the plant journal

SEB Society for Experimental Biology

The Plant Journal (2017) **89,** 181–194 doi: 10.1111/tpj.13385

RESOURCE

Draft genome assembly and annotation of *Glycyrrhiza uralensis*, a medicinal legume

Keiichi Mochida[1,2,3,†], Tetsuya Sakurai[1,4,†], Hikaru Seki[1,5], Takuhiro Yoshida[1], Kotaro Takahagi[1,3], Satoru Sawai[1,‡], Hiroshi Uchiyama[6], Toshiya Muranaka[1,5] and Kazuki Saito[1,7,*]

[1]*RIKEN Center for Sustainable Resource Science, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan,*

[2]*Institute of Plant Science and Resources, Okayama University, Chuo 2-20-1, Kurashiki, Okayama 710-0046, Japan,*

[3]*Kihara Institute for Biological Research, Yokohama City University, 641-12 Maioka-cho, Totsuka-ku, Yokohama, Kanagawa 244-0813, Japan,*

[4]*Research and Education Faculty, Multidisciplinary Science Cluster, Interdisciplinary Science Unit, Kochi University, 200 Otsu, Monobe, Nankoku, Kochi 783-8502, Japan,*

[5]*Department of Biotechnology, Graduate School of Engineering, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan,*

[6]*Department of Applied Biological Science, College of Bioresource Sciences, Nihon University, 1866 Kameino, Fujisawa, Kanagawa, Japan, and*

[7]*Graduate School of Pharmaceutical Sciences, Chiba University, 1-8-1 Inohana, Chuo-ku, Chiba 260-8675, Japan*

Received 14 June 2016; revised 8 September 2016; accepted 26 September 2016; published online 24 October 2016.

*For correspondence (e-mail kazuki.saito@riken.jp).

†These authors contributed equally to this work.

‡Present address: Department of Biotechnology, Graduate School of Engineering, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan.

[Correction added on 30 July 2022, after first online publication: The copyright line was changed.]

SUMMARY

Chinese liquorice/licorice (*Glycyrrhiza uralensis*) is a leguminous plant species whose roots and rhizomes have been widely used as a herbal medicine and natural sweetener. Whole-genome sequencing is essential for gene discovery studies and molecular breeding in liquorice. Here, we report a draft assembly of the approximately 379-Mb whole-genome sequence of strain 308-19 of *G. uralensis*; this assembly contains 34 445 predicted protein-coding genes. Comparative analyses suggested well-conserved genomic components and collinearity of gene loci (synteny) between the genome of liquorice and those of other legumes such as *Medicago* and chickpea. We observed that three genes involved in isoflavonoid biosynthesis, namely, 2-hydroxyisoflavanone synthase (*CYP93C*), 2,7,4′-trihydroxyisoflavanone 4′-*O*-methyltransferase/isoflavone 4′-*O*-methyltransferase (*HI4OMT*) and isoflavone-7-*O*-methyltransferase (*7-IOMT*) formed a cluster on the scaffold of the liquorice genome and showed conserved microsynteny with *Medicago* and chickpea. Based on the liquorice genome annotation, we predicted genes in the P450 and UDP-dependent glycosyltransferase (UGT) superfamilies, some of which are involved in triterpenoid saponin biosynthesis, and characterised their gene expression with the reference genome sequence. The genome sequencing and its annotations provide an essential resource for liquorice improvement through molecular breeding and the discovery of useful genes for engineering bioactive components through synthetic biology approaches.

Keywords: *Glycyrrhiza uralensis*, liquorice/licorice, genome sequence, PRJDB3943, synteny, flavonoid, P450, UDP-dependent glycosyltransferases.

INTRODUCTION

The genus *Glycyrrhiza*, which includes liquorice/licorice, contains perennial herbal legumes that are frequently used in Chinese traditional medicine (Kao *et al.*, 2014). The roots and rhizomes of liquorice have long been used as a herbal medicine and natural sweetener (Asl and Hosseinzadeh, 2008). Experimental and clinical studies have demonstrated that liquorice has a wide range of pharmacological properties such as anti-inflammatory, antiviral,

© 2016 The Authors.
The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

antimicrobial, antioxidative, antidiabetic, antiasthma and anticancer activities as well as immunomodulatory, gastroprotective, hepatoprotective, neuroprotective and cardioprotective effects (Hosseinzadeh and Nassiri-Asl, 2015). Flavonoids and triterpene saponins are major bioreactive ingredients in liquorice. Nearly 300 flavonoids and more than 20 triterpenoids have been isolated from liquorice species (Wang *et al.*, 2015). The flavonoids of liquorice comprise various types including flavanones, chalcones and isoflavones, which possess pharmacological activities such as anticancer, antioxidative, antimicrobial, antiviral and hepatoprotective effects (Sumner *et al.*, 2015; Ji *et al.*, 2016). Glycyrrhizin is the major component of the triterpene saponins of liquorice, and it has been used as a natural sweeter. Pharmacologically, liquorice has been shown to have anti-inflammatory, anticancer, antiallergic and antiviral activities (Tang *et al.*, 2015). Despite the commercial importance of liquorice and the increasing demand for it, the lack of whole-genome information has inhibited the enhancement of bioactive traits and productivity through molecular breeding.

Each plant species has evolved specific metabolic pathways to produce specialised secondary metabolites, which has resulted in a rich chemical diversity of natural products (an estimate of as many as 1 060 000 metabolites has been given; Afendi *et al.*, 2012) in the plant kingdom (Yonekura-Sakakibara and Saito, 2009). The vast chemical diversity in plants is mainly because of the formation of chemical scaffolds and their subsequent modifications by the enzymes involved in these steps (Sumner *et al.*, 2015). The huge superfamily of cytochrome P450 monooxygenases (CYPs) found in angiosperms (about 1% of all protein-coding genes annotated in angiosperm genomes) catalyses a broad range of steps in metabolic pathways, in particular in the modification of chemical scaffolds of plant secondary metabolites. In our previous transcriptome analysis of Chinese liquorice (*Glycyrrhiza uralensis*), we identified 261 non-redundant contigs of transcripts putatively encoding P450s expressed in the roots (Ramilowski *et al.*, 2013). A substantial collection of genes involved in various aspects of plant secondary metabolism will provide a useful gene resource for designing non-natural pathways, as well as enzymes for the generation of novel compounds in synthetic biology (Renault *et al.*, 2014). Primary whole-genome information has enabled the discovery of genes associated with phytochemical properties in a diverse range of plants, and has been useful for decoding their functions by integration of multiple spectra in omics such as the transcriptome and metabolome (Fukushima *et al.*, 2009; Rai and Saito, 2016).

With the recent progress in whole-genome sequencing in different plant species, an increasing number of examples of clustered organisation of genes involved in particular metabolic pathways have been identified in plant genomes (Ober, 2005; Boycheva *et al.*, 2014; Nutzmann and Osbourn, 2014). These metabolic gene clusters are a notable feature of plant genomes that can be exploited to develop strategies for mining new genes, pathways and metabolites (Nutzmann *et al.*, 2016). In legumes, whole-genome sequencing of several species have been performed, for example in soybean (*Glycine max*) (Schmutz *et al.*, 2010), *Medicago truncatula* (Young *et al.*, 2011), chickpea (*Cicer arietinum*) (Jain *et al.*, 2013; Varshney *et al.*, 2013), *Lotus japonicus* (Sato *et al.*, 2008) and pigeonpea (*Cajanus cajan*) (Varshney *et al.*, 2012).

Comparative analyses of the sequenced legume genomes have illustrated that genic regions often share conserved and relatively extensive synteny (Cannon *et al.*, 2009). Therefore, genome-scale comparative studies in legumes should be useful for pathway discovery through systematic mining of metabolic gene clusters.

Here, we report the first draft genome assembly and annotations of *G. uralensis.* To obtain the whole-genome structure of a liquorice species we sequenced the genome of *G. uralensis*, the most widely used liquorice species in the pharmaceutical industry, through a combination of long-read and short-read sequencing. To characterise the functional and structural properties of the *G. uralensis* genome, we performed a comparative analysis with the published sequences of legume species such as *Medicago* and chickpea. Our genome-scale exploration of the genes involved in isoflavonoid biosynthesis illustrated a clustered organisation for three genes in that pathway in the liquorice genome as well as in the syntenic genomic region in *Medicago* and chickpea. Additionally, we performed a genome-wide prediction of P450- and UGT-encoding genes, and produced a list of the predicted genes with their gene expression patterns. The draft genome of *G. uralensis* offers an indispensable information resource to facilitate gene discovery and genome-based breeding in liquorice, as well as to understand the chemical diversity that has evolved in legumes.

RESULTS

Genome sequencing, assembly and scaffolding

The 308-19 ($2n = 16$) strain of *G. uralensis* was used for genome sequencing (400.95 $\pm$ 4 Mb) (see Figure S1a,b in the Supporting Information); this strain is held at the Takeda Garden for Medicinal Plant Conservation, Kyoto, Japan. We sequenced and assembled the genome by a hybrid assembly strategy using short reads from Illumina and long reads from Pacific Biosciences sequencing (Figure S2). In total, 380.53 Gb of raw Illumina reads were generated by sequencing libraries with three different insert sizes of 190 bp, 3 and 10 kb. After filtering short and low-quality reads, we obtained 327.53 Gb of high-quality sequences. We estimated the genome size of the strain

© 2016 The Authors.

based on flow cytometry and a *k*-mer analysis by using the Illumina paired-end reads. The flow cytometric analysis gave an estimated genome size of 400.95 Mb (±4 Mb) (Figure S1b), while the *k*-mer analysis gave an estimated size of 396.98 Mb based on the homozygous *k*-mer peak at about 635-fold coverage (Figure S3). Therefore, compared with the estimated genome size, the short-read sequence dataset covered a depth of approximately 817× (Table S1).

