1 Duplications and functional specialization force distinct evolution of 2 isoflavonoid biosynthetic genes in legumes

3 Yin Shan Jiao^{a, 1}, Yu Zhao^b, Wen Feng Chen^{a,1}

4 a. State Key Laboratory of Agrobiotechnology, Beijing 100193; College of Biological
5 Sciences and Rhizobium Research Center, China Agricultural University, Beijing 100193,
6 China.

7 b. College of Biological Sciences; China Agricultural University, Beijing 100193, China.
8 Email: yu1.zhao@wur.nl

9
10
11 1. Authors for correspondence:

12 Yinshan Jiao Tel: +86 18810803969 Email: jiaoyinshan@126.com, or Wenfeng Chen Tel: +86
13 62734009 Email: chenwf@cau.edu.cn

14
15
16
17
18
19 **Running title:** distinct evolution route of isoflavonoid biosynthetic genes in legumes

20 **Highlight:** Isoflavonoids are almost exclusive to the legume family. We firstly revealed that
21 gene duplications and functional specialization played essential roles in evolutionary origin of
22 legume-specific isoflavonoid biosynthesis.

23 **Key words:** Gene duplication, Isoflavonoid, CHS, CHR, IFS, IFR
24
25
26
27

Abstract

Isoflavonoids are specialized plant metabolites, almost exclusive to legumes, and synthesized by the phenylpropanoid pathway. Leguminous plants produce 5-deoxyflavonoids and 5-deoxyisoflavonoids that act in symbiosis with nitrogen-fixing bacteria and involved in plant pathogen and stress response. However, little is known about evolutionary origin of legume-specific isoflavonoid biosynthesis pathway. Here, we explored the genome-wide analysis of key genes: chalcone synthase (CHS), chalcone reductase (CHR), isoflavone synthase (IFS) and isoflavone reductase (IFR), encoding enzymes involved in the biosynthesis of (iso) flavonoids in legumes and nonlegumes. Among them, *CHS*, *CHR* and *IFR* comprise multigene families, underling the significant role of gene duplication in the evolutionary. Most duplications of *CHS* were highly the conventional leguminous type, whereas some were grouped with nonleguminous *CHS* genes. We also found that *CHR* homologs in soybean and *Sesbania rostrata* previously reported were ambiguous and should be re-identified. Phylogenetic analysis and protein sequences alignment indicated that IFSs in legumes are highly conserved. Intriguingly, unlike other IFRs in legumes, IFR-like homologs in *Sophora flavescens* and *Lupinus angustifolius* shared high sequence similarity and protein structures with homologs in nonlegumes. Overall, these results offer reasonable gene annotations and comparative analysis and also provided a glimpse into evolutionary route of legume-specific isoflavonoid biosynthesis.

Highlight: Isoflavonoids are specialized plant metabolites, almost exclusive to legumes. We firstly provide evidence that evolutionary origin of legume-specific isoflavonoid biosynthesis may be driven by gene duplications and functional specialization.

Key words:

Isoflavonoids, legume, evolution, gene duplication, CHS, CHR, IFS, IFR

58 Introduction

59 Flavonoids including flavones, flavanols and condensed tannins are plant secondary
60 metabolites and ubiquitous in nature. More than 9,000 different flavonoids have been
61 identified in plants, and a particular subset of them is involved in mediating host specificity in
62 legumes (Ferrer *et al.*, 2008; Perret *et al.*, 2000). Isoflavonoids, a class of flavonoids, form a
63 group of distinct secondary metabolites distributed almost exclusively in legume family.
64 Many legumes produce specific flavonoids that only induce Nod factor production in
65 homologous rhizobia, and therefore act as important determinants of host range (Liu and
66 Murray, 2016). Daidzein and genistein, two isoflavonoids produced by soybean (*Glycine*
67 *max*), are crucial signaling molecules that induce *Bradyrhizobium japonicum* to secrete Nod
68 factors to establish the symbiosis. *Medicago truncatula*, a model legume that has symbiosis-
69 specific relationship with rhizobia, is only nodulated by either *Sinorhizobium meliloti* or
70 *Sinorhizobium medicae* (Andrews and Andrews, 2017). On the contrary, more than 30
71 rhizobial species can establish symbiosis with *Sophora flavescens*, the most broad-spectrum
72 host (Jiao *et al.*, 2015). Several studies have shown that the production and release of
73 flavonoids is central to how host-symbiont specificity is achieved (Den Herder *et al.*, 2007;
74 Perret *et al.*, 2000). However, knowledge about genetic gene resources involved in the
75 biosynthetic pathways of flavonoids in legume is limited and evolutionary comparative analysis
76 are lacking.

77 Vestitol, an isoflavan phytoalexin produced by *Lotus japonicus*, is postulated to be
78 synthesized as part of the phenylpropanoid pathway (Fig. 1). It has multiple branches
79 common to non-legumes as well as legumes, which provide general flavonoids such as
80 condensed tannins and isoflavonoids including glyceollins, medicarpin and vestitol. In *L.*
81 *japonicus*, the cDNAs encoding 10 enzymes involved in the biosynthesis of vestitol have been
82 identified, except 7,20-dihydroxy-40-O-methoxyisoflavanol dehydratase. Chalcone synthase
83 (CHS) is a key enzyme that catalyzes the first committed step towards flavonoid biosynthesis.
84 Based on phylogenetic analysis, legumes have had a further expansion of one branch of the
85 CHS gene family (Zavala and Opazo, 2015). Isoliquiritigenin is produced by a legume-
86 specific enzyme, chalcone reductase (CHR), acts on an intermediate of the CHS reaction,
87 yielding deoxychalcone from the coupled catalytic action with CHS (Wang, 2011). CHR
88 belongs to the aldo-keto reductase (AKR) superfamily (Jez *et al.*, 1997). All members of this
89 superfamily contain a catalytic tetrad of Asp-53, Tyr-58, Lys-87, and His-120 and a common
90 NAD(P)H binding site that is located in a deep, large and hydrophobic pocket at the C-

terminus end (Bomati *et al.*, 2005). Recently, Caroline J. Sepiol *et al.* performed a genome-wide search of AKR gene family members and identified 14 CHR genes in soybean (Sepiol *et al.*, 2017). Up to now, CHR enzymes have been identified in various leguminous plant species, including *Medicago sativa* (Ballance and Dixon, 1995), *Sesbania rostrata* (Goormachtig *et al.*, 1999), *Pueraria montana* var. *lobata* (He *et al.*, 2011), *Glycyrrhiza glabra* (Hayashi *et al.*, 2013), and *L. japonicus* (Shimada *et al.*, 2007a).

Isoflavone synthase (IFS), a key cytochrome P450 enzyme, is the first step in the branch of the phenylpropanoid pathway that commits metabolic intermediates to the synthesis of isoflavones (Wang, 2011). *Arabidopsis* possesses 244 putative P450 genes (and 28 pseudogenes), but none of them has IFS activity for synthesizing isoflavone (Bak *et al.*, 2011). In soybean, IFS1 and IFS2 were identified by using a soybean EST collection (Jung *et al.*, 2000). Introducing soybean IFS1 into *Arabidopsis thaliana* resulted in production of genistein (Jung *et al.*, 2000), indicating its central role in the biosynthesis of isoflavonoids and extremely high substrate specificity. Interestingly, a recent study has shown that both IFS1 and IFS2 show positive interaction with upstream enzymes involved in isoflavonoid biosynthesis in soybean (Dastmalchi *et al.*, 2016). Isoflavone reductase (IFR) was first identified as a key enzyme involved in the latter part of the medicarpin biosynthetic pathway (Guo *et al.*, 1994; Paiva *et al.*, 1991). IFR converts 2'-hydroxyformononetin to (3R)-vestitone (Fig. 1). In *M. truncatula*, MtIFR was first identified as a key enzyme involved in the biosynthetic pathway of the medicarpin, a major isoflavonoid phytoalexin (Paiva *et al.*, 1991). Similar function of IFR homologs were also reported in chickpea (*Cicer arietinum*) and pea (*Pisum sativum*) (Daniel *et al.*, 1990; Sun *et al.*, 1991). Recently, a novel member of the soybean isoflavone reductase gene family *GmIFR* was identified (Cheng *et al.*, 2015). Over expression of *GmIFR* transgenic soybean exhibited enhanced resistance to *Phytophthora*. Although isoflavonoid phytoalexins are restricted primarily to legumes, IFR-like proteins have been isolated from various nonlegumes such as maize (*Zea mays*) (Petrucchio *et al.*, 1996), rice (*Oryza saliva*) (Kim *et al.*, 2003) and tobacco (*Nicotiana tabacum*) (Shoji *et al.*, 2002). Here, we screened the available genomic and transcriptional resources of legumes distributed in three phylogenetic lineages to identify candidate genes encoding the four key enzymes: CHS, CHR, IFS and IFR. Phylogenetic relationship, comparative analysis of nucleotide and protein sequences and evolutionary processes of these four gene homologs in legumes and nonlegumes, were investigated. We demonstrate that: (a) similar to *CHI*, *CHS* also has legume-specific type and evolved more rounds of gene duplications compared to those in

nonlegumes; (b) *CHR* homologs in soybean and *CHR* of *Sesbania rostrata* previously reported are ambiguous and should be re-identified; (c) IFSs share high levels of Sequence identity, indicating highly conserved during proteins evolution in plants; (d) *IFR-like* genes in *S. flavescentis* and *Lupinus angustifolius* were clustered with homologs in nonlegume, sharing high sequence similarity and similar protein structures. These findings suggest the functional diversification of the multigene families, due to gene duplications in legume. Overall, our results offer reasonable gene annotations and comparative and functional analysis of legumes, as well as insights into the genetic evolution of the legume-specific biosynthetic genes of isoflavonoids.

