Comparison of buckwheat genomes reveals the genetic basis of metabolomic divergence and ecotype differentiation

Ming He[1,2,11], Yuqi He[1,11], Kaixuan Zhang[1,11], Xiang Lu[1,3,11], Xuemei Zhang[4,11], Bin Gao[1], Yu Fan[1,3], Hui Zhao[1], Rintu Jha[1], Md. Nurul Huda[1], Yu Tang[1], Junzhen Wang[5], Weifei Yang[4], Mingli Yan[6], Jianping Cheng[3], Jingjun Ruan[3], Ehsan Dulloo[7], Zongwen Zhang[7], Milen I Georgiev[8,9], Mark A. Chapman[10], Meiliang Zhou[1*]

[1] Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, National Crop Genebank Building, Zhongguancun South Street No. 12, Haidian District, Beijing, 100081, China.

[2] Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, 100081, China

[3] College of Agriculture, Guizhou University, Guiyang 550025, China.

[4] Annoroad Gene Technology (Beijing) Co., Ltd, Beijing 100176, P. R. China.

[5] Research Station of Alpine Crop, Xichang Institute of Agricultural Sciences, Liangshan 616150, Sichuan, China.

[6] Crop Research Institute, Hunan Academy of Agricultural Sciences, Changsha, 410125. China

[7] The Alliance of Bioversity International and CIAT, Via di San Domenico, 1 00153 Rome, Italy

[8] Group of Plant Cell Biotechnology and Metabolomics, The Stephan Angeloff Institute of Microbiology, Bulgarian Academy of Sciences, 4002 Plovdiv, Bulgaria.

[9] Center of Plant Systems Biology and Biotechnology, 4002 Plovdiv, Bulgaria.

[10] Biological Sciences, University of Southampton, Life Sciences Building 85, Highfield Campus, Southampton, SO17 1BJ, UK.

[11] These authors contributed equally to this work.

[*] **Correspondence and requests for materials should be addressed to M. Z. (email: zhoumeiliang@caas.cn).**

Summary

- Golden buckwheat (*Fagopyrum dibotrys* or *F. cymosum*) and Tartary buckwheat (*F. tataricum*) belong to the Polygonaceae and the *Fagopyrum* genus is rich in flavonoids. Golden buckwheat is a wild relative of Tartary buckwheat, yet golden buckwheat is a traditional Chinese herbal medicine and Tartary buckwheat is a food crop. The genetic basis of adaptive divergence between these two buckwheats is poorly understood.
- Here, we assembled a high-quality chromosome-level genome of golden buckwheat and found a 1-to-1 syntenic relationship with the chromosomes of Tartary buckwheat. Two large inversions were identified that differentiate golden buckwheat and Tartary buckwheat.
- Metabolomic and genetic comparisons of golden buckwheat and Tartary buckwheat indicate an amplified copy number of *FdCHI*, *FdF3H*, *FdDFR*, and *FdLAR* gene families in golden buckwheat, and a parallel increase in medicinal flavonoid content. Resequencing of 34 wild golden buckwheat accessions across the two morphologically distinct ecotypes identified candidate genes, including *FdMYB44* and *FdCRF4*, putatively involved in flavonoid accumulation and differentiation of plant architecture, respectively.
- Our comparative genomic study provides abundant genomic resources of genomic divergent variation to improve buckwheat with excellent nutritional and medicinal value.

Keywords: golden buckwheat, genome, comparative genomics, flavonoids biosynthesis, differentiation

Introduction

Golden buckwheat (*Fagopyrum dibotrys* or *F. cymosum*) is a famous traditional Chinese herbal medicine with edible seeds and leaves, rich in flavonoids with antioxidant and antidiabetic activity (Li et al., 2021; Jing et al., 2016). According to bencaoshiyi, an ancient monograph on Chinese herbal medicine, the rhizome of

golden buckwheat was used as folk medicine to treat those with lung diseases, bacterial dysentery, abdominal pain and rheumatism. Recent studies report its antioxidant, anti-inflammatory, analgesic, antipyretic and anti-proliferation effects (Chen and Li, 2016). WeiMaiNing (WMN), a non-toxic Chinese herbal medicine prepared from dried golden buckwheat rhizomes, has been used to treat different types of cancer (Chan, 2003; Ke et al., 2021).

Several studies have shown that the rhizome of golden buckwheat contains multiple bioactive substances that are effective in the treatment of inflammation, cancer and diabetes (Wang et al., 2005). Among them, proanthocyanidins (condensed tannins) are the main active ingredients of WMN, with strong antioxidant, anti-tumor, and immunomodulatory properties and participate in tackling fatal diseases (Li et al., 2021; Chen and Li, 2016). Epicatechin (a monomer of proanthocyanidins), is a pharmaceutically relevant flavonoid with significant antioxidant activity, and is considered as biomarker for evaluation of golden buckwheat quality according to the Chinese pharmacopoeia (Chen and Li, 2016; Musial et al., 2020). Quercetin is a flavanol with strong antioxidant activity, and largely utilized to treat a variety of diseases like inflammation, depression, diabetes, cancer and cardiovascular disorders (Xu et al., 2019). The extraction of these compounds from golden buckwheat appears a promising alternative for treatment of multiple ailments. Thus, it is of particular importance to breed golden buckwheat varieties with increased content of value-added molecules.

Golden buckwheat is the only outbreeding perennial species in the cymosum group of *Fagopyrum*, and recognized as the primitive type and wild ancestor of the closely related cultivated species, Tartary buckwheat and common buckwheat (Ohnishi, 1998; Cheng et al., 2020). Due to its relatively large seeds and abundant flavonoids, golden buckwheat is considered an ideal wild relative for cultivated buckwheat improvement (Jing et al., 2016; Wang et al., 2018). Spontaneous hybridization between golden buckwheat and Tartary and common buckwheat, demonstrate that cultivated buckwheat improvement can take place through interspecific crossing (Cheng et al., 2020; Chen, 2016). Here, we report a high-quality genome assembly of golden

buckwheat and characterize the mechanism regulating flavonoid biosynthesis in golden buckwheat. Resequencing of 34 wild accessions identified two ecotype groups and quantified their genomic diversity, identifying further candidate genes underlying important traits. Our comprehensive genomic analyses provide insights into the evolution of flavonoid synthesis genes in buckwheat and will hopefully facilitate biological discovery and pharmacological plant genetic improvement in golden buckwheat as well as other plant species.

Materials and Methods

Plant Materials and Phenotyping

The golden buckwheat (*Fagopyrum dibotrys* or *F. cymosum*) accession assembled in this study was collected in Luoji Mountain (Sichuan province, China) and then transplanted in the Liangshan Autonomous Prefecture, Sichuan province, and in Changping District, Beijing. The sample was a diploid accession analyzed by flow cytometry analysis (Fig. S2). The resequencing accessions were collected from Sichuan, Yunnan and Guizhou province in China. All the samples were transplanted and propagated in Liangshan Autonomous Prefecture. Young leaves were used for flow cytometry analysis. For phenotyping, three plants from each accession were used for the measurement of plant height, internode number, stem diameter, basal branch number, branch number, and 1000-grain weight in the middle of July in Liangshan Autonomous Prefecture.

Whole Genome Sequencing

Genomic DNA was extracted from young leaves using a CTAB extraction method (Doyle and Doyle, 1987). The short paired-end reads were sequenced on MGISEQ-2000 platform. Genomic DNA was randomly sheared by ultrasonic high-performance processing system (Covaris). The 350 bp fragments were screened for terminal repair to single-strand isolation and cyclization treatment. The original data obtained by sequencing was converted into raw sequence data (RAW reads) by Base Calling, and the data stored in FASTQ file format. The raw data was filtered using SOAPnuke1.5.6 (https://github.com/BGI-flexlab/SOAPnuke) with parameters:

-n 0.01 -l 20 -q 0.1 -i -Q 2 -G -M 2 -A 0.5 –d, generating 153 Gb clean reads finally. High molecular weight genomic DNA were extracted, and 20 kb fragments were selected with BluePippin. Nanopore libraries were constructed with Library Perpare Kit provided by Oxford Nanopore Technology, and sequenced on ONT Promethion platform. Finally, 110 Gb long reads were generated, about 101X coverage of golden buckwheat genome.

Estimation of the Genome Size and Heterozygosity

The golden buckwheat genome size was estimated with k-mer frequency distribution. K-mers with 17-31 bp were counted with jellyfish (V2.1.4, Marcais and Kingsford, 2011), and then GenomeScope (http://qb.cshl.edu/genomescope/) was used to estimate genome size and heterozygosity based on the k-mer frequency with default parameters.

Genome Assembly and Quality Assessments

The 110 Gb Nanopore long reads were self-corrected and assembled with NECAT (https://github.com/xiaochuanle/NECAT). The assembled contigs were polished through three rounds of long reads by Racon (https://github.com/lbcb-sci/racon). To improve the base accuracy of the contigs, Pilon v1.22 (https://github.com/broadinstitute/pilon/) (Walker et al., 2014) was used to polish contigs using short reads with default parameters. The primary assembled genome size was 1.59 Gb, larger than estimated genome size, 1.08 Gb. Purge Haplotigs was employed to remove the redundant sequences caused by heterozygote with default parameters (Roach et al., 2018). The completeness of assembled genome was assessed by BUSCO (Simao et al., 2015), with the library of embryophyta_odb10.

Hi-C Sequencing and Scaffolding

Young leaves were collected for Hi-C library construction. The sample was fixed with formaldehyde, lysed, biotin-labeled, and purified (Lieberman-Aiden et al., 2009; Van Berkum et al., 2010), and then sent for sequencing on an MGISEQ-2000 platform. We removed the adapters and low-quality bases using fastp (v0.12.6; Chen et al., 2018) with default parameters and then aligned the Hi-C data to the assembled contigs using BWA-MEM (Houtgast et al., 2018). Hi-C-Pro pipeline (Van Berkum et al., 2010)

were used to identify the Valid Interaction Pairs with default parameters. The contigs were anchored into pseudomolecule with 3d-dna (v180922) (Dudchenko et al., 2017; https://github.com/theaidenlab/3d-dna). In total, 1.08 Gb of sequence was anchored into 8 pseudomolecules. The heatmap for Hi-C interaction was processed using 3d-dna visualize module.

Genome Annotation

For repeat analysis, we used *ab initio* prediction and homology-based approach. RepeatMasker (v4.0.7; http://www.repeatmasker.org) was used to identify homologous sequences based on the RepBase (v21.12) library (http://www.girinst.org /repbase). LTR_FINDER (v1.06; Xu and Wang, 2007; http://tlife.fudan. edu.cn/ltr_finder/) was used to indentify LTR retrotransposon, and RepeatModeler (http://www.repeatmasker.org/RepeatModeler/) were used to build *ab initio* prediction repeat library. LTRs and the de novo repeat library were combined, and subsequently used to screen the golden buckwheat genome using RepeatMasker.

For gene structure annotation, we combined RNA-seq evidence, *ab initio* and homology-based methods. The *ab initio* prediction was performed using Augustus (Stanke and Waack, 2003) (version 3.2.3; http://bioinf.unigreifswald.de/augustus/) with parameters: -strand = both –genemodel = partial -gff3 = on –species = maize5. For the RNA-seq used in genome annotation, we collected roots, stems, tubers, and leaves, obtaining 229,501,400 clean reads. RNA-seq reads were mapped to the genome by HISAT2 (v2.1) with default parameters (Kim et al., 2015), and the transcripts were assembled from the read alignments using Cufflinks (Trapnell et al., 2010). Proteins from Tatary buckwheat (Zhang et al., 2017) and common buckwheat (http://buckwheat.kazusa.or.jp/) were used as protein homology evidence. Finally, the high-confidence gene models in golden buckwheat genome were predicted using MAKER pipline (version 2.31.10) (Campbell et al., 2014). The functional annotation of the predicted proteins was performed by eggNOG-mapper (Huerta-Cepas et al., 2019; Huerta-Cepas et al., 2017) and went through BLASTP against different database, containing GO, KEGG, DOG, NR, and swissprot databases.

