17Arg) affecting a conserved NADPH-binding residue and a ~ 20 kb deletion removing the entire gene. Metabolite profiling and population-scale analyses further supported the central role of DFR in regulating flavonoid accumulation in ginseng berries.

Conclusion

Our findings demonstrate that pangenome analysis can effectively uncover structural variants underlying metabolic trait diversity in complex plant genomes. This study advances understanding of berry colour variation in ginseng and provides a valuable genomic framework for trait discovery and molecular breeding in this genetically and experimentally challenging medicinal crop.

Background

Panax ginseng C. A. Meyer is a high-value medicinal plant cultivated primarily in East Asia [1]. Despite its long history of medicinal use and substantial economic importance [2, 3], ginseng remains genetically underexplored compared with many major crop species. This gap largely reflects the biological and cultivation constraints associated with the species. *P. ginseng* is a slow-growing perennial that typically requires four years to reach reproductive maturity, making breeding and experimental studies inherently time-consuming [4]. In addition, the species is highly shade-adapted and sensitive to environmental conditions, requiring specialized cultivation under artificial shade structures that mimic forest canopy environments (Fig. 1a). These biological limitations are further compounded by the complexity of the ginseng genome. *P. ginseng* is an allotetraploid species (2n = 4x = 48) with an estimated genome size of approximately 3.6 Gbp and has experienced two relatively recent whole-genome duplication events [5, 6]. Consequently, the genome contains extensive gene redundancy, large-scale segmental duplication, and a high proportion of repetitive elements [7], posing substantial challenges for genetic analysis and breeding. Together, these biological and genomic constraints have limited the development of comprehensive genomic resources for *P. ginseng*, placing it among economically important medicinal plants that remain comparatively underexplored at the genetic level.

Understanding genomic diversity is particularly important for medicinal crops such as ginseng because variation in specialized metabolic pathways can directly influence the accumulation of bioactive compounds. Structural genomic variation—including insertions, deletions, inversions, copy number variation, and presence–absence variation—represents a major source of genetic diversity that can alter gene content or gene function. Increasing evidence from plant genomics

indicates that structural variants affecting genes involved in specialized metabolic pathways can drive phenotypic diversity in crops [8]. A major challenge in addressing this question is that structural variants are often poorly captured by analyses based on a single reference genome. Pangenome approaches provide a framework for representing the genomic diversity of a species by integrating sequences from multiple individuals into a shared genomic resource. Such approaches have revealed extensive gene content variation and structural polymorphisms in several crop species and have provided important insights into trait diversity and adaptation [9, 10]. Nevertheless, comprehensive pangenome resources remain limited for many medicinal and underexplored crops, including *P. ginseng*.

One conspicuous but poorly understood example of phenotypic diversity in ginseng is berry colour. Mature ginseng berries exhibit distinct colour phenotypes, most commonly red and yellow, with less frequent pink and orange forms. In addition to their visual differences, ginseng berries contain diverse bioactive compounds, including ginsenosides, flavonoids, and polysaccharides, and have attracted increasing attention due to their pharmacological potential [11–13]. Pigmentation in plant fruits is typically associated with flavonoid and anthocyanin biosynthesis, pathways that are closely linked to specialized metabolism [14]. Differences in berry colour may therefore reflect underlying variation in metabolic pathways. However, despite this conspicuous phenotypic variation and its potential metabolic significance, the genetic mechanisms underlying berry colour differentiation in *P. ginseng* remain unknown.

Here we construct a chromosome-scale pangenome of *P. ginseng* based on sixteen genetically diverse accessions representing major lineages within the species. Using this resource, we characterize genome-wide structural variation and investigate its contribution to phenotypic diversity. Focusing on berry colour variation, we identify causal genetic variants disrupting the anthocyanin biosynthesis pathway and demonstrate how structural genomic variation shapes metabolic traits in this medicinal crop. Our results provide both a genomic framework for studying trait diversity in *P. ginseng* and a broader example of how pangenome approaches can accelerate genetic discovery in underexplored crop species.

Results

Chromosome-level genomes for sixteen ginseng accessions

To investigate genomic variation underlying phenotypic diversity in *P. ginseng*, we constructed a chromosome-scale pangenome based on sixteen genetically diverse accessions. These accessions were selected from a population-scale diversity panel genotyped with a custom SNP array [15] to capture the breadth of genetic variation across the species. The panel spans four major phylogenetic clusters and includes accessions originating from Korea (n = 11), China (n = 4), and Russia (n = 1), representing the principal reservoirs of diversity within the global *P. ginseng* gene pool (Fig. 1a-c). In addition to maximizing genetic diversity, the panel also captured variation in berry colour phenotypes. Owing to the challenges of maintaining ginseng plants and the restricted availability of germplasm, most accessions correspond to genetically fixed cultivars maintained as adventitious root cultures for long-term preservation (**Additional file 2: Table S1**). Most accessions were maintained as adventitious roots to facilitate long-term preservation (Fig. 1d).

High-fidelity (HiFi) long-read sequencing was performed for total of 16 ginseng accessions, generating coverage depths of 11x to 34x (average 22.8x) (**Additional file 2: Table S2**). HiFi reads are assembled into contig-level assemblies with N50 values ranging from 3.4 Mb to 189.9 Mbp (average 74.9 Mbp). Assemblies with sufficient HiFi coverage (≥ 25x) showed high contiguity (N50 > 80 Mbp), whereas those with lower coverage (< 15x) were markedly more fragmented (N50 ≤ 13 Mbp) and exhibited reduced consensus quality (QV < 50). Despite this, BUSCO-based completeness scores remained consistently high across all assemblies (≥ 99%; average 99.2%) (**Additional file 2: Table S3**). Genome-wide heterozygosity was low (mean 0.13%), consistent with predominant self-fertilization in *P. ginseng*. Chromosome-scale pseudomolecules were generated for nine accessions using Hi-C data, while the remaining seven assemblies were scaffolded using a reference-guided approach anchored to the Geumsun genome. Final genome sizes ranged from 3.43 to 3.55 Gbp, with chromosome anchoring rates between 96.9% and 99.2% (**Additional file 2: Table S2**). Overall, our assemblies exhibit higher quality than previously published genomes, with markedly higher QV values across all samples (Table 1, S2).

Table 1
Assembly statistics and quality comparison between this study and previously published *P. ginseng* genomes.

No.	Sample name	Germplasm type	Country	Contig N50 (Mbp)	Genome size (Gbp)	Assembly quality (QV)	No. of telomeres	Genome BUSCO (%)	Gene BUSCO (%)
1	TYP	Collection	Korea	189.9	3.53	61.1	39	99.3	98.4
2	BioFD&C	Collection	Korea	77.7	3.54	60.1	40	99.3	98.4
3	Geumsun	Cultivar	Korea	91.7	3.51	62.5	41	99.2	98.4
4	pRussia	Wild collection	Russia	87.5	3.52	62.4	39	99.2	98.5
5	Cheonryang	Cultivar	Korea	103.7	3.54	62.0	41	99.2	98.5
6	G03009	Breeding line	Korea	89.8	3.51	61.3	46	99.2	98.4
7	Gumpoong	Cultivar	Korea	59.5	3.45	62.8	35	99.1	98.4
8	Chunpoong	Cultivar	Korea	87.1	3.53	61.3	40	99.2	98.5
9	G03055	Breeding line	Korea	110.9	3.54	40.2	44	99.3	98.4
10	HamyangG2020	Wild-simulated ginseng	Korea	76.8	3.55	60.7	38	99.2	98.4
11	NYP	Collection	Korea	85.0	3.48	63.7	42	99.3	98.4
12	Yunpoong	Cultivar	Korea	104.2	3.54	64.6	44	99.3	98.4
13	Jianbt	Collection	China	13.3	3.48	30.2	32	99.3	98.5
14	Kuandian	Collection	China	4.8	3.43	57.9	26	99.0	98.4
15	Fengcheng [36]	Wild-simulated ginseng	China	3.4	3.46	56.7	29	99.3	98.4
16	Jianxkh [36]	Wild-simulated ginseng	China	12.6	3.48	46.4	36	99.3	98.5

To assess variation in gene content, we constructed a single-copy pangenome comprising the 16 newly assembled genomes and two previously published references [16, 17]. After redundancy filtering, the resulting non-redundant dataset contained 985,320 single-copy sequences totaling 278.4 Mbp. Of these, 114.7 Mbp were shared among all accessions, defining the core genome, while 163.7 Mbp represented presence–absence variable (PAV) clusters absent from one or more accessions. Further, 144.0 Mbp consisted of unique clusters restricted to individual genomes, representing accession-specific or potentially novel gene fragments. The accumulation curve of single-copy sequences did not reach saturation, suggesting that the *P. ginseng* pangenome remains incompletely sampled and that additional genomes would likely reveal further sequence diversity (Fig. 1e).