Methods

Bioinformatics analysis

As no genome are available for *S. flavescentis*, we constructed a *de novo* transcriptome assembly by using newly generated root RNA-seq data (unpublished). These *CHS*, *CHR*, *IFS* and *IFR* homologs in *S. flavescentis* were identified by performing BLASTN from the BLAST+ package (Camacho *et al.*, 2009) against our RNA-seq database. For each gene performance analysis, sequences of homolog genes reported in other legumes were used as queries and retrieved from NCBI database. For example, to identify *IFR* candidate genes of *S. flavescentis*, *IFR* genes from *Medicago sativa* (accession number X58078.1) (Paiva *et al.*, 1991) and *Pisum sativum* (accession number S72472.1) (Paiva *et al.*, 1994) were blasted against our RNA-seq library.

Genome-wide analyses of *CHS* homologs in *M. truncatula*, *L. japonicus*, *G. max* and related nonlegumes have been identified (Zavala and Opazo, 2015). In our study, these sequences were re-conformed and located on the precise position of chromosome in legumes (https://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/). *CHR*, *IFS* and *IFR* homologs in *M. truncatula*, *L. japonicus*, *G. max*, *V. unguiculata*, *P. vulgaris*, *C. cajan* and *V. angularis* were mainly obtained from PlantGDB database (<http://www.plantgdb.org/>) and Legume information system (<https://legumeinfo.org/>). Phytozome, annotated with KOG, KEGG, ENZYME and pathways, was also included to attain broader coverage. In *Lupinus angustifolius*, homologs of *CHR* and *IFS* were blasted against recent released reference genome sequence (Hane *et al.*, 2017). As reference analysis, related homologs of nonlegumes were retrieved from GenBank. All genes and accession numbers used were listed in Table 1.

Sequence similarity and phylogenetic analysis

Evolutionary similarity of *CHSs* nucleotide sequences between legume-specific and nonlegume groups was conducted in MEGA7 (Kumar *et al.*, 2016). The estimates were obtained by a bootstrap procedure with 1000 replicates. Analyses were conducted using the Maximum Composite Likelihood model. All positions containing gaps and missing data were eliminated. The nucleotide sequences of *CHS*, *CHR*, *IFS* and *IFR* genes were aligned manually. The alignment lengths were 1173 bp for *CHSs*, 1077 bp for *CHR*, 1620 bp for *IFSs* and 1412 bp for *IFRs*, respectively. The evolutionary relationship of these four genes were inferred by using the Maximum Likelihood (ML) method based on the Tamura-Nei model (Tamura and Nei, 1993) integrated in MEGA 7. Initial trees for the heuristic search were obtained automatically by applying Neighbor-Joining and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value.

Comparative analysis of protein sequence

The protein sequences of IFS and IFR homologs in all studied species were translated from coding sequences. The alignment was conducted using ClustalX 2.1 and checked manually. As we did not obtain the whole coding sequence of IFS2 of *S. flavescens*, it was excluded in the further protein sequence alignment. The secondary structure elements in IFR structure were predicted according to the previous result (Wang *et al.*, 2006).

Results

Identification of *CHS*, *CHR*, *IFS* and *IFR* homologs

We focus on four essential enzymes, chalcone synthase (CHS), chalcone reductase (CHR), isoflavone synthase (IFS) and isoflavone reductase (IFR), involved in the pathway of isoflavonoid biosynthesis in legumes. In order to shed more light on evolutionary route of four enzymes in legumes belonging to different phylogenetic Papilionoideae tribes, genomic resources of the inverted repeat loss clade (IRLC) of the Hologalegina (*Medicago truncatula*), robinoids (*Lotus japonicus*), millettoids (*Glycine max*, *Phaseolus vulgaris* and *Vigna angularis*) and basal genistoids (*Lupinus angustifolius* and *Sophora flavescens*) were investigated in our study. As no genome data of *Sophora flavescens* is available, RNA sequencing (RNA-seq) and de novo assembly were performed to obtain the repertoire of isoflavonoids biosynthesis genes in *S. flavescens*. Total 137828 unigenes representing the

total root transcriptome of *S. flavescens* were generated by RNA-seq (unpublished data). Meanwhile, homologs of *CHS*, *CHR* and *IFR* in some nonlegumes from public genomic data were also identified for comparative analysis.

Multiple duplications of *CHS* homologs in legumes

Similar to *CHS* genes (*CHSs*) reported in other legumes (Zavala and Opazo, 2015), *CHSs* in *S. flavescens* also have many copies that *CHR1* (c32443 g1), *CHR2* (c34542 g1), *CHR3* (c51007 g1), *CHR4* (c52952 g2), *CHR5* (c63917 g4) were identified from root library (Table 1). To compare phylogenetic evolution of the *SfCHSs* with those in other legumes and nonlegumes, we download *CHSs* sequences that have been annotated (Zavala and Opazo, 2015) and further conformed the *CHSs* positions in genomes of different plant species by blasting (<http://www.plantgdb.org/>). The phylogenetic tree was reconstructed by using maximum likelihood (ML) method based on nucleotide sequences (Fig. 2). It showed that the copy number of putatively functional *CHSs* in legumes is higher than those in nonlegumes. For example, 18 *CHSs* and 1 *CHS* copy were present in *M. truncatula* and *A. thaliana*, respectively (Fig. 2). The result suggested that *CHSs* in legumes underwent more duplication events than those in nonlegumes. Meanwhile, most *CHSs* were phylogenetically grouped into a divergent legume-specific clade, named type II (Fig. 2). Intriguingly, *SfCHS4*, together with *CHS14* in *G. max*, *CHS3* in *L. japonicus* and *CHS18* in *M. truncatula*, was distributed to one of ancestral nonlegumious lineages (type I) that also contained two *CHS* homologs in *A. thaliana* and *A. lyrata*. Further, nucleotide sequences identities of representative *CHSs* in both legume-specific and nonlegume groups were calculated in every combination (Table S1). The high sequence identities (about 80%) within the legume-specific type II *CHSs* showed a close orthologous relationship each other. In contrast, the nucleotide identities between the two types of *CHSs* in legumes were significantly lower 75%. Taken together, these observations supported the hypothesis that the evolution of *CHSs* in legumes was derived from an ancestral repertoire presented in the last nonlegumious ancestor (Fig. 1).

CHR homologs in legumes as AKR4

Three putative *CHR* genes (*CHRs*) in *S. flavescens*, named *SfCHR1-SfCHR3*, were identified using BLAST against our RNA-seq library. To figure out copy number variations and phylogenetic relationship of *CHRs* among different legume species, *CHR* homologs in *L. japonicus*, *M. truncatula* and *G. max* were searched from Legume Genome Database

(<https://legumeinfo.org/>). Besides, other related genes such as genes encoding codeinone reductases (COR) were selected as reference. These nucleotide sequences in all studied species were aligned for building phylogenetic tree (Fig. 3). The result showed that *CHRs* in legumes were grouped together in a distinguished subclade of ARK4 family, indicating highly conserved in the evolutionary history. *COR* homologs in nonlegumes were out-grouped from clade of *CHR*, constituting another subclade of ARK4 family. Further analysis showed *CHRs* from these four legume species fell into 3 distant subclades (Fig. 3). In clade I, *MtCHR2-MtCHR4*, located on chromosome 5, shared high similarity at nucleic acid lever (Table 1). Besides *SfCHR1*, *GmCHR1-GmCHR2* and *LjCHR3-LjCHR6* were gathered together, respectively. Some *CHR* homologs such as *SfCHR2* and *MtCHR1* were placed into divergent subclades belonging to Clade II. *LjCHR1*, distributed in clade II, showed catalytic activity in a previous study (Shimada *et al.*, 2007a), giving evidence that other *CHRs* in this clade may also encode active enzymes. Notably, clade III as an early branch of *CHRs* was firstly investigated. These data suggested that *CHRs* in legumes were possibly duplicated twice before the speciation of Papilionoideae in the evolutionary processes.

Recently, *CHRs* in *G. max* were investigated and have 11 genes copy, which distributed in multiple clades of AKR family (Sepiol *et al.*, 2017). However, *CHR9A*, *CHR9C*, *CHR9D*, *CHR12*, *CHR16A* and *CHR16B* from *G. max* were distantly placed out of the *CHRs* clade in our study. To further conform the identities of these 6 *CHRs*, we blasted them in the NCBI database. We found that *CHR9D* and *CHR12* were annotated as codeinone reductase-like protein and shared more than 85 % nucleotide sequence similarity with known *COR* homologs in legumes. More supersizing, *CHR9A*, *CHR9C*, *CHR16A* and *CHR16B* have highest similarity with genes coding methylecgonone reductase (*MecgoR*) in *C. arietinum* and *L. angustifolius*. Furthermore, similar case was also observed in *S. rostrata* (Goormachtig *et al.*, 1999) that *SrCHR* gene (AJ223291.1) was clustered into *MecgoR* group (Fig. 3). Based on our blasting results and phylogenetic analysis, this gene was supposed to be a *MecgoR*-like gene. Overall, our study provides strong evidences for identifies of *CHRs* in AKR4 family and helps us to distinguish phylogenetic relationship between *CHRs* and other closed genes like *MecgoRs* and *CORs* in legumes.

Conserved evolution of IFS homologs in legumes

Isoflavone synthase (IFS), also named as 2-HIS, introduces the isoflavonoid biosynthetic pathway in legumes (Wang, 2011). Compared with many copies of *SfCHR* and *SfCHS*, only two *IFS* candidate genes, *SfIFS1* and *SfIFS2*, were identified from the same root RNA-seq

library. The blast analysis in NCBI database showed that *SfIFS1* and *SfIFS2* have 80% and 72.6% at nucleotide level with well-studied *GmIFS1* and *GmIFS2*, respectively. We further identified *IFS* homologs (*IFSs*) from genomes of *P. vulgaris*, *V. angularis*, *G. max*, *M. truncatula*, *C. arietinum* and *L. japonicus* (Table 1). A *IFS* phylogenetic tree was then constructed and showed that most homologous of legumes, named type I, were grouped together, forming a main phylogenetic clade (Fig. 4). Meanwhile, five *IFS* candidate genes including *SfIFS2* and four *IFSs* in *L. angustifolius* and *L. luteus*, named type II, were distantly grouped in an early-branching clade. To further figure out the evolutionary relatedness of two types of *IFSs*, all encoded *IFS* protein sequences were aligned and showed a high degree of similarity (Fig. 5), ranging from 70.0% to 100.0%, indicating that the evolution route of *IFSs* in legume was distinctly conserved. This observation was also supported by motif analysis that all studied *IFSs* shared conserved “T” helix, FSAGTDST, proposed to be involved in oxygen binding by P450 proteins. In addition, the heme-iron ligand signature, FGSGRRMCPG, were also exhibited at the C-terminal sequences of *IFSs* (Steele et al., 1999). We also found two obvious sequences distinctions among these proteins; one particular region was that two high conserved residues in type II of *IFSs* (20 P and 55 L) was lost in corresponding region of type I; another one showed high variation of amino acid sequences from position 423 to 435 of type I but highly conserved in type II. Further studies will be needed to investigate the functional difference of two types of *IFSs*.