First, we used the ALLPATHS-LG assembler (Gnerre *et al.*, 2011) to assemble de novo the 212.03-Gb filtered Illumina short-read sequences, and generated a genomic assembly (ALLPATH-LG scaffolds) showing a total scaffold length of 381.79 Mb and a scaffold *N50* of 137.28 kb. We then filtered low-quality scaffolds using the REAPR program (Hunt *et al.*, 2013) in comparison with all of the Illumina short-read sequences. Additionally, we generated 6.76 Gb sequences, representing 16.86× genome coverage, by the Pacific Biosciences sequencing platform (PacBio RS II) (Table S1). After correcting the sequencing error of the PacBio reads using the Sprai program with the Illumina reads, we yielded 4.52 Gb of error-corrected long reads, covering approximately 11.27× depth of the estimated size of the liquorice genome. By using the error-corrected long reads, we improved the gap-filled assembly using the PBJelly2 tool (English *et al.*, 2012) and generated 12 528 scaffolds with an *N50* of 109 kb. As a result of this hybrid assembly approach, our final scaffold represented a *G. uralensis* genome with a total length of 379 Mb (>1 kb) (Table 1). Compared with the estimated genome size, the final scaffold of 379 Mb represented 94.5% of the *G. uralensis* genome.

Table 1 Assembly and annotation of the *Glycyrrhiza uralensis* genome

	Scaffold ≥1 kb	(≥5 kb)
Assembly features		
Number of scaffolds	12 528	(6921)
Total length	378 856 821 bp	(366 562 450 bp)
Longest scaffold	1 116 388 bp	(1 116 388 bp)
N50 (scaffolds)	109 269 bp	(114 660 bp)
GC content	35.34%	(35.32%)
Number of contigs	72 148	
Longest contig	74 557	
N50 (contigs)	7324	
Gene models		
Number of gene loci	34 445	
Number of gene models	38 135	
Mean transcript length	1359.88 bp	
Mean coding sequence length	1096.51 bp	
Mean number of exons per gene	5.31	
Mean exon length	253.49 bp	
Mean intron length	557.25 bp	
CEGMA		
Complete (%) to CEGs by the CEGMA pipeline	90.32%	
Partial (%) to CEGs by the CEGMA pipeline	97.58%	

Evaluation of the genome scaffolds of *G. uralensis*

To confirm the coverage of the deciphered sequence in the final scaffold dataset, we mapped all of the Illumina reads to the scaffolds. More than 95% of the Illumina reads were mapped to the final scaffold (Table S2). We also mapped expressed sequence tags (ESTs) and RNA sequencing (RNA-seq) reads available in *G. uralensis* to evaluate coverage of the genic region in the scaffolds. More than 95% of the ESTs and RNA-seq reads were mapped to the scaffolds (Table S3). We also searched for the presence of 248 conserved core eukaryotic genes (CEGs) using the CEGMA program (Parra *et al.*, 2007), and found sufficient hits for 224 (90.32%) full-length CEG proteins and 242 partial (97.58%) CEG proteins (Table 1). The results of the quality assessments suggest that our final scaffolds cover most of the *G. uralensis* genome and gene space in the deciphered sequence, and provide a draft genome sequence of this liquorice species.

Genome characterisation and functional gene annotations

The *G. uralensis* draft genome possesses a GC content of 35.34%, which is slightly higher than that of other sequenced legume genomes (Figure S4). We surveyed the draft genome for repetitive DNA using RepeatModeler and RepeatMasker (http://www.repeatmasker.org/). A total of 161 Mb (36.48% of the draft genome) was identified as transposable elements and unclassified repeats, excluding low-complexity sequences (Table S4). We predicted the gene structure on the draft genome using the MAKER-P pipeline (Campbell *et al.*, 2014) with the available RNA-seq reads and the ESTs available in liquorice (Table S5) and the deduced proteome datasets of genome-sequenced legume species. The MAKER-P analysis resulted in a set of protein-coding sequences (CDSs) putatively comprising 34 445 gene loci and 38 135 gene models (Table 1). The *k*-mer distribution of the sequence reads (Figure S3) was typically bimodal, which suggests heterogeneity of the *G. uralensis* genome. We assessed its genomic heterogeneity, and identified 1 342 713 filtered heterogenic regions, indicating one variant per 279 bases biased on intergenic regions (Figure S5, Table S6).

In the functional gene annotation, 82.3% of the protein-coding genes of *G. uralensis* showed significant homology to entries of known protein databases and deduced proteomes in the sequenced legume genome. Using a conserved protein domain search, 67.5% of the protein-coding genes showed significant hits. With these search results, 50.6% and 30.0% of those were assigned to at least one term in the Gene Ontology and KEGG (Kyoto Encyclopedia of Genes and Genomes) Ontology databases, respectively

© 2016 The Authors.
The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd.,

(Tables S7 and S8, Data S1). Furthermore, 2224 genes were classified to genes putatively encoding transcription factors (TFs), corresponding to 60 TF families in plants with specific DNA-binding domains (Mochida *et al.*, 2010) (Table S9).

Comparative analysis of the liquorice genome with other sequenced genomes

For genome-scale comparison between *G. uralensis* and genome-sequenced legumes, we attempted to virtually order the *G. uralensis* scaffolds based on the discovered microsynteny with genomes of *M. truncatula* (Young *et al.*, 2011) and *C. arietinum* (Varshney *et al.*, 2013). Synteny prediction using MCscan showed 1811 and 1602 blocks, which possess at least two reciprocal best-hit gene pairs, collinearly ordered in *G. uralensis*–*M. truncatula* and *G. uralensis*–*C. arietinum*, respectively. In total, the scaffolds covering 84.1% of whole sequence and containing 98.8% of the annotated genes were virtually ordered based on the microsynteny- and sequence similarity-based mapping in the *G. uralensis*–*M. truncatula* comparison (Tables S1 and, S11, Figure 1a,b). The virtually ordered scaffolds represented chromosome-level genomic properties of the *G. uralensis* genome (Figure 1a).

To provide the plastome sequence resource of *G. uralensis* and verify its phylogenetic relationship in Fabaceae, we attempted to reconstruct a pseudomolecule of the chloroplast genome in *G. uralensis* (Figure S6). Based on 58

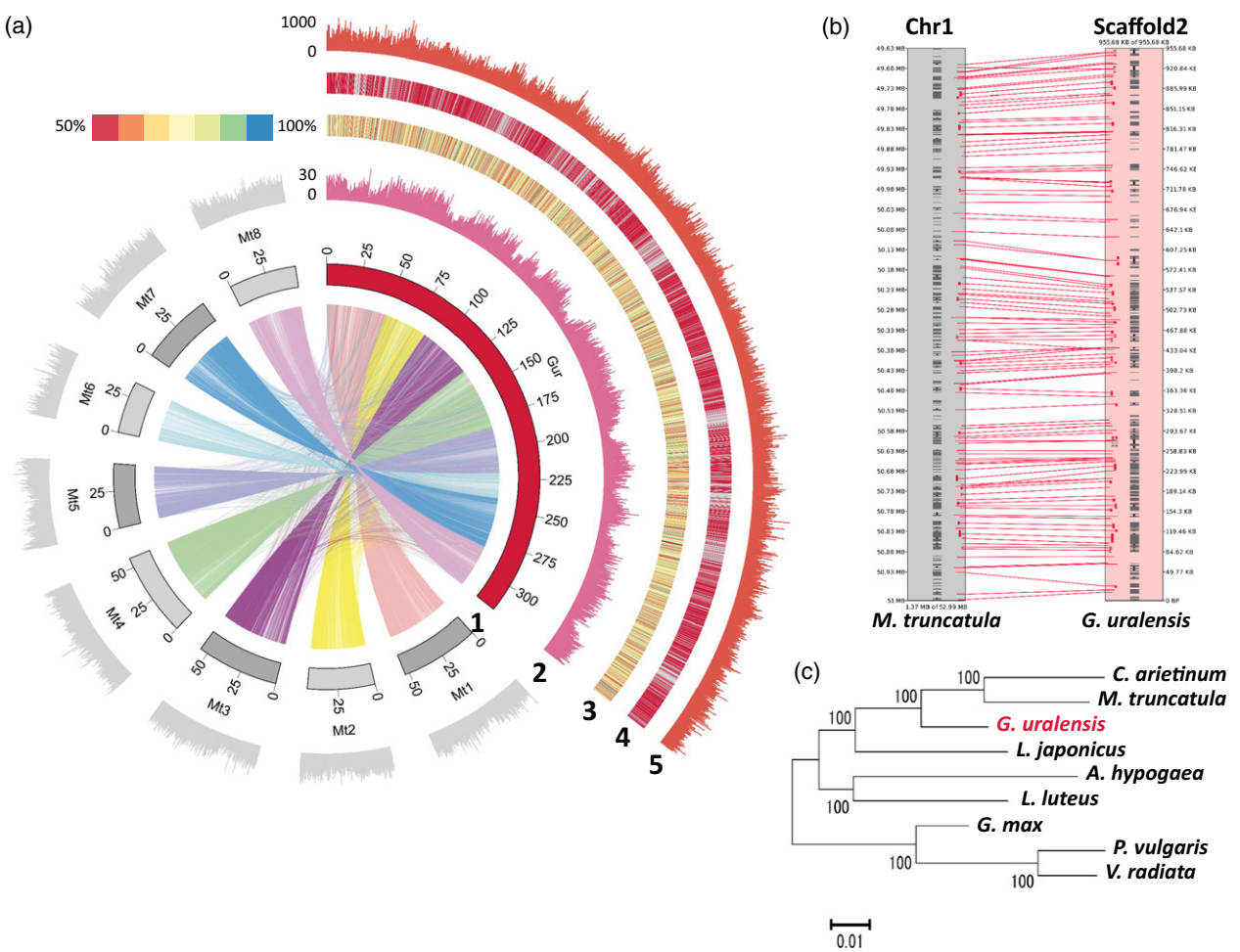


Figure 1. Whole-genome structure of *Glycyrrhiza uralensis*.
(a) A circular representation of the virtual order of the *G. uralensis* scaffold based on the conserved synteny with the *Medicago* genome. Corresponding homologues between *G. uralensis* and *Medicago* based on a reciprocal best hit of proteins are represented with coloured lines corresponding to each *Medicago* chromosome. (1) Chromosomes of *Medicago* (grey) and virtually ordered scaffolds of *G. uralensis* (red). (2) Gene density per 100 kb. (3) Heatmap representing sequence identity of putative homologues between *G. uralensis* and *Medicago*. (4) Virtual ordered scaffold of *G. uralensis* based on the conserved synteny (red) and homology search (grey) with the *Medicago* genome. (5) Heterogeneity of the *G. uralensis* genome figured by a polymorphism density (alternative nucleotide = 50%, sequence depth ≥50, per 100 kb).
(b) A close-up example of a conserved synteny between *G. uralensis* scaffold 2 and *Medicago* chromosome 1 drawn using SyMap.
(c) A phylogenetic tree of nine legume species including *G. uralensis* based on proteins of 58 genes annotated on the plastome sequence of each species.

