Multi-tissue integration of transcriptomic and specialized metabolite profiling provides tools for assessing the common bean (*Phaseolus vulgaris*) metabolome

Leonardo Perez de Souza¹, Federico Scossa¹,², Sebastian Proost¹, Elena Bitocchi³, Roberto Papa³, Takayuki Tohge¹,⁴,* and Alisdair R. Fernie¹,*

¹Max-Planck-Institute of Molecular Plant Physiology, Am Mühlenberg 1, Potsdam-Golm 14476, Germany,
²Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria, CREA-OFA, Via di Fioranello 52, 00134 Rome, Italy,
³Department of Agricultural, Food, and Environmental Sciences, Università Politecnica delle Marche, 60131 Ancona, Italy, and
⁴Graduate School of Biological Sciences, Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan

Received 11 April 2018; revised 15 November 2018; accepted 23 November 2018; published online 27 November 2018.

*For correspondence (e-mails fernie@mpimp-golm.mpg.de and tohge@bs.naist.jp).

SUMMARY

Common bean (*Phaseolus vulgaris* L.) is an important legume species with a rich natural diversity of landraces that originated from the wild forms following multiple independent domestication events. After the publication of its genome, several resources for this relevant crop have been made available. A comprehensive characterization of specialized metabolism in *P. vulgaris*, however, is still lacking. In this study, we used a metabolomics approach based on liquid chromatography-mass spectrometry to dissect the chemical composition at a tissue-specific level in several accessions of common bean belonging to different gene pools. Using a combination of literature search, mass spectral interpretation, ¹³C-labeling, and correlation analyses, we were able to assign chemical classes and/or putative structures for approximately 39% of all measured metabolites. Additionally, we integrated this information with transcriptomics data and phylogenetic inference from multiple legume species to reconstruct the possible metabolic pathways and identify sets of candidate genes involved in the biosynthesis of specialized metabolites. A particular focus was given to flavonoids, triterpenoid saponins and hydroxycinnamates, as they represent metabolites involved in important ecological interactions and they are also associated with several health-promoting benefits when integrated into the human diet. The data are presented here in the form of an accessible resource that we hope will set grounds for further studies on specialized metabolism in legumes.

Keywords: common bean, *Phaseolus vulgaris*, specialized metabolism, metabolomics, transcriptomics, natural diversity.

INTRODUCTION

Common bean (*Phaseolus vulgaris* L.) is the main grain legume for direct human consumption and the second most economically relevant legume after soybean (*Glycine max*) (Broughton *et al.*, 2003). It represents a rich source of protein, vitamins, minerals, and fiber, particularly for poor regions of Africa and Latin America (Broughton *et al.*, 2003; Bitocchi *et al.*, 2016). *Phaseolus vulgaris* shows a very specific evolutionary history for which it has emerged as a model species to study the phenotypic changes associated with its domestication (Bitocchi *et al.*, 2017). The wild forms of *P. vulgaris* are distributed over a large area of Latin America, ranging from northern Mexico to northwestern Argentina, with the presence of at least three well differentiated eco-geographical gene pools: the Mesoamerican, the Andean and the Inca gene pool. The Andean gene pool is believed to have diverged from the original Mesoamerican pool approximately 165 000 years ago (Schmutz *et al.*, 2014), before the process of domestication that later took place independently for each gene pool around 8000 years ago (Gepts *et al.*, 1986; Mamidi *et al.*, 2014).
The Plant Journal, 2019, 97, 1132–1153

Secondary metabolism of common beans

1133

The populations of the Inca gene pool are instead distributed in a limited range across northern Peru and Ecuador; they also originate from migration of Mesoamerican wild forms, but no domesticated forms for this gene pool have been discovered so far (Bitocchi et al., 2012, 2013).