Evolutional analysis and characterization of protein sequences between IFR and IFR-like (IRL) in legumes

The obtained cDNA clone *SfIFR* and *SfIRL* were 1539 and 1693 nucleotides in length, encoding proteins of 309 and 320 amino acids, respectively. The analysis of conserved protein domains (<http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>) showed that *SfIFR* contained an N-terminus characteristic of NAD(P)-binding domain and a small C-terminal domain presumed to be involved in substrate binding. On the other hand, *SfIRL* belonged to NmrA-like family and displayed a Rossmann-fold NAD(P)H/NAD(P)⁺ binding (NADB) domain, which was consistent with the observed protein structure of IFR in *M. sativa* (Wang et al., 2006).

To gain insight into the phylogenetic diversity of *IFR* genes (*IFRs*) in legumes and their related genes in nonlegumes, database searches were firstly conducted. A considerable number of genes sharing high nucleotide sequence identity with *SfIFR* and *SfIRL* were identified, many of which were annotated as encoding isoflavone reductase-like proteins. ML

phylogenetic tree indicated that *SfIFR* grouped with most leguminous *IFRs* together, which constituted a monophyletic legume-group (Fig. 6). we further observed *PCBER-like* genes from *G. max*, *M. truncatula* and *Trifolium pretense* were grouped together, forming a closed sister clade with the legume-specific clade. Interestingly, *SfIRL*, *IRLs* in *L. angustifolius* and *T. pretense* were evolved from same early-branch with *IRLs* in nonlegumes such as *Gossypium raimondii*. In order to investigate potential divergence between these four unusual *IRLs* in legumes and *IFRs* in legume-specific clade, amino acid sequences alignment of representative *IFRs*, *PCBERs* and *IRLs* were performed. The secondary structure elements predicted in the *IFR* structure were shown in Fig. 7. We observed a conserved G17-X-X-G20-X-X-G23 fingerprint region presented in the N terminus (loop $\beta 1-\alpha 1$) of all *IFRs* and related proteins. The largest difference between *IFRs* in Legume-group and *IFRs* in nonlegume-group was observed in the region between strand $\beta 2$ and helix $\alpha 2$. Compared with *SfIFR* and other leguminous homologs, *SfIFL*, *LaIRL1* and *LaIRL2* have ten less amino acid residues deletion in this region as well as in *IRLs* in nonlegumes (Fig. 7).

Discussion

We investigated the evolutionary history of key enzymes involved in the legume-specific isoflavonoid biosynthesis in *Sophora flavescens* and other representative species that located in divergent phylogenetic lineages of Papilionoideae, a subfamily of Fabaceae. Phylogenetic analysis revealed that *CHSs* and *CHRs* composed multigene subfamilies and formed divergent gene subclades in many cases. *IFSs*, only found in legumes, were distinctly conserved in the evolutionary history. We also found that some *CHS* and *IFR* paralogous copies in legume existed as non-leguminous type. Based on these results, we propose that gene duplication events and neofunctionalization of some key enzymes in the isoflavonoid biosynthesis is a major driving force for the evolutionary gain of a legume-specific function such as inducer of Rhizobia nodulation. This hypothesis can be convinced by the observation that two rounds of duplication in the MtLYK3-LjNFR1a clade are ancestral to evolutionary gain of a legume-specific nod factor receptor (De Mita et al., 2014).

The phylogeny of chalcone synthase gene (*CHS*) in *Rosids* has been studied (Shimada *et al.*, 2007b; Zavala and Opazo, 2015). However, none of these studies was specifically focused on the evolutionary origin of legume-type of *CHS*. Instead, they only considered multiple duplication processes occurred in the legume-specific clade of *CHSs*. In *L. japonicas*, *CHS1*

was phylogenetically distant with other *CHS* homologs in legumes (Shimada et al., 2007b). Here, we found the same case also existed in *S. flavescens*, *G. max*, and *M. truncatula* (Fig. 2). These finding suggested that legume-specific *CHS*s might be derived from the ancestral *CHS* genes in nonlegume-clade. Additional evidence supporting this hypothesis has emerged from the study about two types of chalcone isomerase (*CHI*) genes, involved in the biosynthesis of general flavonoids and legume-specific 5-deoxy(iso)flavonoids in *L. japonicus* (Shimada et al., 2003). Therefore, it is likely that type I *CHS* in nonlegumes may only function in the production of 6-deoxychalcone, whereas legume-specific type II *CHS* may involve production of both 6-deoxychalcone and 6-hydroxychalcone.

CHRs, belong to the aldo-ketoreductase (*AKR*) superfamily, play key roles in the biosynthesis of deoxychalcone-derived isoflavonoids such as soybean phytoalexins. *CHRs* belonging to three subclades in legumes were identified in our study, which was inconsistent with results that showed that only two subclades were observed (Sepiol et al., 2017). It was supposed that two independent duplication events occurred in the evolution process of *CHRs* in Papilionoideae (Fig. 3). The ancestral *CHR* in clade III was possibly duplicated once before the speciation of clade I and II, and the subsequent duplication event caused the paralogous genes in the two subclades. In our study, some *CHR-like* genes in *G. max* and *S. rostrata* were re-identified as *COR-like* and *MecgoR-like* genes, showing that *CHRs*, *CORs* and *MecgoRs* were very close to each other. Further study is required to illustrate the biochemistry and functional divergence of *CHRs*, *CORs* and *MecgoRs*, which will provided more evidences for gene classification.

Isoflavone synthase (*IFS*) is the key enzyme for biosynthesis of isoflavones in legumes. By introducing soybean *IFS1* into *A. thaliana* which had no ability to synthesize isoflavone, genistein accumulated at a different level (Liu et al. 2002 and 2000). Based on our results, the duplication event of *IFS* genes in legume was conserved. Exceptionally, we also found that three and two *IFS* homologs identified in *L. angustifolius* and *S. flavescens* were classified into type I and II groups, which placed in distant phylogenetic clades. P450s and NADPH-P450 reductases are dissociated and independently anchored on the outer face of the endoplasmic reticulum (ER) by amino-terminal hydrophobic anchors (Werck-Reichhart and Feyereisen, 2000). At the N-terminal sequence, targeting sequence such as LLELAIGLVVLALFLHLR (*GmIFS1*) was supposed to be required for proper orientation in the ER membrane (Yamazaki et al., 1993). In our study, conservation analysis revealed that amino acid 20P of *IFSs*, near ER region, was highly conserved in Type *IFS-I* while was lost in

type *IFS-II*. Further investigation is required to shed more light on evolutionary mechanism and enzymatic activity of two IFS versions in legumes.

The most intriguing and unique feature of *IFR* homologs was that some *IFRs* present in *S. flavescens*, *L. angustifolius* and *T. pretense* had close relationship with *IRLs* in nonlegumes. *SfIRL*, *TpIRL*, *LaIRL1* and *LaIRL2* were classified into nonlegume group based on phylogenetic analysis as well as amino acid alignment. We also observed the missing residues at the 5' region of *IFRs*, *PCBERs* and *IRLs* in all studied plants. This observation was consistent with the result that *IFRs* were comparable with *PCBERs* and were deduced to have the same overall basic structure, a continuous NADPH-binding domain and a smaller substrate-binding domain (Min *et al.*, 2003). The highly structure similarity suggested that *SfIRL* and *LaIRLs* may play similar function with *IRLs* in nonlegumes. However, the function information of these *IRL* proteins present in non-legume plants are not available. Because both *IFR* homologs of *L. angustifolius* were *IRL* type and we did not find any legume-specific homologs in *L. angustifolius* from available genome and RNA-seq data. This raises the question of whether this key step of isoflavonoid biosynthetic pathway in *L. angustifolius* is different from that in other legume species. However, our results can not rule out the functional possibility of isoenzymes. Further studies about biochemical function difference of two types *IFR* homologs in *S. flavescens* are needed. Overall, our results suggest that evolution of legume specific *IFRs* and homologs in nonlegumes may be forced from common ancestral gene repertoire. Considering the fact that the dominated role of isoflavonoids in the process of symbiosis between rhizobia and legumes, more efforts are needed to investigate isoflavone components secreted by *L. angustifolius* and other plants when interaction with rhizobium or pathogens.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (31770039). Y. J. and Y. Z. acknowledge support from China Scholarship Council. The authors declare that they have no conflict of interest.

Author Contribution

Y. J. designed the research, and collected and analysed the data. Y. Z. assisted in developing the research design. Y. J wrote the manuscript and W. C. edited the manuscript.

