Genome and systems biology of *Melilotus albus* provides insights into coumarins biosynthesis

Fan Wu1,†, Zhen Duan1,†, Pan Xu1,†, Qi Yan1, Minghui Meng1, Mingshu Cao2, Chris S. Jones3, Xifang Zong1, Pei Zhou1, Yi Meng Wang1, Kai Luo1, Shengsheng Wang1, Zhuanzhuan Yan1, Penglei Wang1, Hongyan Di1, Zifeng Ouyang1, Yanrong Wang1 and Jiuy Zhang1.*

1State Key Laboratory of Grassland Agro-ecosystems, Key Laboratory of Grassland Livestock Industry Innovation, Ministry of Agriculture and Rural Affairs, Engineering Research Center of Grassland Industry, Ministry of Education, College of Pastoral Agriculture Science and Technology, Lanzhou University, Lanzhou, China
2Grasslands Research Centre, AgResearch Limited, Palmerston North, New Zealand
3Feed and Forage Development, International Livestock Research Institute, Nairobi, Kenya

Summary

*Melilotus* species are used as green manure and rotation crops worldwide and contain abundant pharmacologically active coumarins. However, there is a paucity of information on its genome and coumarin production and function. Here, we reported a chromosome-scale assembly of *Melilotus albus* genome with 1.04 Gb in eight chromosomes, containing 71.42% repetitive elements. Long terminal repeat retrotransposon bursts coincided with declining of population sizes during the Quaternary glaciation. Resequencing of 94 accessions enabled insights into genetic diversity, population structure, and introgression. *Melilotus officinalis* had relatively larger genetic diversity than that of *M. albus*. The introgression existed between *M. officinalis* group and *M. albus* group, and gene flows was from *M. albus* to *M. officinalis*. Selection sweep analysis identified candidate genes associated with flower colour and coumarin biosynthesis. Combining genomics, BSA, transcriptomics, metabolomics, and biochemistry, we identified a β-glucosidase (BGLU) gene cluster contributing to coumarin biosynthesis. *MaBGLU1* function was verified by overexpression in *M. albus*, heterologous expression in *Escherichia coli*, and substrate feeding, revealing its role in scopoletin (coumarin derivative) production and showing that nonsynonymous variation drives BGLU enzyme activity divergence in *Melilotus*. Our work will accelerate the understanding of biologically active coumarins and their biosynthetic pathways, and contribute to genomics-enabled *Melilotus* breeding.

Keywords: sweet clover, genome evolution, population genetics, coumarin biosynthesis, BGLU.

Introduction

There is growing recognition that improving the environmental performance of agriculture and livestock systems and establishing sustainable levels of the consumption of crop and animal-sourced foods are essential for the sustainability of the global food system (Foley et al., 2011; Herrero et al., 2013; Tester and Langridge, 2010). Deploying nutrients more efficiently is considered a promising path to environmentally sustainable agriculture and global food security (Tilman, 2011). Legume plants play an important role in delivering this sustainability (Stagnari et al., 2017), which are as legumes perform well in conservation and intercropping systems, important in low productivity farming systems (Stagnari et al., 2017). For example, leguminous trees of the genera *Sesbania* and *Cajanus* have been interplanted or rotated with maize crops and used as fallows, resulting in a two- to fourfold increase in maize yields in Africa (Sanchez, 2002).

*Melilotus*, commonly referred to as sweet clover or wild alfalfa, comprises approximately 19 annual or biennial species, and *Melilotus albus* and *Melilotus officinalis* are the most common species, although the taxonomy of these two species remains controversial (Darbyshire and Small, 2018). *Melilotus* spp. are widely distributed in temperate Europe, the Mediterranean, subtropical Asia, and North Africa. They have adapted to extremely varied environments including drought, cold, and saline conditions and are widely used as fodder, green manure, and soil conservation crops (Stickler and Johnson, 1959; Zhang et al., 2018b). Consequently, *Melilotus* has the potential to contribute to the environmental sustainability of agriculture and to global food security. For example, sweet clover was considered the king of green manures and grazing legumes in the South and Midwest of the United States in the first half of the 20th century (Clark, 2007). It was also used as a cover crop in the plains region of the United States before nitrogen fertilizer became widely available (Clark, 2007). *Melilotus albus* has also been reported to have a superior yield and nitrogen-fixation ability to that of *Medicago sativa* (Mcewen and Johnston, 1985).

Coumarins are broadly distributed across the plant kingdom, occurring in over 150 species from 74 families (Stringlis et al., 2019; Venugopala et al., 2013). Coumarins contribute to the phytoalexin and allelopathic properties of plants and are involved in processes such as defense against phytopathogens, response to abiotic stresses, regulation of oxidative stress, and hormonal regulation (Bourgaud et al., 2006; Stringlis et al., 2019). In addition to *in planta* functions, coumarins or their derivatives are also known for their pharmacological properties, such as...
anticoagulant (Lei et al., 2015; Sun et al., 2020; Wadelius and Pirzamabadi, 2007), antiviral (Hassan et al., 2016), and anti-
cancer (Bello et al., 2005; Duru et al., 2015; Lata et al., 2015; 
Rehman et al., 2013) agents. Melilotus is a well-known medicinal 
plant used for coumarin extraction because of its relatively high 
coumarin content. There was 0.06–0.75% of dry matter in 15 
Melilotus species (Nair et al., 2010) compared to 0.02–0.13% in 
Santolina insularis (Sacchetti et al., 1997) and 0.07–0.52% in 
Mikania glomerata at the flowering stage. The high content of 
coumarins in Melilotus may, however, act as a deterrent to 
grazing animals. Thus, reducing coumarin content has been a 
target for improving grazing animals. Thus, reducing coumarin content has been a 

genetic exchange between Medicago and Medicago officinalis, whole 

Genome, assembly, and annotation

We sequenced the genome of M. albus (2n = 2x = 16) using the 

near-isogenic line Ma46 (Figure S1). The genome size was 
estimated to be 1.15 Gb based on k-mer analysis (Figure S2), 

which is similar to a previous report (1.00–1.25 Gb) (Hirsch et al., 

1998). An additional 125.91 Gb of PacBio long reads were 

assembled using Falcon (ver 0.3.0). This resulted in a 1.05 Gb 
genome consisting of 283 contigs with an N50 of 7.49 Mb and a 

GC content of 35.77% (Table S1). Furthermore, we used 

61.77 Gb of clean reads from a Hi-C sequencing library to assist 

in assembly correction. A total of 239 contigs covering 1.04 Gb 

(99.40%) of the assembled genome were anchored to 8 
pseudochromosomes (Figure S3, Table S2). To further confirm 

the completeness of the assembly, we performed BUSCO 

assessments and found that 1343 (96.67%) of the 1375 total 

conserved genes were present in the M. albus genome (Table S3). 

We mapped the mRNA sequences of M. albus to the assembled 
genome, and the alignment rate reached 98.28%. In addition, 

the high-quality short reads generated on the Illumina platform 

were mapped onto the assembled genome, resulting in a 

mapping rate of 99.14% (Table S4).

A total of 41 910 genes, with an average gene length of 

3147 bp, exon length of 261 bp (~4.4 exons per gene), and 

intronic length of 587 bp (~3.4 introns per gene), were predicted 
in the genome (Table S5, Figure S4). Approximately 98.84% 

(39 783) of the predicted protein-coding genes were annotated 

(Table S6). A higher gene density (39 816, 95%) was distri-
buted towards both arms of the pseudomolecules (Figure 1a). 
Noncoding RNA genes, including 925 tRNAs, 481 rRNAs, 3154 
rRNAs, and 299 miRNA precursors, were also annotated 
(Table S7).

