A highly contiguous genome assembly of red perilla (*Perilla frutescens*) domesticated in Japan

Keita Tamura[1,2], Mika Sakamoto[3], Yasuhiro Tanizawa[3], Takako Mochizuki[3], Shuji Matsushita[4], Yoshihiro Kato[5], Takeshi Ishikawa[5], Keisuke Okuhara[1,6], Yasukazu Nakamura[3], Hidemasa Bono[1,2*]

[1]Laboratory of Genome Informatics, Graduate School of Integrated Sciences for Life, Hiroshima University, 3-10-23 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-0046, Japan

[2]Laboratory of BioDX, Genome Editing Innovation Center, Hiroshima University, 3-10-23 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-0046, Japan

[3]Genome Informatics Laboratory, Department of Informatics, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

[4]Agricultural Technology Research Center, Hiroshima Prefectural Technology Research Institute, 6869 Hara, Hachihonmatsucho, Higashi-Hiroshima, Hiroshima 739-0151, Japan

[5]Mishima Foods Co., Ltd., 3-10-7 Kanonshinmachi, Nishi-ku, Hiroshima City, Hiroshima 733-0036, Japan

[6]PtBio Inc., 3-10-23 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-0046, Japan

Running Head: Genome sequencing of *Perilla frutescens*

*Correspondence: bonohu@hiroshima-u.ac.jp; Tel.: +81-82-424-4013

Summary

Perilla frutescens (Lamiaceae) is an important herbal plant with hundreds of bioactive chemicals, among which perillaldehyde and rosmarinic acid are the two major bioactive compounds in the plant. The leaves of red perilla are used as traditional Kampo medicine or food ingredients. However, the medicinal and nutritional uses of this plant could be improved by enhancing the production of valuable metabolites through the manipulation of key enzymes or regulatory genes using genome editing technology. To this end, the construction of a high-quality reference genome sequence is necessary. Here, we generated a high-quality genome assembly of red perilla domesticated in Japan. A near-complete chromosome level assembly of *P. frutescens* was generated contigs with N50 of 41.5 Mb from PacBio HiFi reads. The contigs were ordered into 20 pseudochromosomes by Omni-C chromosome conformation capture technique and alignment with the previous genome assembly of the species. Additionally, 99.2% of the assembly was anchored into 20 pseudochromosomes, among which seven pseudochromosomes consisted of one contig, while the rest consisted of less than six contigs. Gene annotation and prediction of the sequences, including PacBio isoform sequencing (Iso-Seq) and RNA sequencing (RNA-Seq) data, as well as BRAKER2 successfully predicted 86,258 gene models, including 76,825 protein-coding genes. Further analysis showed that potential targets of genome editing for the engineering of anthocyanin pathways in *P. frutescens* are located on the late-stage pathways. Overall, our genome assembly could serve as a valuable reference for selecting target genes for genome editing of *P. frutescens*.

Keywords

Perilla frutescens, genome assembly, PacBio HiFi reads.

Significance statement

We sequenced and assembled the allotetraploid genome of red perilla (*Perilla frutescens*) ($2n = 4x = 40$) domesticated in Japan using PacBio HiFi reads, Omni-C chromosome conformation capture technique, and alignment with the previous genome assembly. Seven of the 20 pseudochromosomes consisted of one contig, and the rest consisted of less than six contigs, indicating that our genome assembly is highly contiguous and could serve as a good reference genome sequence of *P. frutescens*.

INTRODUCTION

Perilla frutescens is an annual herbal plant belonging to the Lamiaceae family, and it is widely cultivated in Asian countries (Ahmed, 2019). *P. frutescens* is an allotetraploid ($2n = 4x = 40$) species, and *P. citriodora* ($2n = 20$) is believed to be one of the diploid genome donors (Nitta *et al.*, 2005). There are two chemotypes of perilla plants based on the content of anthocyanins: red and green perilla. Red perilla ("aka-shiso" in Japanese) is an anthocyanin-rich variety with dark red or purple leaves and stems, while green perilla ("ao-jiso" in Japanese) is an anthocyanin deficient variety with green leaves and stems (Saito and Yamazaki, 2002). Both green and red perilla leaves are often used as a material for cooking. Particularly, the leaves of red perilla are used as traditional Kampo medicine "soyou" to treat stomach problems (Ueda *et al.*, 2002; Nitta *et al.*, 2005; Ahmed, 2019; Deguchi and Ito, 2020), and the seeds are also used to produce oil. Perilla seed oil is a rich source of α-linolenic acid, and its potential health benefits have been reported (Longvah *et al.*, 2000; Hashimoto *et al.*, 2021).

Thus far, hundreds of bioactive compounds have been identified in *P. frutescens* (Ahmed, 2019; Hou *et al.*, 2022), among which perillaldehyde (monoterpenenoid) and rosmarinic acid (phenylpropanoid) are major phytochemicals (Yoshida *et al*., 2021). Perillaldehyde has been shown to possess anti-inflammatory (Uemura *et al.*, 2018), antidepressant (Ji *et al.*, 2014), antifungal, and antibacterial activities (Sato *et al.*, 2006; Tian *et al.*, 2016); additionally, rosmarinic acid possesses antiviral, antibacterial, and anti-inflammatory activities (Petersen and Simmonds, 2003). Several enzymes for the biosynthesis of these compounds have been identified in *P. frutescens*. Perillaldehyde appears to be biosynthesized by the hydroxylation and subsequent oxidation of the C-7 position of limonene. Limonene synthase and a cytochrome P450 monooxygenase (P450) catalyzing the two-step oxidation of limonene at C-7 position have been cloned and characterized in *P. frutescens* (Yuba *et al.*, 1996; Fujiwara and Ito, 2017). Rosmarinic acid is proposed to be biosynthesized from 4-coumaroyl-CoA and 4-hydroxyphenyllactic acid (Trócsányi *et al.*, 2020). Rosmarinic acid synthase, which is the first specific enzyme for rosmarinic acid biosynthesis, catalyzes the ester formation step of these two compounds. After 4-coumaroyl-4′-hydroxyphenyllactic acid formation, P450 enzymes belonging to the CYP98A family member are known to catalyze the final hydroxylation steps leading to rosmarinic acid production (Petersen and Simmonds, 2003; Trócsányi *et al.*, 2020). These enzymes have been cloned and characterized from several plant species, including *Coleus scutellarioides* (Lamiaceae); however, none has been identified in perilla plants.