© 2016 The Authors.

genes in the chloroplast genome, together with their counterparts commonly annotated in eight other legume species, the phylogenetic relationships in legume plants were represented (Figure 1c). A phylogenetic tree showed *G. uralensis* neighbouring *M. truncatula* and *C. arietinum*, which is consistent with the high rate of synteny block identification and sequence similarity search results in the genome-scale comparison of *G. uralensis* with these legumes.

We examined orthologue distributions of *G. uralensis* with other legumes and major dicotyledonous plants. The *G. uralensis* genes were grouped to 15 401 and 16 590 orthologous genes clusters with dicot species and legumes, respectively (Figure 2a,b). Among these clusters, 12 370 and 9889 clusters included commonly shared genes in all five legume species (Figure 2a) and in five dicot species, respectively (Figure 2b).

Genes involved in liquorice flavonoid biosynthesis

Flavonoids are a valuable metabolite group produced in liquorice plants. Therefore, we annotated genes involved in flavonoid biosynthesis in the *G. uralensis* genome and compared them with those in other plant genomes annotated in KEGG. Distribution patterns of genes encoding enzymes in secondary metabolite pathways represented the enrichment of genes involved in flavonoid biosynthesis, and particularly isoflavonoid biosynthesis pathways in legumes, including *G. uralensis* (Figure 2c, Data S2). We annotated genes encoding enzymes involved in flavonoid biosynthesis, some of those, such as chalcone synthase, chalcone isomerase and *trans*-cinnamate 4-monooxygenase, represented an increase of the copy number through tandem gene duplication in the *G. uralensis* genome (Table S12). We annotated isoflavonoid biosynthesis

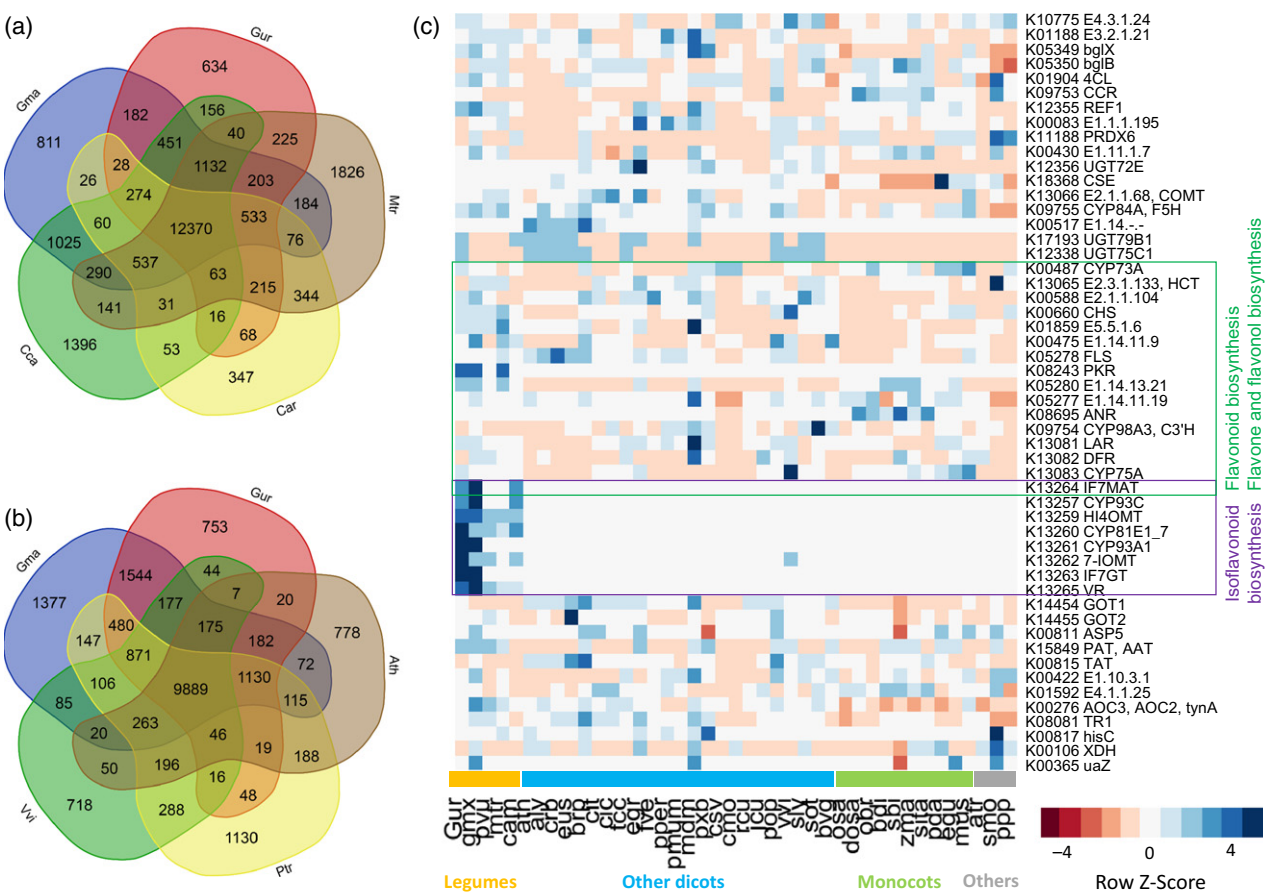


Figure 2. Gene families in *Glycyrrhiza uralensis*.
(a) Distribution of orthologous gene families in *G. uralensis* (Gur), and four sequenced legume species, soybean (Gma), *Medicago* (Mtr), chickpea (Car) and pigeonpea(Cca).
(b) Distribution of orthologous gene families in *G. uralensis* and four sequenced representative dicot plants, Arabidopsis (Ath), soybean (Gma), poplar (Ptr) and grape (Vvi).
(c) Distribution of *G. uralensis* genes involved in secondary metabolite biosynthesis classified into KEGG Ontology terms compared with those in 40 plant species. Each of alphabet codes corresponds to the species code used in KEGG. The raw data with full descriptions of KEGG Ontology terms and full species name is provided in Data S2.

© 2016 The Authors.

related genes corresponding to eight functional groups in KEGG. Three of five genes putatively encoding CYP81E1_7, isoflavone 2′-hydroxylase (EC:1.14.13.89) were located on the same scaffold as neighbouring genes (Table S13). We also observed genes putatively encoding CYP93C 2-hydroxyisoflavanone synthase (EC:1.14.13.136), 2,7,4′-trihydroxyisoflavanone 4′-*O*-methyltransferase/isoflavone 4′-*O*-methyltransferase (EC:2.1.1.212 2.1.1.46) and isoflavone-7-*O*-methyltransferase (EC:2.1.1.150), which were located on the same scaffold (Scaffold00721), suggesting that the organisation of gene clusters is a functional component for isoflavonoid biosynthesis (Table S13). The scaffold shared a microsynteny block with part of chromosome 4 in *M. truncatula* and chromosome 6 in *C. arietinum*, suggesting conserved genomic organisation including the functional gene cluster (Figure 3). These three enzymes, the genes for which are clustered in the genomes of *G. uralensis*, *C. arietinum* and *M. truncatula*, catalyse the closely located reactions of the entry steps of the isoflavonoid biosynthetic pathway (Figure S7), providing a further insight into the evolution and genetic organisation of the genes for isoflavonoid biosynthesis in the Fabaceae.

Genes involved in glycyrrhizin biosynthesis in liquorice

Glycyrrhizin is a triterpenoid-saponin produced in liquorice roots. It is synthesised via the cytosolic mevalonic acid pathway for the production of 2,3-oxidosqualene, which is subsequently cyclised to β-amyrin by β-amyrin synthase (bAS). Then, β-amyrin undergoes a two-step oxidation at the C-11 position and a three-step oxidation at the C-30 position followed by glycosylation reactions at the C-3 hydroxyl group to synthesise glycyrrhizin, as shown in Figure 4(a) (Seki *et al.*, 2008, 2011). The superfamilies of P450s and UDP-dependent glycosyltransferases (UGTs) are key enzymes contributing to the structural diversity of triterpenoid saponins such as glycyrrhizin (Sawai and Saito, 2011; Seki *et al.*, 2015). Therefore, we annotated genes encoding the conserved protein domains for P450 and UGT in the *G. uralensis* genome, and identified 257 and 91 genes (422 and 274 genes in the *M. truncatula* genome) encoding P450 and UGT domains, respectively (Data S3 and S4).