Comparative genomic and evolutionary analysis

To track the evolutionary history of M. albus, we examined the 
genomic relationships between M. albus and 12 other sequenced 
plant species. A phylogenetic tree was constructed based on a 
catenated alignment of 600 single-copy gene families (Figure 
1c). The phylogenetic tree showed that M. albus (Ma) is most 
closely related to Medicago truncatula (Mt) and Trifolium pratense 
(Tp), which belong to the same Trifolieae tribe, and speciation between Ma and Mt occurred ~14.16 million years ago 
(Mya).

A total of 36 792 M. albus genes were clustered into 21 129 
genome families, among which 2039 were expanded and 1586 
contracted relative to the other 12 species (Figure 1c). The genes in the expanded families of M. albus were enriched in functional 
categories that include plant–pathogen interaction, phenyl-
propanoid biosynthesis, plant hormone signal transduction, and 
the upstream pathway of plant hormone signal transduction 
(Figure S5). Phenylpropanoid metabolism plays a pivotal role in 
the production of secondary metabolites such as coumarins, 
linin, and flavonoids. These secondary metabolites contribute to 
the stability and robustness of plants when challenged with 
mechanical or environmental damage (Vogt, 2010). In the 
phenylpropanoid biosynthesis pathway, BGLU gene subfamilies 
that are associated with coumarin accumulation were found to be 
expanded (Figure S5a). In addition, the hydroxy cinnamoyl 
transferase (HCT) and cytochrome p450 subfamilies were 
expanded in M. albus relative to M. truncatula and may be 
involved in the biosynthesis of caffeoyl-CoA and feruloyl-CoA 
(Figure S5a). Gene ontology (GO) enrichment analysis of the 
expanded genes showed that they were significantly enriched in 
the terms transferase activity (GO: 0016740), stress response 
(GO: 0006950), response to phenylalanine (GO: 0080053), and 
phenylalanine ammonia-lyase activity (GO: 0045548) (Figure S6a). 
The expanded family members responded to drought stress, and 
the expression patterns of different members were different 
(Figure S7).

We also investigated the physiological and transcriptomic 
effects of drought stress conditions on both the shoots and roots 
of M. albus. The contents of H2O2, malonaldehyde (MDA), 
proline, and soluble sugar in shoots showed continuous increases 
over 48 h under drought stress, and the soluble protein content 
decreased (Figure S6b). The coumarin (secondary metabolism) 
content was increased by drought treatment (Figure S6c). A 
weighted correlation network analysis (WGCNA) identified four 
modules associated with the response to drought (Figures S6d, 
S8, and S9). GO and KEGG enrichment analyses of the genes 
from the four modules showed that they were related to 
metabolism of secondary metabolites, amino acids, and hor-
mones and glycolysis (Figures S6e and S10). Furthermore, there 
were correlations among these terms (Figure S6e), suggesting 
that they coordinate on the drought resistance of sweet clover.
Repetitive sequence underlying dynamic evolution

The expansion of the *M. albus* genome has been mainly caused by the proliferation of TEs, because *M. albus* did not undergo independent whole-genome duplication (WGD) (Figure 1b).

Approximately 71.42% of the *Ma* genome was annotated as repeated sequences. For comparison, the TE contents of *Mt*, *Ca*, and *Glycine max* (*Gm*) are 43.61%, 54.03%, and 53.41%, respectively (Table S8). The abundance of long terminal repeat retrotransposon (LTR-RT) Ty3-gypsy and Ty1-copia elements
differed notably among the four genomes – Ma (34.14%, 20.85%), Mt (8.98%, 8.23%), Ca (11.92%, 15.74%), and Gm (24.82%, 16.03%) (Table S8). Phylogenetic reconstructions based on the reverse transcriptase domains of the LTR-RTs in these four species revealed that most of the Gypsy and Copia elements in Ma were clustered together, with other distinct small clades containing Ma, Mt, Ca, and Gm elements. Thus, Ma has its own specific LTR-RTs, and minor diversity is encountered in the other three species (Figure 2a).

The analysis of the insertion times of intact LTR sequences in Ma, Mt, Ca, and Gm revealed distinct evolutionary patterns (Figure 2b). Intact LTRs were most abundant in Ma, in which bursts of insertions were observed to have occurred more recently and lasted longer than in other species (Figure 2b). Interestingly, the four species all displayed recent bursts. The number of LTRs was significantly greater in Ma than in the other three species, and the proportion of DNA transposons was slightly lower (Figure 2c). Based on the Jukes-Cantor distance, Ma experienced a single recent burst, while Mt, Ca, and Gm experienced two bursts (Figure 2c). Further analysis showed that most existing intact LTRs in Ma were inserted after its divergence from Mt, in accordance with the finding that the burst time of the intact LTRs (0.05–0.85 Mya) was notably more recent than the estimated Ma divergence time (14.16 Mya) from Mt (Figures 1c and 2b). These results show that M. albus has undergone specific LTR insertion events, resulting in its relatively larger genome size.

Notably, the numbers and average lengths of both solo-LTRs and intact LTRs in Ma are greater than those in the other three species (Table S9). The length distribution of the intact LTR sequences among the four species shows the greatest length in Ma, with a peak at approximately 6000 bp, while the peaks in the other three species occur at approximately 2000 bp (Figure 2d). These results are consistent with the relatively young age of the intact LTRs in Ma, and they may therefore be highly dynamic. Ma and Mt share a high degree of collinearity (Figure S11), but the number of intact LTR-RTs is far greater in Ma (>99) (Table S9).

To evaluate the effect of LTR insertion in Ma, we investigated the genes in the regions around LTRs in Ma and their homologous genes in Mt. The following two types of genes were distinguished (Figure S12): type I genes exhibited overlap between the genes

Figure 2 Comparative analysis of repeat sequences in four legume species. (a) Phylogenetic analysis of Copia and Gypsy elements in Melilotus albus (red), Cicer arietinum (blue), Glycine max (green), Medicago truncatula (yellow), and Trifolium pratense (pink) using conserved protein domains. (b) Analysis of intact LTR numbers and insertion times in M. albus. (c) Distribution of sequence divergence among four types of TEs from M. albus, M. truncatula, C. arietinum, and G. max. The right y-axis represents the genome percentage of M. albus TEs. (d) Comparison of intact LTR length among M. albus, M. truncatula, C. arietinum, and G. max. (e) Structure of LTRs and the corresponding genes (MaTLP and MaSTKc). The grey box represents a 5-bp target site duplication, and red triangles represent dinucleotide palindromic motifs. (f) RNA abundance (FPKM) and relative expression levels of MaTLP and MaSTK in roots under drought stress. The patterns of FPKM values and relative expression levels under drought stress were similar. D-0 h, D-3 h, D-24 h: roots at 0, 3, and 24 h of drought treatment, respectively.
and LTRs, and type II genes were located inside LTRs. A total of 1441 type I genes were identified in Ma and 1251 in Mt. KEGG enrichment analysis showed that type I genes in Ma and Mt were enriched in particular functional categories (Figure S13). RNA transport, RNA degradation, and steroid biosynthesis pathways were enriched in Ma and Mt, while plant-pathogen interaction, phenylpropanoid biosynthesis, fatty acid degradation, and diterpenoid biosynthesis pathways were only enriched in Ma. Thus, there may be different mechanisms of adaptive evolution in Ma and Mt. A total of 724 type II genes were identified, among which the most common functional category was transcriptase. Only a small proportion of these genes were expressed (FPKM value >1), consistent with most of these genes being silenced by LTRs (Table S10). The silencing of these transcriptase genes could influence the stability of *Melilotus* genes. There were also 73 type II genes in the categories of stress tolerance, ATP synthase, and phosphodiesterase (Table S11). Interestingly, a gene encoding a thaumatin-like protein (TLP, involved in the response to environmental stress (Liu et al., 2010)) was not expressed in the leaves but was significantly upregulated (fold change > 20) in the roots under moderate drought stress (Figure 2e and f). The expression level of the serine/threonine kinase (STK) gene was also upregulated under moderate drought stress (Figure 2e and f). These findings provide evidence that the expanded LTRs in *Melilotus* may play an important role in the responses to environmental stress.