Structural variation landscape

Approximately 84% of assembled sequences were masked as transposable elements (TEs), predominantly long terminal repeat (LTR) retrotransposons (**Additional file 1: Fig. S1a**). Among LTR elements, Gypsy-type retrotransposons were the most abundant, occupying 1.66–1.80 Gbp (46–50% of the total genome), followed by Copia (7–9%) and unclassified LTRs (22–

25%) (**Additional file 2: Table S4**). On average, each assembly contained 33,568 intact full-length LTR retrotransposons. Total TE content was strongly correlated with genome assembly size ($R^2 \approx 0.91$) (**Additional file 1: Fig. S1b**), with Gypsy elements contributing most strongly to this relationship, indicating that genome size variation among accessions is driven primarily by differential repeat content rather than expansion of gene space (**Additional file 1: Fig. S1c**).

Pairwise comparison of 15 pangenome assemblies against the Geumsun reference uncovered extensive presence–absence variation (PAV) below 1 Mbp. In total, 238,297 PAVs were identified, with each accession harboring tens of thousands of structural variants (**Additional file 2: Table S5**). Insertions and deletions were the most frequent classes, with about 6,000–9,000 events per genome affecting up to ~ 44 Mbp. Tandem expansions and contractions were less frequent but often spanned tens of megabases. Repeat-associated variants, particularly repeat contractions, had the largest cumulative impact on genome architecture, exceeding 100 Mbp in several accessions, such as 105.4 Mbp in Fengcheng and 108.1 Mbp in Jianxkh (**Additional file 1: Fig. S1d**). Across genomes, total repeat contraction length was negatively correlated with final assembly size, indicating that contraction of repetitive elements is the primary driver of genome size reduction in *P. ginseng* (**Additional file 1: Fig. S1e**). The G03055 accession showed the negligible amount of PAV, likely reflecting its genetic similarity to the reference, as also indicated by clustering analysis (Fig. 1b). Except G03055, the total number of structural variants ranged from 13,333 in NYP to nearly 19,685 in HamyangG2020, with cumulative affected sequence lengths varying from ~ 83.0 Mbp in NYP to ~ 260.9 Mbp in Fengcheng.

As transcriptomic data was unavailable, gene annotations represent computational predictions supported by genome sequence and homology-based evidence rather than experimentally validated transcripts. Nonetheless, a pan-gene projection [18, 19] approach enabled comparative analysis across accessions. Across the 16 pangenome assemblies, we identified a total of 127,642 predicted pan-genes. Individual genomes contained between 102,325 and 108,433 predicted genes, and BUSCO analysis indicated high completeness of the annotated gene sets (98.4–98.5%), demonstrating the robustness of gene prediction (**Additional file 2: Table S6**). Clustering of annotated genes identified a total of 36,700 gene families, of which 34,031 were present in all 16 accessions, 1,550 were present in multiple but not all accessions, and 1,119 were unique to a single genome.

Large chromosomal rearrangement

Whole-genome synteny analysis across nine chromosome-scale ginseng assemblies revealed multiple chromosomal rearrangements, including both inter- and intrachromosomal events that contribute to genomic differentiation. In total, 54 inversions larger than 1 Mbp were identified (**Additional file 1: Fig. S2**). Among these, a conspicuous rearrangement on chromosome 10—comprising two consecutive inversions of approximately 30 Mbp and 89 Mbp—segregated among accessions at a frequency of 6:3 (Yunpoong, G03009, Chunpoong, TYP, Jianxkh, NYP and Gumpoong, Biofdnc, Geumsun) (Fig. 2a). The 89 Mbp inversion contained a smaller, nested 10 Mbp inversion (Fig. 2b). PCR markers spanning the inversion breakpoints were developed to validate this complex structural architecture (Fig. 2b, **Additional file 2: Table S7**). The representative cultivars Geumsun (inverted haplotype relative to the ancestral orientation) and Yunpoong (non-inverted haplotype) produced amplification patterns consistent with the genome assemblies, validating the complex structural variation in chromosome 10 (Fig. 2c).

To further understand the evolutionary context of this rearrangement, ancestral haplotype block analysis was performed using SNP datasets derived from 114 resequenced accessions spanning diverse ginseng habitats, together with the 16 pangenome genotypes (**Additional file 2: Table S8**). The chromosome 10 haplotype block distinguished into two haplotype clusters corresponding to the inverted and non-inverted orientations (Fig. 2d). Accessions carrying the inverted orientation exhibited long, homogeneous haplotype blocks, indicative of a derived state, whereas non-inverted accessions showed much colorful haplotypes consistent with an ancestral configuration. Both haplotype orientations were present across Korean, Chinese, and Japanese accessions, suggesting that the inversion predates regional diversification and persisted through cultivation.

Additionally, two inter-chromosomal translocations were identified, one in TYP and the other in Gumpoong (Fig. 2a). Both TYP and Gumpoong were sequenced from adventitious-roots initiated in 2016 and since maintained in long-term tissue-culture. By contrast, other accessions were induced more recently or were sequenced directly from leaf tissue. These rearrangements are therefore likely the result of somaclonal variation [20, 21], a well-known phenomenon in plant tissue culture involving replication stress, transposon activation, and chromosomal mis-segregation. Consistent with this, reciprocal and non-reciprocal translocation signatures were identified in Hi-C contact map of TYP (**Additional file 1: Fig. S3**). Moreover, chromosome counting confirmed that the TYP sequencing sample carried 45 chromosomes (2n = 45), and this aneuploid state was independently supported by abnormal Hi-C interaction patterns and corresponding HiFi read-depth deviations (**Additional file 1: Fig. S4**).

Distinct genetic alterations of DFR underlying yellow berry

Berry color in ginseng is a phenotypic trait with both pharmacological and commercial significance, primarily segregating into red and yellow types, with red being the majority (Fig. 3a). To identify whether colour variation reflected differences in anthocyanin or carotenoid accumulation, we quantified pigment content in ripe berries of the red-berried cultivar Geumsun and the yellow-berried cultivar Gumpoong. Anthocyanin levels differed markedly between the two: Gumpoong showed near-background levels (0.0046 ± 0.0033 μg/mg DW), whereas Geumsun accumulated substantially higher amounts (0.0714 ± 0.0451), approximately 15-fold difference (**Additional file 2: Table S9**). Meanwhile, carotenoid concentrations were also higher in Geumsun (0.0737 ± 0.0100) than in Gumpoong (0.0381 ± 0.0054). These results indicate that the yellow berry phenotype primarily results from the loss of anthocyanin biosynthesis. Carotenoids are present in both berry types, but in red berries their visual contribution is masked by the accumulation of anthocyanins.

To identify the genetic basis of pigment variation, we examined genes in flavonoid biosynthetic pathway across 16 pangenomes, comprising 14 red-berry and two yellow-berry genotypes (**Additional file 2: Table S1**). The analysis encompassed nine key enzymes spanning the phenylpropanoid to anthocyanin biosynthetic steps: phenylalanine ammonia-lyase (*PAL*), cinnamate 4-hydroxylase (*C4H*), 4-coumarate–CoA ligase (*4CL*), chalcone synthase (*CHS*), chalcone isomerase (*CHI*), flavanone 3-hydroxylase (*F3H*), flavonoid 3'-hydroxylase (*F3'H*), dihydroflavonol-4-reductase (*DFR*), and anthocyanidin synthase (*ANS*). Gene copy number varied substantially across the pathway. The *PAL* family exhibited the highest copy number with up to 12 copies per genome, followed by two to six copies of *C4H, 4CL, CHS, CHI, F3'H* and *ANS (*Fig. 3b, **Additional file 2: Table S10**). In multi-copy families, the presence of additional intact paralogs likely compensates for impaired mutated genes, thereby maintaining overall pathway activity. In contrast, *F3H* and *DFR* were present as single-copy genes.

A notable gene was *DFR*. In contrast to the conserved single-copy *F3H*, *DFR* exhibited variation specifically in yellow-berry accessions. In Gumpoong, *DFR* carried a missense mutation, whereas in HamyangG2020 the gene was entirely absent. In the yellow-berry cultivar Gumpoong, a nonsynonymous SNP replaced glycine with arginine at position 17, whereas in HamyangG2020 a ~ 20 kb deletion eliminated the genomic region encompassing *DFR* (Fig. 3c). We identified two *DFR* homeologous copies located on chromosomes 1 and 13, consistent with the allotetraploid origin of *P. ginseng* (**Additional file 1: Fig. S5**). The chromosome 1 copy harbors a one-base-pair insertion that introduces a frameshift and premature stop codon, rendering it nonfunctional in both red- and yellow-berry genotypes. Consequently, the chromosome 13 copy represents the only functional *DFR* gene in the analyzed accessions. This lack of functional redundancy explains why mutations affecting the chromosome 13 copy alone abolish anthocyanin biosynthesis and give rise to the yellow berry phenotype.