To verify expression of the genes encoding P450s and UGTs, we mapped the RNA-seq reads that we previously sequenced in a *de novo* transcriptome analysis of *G. uralensis*, and estimated the abundances of transcripts for each gene with the structural gene annotation. We examined the expression patterns of genes encoding lupeol synthase (LUS), cycloartenol synthase (CAS), bAS, β-amyrin 11-oxidase (CYP88D6) and 11-oxo-β-amyrin 30-oxidase (CYP72A154) based on their fragments per kilobase of exon per million mapped sequence reads (FPKM) values in the four different *G. uralensis* samples. As we

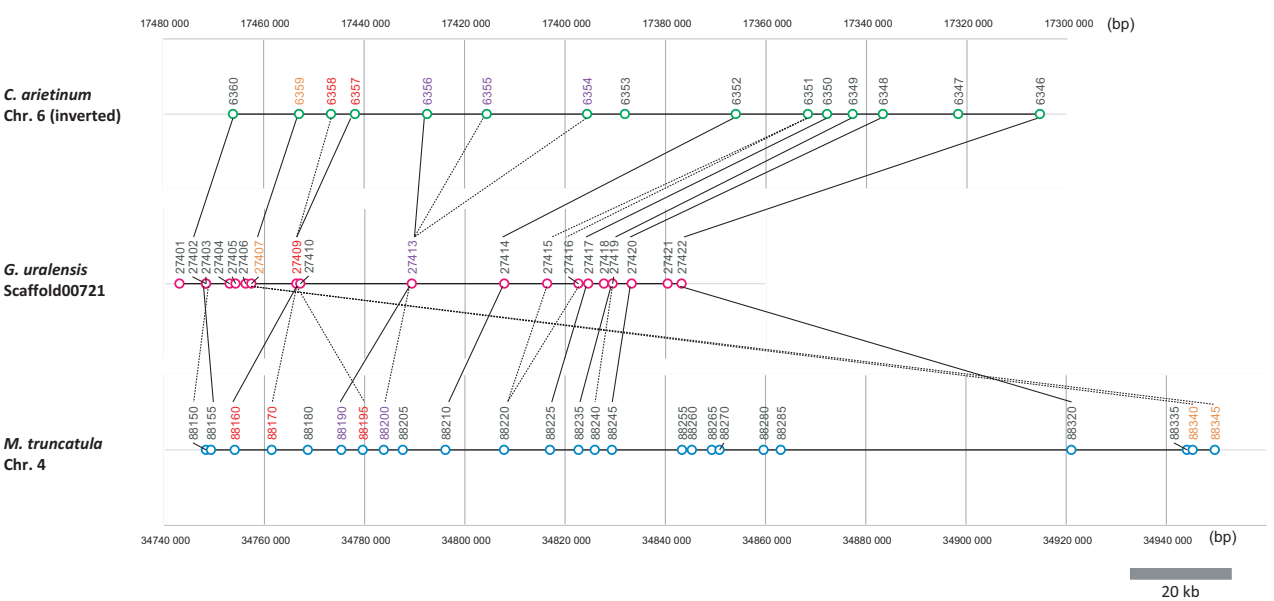


Figure 3. A synteny block.
A synteny block among *Cicer arietinum* chromosome 6 (inverted), *Glycyrrhiza uralensis* scaffold00721 and *Medicago truncatula* chromosome 4, which includes genes involved in isoflavonoid biosynthesis (*7-IOMT*, *CYP93C* and *HI4OMT*). Each circle indicates an annotated gene present in three genomes. The identifiers for annotated genes include a prefix (Ca_0, Scaffold00721s000, or Medtr4 g0) and the numerical suffix given above each circle. Orange, red and purple numbers correspond to genes encoding 7-IOMT, CYP93C and HI4OMT, respectively. These genes were annotated by KAAS in KEGG for *G. uralensis*, with homologues identified in *C. arietinum* and *M. truncatula*. The solid lines indicate pairs of genes identified as reciprocal best hits. The dashed lines indicate homologous genes identified in a single directional blast search (BLASTp identity ≥60%).

© 2016 The Authors.

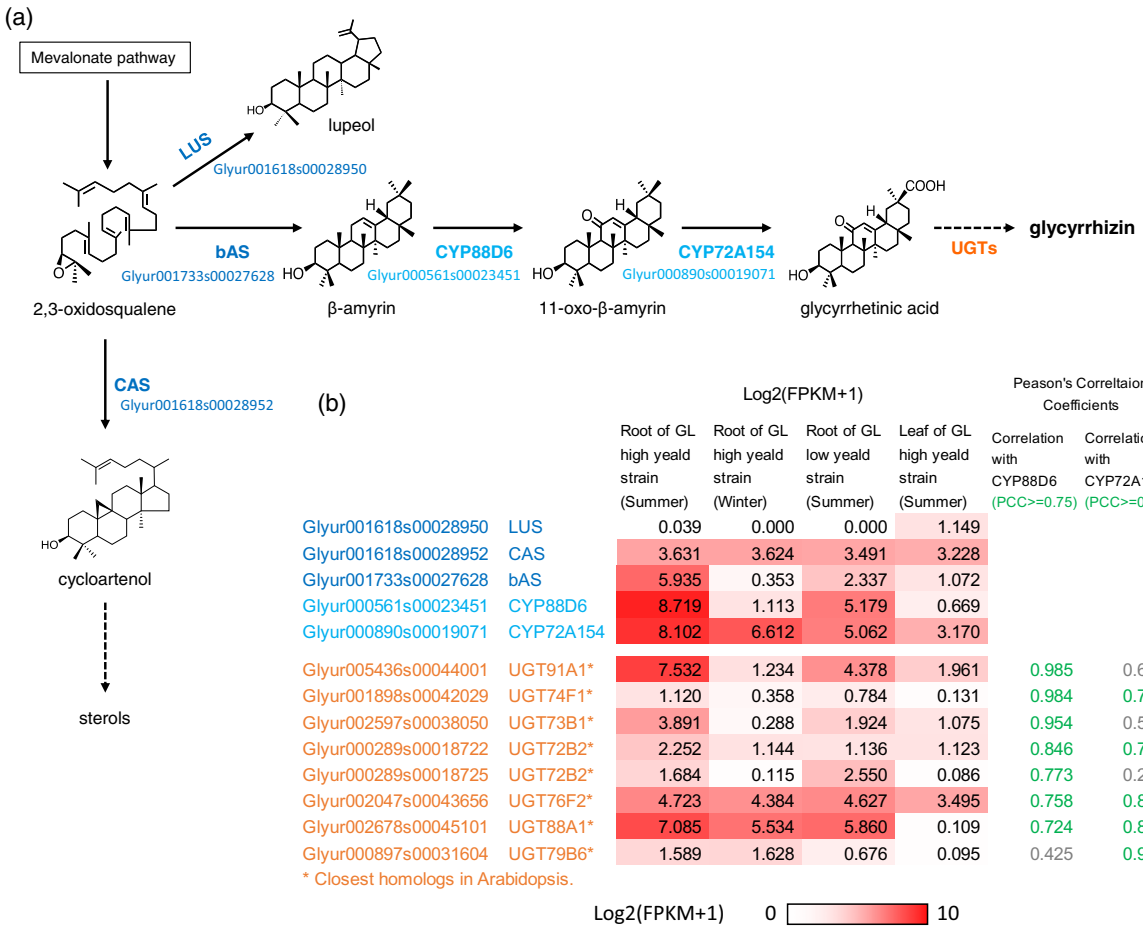


Figure 4. Glycyrrhizin biosynthetic pathway involving P450s and UDP-dependent glycosyltransferases (UGTs) in *Glycyrrhiza uralensis*.
(a) Proposed glycyrrhizin biosynthetic pathway in *G. uralensis*: LUS, lupeol synthase; bAS, β-amyrin synthase; CAS, cycloartenol synthase; CYP88D6, β-amyrin 11-oxidase; CYP72A154, 11-oxo-β-amyrin 30-oxidase; UGTs, UDP-glucuronosyltransferases. The dotted arrows signify undefined multiple steps.
(b) Expression patterns of genes encoding LUS, bAS, CAS, CYP88D6 and CYP72A154, and eight UGT genes highly correlated with the expression pattern of the *CYP88D6* and/or *CYP72A154* gene monitored in four different RNA sequencing samples of *G. uralensis*. The heatmap shows $\log_2$ (FPKM + 1) values (FPKM, fragments per kilobase pair of transcript per million fragments mapped) with a colour gradient from white to red. The Pearson's correlation coefficients show correlation between the expression pattern of each UGT gene and those of *CYP88D6* or *CYP72A154*; those that are >0.75 are highlighted in green.
(c), (d) Genomic distribution of the genes encoding: (c) P450s (blue lines) and (d) UGTs (orange lines) in the genomes of *G. uralensis* (red) and *M. truncatula* (grey) as indicated in Figure 1(a).

© 2016 The Authors.
The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd.,

expected, expression of *bAS*, *CYP88D6* and *CYP72A154* was consistent with the glycyrrhizin productivity of the samples (Ramilowski *et al.*, 2013). Based on the correlation of expression patterns between genes encoding UGTs and that of CYP88D6 or CYP72A154, we listed UGT genes that showed high values of Pearson's correlation coefficient (≥0.75) (Figure 4b).

We further classified the P450 and the UGT genes of *G. uralensis* into putative families based on the closest homologues of the subfamily genes annotated in Arabidopsis (http://www.p450.kvl.dk/), and compared those with *M. truncatula* (Figure 4c,d). Although the total numbers of P450 genes in *G. uralensis* and *M. truncatula* were not significantly different, there were differences in the number of genes in particular families, such as CYP71 (*P*-value = 6.39×10^{-6}) and CYP96 (*P*-value = 0.008288) (Figure S8). The total numbers of UGT genes in *G. uralensis* and *M. truncatula* were significantly different (91 and 274 in *G. uralensis* and *M truncatula*, respectively; *P*-value of Fisher's exact test = 7.871×10^{-10}). Some families showed significant differences between the two species, such as UGT73 (7 and 65; *P*-value = 1.019×10^{-8}) and UGT85 (7 and 53; *P*-value = 1.937×10^{-6}).

We found tandemly or closely located gene clusters on 39 scaffolds containing 95 P450 genes and 14 scaffolds containing 30 UGT genes (Data S3 and S4). This clustering suggests the expansion of P450s and UGTs through tandem gene duplication, which has possibly contributed to archiving the structural diversity of triterpenoids in *G. uralensis*. However, no evidence of the gene cluster for glycyrrhizin biosynthesis was obtained, despite the fact that *LUS* and *CAS* are tandemly located on the Glyur001618 scaffold (Figure 4a).