References

- Amore R, Kotter P, Kuster C, Ciriacy M, Hollenberg CP.** 1991. Cloning and Expression in *Saccharomyces-Cerevisiae* of the Nad(P)H-Dependent Xylose Reductase-Encoding Gene (Xyl1) from the Xylose-Assimilating Yeast *Pichia-Stipitis*. *Gene* **109**, 89-97.
- Andrews M, Andrews ME.** 2017. Specificity in Legume-Rhizobia Symbioses. *Int J Mol Sci* **18**.
- Bak S, Beisson F, Bishop G, Hamberger B, Hofer R, Paquette S, Werck-Reichhart D.** 2011. Cytochromes p450. *Arabidopsis Book* **9**, e0144.
- Ballance GM, Dixon RA.** 1995. Medicago-Sativa Cdnas Encoding Chalcone Reductase. *Plant Physiology* **107**, 1027-1028.
- Bomati EK, Austin MB, Bowman ME, Dixon RA, Noel JP.** 2005. Structural elucidation of chalcone reductase and implications for deoxychalcone biosynthesis. *Journal of Biological Chemistry* **280**, 30496-30503.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.** 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421.
- Cheng Q, Li N, Dong L, Zhang D, Fan S, Jiang L, Wang X, Xu P, Zhang S.** 2015. Overexpression of Soybean Isoflavone Reductase (GmIFR) Enhances Resistance to *Phytophthora sojae* in Soybean. *Front Plant Sci* **6**, 1024.
- Daniel S, Tiemann K, Wittkamp U, Bless W, Hinderer W, Barz W.** 1990. Elicitor-Induced Metabolic Changes in Cell-Cultures of Chickpea (*Cicer-Arietinum* L) Cultivars Resistant and Susceptible to *Ascochyta-Rabiei* .1. Investigations of Enzyme-Activities Involved in Isoflavone and Pterocarpan Phytoalexin Biosynthesis. *Planta* **182**, 270-278.
- Dastmalchi M, Bernards MA, Dhaubhadel S.** 2016. Twin anchors of the soybean isoflavonoid metabolon: evidence for tethering of the complex to the endoplasmic reticulum by IFS and C4H. *Plant J* **85**, 689-706.
- De Mita S, Streng A, Bisseling T, Geurts R.** 2014. Evolution of a symbiotic receptor through gene duplications in the legume-rhizobium mutualism. *New Phytologist* **201**, 961-972.
- Den Herder J, Vanhee C, De Rycke R, Corich V, Holsters M, Goormachtig S.** 2007. Nod factor perception during infection thread growth fine-tunes nodulation. *Mol Plant Microbe Interact* **20**, 129-137.
- Ferrer JL, Austin MB, Stewart C, Jr., Noel JP.** 2008. Structure and function of enzymes involved in the biosynthesis of phenylpropanoids. *Plant Physiol Biochem* **46**, 356-370.
- Gang DR, Kasahara H, Xia ZQ, Vander Mijnsbrugge K, Bauw G, Boerjan W, Van Montagu M, Davin LB, Lewis NG.** 1999. Evolution of plant defense mechanisms - Relationships of phenylcoumaran benzylic ether reductases to pinoresinol-lariciresinol and isoflavone reductases. *Journal of Biological Chemistry* **274**, 7516-7527.
- Goormachtig S, Lievens S, Herman S, Van Montagu M, Holsters M.** 1999. Chalcone reductase-homologous transcripts accumulate during development of stem-borne nodules on the tropical legume *Sesbania rostrata*. *Planta* **209**, 45-52.
- Guo L, Dixon RA, Paiva NL.** 1994. Conversion of vestitone to medicarpin in alfalfa (*Medicago sativa* L.) is catalyzed by two independent enzymes. Identification, purification, and characterization of vestitone reductase and 7,2'-dihydroxy-4'-methoxyisoflavanol dehydratase. *J Biol Chem* **269**, 22372-22378.
- Haas BJ, Volfovsky N, Town CD, Troukhan M, Alexandrov N, Feldmann KA, Flavell RB, White O, Salzberg SL.** 2002. Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biol* **3**, RESEARCH0029.
- Hane JK, Ming Y, Kamphuis LG, Nelson MN, Garg G, Atkins CA, Bayer PE, Bravo A, Bringans S, Cannon S, Edwards D, Foley R, Gao LL, Harrison MJ, Huang W, Hurgobin B, Li S, Liu CW, McGrath A, Morahan G, Murray J, Weller J, Jian J, Singh KB.** 2017. A comprehensive draft genome sequence for lupin (*Lupinus angustifolius*), an emerging health food: insights into plant-microbe interactions and legume evolution. *Plant Biotechnology J* **15**, 318-330.
- Hayashi H, Fujii I, Iinuma M, Shibano M, Ozaki K, Watanabe H.** 2013. Characterization of a Glycyrrhizin-Deficient Strain of *Glycyrrhiza uralensis*. *Biological & Pharmaceutical Bulletin* **36**, 1448-1453.
- He XZ, Blount JW, Ge SJ, Tang YH, Dixon RA.** 2011. A genomic approach to isoflavone biosynthesis in kudzu (*Pueraria lobata*). *Planta* **233**, 843-855.

- Hibi N, Higashiguchi S, Hashimoto T, Yamada Y. 1994. Gene-Expression in Tobacco Low-Nicotine Mutants. *Plant Cell* **6**, 723-735.
- Iida K, Fukami-Kobayashi K, Toyoda A, Sakaki Y, Kobayashi M, Seki M, Shinozaki K. 2009. Analysis of Multiple Occurrences of Alternative Splicing Events in Arabidopsis thaliana Using Novel Sequenced Full-Length cDNAs. *DNA Research* **16**, 155-164.
- Jez JM, Bennett MJ, Schlegel BP, Lewis M, Penning TM. 1997. Comparative anatomy of the aldo-keto reductase superfamily. *Biochemical Journal* **326**, 625-636.
- Jiao YS, Liu YH, Yan H, Wang ET, Tian CF, Chen WX, Guo BL, Chen WF. 2015. Rhizobial Diversity and Nodulation Characteristics of the Extremely Promiscuous Legume *Sophora flavescens*. *Mol Plant Microbe Interact* **28**, 1338-1352.
- Jung W, Yu O, Lau SM, O'Keefe DP, Odell J, Fader G, McGonigle B. 2000. Identification and expression of isoflavone synthase, the key enzyme for biosynthesis of isoflavones in legumes. *Nat Biotechnol* **18**, 208-212.
- Kim ST, Cho KS, Kim SG, Kang SY, Kang KY. 2003. A rice isoflavone reductase-like gene, OsIRL, is induced by rice blast fungal elicitor. *Molecules and Cells* **16**, 224-231.
- Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution* **33**, 1870-1874.
- Liu CW, Murray JD. 2016. The Role of Flavonoids in Nodulation Host-Range Specificity: An Update. *Plants (Basel)* **5**.
- Min T, Kasahara H, Bedgar DL, Youn B, Lawrence PK, Gang DR, Halls SC, Park H, Hilsenbeck JL, Davin LB, Lewis NG, Kang C. 2003. Crystal structures of pinoresinol-lariciresinol and phenylcoumaran benzylic ether reductases and their relationship to isoflavone reductases. *J Biol Chem* **278**, 50714-50723.
- Paiva NL, Edwards R, Sun YJ, Hrazdina G, Dixon RA. 1991. Stress responses in alfalfa (*Medicago sativa* L.) 11. Molecular cloning and expression of alfalfa isoflavone reductase, a key enzyme of isoflavonoid phytoalexin biosynthesis. *Plant Mol Biol* **17**, 653-667.
- Paiva NL, Sun Y, Dixon RA, VanEtten HD, Hrazdina G. 1994. Molecular cloning of isoflavone reductase from pea (*Pisum sativum* L.): evidence for a 3R-isoflavanone intermediate in (+)-pisatin biosynthesis. *Arch Biochem Biophys* **312**, 501-510.
- Perret X, Staehelin C, Broughton WJ. 2000. Molecular basis of symbiotic promiscuity. *Microbiol Mol Biol Rev* **64**, 180-201.
- Petrucchio S, Bolchi A, Foroni C, Percudani R, Rossi GL, Ottonello S. 1996. A maize gene encoding an NADPH binding enzyme highly homologous to isoflavone reductases is activated in response to sulfur starvation. *Plant Cell* **8**, 69-80.
- Ralph SG, Chun HJE, Cooper D, Kirkpatrick R, Kolosova N, Gunter L, Tuskan GA, Douglas CJ, Holt RA, Jones SJM, Marra MA, Bohlmann J. 2008. Analysis of 4,664 high-quality sequence-finished poplar full-length cDNA clones and their utility for the discovery of genes responding to insect feeding. *Bmc Genomics* **9**.
- Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y, Wu JZ, Niimura Y, Cheng ZK, Nagamura Y, Antonio BA, Kanamori H, Hosokawa S, Masukawa M, Arikawa K, Chiden Y, Hayashi M, Okamoto M, Ando T, Aoki H, Arita K, Hamada M, Harada C, Hijishita S, Honda M, Ichikawa Y, Idonuma A, Iijima M, Ikeno M, Ito S, Ito T, Ito Y, Iwabuchi A, Kamiya K, Karasawa W, Katagiri S, Kikuta A, Kobayashi N, Kono I, Machita K, Maehara T, Mizuno H, Mizubayashi T, Mukai Y, Nagasaki H, Nakashima M, Nakama Y, Nakamichi Y, Nakamura M, Namiki N, Negishi M, Ohta I, Ono N, Saji S, Sakai K, Shibata M, Shimokawa T, Shomura A, Song JY, Takazaki Y, Terasawa K, Tsuji K, Waki K, Yamagata H, Yamane H, Yoshiki S, Yoshihara R, Yukawa K, Zhong HS, Iwama H, Endo T, Ito H, Hahn JH, Kim HI, Eun MY, Yano M, Jiang JM, Gojohori T. 2002. The genome sequence and structure of rice chromosome 1. *Nature* **420**, 312-316.
- Sepiol CJ, Yu JJ, Dhaubhadel S. 2017. Genome-Wide Identification of Chalcone Reductase Gene Family in Soybean: Insight into Root-Specific GmCHRs and Phytophthora sojae Resistance. *Frontiers in Plant Science* **8**.
- Shimada N, Akashi T, Aoki T, Ayabe S. 2000. Induction of isoflavonoid pathway in the model legume *Lotus japonicus*: molecular characterization of enzymes involved in phytoalexin biosynthesis. *Plant Sci* **160**, 37-47.
- Shimada N, Aoki T, Sato S, Nakamura Y, Tabata S, Ayabe S. 2003. A cluster of genes encodes the two types of chalcone isomerase involved in the biosynthesis of general flavonoids and legume-specific 5-deoxy(iso)flavonoids in *Lotus japonicus*. *Plant Physiology* **131**, 941-951.
- Shimada N, Nakatsuka T, Nishihara M, Yamamura S, Shin-ichi A, Aoki T. 2006. Isolation and characterization of a cDNA encoding polyketide reductase in *Lotus japonicus*. *Plant tissue culture letters* **23**, 509-513.
- Shimada N, Sato S, Akashi T, Nakamura Y, Tabata S, Ayabe S, Aoki T. 2007a. Genome-wide analyses of the structural gene families involved in the legume-specific 5-deoxyisoflavonoid biosynthesis of *Lotus japonicus*. *DNA Res* **14**, 25-36.
- Shimada N, Sato S, Akashi T, Nakamura Y, Tabata S, Ayabe SI, Aoki T. 2007b. Genome-wide analyses of the structural gene families involved in the legume-specific 5-deoxyisoflavonoid biosynthesis of *Lotus japonicus*. *DNA Research* **14**, 25-36.
- Shoji T, Wenz R, Iwase T, Nakajima K, Yamada Y, Hashimoto T. 2002. Expression patterns of two tobacco isoflavone reductase-like genes and their possible roles in secondary metabolism in tobacco. *Plant Molecular Biology* **50**, 427-440.
- Steele CL, Gijzen M, Qutob D, Dixon RA. 1999. Molecular characterization of the enzyme catalyzing the aryl migration reaction of isoflavonoid biosynthesis in soybean. *Arch Biochem Biophys* **367**, 146-150.
- Sun YJ, Wu QD, Vanetten HD, Hrazdina G. 1991. Stereoisomerism in Plant-Disease Resistance - Induction and Isolation of the 7,2'-Dihydroxy-4',5'-Methylenedioxyisoflavone Oxidoreductase, an Enzyme Introducing Chirality during Synthesis of Isoflavonoid Phytoalexins in Pea (*Pisum-Sativum* L). *Archives of Biochemistry and Biophysics* **284**, 167-173.
- Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* **10**, 512-526.