The structure of common bean populations, with its two independent domestication events and a rich repertoire of wild forms, land races and commercial varieties makes it a highly suitable candidate for forward genetic approaches based on natural variation. Such approaches represent a powerful method for investigation of mechanisms that underlie metabolic diversity and regulation, and ultimately in identifying candidates for the reintroduction of interesting genetic traits that were lost during domestication and breeding (Tanksley and McCouch, 1997; Paran and Zamir, 2003; McCouch, 2004; Fernie and Klee, 2011; Bellucci et al., 2014; Beleggia et al., 2016; Fernie and Tohge, 2017). An interesting target for such approaches is that particularly relevant both for plant fitness and human nutrition is specialized metabolism, which has been demonstrated to be one of the most affected traits by the loss of diversity occurring through the process of domestication (Meyer and Purugganan, 2013). Despite the importance of studying the changes of metabolites during the domestication process, however, the specialized metabolism of beans remains poorly characterized, especially in comparison with the knowledge acquired so far on the (central) metabolism of model legumes. Several studies have in fact focused on the characterization of transcript/metabolite dynamics in root nodules or during salt stress and ammonium accumulation in Lotus japonicus (Desbrosses et al., 2005; Sánchez et al., 2008, 2011; Pérez-Delgado et al., 2013); the metabolic changes during the rhizobial infection process (establishment of nodulation) have also been studied in soybean (G. max) (Brechenmacher et al., 2010); and several bioinformatic resources are also available in Medicago truncatula (Urbanczyk-Wochniak and Sumner, 2007; He et al., 2009; Zhang et al., 2014). Part of the reason behind the lack of large-scale studies of specialized metabolism in Phaseolus spp. lies in its complexity: beans contain several alkaloids, non-protein amino acids (NPAA), cyanogens, peptides, phenolics, polyketides, and terpenoids (Wink, 2013). Two of these classes, isoflavonoids and triterpenoid saponins have received particular attention in other legume species due to their proposed health-related properties. Soybean, a species closely related to the common bean (Schmutz et al., 2014) is well known to accumulate high levels of isoflavonoids, which in turn exhibit a diverse range of biological activities (Martin et al., 2013; Tohge and Fernie, 2017). While the case of soybean is exceptional and isoflavonoids accumulate at much lower levels in other legumes, this interesting class of specialized metabolite is almost exclusively found in the legume subfamily Papilionoideae (Veitch, 2013; Tohge et al., 2017).

The other highly relevant class of metabolites associated with multiple biological activities and often described as being highly abundant in the legumes is the triterpenoid saponins, particularly those derived from β-amyrin. Most of the knowledge regarding the biosynthesis and regulation of these compounds comes from studies in the model legume M. truncatula (Achnine et al., 2005; Naoumkina et al., 2010; Carelli et al., 2011; Fukushima et al., 2011, 2013; Pollier et al., 2011; Biazzi et al., 2015; Mertens et al., 2015). Nevertheless, many genes, particularly UDP-sugar dependent glycosyltransferase (UGTs) and cytochrome P450 monooxygenases (CYPs) were also functionally characterized in several plants with economical relevance and interesting medicinal properties, such as soybean (Shibuya et al., 2006, 2010; Sayama et al., 2012) and Glycyrrhiza uralensis (Seki et al., 2008; Xu et al., 2016; Tamura et al., 2017). A great resource for plant triterpene biosynthesis including structures, enzymes and pathways in multiple plants was recently made available through the TriForC database (Miettinen et al., 2018). In soybean, as well as in common bean, the main saponins are the soyasaponins, compounds derived from the hydroxylation of β-amyrin via CYP93E (Moses et al., 2014).

The specialized metabolism of common bean additionally includes other phenolic compounds, mainly flavonoids, with most of this knowledge based on phytochemical analysis of seed coats, sprouts and pods, which are the most commonly consumed parts of the plant (Tohge et al., 2017). Importantly, to date, only a few studies on common bean have attempted to investigate the genetic basis underlying their accumulation with most of these studies being related to flavonoids (Tohge et al., 2017). A couple of enzymes, namely chalcone synthase (PvCHS) (Ryder et al., 1984, 1987) and chalcone isomerase (PvCHI) (Dixon et al., 1982; Blyden et al., 1991), were characterized in relation to their possible roles in isoflavonoid biosynthesis. Through the work of several groups, Mendelian traits linked to seed coat color and flavonoid accumulation were described in different varieties of beans (Tohge et al., 2017), and more recently genes involved in anthocyanin biosynthesis in the pods were identified and characterized (Hu et al., 2015).