**Phylogeny of *Melilotus* species and evolutionary history of *M. albus* and *M. officinalis***

We examined the relationships and divergence among populations of *Melilotus* by resequencing a collection of 94 *Melilotus* accessions from various geographic regions (Table S12). Based on 4 999 288 SNPs and 606 387 InDels identified among the accessions, we constructed a neighbour-joining phylogenetic tree (*Melilotus archiducis* as an outgroup) (Figure 3a) and performed a population structure analysis (Figure 3b). Three major groups were formed, with one group mainly including *M. albus* accessions, one mainly including *M. officinalis* accessions, and another including the accessions of 16 species with a few discernible exceptions (Figure 3a). This grouping is further supported by a model-based analysis of population admixture with K = 3 (Figures 3b and S14), as well as a principal component analysis (PCA) (Figure 3c). Most of *M. officinalis* and *M. albus* formed respective compact clusters, indicating limited within-group variance compared to the accessions of 16 species (Figure S15). The taxonomic and genetic relationships between *M. albus* and *M. officinalis* remain controversial (Darbyshire and Small, 2018). Regardless of which methods were used, *M. albus* and *M. officinalis* accessions were not clearly distinguished. The chloroplast genes of 94 accessions were also used to construct the phylogenetic tree, and similar results were obtained (Figure S16).

To test for the genomic introgression that may exist between *M. officinalis* and *M. albus*, we conducted D-statistics-based phylogenetic analysis of the four groups. The *M. officinalis*, *M. albus*, other, and outgroup groups were set as P2, P3, P1, and O, respectively (Figure 3d). The tree with the ABBA pattern was counted 101 010 times, while that with the BABA pattern was counted 96 813 times. Furthermore, the D-statistic was greater than zero (2.1%), and the Z-score was 6.9 (Figure 3d). These results indicated that significant gene flow existed between P2 (*M. officinalis* group) and P3 (*M. albus* group). To further determine the direction of gene flow, we conducted a population splitting and migration analysis using Treemix (Pickrell and Pritchard, 2012), which indicated the occurrence of highly weighted gene flow from *M. albus* to *M. officinalis* and other *Melilotus* species (Figure 3e). In addition, *M. officinalis* and *M. albus* share some native locations, and both were introduced to similar places (Figure S17). However, we found that the mixed accessions and the adjacent accessions were not of the same geographic origin. Therefore, we speculated that gene introgression existed in *M. officinalis* and *M. albus*, which may explain why *M. albus* and *M. officinalis* were not clearly separated by molecular markers (Wu et al., 2016; Yan et al., 2017; Zhang et al., 2018b). Climate oscillations dramatically affect genetic diversity and effective population sizes (Ne). We inferred that *M. albus* and *M. officinalis* experienced bottlenecks mirrored by continual Ne declines in the last 0.2 Mya (Figure 3f), coinciding with the period of the Quaternary glaciation (0.01–2 Mya). The LTR burst time also coincided with the significant decline in *Melilotus* population sizes (Figure 2b).

The linkage disequilibrium (LD) decay reached half of the maximum average (r²) at 400 bp in *M. albus* and 300 bp in *M. officinalis* (Figure 3g), indicating fast decay, which is typical of outcrossing species, and *M. officinalis* showed higher genetic diversity than *M. albus*. The Tajima’s D values of *M. albus* and *M. officinalis* were estimated to examine their possible selection patterns, and the Tajima’s D values of *M. officinalis* (0.96) were larger than those of *M. albus* (0.16) (Table S13), indicating that *M. albus* experienced stronger positive selection. To examine the selective signatures during adaptive evolution in *Melilotus*, we identified candidate selective regions and genes in *M. albus* and *M. officinalis*. The analysis of selective sweeps based on x and Fst identified 105 selected regions containing 754 genes in *M. officinalis* and 109 selected regions containing 694 genes in *M. albus* (Figure S18, Supplementary data 1). The functional annotation of identified genes revealed that the selected genes were mainly involved in environmental adaptation and metabolism (Figure S19). Phenylalanine metabolism pathway genes, including one BGLU, four COMTs, four HCTs, and seven UGTs, were selected during the evolution of *M. albus* relative to *M. officinalis*.

Based on Fst for the white to yellow flower accessions (Figure 4a), we identified a 2 Mb region on Chr7 harbouring one cytochrome P450 gene, one UDP-glucose flavonoid 3-O-glucosyltransferase (UGFT) gene (Chen et al., 2011), and six Agamous-like (AGL) genes (Zhang et al., 2020) with tandem duplication (Figure 4b); 1 Mb region on Chr1 containing one ethylene-responsive transcription factor (ERF) gene, 1 Mb region on Chr8 containing one auxin transporter-like gene, one Myb-like gene, and three genes with F-box domain, which may relate to flower color (Figure 4a). To test the expression of these genes, we randomly picked one accession with white flower and one accession with yellow flower for q-PCR analysis. Compared to white flower accession, yellow flower accession had highly expression of ERF, AGL80_3, AGL80_4, and AGL80_5, and lowly expression of UGFT and P450 (Table S14). We also investigated the nucleotide diversity (π) between high-coumarin-content and low-coumarin-content accessions. Five genes (4CL1, BGLU46, COMT5, UGT8, and UGT90) related to the coumarin biosynthesis pathway were highly differentiated (Figure 4c).

**Metabolite profiling of NILs and evolution of coumarin biosynthesis**

A metabolomic analysis of *M. albus* near isogenic lines (NILs) (Ma46, low coumarin content; Ma49, high coumarin content)
was conducted to identify coumarins (Figure 5a). Two isomers of coumaric acid (C₉H₈O₃) were detected as peaks eluted at 3.95 min and 4.56 min (with \(m/z\) [M⁺/C₀H₁₀O₃] = 163.0399), presumably corresponding to p-coumaric acid and o-coumaric acid. Both types of coumaric acids showed high levels in the Ma49 lines. CP349.088_3.94 (denotes positive ion mode CP, m/z and retention time) was identified as a coumaroyl glucoside (C₁₅H₁₈O₈, [M⁺/Na⁺]), which was also detected in negative ion mode as CN325.0920_3.95 ([M⁻/H⁻]). Coumaroyl glucoside exhibited elevated levels in Ma49 (Figure 5b). CP147.044_5.21 was identified as a coumarin (C₉H₆O₂) showing high levels in the Ma49 lines (Figure 5b). These differences suggest that the accumulation of coumarins in Ma is regulated via glycosylation or deglycosylation. Therefore, BGLU/UGT plays an important role in coumarin biosynthesis and the divergence between Ma46 and Ma49. Interestingly,
umbelliferone (C_{10}H_{8}O_{3}, [M+H]^+ = 163.0390) was identified, but no difference in its levels was found between Ma46 and Ma49, indicating that the genetic differences between the isogenic lines were not associated with the production of umbelliferone.

To identify genes involved in coumarins biosynthesis, we constructed the corresponding pathway. Thirteen gene families involved in coumarin biosynthesis were identified in *M. albus*, and the number of genes in each family was compared among the four species, as shown in Figure 5c. Genes encoding phenylalanine ammonia-lyase (PAL), which is the first committed enzyme in the phenylpropanoid pathway and is responsible for catalysing the transformation of phenylalanine to trans-cinnamic acid, showed expansion in Ma relative to Mt (8 genes in Ma and 6 in Mt). Modified derivatives of coumarin are derived from p-coumaric acids. Scopoletin is glycosylated by UDP-glycosyltransferase (UGT) to form scopolin (Li et al., 2001), and BGLU hydrolyses scopolin to transform it back into scopoletin (Ahn et al., 2010). However, the genes involved in scopoletin biosynthesis in *M. albus* remain uncharacterized. A total of 206 UGTs and 49 BGLUs were identified in Ma, while 229, 157, and 57 UGTs, and 53, 48, and 24 BGLUs were identified in Mt, Gm, and Lj, respectively (Figures 5a and S20). Other gene families related to coumarin biosynthesis, including C3H, C4H, C2'H, C3'H, 4CL, HCT, COMT, CCoAOMT, F6'H, COSY, and SBH (Figures 5c and S20), were also identified in the four species. There was little difference among the four species in terms of gene copy numbers (Figure 5c), meaning that the high coumarin content of sweet clover may not result from the size of the gene family.