Recently, genome editing tools, such as CRISPR-Cas9, have been used for the engineering of plant biosynthetic pathways (Nishida and Kondo, 2021). For instance, deletion of the autoinhibitory domain of glutamate decarboxylase 3 (*GAD3*) using CRISPR-Cas9 technology promoted the accumulation of GABA (γ-aminobutyric acid) in tomato fruit, and this product is already commercially available (Nonaka *et al.*, 2017; Ezura, 2022). Additionally, silencing of the potato sterol side chain reductase 2 (*SSR2*) by transcription activator-like effector nucleases (TALEN) suppressed the accumulation of toxic steroidal glycoalkaloids (Sawai *et al.*, 2014). Therefore, genome sequencing and gene annotation could facilitate the identification of target genes to enhance desired traits, such as higher contents of valuable compounds and lower contents of unwanted compounds. Additionally, the genome of edited plants could be compared with that of the reference genome to

identify the potential risk of off-target changes (Graham *et al.*, 2020; Sturme *et al.*, 2022). Long-read DNA sequencing technologies have emerged as powerful tools to obtain high-quality whole genome sequences (Logsdon *et al.*, 2020). Recently developed PacBio technology has facilitated the generation of high-fidelity (HiFi) reads from circular consensus sequencing (CCS), with long-read (> 10 kb) and high accuracy (> 99%) (Logsdon *et al.*, 2020). In plant genome sequencing, highly contiguous near-chromosomal level sequences have been generated using HiFi read-only (Sharma *et al.*, 2022).

Here, we generated a highly contiguous genome assembly of red perilla (*P. frutescens*) domesticated in Japan using PacBio HiFi reads. Functional annotation of identified genes was obtained using a systematic functional annotation workflow optimized for plants. It is anticipated that the highly contiguous genome assembly obtained in this study would promote the development of perilla varieties with desirable traits.

RESULTS

De novo assembly of red perilla cultivar Hoko-3

We performed *de novo* assembly of the red perilla cultivar Hoko-3 (Figure 1) from 72.4 Gb (57.5× coverage) of PacBio HiFi reads using Hifiasm (Cheng *et al.*, 2021; Cheng *et al.*, 2022). Hi-C (Omni-C) integrated assembly of Hifiasm was performed by combining the Omni-C reads and "--primary" option to generate primary and alternate contigs, as well as fully phased haplotype 1 and 2 contigs. We specified the homozygous coverage in the parameter setting of Hifiasm (--hom-cov 55) because the default setting could misidentify the coverage threshold for homozygous reads. The Hifiasm outputs generated 317 primary contigs and 14,150 alternate contigs (Table S1). *K*-mer evaluation using Merqury (Rhie *et al.*, 2020) showed that fully phased haplotype 1 and 2 were almost identical (Figure S1); therefore, we did not distinguish the fully phased haplotypes for further analysis. Merqury analysis indicated high base accuracy (QV = 60.1) and completeness (98.3%) of the primary contigs (Table S1).

Additionally, the Omni-C reads were mapped to the primary contigs, and then scaffolding was performed using SALSA2 (Ghurye *et al.*, 2019) to construct pseudochromosomes. A total of 298 scaffolds were generated from the 319 primary contigs (two contigs were broken during the scaffolding), among which 25 were longer than 1 Mb (Table S2). The 25 scaffolds (> 1 Mb) were aligned against the previously assembled chromosome-scale genome of green perilla cultivar PF40 (Zhang *et al.*, 2021) using MUMmer4 (Marçais *et al.*, 2018) (Figure S2). Among the longest 20 scaffolds (scaffold_1–20), 19 (except for scaffold_19) covered each chromosome of the PF40 genome. Scaffold_19 partially covered chromosome 15 of the PF40 genome, and two other scaffolds (scaffold_21 and 22) partially aligned with chromosome 15 of the PF40 genome (Figure S2). Similarly, Hi-C contact map indicated that the three scaffolds (scaffold_19, 21, and 22) corresponded to the same chromosome (Figure S3); therefore, the three scaffolds were combined based on their partial alignment to chromosome 15 of the PF40 genome. We renamed and sorted the scaffolds based on the alignment with the PF40 genome to construct 20 pseudochromosomes (Figure 2). After removal of scaffolds predicted to be derived from mitochondria or chloroplast genome, we obtained 71 scaffolds with N50 of 63.3 Mb (Pfru_yukari_1.0; Table 1). The N50 value of the scaffolds was almost similar to that of the PF40 assembly; however, the N50

122 value of the contigs was 41.5 Mb, which was ten times more than that of the PF40 assembly (Table 1). Each of
123 the 20 pseudochromosomes consisted of less than six contigs, of which seven pseudochromosomes consisted
124 of only one contig, indicating a highly contiguous genome assembly (Table S3). Additionally, 99.2 % of the
125 assembly was assigned to 20 pseudochromosomes (Table S3). Completeness of the genome assembly evaluated
126 with benchmarking universal single-copy orthologs (BUSCO) (Manni *et al.*, 2021) showed that the assembly
127 achieved almost complete coverage of the BUSCO core gene sets (99.5% completeness) (Table 1).
128