Unterlinner B, Lenz R, Kutchan TM. 1999a. Molecular cloning and functional expression of codeinone reductase: the penultimate enzyme in morphine biosynthesis in the opium poppy *Papaver somniferum*. *Plant J* **18**, 465-475.

Unterlinner B, Lenz R, Kutchan TM. 1999b. Molecular cloning and functional expression of codeinone reductase: the penultimate enzyme in morphine biosynthesis in the opium poppy *Papaver somniferum*. *Plant Journal* **18**, 465-475.

Wang X. 2011. Structure, function, and engineering of enzymes in isoflavonoid biosynthesis. *Funct Integr Genomics* **11**, 13-22.

Wang XQ, He XZ, Lin JQ, Shao H, Chang ZZ, Dixon RA. 2006. Crystal structure of isoflavone reductase from alfalfa (*Medicago sativa* L.). *Journal of Molecular Biology* **358**, 1341-1352.

Welle R, Schroder G, Schiltz E, Grisebach H, Schroder J. 1991. Induced Plant-Responses to Pathogen Attack - Analysis and Heterologous Expression of the Key Enzyme in the Biosynthesis of Phytoalexins in Soybean (*Glycine-Max* L. Merr Cv Harosoy-63). *European Journal of Biochemistry* **196**, 423-430.

Werck-Reichhart D, Feyereisen R. 2000. Cytochromes P450: a success story. *Genome Biol* **1**, REVIEWS3003.

Yamazaki S, Sato K, Suhara K, Sakaguchi M, Mihara K, Omura T. 1993. Importance of the proline-rich region following signal-anchor sequence in the formation of correct conformation of microsomal cytochrome P-450s. *J Biochem* **114**, 652-657.

Zavala K, Opazo JC. 2015. Lineage-Specific Expansion of the Chalcone Synthase Gene Family in Rosids. *Plos One* **10**.

520

521

522

523

524 Table 1. Summary of four key genes involved in (iso)flavonoid biosynthesis in legumes and related nonlegumes.

Species	Gene name	Accession	Location	Position/strand	Reference	Clade
Chalcone synthase (CHS)						
<i>Sophora flavescens</i>	<i>CHS1</i>	MH402000	c32443 g1 [*]	NK	this study	Type II
	<i>CHS2</i>	NK	c34542 g1 [*]	NK	this study	Type II
	<i>CHS3</i>	MH402001	c51007 g1 [*]	NK	this study	Type II
	<i>CHS4</i>	MH402002	c52952 g2 [*]	NK	this study	Type I
	<i>CHS5</i>	MH402003	c63917 g4 [*]	NK	this study	Type II
<i>Medicago truncatula</i>	<i>CHS1</i>	MTR_1g097910	chr01	44132370..44133661/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS2</i>	MTR_1g098140	chr01	44128165..44129456/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS3</i>	MTR_1g098150	chr01	44140792..44142083/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS4</i>	MTR_2g058470	chr02	24163030..24165622/-	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS5</i>	MTR_3g083910	chr03	37858604..37859908/-	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS6</i>	MTR_3g083920	chr03	37862630..37863898/-	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS7</i>	MTR_3g086260	chr03	39029509..39030953/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS8</i>	MTR_5g007720	chr05	1477586..1478881/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS9</i>	MTR_5g007730	chr05	1491587..1493050/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS10</i>	MTR_5g007740	chr05	1499142..1500122/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS11</i>	MTR_5g007760	chr05	1504351..1505859/+	(Zavala and Opazo, 2015), this study	Type II

<i>Lotus japonicus</i>	<i>CHS12</i>	MTR_5g007770	chr05	1509044..1509941/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS13</i>	MTR_7g016700	chr07	5283860..5285190/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS14</i>	MTR_7g016720	chr07	5288937..5290267/-	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS15</i>	MTR_7g016780	chr07	5302048..5303378/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS16</i>	MTR_7g016800	chr07	5309685..5311279/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS17</i>	MTR_7g016820	chr07	5314502..5315993/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS18</i>	MTR_7g084300	chr07	32530242..32532033/-	(Zavala and Opazo, 2015), this study	Type I
	<i>CHS1</i>	LJ1G026890	chr01	25659565..25661173/-	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS2</i>	LJ1G026900	chr01	25665922..25666044 /-	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS3</i>	LJ1G041030	chr01	39252459..39253152 /+	(Zavala and Opazo, 2015), this study	Type I
	<i>CHS4</i>	LJ2G029320	chr02	26785757..26786912 /+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS5</i>	LJ2G029340	chr02	26791950..26794294/-	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS6</i>	LJ2G029350	chr02	26797268..26803064/-	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS7</i>	LJ2G029740	chr02	27058464..27058923/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS8</i>	LJ2G029780	chr02	27076189..27076779/-	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS9</i>	LJ2G029790	chr02	27078451..27078875/+	(Zavala and Opazo, 2015), this study	Type II

<i>Glycine max</i>	<i>CHS10</i>	LJ2G029890	chr02	27145235..27146012/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS11</i>	LJ3G007130	chr03	5935036..5935996/-	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS12</i>	LJ3G007200	chr03	5958751..5959497/-	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS13</i>	LJ4G027870	chr04	24212737..24216525/-	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS1</i>	Glyma01G22880	chr01	29150207..29151975/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS2</i>	Glyma01G43880	chr01	54742783..54744397/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS3</i>	MH402004	chr02	13039519-13041358/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS4</i>	Glyma05G28610	chr05	34421004-34422307/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS5</i>	Glyma08G11520	chr08	8380515..8381802/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS6</i>	Glyma08G11530	chr08	8384488..8385776/-	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS7</i>	Glyma08G11610	chr08	8462794..8464099/-	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS8</i>	Glyma08G1162	chr08	8468623..8469911/-	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS9</i>	Glyma08G11630	chr08	8472178..8473466/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS10</i>	Glyma08G11635	chr08	8476151..8477438/-	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS11</i>	Glyma08G11650	chr08	8485475..8486261/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS12</i>	Glyma09G08780	chr09	8193759..8195572/-	(Zavala and Opazo, 2015), this study	Type II

	<i>CHS13</i>	Glyma11G01350	chr11	793736..795376/+	(Zavala and Opazo, 2015), this study	Type II
	<i>CHS14</i>	Glyma19G27930	chr19	35280638..35283188/-	(Zavala and Opazo, 2015), this study	Type I
<i>Theobroma cacao</i>	<i>CHS1</i>	TC04G021630	Scaffold04	28274678..28277325/-	(Zavala and Opazo, 2015), this study	N-G
	<i>CHS2</i>	TC00G003270	Scaffold08	8591259..8592562/+	(Zavala and Opazo, 2015), this study	N-G
<i>Ricinus communis</i>	<i>CHS</i>	RC29904G00940	Sca29904	624954..626268/-	(Zavala and Opazo, 2015), this study	N-G
<i>Fragaria vesca</i>	<i>CHS1</i>	FV2G30500	chr02	20621126..20622366/+	(Zavala and Opazo, 2015), this study	N-G
	<i>CHS2</i>	FV2G30510	chr02	20623449..20624718/+	(Zavala and Opazo, 2015), this study	N-G
	<i>CHS3</i>	FV7G02590	chr07	1803204..1804580/-	(Zavala and Opazo, 2015), this study	N-G
	<i>CHS4</i>	FV7G02600	chr07	1807867..1809208/-	(Zavala and Opazo, 2015), this study	N-G
<i>Malus domestica</i>	<i>CHS1</i>	MD09G015370	chr09	15921955..15924143/-	(Zavala and Opazo, 2015), this study	N-G
	<i>CHS2</i>	MD09G015400	chr09	15935641..15936986/+	(Zavala and Opazo, 2015), this study	N-G
	<i>CHS3</i>	MD00G182230	Scaffold 182230	2813..4763/+	(Zavala and Opazo, 2015), this study	N-G
	<i>CHS4</i>	MD00G258370	Scaffold 258370	10019..11824/-	(Zavala and Opazo, 2015), this study	N-G
	<i>CHS5</i>	MD00G276470	Scaffold 276470	2287..4677/+	(Zavala and Opazo, 2015), this study	N-G
<i>Arabidopsis lyrata</i>	<i>CHS</i>	LOC9309706	Scaffold 6	5616928..5618205/+	(Zavala and Opazo, 2015), this study	N-G
<i>Arabidopsis thaliana</i>	<i>CHS</i>	AT5G13930	chr05	4488762..4490035/+	(Zavala and Opazo, 2015), this study	N-G