DFR determines berry colour in Panax ginseng

As *DFR* catalyzes an NADPH-dependent reduction of dihydroflavonols to leucoanthocyanidins [22, 23], the absence of a redundant paralog renders pigment formation highly sensitive to *DFR* disruption in chromosome 13. For the SNP mutation, the substituted Gly17 residue lies within the conserved NADPH-binding domain of DFR proteins (**Additional file 1: Fig. S6**),

and replacement by a bulky, positively charged arginine is predicted to impair cofactor binding and enzymatic activity (Fig. 4a).

Metabolite profiling of mature berries provided functional support for these genetic observations (**Additional file 2: Table S11**). As naringenin is an upstream precursor of the anthocyanin pathway, it was not detected in either mature berry type. In contrast, pronounced differences were observed in downstream metabolite composition. Gumpoong accumulated high levels of dihydroflavonols, including dihydrokaempferol (5.3829 mg/kg DW) and dihydroquercetin (4.1559), whereas Geumsun contained only trace dihydrokaempferol (0.0543) and no detectable dihydroquercetin. In the other hand, Geumsun showed strong accumulation of pelargonidin-3-O-glucoside (21.9458), whereas Gumpoong showed complete absent. Also, Geumsun produced none Cyanidin-3-O-glucoside, while Gumpoong retained only a barely detectable trace (0.0548) (Fig. 4b). Together, these data demonstrate that *DFR* loss-of-function in yellow berries creates a metabolic bottleneck, preventing efficient reduction of dihydroflavonols and causing their accumulation. Metabolite profiling further indicated that, in red berries, metabolic flux is predominantly channeled toward the pelargonidin branch, likely due to rapid reduction of dihydrokaempferol by DFR. In contrast, in yellow berries, the loss of DFR activity appears to enable F3'H-dependent hydroxylation of dihydrokaempferol, resulting in the accumulation of both dihydrokaempferol and dihydroquercetin.

Molecular markers targeting the two distinct *DFR* variants—a missense SNP and a ~ 20 kb deletion—were developed to survey yellow-berried accessions in a broader panel (**Additional file 1: Fig. S7a, Additional file 2: Table S12, S13**). Of the 20 accessions, 18 carried the SNP allele, whereas the remaining two harbored the deletion variant (**Additional file 1: Fig. S7b, Additional file 2: Table S13**). These results are consistent with two independent loss-of-function variants at the same locus underlying the yellow-berry phenotype in ginseng. To extend validation at the population level, a k-mer–based genome-wide association study (GWAS) was conducted with 114 accessions (84 red and 30 yellow) and 16 pangenome accessions (14 red and 2 yellow). A single, continuous association peak was detected on chromosome 13, precisely overlapping the *DFR* locus (Fig. 4c, **Additional file 2: Table S8**). The concordance of sequence variation, metabolic profiles, and population-level associations provides strong support for *DFR* as a major genetic determinant of anthocyanin accumulation and berry colour variation in *P. ginseng*.

Discussion

Structural diversity in the ginseng genome

This study provides genomic insight into the genetic basis of berry colour variation in *P. ginseng* and demonstrates how pangenome analysis can reveal causal structural variants underlying metabolic trait diversity. Using this framework, we identified independent loss-of-function variants in the *DFR* gene that disrupt anthocyanin biosynthesis and give rise to the yellow berry phenotype. The chromosome-scale pangenome generated here reveals extensive structural diversity within *P. ginseng*, highlighting the importance of capturing genome variation beyond a single reference for understanding trait diversity in this species. Comparative analysis across the 16 assemblies identified numerous large-scale rearrangements, including inversions and translocations, illustrating the extent of structural variability within the species. While some rearrangements were shared among accessions and likely represent genuine heritable polymorphisms, others were detected exclusively in assemblies derived from long-term tissue culture. These observations highlight an important consideration when interpreting structural variation in plant genomes: culture-induced changes can introduce chromosomal alterations that do not necessarily reflect natural population diversity.

Large chromosomal inversions in plants have been associated with adaptive and agronomic traits through their effects on local recombination landscapes and the regulatory environment of nearby genes [24, 25]. For example, in barley, an inversion was detected exclusively among domesticated accessions, and the inverted haplotype was confined to genotypes of Western geographical origin, occurring approximately 440 kb from the *HvCEN* gene, a locus involved in flowering time adaptation [26]. In *P. ginseng*, the chromosome 10 region displays a more complex architecture, consisting of two consecutive inversions with an additional nested inversion. Although no genes reside exactly at the inversion breakpoints,

several genes are located in close proximity and may experience altered genomic context or regulatory influences as a result of this structural reorganization. The scale and stability of the chromosome 10 inversion underscore the importance of large structural variants as potential contributors to phenotypic diversity. Integrating high-resolution phenotyping with structural genomic data will be necessary to determine whether this rearrangement has functional consequences in ginseng biology.

Recurrent disruption of DFR underlying berry colour variation

This study provides genomic insight into the genetic basis of berry colour variation in *P. ginseng*. We identified two independent loss-of-function alleles of *DFR*: a nonsynonymous SNP affecting a conserved residue within the NADPH-binding domain and a large structural deletion that removes the *DFR* locus entirely. Both mutations abolish anthocyanin biosynthesis and give rise to a yellow-berry phenotype, yet they represent distinct mutational events. This convergence demonstrates that phenotypically similar accessions can carry different underlying causal genotypes, an important consideration for germplasm characterization and the development of precise genetic tools.

Independent support for the involvement of *DFR* in berry colour determination comes from a recent study in *P. quinquefolius*, which implicated the same gene through transcriptome profiling [27]. While *P. quinquefolius* study provided expression-level evidence and our analysis offers genomic sequence-level evidence, both converge on *DFR*, mutually reinforcing its central role in determining berry colour in tetraploid *Panax*. Notably, *P. ginseng* and *P. quinquefolius* diverged following an allotetraploidization event, after which the *P. ginseng* lineage remained in Asia, whereas *P. quinquefolius* migrated to North America [5]. The repeated association of yellow-berry phenotypes with *DFR* in these evolutionarily distinct lineages suggests either that a yellow-berried ancestor may have existed prior to divergence or that *DFR* is a recurrent target of inactivation, reflecting its sensitivity and evolutionary lability within the anthocyanin biosynthetic pathway.

Similar patterns of recurrent *DFR* disruption have been reported in other plant species [28], especially in yellow onion [29], in which independent *DFR* alleles underlie colour variation. Also, many studies have identified *DFR* as the causal gene underlying colour variation, with diverse types of mutations reported across species [30–33]. The repeated involvement of this single locus is striking in light of *DFR*'s central biochemical role within the flavonoid biosynthetic pathway. Together, these observations indicate that, within flavonoid biosynthetic networks, *DFR* constitutes a genetically labile node whose inactivation yields large phenotypic effects with minimal pleiotropic cost.

Metabolic consequences of DFR disruption

Pigment quantification showed that both red and yellow berries contain measurable levels of carotenoids (**Additional file 2: Table S9**). Berry colour differences therefore arise from the presence or absence of anthocyanins, which determines whether carotenoids are visually masked or remain the dominant visible pigment (**Fig. S8**). Disruption of *DFR* markedly alters flavonoid metabolite profiles. In yellow-berried accessions, loss of *DFR* activity creates a metabolic bottleneck that leads to the accumulation of dihydroflavonols, whereas red-berried accessions efficiently convert these intermediates into anthocyanin end products.

Dihydroquercetin, also called taxifolin, is a widely distributed plant flavonoid with antioxidant, anti-inflammatory and cytoprotective properties and is generally regarded as safe for human consumption [34, 35]. The substantial accumulation of taxifolin and dihydrokaemferol in yellow berries therefore suggests that these accessions represent a metabolically distinct chemotype of *P. ginseng*. Although functional and nutritional assays were beyond the scope of this study, the altered flavonoid composition implies that yellow-berried genotypes may offer advantages as a source of bioactive compounds that are largely depleted in red-berried forms. Their predictable accumulation of taxifolin, driven by naturally occurring *DFR* loss-of-function alleles, provides a clear genetic entry point for the targeted exploitation or breeding of ginseng varieties optimized for specific flavonoid profiles.

Conclusions

In this study, we constructed a chromosome-scale pangenome of *P. ginseng* using sixteen genetically diverse accessions and characterized genome-wide structural variation in this complex allotetraploid medicinal crop. Comparative analyses identified *DFR* as the major genetic determinant of berry colour variation, with two independent loss-of-function alleles disrupting anthocyanin biosynthesis and altering flavonoid metabolism. These findings provide one of the first clear examples linking structural genomic variation to phenotypic trait diversity in *P. ginseng*.