We then explored co-localisation of P450 and UGT genes in the *G. uralensis* genome. We found eight scaffolds that possess genes encoding UGT(s) and P450 localised on the same scaffold of *G. uralensis* (Table 2). Five of these clusters were also observed in the corresponding putative homologous genes annotated in the *M. truncatula* genome. One of them, the Glyur002597s00038051 and Glyur002597s00038050 pair (closely homologous to *CYP72A9* and *UGT73B1* in Arabidopsis, respectively), showed a well-correlated gene expression pattern in the RNA-seq data (PCC = 0.978) (Table 3). The putative counterparts of P450 and UGT in *M. truncatula*, Medtr4 g031820 (*CYP72A61v2*) and Medtr4 g031800 (*UGT73K1*), are also tightly co-localised on *Medicago* chromosome 4 (Naoumkina *et al.*, 2010). It was reported that these genes are mainly expressed in the *M. truncatula* root, which is consistent with the accumulation of soyasapogenol B conjugates in *M. truncatula*, suggesting that these enzymes might be involved in the same triterpenoid biosynthetic pathway. The genomic co-localisation of Glyur002597 s00038050 and Glyur002597s00038051 as well as their correlated and preferential expression in root tissues in *G. uralensis* suggests that these P450 and UGT genes might be involved in triterpene saponin biosynthesis in *G. uralensis*, and might share genomic structure as well as cellular networks in soyasaponin biosynthesis of both *M. truncatula* and *G. uralensis*.

DISCUSSION

Only a small fraction of the immensely diverse metabolism of plants has been explored for the production of new medicines and other products important for the well-being of humans (De Luca *et al.*, 2012). The isolation of key

Table 2 P450-UGT co-localization in the *Glycyrrhiza uralensis* scaffolds

Scaffold	*G. uralensis* gene ID	P450/UGT[a]	Best hit in *Medicago truncatula*
Glyur000026	Glyur000026s00001872	*CYP77B1*	Medtr4g131830
	Glyur000026s00001957	*UGT91A1*	Medtr2g008226
Glyur000042	Glyur000042s00005099	*CYP81D8*	Medtr5g016440
	Glyur000042s00005102	*UGT84B2*	Medtr5g016480
Glyur000079	Glyur000079s00008901	*UGT72E1*	Medtr8g074550
	Glyur000079s00008902	*UGT72E1*	Medtr5g019580
	Glyur000079s00008938	*CYP90B1*	Medtr5g020020
Glyur000243	Glyur000243s00013133	*CYP97A3*	Medtr7g079440
	Glyur000243s00013134	*UGT80A2*	Medtr7g079430
Glyur000514	Glyur000514s00017574	*UGT80B1*	Medtr1g098640
	Glyur000514s00017578	*CYP83B1*	Medtr1g023720
Glyur000817	Glyur000817s00031370	*UGT71C3*	Medtr5g070090
	Glyur000817s00031375	*CYP86A1*	Medtr5g070010
Glyur001090	Glyur001090s00029967	*UGT72B1*	Medtr8g006260
	Glyur001090s00029969	*UGT72B1*	Medtr8g006270
	Glyur001090s00029972	*CYP51G1*	Medtr8g006450
Glyur002597	Glyur002597s00038050	*UGT73B1*	Medtr4g031800
	Glyur002597s00038051	*CYP72A9*	Medtr4g031820

[a]Closest homologues in Arabidopsis.

© 2016 The Authors.

Table 3 Gene expression patterns of the putative counterparts of *CYP72A61v2* and *UGT73K1* in *Glycyrrhiza uralensis*

		Log_2 (FPKM + 1)					
G. uralensis gene ID	P450/UGT[a]	Root of GL high-yield strain (summer)	Root of GL high-yield strain (winter)	Root of GL low-yield strain (summer)	Leaf of GL high-yield strain (summer)	*Medicago truncatula* gene ID	Gene name
Glyur002597s00038051	*CYP72A9*	9.327	0.937	6.113	2.304	Medtr4g031820	*CYP72A61v2*
Glyur002597s00038050	*UGT73B1*	3.891	0.288	1.924	1.075	Medtr4g031800	*UGT73K1*

Expression was measured as fragments per kilobase pair of transcript per million fragments mapped (FPKM). The colour gradient from white to red was represented based on log_2 (FPKM + 1) values.
GL, glycyrrhizin.
[a]Closest homologues in Arabidopsis.

enzymes for undefined hydroxylation and glycosylation steps in the modification of secondary metabolites, such as triterpenoid saponins, in legumes will provide invaluable resources for engineering new artificial metabolic processes using synthetic genomic methods and for designing new drugs using combined biochemical approaches (Arendt *et al.*, 2016).

We deciphered 379 Mb (about 95%) of the genome and constructed 38 135 high-confidence gene models in *G. uralensis*. Deep genome sequencing enabled us to determine a high-confidence polymorphic site in the heterogenic *G. uralensis* genome. The establishment of a strategy for the sustainable use of liquorice is an urgent issue because of the recent increase in commercial use of of liquorice (Hayashi and Sudo, 2009). The reference sequence of *G. uralensis* will be an essential tool to enhance molecular breeding of the species. The draft genome of *G. uralensis* obtained in the present study will provide the first reference genome resource, which should be valuable for the identification, isolation and editing of useful genes to improve agronomic and medicinal traits in liquorice. Population sequencing methods with the draft genome will enable us to identify key genetic factors in association with important traits such as the chemical properties of bioactive compounds, rhizome and root productivity, environmental stress tolerances and disease resistance. Since three species of *Glycyrrhiza* (*G. uralensis*, *Glycyrrhiza glabra* and *Glycyrrhiza inflata*) are generally used for medicine and as a source of glycyrrhizin, it would be beneficial to determine the genome sequences of the two other species. In particular, a comparison of the genome sequences of non-glycyrrhizin-producing *Glycyrrhiza* species, such as *Glycyrrhiza macedonica* and *Glycyrrhiza echinata*, would enhance the molecular and genomic understanding of the production and accumulation of glycyrrhizin, which is an economically important plant product. Glycyrrhizin is a valuable medicine and sweetener, and this saponin possibly functions as a defence compound against the attack of herbivores, insects and microorganisms as demonstrated in some cereal species (Osbourn, 2003).

Our data strongly suggest shared synteny of genes in liquorice and related legume species. Comparative genomic analyses with sequenced legumes such as soybean (Schmutz *et al.*, 2010), chickpea (Varshney *et al.*, 2013), pigeonpea (Varshney *et al.*, 2012), *L. japonicus* (Sato *et al.*, 2008) and *M. truncatula* (Young *et al.*, 2011) will provide valuable data to identify candidates of useful genes and elucidate their biological functions. This will also elucidate the evolutionary history, diversification, and adaptation to global environments in the family Fabaceae (Chae *et al.*, 2014).

Recently, several examples of gene clusters of plant secondary metabolites have been reported (Nutzmann and Osbourn, 2014). In the *G. uralensis* genome, the genes coding for three enzymes involved in isoflavonoid biosynthesis have been located in the same scaffold. The first gene encodes the enzyme CYP93C (2-hydroxyisoflavanone synthase) (EC:1.14.13.136) and catalyses the first step of isoflavonoid biosynthesis, which is the migration of the B ring from position 2–3, and converts liquiritigenin, a flavanone, to 2,7,4′-trihydroxyisoflavanone, an isoflavanone. The second gene in the cluster encodes the enzyme HI4OMT (2,7,4′-trihydroxyisoflavanone 4′-*O*-methyltransferase/isoflavone 4′-*O*-methyltransferase) (EC:2.1.1.212 2.1.1.46), which typically converts daidzein to formononetin or genistein to biochanin A. The third gene encodes 7-IOMT (isoflavone-7-*O*-methyltransferase) (EC:2.1.1.150), which catalyses the methylation of daidzein to isoformononetin or genistein to prunetin. Although the close localisation of two genes encoding *O*-methyltransferases could be because of local gene duplication, the clustering of the P450 gene encoding the entry enzyme CYP93C of the isoflavonoid pathway together with two *O*-methyltransferases is intriguing considering the evolution of the isoflavonoid pathway. The same cluster containing these three enzyme genes was also found in the region spanning 220 kb in chromosome 4 of the *M. truncatula* genome and 200 kb in

© 2016 The Authors.

chromosome 6 of the *C. arietinum* genome. The results of the comparative genome analysis among the three Fabaceae species suggested that the genomic regions, including genes, involved in the entry steps of the isoflavonoid biosynthetic pathway might share a conserved microsynteny. Through the use of whole-genome sequence information and extensive compilation and comparison of microsyntenic regions, including metabolic gene clusters, insights will be obtained into conserved and diverse chemical properties in legumes. There have been no reports of gene clusters for the general flavonoid biosynthetic pathway in plants, possibly because the general flavonoid pathway is relatively old in the evolutionary history of flowering plants. However, since the isoflavonoid pathway is presumed to have emerged relatively recently, along with the evolution of the Fabaceae, the clustering of the pathway genes is still preserved or exhibits fitness advantages obtained during evolution.

There is evidence of that gene clusters for biosynthesis of triterpenes have accumulated in several species (Nutzmann and Osbourn, 2014). In Fabaceae species, a triterpene-producing gene cluster containing *OSC*, *AMY2* and two P450s (*CYP88D5* and *CYP71D353*) was reported in *L. japonicus* (Krokida *et al.*, 2013). AMY2 biosynthesises the triterpene dihydrolupeol, which can be converted to 20-hydroxylupeol and 20-hydroxybetulinic acid by CYP71D353. However, thus far, the role of CYP88D5 is unknown (Krokida *et al.*, 2013). In *M. truncatula*, two *CYP72A61*s and *UGT73K1*, likely to be associated with the production of triterpene saponins, are tightly linked on chromosome 4 (Naoumkina *et al.*, 2010). Soyasapogenol B is a β-amyrin derivative with hydroxyl groups at positions C-22 and C-24. CYP72A61 has been shown to catalyse C-22 hydroxylation of 24-hydroxy-β-amyrin to produce soyasapogenol B (Fukushima *et al.*, 2013). UGT73K1 is known to glycosylate multiple sapogenins including soyasapogenol B (Achnine *et al.*, 2005). In this study, we observed tightly localised gene pairs for P450 (Glyur002597s00038051) and UGT (Glyur002597s00038050) genes on the same scaffold of the *G. uralensis* genome, each of which might be a putative orthologous gene encoding *CYP72A61*s and *UGT73K1* in *M. truncatula,* respectively. This suggests the conservation of genomic and metabolic collinearity in soyasaponin biosynthesis between the two species. In the *Medicago* genome, although an *OSC* encoding β-amyrin synthase is also located on the same chromosome, it is not tightly linked to *CYP72A61s* and *UGT73K1* (Naoumkina *et al.*, 2010). Thus, the gene clusters for triterpene biosynthesis in the Fabaceae are not necessarily representative as gene clusters in oat (*Avena* spp.), which contains five genes encoding an OSC, a P450, a UGT and two other enzymes for the biosynthesis of the triterpene saponin, avenacin (Nutzmann and Osbourn, 2014). With the rapid increase in newly available genome assemblies in various plant families including Fabaceae (Kang *et al.*, 2014; De Vega *et al.*, 2015; Bertioli *et al.*, 2016; Hane *et al.*, 2016), reconstruction of ancestral genomes in palaeogenomics (Salse *et al.*, 2009; Murat *et al.*, 2015) might provide clues to aid our understanding of the evolutionary history of each metabolic gene cluster and its association with the emergence and evolution of new metabolic pathways in legumes.