Besides the studies mentioned above, which have mostly focused on selected classes of specialized metabolites in model legume species (L. japonicus, M. truncatula and G. max), as yet no major effort was invested in developing an extensive characterization of common bean-specialized metabolism at the whole organism level. We believe that such a description would represent an important resource for further improvement of this important crop. Additionally, characterizing the extent of the differences in common bean metabolism, given the complex origin of this species with two parallel domestication
events, has the potential to shed light on mechanisms of adaptation and metabolic diversification. One of the greatest challenges in providing such a description of any species, widely discussed within the metabolomics community is that of metabolite annotation. While several tools provide interesting pipelines for metabolite annotation, most of these tools rely on complex experimental design, data acquisition (MS^n) and/or data analysis (Perez de Souza et al., 2017). In this study we made use of an alternative approach based on high-resolution mass accuracy, data-independent MSMS acquisition (DIA) and molecular network analysis to describe *P. vulgaris* metabolism within a tissue and accession-specific level. Furthermore, the integration of transcriptomics data allowed a thorough interpretation of genetic mechanisms underlying metabolites accumulation (Cavill et al., 2016). We illustrate the power of this integrative approach by using flavonoid, phenylpropanoid and triterpenoid metabolism as case studies, however we additionally provide the entire metabolomics and transcriptomics datasets such that they can be interrogated by researchers interested in any other of the metabolic pathways which we measure. In addition this web resource (http://www.mpimp-golm.mpg.de/719693/Bioinformatik-Tools) will ultimately be updated to contain both primary and lipophilic metabolites in order to provide coverage as broad as possible for metabolism within this species.

**RESULTS**

Metabolic diversity across tissues and accessions of *P. vulgaris*

Five different tissues (seedlings, roots, leaves, flowers and pods), from three distinct *P. vulgaris* accessions including a Mesoamerican wild (MW), a Mesoamerican domesticated (MD) and an Andean domesticated (AD) varieties were harvested, snap frozen in liquid nitrogen, ground, extracted and analyzed using LC-MS. Mass spectral signals from positive and negative ionization modes were manually analyzed for each sample in order to exclude commonly known adduct and fragment ions and reduce signal redundancy. In this manner only those features (mass to charge ratio, m/z, signals) most likely to represent [M–H]− and/or [M+H]+ quasi-molecular ions were selected. In total, 850 putative metabolites were consistently detected and areas from the negative ionization mode, as well as data independent MSMS data from both ionization modes, were subsequently extracted from the results obtained by processing mass chromatograms with MS-DIAL software v.3.08 (Table S1) (Pluskal et al., 2010). The results are provided in Table S1.

Principal component analysis (PCA) using singular value decomposition was performed on the basis of peak areas. The combined datasets show the dispersed nature of the data in the multidimensional space: the first principal component (PC1), which accounts for 37.9% of the variance separates the roots datasets from all remaining samples, while the PC2 (10.6%) effectively separates the other four tissues (Figure 1a). Tissue specificity is clearly the main source of variance in the metabolite dataset. However, there is still a significant proportion of variance represented by further PCs, which can only be explained considering accession specificity. The PC3 for instance separates most Andean tissues with the exception of the leaves away from the Mesoamerican samples (Figure S1). PCA calculated using each individual genotype (Figures S2–S4) exhibits a similar profile across the three genotypes with approximately 80% of variance explained by PCs 1–3; PC1 separates seedlings from the other tissues, PC2 separates all tissues well, and PC3 separates the pods from the other tissues. When the PCA was calculated separately for each tissue (flower, leaf, pod, root and seedling), we observed a clear separation of the three accessions along the first and second PCs (PC1 and PC2, Figure 1b–f). In particular, the Mesoamerican genotypes (MW and MD) clustered away with respect to the Andean accession along the PC1 in the flower, leaf, pod and root datasets, with the domesticated accession (MD) being further separated from the wild (MW) along the PC2 (Figure 1b–e). The seedling dataset was instead the one in which the three accessions had the highest degree of dispersion, and their separation was evident along PC1 (Figure 1f) between the wild (MW) and the two domesticated accessions MD and AD; the two domesticated accessions subsequently separated on PC2 (Figure 1f).