Phylogenetic Analysis and Gene Family Analysis

For the evolutionary history analysis of golden buckwheat, nine Eudicotyledons, comprising Tartary buckwheat (http://www.mbkbase.org/Pinku1/), *Beta vulgaris* (https://www.ncbi.nlm.nih.gov/genome/?term=Beta+vulgaris), *Beta patula* (http://bvseq.boku.ac.at/ Genome/Download/Bpat/), *Carica papaya* (Phytozome v12.1), *Solanum lycopersicum* (Phytozome v12.1), *Solanum tuberosum* (ftp://ftp.ncbi.nlm.nih.gov/genomes/Solanum_tuberosum/), *Arabidopsis thaliana* (Phytozome v12.1), *Populus trichocarpa* (Phytozome v12.1), *Vitis vinifera* (Phytozome v12.1), *Prunus mume* (https://github.com/lileiting/prunusmumegenome), and two monocotyledons, *Oryza sativa* (https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=Osativa) and *Zea mays* (https://www.ncbi.nlm.nih.gov/genome/?term=Zea+mays) were used for orthogroup and clustering using Orthofinder v2.3.3 (Emms and Kelly, 2015) with default parameters. We constructed a phylogenetic tree with the single copy orthologs using the maximum likelihood approach in RaxML (v8; Stamatakis, 2014) with JTT+I+G model of sequence evolution and 1,000 bootstrap replicates. Species divergence times were estimated by PAML (v4.5) and calibrated from TimeTree database (http://www.timetree.org/). Expanded and contracted gene families were determined by Computational Analysis of gene Family Evolution (CAFÉ v4.2.1) (De Bie et al., 2006). Transcriptions factors were predicted by comparing the 12 species to PlantTFDB (http://planttfdb.cbi.pku.edu.cn/) with iTAK (Zheng et al., 2016).

Genomic Comparisons and Whole Genome Duplication Analysis

The genomes of golden buckwheat and Tartary buckwheat were screened to identify syntenic blocks by McscanX (https://github.com/tanghaibao/jcvi/wiki/MCscan, Python-version; Tang et al., 2008) with default parameters. For a block to be identified it must contain at least five collinear gene pairs. The synonymous substitution values (Ks) were calculated using KaKs_calculator (version 2.0) with the PAML yn00 model (http://abacus.gene.ucl.ac.uk/software/paml.html). The self-collinearity of golden buckwheat was performed by jcvi (v0.84, https://github.com/tanghaibao/jcvi).

Chromosome Structural Variation Analysis

For the inversion analysis, we aligned the genomes of golden buckwheat and Tartary buckwheat using minimap2 (v2.17) with parameters: -ax, asm5, --eqx. The variations were measured using syri with parameters: -k, -F, S and annotated using ANNOVAR (Goel et al., 2019; Wang et al., 2010).

For the insertion, deletion and duplication analysis, the genome alignment was performed by nucmerV4.0 (Marcais et al., 2018) and structural variations were identified using Assemblytics (Nattestad and Schatz, 2016) with parameters: unique length required: 500, min size: 50, max size: 1000000 and annotated using ANNOVAR.

RNA Sequencing Data Analysis

Seedlings were grown on MS plates in the dark, and in red light and blue light for ten days. Three replicates of each were collected and RNA-seq carried out using an Illumina sequencing platform by Annoroad Company. The adaptor and low-quality sequences were removed by Trimmomatic (v0.33, default paramaters) and leaving a total of 341,721,934 clean reads which were mapped to the genome using HISAT2 v2.1.0 (Kim et al., 2015). The mapped reads referring to each transcript were assembled and merged by Cufflinks v2.2.1 (Trapnell et al., 2010). FPKM values were calculated by Htseq-count v0.6.0 (Anders et al., 2015), with FPKM > 0.1 was considered as expressed gene. Differentially expressed genes (DEGs) were identified using DESeq2 v1.20.0 (Love et al., 2014), with $q \leq 0.05$ and $|\log2\_ratio| \geq 1$.

For the different tissue RNA-seq, we collected old leaves, young leaves, flowers, steam, rhizomes, and roots from plants growing in Changping base, Beijing in mid-July and RNA-seq performed as above, obtaining 856,753,554 clean reads. The reads were analyzed similar to the above description. The WGCNA was performed as reported in Kang et al., 2021. The co-expression network was constructed with a soft-thresholding power of 19 and default parameters. The minimum module size and the minimum height for merging modules were set to 30 and 0.27, respectively. The co-expression network analysis was visualized using Cytoscape.

Sequencing and Variant Calling

Genomic DNA was used to construct a sequencing library according to vendor-provided instructions (Illumina) as reported (Zhang et al., 2021). The library was sequenced on Illumina NovaSeq 6000 platform by Annoroad in paired-end mode and 350bp insert size. Using the manufacturer's adapter sequences, we used Trimmomatic v0.33 to remove adapters to generate clean reads.

The clean paired-end reads were mapped to the golden buckwheat genome using the Burrows-Wheeler Aligner program (BWA 0.7.5a; Houtgast et al., 2018) with default parameters. We removed the low mapping quality reads (MQ < 30) according to mapping coordinates in samtools (0.1.19) (Li et al., 2009). SNPs and indels (1-50 bp) were called using GATK (v3.4-46-gbc02625) UnifiedGenotyper module for diploids with -stand_call_conf 50-stand_emit_conf 10-dcov 1000 (McKenna et al., 2010). For SNPs we used the population filter: (a) QUAL > 30.0; (b) QD > 5.0; (c) FS < 60.0; (d) MQ 0 ≥ 4 && ((MQ0/(1.0*DP)) > 0.1); (e) DP > 5. We chose Hard Filtering instead of the Variant Recalibration (VQSR) method to filter our variants callset, following GATK best practice. All SNPs and indels were assigned to specific genomic coordinates and the corresponding genes using ANNOVAR (Wang et al., 2010) based on golden buckwheat genome annotations. To ensure that SNPs called from the whole-genome resequencing data are reasonable, site frequency spectrum (SFS) (Xue and Hickerson, 2015) was applied with the callset at population level based on MAF > 0.05 and missing rate < 0.1.

Phylogenetic Analysis

The golden buckwheat resequencing panel neighbor-joining tree was constructed with 300,000 filtered SNPs by TreeBest 1.9.2 with 100 bootstrap replicates (Li et al., 2006). The principal component analysis was carried out by SNPRelate (1.18.1) with default parameters (Zheng et al., 2012). The eigenvectors were extracted to create a plot in two dimensions.

Identification of Genomic Differentiation (F_{ST})

Genetic differentiation (F_{ST}) was determined between the decumbent group and the erect group in 200 kb window and a step size of 10 kb using PopGenome (Pfeifer et al., 2013). Windows corresponding to the top 5% of F_{ST} values were assumed to be

those with the strongest differentiation and explored further.

Metabolome Analysis

The samples of golden buckwheat root, golden buckwheat tuber and Tartary buckwheat root were collected from mature plants. The samples were ground into a powder using a mixer mill (MM 400, Retsch). The sample powder (0.2 g) was extracted with 10 ml of 80% methanol. The mixture was treated with ultrasound for 45 min and centrifuged at 10,000 g for 10 min. The extracts were diluted with the initial mobile phase and passed through a 0.22 μm hydrophobic PTFE needle filter. The extracts were analyzed using the HPLC-qToF-MS system (Agilent G6500 Series HPLC-QTOF, Agilent MassHunter Workstation Date Acquisition, and Qualitative Analysis B.07.00). The chromatographic conditions were as follows: water (0.1% formic acid), methanol solution (0.1% formic acid); flow rate, 0.5 ml/min; temperature, 40 ℃; injection volume, 5 μl; and detection wavelength, 350 nm. The Electrospray ionization (ESI) Turbo Ionspray interface operated in positive ion mode was used to perform spectrometry analysis. The ESI source conditions were as follows: gas, N2; gas temperature, 300 ℃; gas flow rate, 8.0 L/min; curtain gas, 35 psi; capillary voltage, 3.5kV; capillary outlet voltage, 175V; cone hole voltage, 65V; collision energy, 10V, 30V, 50V; scanning range of MS, m/z 100 ~ 1000. The Qualitative Analysis B.07.00 was used to control instruments and collect data. The software EZinfo V3.0.1 was used for data processing and multivariate analysis.

Measurement of the Flavonoids Content

As reported (Zhang et al., 2021), seedlings from the different light treatments were ground and filtered after dried at 105 ℃. The powder (0.2 g) was extracted with ultrasound for 45 min in 20 ml of 80% methanol at 50 ℃. The solution was filtered through a 0.22 um hydrophobic PTFE needle filter and analyzed using HPLC (Agilent G6500 Series HPLC-QTOF). The CAS number of kaempferol, quercetin, rutin, epicatechin, and cyaniding was 520-18-3, 117-39-5, 153-18-4, 490-46-0, and 528-58-5, respectively. The contents were calculated by comparing the HPLC peak area with authentic standards (Sigma-Aldrich).

Transgenic Hairy Root Culture

The *FdFRS1* CDS sequence was cloned and constructed into the pCAMBIA 1307 vector with the following primers: F: 5'-acgggggactcttgaccatggATGGAAAATCAAC CCGAAATAGAC-3' and R: 5'-aagttcttctcctttactagtTCAATGTTTCAGTATCAAGTG AACGT-3'. The vectors were transformed into *Agrobacterium* A4 to generate transgenic hairy roots as reported (Zhou et al., 2017).

Results

Genome Assembly and Annotation of Golden Buckwheat

We sequenced and assembled a diploid accession of golden buckwheat from Luoji Mountain (Sichuan province, China) (**Fig. 1a-f**), using a combination of MGISEQ-2000, Oxford Nanopore Technology and Hi-C technology (**Table S1**). The estimated genome size of golden buckwheat was 1.08 Gb and heterozygous ratio was 1.02% (**Table S2**). For high heterozygous, the primary assembled genome size was 1.59 Gb (**Table S3**). After removed the redundant sequences, the final genome size was 1.08 Gb, with a contig N50 of 2.84 Mb, matching the genome size based on flow cytometry analysis (1.07 Gb) and k-mer analysis (1.08 Gb; **Fig. S1-4 and Table S3**). The contigs was anchored on 8 pseudomolecules with HiC sequencing data (**Table S4**). BUSCO assessment indicated that about 92.9% of the core conserved plant genes were complete in assembled golden buckwheat using embryophyta_odb10 database (**Table S5**). A total of 38,919 genes were predicted using RNA-seq, *ab initio* prediction and homology-based methods (**Fig. 1g and Table S6-8**). We evaluated the completeness of predicted genes with embryophyta_odb10, and 91.1% BUSCOs were complete (**Table S9**).

Repetitive sequences accounted for the largest proportion of the golden buckwheat genome; in total, 0.74 Gb (68.21%) of the genome was identified as repeats (**Table S10**). The long terminal repeat retrotransposons (LTR-RTs) are the most abundant repeat type, occupying 48.35% of the genome. The total length of the largest LTR-RT superfamilies, *Gypsy* and *Copia*, was about 0.5 Gb, or 46.23% of the genome. The frequency distribution of the *Gypsy* and *Copia* insertion times showed a burst at 0.8~1.0 Mya, compared with a more scattered distribution over the past 1 Mya in

Tartary buckwheat (**Fig. S5**). The golden buckwheat genome is more than twice the size of the closely related Tartary buckwheat genome (0.48 Gb) (Zhang et al., 2017), but the number of genes is similar (38,919 vs. 33,366 in Tartary buckwheat) collectively indicating that the large-scale amplification of repetitive sequences is the main reason for the difference in genome size between golden buckwheat and Tartary buckwheat.

Genomic Structure Variation and Whole-genome Duplication

Golden buckwheat is a wild relative of Tartary buckwheat (Ohnishi and Zhou, 2018), despite the significantly different genome sizes. We aligned the two genomes and detected a 1-to-1 syntenic relationship at the whole-chromosome level, including 114 blocks and 20,676 collinear genes, which further confirmed that golden buckwheat has a diploid genome (**Fig. 2a**).

The Tartary buckwheat genome showed more deletions compared with golden buckwheat genome since their divergence (**Fig. S6**), further consistent with the larger genome size of golden buckwheat. Two larger inversions were identified in chromosome (Chr) 3 (29.5 Mb) and Chr4 (23.8 Mb), containing 1,673 genes and 759 genes, respectively (**Fig. 2b and Fig. S7-8, and Table S11-12**). The GO enrichment analysis of the 200 genes near the breakpoint regions of Chr3 showed that the GO terms of 'shoot apical meristem development', 'carbohydrate metabolic process' and 'metabolic process' were significantly enriched (**Fig. 2c**). Nine flavonoid 3',5'-hydroxylases were identified in the genes annotated with the term 'metabolic process' (**Table S12**), which play a crucial role in anthocyanin and catechin biosynthesis (Seitz et al., 2015; Jin et al., 2018). Genes near the breakpoint regions of Chr4 were mainly enriched in the GO terms related to the cell wall biosynthesis, including pectin and polysaccharide biosynthesis, and cell wall organization (**Fig. S9**).