Furthermore, the combination of comparative genomics, gene co-expression analysis and combined biosynthesis with heterologous expression in yeast has successfully identified enzymes involved in triterpenoid biosynthesis, and generated natural and rare triterpenoids that do not accumulate *in planta*, which will enable us to construct a natural–unnatural triterpenoid library (Fukushima *et al.*, 2013; Arendt *et al.*, 2016). The *G. uralensis* genome sequence will also contribute to investigating natural enzyme variants and orthologues. With the use of enzyme resources to diversify biosynthetic processes, the combined biosynthesis approach to creating a library of rare natural or novel bioactive compounds will be a new tool for enhancing drug discovery.

EXPERIMENTAL PROCEDURES

Whole-genome sequencing and assembly

Chromosome observation. For DNA sample preparation we used leaf tissues harvested from a specific strain 308-19 ($2n = 2x = 16$) of *G. uralensis* that has been preserved in Takeda Garden for Medicinal Plant Conservation, Kyoto, Japan. Chromosome numbers of strain 308-19 were observed in root tip cells using the conventional squashing method (Figure S1a). The root tips were pre-treated with 2 mM 8-hydroxyquinoline for 3 h at 24°C before they were fixed in ethanol-acetic acid (3:1) at 4°C for 24 h. Then they were macerated with 1 N HCl at 60°C for 10 min and stained with lactopropionic orcein for 30 min at 24°C.

Genomic DNA sample preparation. Genomic DNA of *G. uralensis* strain 308-19 was extracted from frozen leaf samples using Nucleon PhytoPure (GE Healthcare Life Sciences, http://www.gelifesciences.com/) and NucleoBond AXG100 DNA isolation columns with NucleoBond Buffer Set IV (MACHEREY-NAGEL, http://www.mn-net.com/) for Illumina and PacBio sequencing, respectively. Genomic DNA was quantified using the Quant-iT dsDNA BR Assay Kit (Invitrogen, http://www.invitrogen.com/) and assessed by agarose gel electrophoresis.

Whole-genome shotgun sequencing. Paired-end and mate-pair sequencing were performed on an Illumina platform, Hiseq 2000 sequencer. The paired-end and mate-pair sequencing libraries were constructed according to the standard protocols for the TruSeq Nano DNA LT Sample Prep Kit (Illumina, http://www.illumina.com/) and the Nextera Mate Pair Sample Preparation Kit (Illumina), respectively. The sequencing libraries included short paired-end libraries with 190-bp insert lengths, and mate-pair libraries with 3-kb and 10-kb jumping distances. Clonal clusters of each library were generated using cBot with the TruSeq PE Cluster Kit v.3 – cBot – HS (Illumina) and sequenced using a Hiseq 2000

© 2016 The Authors.
The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd.,

sequencer with a TruSeq SBS Kit v.2 – HS (Illumina, USA) by the paired-end sequencing method for 100-bp lengths. Sequence data were analysed using Hiseq Control Software (HCS) v.2.0.12, Real Time Analysis (RTA) v.1.17.21.3, and Consensus Assessment of Sequence and Variation (CASAVA) v.1.8.2 (Illumina). A total of 292.28 Gb with paired-end and 88.25 Gb with mate-pair sequencing were generated for the Illumina-based short-read sequencing. Raw sequencing reads were trimmed and filtered according to quality using default parameters of the Fastx-Toolkit for fastq data (http://hannonlab.cshl.edu/fastx_toolkit/).

A sequencing library for a PacBio sequencer was constructed using the SMRTBell Template Prep Kit 1.0 (Pacific Biosciences, http://www.pacb.com/) according to the manufacturer's instruction. The SMRTBell Template sequencing primer with DNA polymerase was applied to the SMRT Cell for the sequencing reaction. The P4 DNA polymerase with C2 chemistry (P4-C2; DNA/Polymerase Binding Kit P4, DNA Sequencing Reagent 3.0) was used in the sequencing reaction using the PacBio RS II sequencer (Pacific Biosciences). The sequence data were analysed using SMRT Analysis v.2.2.0 (Pacific Biosciences).

Genome size estimation of strain 308-19 of G. uralensis. By a flow cytometry-based method, genome size was estimated according to the procedure of Pfosser *et al.* (1995). Small pieces of *G. uralensis* strain 308-19 leaflets and Arabidopsis (Col 0) leaves were chopped simultaneously with a razor blade in ice-cold 2-amino-2-(hydroxymethyl)-1,3-propanediol (TRIS)-$MgCl_2$ buffer (0.2 M TRIS-HCl, 4 mM $MgCl_2$, 0.5% Triton-X 100, pH 7.5) including propidium iodide (50 $\mu g\ ml^{-1}$) and ribonuclease (50 $\mu g\ ml^{-1}$). After incubation for 5 min, the suspension was filtered through a 35-μm nylon mesh. Relative DNA contents of isolated nuclei were analysed with a flow cytometer (FACSCalibur system, Becton Dickinson, http://www.bd.com/). The haploid genome size of *G. uralensis* was estimated in comparison with the 135-Mb *A. thaliana* genome. To estimate the genome-size based on the sequence reads, *k*-mer frequencies of the Illumina paired-end read data were calculated using the K-mer Analysis Toolkit (KAT) (v.1.0, https://documentation.tgac.ac.uk/display/KAT/KAT+Home).

Whole-genome sequence assembly. The cleaned Illumina reads of paired-end and mate-pair sequencing were assembled using the ALLPATHS-LG assembler (Gnerre *et al.*, 2011) with a default parameter setting for *de novo* assembly and scaffolding. The REAPR program (Hunt *et al.*, 2013) was used to evaluate and assemble the ALLPATH-LG scaffolds and to filter out low-quality scaffolds through a process of mapping all of the cleaned Illumina reads to the scaffolds with a default parameter setting. The cleaned PacBio long reads were subjected to an error correction process using the Sprai program (http://zombie.cb.k.u-tokyo.ac.jp/sprai/) with the cleaned Illumina reads. The final scaffolds were generated by the PBJelly2 tool (English *et al.*, 2012) using the error-corrected long reads with one iteration (Figure S3).

Quality assessment of the genome scaffold. To assess the genome coverage of the constructed scaffold, all paired-end Illumina reads were mapped to the final scaffolds using the bwa program (0.7.12-r1039) (Li and Durbin, 2009) with the MEM algorithm and a default parameter setting (Table S2). Publicly available transcriptome sequences, ESTs, full-length cDNAs and RNA-seq datasets were also mapped to assess coverage of genes on the scaffolds. The ESTs and cDNAs of liquorice (Sudo *et al.*, 2009) were retrieved from the NCBI GenBank and mapped to the scaffold using the GMAP program (Wu and Watanabe, 2005) with a default parameter setting. The RNA-seq datasets (Ramilowski *et al.*, 2013) were downloaded from NCBI SRA(DRP000996) and were mapped to the scaffold using the tophat2 program (Kim *et al.*, 2013) with a default parameter setting (Table S3). The mapping results from analysis of the RNA-seq data were also used to quantify gene expression with the gene structural annotation using Cufflinks v.0.9.0 (Trapnell *et al.*, 2012). The CEGMA pipeline (Parra *et al.*, 2007) was used to assess coverage of genes that are likely to be present/absent in the scaffold.

Characterisation of the *G. uralensis* genome

GC content. The GC content of the *G. uralensis* genome was calculated in each 100-kb window and compared with those in four sequenced legume genomes: *M. truncatula*, *C. cajan*, *G. max* and *C. arietinum*. bedtools (2-2.25.0) (Quinlan and Hall, 2010) was used for the window analysis.

Heterogeneity assessment. To assess heterogeneity of the *G. uralensis* genome, the paired-end reads of Illumina short reads were mapped to the scaffolds using bwa with the default parameter settings. Using samtools (v.0.1.19) (http://samtools.sourceforge.net/) and bcftools (v.0.1.19), heterogeneous regions were identified on the *G. uralensis* genome with thresholds of DP (number of reads) $\geq$ 50 and AF1 (the site allele frequency of the strongest non-reference allele) = 0.5. Relationships of genomic structure and the heterogeneous regions were examined using SnpEff (v.4.2) (Cingolani *et al.*, 2012).

Repeat sequence annotation. RepeatModeler (v.1.0.8) was used to identify *de novo* repetitive sequences in the *G. uralensis* genome based on a self-blast search. RepeatMasker (v.4.0.5) was used to search known repetitive sequences using a cross-match program with a Repbase-derived RepeatMasker library (v.20140131) and the *de novo* repetitive sequences were constructed using RepeatModeler.

Gene structural annotation. The MAKER-P pipeline (Campbell *et al.*, 2014) was used to predict gene structure in the *G. uralensis* scaffolds. To run the MAKER-P first pass, we generated SNAP hidden Markov models (HMMs) with the CEGMA (Parra *et al.*, 2007) result file from the *G. uralensis* scaffolds and obtained GeneMark HMMs from the GeneMark result (Lukashin and Borodovsky, 1998). The MAKER-P first pass was executed with these HMMs, the transcriptome sequence information available in *G. uralensis* and the protein sequences of chickpea, pigeonpea, soybean and *Medicago* from Uniprot. Subsequently, we converted SNAP (Lomsadze *et al.*, 2005) and Augustus (Stanke *et al.*, 2006) HMMs via the first-pass result. Then, we ran the MAKER-P second pass with the SNAP and Augustus HMMs, the transcriptome sequences and the protein sequences from Uniprot. Finally, we extracted the genes with an annotation edit distance (AED) score of less than 1.