129 **Annotation of the Hoko-3 genome**

130 The length of repetitive sequences in the Hoko-3 genome was 866.7 Mb (68.84% of the genome) (Table 2).
131 Long terminal repeat (LTR) elements accounted for 37.07% of the genome, with Copia and Gypsy constituting
132 14.01 and 14.92%, respectively (Table 2). Gene annotation of the Hoko-3 genome was performed by merging
133 the gene models generated by Iso-Seq and RNA-Seq data, and gene prediction using BRAKER2 (Brůna *et al.*,
134 2021) in this order. A total of 86,258 gene models were predicted with the 98.7% BUSCO complete data (Table
135 3). Among each of the three different annotation methods, RNA-Seq alone gave 97.0% of BUSCO completeness
136 and BRAKER2 alone gave 98.5% of BUSCO completeness (Table S4); however, the combination of the Iso-
137 Seq and RNA-Seq gave higher BUSCO completeness (97.7%), and addition of the predicted model from
138 BRAKER2 achieved 98.7% BUSCO completeness (Table 3). The gene models were subjected to
139 Fanflow4Plants, designed for the functional annotation of plant species based on Fanflow4Insects (Bono *et al.*,
140 2022). Among the 86,258 gene models, 76,825 gene models were predicted as protein-coding genes, among
141 which 72,983 gene models were annotated to at least one of the reference sequences in GGSERACH or pfam
142 domain by HMMSCAN (Table 4).
143

144 **Identification of the genes related to specialized metabolites in Hoko-3**

145 As a practical example of genome editing target selection, enzyme coding genes in the anthocyanin biosynthetic
146 pathway were studied. The major anthocyanin in *P. frutescens* is malonylshisonin, which is a glycosylated form
147 of cyanidin (Saito and Yamazaki, 2002). After the curation of the genes, the number of enzyme coding genes
148 in the anthocyanin biosynthetic pathway in the genome were listed, including putative isoforms (Figure 3).
149 There were multiple copies of enzyme genes upstream of the pathway in the genome, but downstream enzyme
150 genes were encoded in only a few locations in the genome. Based on this observation, it could be concluded
151 that genes downstream of the pathway are most likely targets for genome editing to engineer the anthocyanin
152 biosynthetic pathway in *P. frutescens*.
153

154 **DISCUSSION**

155 Here, we generated a chromosome-level genome assembly of *P. frutescens* domesticated in Japan, using PacBio
156 HiFi reads. Seven of the 20 pseudochromosomes were composed of only one contig, and the other
157 pseudochromosomes were composed of not more than five contigs (Table S3), indicating that the contigs
158 generated from HiFi reads achieved a near-complete chromosome level. Recently, near-complete chromosome

159 level assembly of *Macadamia jansenii* genome was generated from HiFi reads, with eight of the 14
160 pseudochromosomes represented by a single large contig (Sharma *et al*., 2022), which is comparable to our
161 assembly. Although it is difficult to construct complete chromosomal-level genome assembly from HiFi read-
162 only, it is now possible to obtain near-complete chromosome-level assembly simply by running a HiFi read
163 assembler, including Hifiasm.

164 The number of gene models annotated in this study by combining two evidence-based annotations (Iso-
165 Seq and RNA-Seq) and the gene prediction method (BRAKER2) was 86,258 (Table 3), which is almost twice
166 the previously assembled *P. frutescens* genome (43,527 genes) (Zhang *et al*., 2021) and close to the number of
167 genes reported in another Lamiaceae tetraploid species *Salvia splendens* (88,489 genes) (Jia *et al*., 2021). The
168 gene models generated in the present study achieved extremely high BUSCO completeness (98.7%) (Table 3),
169 indicating that the models could be valuable resources for gene functional analysis of *P. frutescens*. Furthermore,
170 an annotation system named Fanflow4Plants was developed based on the Fanflow4Insects for the functional
171 annotation of gene models (Bono *et al*., 2022). Only well-curated protein datasets were used as references in
172 this system to obtain reliable functional annotations, including protein sequences of three plant species
173 (Arabidopsis, rice, and tomato) and two mammalian species (human and mouse), as well as UniProtKB/Swiss-
174 Prot. Overall, 72,339 of 76,825 (94.2%) protein-coding genes were functionally annotated to at least one of the
175 reference sequences (Table 4).

176 Since *P. frutescens* is a rich source of several metabolites (Ahmed, 2019; Hou *et al*., 2022), metabolic
177 engineering could be used to enhance the biosynthesis and accumulation of valuable compounds in this species.
178 Although genome editing of *P. frutescens* has not yet been reported, recent advances have shown that genome
179 editing could be done using *Agrobacterium*-mediated transformation (Kim *et al*., 2004). In the present study,
180 potential targets of genome editing to manipulate the anthocyanin biosynthetic pathway were identified (Figure
181 3). As anthocyanin and rosmarinic acid share the upstream biosynthetic pathway towards 4-coumaroyl-CoA, it
182 could be possible to change the metabolic flux into the biosynthesis of rosmarinic acid by knocking down the
183 specific pathway for anthocyanin biosynthesis. Similar approaches could be used to identify target genes to
184 enhance the biosynthesis of perillaldehyde or other beneficial compounds by examining the functional
185 annotation of this genome assembly. Additionally, further analysis showed that the *P. frutescens* Hoko-3
186 cultivar possessed a highly homozygous genome (Figure S1), which could be due to the fact that *P. frutescens*
187 is a self-fertilizing crop (Sa *et al*., 2012). This homozygosity would be beneficial for the selection of unique
188 targets for genome editing. Overall, our genome assembly and annotation could serve as a unique resource for
189 future genome editing studies of *P. frutescens*.
190
191

Experimental procedures

Sample preparation and genome sequencing

DNA sample for genome sequencing was isolated from the young leaves of hydroponically grown *P. frutescens* cultivar (Hoko-3) using a Genomic-tip kit (Qiagen, Hilden, Germany). Library preparation was performed using SMRTbell Express Template Prep Kit 2.0 (PacBio, Menlo Park, CA, USA), and reads longer than 20 kb were collected using BluePippin (Sage Science, Beverly, MA, USA). Thereafter, the libraries were sequenced using Sequel IIe instrument (PacBio), and HiFi reads were selected from the circular consensus reads generated.