Beyond elucidating the genetic basis of berry colour, this work establishes a valuable genomic resource for investigating genetic and metabolic diversity in ginseng. The high-quality pangenome assemblies generated here reveal extensive structural variation that is difficult to capture using single-reference approaches and provide a framework for future studies of trait-associated variation in this species. Given the long generation time, slow growth, and environmental sensitivity of ginseng, genomic resources and molecular markers such as those identified in this study may facilitate more efficient germplasm characterization and marker-assisted breeding. More broadly, our results demonstrate how pangenome-scale analyses can accelerate genetic discovery in medicinal and other genomically underexplored crop species.

Methods

Pangenome accessions in diversity space

To capture representative genomic diversity within *Panax ginseng*, pangenome accessions were selected based on population structure inferred from a previously developed high-density SNP array applied to over 1,000 genotypes [15]. Principal component analysis (PCA) of the genotyping data was used to visualize and quantify genetic variation across the diversity panel.

Sixteen accessions were chosen to span the breadth of genetic diversity observed in the PCA plot, minimizing redundancy and maximizing genomic representation. These accessions included 11 from Korea, 4 from China, and 1 from Russia. Selected individuals were subsequently used for whole-genome sequencing and pangenome construction.

Plant materials

A total of 16 *P. ginseng* individuals were used for pangenome analysis. Of these, 11 were from Korea, 4 from China, and 1 from Russia (**Additional file 2: Table S1**). The 11 Korean accessions and the Russian accession were induced to form adventitious roots under in vitro conditions to enable clonal propagation from a single individual, thereby minimizing heterozygosity and individual-level bias in downstream genomic analyses. These adventitious roots were used as the source material for genomic DNA extraction and sequencing. In contrast, leaf tissues were collected directly from the 4 Chinese individuals for DNA isolation. Among the four Chinese accessions, genome assemblies for two accessions were available from previous studies and were included in the present pangenome analysis [36].

PacBio HiFi sequencing

HiFi circular consensus sequencing (CCS) reads were generated by PacBio Sequel IIe instrument (Pacific Biosciences) following the manufacturer's instructions. Per genotype, about 90 Gbp were sequenced to obtain an approximate pseudohaploid genome coverage of about 20-fold.

Hi-C library preparation and Illumina sequencing

In situ Hi-C libraries were prepared from ginseng adventitious roots based on a previously published protocol [37]. For this, adventitious root tissues from the 11 Korean and 1 Russian ginseng individuals were used. In addition, one Chinese individual used leaf tissue collected directly from a plant leaf tissue. Sequencing and Hi-C raw data processing was performed as described before [37, 38]. Libraries were sequenced using a 151 bp paired-end mode on an Illumina NovaSeq platform (Illumina). Per genotype, about 36 Gbp were sequenced to obtain an approximate pseudohaploid genome coverage of about 10-fold.

Genome sequence assembly and validation

PacBio HiFi reads were assembled using hifiasm (v.0.11-r302) [39]. Pseudomolecule construction was done with the TRITEX pipeline [40]. Chimeric contigs and orientation errors were identified through manual inspection of Hi-C contact matrices. For the genotypes with lack of Hi-C data, primary contigs were assembled with hifiasm and subsequently scaffolded in a reference-guided manner using RagTag [41], with the Geumsun as a reference. Genome completeness and consensus accuracy were evaluated using Merqury (v.1.3) [42] and BUSCO using the embryophyta_odb10 database [43]. Levels of duplication and heterozygosity were assessed with Merqury and FindGSE (v.1.94) [44]. Further, we estimated heterozygosity in the HiFi reads with a k-mer approach.

Single copy pangenome construction

The single-copy regions in each chromosome-level assembly were identified by filtering 31-mers occurring more than once in the genomic regions by BBDuk (BBMap_37.93, https://jgi.doe.gov/data-and-tools/software-tools/bbtools). BBMap was used to count k-mer occurrences in each genome with the parameter –mincount 2. Then, non-unique genomic regions (that is, those composed of k-mers occurring at least twice) were masked by BBDuk on the basis of k-mer counts. Single-copy regions extracted in BED format and their sequences (with the command 'bedtools complement') were retrieved using BEDTools (v.2.29.2) [45]. The single-copy sequences were clustered using MMseqs2 (Many-against-Many sequence searching) [46] with the parameters '--cluster-mode' and setting over 95% sequence identity. A representative from each cluster (the largest in a cluster) was selected to estimate the pangenome size.

Structural variant calling

Reciprocal genome alignment, in which each of the pangenome assemblies was aligned to the Geumsun assembly with the latter acting either as alignment query or reference, was done with Minimap2 (v.2.20) [47]. From the resultant two alignment tables, indels were called by Assemblytics (v.1.2.1) [48] and only deletions were selected in both alignments to convert into presence/absence variants relative to the Geumsun reference genome. Further, balanced rearrangements (inversions, translocations) were scanned for with SyRI [49].

Annotation of TEs

Annotation of the genome repeat sequences was performed using Extensive de novo TE Annotator software [50]. The first step involved conducting de novo identification of repetitive elements, where individual TE libraries were constructed with parameters '--sensitive 1 --anno 1' and RepeatMasker v4.1.1 (http://www.repeatmasker.org) was used to mask the genome. These libraries served as crucial inputs for the subsequent step of structural annotation of the TEs.

Gene annotation and pan-gene projection

Gene annotation was performed independently for each of the 16 *P. ginseng* genome assemblies using BRAKER3 [51]. As no RNA-seq data were available, gene prediction relied on ab initio modeling supported by protein homology evidence from closely related species. Predicted gene models from all accessions were combined and clustered to construct a non-redundant source gene set (https://github.com/GeorgHaberer/gene_projection). Protein sequences were clustered using CD-HIT [52] to remove redundancy, requiring 100% sequence identity and allowing a maximum length difference of 4 bp, generating a representative pan-gene source model for downstream projection and comparative analyses. The pan-gene source models were aligned to each of the 16 genome assemblies using sequence similarity–based mapping. Multiple filtering steps were applied to remove partial, low-confidence, or ambiguous projections.

Gene presence-absence variation

Gene presence–absence variation analysis was conducted by clustering every annotated coding sequence of 16 genomes into gene families using MMseqs2 [53]. Clustering was performed with a minimum sequence identity threshold of 0.90 and a coverage requirement of 0.20 based on sequence similarity. Gene families were classified to core, accessory, and unique according to their distribution across accessions.

Ancestral haplotype block analysis

To infer ancestral haplotype group (AHG) structures across ginseng accessions, we applied IntroBlocker [54] using whole-genome SNP data. Resequencing reads from 99 accessions and PacBio HiFi long-reads from 16 pangenome were mapped to the Geumsun reference genome using BWA-MEM [55] with default parameters. SNPs were called using VCFtools [56] and filtered with bcftools [57] to ensure high-confidence variants, retaining only biallelic sites with a Phred quality score > 20, missing data < 10%, heterozygosity < 10%, and minor allele frequency (MAF) > 5%. The filtered SNP dataset was used as input for IntroBlocker, with the ancestral haplotype grouping (AHG) bin size set to 3 Mb.

Cytogenetic analysis

Root tips from hairy root cultures were freshly harvested and treated with nitrous oxide at 10 bars for five hours, fixed in Carnoy's solution (3:1 ethanol - acetic acid) for at least two days and stored in 70% ethanol at 4°C until use. Chromosome spreads were prepared according to Peniton, et al. [58]

Pigmnet analysis sample preparation

For pigment analysis, pericarp tissues were collected from mature berries of two ginseng cultivars showing distinct berry colors: Geumsun (red berry) and Gumpoong (yellow berry). Three independent biological replicates were prepared for each cultivar. Samples were freeze-dried and ground into fine powder before extraction.

Anthocyanin extraction and quantification

Anthocyanins were extracted following a modified methanol-formic acid protocol. For each sample, 200 mg of dry powder was mixed with 1 mL of extraction solvent composed of 50% methanol containing 3% formic acid (v/v). The mixture was incubated overnight (≥12 h) at -20°C in darkness to prevent pigment degradation. After extraction, samples were centrifuged to remove debris, and the absorbance of the supernatant was measured using a spectrophotometer at 532 nm. Relative anthocyanin content was expressed in arbitrary units normalized per milligram of dry weight.

Carotenoid extraction and quantification

Carotenoids were extracted from 40 mg of dry powder with 4 mL of 80% acetone. Samples were subjected to ultrasonic extraction for 15 min and centrifuged to obtain the supernatant. For measurement, absorbance was recorded at 663.2 nm, 646.8 nm, and 470 nm using a UV spectrophotometer. Pigment concentrations (μg/mL) were first estimated using the following equations:

Chla = 12.25A663.2-2.79A646.8

Chlb = 21.5A646.8-5.1A663.2

Carotenoid = (1000A470-1.82Ca-85.02Cb)/198

Identification of flavonoid biosynthesis genes and their variations

Genes involved in the flavonoid biosynthesis pathway were identified from reference annotations and KEGG orthology [59]. Their presence, completeness, and coding sequence validity were manually inspected across all 16 genome assemblies using the corresponding GFF annotation files. Open reading frames (ORFs) were evaluated for frameshifts, premature stop codons, and truncations to identify potential loss-of-function alleles. In addition, structural variants overlapping flavonoid pathway genes were retrieved from Assemblytics results and manually inspected for functional impact.