Whole-genome duplication (WGD) is one of the most important forces driving the expansion of genome size. Therefore, we analyzed the paralogous syntenic genes based on self-comparison and found 2,953 and 5,184 collinear genes in golden buckwheat and Tartary buckwheat, respectively (**Table S13**). The Ks peak for putative

orthologous gene pairs indicates the speciation event and genes corresponding to this Ks peak (i.e., gene pairs with Ks<0.3, **Table S13**) were significantly enriched in metabolic and development processes, including lipid metabolism, flavonoid metabolism, fruit development, as well as lateral root formation (**Fig. 2d and e**). Only one Ks peak of the paralogous syntenic gene pairs was identified within both golden buckwheat and Tartary buckwheat, at Ks = ~0.99. This suggests an ancient WGD event at around 71.1 Mya predating the speciation event, without recent WGDs (**Fig. 2d**), which was supported by self-synteny analysis (**Fig. S10**).

We analyzed the different origins of gene duplicates and found a similar distribution of duplicate type in both the golden buckwheat and Tartary buckwheat genomes, with three main types of dispersed duplication, WGD, and transposed duplication, although golden buckwheat showed a higher proportion of dispersed duplicate type (**Fig. S11**). Gene duplicates that originated through dispersed duplication, proximal duplication, and tandem duplication were significantly enriched in GO terms related to flavonoid biosynthesis process (**Fig. S12-14**). Tandem gene duplicates were also enriched in cell wall biosynthesis, as reported in *Miscanthus lutarioriparius* (Miao et al., 2021). Transposed duplicates were enriched in GO terms related to protein phosphorylation modification and response to biotic and abiotic stress (**Fig. S15**). The WGD genes were mainly enriched in abiotic and biotic stress and development-related terms (**Fig. S16**).

Genome Evolution of Golden Buckwheat

To reveal the evolutionary history of the golden buckwheat genome, we analyzed gene orthologues among golden buckwheat, Tartary buckwheat, nine other Eudicotyledons, and two monocotyledons, and characterized 37,025 genes into 14,549 families (**Table S14**). Through comparing the gene families among Caryophyllales (golden buckwheat, Tartary buckwheat, *Beta vulgaris*, and *B. patula*), 10,338 shared gene families and 7,277 one-to-one orthogroup gene families between golden buckwheat and Tartary buckwheat were identified (**Table S15**). Comparing golden buckwheat to Tartary buckwheat revealed 8 plant self-incompatibility protein

S1s unique to golden buckwheat (**Table S16**), coinciding with the self-incompatibility of golden buckwheat (Zhou et al., 2018). We constructed a phylogenetic tree with 344 single-copy genes shared among the 13 species, revealing that Caryophyllales (buckwheat and sugar beet) were grouped on a branch outside of Brassicales and Rosales and diverged from other dicotyledons at about 138 Mya (**Fig. 3a**).

The evolution of gene families can play important roles in plant phenotypic diversification. In golden buckwheat, we identified 739 expanded and 528 contracted gene families from the 14,549 gene families (**Table S17**). The GO enrichment analysis of the expanded gene families revealed that they were mainly enriched in regulation of cellular, biosynthetic, and metabolic processes (**Fig. 3b**). Among the genes of expanded gene families in golden buckwheat, we identified 85 cytochrome P450 (CYP) enzymes (FdCYPs; **Table S18**), which are involved in secondary metabolism and hormone metabolism (Bak et al., 2011). We further found that 54 out of these 85 FdCYPs possessed a flavonoid 3'-monooxygenase domain (**Fig. S17 and Table S18**) and 12 were related to Tartary buckwheat FtF3’5’H (**Fig. 3c**), key factors regulating anthocyanin and catechin biosynthesis in other plants (Seitz et al., 2015; Jin et al., 2018). Some of these 54 FdCYPs were differently expressed under red light and blue light treatments, consistent with the change of flavonoids accumulation (**Fig. 3d-3f**).

Several families of MYB-related genes were expanded, which comprised 52 golden buckwheat *FdMYB* genes. Twenty-four of these clustered with a single Arabidopsis MYB, AtMYB91 (**Fig. S18**), which plays critical role in growth immune response and asymmetric leaf formation (Byrne et al., 2000; Nurmberg et al., 2007). Among the expanded gene families, we also found 55 *FdLRRs* (Leucine rich repeats) and 11 *FdTIFYs* (transcription factors possessing a TIFY motif) (**Table S17**), both of which play crucial roles in plant growth and development process and biotic and abiotic stress processes (Liu et al., 2017; Xie et al., 2019).

We found the copy number of the *FAR* transcription factor genes to be expanded in golden buckwheat compared with that of Tartary buckwheat (**Table S19**), which play critical role in plant growth and development, and metabolic processes

(Fernandez-Calvo et al., 2020; Lin and Wang, 2004). We found that 13 *FdFARs* were clustered on a branch with the single Arabidopsis *AtFRS1* (**Fig. 3g**), suggesting this subgroup of *FAR* genes has a significant role in golden buckwheat. Overexpression of one of these, *FD04G006530* (named *FdFRS1*), in hairy root culture elevated the content of rutin (**Fig. 3h and Fig. S19 and 20**), an important flavonoid with potential health benefits (Ganeshpurkar and Saluja, 2017).

Flavonoid Biosynthesis Genes and Metabonomic Analysis of Golden Buckwheat and Tartary Buckwheat

Golden buckwheat and Tartary buckwheat both accumulate value-added molecules, especially flavonoids, such as procyanidins, epicatechin, quercetin, and rutin (Joshi et al., 2020; Zhang et al., 2021). However, only the rhizomes of the golden buckwheat, and not the root of the Tartary buckwheat, are used in traditional Chinese medicine. To explore the genetic basis, we performed metabolomics analysis and compared the flavonoid biosynthesis genes between golden buckwheat and Tartary buckwheat.

In total, 491 annotated metabolites were tentatively identified in golden buckwheat root (GBR), golden buckwheat tuber (GBT) and Tartary buckwheat root (TBR; **Fig. 4a and Table S20**). A principal component analysis of all the metabolites revealed independent grouping of the GBR, GBT and TBR samples (**Fig. 4b**), which was further supported by correlation analysis (**Fig. S21**). We found that the content of 304, 331, and 357 metabolites were significantly different between TBR and GBT, TBR and GBR, and GBT and GBR ($p < 0.05$), respectively (**Fig. 4c, Fig. S22 and Table S20**) which included the flavonoids, rutin, quercetin, astragalin, jaceosidin, procyanidins, and catechin. Overall, flavonoid content was higher in golden buckwheat than that in Tartary buckwheat. The content of rutin, quercetin, and jaceosidin were highest in GBT, and astragalin, procyanidin B1, procyanidin A2, and catechin were highest in GBR (**Fig. 4d-4g and Fig. S23**). We found significant differences in the content of specific flavonoids between GBR and GBT, the GBT accumulated more rutin, quercetin, jaceosidin, taxifolin, and catechol than GBR, which appeared richer in proanthocyanidins (**Fig. S22 and 23**).

Based on the KEGG database and previous reports (Zhang et al., 2017; Katsu et al., 2017; Liu et al., 2013), we examined the generalized flavonoid metabolic pathway (**Fig. 4h**). By combing gene functional annotation and homology searching, 169 genes belonging to 16 gene families involved in flavonoid biosynthesis in golden buckwheat were identified, which was almost 2.5-fold the number of genes in Tartary buckwheat (71 genes; **Table S21**). Almost all of the genes related to flavonoid biosynthesis displayed variable expression levels across six tissues (**Fig. 4h**) and under different light treatments (**Fig. S24**).

Examining gene copy numbers, the *FdCHI*, *FdF3H*, *FdDFR*, and *FdLAR* families showed the greatest gene family expansion, with six times more copies in golden buckwheat than Tartary buckwheat (**Table S21**), which were involved in the accumulation of proanthocyanidin, quercetin, anthocyanin and catechin (Jiang et al., 2015; Reuben et al., 2013; He et al., 2021). Expansion of *FdCHI*, *FdF3H*, *FdDFR*, and *FdLAR* families might be pivotal for the higher content of procyanidins and catechin in golden buckwheat. Analyzing their positions in the golden buckwheat genome revealed that 93 out of these 169 genes (55.03%) were in clusters separated by at most 500 kb, likely resulting from tandem duplication, compared with 29 out of 71 genes (40.85%) in Tartary buckwheat (**Fig. S25 and 26**).

To elucidate co-expression networks of genes related to flavonoid biosynthesis in golden buckwheat, we performed weighted correlation network analysis with the differentially expressed genes across six tissues and obtained 14 modules (**Fig. S27 and 28**). Some key flavonoid synthesis related genes were grouped into two modules (grey60 and brown; **Table S22**), which were closely associated with tuber and root, respectively (**Fig. S29**). We carried out homology searching with *Arabidopsis* and identified some candidate genes that regulate flavonoids accumulation in the two modules (**Fig. S30**), including *FD03G014500* (*REF1*, involved in formation of ferulic acid and sinapic acid), *FD05G020950* (*COMT*, encoding a flavonol 3-*O*-methyltransferase), *FD06G015940* (*NINJA*, involved in glucosinolate biosynthesis), *FD05G043040* (*JAR1*, encoding a jasmonate-amido synthetase), *FD04G011550* (*MYC4*, activating jasmonate response), *FD02G039370* (*TIR1*,

encoding an auxin receptor), and *FD04G014950* (*GH9C2*, encoding a glycosyl hydrolase). These results provide insights on potentially candidate genes for further study of flavonoid metabolism in buckwheat.

Population Structure and Genetic Diversity Analysis of Golden Buckwheat

Golden buckwheat varies greatly in morphology with two main ecotypes, the erect growth type (EGT) and the decumbent growth type (DGT) (Zhou et al., 2018). All buckwheat from high altitudes, including the Tibetan Plateau and Yunnan-Guizhou plateau in China, were of the DGT type, while both DGT and EGT were found at lower altitudes, *i.e.*, Yun Nan, Si Chuan and Gui Zhou in China. High altitude accessions appear to accumulate greater amounts of flavonoids than those from lower altitudes (Cheng et al., 2019). We collected 34 diploid golden buckwheat accessions from Yun Nan, Si Chuan and Gui Zhou, comprising 17 EGTs and 17 DGTs (**Fig. S31-32 and Table S23**). The DGTs displayed a significantly greater flavonoid content, while EGTs displayed overall greater height, branching, stem diameter and grain weight (**Fig. 5a and Fig. S33**).

We performed resequencing for the 34 samples, generating 26.75 Gb of sequence data, with an average sequencing depth of 6.25× (**Table S23**). By aligning to the reference genome generated above, we identified 2,201,413 SNPs and 254,271 indels (insertions and deletions, 1-50 bp in length; **Table S24**). Most of the SNPs and indels were located in the intergenic regions, only 12.03% of SNPs were presented in the coding sequence, consisting of 136,969 synonymous SNPs and 125,173 nonsynonymous SNPs. The ratio of nonsynonymous to synonymous SNPs (0.91) was lower than that of Tartary buckwheat (1.69) (Zhang et al., 2021). According to the phylogenetic analysis, the two ecotypes form distinct clades, which is further supported by principal component analysis (**Fig. 5b and c**). To explore the genetic mechanisms underlying growth and metabolic differentiation between two ecotypes, pairwise fixation statistic (F_{ST}) was calculated in windows across the genome. We identified 361 genomic regions (46.8 Mb) with the greatest genetic divergence, which contained 2,200 genes (top 5% F_{ST}, **Fig. 5d and Table S25**). Among these genes, we

found a series of putative *Arabidopsis* orthologues that are involved in the regulation of growth, development and metabolism, including *FD01G021970* (*KOB1*, encoding a glycosyltransferase), *FD01G027140* (*CRF4*, a member of the ERF/AP2 transcription factor family), *FD02G037460* (*F5H*, encoding a ferulate 5-hydroxylase), *FD03G046410* (*MYB44*, a member of the R2R3 factor MYB gene family; **Fig. S34a**), *FD03G003530* (*UGT78D2*, encoding an anthocyanidin 3-*O*-glucosyltransferase), and other related genes (**Fig. 5d**).