Gene functional annotation. To predict gene functions, homology searches were conducted using BLASTp search with parameter settings of –e 1e–5 –F F against entries of known proteins in a database (NCBI nr) and/or deduced proteome sequences in the sequenced plant genomes. The KAAS web server (http://www.genome.jp/tools/kaas/) (Moriya *et al.*, 2007) was used to classify *G. uralensis* genes into metabolic pathways in KEGG. In mapping to the KEGG pathway using KAAS, "ath, cit, tcc, gmx, fve, csv, vvi, sly, osa" as search organisms were used with the

© 2016 The Authors.
The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd.,

single directional best hit method. InterPro search to identify conserved protein domains was used to classify *G. uralensis* genes into functional categories defined by the Gene Ontology.

Transcription factors. To identify putative TF-encoding genes in the *G. uralensis* genome, 51 HMMs of Pfam and 11 originally created using HMMbuild of the HMMER package (http://hmmer.janelia.org/) were applied, which corresponded to a total of 60 TF families (Mochida *et al.*, 2010). The deduced proteome data of annotated genes in *G. uralensis* were subjected to a profile search in the HMM dataset using Pfam-HMM with thresholds of the *E*-value set at $<1 \times 10^{-5}$.

Comparative analysis of the *G. uralensis* genome with legume species

Synteny prediction. We generated putative orthologous pairs between *G. uralensis* and *M. truncatula*, and *G. uralensis* and *C. arietinum* according to the best hit found by reciprocal BLASTp searches with parameter settings of –e 1e–5 –F F in deduced proteome datasets. For *G. uralensis* and *M. truncatula*, 17 304 reciprocal best hit (RBH) pairs were identified and for *G. uralensis* and *C. arietinum*, 14 614 RBH pairs were identified. Within the RBH pairs, synteny blocks were identified using a toolset for generating the synteny correspondences, MCScan (v.0.8) (Tang *et al.*, 2010) with parameter settings of –u 10 –s 2 –k 50 –e 10 and SyMap (v.4.2) (Soderlund *et al.*, 2011) to visualise details of microsynteny. To order the *G. uralensis* scaffolds, we used the 1492 synteny block alignments of *G. uralensis* scaffolds along with *M. truncatula* chromosomes. Additionally, we allocated 3011 scaffolds containing 6755 genes based on the RBH pairs with *M. truncatula* genes.

Orthologue identification. Orthologous gene clusters for the genes of *G. uralensis* with other legumes and major dicotyledonous plants were found using the OrthoMCL program (Li *et al.*, 2003) with default parameter settings following an all-against-all BLASTp search with threshold *E*-values less than 1×10^{-5}. The representative proteome datasets of *Arabidopsis thaliana* (TAIR10), *G. max*, *Vitis vinifera* and *Populus trichocarpa* (Phytozome v.11.0) were subjected to orthologue analysis among major dicotyledonous plants. The representative proteome datasets of *M. truncatula*, *C. arietinum*, *G. max* and *C. cajan* (Phytozome v.11.0) were subjected to analysis among legume species. The OrthoMCL program was used to identify orthologue clusters in legume species and eudicots with the default parameter settings.

Glycyrrhiza uralensis chloroplast genome. To construct a pseudomolecule of the *G. uralensis* chloroplast genome, the PacBio read dataset was aligned on the chloroplast genome sequence of *G. glabra* (KF201590) using the BLASR (Chaisson and Tesler, 2012) program (minPctIdentity 85, minReadLength 500). Subsequently, we assembled the aligned PacBio reads using the PBcR pipeline (Koren *et al.*, 2012) (length 500, partitions 200, genomeSize = 150000). The gene structural annotation of the assembled chloroplast genome was determined using DOGMA (Wyman *et al.*, 2004) and the OrganellarGenomeDRAW program (Lohse *et al.*, 2013). The protein sequences of genes annotated on the plastome sequence were used to represent the phylogenetic relationships, together with their counterparts commonly annotated in eight other legume species: *Arachis hypogaea* (NC_026676), *C. arietinum* (NC_011163), *Glycine max* (NC_007942), *L. japonicus* (NC_002694), *Lupinus luteus* (NC_023090), *M. truncatula* (NC_003119), *Phaseolus vulgaris* (EU196765) and *Vigna radiata* (NC_013843). The phylogenetic analysis was conducted in MEGA6 with Muscle for building the sequence alignment and the maximum likelihood method for inferring the phylogenetic tree (Tamura *et al.*, 2013).

Genes involved in flavonoid and isoflavonoid biosynthesis. Genes putatively involved in flavonoid and isoflavonoid biosynthesis were annotated based on the KAAS web service in KEGG. The synteny between scaffold 00721 of the *G. uralensis* genome and chromosome 4 of the *Medicago* genome was determined based on gene pairs identified by reciprocal and single-directional BLAST searches.

Genes encoding P450 and UGT. Genes putatively encoding P450 and UGT in the *G. uralensis* genome and in the *Medicago* genome were searched for using the HMMER program (*E*-value set at $<1 \times 10^{-5}$) against the HMM data, PF00067 for the P450 domain and PF00201 for the UGT domain, from the Pfam database.

ACCESSION CODES

The raw data of the *G. uralensis* genome sequence have been deposited at the DNA Data Bank of Japan (DDBJ) under the Bioproject ID PRJDB3943. The genome assembly and the annotation datasets are available for downloading at http://ngs-data-archive.psc.riken.jp/Gur-genome/download.pl.

ACKNOWLEDGEMENTS

We thank Takeda Pharmaceutical Co. Ltd, Takeda Garden for Medicinal Plant Conservation, Kyoto, for providing *G. uralensis,* strain 308-19. We also thank Yuji Sawada for his valuable comments on this work. This study was supported, in part, by the Special Fund from the Director of RIKEN Yokohama Institute, by Grants-in-Aid for Scientific Research of the Japan Society for the Promotion of Science (JSPS) and by Strategic Priority Research Promotion Program of Chiba University. The authors declare no competing financial interests.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Somatic chromosomes and genome size of *Glycyrrhiza uralensis*.

Figure S2. Assembly pipeline for constructing a *Glycyrrhiza uralensis* draft genome.

Figure S3. Distribution pattern of 27 mers in the whole-genome Illumina reads.

Figure S4. GC content of sequenced genomes for *Glycyrrhiza uralensis* and other legumes.

Figure S5. The distribution of the heterogenic regions classified to genomic structures in the *Glycyrrhiza uralensis* genome.

Figure S6. The chloroplast genome of *Glycyrrhiza uralensis*.

Figure S7. The KEGG pathway of isoflavonoid biosynthesis.

Figure S8. Distribution of genes putatively encoding P450 and UGT.

Table S1. Summary of the sequencing information for *Glycyrrhiza uralensis* genome analysis.

Table S2. Summary of Illumina reads mapping to the scaffolds.

© 2016 The Authors.

Table S3. Summary of transcripts mapping to the scaffolds.

Table S4. Summary of repetitive sequences in the *Glycyrrhiza uralensis* genome.

Table S5. Transcript evidence sources for gene prediction using the MAKER-P pipeline.

Table S6. Summary of variant analysis in the *Glycyrrhiza uralensis* genome.

Table S7. Functional annotation of the predicted genes from the *Glycyrrhiza uralensis* genome.

Table S8. Summary of the Gene Ontology annotation.

Table S9. Number of genes encoding transcription factors in the *Glycyrrhiza uralensis* genome.

Table S10. Summary of synteny analysis of the *Glycyrrhiza uralensis* (Gur) genome with the *Medicago truncatula* (Mtr) and *Cicer arietinum* (Car) genome.

Table S11. Summary of virtually ordered *Glycyrrhiza uralensis* scaffolds with the *Medicago truncatula* genome.

Table S12. Genes involved in flavonoid biosynthesis in *Glycyrrhiza uralensis.*

Table S13. Genes involved in isoflavonoid biosynthesis in *Glycyrrhiza uralensis.*

Data S1. Functional annotations of gene models in the *Glycyrrhiza uralensis* genome.

Data S2. Comparison of genes encoding enzymes involved in biosynthesis of other secondary metabolites annotated in KEGG.

Data S3. Genes encoding the P450 domain (Pfam: PF00067) in *Glycyrrhiza uralensis* and *Medicago truncatula.*

Data S4. Genes encoding the UGT domain (Pfam: PF00201) in *Glycyrrhiza uralensis* and *Medicago truncatula*.

REFERENCES

Achnine, L., Huhman, D.V., Farag, M.A., Sumner, L.W., Blount, J.W. and Dixon, R.A. (2005) Genomics-based selection and functional characterization of triterpene glycosyltransferases from the model legume Medicago truncatula. *Plant J.* **41**, 875–887.

Afendi, F.M., Okada, T., Yamazaki, M. *et al.* (2012) KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol.* **53**, e1.

Arendt, P., Pollier, J., Callewaert, N. and Goossens, A. (2016) Synthetic biology for production of natural and new-to-nature terpenoids in photosynthetic organisms. *Plant J.* **87**, 16–37.

Asl, M.N. and Hosseinzadeh, H. (2008) Review of pharmacological effects of Glycyrrhiza sp. and its bioactive compounds. *Phytother. Res.* **22**, 709–724.

Bertioli, D.J., Cannon, S.B., Froenicke, L. *et al.* (2016) The genome sequences of Arachis duranensis and Arachis ipaensis, the diploid ancestors of cultivated peanut. *Nat. Genet.* **48**, 438–446.

Boycheva, S., Daviet, L., Wolfender, J.L. and Fitzpatrick, T.B. (2014) The rise of operon-like gene clusters in plants. *Trends Plant Sci.* **19**, 447–459.