Omni-C sequencing

Dovetail Omni-C libraries were prepared using Omni-D kit (Dovetail Genomics, CA, USA), according to the manufacturer's protocol. Sequencing was performed using DNBSEQ-G400 (MGI Tech, Shenzhen, China) in a 2 × 150 bp paired-end (PE) setting to obtain 356.34 million PE reads. The obtained reads were processed using fastp v0.23.1 software (Chen *et al.*, 2018) with default settings.

De novo assembly of HiFi reads

HiFi reads were assembled using Hifiasm v0.16.1 (Cheng *et al.*, 2021; Cheng *et al.*, 2022) with the combination of processed Omni-C reads with --hom-cov 55 --primary options (Hi-C integrated assembly).

Quality assessment of genome assemblies

Assembly statistics were obtained using QUAST v5.0.2 (Mikheenko *et al.*, 2018). Genome completeness was evaluated using BUSCO v5.2.2 software (Manni *et al.*, 2021) against embryophyta_odb10 (eukaryota, 2020-09-10) dataset (1,614 total BUSCO groups). *K*-mer based assembly evaluation was performed using Merqury v1.3 (Rhie *et al.*, 2020) and the meryl db (k = 21) generated from the HiFi reads.

Omni-C scaffolding and construction of pseudochromosomes

The processed Omni-C reads were mapped to the primary contigs generated by Hifiasm using BWA-MEM v0.7.17 (Li, 2013), followed by the removal of the 3′-side of the chimeric mapping and merging of the paired BAM files using perl scripts from Arima Genomics mapping_pipeline (https://github.com/ArimaGenomics/mapping_pipeline) and SAMtools v1.12 (Li *et al.*, 2009). The processed BAM file was converted to BED format using BEDtools v2.30.0 (Quinlan and Hall, 2010), then scaffolding was performed using SALSA2 v2.3 (Ghurye *et al.*, 2019) with -e DNASE -p yes options. To draw a Hi-C contact map, the output of SALSA2 was converted to .hic file using convert.sh script equipped with SALSA2, then the generated .hic file was visualized using Juicebox v1.11.08 (Durand *et al.*, 2016). A total of 25 scaffolds of 1 Mb or longer were extracted using SeqKit v2.0.0 (Shen *et al.*, 2016) and aligned with the 20 pseudochromosomes of the PF40 genome (GenBank assembly accession GCA_019511825.2; Zhang *et al.*, 2021) using nucmer sequence aligner with default settings and filtering of the delta alignment file with delta-filter command with -q -r options in MUMmer4 v4.0.0rc1 package (Marçais *et al.*, 2018). The alignment was

229 drawn with a custom R script (original: https://jmonlong.github.io/Hippocamplus/2017/09/19/mummerplots-
230 with-ggplot2/). To create pseudochromosomes based on the alignment with the PF40 genome, scaffold_19 and
231 the reverse complement sequences of scaffold_22 and scaffold_21 were joined in this order with gaps (500 Ns)
232 to create a new scaffold (scaffold_19). The longest 20 scaffolds (scaffold_1–20) of this assembly were renamed
233 following the PF40 genome. Additionally, 10 of the 20 scaffolds (scaffold_2, 3, 6, 8, 13, 14, 15, 16, 17, and 18)
234 were reverse complemented to align with the PF40 genome in the same direction.
235
236 **Removal of organellar sequences**
237 The scaffolds were searched against chloroplast and mitochondrial genome sequences obtained from NCBI
238 RefSeq by blastn v2.12.0 with an E-value cutoff of 1e-5. Scaffolds with 90% or more coverage against one of
239 the reference organellar sequences were labeled as organellar scaffolds and removed from the primary assembly.
240 All of the scaffolds removed in this step were made from a single contig. We also performed the same organellar
241 removal procedure for the alternate haplotigs generated by Hifiasm.
242
243 **Transcriptome analysis**
244 Total RNA for PacBio Iso-Seq was extracted from the mixed sample of leaves, stems, and roots of the Hoko-3
245 cultivar using RNeasy Plant Mini kit (Qiagen, Hilden, Germany). Library preparation was performed using
246 NEBNext Single Cell/Low Input cDNA Synthesis & Amplification Module (New England Biolabs, Ipswich,
247 MA, USA) and SMRTbell Express Template Prep Kit 2.0 (PacBio). Thereafter, the reads were sequenced using
248 the Sequel II instrument (PacBio), and CCS reads were generated using SMRTLink v10.0 (PacBio). The
249 obtained BAM file was processed using IsoSeq v3.4.0 pipeline. Total RNA for RNA-Seq was extracted from
250 leaves of the Hoko-3 cultivar using ISOSPIN Plant RNA kit (Nippon Gene, Tokyo, Japan), and a sequencing
251 library was prepared using NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs) and
252 NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England Biolabs). Sequencing was
253 performed using NovaSeq 6000 (Illumina, San Diego, CA, USA) in a 2 × 150 bp paired-end (PE) setting. Reads
254 from three biological replicates (154.20 million PE reads in total) were combined and used for gene prediction
255 and annotation.
256
257 **Gene prediction and annotation**
258 The processed Iso-Seq reads (high-quality isoforms) were mapped to the assembled genome using minimap2
259 v2.23 (Li, 2018), then collapsed to obtain non-redundant isoforms using Cupcake ToFU scripts on
260 cDNA_Cupcake v28.0.0 (https://github.com/Magdoll/cDNA_Cupcake). The RNA-Seq reads were processed
261 using fastp v0.23.2 with default settings, mapped to the assembled genome using HISAT2 v2.2.1 (Kim et al.,
262 2019), and then transcript models were constructed using StringTie v2.2.1 (Pertea *et al.*, 2015). Coding
263 sequences from these two annotations were identified using GenomeTools v1.6.2 (Gremme *et al.*, 2013), and
264 gene features were obtained using a custom python script. The RNA-Seq reads were mapped to the assembled
265 genome to predict protein-coding genes using BRAKER2 v2.1.6 (Brůna *et al.*, 2021). For the input of

BRAKER2, the assembled genome was repeat masked using RepeatModeler v2.0.2 and RepeatMasker v4.1.2 in Dfam TE Tools Container v1.4 (https://github.com/Dfam-consortium/TETools). These three annotations (Iso-Seq, RNA-Seq, and BRAKER2) were merged in this order by adding the additional features of the second GFF file obtained using BEDtools v2.30.0 and removal of incomplete gene features using a custom python script.