We found an indel located at the end of the first exon and the beginning of the intron of *FD03G046410* (named as *FdMYB44*; **Fig. 5e**) in some accessions. *FdMYB44* is orthologous to *AtMYB44* that has been reported to negatively regulate flavonoid accumulation (Jung et al., 2010). The indel leads to a loss of six amino acids at the intrinsically disordered region (**Fig. S34b**), which performs a crucial role in coordinated cellular responses by mediating dynamic interactions with different partners (Wright and Dyson, 2015). The *FdMYB44* genotype was classified into three haplotypes dependent on this indel. Hap.1 (CTCCGGTTACTG/CTCCGGTTACTG), Het. (CTCCGGTTACTG/-), and Hap.2 (-/-) (**Fig. 5e and Table S26**). Hap.1 was fixed in the DGT clade (100% of 16 accessions) and only present in one accession of the EGT clade. Hap.2 was most common in the EGT clade (75% of 16 accessions) (**Fig. 5a and 5f**). We found the expression level of *FdMYB44* was variable under dark light, red light and blue light treatments (**Fig. 5g**), implying a potential relationship with secondary metabolism. We performed co-expression analysis of *FdMYB44* with other flavonoid biosynthesis genes and found a higher coefficient with some members of the *FdF3'5'H*, *FdF3GT*, *FdRT*, *Fd4CL*, *FdDRF*, and *FdF3H* gene families, which possess MYB binding site at the promoter (**Fig. S35 and Table S27**). Taken together, we suggest that *FdMYB44* might play an important role in regulating flavonoid accumulation and the indel might be critical for the differences in flavonoid content between the two ecotypes.

In addition, we identified a nonsynonymous SNP at Fd1:82879067 (G/A), leading to the loss of the initiation codon ATG of *FD01G027140* (named as *FdCRF4*) (**Fig. S36 and Table S28**), and orthologues of this gene positively regulate root and shoot

growth in cold stress (Zwack et al., 2016). The nonsynonymous SNP yielded three haplotypes, Hap.1 (A/A), Het. (G/A), and Hap.2 (G/G). Hap.1 represents the allele with an intact start codon and was only present in the DGT accessions (9 of 16), whereas Hap.2, lacking the start codon, was fixed in the EGT accessions (**Fig. S36**). *FdCRF4* might be a key gene regulating differentiation of growth phenotypes and requires follow-up.

In summary, a large number of putative candidate genes have been unveiled and these may be useful to improve agronomic and quality traits of golden buckwheat in the future.

Discussion

Buckwheat comprises a small group of species rich in pharmacologically relevant molecules, especially flavonoids, making these dual-purpose, with both considerable edible and medicinal values. Golden buckwheat, belonging to the cymosum group of *Fagopyrum*, has traditionally been used as Chinese herbal medicine to treat a variety of human disorders because the content of value-added molecules is significantly greater than other cultivated species, including Tartary buckwheat and common buckwheat (Zhao et al., 2018). Here, we assembled and annotated the genome of a diploid golden buckwheat with a total length of 1.08 Gb, over twice the size of the closely related Tartary buckwheat (0.48 Gb). Proliferation of TEs, especially LTR-retrotransposons, is the main cause for this difference and not a polyploid event. Genome enlargement caused by LTR amplification is common in plant genome. Bursts of LTR-retrotransposon proliferation in *Oryza australiensis* and Australian cotton have similarly given rise to genome size increases, relative to the closely related *O. sativa* and American cotton, respectively, suggesting LTR proliferation is broadly involved in plant genome size variation across plants (Piegu et al., 2006; Hawkins et al., 2006). Two large inversions were found on chromosome 3 and chromosome 4, relative to Tartary buckwheat. Chromosomal inversions can establish and maintain favorable combinations of alleles, facilitating local adaptation and speciation, and genes in these inversions can underlie ecologically relevant traits,

including flowering time (Todesco et al., 2020), and annual/perennial life-history divergence (Lowry and Willis, 2010). Thus, we speculate that the genome size variation induced by repeats amplification and these large inversions might be responsible for the difference of adaptability, plant architecture between golden buckwheat and Tartary buckwheat as well as the unique adaptability and reproductive isolation for golden buckwheat.

The abundance of flavonoids in the rhizome of golden buckwheat focused our attention on the genetic basis of flavonoid biosynthesis and rhizome development. Chromosomal inversions could modulate gene dosage and the epigenetic environment near breakpoints (Mérot et al., 2020) and the duplications of specific gene families could be involved in divergence of adaptive traits. We found that genes near the inversion breakpoints, multiple duplicate genes and genes unique to golden buckwheat were associated with metabolic processes (including flavonoid biosynthesis), cell wall biosynthesis, root morphological development, and stress response. These results lead us to speculate that these large inversions and other aspects of gene content evolution are associated with the increased flavonoid accumulation and the unique root architecture of golden buckwheat. In addition, *CYP*, a plant gene family catalyzing versatile oxidation process in secondary metabolism (Mizutani and Ohta, 2010), and a well-known transposase-derived transcription factor *FAR*, involved in JA signaling and glucosinolate biosynthesis (Fernandez-Calvo et al., 2020; Liu et al., 2019), were expanded in golden buckwheat. As flavonoid accumulation is regulated by CYP and JA signaling (Shan et al., 2009; Qi et al., 2011; Ni et al., 2020; Sun et al., 2020), we speculated that the expansion of CYP and FAR was responsible for the increased flavonoid accumulation in golden buckwheat. The overexpression of *FdFRS1* resulted in elevated rutin content, helping support the role of this gene in flavonoid biosynthesis in golden buckwheat. Previous research has also found that the expansion of key genes involved in secondary metabolite biosynthesis can be beneficial to the accumulation of bioactive components and is common in plants genome evolution. Some gene families essential for flavonoid biosynthesis, including *FdCHI*, *FdF3H*, *FdDFR* and *FdLAR*, were expanded and appeared in

clusters across the golden buckwheat genome. Similar results were found in camphor tree and azalea, in which the expansions of terpenoid synthase and flavonoids synthesis genes were associated with the mass production of terpenoids and anthocyanins, respectively (Chaw et al., 2019; Yang et al., 2020). This provides further evidence for our suggestion that the expansion of these structural and regulatory genes might be responsible for the increased flavonoid accumulation in golden buckwheat. Co-expression network analysis demonstrated that some of the modules of genes were correlated with specific tissues, including two modules with a high number of genes putatively involved in flavonoid biosynthesis. This potentially provides new possibilities for resolving the genetic basis of tissue-specific flavonoid metabolism regulation in golden buckwheat and other species.

The Himalayan region is the center of origin for Tartary buckwheat (Zhang et al., 2021), raising the possibility of a similar golden buckwheat origin, however, recent overexploitation has dramatically reduced the golden buckwheat wild resources (Chen and Li, 2016). We investigated buckwheat germplasm resources from southwest China. Thirty-four diploid golden buckwheat accessions were collected and resequencing led to the confirmation that the two morphological types are also genetically differentiated. Previous research illustrated that most of the golden buckwheat growing at low altitude are of the erect ecotypes, while decumbent ecotypes usually grow at high altitude (Cheng et al., 2019). By comparing the genomes of these ecotypes, we identified some candidate genes potentially responsible for divergence phenotypes of the two ecotypes. For instance, CRFs are involved in lateral root initiation under cold stress (Zwack et al., 2016; Jeon et al., 2016) and in our study, we found loss of the predicted start codon of a CRF gene in the erect ecotype, which, if this causes loss of expression, or expression of a mutant protein, could prevent this ecotype from inhabiting the colder high altitude environments, restricting its presence to the warm environment at low altitude.

Genes involved in cellulose biosynthesis (e.g., *KOB1*) and syringic lignin biosynthesis (e.g., *F5H*) were also significantly differentiated between the two ecotypes and might be responsible for the difference in plant architecture in the two

ecotypes (Pagant et al., 2002; Wang et al., 2015; Wu et al., 2019). A flavonoid 3-*O*-glucosyltransferase gene *UGT78D2* involved in both flavonoid biosynthesis and shoot polar auxin transport were also divergent between two ecotypes, which could be responsible for not only the different flavonoids content but also the differences in plant height and branching that we identified (Yin et al., 2012; Yin et al., 2014). In addition, differentiation of MYB44 (an indel affecting the coding region and an exon-intron boundary), which is a universal transcription factor involved in flavonoid biosynthesis, was also identified (Wei et al., 2020). The alteration of the intrinsically disordered region might play a role in the different value-added flavonoids accumulation of two distinct ecotypes.

In conclusion, the genome sequence presented in this study lays the foundation for elaboration of the synthesis and regulatory mechanisms underlying pharmaceutically relevant molecules in golden buckwheat and will facilitate the breeding and/or genetic engineering of buckwheat and other medical plant species containing value-added molecules. The resequencing panel comprising broad genetic variation could greatly facilitate the study of golden buckwheat population structure and provide knowledge regarding the better utilization of the natural genetic variation resource for buckwheat breeding in the future.

Acknowledgements

This research was supported by National Key R&D Program of China (2019YFD1001300/2019YFD1001301), National Natural Science Foundation of China (32161143005, 31871536, 31801427, 31901511), European Union Horizon 2020 research and innovation programme, project PlantaSYST (SGA No. 739582 under FPA No. 664620), the BG05M2OP001-1.003-001-C01 project financed by the European Regional Development Fund through the “Science and Education for Smart Growth” Operational Programme, China National Postdoctoral Program for Innovative Talents (BX20200377).

We appreciate the critical reading and helpful comments on the manuscript made by Tao Lin from China Agricultural University, Qiang Gao from Boke Biotech Co.

and Huilong Du from Hebei University. We thank Faliang Li and Junzhen Wang from Research Station of Alpine Crop, Xichang Institute of Agricultural Sciences for their efforts on the preparation of plant samples.

Author's contributions

M.Z., E.D., Z.Z., and M.I.G. designed and managed the project. M.Z. and M.I.G. organized the funding for this research. M.H. and B.G. contributed to the generation of genome assembly data and whole-genome resequencing data. M.H., X.Z., B.G., and W.Y. performed the data analysis and/or figure design. X.L., Y.F., H.Z., R.J., and M.N.H. performed most of experiments. M.H., Y.H., K.Z., M.A.C. and M.Z. wrote the manuscript and finalized the manuscript. Y.T., J.W., M.Y., J.C., J.R., E.D., Z.Z., M.I.G., M.A.C., and M.Z. provided scientific advice. All authors read and approved the paper.

ORCID

Ming He https://orcid.org/0000-0002-6730-4602
Yuqi He https://orcid.org/0000-0001-9920-9653
Kaixuan Zhang https://orcid.org/0000-0001-9562-1675
Xiang Lu https://orcid.org/0000-0003-1792-5926
Xuemei Zhang https://orcid.org/0000-0002-1554-3388
Hui Zhao https://orcid.org/0000-0002-0175-5494
Weifei Yang https://orcid.org/0000-0002-3873-8911
Mark A. Chapman https://orcid.org/0000-0002-7151-723X
Meiliang Zhou https://orcid.org/0000-0001-5175-3684

Data availability

All sequencing data for this project is available from CNGB Nucleotide Sequence Archive (https://db.cngb.org/cnsa) under the accession ID CRA006831. The final chromosome-level genome of golden buckwheat is available in Figshare doi: 10.6084/m9.figshare.19711891.

References

Anders S, Pyl PT, Huber W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166-169.

Bak S, Beisson F, Bishop G, Hamberger B, Höfer R, Paquette S, Werck-Reichhart D. 2011. Cytochromes P450. *Arabidopsis Book* **9**: e0144.

Byrne ME, Barley R, Curtis M, Arroyo JM, Dunham M, Hudson A, Martienssen RA. 2000. Asymmetric leaves1 mediates leaf patterning and stem cell function in Arabidopsis. *Nature* **408**: 967-971.

Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome annotation and curation using MAKER and MAKER-P. *Current protocols in bioinformatics* **48**: 4.11.11-39.

Chan P. 2003. Inhibition of tumor growth in vitro by the extract of *Fagopyrum cymosum* (*fago-c*). *Life Sciences* **72**: 1851-1858.

Chaw S, Liu Y, Wu Y, Wang H, Lin C, Wu C, Ke H, Chang L, Hsu C, Yang H, et al. 2019. Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. *Nature Plants* **5**: 63-73.