**Identification of candidate coumarin biosynthesis genes**

To narrow down the genes related to coumarin biosynthesis and accumulation in *M. albus*, we conducted an RNA-seq analysis of NILs. Gene family members in the coumarin biosynthesis pathway exhibited diverse gene expression patterns between the two groups of NILs and under drought stress (Table S15, Figure S21), indicating the complexity of regulation in this pathway. All differentially expressed genes (DEGs) related to coumarin biosynthesis were subjected to cluster analysis. Some genes in the 11 clusters exhibited higher expression levels in Ma48, Ma49, and RP than in Ma46 and Ma47 (Figure S22), and BGLU1 and COMT1 were highly expressed in all five NILs (FPKM > 100). WGCNA of all expressed genes (FPKM > 0.5) from NILs showed that one module was highly correlated with coumarin content (module 1, cor = 0.85), while another module was negatively correlated with coumarin content (module 2, cor = −0.99) (Figure S23). The coexpression relationships of the genes in the two modules showed that seven COMTs, 12 BGLUs, 58 UGTs, three HCTs, five 4CLs, one C3H, three SBHs, three PALs, four CCoAOMTs, and one COSY were highly correlated with each other (Figure 5d), and most of them belonged to the 11 clusters. Furthermore, the expression levels of BGLU13 and SBH4, SBH2 and UGT, HCT2 and 4CL13, and most UGTs and BGLUs were correlated, indicating that genes upstream and downstream of the coumarin synthesis...
Figure 5  Coumarin biosynthesis in Melilotus albus. (a) Mass spectrum of coumarins from the leaves. Mass spectrum of coumaric acid glucoside (C_{15}H_{18}O_{8}) and its glucoside (C_{6}H_{12}O_{5}) (left), detected in negative ion mode; and coumarin (C_{9}H_{6}O_{2}) (right) detected in positive ion mode. (b) The intensity of the peaks for coumaric acid glucoside, its glucoside, and coumarin in Ma46, Ma49, and qc (qc, a mixed sample of Ma46 and Ma49, as a control). (c) Gene families involved in the coumarin biosynthesis pathway. Abbreviations for the enzymes involved in each catalytic step are shown in bold. The numbers under the enzyme abbreviations are (from left to right) the gene numbers per gene family in M. albus (red), Medicago truncatula, Glycine max and Lotus japonicus. PAL: phenylalanine ammonia-lyase, BGLU: β-glucosidase, UGT: UDP-glycosyltransferase, C4H: cinnamic acid 4-hydroxylase, 4CL: 4-coumarate, HCT: hydroxy cinnamoyl transferase, C2'H: q-coumaroyl CoA 2'-hydroxylase, COSY: coumarin synthase, C3H: 4-coumarate-3-hydroxylase, COMT: caffeic/5'-hydroxyferulic acid O-methyltransferase, CCoAOMT: caffeoyl CoA O-methyltransferase, F6'H: feruloyl-CoA 6'-hydroxylase, S8H: scopoletin 8'-hydroxylase. (d) Coexpression network of coumarin biosynthesis pathway genes. (e) Manhattan plot of the two NILs based on BSA. The annotated genes were related to coumarin biosynthesis and identified with SNPs and InDels between NILs Ma46 and Ma49.
pathway could be coordinately expressed. In addition, we found with high confidence that between Ma46 and Ma49, 18 coumarin pathway genes contained nonsynonymous SNPs or InDelS in their exons, including one 4CL, six BGLUs, one HCT, one PAL, and nine UGTs, based on BSA (Figure 5e, Table S16). The expression of ten of 18 genes was correlated, and eight of the ten UGTs were correlated. Confirmation of the MaBGLU1 enzyme was performed by conducting an enzymatic activity assay of the different translated products in Ma46 and Ma49, which was further supported by the absence of the MaBGLU1 gene in germinated seeds and a 715-bp band in the control (Figure S26). A genomic PCR analysis of each MaBGLU1-overexpressing line (MaBGLU1-OE) showed the expected band in all the transgenic lines except for the control (Figure S27). qRT-PCR analyses showed an obvious increase in relative expression levels in the positive transgenic lines, and the MaBGLU1 expression levels in the roots of OE-4 of NIL-Ma46 and OE-4 of NIL-Ma49 were 8.9 times and 10.5 times higher, respectively, than the control levels (Figure 6g and h). The scopoletin and scopoletin levels were increased in these lines (Figure S28). These results demonstrated the role of the MaBGLU1 gene in regulating coumarin content.

Discussion

We constructed a high-quality chromosome-level genome assembly of *M. albus*. The *M. albus* genome, composed of approximately 71% repetitive sequences, is relatively large in size (1.05 Gb) compared to the reported genomes of Fabaceae species, such as *Trifolium pretense* (309 Mb) (De Vega et al., 2015), *M. truncatula* (360 Mb), *C. arietinum* (738 Mb) (Varshney et al., 2013), *Lotus japonicus* (315 Mb) (Sato et al., 2008), and *G. max* (950 Mb) (Schmutz et al., 2010), except for *Arachis ipaensis* (1512 Mb) (Bertoli et al., 2016). Transposable elements (TEs) constitute a large part of many plant genomes and contribute to adaptive evolution in plants (Lisch, 2013). LTRs underwent a burst in sweet clover from 0.05 to 0.85 Mya, and over the past 0.2 Mya, the population size of sweet clover significantly declined during the Quaternary glaciation period. Comparative genomic analysis with the other species revealed more abundant and longer intact LTRs in *M. albus*. The LTR-associated genes were mainly enriched in stress response pathways, suggesting a relationship between LTR burst and the environmental adaptation of sweet clover.

*Mellilotus albus* and *M. officinalis* are morphologically similar in the vegetative stage. Both species are indigenous to Eurasia and have been introduced to many different areas around the world (Hultén and Fries, 1986). Although both species grow in the same habitat in many regions (Gucker, 2009), strong reproductive barriers exist between them (Darbyshire and Small, 2018). Studies based on biochemistry markers, karyotypes, and DNA microsatellites have only provided weak confirmation of the separation of the two species (Darbyshire and Small, 2018). Other methods applied to distinguish *M. albus* and *M. officinalis*, such as analyses of simple-sequence repeat (SSR) markers (Wu et al., 2016), EST-SSR markers (Yan et al., 2017), and DNA barcodes (Wu et al., 2017), have also failed to provide a clear distinction. There are notable differences between *M. albus* and *M. officinalis* in terms of flower colour and venations of the areolae on the mature pods (Voronchikhin, 1990). In this study, we first provided phylogenomic information that gene introgression has occurred between the two species, and strong gene flow from *M. albus* to *M. officinalis* was detected. Phylogenomic analysis revealed that *M. albus* and *M. officinalis* accessions formed independent clades.
overall, but 11.9% of the *M. albus* accessions clustered in the *M. officinalis* group and 2.8% of the *M. officinalis* accessions clustered in the *M. albus* group.