Functional annotation

Functional annotation of the protein-coding genes on the primary assembly was performed using Fanflow4Plants, designed for the functional annotation of plant species based on Fanflow4Insects (Bono *et al*., 2022). In the functional annotation of these protein coding sequences, these sequences were searched by GGSERACH v36.3.8g in the FASTA package (https://fasta.bioch.virginia.edu/). Functionally well-curated protein datasets of *Arabidopsis thaliana*, UniProtKB/Swiss-Prot, *Oryza sativa*, *Solanum lycopersicum*, *Homo sapiens*, and *Mus musculus* were used as a reference. The sequences were also searched by HMMSCAN in HMMER package v3.3.2 (http://hmmer.org/) against the hidden Markov model (HMM) profile libraries of Pfam database v35.0 (Mistry *et al*., 2021).

Acknowledgements

We thank Masaki Kurao (Hiroshima Prefectural Technology Research Institute) for technical assistance. This work was supported by Hiroshima Prefectural Government, the Center of Innovation for Bio-Digital Transformation (BioDX), an open innovation platform for industry-academia co-creation of JST (COI-NEXT, JPMJPF2010) and JSPS KAKENHI Grant 21K19118 to HB. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

Author contributions

Conceptualization: KO, HB, YN
Methodology: KT, YT, MS, TM
Software: KT, YT, MS, HB
Validation: KT, YT, MS
Formal analysis: KT, YT, MS, HB
Investigation: KT, YT, MS, HB, SM
Resources: KT, YT, MS, HB, SM, YK, TI
Data Curation: KT, YT, MS, HB
Writing - Original Draft: KT
Writing - Review & Editing: MS, YT, TM, SM, YK, TI, KO, YN, HB
Visualization: KT, YT, MS, HB
Supervision: YN, HB
Project administration: HB
Funding acquisition: KO, HB

Conflict of interest

The authors declare no conflict of interest.

Data availability statement

The genome assembly from primary contigs has been deposited in DDBJ under the accession numbers BRKX01000001 to BRKX01000071. A set of haplotigs only sequences have been deposited in DDBJ under the accession numbers BRKY01000001 to BRKY01012627. The raw sequence reads have been deposited in DDBJ under the accession numbers DRR361636 (PacBio HiFi reads), DRR361637 (PacBio Iso-Seq reads), and DRR361638 (Illumina RNA-Seq reads). Gene annotation and functional annotation of the protein coding genes are available at figshare (https://doi.org/10.6084/m9.figshare.20780995). Custom scripts used in this study are available at figshare (https://doi.org/10.6084/m9.figshare.20781466).

319 **Supporting Information**

320 Additional Supporting Information is available at figshare (https://doi.org/10.6084/m9.figshare.20780419).

321

322 **Legends for supporting information**

323 **Table S1.** Statistics of the contigs generated by Hifiasm.

324 **Table S2.** AGP file (scaffolds_FINAL.agp) generated by SALSA2 describing the contig assignment of each
325 scaffold.

326 **Table S3.** Length and the components of 20 pseudochromosomes.

327 **Table S4.** Summary of the gene annotation of each of the three methods.

328 **Figure S1.** Merqury assembly spectrum plots of haplotype 1 (hap1) and haplotype 2 (hap2).

329 **Figure S2.** Dot-plot alignment between the draft scaffolds (indicated as "s") (longest 25 scaffolds; scaffold_1–
330 25) and reference PF40 genome assembly. Blue dots represent +/+ strand alignments and red dots represent +/-
331 strand alignments.

332 **Figure S3.** Hi-C contact map of the draft scaffolds (indicated as "s") generated by SALSA2.

333

334

References

Ahmed, H.M. (2019) Ethnomedicinal, phytochemical and pharmacological investigations of *Perilla frutescens* (L.) Britt. *Molecules*, **24**, 102.

Bono, H., Sakamoto, T., Kasukawa, T. and Tabunoki, H. (2022) Systematic functional annotation workflow for insects. *Insects*, **13**, 586.

Brůna, T., Hoff, K.J., Lomsadze, A., Stanke, M. and Borodovsky, M. (2021) BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. Bioinform.*, **3**, lqaa108.

Chen, S., Zhou, Y., Chen, Y. and Gu, J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. and Li, H. (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods*, **18**, 170–175.

Cheng, H., Jarvis, E.D., Fedrigo, O., Koepfli, K.-P., Urban, L., Gemmell, N.J. and Li, H. (2022) Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.*, **40**, 1332–1335.

Deguchi, Y. and Ito, M. (2020) Rosmarinic acid in *Perilla frutescens* and perilla herb analyzed by HPLC. *J. Nat. Med.*, **74**, 341–352.

Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S. and Aiden, E.L. (2016) Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.*, **3**, 99–101.

Ezura, H. (2022) Letter to the editor: The world's first CRISPR tomato launched to a Japanese market: The social-economic impact of its implementation on crop genome editing. *Plant Cell Physiol.*, **63**, 731–733.

Fujiwara, Y. and Ito, M. (2017) Molecular cloning and characterization of a *Perilla frutescens* cytochrome P450 enzyme that catalyzes the later steps of perillaldehyde biosynthesis. *Phytochemistry*, **134**, 26–37.

Ghurye, J., Rhie, A., Walenz, B.P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A.M. and Koren, S. (2019) Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput. Biol.*, **15**, e1007273.