DFR protein structure prediction

The three-dimensional protein structures of *P. ginseng DFR* genes were predicted using AlphaFold2 (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb). The protein sequences

corresponding to red- and yellow-berry accessions were individually submitted to generate structure models. The resulting predicted structures were visualized and compared using UCSF ChimeraX (version 1.10.1) [60].

Metabolite quantification analysis

Targeted metabolite quantification was performed using a Rapid LC–MS/MS system equipped with a TSQ Altis triple quadrupole mass spectrometer (Thermo Scientific, USA). Metabolites were separated by liquid chromatography and detected in multiple reaction monitoring (MRM) mode under optimized conditions for each compound. Authentic chemical standards, including naringenin, dihydrokaempferol, dihydroquercetin, pelargonidin-3-O-glucoside, and cyanidin-3-O-glucoside, were purchased from MedChemExpress (MCE, USA) and used for metabolite identification and quantification. Standard curves were generated using serial dilutions of each reference compound to enable accurate quantification. Metabolite quantification was performed with three biological replicates of mature berry for the red-berry cultivar Geumsun and the yellow-berry cultivar Gumpoong.

SNP and InDel marker development for DFR gene variation

To facilitate marker-assisted screening of yellow-berry phenotypes, SNP and InDel variants in the *DFR* gene (Chr13) were targeted for marker design. Molecular markers were developed and validated in an extended panel of yellow-berry accessions from the Rural Development Administration (RDA).

To distinguish the structural variants at the *DFR* locus associated with berry color, InDel markers were designed targeting the ~ 20 kb deletion. A shared forward primer was designed in the flanking region immediately upstream of the deleted segment, allowing amplification in both deletion and non-deletion genotypes. Two reverse primers were used to differentiate allele types. DFR_InDel_R1 binds within the deleted region and produces a 157 bp amplicon in genotypes without deletion (intact DFR locus). DFR_InDel_R2 anneals downstream of the deletion breakpoint and amplifies a 235 bp product in genotypes with deletion (deletion-type allele). The marker design enables genotype discrimination by PCR using either primer pair in combination with the shared forward primer. Primer properties (Tm ~ 58–59°C, GC content 45–50%) were optimized to prevent secondary structures and ensure specificity.

To detect the SNP in the *DFR* gene associated with yellow berry phenotype, high-resolution melting (HRM) primers were designed to flank the variant site. The forward primer and reverse primer amplify a short fragment suitable for HRM analysis. Primers were designed to produce a specific amplicon covering the SNP site, with melting temperatures around 58°C and GC contents between 38–50%, avoiding secondary structures.

Declarations

Data availability

All the sequence data collected in this study have been deposited at the Korea BioData Station (K-BDS) [61] under BioProject KAP242104. Accession codes for individual genotypes are listed in the supplementary tables in Additional file 2: Table S2 (pangenome assemblies and associated raw data) and Table S8 (resequencing data for population-level analysis). Gene annotation files are available in Figshare (https://doi.org/10.6084/m9.figshare.31304941). The datasets supporting the conclusions of this article are included within the article and its additional files.

Acknowledgements

The authors thank Nils Stein for assistance with the setup and guidance of the Hi-C experiment.

Funding

This work was carried out with the support of "Cooperative Research Program for Agriculture Science and Technology Development (Project No. PJ015903)" Rural Development Administration, Republic of Korea.

Author contributions

T.J.Y. and J.X. conceived and designed the study. M.M. and T.J.Y. supervised the pangenomic analyses and interpreted the results. W.C. and X.L. coordinated the experiments and sequencing. J.Y.A. and H.S. selected the genotypes and prepared the samples. J.W.L., Y.C.P., Y.U., S.H.M., S.C., D.D., H.H., and Y.X. provided plant materials. Y.S.L. and Y.J.K. supervised the metabolite analysis. J.Y.P. managed the project budget. Y.J.O. assisted the experiments. H.H.K. and N.E.W. performed chromosome counting. H.R., J.K., and M.J. provided critical input on the study. W.C. performed the pangenome analysis and wrote the manuscript. All authors read and approved the final manuscript.

Ethics declarations

Ethics approval and consent to participate

Not applicable**.**

Consent for publication

Not applicable**.**

Competing interests

The authors declare no competing interests.

References

1. Baeg I-H, So S-H. The world ginseng market and the ginseng (Korea). J ginseng Res. 2013;37(1):1.
2. Yun TK. Brief introduction of Panax ginseng CA Meyer. J Korean Med Sci. 2001;16(Suppl):S3.
3. Christensen LP. Ginsenosides: chemistry, biosynthesis, analysis, and potential health effects. Adv Food Nutr Res. 2008;55:1–99.
4. Court WE. Ginseng. The genus Panax2000.
5. Kim NH, Jayakodi M, Lee SC, Choi BS, Jang W, Lee J, et al. Genome and evolution of the shade-requiring medicinal herb Panax ginseng. Plant Biotechnol J. 2018;16(11):1904–17.
6. Waminal NE, Yang T-J, In J-G, Kim HH. Five-color fluorescence in situ hybridization system for karyotyping of Panax ginseng. Hortic Environ Biotechnol. 2020;61(5):869–77.
7. Kim N-H, Choi H-I, Kim KH, Jang W, Yang T-J. Evidence of genome duplication revealed by sequence analysis of multi-loci expressed sequence tag–simple sequence repeat bands in Panax ginseng Meyer. J ginseng Res. 2014;38(2):130–5.
8. Yuan Y, Bayer PE, Batley J, Edwards D. Current status of structural variation studies in plants. Plant Biotechnol J. 2021;19(11):2153–63.
9. Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet. 2019;51(6):1044–51.
10. Song J-M, Guan Z, Hu J, Guo C, Yang Z, Wang S, et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat plants. 2020;6(1):34–45.
11. Lee S-Y, Song S-Y, Lee S-H, Kim G-Y, Park J-W, Bae C-S, et al. Ginseng Berry Juice (GBJ) Regulates the Inflammation in Acute Ulcerative Mouse Models and the Major Bioactive Substances Are Ginsenosides Rb3, Rc, Rd, and Re. Nutrients. 2024;16(7):1031.
12. Zheng Y, Tan H, Chai J, Han L, Zhai C, Lee J, et al. Ginseng fruit rare saponins (GFRS) improved inflammatory response: In vitro and in vivo assessment. Fitoterapia. 2024;179:106244.
13. Guo M, Shao S, Wang D, Zhao D, Wang M. Recent progress in polysaccharides from Panax ginseng CA Meyer. Food Funct. 2021;12(2):494–518.