<i>Arabidopsis thaliana</i>	<i>LAP5</i>	NM_1196513	chr04	16608349..16609720/+	(Zavala and Opazo, 2015), this study	O-G
<i>Fragaria vesca</i>	<i>LAP5</i>	XM_004289785.2	chr02	5091502..5092908/+	(Zavala and Opazo, 2015), this study	O-G
Chalcone Reductase (CHR)						
<i>Sophora flavescens</i>	<i>CHR1</i>	MH450274	c30947 g1 [*]	NK	This study	C-I
	<i>CHR2</i>	MH450275	c60795 g1 [*]	NK	This study	C-II
	<i>CHR3</i>	MH450276	c63937 g2 [*]	NK	This study	C-III
<i>Medicago truncatula</i>	<i>CHR1</i>	MH450283	chr03	37503715..37505830/-	This study	C-II
	<i>CHR2</i>	MH450285	chr05	42874291..42875621/-	This study	C-I
	<i>CHR3</i>	MH450284	chr05	42883433..42884926/-	This study	C-I
	<i>CHR4</i>	MH450286	chr05	42889718..42890852/+	This study	C-I
	<i>CHR5</i>	MH450282	scaffold0374	12824..15980/+	This study	C-III
<i>Lotus japonicus</i>	<i>CHR1</i>	MH450279	chr00	467239..468833/-	This study	C-II
	<i>CHR2</i>	MH450280	chr00	474402..476105/+	This study	C-II
	<i>CHR3</i>	MH450277	chr00	171587274..171589109 /-	This study	C-I
	<i>CHR4</i>	MH450278	chr00	171594238..171595443 /+	This study	C-I
	<i>CHR5</i>	AB263016	chr02	42631044..42632371/-	(Shimada <i>et al.</i> , 2006)	C-I
<i>Glycine max</i>	<i>CHR6</i>	MH450281	chr02	42687381..42688649/-	This study	C-I
	<i>CHR1</i>	Glyma.02G307300	chr02	48163443..48164865/-	This study	C-I
	<i>CHR2</i>	Glyma.14G005700	chr14	463009..64543/+	This study	C-I
	<i>CHR3</i>	Glyma.15G189300	chr15	20075936..20082802/-	This study	C-III
	<i>CHR4</i>	Glyma.18G285800	chr18	56611460..56612978/-	This study	C-II
	<i>CHR5</i>	Glyma.20G031100	chr20	3790362..3793599/+	This study	C-III
	<i>COR1</i>	Glyma.09G229500	chr09	45341437..45343797/+	(Sepiol <i>et al.</i> , 2017), this study	O-G
	<i>COR2</i>	Glyma.12G006800	chr12	508878..511003/-	(Sepiol <i>et al.</i> , 2017), this study	O-G
	<i>MECGOR1</i>	Glyma.09G169700	chr09	39437065..39439064/-	(Sepiol <i>et al.</i> , 2017), this study	O-G

	(CHR9A)					
	MECGOR2	Glyma.16G219500	chr09	37672779..37675776/-	(Sepiol <i>et al.</i> , 2017), this study	O-G
	(CHR9C)					
	MECGOR3	Glyma.16G219400	chr16	37677371..37670638/-	(Sepiol <i>et al.</i> , 2017), this study	O-G
	(CHR16A)					
	MECGOR4	Glyma.16G219500	chr16	39453516..39456085/-	(Sepiol <i>et al.</i> , 2017), this study	O-G
	(CHR16B)					
<i>Lupinus</i>	CHR	XM_019573161.1	NK	NK	This study	C-III
<i>angustifolius</i>	MecgoR	XM 019577645.1	NK	NK	This study	O-G
<i>Papaver</i>	COR1	AF108435	NK	NK	(Unterlinner <i>et al.</i> , 1999a)	O-G
<i>somniferum</i>						
<i>Arabidopsis</i>	COR1	AT1G59960	chr01	22071217..22073267/-	(Iida <i>et al.</i> , 2009)	O-G
<i>thaliana</i>	COR2	AT1G59950	chr01	22068033 – 22070588/-	(Unterlinner <i>et al.</i> , 1999b; Welle <i>et al.</i> , 1991)	O-G
<i>Sesbania rostrata</i>	MECGOR	AJ223291	NK	NK	(Goormachtig <i>et al.</i> , 1999)	O-G
	(SrCHR)					
<i>Cicer arietinum</i>	MecgoR	XM 004513280.2	NK	NK	This study	O-G
<i>Pichia stipitis</i>	XYL1	X59465	NK	NK	(Amore <i>et al.</i> , 1991)	O-G
Isoflavone synthase (IFS)						
<i>Sophora flavescens</i>	IFS1	MH538346	c33110 g1*	NK	This study	Type I
	IFS2	MH538347	c61236 g1*	NK	This study	Type II
<i>Medicago</i>	IFS1	MH450243	chr04	34779643..34782159/+	This study	Type I
<i>truncatula</i>	IFS2	MH450242	chr04	34761467..34764335/-	This study	Type I
	IFS3	MH450241	chr04	34754100..34756270/-	This study	Type I
<i>Lotus japonicus</i>	IFS1	AB024931	chr04	7949869..7951739	(Shimada <i>et al.</i> , 2000)	Type I
				7949536..7951837		
	IFS2	AB279984	chr04	8197652..8199484	(Shimada <i>et al.</i> , 2007a)	Type I
	IFS3	AP009071	chr04	8165351..8167491	(Shimada <i>et al.</i> , 2007a)	Type I
	(pseudogene)			8165668..8167322		
	)					

<i>Glycine max</i>	<i>IFS1</i>	MH450245	chr07	37166660..37169027/+	This study	Type I
	<i>IFS2</i>	MH450244	chr13	28761503..28763755/+	This study	Type I
<i>Phaseolus vulgaris</i>	<i>IFS1</i>	MH450246	chr03	6546368..6548523/+	This study	Type I
	<i>IFS2</i>	MH450248	chr03	11641754..11643653/-	This study	Type I
	<i>IFS3</i>	MH450247	chr03	6570289..6571942/- 6557380..6559219/+	This study	Type I
<i>Cicer arietinum</i>	<i>IFS1</i>	NM_001309700.1	chr06	13444140..13446195	This study	Type I
	<i>IFS2</i>	XM_004505052.2	chr06	13449014..13450873	This study	Type I
<i>Vigna angularis</i>	<i>IFS1</i>	XM_017567230.1	scaffold_349	376090..378297	This study	Type I
	<i>IFS2</i>		scaffold_126	924476..926290	This study	Type I
<i>Lupinus angustifolius</i>	<i>IFS1</i>	MH450250	NLL-06	6732977..6734473	This study	Type I
	<i>IFS2</i>	MH450249	NLL-06	6732977..6734473	This study	Type II
	<i>IFS3</i>	MH450251	NLL-19	16434284..16437521	This study	Type II
<i>Lupinus luteus</i>	<i>IFS1</i>	FJ539089	NK	NK	NCBI database	Type II
	<i>IFS2</i>	FJ539090	NK	NK	NCBI database	Type II
<i>Arabidopsis thaliana</i>	<i>IFS-like</i>	AY08662	chr3	9573360-9575148/-	(Haas <i>et al.</i> , 2002)	O-G
Isoflavone reductase (IFR)						
<i>Sophora flavescens</i>	<i>IFR</i>	MH450252	c48179 g1*	NK	This study	L-G
	<i>IRL</i>	MH450253	c41667 g1*	NK	This study	N-G
<i>Medicago truncatula</i>	<i>IFR1</i>	MH450259	chr05	7953179..7954887/+ 7953059..7955273/+	This study	L-G
	<i>IFR2</i>	MH450260	chr05	7965936..7967985/+	This study	L-G
<i>Lotus japonicus</i>	<i>IFR1</i>	MH450268	chr02	26770357..26772474/-	This study	L-G
	<i>IFR2</i>	MH450269	chr02	26794155..26803442/-	This study	L-G
<i>Cajanus cajan</i>	<i>IFR1</i>	XM_020361274.1	chr06	17506643..17507472	This study	L-G
	<i>IFR2</i>	XM_020362544.1	chr06	17612904..17614932	This study	L-G
	<i>IFR3</i>	XM_020362569.1	chr06	17629943..17631543	This study	L-G
	<i>IFR4</i>	XM_020362340.1	chr06	17642007..17643967	This study	L-G