Graham, N., Patil, G.B., Bubeck, D.M., et al. (2020) Plant genome editing and the relevance of off-target changes. *Plant Physiol.*, **183**, 1453–1471.

Gremme, G., Steinbiss, S. and Kurtz, S. (2013) GenomeTools: A comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **10**, 645–656.

Hashimoto, M., Matsuzaki, K., Hossain, S., et al. (2021) Perilla seed oil enhances cognitive function and mental health in healthy elderly Japanese individuals by enhancing the biological antioxidant potential. *Foods*, **10**, 1130.

Hou, T., Netala, V.R., Zhang, H., Xing, Y., Li, H. and Zhang, Z. (2022) *Perilla frutescens*: A rich source of pharmacological active compounds. *Molecules*, **27**, 3578.

Ji, W.-W., Wang, S.-Y., Ma, Z.-Q., et al. (2014) Effects of perillaldehyde on alternations in serum cytokines and depressive-like behavior in mice after lipopolysaccharide administration. *Pharmacol. Biochem. Behav.*, **116**,

1–8.

Jia, K.-H., Liu, H., Zhang, R.-G., et al. (2021) Chromosome-scale assembly and evolution of the tetraploid *Salvia splendens* (Lamiaceae) genome. *Hortic. Res.*, **8**, 177.

Kim, D., Paggi, J.M., Park, C., Bennett, C. and Salzberg, S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**, 907–915.

Kim, K.-H., Lee, Y.-H., Kim, D., Park, Y.-H., Lee, J.-Y., Hwang, Y.-S. and Kim, Y.-H. (2004) *Agrobacterium*-mediated genetic transformation of *Perilla frutescens*. *Plant Cell Rep.*, **23**, 386–390.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, *1303.3997*.

Li, H. (2018) Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.

Li, H., Handsaker, B., Wysoker, A., et al. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Logsdon, G.A., Vollger, M.R. and Eichler, E.E. (2020) Long-read human genome sequencing and its applications. *Nat. Rev. Genet.*, **21**, 597–614.

Longvah, T., Deosthale, Y.G. and Uday Kumar, P. (2000) Nutritional and short term toxicological evaluation of Perilla seed oil. *Food Chem.*, **70**, 13–16.

Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A. and Zdobnov, E.M. (2021) BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.*, **38**, 4647–4654.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L. and Zimin, A. (2018) MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.*, **14**, e1005944.

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. and Gurevich, A. (2018) Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*, **34**, i142–i150.

Mistry, J., Chuguransky, S., Williams, L., et al. (2021) Pfam: The protein families database in 2021. *Nucleic Acids Res.,* **49**, D412–D419.

Nishida, K. and Kondo, A. (2021) CRISPR-derived genome editing technologies for metabolic engineering. *Metab. Eng.*, **63**, 141–147.

Nitta, M., Lee, J.K., Kang, C.W., Katsuta, M., Yasumoto, S., Liu, D., Nagamine, T. and Ohnishi, O. (2005) The distribution of *Perilla* species. *Genet. Resour. Crop Evol.*, **52**, 797–804.

Nonaka, S., Arai, C., Takayama, M., Matsukura, C. and Ezura, H. (2017) Efficient increase of ɣ-aminobutyric acid (GABA) content in tomato fruits by targeted mutagenesis. *Sci. Rep.*, **7**, 7057.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T. and Salzberg, S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.

Petersen, M. and Simmonds, M.S.J. (2003) Rosmarinic acid. *Phytochemistry*, **62**, 121–125.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Rhie, A., Walenz, B.P., Koren, S. and Phillippy, A.M. (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.*, **21**, 245.

Sa, K.J., Kim, J.A. and Lee, J.K. (2012) Comparison of seed characteristics between the cultivated and the weedy types of *Perilla* species. *Hortic. Environ. Biotechnol.*, **53**, 310–315.

Saito, K. and Yamazaki, M. (2002) Biochemistry and molecular biology of the late-stage of biosynthesis of anthocyanin: lessons from *Perilla frutescens* as a model plant. *New Phytol.*, **155**, 9–23.

Sato, K., Krist, S. and Buchbauer, G. (2006) Antimicrobial effect of *trans*-cinnamaldehyde, (–)-perillaldehyde, (–)-citronellal, citral, eugenol and carvacrol on airborne microbes using an airwasher. *Biol. Pharm. Bull.*, **29**, 2292–2294.

Sawai, S., Ohyama, K., Yasumoto, S., et al. (2014) Sterol side chain reductase 2 is a key enzyme in the biosynthesis of cholesterol, the common precursor of toxic steroidal glycoalkaloids in potato. *Plant Cell*, **26**, 3763–3774.

Sharma, P., Masouleh, A.K., Topp, B., Furtado, A. and Henry, R.J. (2022) *De novo* chromosome level assembly of a plant genome from long read sequence data. *Plant J.*, **109**, 727–736.

Shen, W., Le, S., Li, Y. and Hu, F. (2016) SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One*, **11**, e0163962.

Sturme, M.H.J., Berg, J.P. van der, Bouwman, L.M.S., De Schrijver, A., Maagd, R.A. de, Kleter, G.A. and Battaglia-de Wilde, E. (2022) Occurrence and nature of off-target modifications by CRISPR-Cas genome editing in plants. *ACS Agric. Sci. Technol.*, **2**, 192–201.

Tian, J., Wang, Y., Lu, Z., Sun, C., Zhang, M., Zhu, A. and Peng, X. (2016) Perillaldehyde, a promising antifungal agent used in food preservation, triggers apoptosis through a metacaspase-dependent pathway in *Aspergillus flavus*. *J. Agric. Food Chem.*, **64**, 7404–7413.

Trócsányi, E., György, Z. and Zámboriné-Németh, É. (2020) New insights into rosmarinic acid biosynthesis based on molecular studies. *Curr. Plant Biol.*, **23**, 100162.