14. Grotewold E. The genetics and biochemistry of floral pigments. Annu Rev Plant Biol. 2006;57(1):761–80.
15. Cho W, Jang W, Shim H, Kim J, Oh Y, Park JY, et al. High-resolution genetic map and SNP chip for molecular breeding in Panax ginseng, a tetraploid medicinal plant. Hortic Res. 2024;11(12):uhae257.
16. Song Y, Zhang Y, Wang X, Yu X, Liao Y, Zhang H, et al. Telomere-to-telomere reference genome for Panax ginseng highlights the evolution of saponin biosynthesis. Hortic Res. 2024;11(6):uhae107.
17. Wang Z-H, Wang X-F, Lu T, Li M-R, Jiang P, Zhao J, et al. Reshuffling of the ancestral core-eudicot genome shaped chromatin topology and epigenetic modification in Panax. Nat Commun. 2022;13(1):1902.
18. Jayakodi M, Lu Q, Pidon H, Rabanus-Wallace MT, Bayer M, Lux T, et al. Structural variation in the pangenome of wild and domesticated barley. Nature. 2024;636(8043):654–62.
19. Feng J-W, Pidon H, Cuacos M, Lux T, Himmelbach A, Haghi R, et al. A haplotype-resolved pangenome of the barley wild relative Hordeum bulbosum. Nature. 2025;645(8080):429–38.
20. Bairu MW, Aremu AO, Van Staden J. Somaclonal variation in plants: causes and detection methods. Plant Growth Regul. 2011;63(2):147–73.
21. Krishna H, Alizadeh M, Singh D, Singh U, Chauhan N, Eftekhari M, et al. Somaclonal variations and their applications in horticultural crops improvement. 3 Biotech. 2016;6(1):54.
22. Li Y, Liu X, Cai X, Shan X, Gao R, Yang S, et al. Dihydroflavonol 4-reductase genes from Freesia hybrida play important and partially overlapping roles in the biosynthesis of flavonoids. Front Plant Sci. 2017;8:428.
23. Miosic S, Thill J, Milosevic M, Gosch C, Pober S, Molitor C, et al. Dihydroflavonol 4-reductase genes encode enzymes with contrasting substrate specificity and show divergent gene expression profiles in Fragaria species. PLoS ONE. 2014;9(11):e112707.
24. Crow T, Ta J, Nojoomi S, Aguilar-Rangel MR, Torres Rodriguez JV, Gates D, et al. Gene regulatory effects of a large chromosomal inversion in highland maize. PLoS Genet. 2020;16(12):e1009213.
25. Hu H, Scheben A, Wang J, Li F, Li C, Edwards D, et al. Unravelling inversions: Technological advances, challenges, and potential impact on crop breeding. Plant Biotechnol J. 2024;22(3):544–54.
26. Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, et al. The barley pan-genome reveals the hidden legacy of mutation breeding. Nature. 2020;588(7837):284–9.
27. Lei X-J, Zhao J, Rong J-B, Zhang M-Y, Jia W-H, Zhang J et al. Telomere-to-Telomere Genome Assembly of Yellow-Fruited Allotetraploid American Ginseng (Panax quinquefolius L.) Provides Insights into Flavonoid Biosynthesis. Hortic Res. 2025:uhaf198.
28. Hisatomi Y, Yoneda Y, Kasahara K, Inagaki Y, Iida S. DNA rearrangements at the region of the dihydroflavonol 4-reductase gene for flower pigmentation and incomplete dominance in morning glory carrying the mutable flaked mutation. Theor Appl Genet. 1997;95(4):509–15.
29. Song S, Kim C-W, Moon JS, Kim S. At least nine independent natural mutations of the DFR-A gene are responsible for appearance of yellow onions (Allium cepa L.) from red progenitors. Mol Breeding. 2014;33(1):173–86.
30. Jiang J, Huang H, Gao Q, Li Y, Xiang H, Zeng W, et al. Effects of editing DFR genes on flowers, leaves, and roots of tobacco. BMC Plant Biol. 2023;23(1):349.
31. Luo P, Ning G, Wang Z, Shen Y, Jin H, Li P, et al. Disequilibrium of flavonol synthase and dihydroflavonol-4-reductase expression associated tightly to white vs. red color flower formation in plants. Front Plant Sci. 2016;6:1257.
32. Shin DH, Choi M-G, Kang C-S, Park C-S, Choi S-B, Park Y-I. Overexpressing the wheat dihydroflavonol 4-reductase gene TaDFR increases anthocyanin accumulation in an Arabidopsis dfr mutant. Genes Genomics. 2016;38(4):333–40.
33. Wong DC, Wang Z, Perkins J, Jin X, Marsh GE, John EG, et al. The road less taken: Dihydroflavonol 4-reductase inactivation and delphinidin anthocyanin loss underpins a natural intraspecific flower colour variation. Mol Ecol. 2025;34(15):e17334.

34. Liu F, Ma Y, Xu Y. Taxifolin shows anticataractogenesis and attenuates diabetic retinopathy in STZ-diabetic rats via suppression of aldose reductase, oxidative stress, and MAPK signaling pathway. Endocrine, Metabolic & Immune Disorders-Drug Targets (Formerly Current Drug Targets-Immune. Endocr Metabolic Disorders). 2020;20(4):599–608.
35. Wang Y-H, Wang W-Y, Chang C-C, Liou K-T, Sung Y-J, Liao J-F, et al. Taxifolin ameliorates cerebral ischemia-reperfusion injury in rats through its anti-oxidative effect and modulation of NF-kappa B activation. J Biomed Sci. 2006;13(1):127–41.
36. Liao X, Xi Y, Liao B, Xiao S, Guo S, Hu H et al. Chromosome-level genome of wild-simulated Panax ginseng identifies SNP markers for germplasm and medicinal quality evaluation. J Adv Res. 2026.
37. Himmelbach A, Walde I, Mascher M, Stein N. Tethered chromosome conformation capture sequencing in Triticeae: a valuable tool for genome assembly. Bio-protocol. 2018;8(15):e2955–e.
38. Himmelbach A, Ruban A, Walde I, Šimková H, Doležel J, Hastie A, et al. Discovery of multi-megabase polymorphic inversions by chromosome conformation capture sequencing in large-genome plant species. Plant J. 2018;96(6):1309–16.
39. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18(2):170–5.
40. Marone MP, Singh HC, Pozniak CJ, Mascher M. A technical guide to TRITEX, a computational pipeline for chromosome-scale sequence assembly of plant genomes. Plant Methods. 2022;18(1):128.
41. Alonge M, Lebeigle L, Kirsche M, Aganezov S, Wang X, Lippman ZB et al. Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing. BioRxiv. 2021:2021.11. 18.469135.
42. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21(1):245.
43. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38(10):4647–54.
44. Sun H, Ding J, Piednoël M, Schneeberger K. findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. Bioinformatics. 2018;34(4):550–7.
45. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
46. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026–8.
47. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinforma Oxf Engl. 2018;34:3094–100.
48. Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics. 2016;32(19):3021–3.
49. Goel M, Sun H, Jiao W-B, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 2019;20(1):277.
50. Ou S, Su W, Liao Y, Chougule K, Agda JR, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20(1):275.
51. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Genome Res. 2024;34(5):769–77.
52. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–2.
53. Kallenborn F, Chacon A, Hundt C, Sirelkhatim H, Didi K, Cha S et al. GPU-accelerated homology search with MMseqs2. Nat Methods. 2025:1–4.

54. Wang Z, Wang W, Xie X, Wang Y, Yang Z, Peng H, et al. Dispersed emergence and protracted domestication of polyploid wheat uncovered by mosaic ancestral haploblock inference. Nat Commun. 2022;13(1):3891.
55. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997. 2013.
56. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
57. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):giab008.
58. Peniton EA Jr, Waminal NE, Yang T-J, Kim HH. Cell cycle synchronization in Panax ginseng roots for cytogenomics research. Hortic Environ Biotechnol. 2022;63(1):137–45.
59. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.
60. Meng EC, Goddard TD, Pettersen EF, Couch GS, Pearson ZJ, Morris JH, et al. UCSF ChimeraX: Tools for structure building and analysis. Protein Sci. 2023;32(11):e4792.
61. Lee B, Hwang S, Kim P-G, Ko G, Jang K, Kim S, et al. Introduction of the Korea BioData Station (K-BDS) for sharing biological data. Genomics Inf. 2023;21(1):e12.