The standard scores (z-scores), defined as the number of standard deviations a value (metabolite within a sample) away from the respective variable mean (metabolite across all samples), were subsequently used for hierarchical clustering analysis (HCA). As the PCA already showed, clustering allowed the discrimination of four different tissues into clear groups (row clusters in Figure 2a). From HCA we could also identify sets of metabolites, specifically high-accumulating, in each distinct tissue/species combination. We defined these sets as ‘modules’, and they were represented as column clusters in Figure 2a). For example, module 01 (M01, dark blue) contains the profile, across all samples, of those metabolites specifically high-accumulating in the pods of the MD accession. The roots modules contained the highest number of metabolites with respect to all other tissue-specific modules; the profiles of the metabolites high-accumulating in flower and seedlings clearly differentiate the Andean genotypes from the Mesoamerican genotypes. The leaf module, as would perhaps be anticipated, contains a smaller number of metabolites whose profiles are instead largely similar across the three genotypes.
Metabolite annotation

In order to gain insights into the structural relationships between the specialized metabolites detected in the various datasets, we used an integrative network approach for metabolite annotation. The approach is based on the reconstruction of the potential biochemical conversions occurring in the set of detected features (Schwahn et al., 2014), together with MSMS spectral similarity analysis (Li et al., 2016). First, a network of biologically meaningful transformations was obtained by iteratively calculating m/z and retention time (rt) shifts for each pair of putative metabolites. These values were then compared with a table consisting of m/z differences associated with common reactions in specialized metabolism (e.g. glycosylation, oxygenation) and to the expected rt shifts resulting from these modifications (Table S2 modified from Morreel et al. 2014). Additionally, networks of MSMS spectral similarity (cut-off 50%) for all multiple collision energies were calculated using MS-FINDER molecular networking function (Tsugawa et al., 2016; Lai et al., 2017). Finally, two independent networks were obtained by applying an absolute correlation cut-off of 0.8 to the biochemical network and by calculating the intersection of the biologically oriented network against the union of all MSMS spectra similarity networks using Cytoscape (Shannon et al., 2003). Both of these networks were subjected to clustering by the community clustering algorithm (GLay) within Cytoscape app clusterMaker2 (Morris et al., 2011), in order to facilitate further investigation of promising metabolite/metabolite relationships within subclusters. Subclusters will be identified throughout the text by the prefix ‘B’ for the biological network and ‘MS’ for the spectra similarity network. Moreover, we validated our approach against MetFamily, an open tool for metabolite identification based on MS1 intensities and MSMS spectra analysis (Treutler et al., 2016). Overall, the results of both methods agreed, nevertheless the approach described here achieved a better discrimination of metabolite classes facilitating data interpretation (Table S1).

![Figure 1. PCA plot of the complete metabolite dataset (a) and of individual tissue-specific metabolite datasets (b) flowers, (c) leaves, (d) pods, (e) roots and (f) seedlings. The different plots represent combinations of the first and second principal components (PC). The data points represent the different bean accessions that are shaped according to the tissue (flower, leaf, pod, root or seedling) and colored according to the genotype (Andean domesticated (AD), Mesoamerican domesticated (MD) and Mesoamerican wild (MW)). (a) The combination of first PC efficiently separates the root data (of all accessions) from all remaining tissues. All remaining tissues are further separated across the second PC. In flowers (b), leaves (c), pods (d) and roots (e), the first PC separates Mesoamerican from Andean accessions and the second PC further separates Mesoamerican accessions. In the PCA plot of the seedlings (f), the first PC separates the wild from domesticated accessions while the second PC further separates all the accessions.](image-url)