<i>Glycine max</i>	<i>IFR5</i>	XM_020362448.1	chr06	17650555..17651982	This study	L-G
	<i>IFR1</i>	MH450261	chr01	50980302..50982573/-	This study	L-G
	<i>IFR2</i>	MH450262	chr01	54334961..54337099/-	This study	L-G
	<i>IFR3</i>	MH450263	chr11	5273807..5275911/+	This study	L-G
<i>vigna_unguiculata</i>	<i>IFR4</i>	MH450264	chr11	5278847..5280644/+	This study	L-G
	<i>IFR1</i>	MH450257	chr02	28054265..28056147	This study	L-G
	<i>IFR2</i>	MH450256	chr02	28065073..28066748	This study	L-G
<i>Phaseolus vulgaris</i>	<i>IFR3</i>	MH450255	chr02	28070818..28072964	This study	L-G
	<i>IFR1</i>	MH450265	chr02	3270612..3272213/-	This study	L-G
	<i>IFR2</i>	MH450266	chr02	3278401..3280020/-	This study	L-G
	<i>IFR3</i>	MH450270	chr02	3287178..3288725/-	This study	L-G
<i>Trifolium pratense</i>	<i>IFR4</i>	MH450271	chr02	3309973..3311583/+	This study	L-G
	<i>IFR</i>	MH450267	chr02	12654522..12657225/+	This study	L-G
	<i>PCBER-like</i>	MH450272	chr02	12674476..12676095/+	This study	O-G
<i>Medicago truncatula</i>	<i>IRL</i>	MH450273	chr07	5570822..5572610/+	This study	N-G
	<i>PCBER-like</i>	XM_003612001.3	chr05	7941706..7943809/+	This study	O-G
<i>Glycine max</i>	<i>PCBER1-like</i>	Glyma.11G070200.1	chr11	5258153..5262782/+	This study	O-G
	<i>PCBER2-like</i>	Glyma.01G172900.1	chr01	51020276..51025649/-	This study	O-G
<i>Nicotiana tabacum</i>	<i>PCBER2-like</i>	D28505	NK	NK	(Hibi <i>et al.</i> , 1994)	O-G
<i>Pinus taeda</i>	<i>PCBER</i>	AF242490	NK	NK	(Gang <i>et al.</i> , 1999)	O-G
<i>Lupinus angustifolius</i>	<i>IRL1</i>	MH450258	NK	NK	This study	N-G
	<i>IRL2</i>	MH450254	NK	NK	This study	N-G
<i>Malus domestica</i>	<i>IRL</i>	KX577776	scaffold274	274:3020..4944/-	This study	N-G
<i>Gossypium</i>	<i>IRL</i>	XM_012610902	NK	NK	This study	N-G

<i>raimondii</i>						
<i>Populus</i>	<i>IRL</i>	ABK94513	NK	NK	(Ralph <i>et al.</i> , 2008)	N-G
<i>trichocarpa</i>						
<i>Arabidopsis</i>	<i>IRL</i>	AAG12677	chr01	28251626..28253686/+	This study	N-G
<i>thaliana</i>						
<i>Zea mays</i>	<i>IRL</i>	U33318	NK	NK	(Petrucco <i>et al.</i> , 1996)	N-G
<i>Oryza sativa</i>	<i>IRL</i>	BAB16910	chr01	327128..328450/-	(Sasaki <i>et al.</i> , 2002)	N-G

Location is given as a chromosome or scaffold identifier when available; position and strand are given relative to this chromosome or scaffold. NK, not known; N-G, gene homologs classified into the nonlegume group; L-G, gene homologs classified into the legume group. O-G, genes used as out-group; *IRL*, *IFR-like*; *, genes blasted from our RNA-seq data.

528

529

530

531

532

Figure legends

Fig.1 General and legume-specific (iso)flavonoid pathway in leguminous plants. Similar to CHIs, CHSs also have two types, legume-specific type (type II) and nonlegume type (type I). The schematic of biosynthetic pathway leading to isoflavonoids is shown in pink. The dotted arrow indicates speculative steps while multiple arrows indicate two or more steps in the pathway. CHR, chalcone reductase; CHS, chalcone synthase; CHI, chalcone isomerase; IFS, isoflavone synthase.

Fig. 2 phylogenetic analysis suggested the origin of the legume-specific type II of CHSs. Maximum likelihood phylogenetic tree was constructed based on nucleotide sequences. Most of CHSs in legumes were grouped together, named type II. *CHS4* of *S. flavescens*, *CHS14* of *G. max*, *CHS3* of *L. japonicus* and *CHS18* of *M. truncatula* were classified into the type I group, closed with homologs in nonlegumes. Nude and gray colours indicate genes identified in legume and nonlegume, respectively. Black rhombus indicate CHSs of *S. flavescens*. *LAP5* of *F. vesca* and *A. thaliana* were used as an out-group. Only bootstrap values greater than 50% are shown at branch nodes. Bar, 10% nucleotide substitution per site.

Fig. 3 CHRs in legumes belonging to AKR4 Family were grouped together. The phylogenetic analysis showed that these CHRs in legumes were distributed into three subclades, clade I, II and III, suggesting that CHRs were possibly duplicated twice (indicated by five-pointed star) before the speciation of Papilionoideae in the evolutionary processes. Some CHR homologs in *G. max* and CHR of *S. rostrata* were re-identified based on phylogenetic evolution. Three colours from up to down indicate chalcone reductases, codeinone reductases and methylecgonone reductase of AKR4 subfamily members, respectively. Dotted branch indicates phylogenetic clades except for CHRs. Black rhombus indicate CHRs of *S. flavescens*. *XYL1* of *P. stipitis* was used as an out-group. Only bootstrap values greater than 50% are shown at branch nodes. Bar, 20% nucleotide substitution per site.

Fig. 4 Maximum likelihood phylogenetic tree of IFSs in legume. Most IFS homologs of legumes, named type I (grey dotted box), were grouped together, forming a main phylogenetic clade. In addition, five IFS candidate genes in *S. flavescens*, *L. angustifolius* and *L. luteus*, were distantly grouped in an early-branching clade, named type II (green dotted box). Black rhombus indicate IFSs of *S. flavescens*. IFS-like gene of *A. thaliana* was used as

an out-group. Only bootstrap values greater than 50% are shown at branch nodes. Bar, 5% nucleotide substitution per site.

Fig. 5 Phylogenetic analysis of *IFR* homologs in legumes and nonlegumes. The deduced nucleotide sequences of putative *IFRs* were aligned and the Maximum Likelihood (ML) tree was constructed using MEGA7 software. The orange branch indicates phylogenetic position of most legumious *IFRs*. *PCBER* candidate genes (in purple branch) in legumes may be evolved from the same ancient clade with *IFRs*. *IFR-like* (*IRL*) genes of *L. angustifolius*, *T. pratense* and *S. flavescentis* were clustered together with *IFR* homologs in nonlegumes such as *IFR-like* gene in *G. raimondii* (in blue branch). Black rhombus indicate *IFRs* of *S. flavescentis*. *IFS-like* genes in *A. thaliana*, *Z. mays* and *O. sativa* were used as an out-group. Only bootstrap values greater than 50% are shown at branch nodes. Bar, 10% nucleotide substitution per site.

Fig. 6 Sequences alignment of *IFSs*. Protein Sequences were from type I group (above the dotted line) including *S. flavescentis* (SfIFS1), *G. max* (GmIFS1 and GmIFS2), *V. angularis* (VaIFS1 and VaIFS2), *L. japonicus* (LjIFS1 and LjIFS2), *M. truncatula* (MtIFS1 and MtIFS2) and *L. angustifolius* (LaIFS1); type II group (below the dotted line) including *L. angustifolius* (LaIFS2 and LaIFS3) and *L. luteus* (LIIFS1 and LIIFS2). Identical residues are highlighted. *, two conserved residues lost in type II of *IFSs*. Black dotted box indicates the drastic variation of amino acid sequences in *IFSs*. Red dotted line indicates the heme-iron ligand signature of *IFSs*.

Fig. 7 A notable distinction between *IFRs* and related *IFR-like* homologs in legumes and nonlegumes was observed in the region between strand $\beta 2$ and helix $\alpha 2$ (black dotted box). Sequence alignment of *IFRs* from *S. flavescentis* (SfIFR), *M. truncatula* (MtIFR1 and MtIFR2), *G. max* (GmIFR1, GmIFR2, GmIFR3 and GmIFR4), *P. vulgaris* (PvIFR1, PvIFR2, PvIFR3 and PvIFR4), *T. pratense* (TpIFR), *L. japonicus* (LjIFR1 and LjIFR2); *PCBER-like* proteins in *T. pratense* (TpPCBER-like) and *M. truncatula* (MtPCBER-like); *IFR-like* (*IRL*) proteins from *L. albus* (LaIRL1 and LaIRL2), *S. flavescentis* (SfIRL), *A. thaliana* (AtIRL), *P. trichocarpa* (PtIRL), *M. domestica* (MdIRL), *Z. mays* (ZmIRL) and *O. sativa* (OsIRL). The secondary structure elements observed in the *IFR* structure are shown above the alignment.