Figures

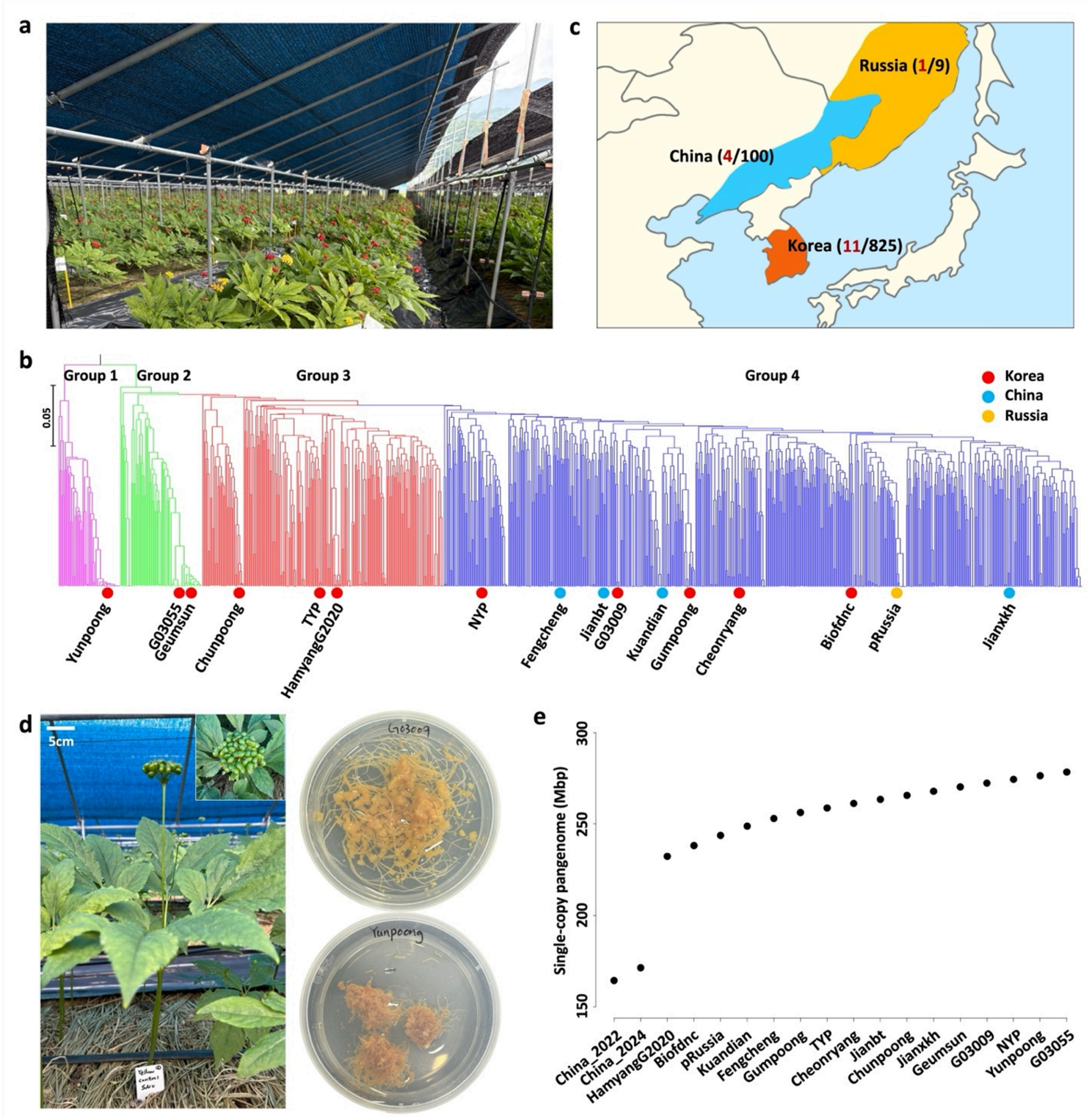

Figure 1

Selection of genotypes capturing genomic diversity in *P. ginseng*. **a,** Ginseng cultivation under an artificial shade structure. **b,** Phylogenetic clustering of the diversity panel resolves four major genetic groups (indicated by coloured branches). The 16 pangenome accessions are marked with circles coloured by country of origin. **c,** Geographic distribution of the 934 ginseng individuals genotyped using the SNP array. Numbers in parentheses indicate the total number of accessions sampled from each country, with red numbers denoting accessions selected for pangenome construction. **d,** Representative images of a ginseng plant (left) and adventitious root cultures (right) used for germplasm maintenance. **e,** Accumulation curve of single-copy sequences across the 18 ginseng genome assemblies. Assemblies were added sequentially in order of decreasing

accession-specific sequence content, illustrating continued expansion of the single-copy pangenome. The upward shift relative to the two previously published Chinese reference genomes (2022 and 2024) indicates that the pangenome approach captures a substantial number of novel sequences.

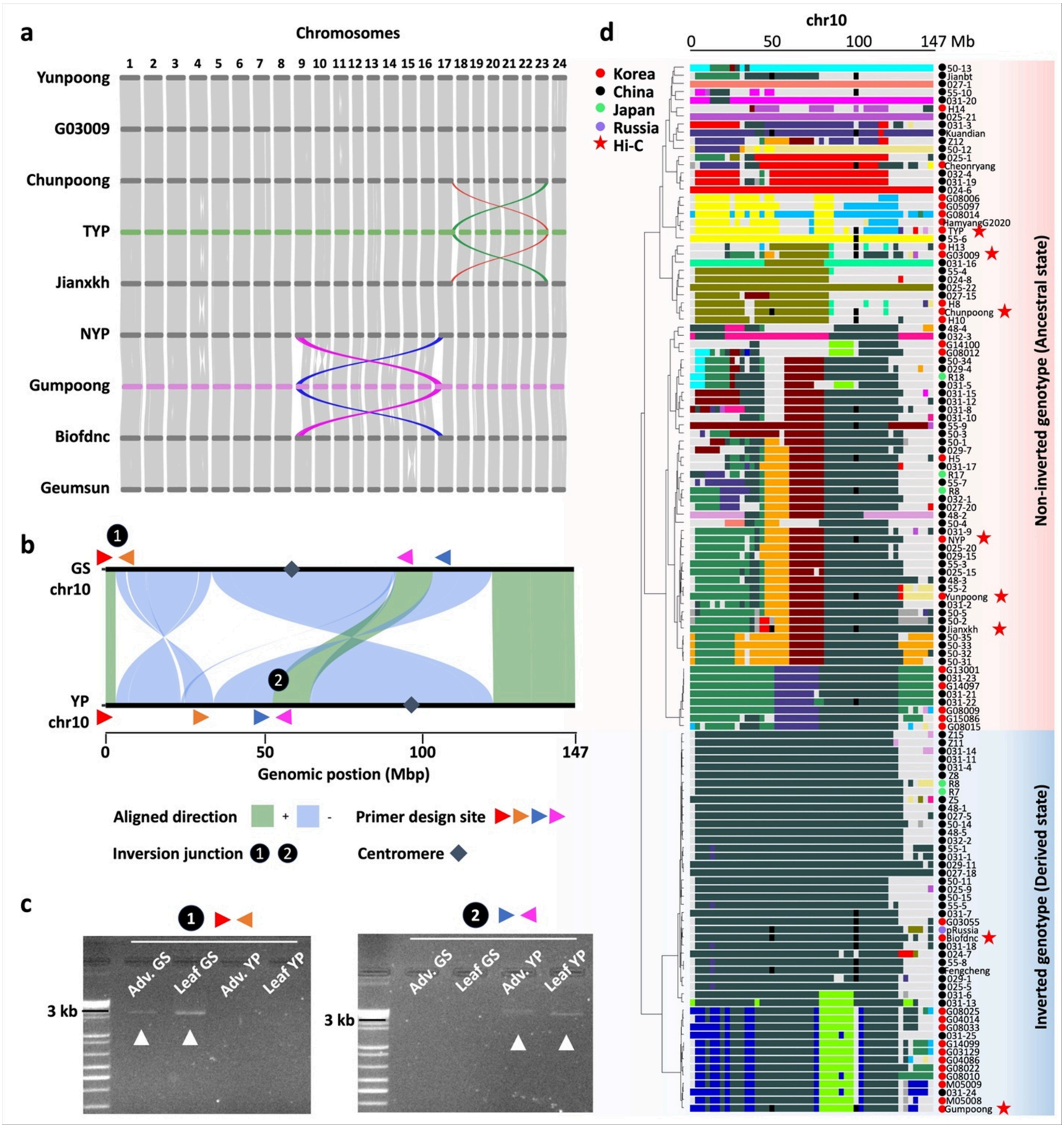

Figure 2

Large-scale chromosomal rearrangements and inversion in ginseng. **a,** Genome-wide synteny and inter-chromosomal rearrangements in ginseng. Whole-genome synteny across nine chromosome-scale ginseng assemblies. Coloured ribbons emphasize the inter-chromosomal rearrangements in specific accessions. **b,** Schematic representation of the chromosome

10 rearrangement, consisting of two consecutive inversions and also nested inversion. The positions of the two inversion junctions (□and □) used for PCR marker design are indicated. **c,** PCR validation of the chromosome 10 inversion using primers spanning two inversion junctions. Representative genotypes carrying the inverted (Geumsun, GS) and non-inverted (Yunpoong, YP) haplotypes were analysed using both adventitious root (Adv.) and leaf tissues. Identical amplification patterns across tissues confirm that the inversion represents a bona fide genomic feature rather than a tissue culture–derived artifact. **d,** Haplotype structure and inversion polymorphism on chromosome 10 inferred from SNP data. Accessions cluster into two distinct haplotype configurations corresponding to the non-inverted (ancestral) and inverted (derived) orientations, based on chromosome-level genome assemblies of nine genotypes supported by Hi-C data.