© 2018 The Authors. The Plant Journal published by John Wiley & Sons Ltd and Society for Experimental Biology, The Plant Journal, (2019), 97, 1132–1153
accumulating within the roots, as indicated by the HCA module color (shades of green). Following this approach, we obtained 34 and 48 subnetworks containing between 33 and 70 metabolites for the biological and MSMS similarity networks respectively. In parallel with the network-based approach we also used MS-FINDER to match all MSMS spectra against multiple databases including MassBank (Horai et al., 2010), METLIN (Smith et al., 2005) and KNApSAcK (Afendi et al., 2012). Top hits for every metabolite were manually curated taking into consideration the metabolites and metabolic pathways previously described in common beans as well as other legumes, and a deeper investigation of MSMS spectra. Metabolite annotations were further integrated into the metabolic network and were used to assign putative structures for other unknown metabolites as well as to propose the metabolic pathways further described in this report. A summary of all annotations including subnetwork information is provided in Table S1. A high proportion of subnetworks was populated by putative saponins, flavonoids and phenylpropanoids, mostly hydroxycinnamoyl derivatives. However, it is important to note that approximately 40% of the analytes captured within the network lacked a structural annotation. Moreover, in Table S1, we distinguished annotation of those metabolites for which we could verify clear diagnostic peaks to confirm the proposed structures by including a compound name. However, we should stress that this does not exclude the possibility of any positional isomers of the proposed compound, which could exhibit indistinguishable mass spectra. Spectra matching with high score to database metabolites, but lacking clear peaks for the unambiguous assignment (e.g., aglycone fragment), were assigned only to the compound class.

It is also important to note here that mass spectrometry alone is unable to provide unambiguous elucidation of molecular structure by itself. Still, this integrative approach provides a high degree of confidence for the annotation of a large number of compounds, unlikely to be achievable by any other technique, particularly in relation to metabolites classes (e.g. flavonoids, saponins). In order to assign a tentative annotation to the unknown metabolites, we manually curated all individual subnetworks, also with the aim of checking how plausible the predicted pathways were. It is important to stress, in fact, that our network-
based approach may include artificial connections between metabolites. This caveat is particularly evident for the biochemical network, as some of the differences (biochemical conversions) considered in our approach may be present among related metabolites without necessarily representing a ‘true’ biochemical modification. In such cases, a careful inspection of the mass spectra fragments is necessary to properly annotate metabolites. A good example of this discrepancy can be observed in subnetwork B12 from the biochemical network, for which the raw data are represented in Figure 3(a). The metabolites included in this subnetwork were annotated as a group of glycosides of the flavonol aglycones in *P. vulgaris* (Figure 3b). Figure 3(c) represents the proposed pathway for flavonol biosynthesis in common beans, including annotated metabolites from subnetwork B12 (highlighted in red) and additional intermediates that were either absent (highlighted in gray) or present in multiple different subnetworks (highlighted in blue). This particular pathway is a great proof of concept as it is well characterized and conserved among higher plants, facilitating its reconstruction and identification of proposed steps. The details are extensively discussed in the literature (Saito et al., 2013; Tohge et al., 2013, 2017). In summary it starts with the production of naringenin chalcone from *p*-coumaryl-CoA and malonyl-CoA via chalcone synthase (CHS). This structure is modified by several cytochrome P450s (CYP) and 2-oxoglutarate-dependent dioxygenases (2-OGD) including flavanone 3-hydroxylase (F3H), flavonoid-3′-hydroxylase (F3′H), flavonol synthase (FLS), anthocyanidin synthase (ANS), dihydroflavonol reductase (DFR) and isoflavone synthase (IFS) leading to the several subclasses of flavonoids. Finally, aglycones are further decorated in a series of reactions, some of the most common include the addition of sugars and acyl side chains by glycosyltransferases and acyltransferases. In *P. vulgaris*, two main flavonol aglycones are detected (Hungria et al., 1991) (kaempferol and quercetin, Figure 3b), both of which could be detected in our reconstructed pathway. Furthermore, flavonols representing isorhamnetin (methylated-quinertin) are also detected. In Arabidopsis and other dicots, aglycones are first glycosylated at the C-3 position, possibly followed by glycosylations of C-5 and C-7 positions (Figure 3b). Further glycosyl units can also be added to the sugars directly linked to the aglycone to generate highly decorated flavonoids. Based on this general pathway structure and flavonoid structures previously described in common beans (Tohge et al., 2017), we propose the pathway in Figure 3(c). By comparing the structure of both raw and curated networks, it can be observed that the oxygenation conversions linking two metabolites in Figure 3(a) actually correspond to a single biochemical step: the basal oxygenation of the kaempferol aglycone giving rise to the parallel quercetin branch, as represented at the top of Figure 3(c).
Additionally, structural differences resulting from multiple modifications to a single precursor are often identified and interpreted as a single modification. In the network of Figure 3(a), this option is the case for the addition of a deoxyhexose, which actually corresponds to two separated steps, an oxygenation and addition of hexose, both also represented in the network. By investigating the fragmentation pattern, however, we can infer the real biochemical difference. It is also relevant to highlight here the splitting of this pathway in several subclusters within the biological network. This fact is no surprise, as we included the correlation cut-off exactly to reduce the complexity of the clusters and facilitate interpretation. Indeed, when not applying this correlation cut-off, all these metabolites cluster together but also include a multitude of spurious connections within the resulting network.