Ueda, H., Yamazaki, C. and Yamazaki, M. (2002) Luteolin as an anti-inflammatory and anti-allergic constituent of *Perilla frutescens*. *Biol. Pharm. Bull.*, **25**, 1197–1202.

Uemura, T., Yashiro, T., Oda, R., Shioya, N., Nakajima, T., Hachisu, M., Kobayashi, S., Nishiyama, C. and Arimura, G. (2018) Intestinal anti-inflammatory activity of perillaldehyde. *J. Agric. Food Chem.*, **66**, 3443–3448.

Yoshida, H., Nishikawa, T., Hikosaka, S. and Goto, E. (2021) Effects of nocturnal UV-B irradiation on growth, flowering, and phytochemical concentration in leaves of greenhouse-grown red perilla. *Plants*, **10**, 1252.

Yuba, A., Yazaki, K., Tabata, M., Honda, G. and Croteau, R. (1996) cDNA cloning, characterization, and functional expression of 4S-(–)-Limonene Synthase from *Perilla frutescens*. *Arch. Biochem. Biophys.*, **332**, 280–287.

Zhang, Y., Shen, Q., Leng, L., Zhang, D., Chen, Sha, Shi, Y., Ning, Z. and Chen, Shilin (2021) Incipient diploidization of the medicinal plant *Perilla* within 10,000 years. *Nat. Commun.*, **12**, 5508.

446 **Tables**

447 **Table 1.** Statistics of the genome assembly of red perilla cultivar Hoko-3 (Pfru_yukari_1.0) in comparison with
448 the previous assembly of the PF40 genome (ICMM_Pfru_2.0)

	Pfru_yukari_1.0	ICMM_Pfru_2.0
Total sequence length	1,258,994,547	1,234,370,464
No. of pseudochromosomes	20	20
No. of scaffolds	71	1,465
Scaffold N50	63,334,402	62,644,896
Scaffold L50	9	10
No. of contigs	94	2,228
Contig N50	41,459,669	2,738,655
Contig L50	11	137
Complete BUSCO[a] (%)	99.5	99.3
Complete and single-copy (%)	4.1	6.3
Complete and duplicated (%)	95.4	93.0

449 [a]embryophyta_odb10 (eukaryota, 2020-09-10) dataset (1,614 total BUSCO groups)

450

451 **Table 2.** Summary of repetitive elements in Pfru_yukari_1.0

	Length occupied (bp)	% of whole genome
Retroelements	474,789,418	37.71
LINEs	8,042,571	0.64
LTR elements	466,746,847	37.07
Copia	176,347,151	14.01
Gypsy	187,848,187	14.92
Others	102,551,509	8.15
DNA transposons	50,983,076	4.04
RC/Helitron	7,435,521	0.59
Unclassified	313,615,583	24.91
Total interspersed repeats	846,823,598	67.26
Small RNA	2,844,362	0.23
Low complexity	2,121,838	0.17
Simple repeats	14,903,442	1.18
Total	866,693,240	68.84

452

453 **Table 3.** Summary of the annotated genes

	Iso-Seq	Iso-Seq RNA-Seq	Iso-Seq RNA-Seq BRAKER2
No. of genes	19,452	53,541	86,258
No. of transcripts	33,869	88,890	121,996
Complete BUSCO[a]	66.7%	97.7%	98.7%
Complete and single-copy	26.5%	5.9%	5.3%
Complete and duplicated	40.2%	91.8%	93.4%

454 [a]embryophyta_odb10 (eukaryota, 2020-09-10) dataset (1,614 total BUSCO groups)

455

456

457 **Table 4.** Protein-level annotation of Pfru_yukari_1.0

Annotation Category	Annotation Level	Gene count
Protein homolog from tophit	Arabidopsis	54,262
	Rice	55,694
	Tomato	59,044
	Human	38,255
	Mouse	36,099
	UniProtKB/Swiss-Prot	52,566
	At least one of the above	72,339
No protein homolog	Protein domain	644
Total genes with protein-level annotation		72,983
Hypothetical protein		3,842
Total		76,825