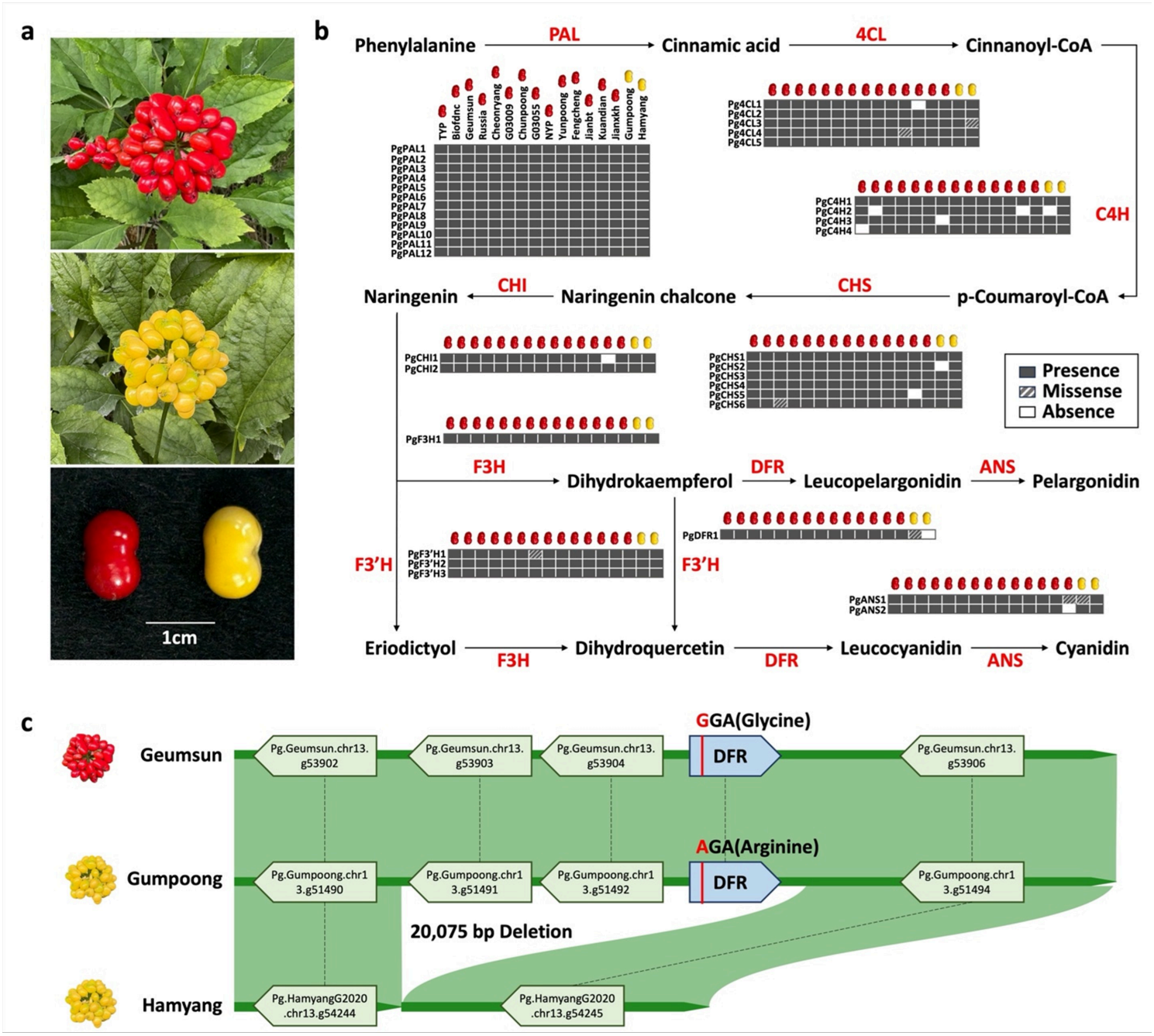

Figure 3

Flavonoid pathway variation underlying red and yellow berry colour in *P. ginseng*. **a,** Representative images of red- and yellow-berried ginseng accessions, illustrating the contrasting pigmentation phenotypes. **b,** Variation of genes in the flavonoid biosynthetic pathway across red- and yellow-berry pangenome accessions. For each enzymatic step, rows represent

individual genes and columns represent genotypes grouped by berry colour. **c,** Structural variation at the *DFR* locus associated with berry colour. Schematic representation of genomic variation at the *DFR* locus on chromosome 13. In the red-berry cultivar Geumsun, *DFR* is intact. In the yellow-berry cultivar Gumpoong, a missense mutation (Gly17Arg) alters the *DFR* coding sequence, whereas in HamyangG2020, a ~20 kb deletion removes the *DFR* locus entirely.

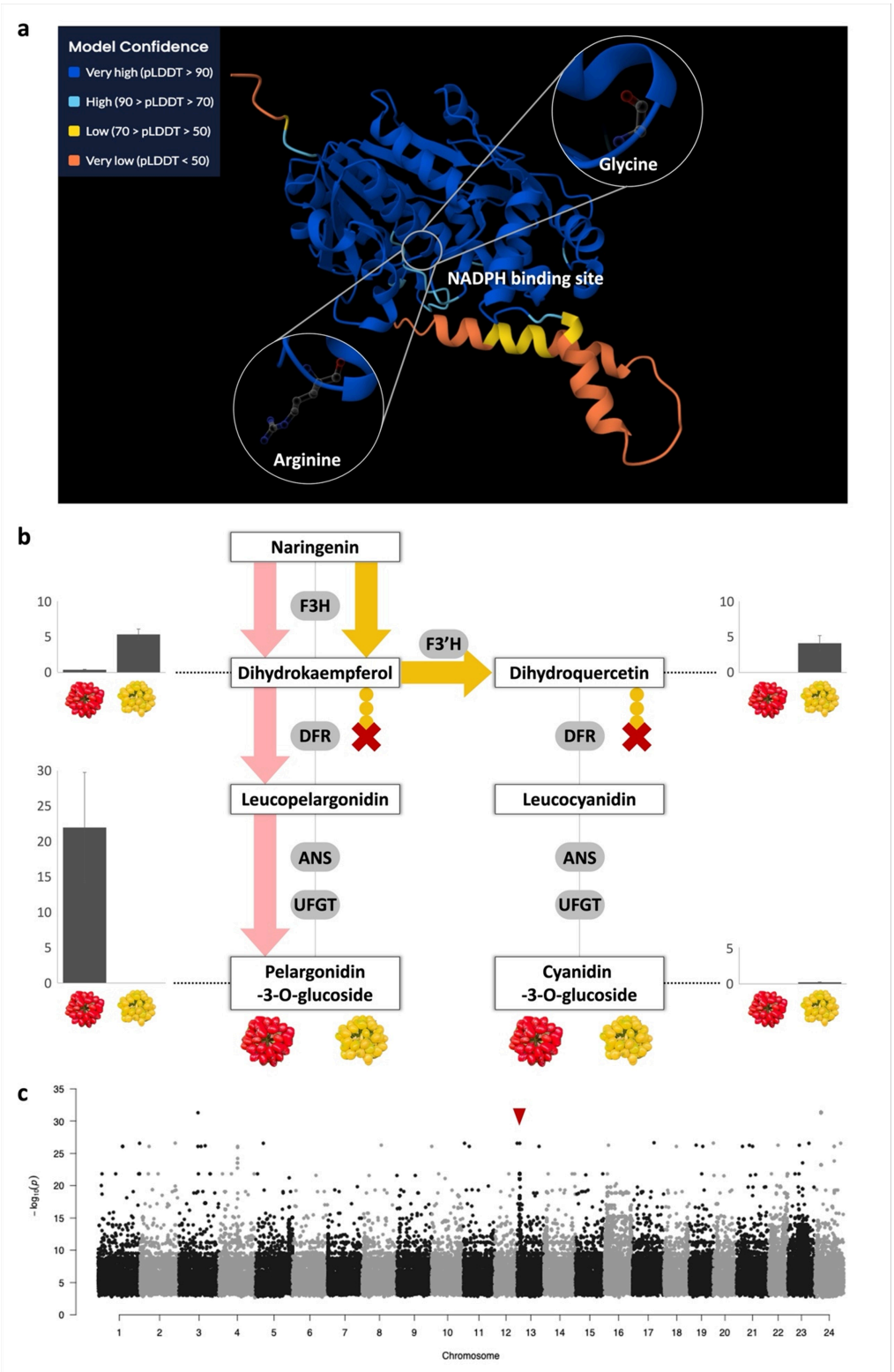

Figure 4

Structural and metabolic consequences of DFR Gly17Arg variation in ginseng berries. **a,** Predicted structural impact of the DFR Gly17Arg substitution identified in Gumpoong. Shown is the predicted three-dimensional structure of DFR, with the

NADPH-binding site highlighted. The Gly17 residue, located within the conserved NADPH-binding region, is replaced by arginine in the yellow-berry genotype. **b,** LC–MS/MS analysis revealed distinct accumulation patterns within the flavonoid pathway between red and yellow berries, with arrows depicting the principal direction of metabolite flow. **c,** K-mer–based genome-wide association analysis of berry color variation in ginseng. Each point represents a k-mer, with the y-axis indicating -$\log_{10}$(P) significance values. A continuous association signal is observed on chromosome 13 (red arrowhead), corresponding to the *DFR* locus identified as functionally disrupted in yellow-berry accessions.**Fig. 4.** Structural and metabolic consequences of DFR Gly17Arg variation in ginseng berries. **a,** Predicted structural impact of the DFR Gly17Arg substitution identified in Gumpoong. Shown is the predicted three-dimensional structure of DFR, with the NADPH-binding site highlighted. The Gly17 residue, located within the conserved NADPH-binding region, is replaced by arginine in the yellow-berry genotype. **b,** LC–MS/MS analysis revealed distinct accumulation patterns within the flavonoid pathway between red and yellow berries, with arrows depicting the principal direction of metabolite flow. **c,** K-mer–based genome-wide association analysis of berry color variation in ginseng. Each point represents a k-mer, with the y-axis indicating -$\log_{10}$(P) significance values. A continuous association signal is observed on chromosome 13 (red arrowhead), corresponding to the *DFR* locus identified as functionally disrupted in yellow-berry accessions.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx
- Additionalfile2.xlsx