0.05

	10	20	30	40	50	60	70	80	90	100	110	1																																																																																																									
SfIFS1	M	L	V	E	V	A	I	S	L	L	V	I	A	L	L	I	H	L	R	P	T	P	T	A	K	S	K	A	L	R	H	L	P	N	P	P	S	P	K	P	R	L	P	F	V	G	H	L	H	L	D	N	P	L	L	H	S	L	I	K	L	S	K	K	Y	G	P	-	L	F	S	L	Y	F	G	S	M	P	T	I	V	A	S	T	P	D	L	F	K	L	F	L	Q	T	H	E	A	S	F	N	T	R	F	Q	T	S	A	I	R	R	L	T	Y	D	N	S	
GmIFS1	M	L	E	L	A	L	G	L	F	V	L	A	L	F	L	H	L	R	P	T	P	S	A	K	S	K	A	L	R	H	L	P	N	P	P	S	P	K	P	R	L	P	F	I	G	H	L	H	L	K	D	K	L	L	H	Y	A	L	I	D	L	S	K	K	H	G	P	-	L	F	S	L	S	F	G	S	M	P	T	V	V	A	S	T	P	E	L	F	K	L	F	L	Q	T	H	E	A	S	F	N	T	R	F	Q	T	S	A	I	R	R	L	T	Y	D	N	S	
GmIFS2	M	L	E	L	A	L	G	L	L	V	L	A	L	F	L	H	L	R	P	T	P	T	A	K	S	K	A	L	R	H	L	P	N	P	P	S	P	K	P	R	L	P	F	I	G	H	L	H	L	K	D	K	L	L	H	Y	A	L	I	D	L	S	K	K	H	G	P	-	L	F	S	L	Y	F	G	S	M	P	T	V	V	A	S	T	P	E	L	F	K	L	F	L	Q	T	H	E	A	S	F	N	T	R	F	Q	T	S	A	I	R	R	L	T	Y	D	S		
VaIFS1	M	L	E	I	T	I	G	L	L	V	L	A	L	F	L	H	L	R	P	T	P	T	A	K	S	K	A	L	R	H	L	P	N	P	P	S	P	K	P	R	L	P	F	I	G	H	L	H	L	K	D	K	L	L	H	Y	A	L	I	D	L	S	K	K	Y	G	P	-	L	Y	S	L	Y	F	G	S	M	P	T	V	V	A	S	S	P	E	L	F	K	L	F	L	Q	T	H	E	A	A	S	F	N	T	R	F	Q	T	S	A	I	R	R	L	T	Y	D	N	S
VaIFS2	M	L	E	I	A	V	G	L	L	V	L	V	L	F	L	H	L	R	P	T	P	T	A	K	S	K	A	L	R	H	L	P	N	P	P	S	P	K	P	R	L	P	F	V	G	H	L	H	L	K	D	Q	L	L	H	S	L	I	T	L	S	Q	R	Y	G	P	-	L	Y	S	L	Y	F	G	S	M	P	T	V	V	A	S	T	P	E	L	F	K	L	F	L	Q	T	H	E	A	A	S	F	N	T	R	F	Q	T	S	A	I	K	R	L	T	Y	D	N	S	
LjIFS1	M	L	V	E	I	A	L	A	L	A	L	A	L	F	L	Q	F	R	P	T	P	T	A	K	S	K	A	L	R	H	L	P	N	P	P	S	P	K	P	R	L	P	F	V	G	H	L	H	L	D	Q	P	L	L	H	N	S	L	I	K	L	G	E	K	Y	G	P	-	L	Y	T	L	Y	F	G	S	M	P	T	V	V	A	S	T	P	E	L	F	K	L	F	L	Q	T	H	E	A	S	F	S	T	R	F	Q	T	S	A	I	R	R	L	T	Y	D	N	S	
LjIFS2	M	L	V	E	L	A	L	A	L	A	I	A	L	F	L	H	L	R	P	T	P	T	A	K	S	K	A	L	R	H	L	P	N	P	P	S	P	K	P	R	L	P	F	V	G	H	L	H	L	D	Q	P	L	L	H	S	L	I	K	L	G	E	R	Y	G	P	-	L	Y	S	L	Y	F	G	S	M	P	T	V	V	A	S	T	P	E	L	F	K	L	F	L	Q	T	H	E	A	S	F	N	T	R	F	Q	T	S	A	I	R	R	L	T	Y	D	N	S		
MtIFS1	M	L	V	E	L	A	V	T	L	L	L	I	A	L	F	L	H	L	R	P	T	P	T	A	K	S	K	A	L	R	H	L	P	N	P	P	S	P	K	P	R	L	P	F	I	G	H	L	H	L	D	N	P	L	L	H	T	L	I	K	L	G	K	R	Y	G	P	-	L	Y	T	L	Y	F	G	S	M	P	T	V	V	A	S	T	P	D	L	F	K	L	F	L	Q	T	H	E	A	S	F	N	T	R	F	Q	T	S	A	I	S	R	L	T	Y	D	N	S	
MtIFS2	M	L	V	E	L	A	V	T	L	L	V	I	A	L	F	I	H	L	R	P	T	P	T	A	K	S	K	A	L	R	H	L	P	N	P	P	S	P	K	P	R	L	P	F	I	G	H	L	H	L	D	N	P	L	L	H	T	L	I	K	L	G	K	R	Y	G	P	-	L	Y	T	L	Y	F	G	S	M	P	T	V	V	A	S	T	P	D	L	F	K	L	F	L	Q	T	H	E	A	S	F	N	T	R	F	Q	T	S	A	I	S	R	L	T	Y	D	N	S	
MtIFS3	M	L	V	E	L	A	V	T	L	L	L	I	A	L	F	L	H	L	R	P	T	P	T	A	K	S	K	A	L	R	H	L	P	N	P	P	S	P	K	P	R	L	P	S	I	G	H	L	H	L	D	N	P	L	L	H	T	L	I	K	L	G	E	R	Y	G	P	-	L	Y	T	L	Y	F	G	S	M	P	T	I	V	A	S	T	P	D	L	F	K	L	F	L	Q	T	H	E	A	S	F	N	T	R	F	Q	T	S	A	I	S	R	L	T	Y	D	N	S	
LaIFS1	M	L	I	E	L	A	L	T	F	L	A	I	A	L	F	I	Y	L	R	P	S	P	S	A	K	S	K	T	L	R	H	L	P	N	P	P	S	P	K	P	R	L	P	F	V	G	H	L	H	L	D	Q	P	L	L	H	Q	S	L	I	K	L	S	E	R	Y	G	P	-	L	Y	S	L	Y	F	G	S	M	P	T	I	V	A	S	T	P	D	L	F	K	L	F	L	Q	T	H	E	A	S	F	N	T	R	F	Q	T	S	A	I	K	R	L	T	Y	N	N	S
LaIFS2	M	L	Q	E	L	I	I	P	F	L	L	L	A	L	Y	I	H	F	H	P	S	-	P	S	K	S	K	H	S	F	P	P	P	S	P	K	L	R	L	P	F	V	G	H	L	H	L	K	L	P	-	L	H	H	S	L	I	S	L	S	K	Q	Y	G	P	-	L	Y	S	L	Y	F	G	S	M	P	T	I	V	S	T	P	Q	L	F	K	L	F	L	Q	T	H	E	A	S	F	N	T	R	F	Q	T	S	A	I	K	R	L	T	Y	G	N	S				
LaIFS3	M	L	E	L	L	I	P	F	L	L	L	V	L	Y	I	G	F	R	P	-	S	P	K	S	N	S	K	R	H	P	P	P	S	P	K	L	R	L	P	F	V	G	H	L	H	L	K	L	P	-	L	H	H	S	L	I	R	L	S	K	Q	Y	G	P	-	L	Y	S	L	Y	F	G	S	M	P	T	I	V	A	S	T	P	E	L	F	K	L	F	L	Q	T	H	E	A	S	F	N	T	R	F	Q	T	S	A	I	K	R	L	T	Y	D	N	S				
LlIFS1	M	L	E	L	L	I	P	F	L	L	L	Y	I	G	F	S	P	-	S	K	S	N	S	K	T	L	H	P	P	P	S	P	K	L	R	L	P	F	V	G	H	L	H	L	K	L	P	-	L	H	H	S	L	I	R	L	S	K	Q	Y	G	P	-	L	Y	S	L	Y	F	G	S	M	P	T	I	V	A	S	T	P	E	L	F	K	L	F	L	Q	T	H	E	A	S	F	N	T	R	F	Q	T	S	A	I	K	R	L	T	Y	D	N	S						
LlIFS2	M	L	E	L	L	I	P	F	L	L	L	Y	I	G	F	S	P	-	S	K	S	N	S	K	T	L	H	P	P	P	S	P	K	L	R	L	P	F	V	G	D	L	H	L	K	L	P	-	L	H	N	S	L	I	R	L	S	K	Q	Y	G	P	-	L	Y	S	L	Y	F	G	S	M	P	T	I	V	A	S	T	P	E	L	F	K	L	F	L	Q	T	H	E	A	S	F	N	T	R	F	Q	T	S	A	I	K	R	L	T	Y	D	N	S						

*

*

	400	410	420	430	440	450	460	470	480	490	500	!																																																																																																								
SfIFS1	P	E	G	A	L	I	L	F	N	V	W	A	V	G	R	D	P	K	Y	W	E	S	P	S	E	F	R	P	E	R	F	I	D	N	S	----	A	D	F	V	D	L	R	G	H	F	Q	L	L	P	F	G	S	G	R	R	M	C	P	G	V	N	L	A	T	A	G	M	A	T	F	L	S	T	L	I	Q	C	F	D	L	Q	P	V	G	P	E	G	Q	I	L	K	G	D	A	K	V	S	M	E	E	R	P	G	L	T	V	P	R	A	H	N				
GmIFS1	P	E	G	A	L	V	L	F	N	V	W	Q	V	G	R	D	P	K	Y	W	D	R	P	S	E	F	R	P	E	R	F	I	E	T	G	----	A	E	G	E	A	G	P	L	D	L	R	G	H	F	Q	L	L	P	F	G	S	G	R	R	M	C	P	G	V	N	L	A	T	S	G	M	A	T	L	L	A	S	I	Q	C	F	D	L	Q	V	L	G	P	Q	Q	I	L	K	G	D	A	K	V	S	M	E	E	R	A	G	L	T	V	P	R	A	H	S		
GmIFS2	P	E	G	A	L	I	L	F	N	V	W	Q	V	G	R	D	P	K	Y	W	D	R	P	S	E	F	R	P	E	R	F	I	E	T	G	----	A	E	G	E	A	G	P	L	D	L	R	G	H	F	Q	L	L	P	F	G	S	G	R	R	M	C	P	G	V	N	L	A	T	S	G	M	A	T	L	L	A	S	I	Q	C	F	D	L	Q	V	L	G	P	Q	Q	I	L	K	G	D	A	K	V	S	M	E	E	R	A	G	L	T	V	P	R	A	H	S		
VaIFS1	P	E	G	A	L	I	L	F	N	V	W	A	V	G	R	D	P	K	Y	W	D	R	P	L	E	F	R	P	E	R	F	I	E	T	G	----	A	E	G	E	A	G	P	L	D	L	R	G	H	F	Q	L	L	P	F	G	S	G	R	R	M	C	P	G	V	N	L	A	T	S	G	M	A	T	L	L	A	S	V	I	Q	C	F	D	L	Q	V	V	G	P	Q	Q	I	L	K	G	D	A	K	V	S	M	E	E	R	G	G	L	T	V	P	R	A	H	N	
VaIFS2	P	E	G	A	L	I	L	F	N	V	W	A	V	G	R	D	P	K	Y	W	D	R	P	S	E	F	R	P	E	R	F	I	E	T	G	----	G	E	G	G	V	G	P	L	D	L	R	G	H	F	Q	L	L	P	F	G	S	G	R	R	M	C	P	G	V	N	L	S	T	S	G	M	A	T	L	L	A	S	V	I	Q	C	F	D	L	Q	V	V	D	P	Q	G	H	I	L	K	G	D	A	K	V	S	M	E	E	R	A	G	L	T	V	P	R	K	H	N
LjIFS1	P	E	G	A	L	V	L	F	N	V	W