The numerous, diverse pharmacological effects of medicinal plants are largely dependent on their secondary plant metabolites (Hussein and El-Anssary, 2019). In plants, the synthesis and accumulation of secondary metabolites are very complex and are affected by both genetics and environmental factors, such as temperature, water, and salinity (Li *et al*., 2020). Plant coumarins are a class of secondary metabolites that are important for plant adaptation and defense (Stringlis *et al*., 2019). Coumarins are derived from the phenylpropanoid pathway (Bourgaud *et al*., 2006; Vanholme *et al*., 2019) and are categorized as simple coumarins, furanocoumarins, phenylcoumarins, and pyrano-coumarins based on their structural and biosynthetic properties (Bourgaud *et al*., 2006; Shimizu, 2014). Nearly 1000 natural coumarins with various biological activities have been isolated (Hussein and El-Anssary, 2019). The accumulation of scopoletin and fraxetin after pathogen infection benefits the growth of native bacteria, thus protecting host *Nicotiana* species from sudden wilt disease (Santhanam *et al*., 2019). Following the exogenous application of the priming agent acibenzolar-S-methyl (ASM) in sunflower, coumarins (scopolin, scopoletin, and ayapin) accumulate in the leaves, and significant amounts of scopoletin

---

**Figure 6** Evolutionary and functional analysis of the MaBGLU gene family. (a) Molecular structures and chromosomal locations of six MaBGLUs. (b) MaBGLU1 alleles with the variation sites identified through BSA. The bases showing variation are shown on the line. (c) The protein sequences of MaBGLU1 NIL-Ma46 and NIL-Ma49. A 2-base insertion in MaBGLU1 of NIL-Ma46 results in a frameshift and premature termination, as indicated by the *. (d) Phylogenetic analysis of the BGLU family in ten legume species and Arabidopsis. Upper right: phylogenetic analysis of the BGLU subfamily in *Melilotus albus*, *Medicago truncatula*, and *Arabidopsis*. (e) Relative expression of six MaBGLUs. (f) LC-MS/MS analysis of scopolin and scopoletin in the presence of recombinant MaBGLU1-Ma49 and the empty vector control. Recombinant MaBGLU1-Ma49 was incubated with scopolin (bottom), and the empty vector control was incubated with scopolin and scoepletin (top). Product formation was analysed by LC-MS/MS. (g) The relative expression level of MaBGLU1 in NIL-Ma46 overexpressing MaBGLU1-Ma49. (h) The relative expression levels of MaBGLU1 in NIL-Ma49 overexpressing MaBGLU1-Ma49.
are secreted on the leaf surface, which leads to reduced disease (Pieterse et al., 2012; Prats et al., 2002). Sweet clover is the richest known source of coumarins (Hoffmann, 2003). Here, we characterized many genes related to coumarin biosynthesis and observed that the coumarin content of sweet clover increased under drought stress. Our research has paved the way for elucidating the molecular mechanism of the influence of coumarins on stress adaptability using sweet clover as a model.

In sweet clover, we found that gene families related to secondary metabolite biosynthesis and biotic and abiotic stress responses were expanded, especially the BGLU family, which is involved in coumarin biosynthesis. In Arabidopsis, AtBGLU21, AtBGLU22, and AtBGLU23 hydrolyse scopoletin to produce scopoletin (Ahn et al., 2010). The expansion of genes associated with chemical and antibacterial defense pathways may play a role in adaptive evolution (Guan et al., 2016). Although the coumarin biosynthesis pathway has been well described in Arabidopsis (Ahn et al., 2010; Li et al., 2001; Schmid et al., 2014; Siwinska et al., 2018; Tsai et al., 2018; Vanholme et al., 2019), information on the pathway is lacking in Melilotus species. Only the biosynthetic pathways of coumarin (coumaroyl glucoside is converted into coumaric acid by the BGLU enzyme, and cinnamic acid undergoes ortho-(o)-hydroxylation to form o-coumaric acid, followed by spontaneous lactonization to form coumarin) have been studied (Martino et al., 2006; Poulton et al., 1980). 4-Hydroxycoumarin was first extracted from M. officinalis showing spoilage caused by fungi (Sun et al., 2020). In our study, a pathway of simple coumarin biosynthesis in sweet clover was constructed, and candidate genes were identified. Coumaric acid glucoside, o-coumaric acid, and coumarin were identified, and interesting variations were revealed among Melilotus NILs. The nonsynonymous variation of BGLU1 resulted in different enzymatic activities, demonstrating that BGLU1 is a key enzyme regulating scopoletin biosynthesis in M. albus.

In summary, the reference genome of M. albus revealed that the timing of LTR bursts and a population size bottleneck coincided, indicating that transposon amplification may have driven adaptive evolution in M. albus. Based on the joint analyses of transcriptomic, metabolomic, and biochemical assay results, we propose that BGLU1 regulates coumarin biosynthesis and that nonsynonymous mutations are responsible for the divergence between NILs showing high and low coumarin contents. The population analysis of widely distributed M. albus germplasms provides insights into the evolution and population structural differentiation of M. albus and M. officinalis. The genomic resources that we developed in this study offer valuable information that will facilitate efficient germplasm exploration and the improvement of sweet clover for pasture and medicinal uses, especially for the specific production of bioactive molecules in planta.

**Methods**

**Plant materials, DNA library construction, and sequencing**

Four near-isogenic lines (NILs) of M. albus (Ma46–49) and the recurrent parent (RP) were developed and used in this study; Ma48, Ma49, and RP have high coumarin contents and Ma46 and Ma47 have low coumarin contents (Luo et al., 2017; Wu et al., 2018). Ma46 was chosen for whole genome sequencing. Plants were grown in a growth room under a 25:18°C (day:night) cycle with natural light. Leaves, flowers, stems, and roots were collected from a single healthy Ma46 plant at the flowering stage, immediately frozen in liquid nitrogen and stored at −80°C. DNA was isolated from young leaves by the CTAB method and used for both PacBio and Illumina sequencing.

A total of 5 μg extracted DNA was randomly fragmented and DNA fragments of the desired length were gel purified. These DNA samples were end-repaired and ligated to adapters and then the samples were pooled, purified, amplified with primers compatible with the adapter sequences, and used to construct a 270 bp paired-end library. The library was sequenced on an Illumina HiSeq X Ten sequencing platform.

PacBio libraries were constructed following the PacBio manufacturer’s protocol. Genomic DNA was randomly sheared to an average size of 20 kb and then end-repaired using polishing enzymes. After purification, a library with a 20 kb insert size was constructed according to the PacBio standard protocol with the BluePippin size-selection system (Sage Science), and sequencing was conducted on the PacBio RS II platform.

Total RNA was isolated from seeds, stems, roots, and flowers during the flowering stage with three biological replications following sample collection methods (Zhang et al., 2021).

**Genome assembly and assessment of the assembly quality**

The genome size of M. albus was estimated based on k-mer analysis by Jellyfish (Marcais and Kingsford, 2011) using 134.28 Gb of Illumina reads. De novo assembly of the PacBio long reads was conducted using Falcon v 0.3.0. (Chin et al., 2016). The Falcon pipeline was used to correct errors and preassembly (parameters were optimized). Based on the pre-assembly contig N50, we used the following parameters to construct initial contigs: length cutoff = 16 000 and length cutoff_pr = 15 000. The contigs were improved using Arrow (https://github.com/PacificBiosciences/GenomicConsensus). The PacBio long reads were mapped to the assembled contigs with the blasr pipeline (Chaisson and Tesler, 2012). After Arrow correction, all the filtered illumina reads were mapped to the corrected contigs with BWA-mem (Li and Durbin, 2009), and the draft assembly was further polished using 3 runs of Pilon (Walker et al., 2014).

The completeness, coverage, and accuracy of the M. albus assembly was assessed by examining the alignment ratio of illumina short reads, RNA-seq data, and the presence of well-conserved core eukaryotic genes. The illumina short reads were aligned to the M. albus assembly using BWA (Li, 2013). RNA-seq data collected from leaves and roots were mapped to the assembled genome to evaluate the coverage. The completeness of the genome assembly was assessed by BUSCO v.3.0 (http://busco.ezlab.org/)(Simão et al., 2015) using the embryophytaodb10 dataset (http://busco.ezlab.org/datasets/embryophyta_odb10.tar.gz), which contains 1375 protein sequences and orthologous group annotations for major clades.