Following the same framework that proved to work well for the annotation of well known flavonoids, we expanded the characterization to other classes of metabolites. In total, we assigned a putative class for 40% of the metabolites detected and a putative structure to approximately 26%. The results are reported in Table S1. It is relevant to highlight that, despite all the information gathered providing great confidence in the assignment of metabolites to particular classes, the same degree of confidence cannot be assumed for the absolute structures. Therefore, for several metabolites, we provide a putative class, but not a putative structure for the metabolite. This situation is particularly evident for the saponins due to the contribution of several factors including high mass spectra and structural complexity and often low concentrations to obtain reliable MS2 spectra for full characterization.

13C labeling

As previously highlighted, even though our approach should provide a great confidence regarding metabolite annotation, we have to point out that mass spectrometry alone is unable to provide unambiguous elucidation of molecular structures by itself. With this in mind, we performed 13C labeling experiments to validate at least part of our annotations. Upon the choice of labeling substrate, it may seem reasonable to use a substrate that should be metabolized and ubiquitously distributed, such as 13C-glucose or 13CO2. However, our experience with such experiments suggests that, while labeled 13C can quickly incorporate into certain central metabolites, this process tends to be particularly slow for most secondary metabolites. Moreover, the independent incorporation of label into multiple positions of the same compound in the accumulation of all possible isotopomers of labeled/unlabeled carbons significantly complicating data analysis. By contrast, feeding a substrate that is a direct precursor to a specific metabolic pathway, and that contributes to the incorporation of a fixed discrete number of labeled positions, would provide a faster and technically simpler experiment. For this reason, we selected 13C-phenylalanine, as it is a precursor for two of the main classes of specialized metabolites detected in beans, the flavonoids and hydroxycinnamic acid derivatives. Additionally, the phenolic ring is conserved across nearly every further biosynthetic step in these pathways, providing an excellent tracer to be detected by mass spectrometry. Leaf discs of the three genotypes were fed in solution and the incorporation of the label into metabolites was accessed by comparing the ratio between label and unlabeled m/z for treated and control samples (Experimental procedures). The results for metabolites representative of each class are presented in Figure 4. For each compound, a higher value of the (M + 6)/M ratio in the labeled compared with the unlabeled
sample is an indication of the incorporation of the carbons derived from the $^{13}$C$_6$-phenylalanine (as for the metabolites reported in Figure 4, Hydroxycinnamates). When the value of (M + 6)/M was not significantly different between the labeled and the unlabeled samples (as in Figure 4, Sapinos), it denoted an negligible enrichment of the M + 6 isotopic peak and therefore a lack of incorporation of the labeled carbons into the structure of the molecule. The results obtained from the feeding of $^{13}$C$_6$-phenylalanine helped to reinforce our putative metabolite annotations. In fact, most hydroxycinnamates exhibited high enrichment (Figure 4), while the enrichment for flavonoids was significantly lower but still detectable (Figure 4). The clear enrichment in the M + 6 m/z signal for these metabolites provided an easy confirmation for the presence of the phenylalanine-derived phenolic ring in the respective structures. For other classes of metabolites that do not use phenylalanine as a precursor, such as saponins, enrichment was completely undetectable (Figure 4). This lack of labeling provides a nice proof of concept, allowing us to exclude the hypothesis of phenylalanine as an indirect supply of $^{13}$C via alternative metabolic routes such as its degradation in our experimental setup. Furthermore, the difference in labeling between flavonoids and hydroxycinnamates is particularly interesting, and suggests higher flux towards the latter. Such result is no surprise as hydroxycinnamates also serve as precursors for lignin building blocks, a pathway estimated to consume up to 30% of fixed carbon in land plants (Pauly and Keegstra, 2008; Wang et al., 2018). Indeed, the fast and high incorporation of $^{13}$C$_6$-phenylalanine into hydroxycinnamates was recently shown in Arabidopsis (Wang et al., 2018).