Chen C, Li A. 2016. Transcriptome analysis of differentially expressed genes involved in proanthocyanidin accumulation in the rhizomes of *Fagopyrum dibotrys* and an irradiation-Induced mutant. *Frontiers in Physiology* **7**: 100.

Chen Q. 2016. Recent progresses on interspecific crossing of genus *Fagopyrum* Mill. *Proc 13th Intl Symp Buckwheat at Cheongju*: 285-298.

Chen S, Zhou Y, Chen Y, Gu J. 2018. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**: 884-890.

Cheng C, Fan Y, Tang Y, Zhang K, Joshi DC, Jha R, Janovska D, Meglic V, Yan M, Zhou M. 2020. *Fagopyrum esculentum* ssp. ancestrale-a hybrid species between diploid *F. cymosum* and *F. esculentum*. *Frontiers in Plant Science* **11**: 1073.

Cheng C, Zhang K, Tang Y, Shao J, Yan M, Zhou M. 2019. Investigation on wild *Fagopyrum cymosum* resources in Yunnan and analysis of genetic diversity. *Journal of Plant Genetic Resources* **20**: 1438-1446.

De Bie T, Cristianini N, Demuth JP, Hahn MW. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**: 1269-1271.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, et al. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**: 92-95.

Doyle J, Doyle, J. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**:11-15.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**: 157.

Fernandez-Calvo P, Inigo S, Glauser G, Vanden Bossche R, Tang M, Li B, De Clercq R, Nagels Durand A, Eeckhout D, Gevaert K, et al. 2020. *FRS7* and *FRS12* recruit NINJA to regulate expression of glucosinolate biosynthesis genes.

New Phytologist **227**: 1124-1137.

Ganeshpurkar A, Saluja AK. 2017. The pharmacological potential of rutin. *Saudi Pharmaceutical Journal* **25**: 149-164.

Goel M, Sun H, Jiao WB, Schneeberger K. 2019. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology* **20**: 277.

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Research* **16**: 1252-1261.

He Y, Zhang X, Li L, Sun Z, Li J, Chen X, Hong G. 2021. SPX4 interacts with both *PHR1* and *PAP1* to regulate critical steps in phosphorus-status-dependent anthocyanin biosynthesis. *New Phytologist* **230**: 205-217.

Houtgast EJ, Sima VM, Bertels K, Al-Ars Z. 2018. Hardware acceleration of BWA-MEM genomic short read mapping for longer read lengths. *Computational Biology and Chemistry* **75**: 54-64.

Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution* **34**: 2115-2122.

Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, et al. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**: D309-D314.

Jeon J, Cho C, Lee MR, Van Binh N, Kim J. 2016. *CYTOKININ RESPONSE FACTOR2 (CRF2)* and *CRF3* regulate lateral root development in response to cold stress in *Arabidopsis*. *The Plant Cell* **28**: 1828-1843.

Jiang W, Yin Q, Wu R, Zheng G, Liu J, Dixon RA, Pang Y. 2015. Role of a chalcone isomerase-like protein in flavonoid biosynthesis in *Arabidopsis thaliana*. *Journal of Experimental Botany* **66**: 7165-7179.

Jin J, Liu Y, Ma C, Ma J, Hao W, Xu Y, Yao M, Chen L. 2018. A novel F3′ 5′ H allele with 14 bp deletion is associated with high catechin index trait of wild tea plants and has potential use in enhancing tea quality. *Journal of Agricultural and Food Chemistry* **66**: 10470-10478.

Jing R, Li H, Hu CL, Jiang Y, Qin L, Zheng C. 2016. Phytochemical and pharmacological profiles of three *Fagopyrum* Buckwheats. *International Journal of Molecular Sciences* **17**: 589.

Joshi DC, Zhang K, Wang C, Chandora R, Khurshid M, Li J, He M, Georgiev MI, Zhou M. 2020. Strategic enhancement of genetic gain for nutraceutical development in buckwheat: a genomics-driven perspective. *Biotechnology Advances* **39**: 107479.

Jung C, Shim JS, Seo JS, Lee HY, Kim CH, Choi YD, Cheong JJ. 2010. Non-specific phytohormonal induction of *AtMYB44* and suppression of jasmonate-responsive gene activation in *Arabidopsis thaliana*. *Molecules and*

Cells **29**: 71-76.

Kang M, Fu R, Zhang P, Lou S, Yang X, Chen Y, Ma T, Zhang Y, Xi Z, Liu J. 2021. A chromosome-level *Camptotheca acuminata* genome assembly provides insights into the evolutionary origin of camptothecin biosynthesis. *Nature communications* **12**: 3531.

Katsu K, Suzuki R, Tsuchiya W, Inagaki N, Yamazaki T, Hisano T, Yasui Y, Komori T, Koshio M, Kubota S, et al. 2017. A new buckwheat dihydroflavonol 4-reductase (DFR), with a unique substrate binding structure, has altered substrate specificity. *BMC Plant Biology* **17**: 239.

Ke H, Wang X, Zhou Z, Ai W, Wu Z, Zhang Y. 2021. Effect of weimaining on apoptosis and Caspase-3 expression in a breast cancer mouse model. *Journal of Ethnopharmacology* **264**: 113363.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nature methods* **12**: 357–360.

Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, et al. 2006. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Research* **34**: D572-D580.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Li X, Liu J, Chang Q, Zhou Z, Han R, Liang Z. 2021. Antioxidant and antidiabetic activity of proanthocyanidins from *Fagopyrum dibotrys*. *Molecules* **26**: 2417.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289-293.

Lin R, Wang H. 2004. *Arabidopsis FHY3*/*FAR1* gene family and distinct roles of its members in light control of *Arabidopsis* development. *Plant Physiology* **136**: 4010-4022.

Liu P, Du L, Huang Y, Gao S, Yu M. 2017. Origin and diversification of leucine-rich repeat receptor-like protein kinase (LRR-RLK) genes in plants. *BMC Evolutionary Biology* **17**: 47.

Liu Y, Shi Z, Maximova S, Payne MJ, Guiltinan MJ. 2013. Proanthocyanidin synthesis in *Theobroma cacao*: genes encoding anthocyanidin synthase, anthocyanidin reductase, and leucoanthocyanidin reductase. *BMC Plant Biology* **13**: 202.

Liu Y, Wei H, Ma M, Li Q, Kong D, Sun J, Ma X, Wang B, Chen C, Xie Y, et al. 2019. *Arabidopsis FHY3* and *FAR1* regulate the balance between growth and defense responses under shade conditions. *The Plant Cell* **31**: 2089-2106.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**: 550.

Lowry DB, Willis JH. 2010. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biology* **8**: e1000500.

Marcais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: a fast and versatile genome alignment system. *PLoS Computational Biology* **14**: e1005944.

Marcais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**: 764-770.

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 Research* **20**: 1297-1303.

Miao J, Feng Q, Li Y, Zhao Q, Zhou C, Lu H, Fan D, Yan J, Lu Y, Tian Q, et al. 2021. Chromosome-scale assembly and analysis of biomass crop *Miscanthus lutarioriparius* genome. *Nature Communications* **12**: 2458.

Mizutani M, Ohta D. 2010. Diversification of P450 genes during land plant evolution. *Annual Review of Plant Biology* **61**: 291-315.

Musial C, Kuban-Jankowska A, Gorska-Ponikowska M. 2020. Beneficial properties of green tea catechins. *International Journal of Molecular Sciences* **21**: 1744.

Nattestad M, Schatz MC. 2016. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**: 3021-3023.

Ni J, Zhao Y, Tao R, Yin L, Gao L, Strid A, Qian M, Li J, Li Y, Shen J, et al. 2020. Ethylene mediates the branching of the jasmonate-induced flavonoid biosynthesis pathway by suppressing anthocyanin biosynthesis in red Chinese pear fruits. *Plant Biotechnology Journal* **18**: 1223-1240.

Nurmberg PL, Knox KA, Yun BW, Morris PC, Shafiei R, Hudson A, Loake GJ. 2007. The developmental selector *AS1* is an evolutionarily conserved regulator of the plant immune response. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 18795-18800.

Ohnishi O. 1998. Search for the wild ancestor of buckwheat - III. The wild ancestor of cultivated common buckwheat, and of Tatary buckwheat. *Economic Botany* **52**: 123-133.

Ohnishi O, Zhou M. 2018. Annual self-incompatible species. In: Zhou M, Kreft I, Suvorova G, Tang Y, Woo SH, editors. Buckwheat germplasm in the world. London: Academic, 71-80.

Pagant S, Bichet A, Sugimoto K, Lerouxel O, Desprez T, McCann M, Lerouge P, Vernhettes S, Hofte H. 2002. *KOBITO1* encodes a novel plasma membrane protein necessary for normal synthesis of cellulose during cell expansion in *Arabidopsis*. *The Plant Cell* **14**: 2001-2013.

Pfeifer B, Wittelsbuerger U, Ramos-Onsins SE, Lercher M. 2013. PopGenome: an efficient swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution* **31**: 1929-1936.

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, et al. 2006. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Research* **16**: 1262-1269.

Qi T, Song S, Ren Q, Wu D, Huang H, Chen Y, Fan M, Peng W, Ren C, Xie D. 2011. The Jasmonate-ZIM-domain proteins interact with the WD-Repeat/bHLH/MYB complexes to regulate Jasmonate-mediated anthocyanin accumulation and trichome initiation in *Arabidopsis thaliana*. *The Plant Cell* **23**: 1795-1814.

Reuben S, Rai A, Pillai BV, Rodrigues A, Swarup S. 2013. A bacterial quercetin oxidoreductase QuoA-mediated perturbation in the phenylpropanoid metabolic network increases lignification with a concomitant decrease in phenolamides in *Arabidopsis*. *Journal of Experimental Botany* **64**: 5183-5194.

Roach MJ, Schmidt SA, Borneman AR. 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatic* **19**: 460.

Seitz C, Ameres S, Schlangen K, Forkmann G, Halbwirth H. 2015. Multiple evolution of flavonoid 3′, 5′-hydroxylase. *Planta*. **242**: 561-573.

Shan X, Zhang Y, Peng W, Wang Z, Xie D. 2009. Molecular mechanism for jasmonate-induction of anthocyanin accumulation in *Arabidopsis*. *Journal of Experimental Botany* **60**: 3849-3860.

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210-3212.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312-1313.

Stanke M, Waack S. 2003. Gene prediction with a hidden markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**: II215-II225.

Sun W, Ma Z, Liu M. 2020. Cytochrome P450 family: genome-wide identification provides insights into the rutin synthesis pathway in Tartary buckwheat and the improvement of agricultural product quality. *International Journal of Biological Macromolecules* **164**: 4032-4045.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008. Perspective-synteny and collinearity in plant genomes. *Science* **320**: 486-488.

Todesco M, Owens GL, Bercovich N, Legare JS, Soudi S, Burge DO, Huang K, Ostevik KL, Drummond EBM, Imerovski I, et al. 2020. Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* **584**: 602-607.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**: 511-515.

Van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES. 2010. Hi-C: a method to study the three-dimensional architecture of genomes. *Journal of Visualized Experiments* (39): 1869.

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

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**: e164.

Wang KJ, Zhang YJ, Yang CR. 2005. Antioxidant phenolic constituents from *Fagopyrum dibotrys*. *Journal of Ethnopharmacology* **99**: 259-264.

Wang X, Jing Y, Zhang B, Zhou Y, Lin R. 2015. Glycosyltransferase-like protein *ABI8*/*ELD1*/*KOB1* promotes *Arabidopsis* hypocotyl elongation through regulating cellulose biosynthesis. *Plant Cell and Environment* **38**: 411-422.

Wang X, Zhou T, Bai G, Zhao Y. 2018. Complete chloroplast genome sequence of *Fagopyrum dibotrys*: genome features, comparative analysis and phylogenetic relationships. *Scientific Reports* **8**: 12379.

Wei Z, Hu K, Zhao D, Tang J, Huang Z, Jin P, Li Y, Han Z, Hu L, Yao G, et al. 2020. *MYB44* competitively inhibits the formation of the *MYB340-bHLH2-NAC56* complex to regulate anthocyanin biosynthesis in purple-fleshed sweet potato. *BMC Plant Biology* **20**: 258.

Wright PE, Dyson HJ. 2015. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology* **16**: 18-29.