Campbell, M.S., Law, M., Holt, C. *et al.* (2014) MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164**, 513–524.

Cannon, S.B., May, G.D. and Jackson, S.A. (2009) Three sequenced legume genomes and many crop species: rich opportunities for translational genomics. *Plant Physiol.* **151**, 970–977.

Chae, L., Kim, T., Nilo-Poyanco, R. and Rhee, S.Y. (2014) Genomic signatures of specialized metabolism in plants. *Science*, **344**, 510–513.

Chaisson, M.J. and Tesler, G. (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, **13**, 238.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92.

De Luca, V., Salim, V., Atsumi, S.M. and Yu, F. (2012) Mining the biodiversity of plants: a revolution in the making. *Science*, **336**, 1658–1661.

De Vega, J.J., Ayling, S., Hegarty, M. *et al.* (2015) Red clover (Trifolium pratense L.) draft genome provides a platform for trait improvement. *Sci. Rep.* **5**, 17394.

English, A.C., Richards, S., Han, Y. *et al.* (2012) Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE*, **7**, e47768.

Fukushima, A., Kusano, M., Redestig, H., Arita, M. and Saito, K. (2009) Integrated omics approaches in plant systems biology. *Curr. Opin. Chem. Biol.* **13**, 532–538.

Fukushima, E.O., Seki, H., Sawai, S., Suzuki, M., Ohyama, K., Saito, K. and Muranaka, T. (2013) Combinatorial biosynthesis of legume natural and rare triterpenoids in engineered yeast. *Plant Cell Physiol.* **54**, 740–749.

Gnerre, S., Maccallum, I., Przybylski, D. *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl Acad. Sci. USA*, **108**, 1513–1518.

Hane, J.K., Ming, Y., Kamphuis, L.G. *et al.* (2016) A comprehensive draft genome sequence for lupin (Lupinus angustifolius), an emerging health food: insights into plant-microbe interactions and legume evolution. *Plant Biotechnol. J.*, doi: 10.1111/pbi.12615.

Hayashi, H. and Sudo, H. (2009) Economic importance of licorice. *Plant Biotechnol.* **26**, 101–104.

Hosseinzadeh, H. and Nassiri-Asl, M. (2015) Pharmacological effects of Glycyrrhiza spp. and its bioactive constituents: update and review. *Phytother. Res.* **29**, 1868–1886.

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M. and Otto, T.D. (2013) REAPR: a universal tool for genome assembly evaluation. *Genome Biol.* **14**, R47.

Jain, M., Misra, G., Patel, R.K. *et al.* (2013) A draft genome sequence of the pulse crop chickpea (Cicer arietinum L.). *Plant J.* **74**, 715–729.

Ji, S., Li, Z., Song, W. *et al.* (2016) Bioactive constituents of Glycyrrhiza uralensis (Licorice): discovery of the effective components of a traditional herbal medicine. *J. Nat. Prod.* **79**, 281–292.

Kang, Y.J., Kim, S.K., Kim, M.Y. *et al.* (2014) Genome sequence of mungbean and insights into evolution within Vigna species. *Nat. Commun.* **5**, 5443.

Kao, T.C., Wu, C.H. and Yen, G.C. (2014) Bioactivity and potential health benefits of licorice. *J. Agric. Food Chem.* **62**, 542–553.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36.

Koren, S., Schatz, M.C., Walenz, B.P. *et al.* (2012) Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693–700.

Krokida, A., Delis, C., Geisler, K., Garagounis, C., Tsikou, D., Pena-Rodriguez, L.M., Katsarou, D., Field, B., Osbourn, A.E. and Papadopoulou, K.K. (2013) A metabolic gene cluster in Lotus japonicus discloses novel enzyme functions and products in triterpene biosynthesis. *New Phytol.* **200**, 675–690.

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

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

Lohse, M., Drechsel, O., Kahlau, S. and Bock, R. (2013) OrganellarGenomeDRAW–a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res.* **41**, W575–W581.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O. and Borodovsky, M. (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506.

Lukashin, A.V. and Borodovsky, M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26**, 1107–1115.

Mochida, K., Yoshida, T., Sakurai, T., Yamaguchi-Shinozaki, K., Shinozaki, K. and Tran, L.S. (2010) LegumeTFDB: an integrative database of Glycine max, Lotus japonicus and Medicago truncatula transcription factors. *Bioinformatics*, **26**, 290–291.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C. and Kanehisa, M. (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182–W185.

© 2016 The Authors.
The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd.,

Murat, F., Louis, A., Maumus, F., Armero, A., Cooke, R., Quesneville, H., Roest Crollius, H. and Salse, J. (2015) Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome Biol.* **16**, 262.

Naoumkina, M.A., Modolo, L.V., Huhman, D.V., Urbanczyk-Wochniak, E., Tang, Y., Sumner, L.W. and Dixon, R.A. (2010) Genomic and coexpression analyses predict multiple genes involved in triterpene saponin biosynthesis in Medicago truncatula. *Plant Cell*, **22**, 850–866.

Nutzmann, H.W. and Osbourn, A. (2014) Gene clustering in plant specialized metabolism. *Curr. Opin. Biotechnol.* **26**, 91–99.

Nutzmann, H.W., Huang, A. and Osbourn, A. (2016) Plant metabolic clusters - from genetics to genomics. *New Phytol.* **211**, 771–789.

Ober, D. (2005) Seeing double: gene duplication and diversification in plant secondary metabolism. *Trends Plant Sci.* **10**, 444–449.

Osbourn, A.E. (2003) Saponins in cereals. *Phytochemistry*, **62**, 1–4.

Parra, G., Bradnam, K. and Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.

Pfosser, M., Amon, A., Lelley, T. and Heberlebors, E. (1995) Evaluation of sensitivity of flow-cytometry in detecting aneuploidy in wheat using disomic and ditelosomic wheat-rye addition lines. *Cytometry*, **21**, 387–393.

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

Rai, A. and Saito, K. (2016) Omics data input for metabolic modeling. *Curr. Opin. Biotechnol.* **37**, 127–134.

Ramilowski, J.A., Sawai, S., Seki, H., Mochida, K., Yoshida, T., Sakurai, T., Muranaka, T., Saito, K. and Daub, C.O. (2013) Glycyrrhiza uralensis transcriptome landscape and study of phytochemicals. *Plant Cell Physiol.* **54**, 697–710.

Renault, H., Bassard, J.E., Hamberger, B. and Werck-Reichhart, D. (2014) Cytochrome P450-mediated metabolic engineering: current progress and future challenges. *Curr. Opin. Plant Biol.* **19**, 27–34.

Salse, J., Abrouk, M., Bolot, S., Guilhot, N., Courcelle, E., Faraut, T., Waugh, R., Close, T.J., Messing, J. and Feuillet, C. (2009) Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. *Proc. Natl Acad. Sci. USA*, **106**, 14908–14913.

Sato, S., Nakamura, Y., Kaneko, T. *et al.* (2008) Genome structure of the legume, Lotus japonicus. *DNA Res.* **15**, 227–239.

Sawai, S. and Saito, K. (2011) Triterpenoid biosynthesis and engineering in plants. *Front. Plant Sci.* **2**, 25.

Schmutz, J., Cannon, S.B., Schlueter, J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

Seki, H., Ohyama, K., Sawai, S., Mizutani, M., Ohnishi, T., Sudo, H., Akashi, T., Aoki, T., Saito, K. and Muranaka, T. (2008) Licorice beta-amyrin 11-oxidase, a cytochrome P450 with a key role in the biosynthesis of the triterpene sweetener glycyrrhizin. *Proc. Natl Acad. Sci. USA*, **105**, 14204–14209.

Seki, H., Sawai, S., Ohyama, K. *et al.* (2011) Triterpene functional genomics in licorice for identification of CYP72A154 involved in the biosynthesis of glycyrrhizin. *Plant Cell*, **23**, 4112–4123.

Seki, H., Tamura, K. and Muranaka, T. (2015) P450s and UGTs: key players in the structural diversity of Triterpenoid Saponins. *Plant Cell Physiol.* **56**, 1463–1471.

Soderlund, C., Bomhoff, M. and Nelson, W.M. (2011) SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res.* **39**, e68.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. and Morgenstern, B. (2006) AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439.

Sudo, H., Seki, H., Sakurai, N. *et al.* (2009) Expressed sequence tags from rhizomes of Glycyrrhiza uralensis. *Plant Biotechnol.* **26**, 105–107.

Sumner, L.W., Lei, Z., Nikolau, B.J. and Saito, K. (2015) Modern plant metabolomics: advanced natural product gene discoveries, improved technologies, and future prospects. *Nat. Prod. Rep.* **32**, 212–229.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729.

Tang, H., Bowers, J.E., Wang, X. and Paterson, A.H. (2010) Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proc. Natl Acad. Sci. USA*, **107**, 472–477.

Tang, Z.H., Li, T., Tong, Y.G., Chen, X.J., Chen, X.P., Wang, Y.T. and Lu, J.J. (2015) A systematic review of the anticancer properties of compounds isolated from Licorice (Gancao). *Planta Med.* **81**, 1670–1687.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578.

Varshney, R.K., Chen, W., Li, Y. *et al.* (2012) Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **30**, 83–89.

Varshney, R.K., Song, C., Saxena, R.K. *et al.* (2013) Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. *Nat. Biotechnol.* **31**, 240–246.

Wang, L., Yang, R., Yuan, B., Liu, Y. and Liu, C. (2015) The antiviral and antimicrobial activities of licorice, a widely-used Chinese herb. *Acta Pharm. Sin. B*, **5**, 310–315.

Wu, T.D. and Watanabe, C.K. (2005) GMAP: A Genomic Mapping and Alignment Program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.

Wyman, S.K., Jansen, R.K. and Boore, J.L. (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, **20**, 3252–3255.

Yonekura-Sakakibara, K. and Saito, K. (2009) Functional genomics for plant natural product biosynthesis. *Nat. Prod. Rep.* **26**, 1466–1487.

Young, N.D., Debelle, F., Oldroyd, G.E. *et al.* (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, **480**, 520–524.

© 2016 The Authors.
The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd.,