**Analysis of tissue and accession specificity**

One of the main outcomes of our analysis is to access the extent of metabolic divergence across multiple common bean accessions. While we do agree that the current sample size does not allow for a conclusive evaluation of domestication effects over specific metabolic pathways, we reinforce that these samples were selected based on their genetic diversity and should represent a comprehensive blueprint of the main specialized metabolism pathways present in beans and the extent of their differences across different gene pools. Moreover, the natural variation across accessions and tissues provides a starting point to investigate the molecular mechanisms involved in the biosynthesis of these compounds.

It is unpractical to discuss every single metabolite here, but a few metabolite groups are particularly worth mentioning. Most flavonoids identified were classified as flavonol glycosides. A smaller group, however, clearly separating in subclusters B7 and MS15 from the biochemical and MSMS similarity networks, respectively, included several compounds annotated as derivatives of the isoflavones daidzein and genistein (Table S1). In agreement with expectations based on the published literature, these metabolites were accumulating exclusively in the roots and at lower levels in the seedlings (Figure 5) (Dhaubhadel et al., 2003; de Lima et al., 2014). The isoflavonoids were detected in all accessions with the exception of Met 345, annotated as a putative methyl-genistein glucoside that was absent in the Andean variety (Figure 5). Surprisingly, no peaks were assigned to anthocyanins, which are expected to be present in beans. These metabolites were mostly probably excluded from the analysis during the data processing step, as one of the criteria used to assign molecular ion peaks was based on the combination of positive and negative ionization modes, together with the presence of common adducts. Anthocyanins are known to not easily ionize in negative mode, and a targeted analysis would be necessary to identify these compounds. Regarding flavonol glycoside distribution, the major compounds represented in Figure 3(c) were generally accumulating at particularly high levels in flowers. However, there were specific metabolites over-cumulating in all other tissues with the exception of the seedlings. The accession specificity of these compounds was interesting as we can observe a good correlation with glycosylation pattern. Metabolites in the biochemical subnetwork B12, including tri- or tetra-glycosylated flavonols composed of the three possible sugar substituents proposed here, are glucose, rhamnose and xylose. All exhibited a high correlation with the MW accession. These metabolites were nearly completely absent in the MD accession, suggesting that the MD accession lacks or has low levels of glycosyl-transferase activity that leads to these highly glycosylated flavonols. It is also interesting to note that xylose derivatives were absent in most tissues for AD, with the exception of flowers, and few metabolites were absent in MW accession (Figure 3d).