Wu Z, Wang N, Hisano H, Cao Y, Wu F, Liu W, Bao Y, Wang Z, Fu C. 2019. Simultaneous regulation of *F5H* in COMT-RNAi transgenic switchgrass alters effects of COMT suppression on syringyl lignin biosynthesis. *Plant Biotechnology Journal* **17**: 836-845.

Xie S, Cui L, Lei X, Yang G, Li J, Nie X, Ji W. 2019. The *TIFY* gene family in wheat and its progenitors: genome-wide identification, evolution and expression analysis. *Current Genomics* **20**: 371-388.

Xu D, Hu M, Wang Y, Cui Y. 2019. Antioxidant activities of quercetin and its complexes for medicinal application. *Molecules* **24**: 1123.

Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**: W265-268.

Xue A, Hickerson MJ. 2015. The aggregate site frequency spectrum for comparative population genomic inference. *Molecular Ecology* **2**: 6223-6240.

Yang F, Nie S, Liu H, Shi T, Tian X, Zhou S, Bao Y, Jia K, Guo J, Zhao W, et al. 2020. Chromosome-level genome assembly of a parent species of widely cultivated azaleas. *Nature Communications* **11**: 5269.

Yin R, Han K, Heller W, Albert A, Dobrev PI, Zazimalova E, Schaeffner AR. 2014. Kaempferol 3-O-rhamnoside-7-O-rhamnoside is an endogenous flavonol inhibitor of polar auxin transport in *Arabidopsis* shoots. *New Phytologist* **201**: 466-475.

Yin R, Messner B, Faus-Kessler T, Hoffmann T, Schwab W, Hajirezaei MR, von Saint Paul V, Heller W, Schaeffner AR. 2012. Feedback inhibition of the general phenylpropanoid and flavonol biosynthetic pathways upon a compromised flavonol-3-O-glycosylation. *Journal of Experimental Botany* **63**: 2465-2478.

Zhang K, He M, Fan Y, Zhao H, Gao B, Yang K, Li F, Tang Y, Gao Q, Lin T, et al. 2021. Resequencing of global Tartary buckwheat accessions reveals multiple

domestication events and key loci associated with agronomic traits. *Genome Biology* **22**: 23.

Zhang L, Li X, Ma B, Gao Q, Du H, Han Y, Li Y, Cao Y, Qi M, Zhu Y, et al. 2017. The Tartary buckwheat genome provides insights into rutin biosynthesis and abiotic stress tolerance. *Molecular Plant* **10**: 1224-1237.

Zhao J, Jiang L, Tang X, Peng L, Li X, Zhao G, Zhong L. 2018. Chemical composition, antimicrobial and antioxidant activities of the flower volatile oils of *Fagopyrum esculentum*, *Fagopyrum tataricum* and *Fagopyrum Cymosum*. *Molecules* **23**: 182.

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**: 3326-3328.

Zheng Y, Jiao C, Sun H, Rosli HG, Pombo MA, Zhang P, Banf M, Dai X, Martin GB, Giovannoni JJ, et al. 2016. iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Molecular Plant* **9**: 1667-1670.

Zhou M, Sun Z, Ding M, Logacheva MD, Kreft I, Wang D, Yan M, Shao J, Tang Y, Wu Y, et al. 2017. *FtSAD2* and *FtJAZ1* regulate activity of the *FtMYB11* transcription repressor of the phenylpropanoid pathway in *Fagopyrum tataricum*. *New Phytologist* **216**: 814-828.

Zhou M, Tang Y, Deng X, Ruan C, Ding M, Shao J, Tang Y, Wu Y. 2018. Perennial self-incompatible wild *Fagopyrum* species. In: *Buckwheat germplasm in the world*: Elsevier, 61-66.

Zwack PJ, Compton MA, Adams CI, Rashotte AM. 2016. Cytokinin response factor 4 (CRF4) is induced by cold and involved in freezing tolerance. *Plant Cell Reports* **35**: 573-584.

Figure legends

Fig. 1. Golden buckwheat morphology and genomic features. (a) Decumbent growth type of golden buckwheat. **(b)** and **(c)** Erect growth type of golden buckwheat. **(d)** to **(f)** The inflorescences, immature fruit, and rhizome of golden buckwheat. **(g)** The genomic landscape, with the following features indicated: (1) the 8 pseudochromosomes, (2) to (6) the density of TEs, genes, *Gypsy* elements, *Copia* elements, LTRs, respectively, (7) GC content, and (8) the colored links represent syntenic blocks in golden buckwheat genome. The window size for the tracks was 500 kb.

Fig. 2. Genome comparison between golden buckwheat and Tartary buckwheat. (a) Syntenic analysis of golden buckwheat and Tartary buckwheat genomes. **(b)**

Inversions of chromosome 3 and chromosome 4. Horizontal lines indicate golden buckwheat (blue-grey) and Tartary buckwheat (red) chromosomes. The yellow lines indicate the inverted regions. The grey lines indicate syntenic regions. **(c)** Gene Ontology (GO) enrichment of genes involved in the large inversion on chromosome 3 of golden buckwheat. The color of the box represents –log10 (p-value) and the size of the box represents the gene count for each GO term. **(d)** The frequency distributions of Ks of homologous gene pairs in the collinearity block of golden buckwheat (GB) versus Tartary buckwheat (TB), GB versus GB, and TB versus TB, respectively. **(e)** GO enrichment of genes located in the collinearity block of GB versus GB with Ks < 0.3.

Fig. 3. Genome and gene family evolution of golden buckwheat. (a) Phylogenetic tree based on single copy orthologues of two buckwheats, and nine other species showing divergence times based on single-copy gene. The estimated divergence timings were indicated at the internodes. The number of expanded and contracted gene families was shown in blue next to the specie names, indicated by plus and minus, respectively. **(b)** GO enrichment results of golden buckwheat expanded gene families. **(c)** Neighbor-joining tree of expanded FdCYPs with flavonoid 3'-monooxygenase domains. The gene IDs of golden buckwheat, Tartary buckwheat, and Arabidopsis were shown in red, blue, and black, respectively. The branches marked in blue, orange, yellow, and red represented F3H, C4H, F3’H, and F3’5’H families, respectively. **(d)** The expression of expanded *FdCYPs* in seedlings grown under different light conditions. DL, RL, and BL represent dark, red light, and blue light, respectively. Genes were clustered according to expression levels. Genes clustered with F3'5'H in **(c)** were marked in red color. The color of the box represents log10 (FPKM+1). **(e)** Images of golden buckwheat seedlings treated with different light conditions for 10 days. Bar = 1 cm. **(f)** Flavonoid content of 10-day golden buckwheat seedlings grown under different light conditions. **(g)** Neighbor-joining tree of FAR transcription factors. The golden buckwheat, Tartary buckwheat, and Arabidopsis gene IDs were shown in red, blue, and black, respectively. **(h)** Rutin content of three *FdFRS1* overexpression lines in transgenic hairy roots. Data are mean

± SD. ***P < 0.001, student's *t*-test.

Fig. 4. Metabolomic analysis and genes involved in flavonoid biosynthesis. (a) Patterns of golden buckwheat rhizome and Tartary buckwheat root. GBR, golden buckwheat root; GBT, golden buckwheat tuber; TBR, Tartary buckwheat root. **(b)** Principal component analysis of metabolites showing the first two components. **(c)** Volcano plot of the metabolic differences between GBR and TBR. The red and orange dots represent differential metabolites and the grey dots represent metabolites with insignificant differences. The fold change threshold and the P value threshold were set to 2.0 and 0.1, respectively. Irha, isorhamnetin; Jac, jaceosidin; Rut, rutin; Que, quercetin; Tax, taxifolin; Ast, astragalin; Dio, diosmetin; Cat, catechol; Pcat, protocatechualdehyde; Pcy A2, procyanidin A2; Pcy B1, procyanidin B1. **(d)** to **(g)** Boxplots showing the relative content of Cat (catechol), Pcy A2 (procyanidin A2), Pcy B1 (procyanidin B1), and Ast (astragalin). Boxes represent the 25%, 50%, and 75% values; whiskers represent SD; "x" represent maximum and minimum values; square dots display the average value. ***P < 0.001, Student's *t*-test. **(h)** A simplified representation of the biosynthetic pathway of flavonoid metabolism. The expression level of each gene in six tissues was shown in color. Root, R; tuber, T; stem, S; young leaf, YL; old leaf, OL; flower, F. The color of the box represents log10 (FPKM+1). The significantly expanded gene families in golden buckwheat are highlighted in red.

Fig. 5. Population structure and genetic differentiation of golden buckwheat. (a) Qualification of quality and agronomic traits of erect growth type (EGT) and decumbent growth type (DGT) groups. n = 17 for both of group EGT and group DGT. The boxes represent 25%, 50%, and 75% values; whiskers represent SD; "x" represent maximum and minimum values; square dots display the average value. *P < 0.05, ***P < 0.001, Student's *t*-test. **(b)** Neighbor-joining tree of 35 buckwheat germplasms, including 34 golden buckwheat accessions and one Tartary buckwheat accession. Branch colors indicate different groups: group DGT (blue), group EGT (red), and outgroup (black line). **(c)** Principal component analysis of golden buckwheat accessions, showing the first two components. Colors correspond to the neighbor-joining tree grouping in **(b)**. **(d)** Highly divergent genomic regions between

EGT and DGT. The horizontal dashed line indicates the top 5% of F_{ST} and red vertical lines indicate candidate genes in the highly divergent regions. **(e)** Schematic representation of *FdMYB44* genomic sequence. Exons and introns are represented by boxes and lines, respectively. The sequence of indel was indicated by dark text (exon) and grey text (intron). **(f)** Distribution of three haplotypes of *FdMYB44* in EGT and DGT. Hap.1, haplotype 1; Hap.2, haplotype 2; Het., heterozygous. **(g)** Expression of *FdMYB44* under different light conditions. DL, RL, and BL represent dark, red light, and blue light, respectively.

Supporting Information

Fig. S1 The frequency distribution of 25 k-mer.

Fig. S2 Histogram of nuclear DNA contents obtained by flow cytometry analysis.

Fig. S3 Histogram for length distribution of raw Oxford Nanopore reads.

Fig. S4 Genome-wide Hi-C map of golden buckwheat showing genome-wide all-by-all interactions.

Fig. S5 The frequency distribution of LTR insertion time.

Fig. S6 Statistics of structural variations in golden buckwheat genome and Tartary buckwheat genome.

Fig. S7 Collinear relationship between golden buckwheat and Tartary buckwheat.

Fig. S8 Structural variants on chromosomes 1, 2, 5-8.

Fig. S9 Gene Ontology enrichment analysis of genes involved in inversions on chromosome 4 of golden buckwheat.

Fig. S10 Intra-genome dot plot comparison of golden buckwheat.

Fig. S11 Classification of different origins of duplicate genes.

Fig. S12 GO enrichment of duplicate genes of dispersed duplication origin of golden buckwheat.

Fig. S13 GO enrichment of duplicate genes of proximal duplication origin of golden buckwheat.

Fig. S14 GO enrichment of duplicate genes of tandem duplication origin of golden buckwheat.

Fig. S15 GO enrichment of duplicate genes of transpose duplication origin of golden buckwheat.

Fig. S16 GO enrichment of duplicate genes of WGD duplication origin of golden buckwheat.

Fig. S17 Schematic representation of the expanded FdCYP amino acid sequence.

Fig. S18 Neighbor-joining tree of expanded FdMYBs.

Fig. S19 The hairy root transgenic system of golden buckwheat.

Fig. S20 Identification of positive transgenic hairy roots overexpressing *FdFRS1*.

Fig. S21 Correlation heatmap of metabolites in three tissues based on Pearson correlation coefficient.

Fig. S22 Volcano plots of the metabolic differences between golden buckwheat and Tartary buckwheat.

Fig. S23 Boxplots showing the relative content of flavonoids.

Fig. S24 Expression heatmap of key genes in flavonoid biosynthesis pathways with different light treatments.

Fig. S25 Distribution of flavonoid biosynthesis related genes on the chromosomes of golden buckwheat.

Fig. S26 Distribution of flavonoid biosynthesis related genes on the chromosomes of Tartary buckwheat.

Fig. S27 Clustering dendrogram of differential expressed genes.

Fig. S28 The heatmap of the gene network depicts the Topological Overlap Matrix among all differential expressed genes.

Fig. S29 Heat map of correlation between modules and tissues.

Fig. S30 Gene co-expression networks of grey60 module and brown module.

Fig. S31 Histogram of nuclear DNA contents obtained by flow cytometry analysis from decumbent golden buckwheat.