**Hi-C analysis and pseudomolecule construction**

The assembled contigs were anchored to the chromosomescale using a Hi-C proximity-based assembly approach (Burton et al., 2013). The Hi-C library was constructed using fresh young leaves of M. albus collected from the same individual used for PacBio sequencing. Nuclear DNA was cross-linked in situ with formaldehyde, extracted, and then digested with HindIII. The purified DNA was sheared to a size of 300–700 bp, end-repaired with T4 DNA polymerase, and further amplified and purified to construct the
sequencing library. The library was sequenced on an Illumina HiSeq X Ten platform in paired-end 150 bp mode. The Hi-C sequencing data were aligned to the assembled contigs using BWA (Li, 2013) with default parameters, and the quality was then assessed using HIC-Pro v.2.8.0 (Servant et al., 2015). The unique valid interaction pairs were uniquely mapped onto the assembled contigs, and the contigs were divided into equally sized bins (250 kb) to group into 8 chromosome clusters with Lachesis (http://shendurelab.github.io/LACHESIS/) using the following parameters: cluster_min_re_sites = 157, cluster_max_link_density = 2, cluster_noninformative_ratio = 2, order_min_n_res_in_trun = 167, and order_min_n_res_in_shreds = 161.

Repeat annotation
Structural predictions and de novo approaches were used to annotate M. albus genome repetitive sequences.LTR_Finder 1.05 (Xu and Wang, 2007), RepeatScout 1.0.5 (Price et al., 2005), and PILER-DF 2.4 (Edgar and Myers, 2005) with default parameters were used to build the M. albus de novo-based repeat library of consensus sequences. These repeat sequences were classified according to their characteristics and redundancy using PASTEClassifier 1.0 (Seberg and Petersen, 2009) with Repbase (Bao et al., 2015), which was used to build the final repeat library. The final repeat library was used to identify repetitive sequences in the M. albus genome using RepeatMasker 4.0.6 (Taraloga, Groaovac, and Chen, 2009). RepeatProteinMask within the RepeatMasker package was used to search against the TE protein datasets of Repbase to identify the repeat-related proteins in M. albus (Bao et al., 2015). The repeat sequences annotation of C. arietinum, G. max, and M. truncatula used the same methods as above.

Gene annotation
A combination of de novo predictions, homologue-based predictions, and RNA-seq-based predictions were adapted to predict the protein-coding genes of the M. albus genome. De novo predictions were performed using six ab initio gene-finding prediction programs that include Augustus (v.2.4) (Stanke and Waack, 2003), Genscan (v.3.1) (Burge and Karlin, 1997), GeneID (v.1.4) (Blanco et al., 2007), GlimmerHMM (v.1.2) (Majoros et al., 2004), GeneMarks-T (Tang et al., 2015), and SNAP (v.2006–07–28) (Korf, 2004) to detect genes in the M. albus genome with the default parameters. For homologue-based predictions, the protein sequences of five species (Arabidopsis thaliana, G. max, M. truncatula, Vigna angularis, and C. arietinum) were downloaded and mapped to the M. albus genome using TBLASTN with a cut-off e-value of 1e–5, and homologous genes were identified using GeneWise and GeMoMa (v.1.3.1) (Keilwagen et al., 2016). For RNA-seq-based predictions, RNA-seq data were assembled and aligned using PASA v2.0.2 (Haas et al., 2003), and TransDecoder (v2.0) ([https://sourceforge.net/projects/ transdecoder/]) and Cufflinks-2.2.1 were used to predict genes. Finally, all the predictions were integrated using EVM (v.1.1.1) (Haas et al., 2008) with default parameters to produce the nonredundant gene sets of M. albus. tRNAs were annotated by tRNAScan-SEM (Lowe and Eddy, 1997). rRNAs and miRNAs were predicted by Infernal (v1.1) (Nawrocki and Eddy, 2013) based on the Rfam database (Griffiths-Jones et al., 2005) and miRBase (Griffiths-Jones, 2006) with an e-value of 1e–5.

Protein-coding gene functional annotations were based on homologous alignment with BLAST (e-value < 1e–5) against Nr (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/), KOG (ftp://ftp.ncbi.nih.gov/pub/COG/KOG/), KEGG (https://www.kegg.jp/), SwissProt (http://www.uniprot.org), and TrEMBL (https://www.ebi.ac.uk/uniprot). Motifs and domains were annotated using InterProScan (v5.8–49.0) (http://www.ebi.ac.uk/Tools/pfa/interproscan/) to search against InterPro (v29.0) databases. The Gene Ontology (GO) terms (Dimmer et al., 2011) for each gene were obtained using the Blast2GO program (Conesa et al., 2005) based on Nr annotation.

Gene family and phylogenomic analysis
Orthologous and paralogous groups were identified using the annotated genes from 13 species including M. albus, M. truncatula, C. arietinum, L. japonicus, G. max, Cajanus cajan, Phaseolus vulgaris, A. ipaensis, Vigna radiata, T. pratense, A. thaliana, Vitis vinifera, and Fragaria vesca, by Orthofinder (Emms and Kelly, 2019). Gene family expansion and contraction were estimated by comparing the cluster size between the ancestor and each species using CAFÉ (Han et al., 2013). KEGG enrichment analyses were conducted for expanded gene families. The amino acid sequences of single-copy ortho-groups were aligned using MUSCLE (Edgar, 2004). Phylogenetic trees were constructed based on concatenated amino acid sequences of single-copy genes from M. albus and the 12 species by PhyML (Guindon et al., 2010) with the maximum likelihood method (1000 bootstraps) and visualized by Fig-tree ([https://github.com/rambaut/figtree/]). The divergence time among 13 species was estimated by the MCMCtree program implemented in PAML (Guindon et al., 2010). The MCMCtree parameters were set as follows: clock = 2, RootAge ≤ 1.73, model = 7, BDpars = 110, kappa_gamma = 62, alpha_gamma = 11, rgene_gamma = 23.18, sigma2_gamma = 14.5. Calibration times were obtained from the TimeTree database ([http://www.timetree.org/]) by placing soft bounds at the split node of F. Vesca–V. Vinifera (10–114 million years ago [Mya]), A. thaliana–P. vulgaris (106–108 Mya), and L. japonicus–P. vulgaris (53–65 Mya). All the amino acid sequences were aligned; the colinear genes within M. albus and between M. truncatula and M. albus were identified using MCScanX (Wang et al., 2012) with default parameters. Synonymous substitutions per synonymous site (Ks) between colinear genes were estimated using the codeml approach as implemented in PAML (Guindon et al., 2010; Yang, 2007). Fourfold degenerate synonymous sites of each single-copy gene family were used to estimate the molecular clock and divergence time among species.

Analysis of gene families involved in coumarin biosynthesis
Based on previous studies (Bourgaud et al., 2006; Siwinska et al., 2018; Vanholme et al., 2019), we constructed an in silico M. albus coumarin biosynthesis pathway and compiled gene families in the pathway. Relevant Arabidopsis and M. truncatula genes were used as queries, and coding and protein sequences of Arabidopsis and M. truncatula were downloaded from The Arabidopsis Information Resource ([https://www.arabidopsis.org/index.jsp]) and PlantGDB ([http://www.plantgdb.org/]), respectively. The corresponding genes in M. albus were identified based on a BLAST search against the M. albus genome with a cut-off e-value < 1e–5. All the identified sequences with redundant sequences removed were submitted to Pfam ([http://pfam.xfam.org/search#tabview=tab1]) for annotations and subjected to a NCBI Conserved Domain Search ([https://www.ncbi.nlm.nih.gov/ cdd/]) to further confirm their identities. Genes without the
conserved domain were discarded. Gene coding and protein sequences of 4CL, C4H, PAL, CCOMT, COMT, HCT, and C3H from Mt, Gm, and Lj were downloaded from the specialized database ‘Phenylpropanoids and plant defense – an updated genomic perspective’ (http://bioinfo.noble.org/manuscript-support/mpp). Other gene families from Mt, Gm, and Lj were identified based on orthologous gene groups from Orthofinder (Emms and Kelly, 2019). Genes in all these groups were submitted to Pfam and NCBI to confirm the presence of conserved domains as described above. Using the same methods, we identified the BGLU family genes in nine Leguminosae species. BGLU amino acid sequences of Arabidopsis, M. albus, and nine other Leguminosae species were aligned using MUSCLE (Edgar, 2004), and phylogenetic trees were constructed by IQ-TREE (Nguyen et al., 2015) with bootstrap testing (1000 replicates).