458

459

460 **Figure legends**

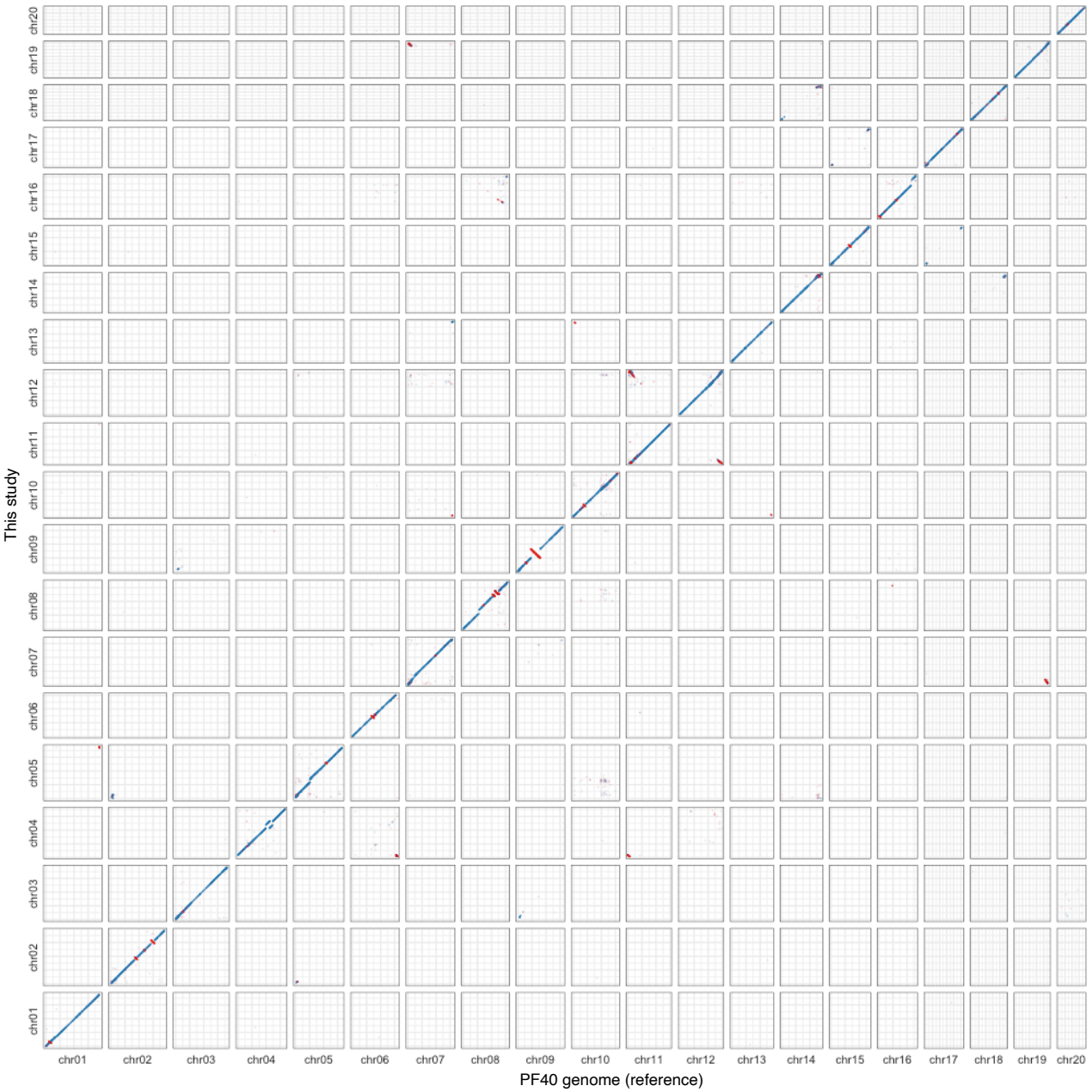
461 **Figure 1.** An image of the *Perilla frutescens* plants (cultivar Hoko-3) used for genome sequencing

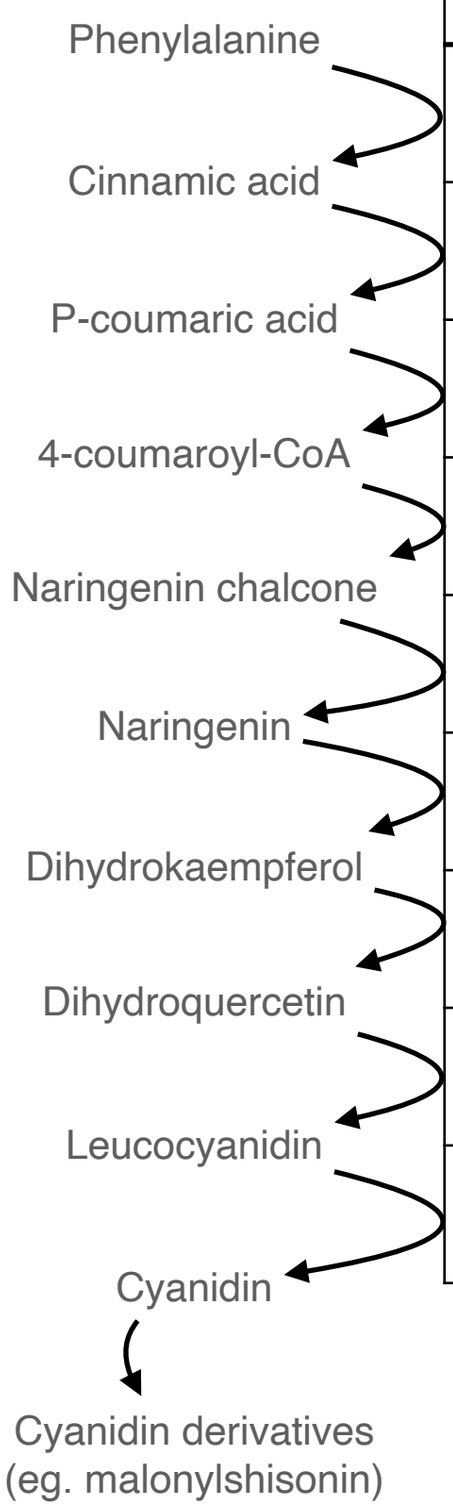
462 **Figure 2.** Dot-plot alignment between this genome assembly and reference PF40 genome assembly. Blue dots
463 represent +/+ strand alignments and red dots represent +/- strand alignments.

464 **Figure 3.** The number of enzyme coding genes in the representative anthocyanin biosynthetic pathway of *P.*
465 *frutescens*. PAL, phenylalanine ammonia-lyase; C4H, trans-cinnamate 4-monooxygenase; 4CL, 4-
466 coumarate:CoA ligase; CHS, chalcone synthase; CHI, chalcone-flavanone isomerase; F3H, flavanone 3-
467 hydroxylase; F3′H, flavonoid 3′-hydroxylase; DFR, dihydroflavonol 4-reductase; LDOX, leucoanthocyanidin
468 dioxygenase.

469

470


This study
chr20
chr19
chr18
chr17
chr16
chr15
chr14
chr13
chr12
chr11
chr10
chr09
chr08
chr07
chr06
chr05
chr04
chr03
chr02
chr01
chr01
chr02
chr03
chr04
chr05
chr06
chr07
chr08
chr09
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
PF40 genome (reference)



Enzyme	locus	isoform	RefUniProtKB
PAL	6	17	PAL1_ARATH
C4H	8	12	TCMO_ARATH
4CL	8	12	4CL1_ARATH
CHS	17	23	CHSY_ARATH
CHI	2	2	CFI1_ARATH
F3H	2	2	FL3H_ARATH
F3'H	4	8	F3PH_ARATH
DFR	1	4	DFRA_ARATH
LDOX	2	3	LDOX_ARATH