Fig. S32 Histogram of nuclear DNA contents obtained by flow cytometry analysis from erect golden buckwheat.

Fig. S33 Qualification of agronomic traits in EGT and DGT.

Fig. S34 Phylogenetic tree analysis of FdMYBs in the differential regions between

DGT and EGT.

Fig. S35 Analysis of MYB binding elements in the promoter of flavonoid biosynthesis genes.

Fig. S36 Variation analysis of *FdCRF4* in different growth types.

Table S1 Summary of sequencing data used for golden buckwheat genome assembly.

Table S2 Genome survey data statistics with k-mer frequency distribution.

Table S3 Summary of genome assembly.

Table S4 Lengths of 8 pseudomolecules.

Table S5 Statistics of BUSCO evaluation of scaffold-genome.

Table S6 RNA-seq data and numbers of expressed genes.

Table S7 Statistics of predicted gene models.

Table S8 Statistics of annotated genes of golden buckwheat using different database.

Table S9 BUSCO evaluation of annotated genes of golden buckwheat.

Table S10 Statistics of repeat sequence predicted with ab initio prediction and homology-based approach.

Table S11 Details of the inversions in golden buckwheat genome compared with Tartary buckwheat.

Table S12 Information of genes in the inversion regions on Chr3 and Chr4.

Table S13 Lists of collinear gene pairs in WGD event.

Table S14 Clustering statistics of gene families.

Table S15 Orthologous gene families among 13 plant species.

Table S16 Self-incompatibility genes in golden buckwheat.

Table S17 A list of expanded and contracted gene families of golden buckwheat.

Table S18 Expression levels and conserved domains of expanded FdCYPs.

Table S19 Number of transcription factors.

Table S20 Relative contents of differential metabolites between golden buckwheat and Tartary buckwheat.

Table S21 Statistics of flavonoid biosynthetic pathway encoding genes.

Table S22 Genes in 14 modules classified by WGCNA.

Table S23 Summary of all accessions sequenced in this study.

Table S24 Distribution of SNPs and indels within various genomic regions.

Table S25 Putative sweep regions and genes experiencing differentiation between EGT and DGT.

Table S26 Variation information of *FdMYB44* (*FD03G046410*) genomic DNA sequence.

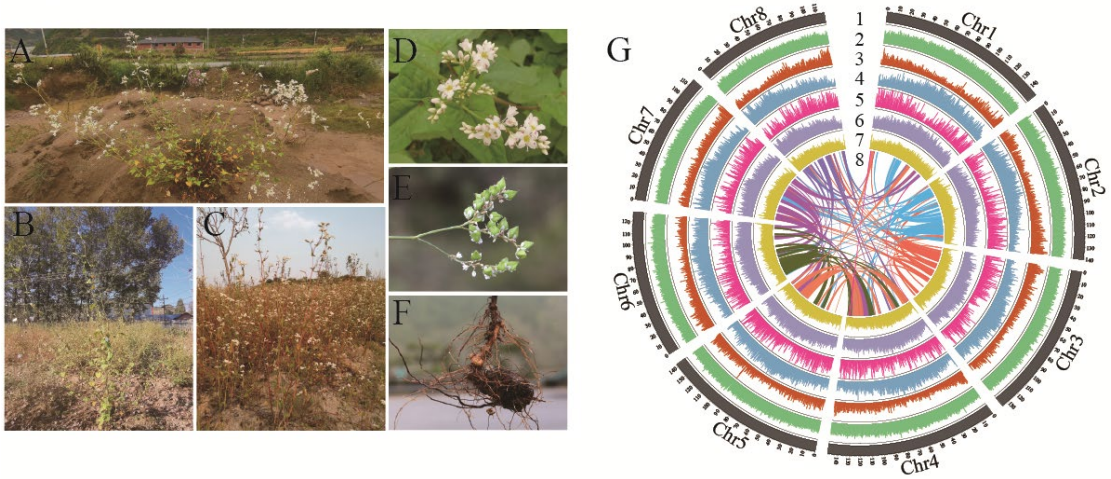
Table S27 Co-expression analysis of FdMYB44 with flavonoid metabolism related genes.

Table S28 Variation information of *FdCRF4* (*FD01G027140*) genomic DNA sequence.

1

2

Figure 1



3

4

Figure 2

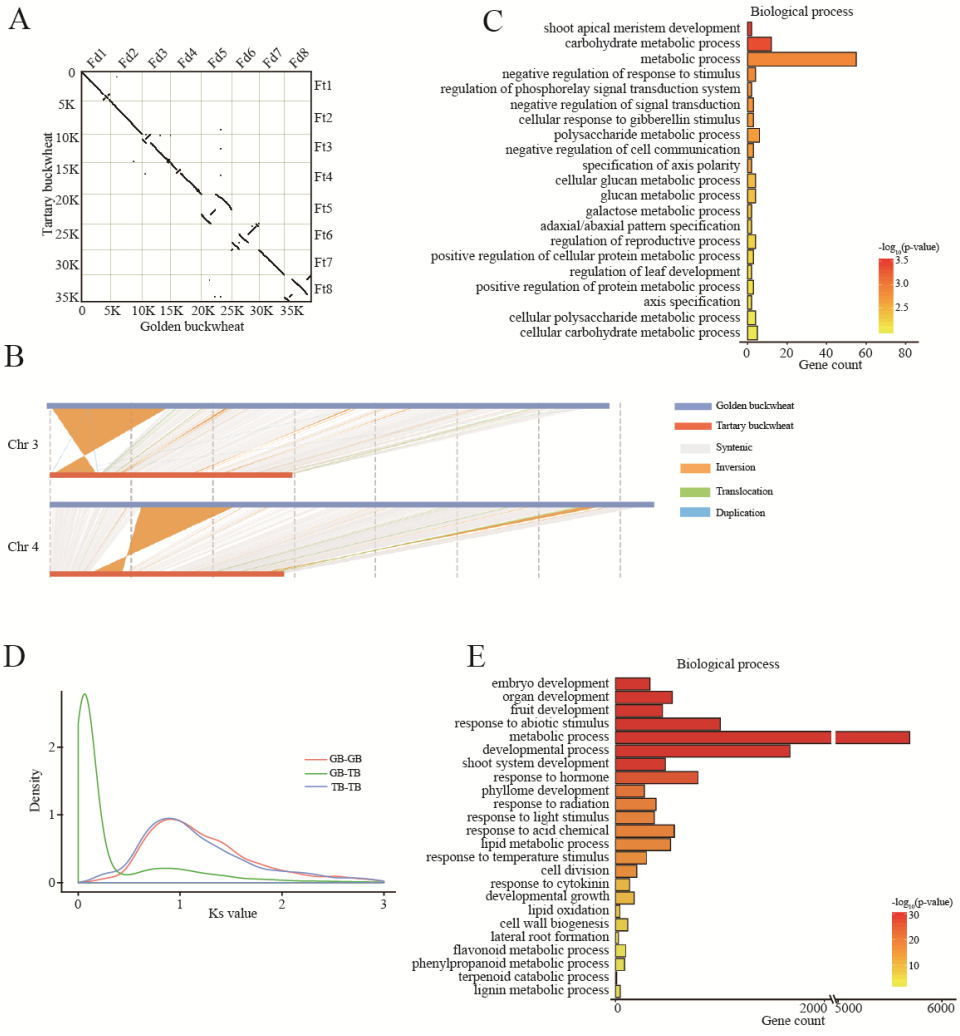

A
Fd1
Fd2
Fd3
Fd4
Fd5
Fd6
Fd7
Fd8
Ft1
Ft2
Ft3
Ft4
Ft5
Ft6
Ft7
Ft8
Tartary buckwheat
Golden buckwheat
0
5K
10K
15K
20K
25K
30K
35K
B
Chr 3
Chr 4
Golden buckwheat
Tartary buckwheat
Syntenic
Inversion
Translocation
Duplication
C
Biological process
shoot apical meristem development
carbohydrate metabolic process
metabolic process
negative regulation of response to stimulus
regulation of phosphorelay signal transduction system
negative regulation of signal transduction
cellular response to gibberellin stimulus
polysaccharide metabolic process
negative regulation of cell communication
specification of axis polarity
cellular glucan metabolic process
glucan metabolic process
galactose metabolic process
adaxial/abaxial pattern specification
regulation of reproductive process
positive regulation of cellular protein metabolic process
regulation of leaf development
positive regulation of protein metabolic process
axis specification
cellular polysaccharide metabolic process
cellular carbohydrate metabolic process
-log10(p-value)
3.5
3.0
2.5
0
20
40
60
80
Gene count
D
Density
Ks value
GB-GB
GB-TB
TB-TB
E
Biological process
embryo development
organ development
fruit development
response to abiotic stimulus
metabolic process
developmental process
shoot system development
response to hormone
phyllome development
response to radiation
response to light stimulus
response to acid chemical
lipid metabolic process
response to temperature stimulus
cell division
response to cytokinin
developmental growth
lipid oxidation
cell wall biogenesis
lateral root formation
flavonoid metabolic process
phenylpropanoid metabolic process
terpenoid catabolic process
lignin metabolic process
-log10(p-value)
30
20
10
0
2000
5000
6000
Gene count