Transcriptome sequencing and analysis

Total RNA of each sample was isolated using the TRIzol reagent (Invitrogen) following the manufacturer’s instructions. cDNA libraries were prepared and sequenced using a paired-end HiSeq2500 with a read length of 125 bp. Clean reads were aligned to the M. albus genome using HISAT2 (Daehwan et al., 2015). The expression values of genes were quantified by the FPKM using StringTie (1.3.1) (Mihaela et al., 2015), and Cuffdiff was employed for differential expression analysis (Trapnell et al., 2010).

A transcriptomic analysis of NILs was performed when no reference genome was available, with nine differentially expressed unigenes predicted to encode the enzymes associated with the coumarin biosynthesis pathway (Luo et al., 2017), seven of which were included in this study (Table S17).

In addition, to study the expression relationships of gene families and identify candidate genes that participate in coumarin biosynthesis, we conducted a WGCNA (Langfelder and Horvath, 2008) analysis to define modules of coregulated genes that varied in five NILs and tissues. A total of 11 modules were finally detected to have highly similar expression, and they were assigned individual colours (Figure S15).

SNPs/InDels detection in M. albus accessions

A collection of 94 Melilotus accessions was sourced, 35 M. albus accessions were collected from the National Gene Bank of Forage Germplasm (NGBFG, China), and 60 other accessions were collected from the National Plant Germplasm System (NPGS, USA). The 94 Melilotus accessions included 42 accessions of M. albus, 36 accessions of M. officinalis, and one accession each of 16 other Melilotus species. All accessions were cultivated in greenhouse in Yuzhong, Lanzhou, China. Genomic DNA was extracted from fresh young leaves using the CTAB method (Murray and Thompson, 1980). A total of 94 paired-end libraries with an average insert size of 250 bp were constructed and then sequenced using the Illumina HiSeq 2500 platform. Medicago archiducis genome sequences (NCBI, SRX9404272) were downloaded and used as the outgroup. Low-quality bases, adaptor sequences, duplications, and potential contaminations of raw reads were filtered, and then clean reads were aligned against the M. albus genome assembled with the BWA (Li, 2013) package using default parameters. SAMtools (Li et al., 2009) was used to filter the unmapped reads, convert them into the BAM format, and sort the mapped reads. GATK (Mckenna et al., 2010) local realignment was performed to refine the read mapping in the presence of the variants. After realignment, SNP calling was carried out by GATK Haplotype Caller program (McKenna et al., 2010) with the following parameters: standard emit confidence (-stand_emit_conf) = 10 and standard call confidence (-stand_call_conf) = 30. To reduce the false discovery rate of SNPs/InDels, raw variants were filtered using Variant Filtration in GATK with the following parameters: removal of sites with quality scores < 30, quality divided by depth (QD) < 2.0, mapping quality (MQ) < 40.0, Fisher’s exact test (FS) > 60.0, minor allele frequency < 0.05, and missing genotype rate < 0.2. SNPs and InDels were annotated based on their chromosomal locations.

Phylogenetic, population structure and genetic diversity analyses

Multiconsensus sequences of SNPs were retrieved in fasta format. The 94 sequences were used to generate maximum-likelihood (ML) phylogenetic trees. ML trees were generated by FastTree (Price et al., 2009) using the ultrafast bootstrap approach (UFBoot) with 1000 replicates. ADMIXTURE was used to examine the population structure with K values ranging from 1 to 10 (Alexander et al., 2009). Principal component analysis (PCA) was also used to evaluate the genetic structure of the populations using the smart pca function in EIGENSOFT (Patterson et al., 2006). The same SNP dataset was transformed into a population allele frequency table according to the sample grouping results from ADMIXTURE when K = 3. Introggression analysis between M. albus and M. officinalis was performed using Dsuite – Fast D-statistics methods (Malinsky et al., 2020). Population split and mixture analysis were inferred with Treemix version 1.13 (Pickrell and Pritchard, 2012). Genome-wide pairwise linkage disequilibrium (LD) was estimated independently for M. albus and M. officinalis. Haploview (Barrett et al., 2005) was used to calculate the squared correlation coefficient (r^2) between pairwise SNPs for M. albus and M. officinalis accesses with the parameters ‘-n -dprime -minMAF 0.05’. The r^2 values were analysed at a 500-kb distance across the whole genome and plotted using the PopLDdecay package (Zhang et al., 2018a). Past evolution of the effective population size of M. albus and M. officinalis was inferred by applying SMC++ (Terhorst et al., 2016). The SMC++ vcf2smc tools were used to generate the input file, and SMC++ inference was performed for the different sets of genotypes using the SMC++ cv command with lower (200 generations) and upper (25 000 generations) time point boundaries and other default options. We considered a mutation rate of 6.5 x 10^-8 per synonymous site per generation.

Genome-wide selective sweep analysis

To identify candidate genomic regions potentially affected by selection, nucleotide diversity (π) and population fixation statistics (Fst) were calculated using VCFTools (Danecek et al., 2011) in a 50-kb sliding window with a step size of 5 kb. π and Fst were used to estimate the degree of pairwise genomic differentiation of candidate genes between pairs of subpopulations. The reduction in diversity (ROD) values were calculated based on the ratio of π for a subpopulation with respect to a control subpopulation. All the output results of ROD and Fst were standardized and transformed into z-scores using a 10-kb sliding window with a 5-kb step size. Candidate genomic regions were
selected based on the top 1% values. Candidate orthologous genes were aligned using DNAMAN.

Gene cloning and subcellular localization assay of *MaBGLU1*

The coding sequence (CDS) of *MaBGLU1* was cloned using the specific primers listed in Table S18 based on sequences from the genome data. To investigate the subcellular localizations of *MaBGLU1*, we constructed recombinant BGLU1 tagged at the N-terminus with RFP, and the CDS was cloned into the pBI121-DsRed2 vector. The specific primers are listed in Table S18. The leaves of 6-week-old *Nicotiana benthamiana* seedlings were selected, and *Agrobacterium tumefaciens* GV3101 carrying pBI121-MaBGLU1-RFP was injected into the leaves. After 2 days, the RFP signals of infiltrated leaves were examined.

Heterologous expression of the *MaBGLU1* gene and β-glucosidase activity assay in vitro

For heterologous expression with an N-terminal His-Tag, the genes were inserted into the pET32a expression vector using the ClonExpress® MultiS One Step Cloning Kit (Vazyme Biotech Co., Ltd, Nanjing) following the provided protocol. Single colonies of BL21 (DE3) cells harbouring the pET32a-*MaBGLU1* vectors were cultured at 37°C in liquid lysogeny broth (LB) medium containing 100 µg/mL ampicillin to an optical density of 0.6–0.8 at 600 nm. Then, isopropyl thio-β-D-galactoside (IPTG) was added, and incubation was continued at 22°C for 18 h with shaking to induce target protein expression. Finally, cells were collected and then disrupted. The supernatant was collected and checked for the presence of the expressed protein using 12% SDS-PAGE.