Figure 3

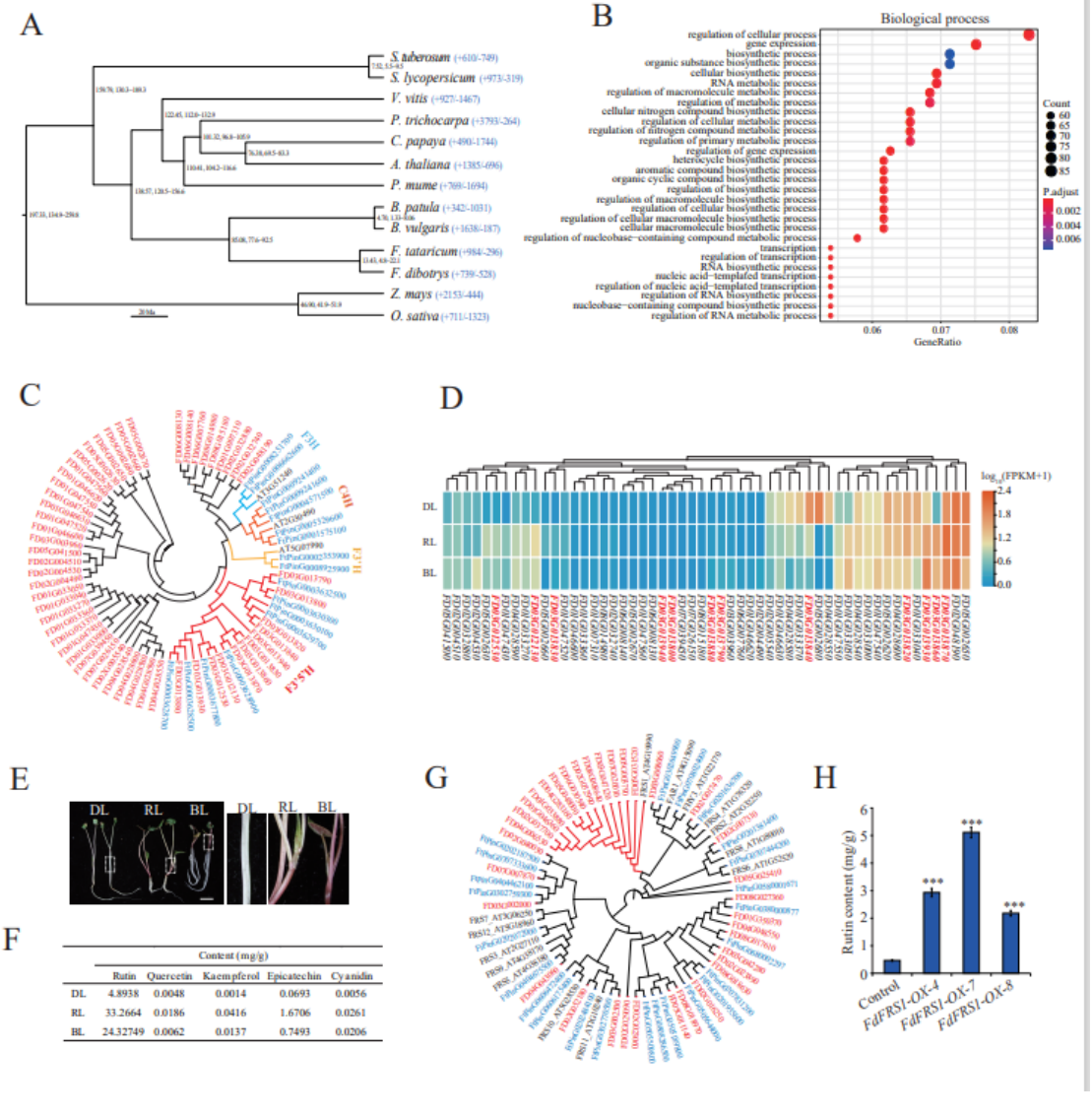
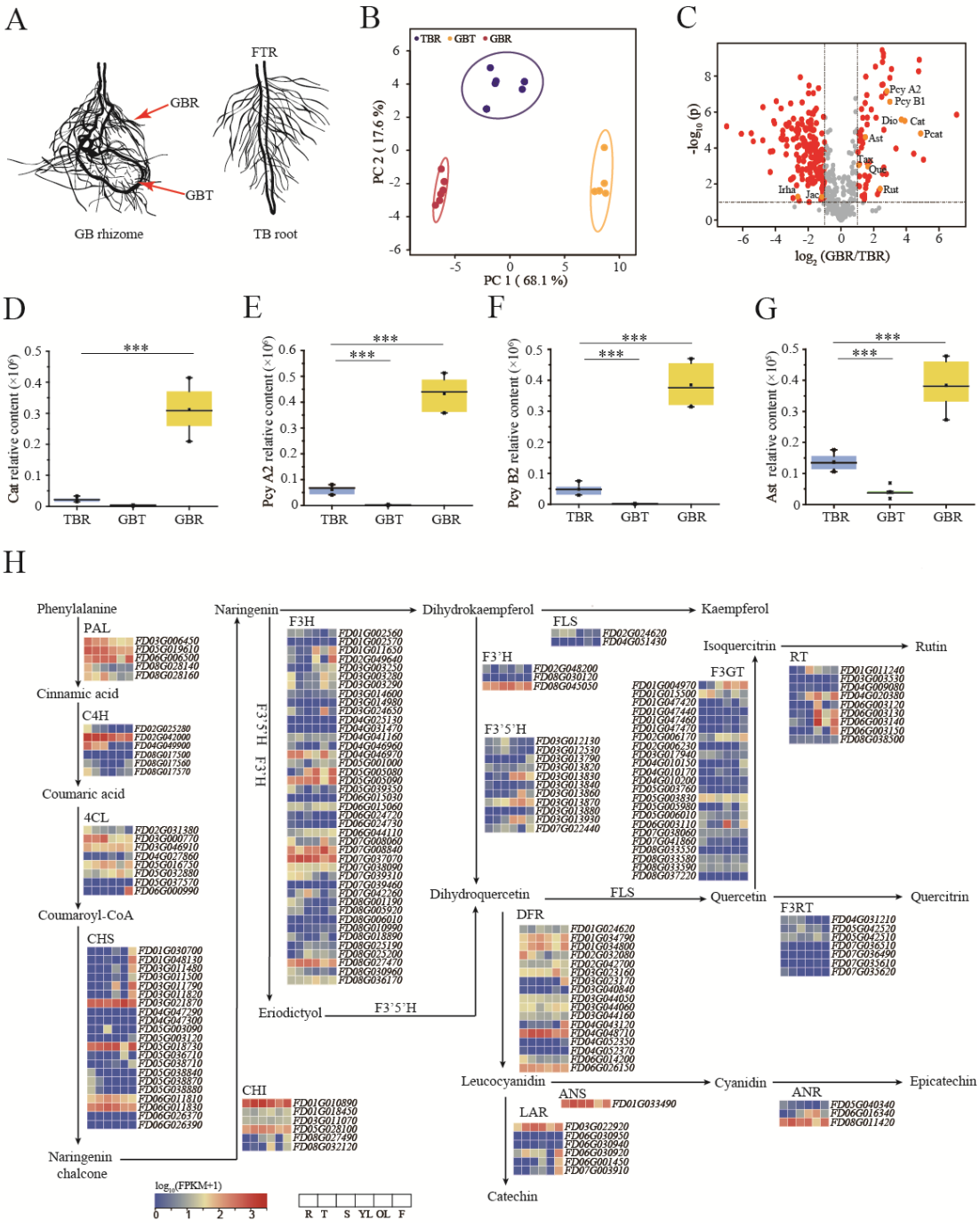

A
S. tuberosum (+610/-749)
S. lycopersicum (+973/-319)
V. vitis (+927/-1467)
P. trichocarpa (+3793/-264)
C. papaya (+490/-1744)
A. thaliana (+1385/-696)
P. mume (+769/-1694)
B. patula (+342/-1031)
B. vulgaris (+1638/-187)
F. tataricum (+984/-296)
F. dibotrys (+739/-528)
Z. mays (+2153/-444)
O. sativa (+711/-1323)
B
Biological process
regulation of cellular process
gene expression
biosynthetic process
organic substance biosynthetic process
cellular biosynthetic process
RNA metabolic process
regulation of macromolecule metabolic process
regulation of metabolic process
cellular nitrogen compound biosynthetic process
regulation of cellular metabolic process
regulation of nitrogen compound metabolic process
regulation of primary metabolic process
regulation of gene expression
heterocycle biosynthetic process
aromatic compound biosynthetic process
organic cyclic compound biosynthetic process
regulation of biosynthetic process
regulation of macromolecule biosynthetic process
regulation of cellular biosynthetic process
regulation of cellular macromolecule biosynthetic process
cellular macromolecule biosynthetic process
regulation of nucleobase−containing compound metabolic process
transcription
regulation of transcription
RNA biosynthetic process
nucleic acid−templated transcription
regulation of nucleic acid−templated transcription
regulation of RNA biosynthetic process
nucleobase−containing compound biosynthetic process
regulation of RNA metabolic process
0.06
0.07
0.08
GeneRatio
Count
60
65
70
75
80
85
P.adjust
0.002
0.004
0.006
C
F3H
C4H
F3'H
F3'5'H
D
DL
RL
BL
log2(FPKM+1)
2.4
1.8
1.2
0.6
0.0
E
DL
RL
BL
DL
RL
BL
F
Content (mg/g)
Rutin Quercetin Kaempferol Epicatechin Cyanidin
DL 4.8938 0.0048 0.0014 0.0693 0.0056
RL 33.2664 0.0186 0.0416 1.6706 0.0261
BL 24.32749 0.0062 0.0137 0.7493 0.0206
G
H
Rutin content (mg/g)
0
1
2
3
4
5
6
***
***
***
Control
FdFRS1-OX-4
FdFRS1-OX-7
FdFRS1-OX-8

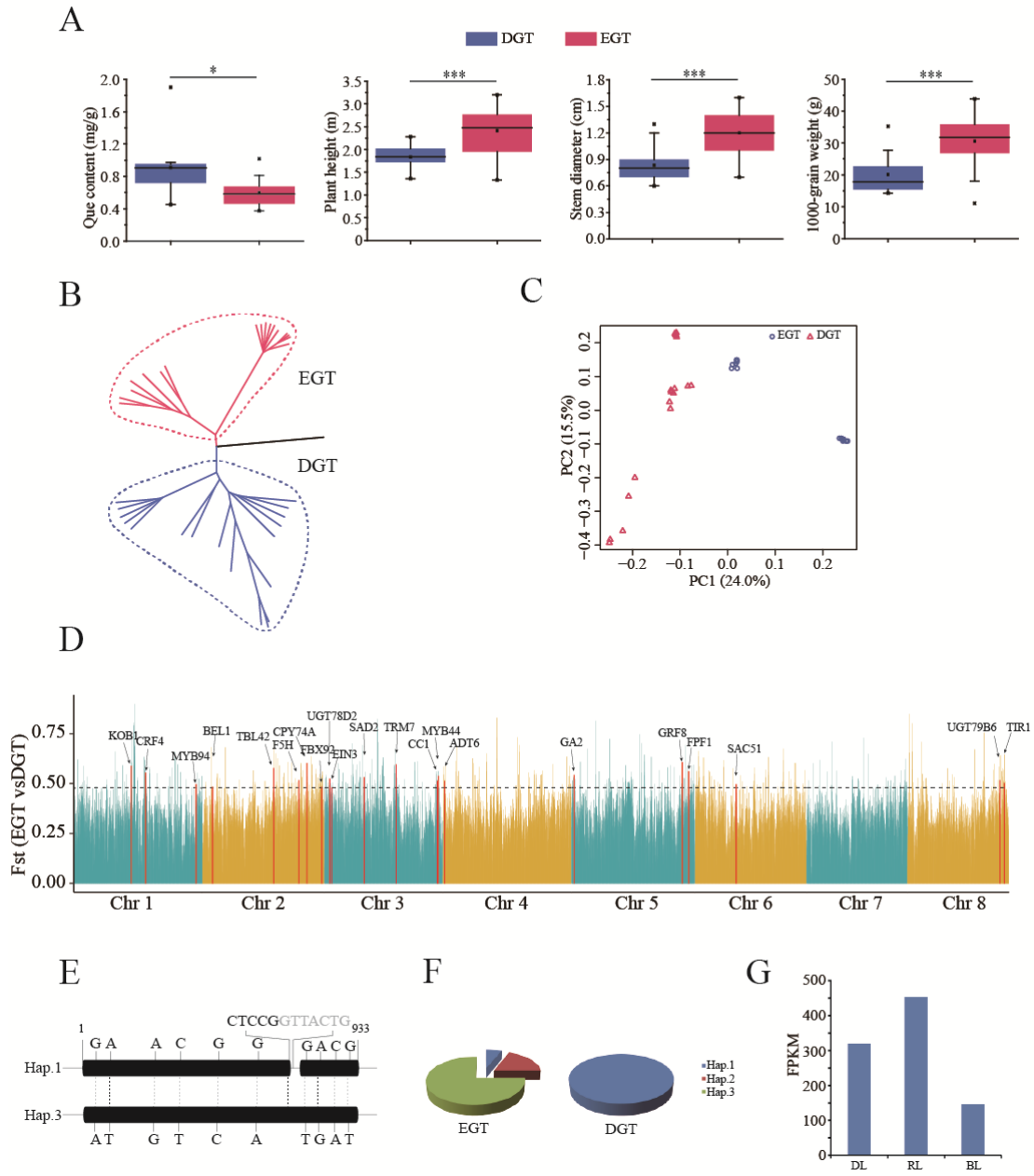
Figure 4


A
GBR
GBT
FTR
GB rhizome
TB root
B
TBR GBT GBR
PC 2 (17.6 %)
PC 1 (68.1 %)
C
Pcy A2
Pcy B1
Dio
Cat
Pcat
Ast
Tax
Qua
Rut
Irha
Jac
-log10 (p)
log2 (GBR/TBR)
D
Cat relative content (×10⁶)
E
Pcy A2 relative content (×10⁶)
F
Pcy B2 relative content (×10⁶)
G
Ast relative content (×10⁵)
TBR GBT GBR
***
H
Phenylalanine
PAL
Cinnamic acid
C4H
Coumaric acid
4CL
Coumaroyl-CoA
CHS
Naringenin chalcone
CHI
Naringenin
F3H
F3'5'H
F3'H
Eriodictyol
Dihydrokaempferol
FLS
Kaempferol
F3'H
F3'5'H
Dihydroquercetin
FLS
Quercetin
Isoquercitrin
F3GT
RT
Rutin
F3RT
Quercitrin
DFR
Leucocyanidin
ANS
Cyanidin
ANR
Epicatechin
LAR
Catechin
log10(FPKM+1)
R T S YL OL F

7

8

Figure 5


A
DGT
EGT
Que content (mg/g)
Plant height (m)
Stem diameter (cm)
1000-grain weight (g)
*
***
***
***
B
EGT
DGT
C
EGT
DGT
PC2 (15.5%)
PC1 (24.0%)
D
Fst (EGT vsDGT)
KOB1
CRF4
MYB94
BEL1
TBL42
CPY74A
F5H
FBX92
UGT78D2
EIN3
SAD2
TRM7
CC1
MYB44
ADT6
GA2
GRF8
FPF1
SAC51
UGT79B6
TIR1
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
E
CTCCGGTTACTG
Hap.1
Hap.3
F
EGT
DGT
Hap.1
Hap.2
Hap.3
G
FPKM
DL
RL
BL