BL21 (DE3) cells harbouring the pET32a-BGLU1 vectors and empty vector were cultured under the same conditions. After IPTG was added for 9 h, the substrate scopolin was added and culture was continued for 9 h. The reaction system was extracted with an equal amount of ethyl acetate. The recovered ethyl acetate was decompomposed and dried and then dissolved in MeOH for high-performance liquid chromatography (HPLC) analysis.

Overexpression of *MaBGLU1* in *M. albus*

The roots of healthy seedlings were cut. Radicals were coated with *Agrobacterium rhizogenes* strain K599 harbouring the pBI121-*BGLU1* plasmid. Then, cocultivation was conducted on BL21 (DE3) cells harbouring the pET32a-BGLU1 vectors and empty vector were cultured under the same conditions. After IPTG was added for 9 h, the substrate scopolin was added and culture was continued for 9 h. The reaction system was extracted with an equal amount of ethyl acetate. The recovered ethyl acetate was decompomposed and dried and then dissolved in MeOH for high-performance liquid chromatography (HPLC) analysis.

RNA extraction and qRT-PCR

Total RNA was extracted from leaves, stems, flowers, roots and seeds of *M. albus* using the TransZol reagent (TransGen Biotech, Beijing). Reverse transcription and cDNA synthesis were performed using the TIANScript II RT Kit (Tiangen, Beijing) from 1 µg of total RNA. Quantitative RT-PCR was performed using Hieff® qPCR SYBR® Green Master Mix (No Rox) (Yeasen Biotech Co., Ltd., Shanghai) on a CFX96 Real-Time PCR Detection System (BioRad, Los Angeles, CA). The Tubulin gene was used as an internal control. All primer sets are listed in Table S19. The expression levels were calculated relative to the controls and determined using the 2−ΔΔCt method. Three biological replicates were used for all analyses.

Other experimental procedures

The other experimental procedures are available in supporting materials and methods of the Supporting Information (Tables S20 and S21).

Acknowledgments

This work was supported by the National Natural Science Foundation of China (32061143035), Gansu Provincial Science and Technology Major Projects (19ZD22NA002), the Special Fund for Agro-scientific Research in the Public Interest (20120304205), the Gansu Provincial Intellectual Property Program (19ZSCQ044), and the 111 Project (B12002). We thank Drs Qiang Qiu, Xiang Yu, and Ulrik P. John for their helpful comments and suggestions on the manuscript.

Conflicts of interest

The authors declare that they have no conflict of interest.

Author contributions

J.Z. and Y.W. managed the project; J.Z., F.W., P.X., and Q.Y. designed research. F.W. and Z.D. wrote the manuscript. J.Z., F.W., K.L., Z.Y., P.W., H.D., and Z.O. collected and planted plant material; F.W. and P.Z. prepared the samples; P.X. processed the raw data, annotated the genome. P.X. and M.M. analysed the resequencing data; F.W., P.X., Q.Y., and X.Z. executed transcriptome sequencing and analysis; F.W., P.Z., and Y.W. conducted the coumarin extracting and analysis; Z.D., F.W., Y.W., and S.W. conducted experiment of gene function verification; J.Z., M.C., and C.S.J. revised the manuscript.

Data availability statement

Data supporting the findings of this work are available within the paper and its Supplementary Information files. The datasets generated and analysed during the current study are available from the corresponding author upon request. The raw PacBio, Illumina, and Hi-C data are available from the National Center for Biotechnology Information Short Read Archive (NCBI SRA) under bioproject PRJNA674670. Raw resequencing reads of *Melilotus* accessions have been deposited in the NCBI SRA under accession number PRJNA759778. The assembled genome sequences and genome annotations have been deposited in the National Genomics Data Center under accession PRJCA006363. Transcriptome sequences of tissues have been submitted to the NCBI SRA under bioproject PRJNA647665.

References

Ahn, Y.O., Shimizu, B., Sakata, K., Gantulga, D., Zhou, C., Bevan, D.R. et al. (2010) Scopolin-hydrolyzing beta-glucosidases in roots of Arabidopsis. Plant Cell Physiol. 51, 132–143.

Alexander, D.H., Novembre, J. and Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, 1655–1664.

Bao, W., Kojima, K.K. and Kohany, O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA, 6, 11.

Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics, 21, 263–265.

Bello, R.I., Gomez-Diaz, C., Lopez-Lluch, G., Forthoffer, N., Cordoba-Pedregosa, M.C., Navas, P. and Villalba, J.M. (2005) Dicoumarol relieves...
Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Figure S1** Khoestome counts (2n = 16) in a *M. albus* cell.

**Figure S2** 4-mer frequency distribution.

**Figure S3** Intercromosomal contact matrix. The intensity of gray represents the normalized count of Hi-C links between 250 kb windows on 8 chromosomes on a logarithmic scale.

**Figure S4** Comparisons of gene structural parameters (the number and length of exons and intron, CDS length, gene length) among the seven legume species.

**Figure S5** KEGG pathway of expanded gene families.

**Figure S6** Physiological, chemical, and gene expression response under drought stress.

**Figure S7** Gene expression heatmap for the expanded family members under drought stress. CKL, H-3L, H-24L: shoot of 0, 3, 24 h of drought treatment, respectively; CKKR, H-3R, H-24R: root of 0, 3, 24 h of drought treatment, respectively.

**Figure S8** Gene coexpression network analysis for genes expressed in root and shoot under drought stress and different tissues.

**Figure S9** A scatterplot of module membership vs gene significance for genotype in the four modules.

**Figure S10** KEGG enrichment analysis of genes in four modules.

**Figure S11** Syntenic of *M. truncatula* and *M. albus*.

**Figure S12** Location relationship between LTR and two types of genes.

**Figure S13** KEGG enrichment analysis of orthologous genes between *M. truncatula* and *M. albus*.

**Figure S14** Cross-validation errors for population structure analysis.

**Figure S15** PCA plots of the first two components for the 94 *Melilotus* accessions.

**Figure S16** Analysis of phylogenetic tree and population structure with K = 3 based on chloroplast genes.

**Figure S17** The distribution of (a) *M. albus* and (b) *M. officinalis* around the world.

**Figure S18** Fst and Δ based selective sweep identification between *M. albus* and *M. officinalis* accessions.

**Figure S19** KEGG annotation of the selected genes identified by Fst and Δ in *Melilotus*.

**Figure S20** Location of coumarin biosynthesis pathway genes on chromosomes.

**Figure S21** Gene expression heatmap for the coumarin biosynthesis pathway family members under drought stress.

**Figure S22** Expression heatmap and cluster analysis of coumarin biosynthesis pathway genes in four NILs and RP.

**Figure S23** Gene coexpression network analysis for genes expressed in shoot with different coumarin content and different tissues.

**Figure S24** Subcellular location of MaBGLU1.

**Figure S25** SDS–PAGE profiles of recombinant BGLU1 from expression in *Escherichia coli*.

**Figure S26** Observation and photo taking of RFP in control and OE-BGLU1 roots under a laser scanning confocal microscope.

**Figure S27** PCR amplification of MaBGLU1 from the gDNAs of each transgenic line using forward and reverse primers specific to 3SS promoter and RFP region, respectively.

**Figure S28** HPLC profiles of scopelin and scopoletin made in MaBGLU1 overexpression hairy roots.

---

© 2021 The Authors. Plant Biotechnology Journal published by Society for Experimental Biology and The Association of Applied Biologists and John Wiley & Sons Ltd., 20, 592–609
**Table S16** Information of genes identified by BSA, these genes belong to coumarin biosynthesis pathway.
**Table S17** The corresponding ID between Unigenes and genes from two versions.
**Table S18** Primer sequences for MaBGLU1.

**Table S19** List of gene-specific primers used in this study.
**Table S20** Conditions of liquid chromatography mobile phase.
**Table S21** Accessions in eight groups.
**Supplementary Data 1** Selective sweep regions of *M. officinalis* and *M. albus*. 