Another major class of compound detected was that of hydroxycinnamate derivatives. Most of these metabolites could be reliably annotated due to their relatively simple structures, easy fragmentation and on the basis of the information retrieved from the $^{13}$C labeling experiment previously described. The MSMS similarity network was particularly efficient in grouping, almost exclusively, compounds annotated as hydroxycinnamates in subnetwork MS30. Furthermore, subnetworks within the biochemical network separated these compounds according to their sample specificity. Subnetwork B39 was composed mostly of p-coumaroyl (all metabolites within this group in Figure 5 except for Met316), feruloyl (Met316, Met541, Met582 and Met635) and sinapoyl (Met552, Met597 and Met648) glucosides (Table S1, Figure 5). These metabolites accumulated highly in flowers and, from all of these, genotype AD exhibited the highest levels (Figure 5). Interestingly, both MD and MW flowers exhibited purple petals.
strongly suggesting the presence of anthocyanins, while AD flowers were white. Studies in Petunia shown that downregulation of caffeoyl-coenzyme A O-methyltransferase (CCoAOMT) involved in the production of feruloyl CoA activated the anthocyanin pathway (Shaipulah et al., 2016). This result suggests a possible influence of CCoAOMT expression in *P. vulgaris* as well. Most of the metabolites in subnetwork B4 were annotated as *p*-coumaroyl (Met139 and Met199), feruloyl (Met200) or sinapoyl galactarates (Met244) (Table S1). These metabolites were highly accumulated in roots and seedlings of all three genotypes with particularly high levels in the roots of AD accession (Figure 5). Finally, an interesting group of compounds within subnetwork MS30 of MSMS similarity included several metabolites annotated as hydroxycinnamates esterified to malate (Met428, Met402, Met411 and Met335) or tartaric acid (Met242 and Met285), including phaselic acid (Met335) that was previously described in beans. These metabolites also exhibited a very distinct accession distribution with tartaric acid derivatives being present only in the MD, while malate derivatives were absent in the same genotype (Figure 5).

The last and largest of the major classes of annotated compounds were the saponins. The unambiguous assignment of specific identities for this class of compounds was severely hindered by the complexity of their structures with a large number of isomers resulting in very similar spectra. Several compounds could be assigned as putative saponins based on the high number of matches to standard spectra and their clear grouping within specific clusters. Therefore, this discussion is restricted to specificity of subcluster distribution rather than trying to speculate on aglycone and substitution pattern. Six main subnetworks from the biochemical network should be highlighted in relation to the saponins. Subnetwork B2 includes several metabolites that accumulate majorly in roots and seedlings at similar levels without exhibiting any clear accession specificity (Figure 5). Subnetwork B3 included metabolites accumulating at much higher levels in roots but also present at lower levels in seedlings and flowers (Figure 5). Subnetwork B11 was similar, with the difference that these
metabolites were also absent in flowers (Figure 5). Subnetwork B23 included metabolites specific for seedlings and that accumulated at much higher levels in the MD genotype (Figure 5). Metabolites within subnetwork B26 were specific to flowers and pods of the MW accession only (Figure 5). Finally, subnetwork B32 included metabolites specific for MD genotype, accumulating in flowers, pods and to a lesser extent roots (Figure 5). Some saponins were assigned a putative structure, most of these were expected to belong to the class of soyasaponins, previously described in soybean and in common bean seeds (Dong et al., 2007). However, as previously mentioned, the complexity of these structures resulted in a much lower confidence to these annotations exclusively based on the mass spectral analysis.

Comparative gene expression analysis
To characterize the general transcriptional divergence and to identify the genes potentially involved in the biochemical conversions described so far, we selected the Andean and the wild Mesoamerican accession for comparing their expression profiles across some of the tissues (leaf, seedling and pod). These two accessions were selected because they are highly divergent on the basis of the amount of SNPs (Bellucci et al., 2014) and also for the availability of whole-genome data for the Andean accession (Schmutz et al., 2014). It is important to highlight that the samples utilized for transcriptomics analysis were obtained from exactly the same material sampled for metabolomics. Flower transcriptomics was performed only for the Andean variety, therefore we excluded this tissue from comparative gene expression analysis and included it only for metabolite/transcript correlation described further. A summary of the data used by DeSeq2 for calculating DE genes as well as the transcripts per million (TPM) normalized data is available in Tables S3 and S4. Data quality was initially assessed by PCA analysis in which we observed a clear separation between tissues as well as genotypes (Figure S5). Interestingly, while metabolomics data split seedlings apart from the other three tissues in the PC1 with 23.8% of the variance, the transcriptomics data show the separation of flowers in the same PC accounting for nearly double the variance (42%). We then calculated DE genes for all pairwise comparisons separately for the two accessions. The results are reported in Table 1. The comparison showing the highest number of DE genes in both the Andean and Mesoamerican accessions is leaf versus seedling.

We then used the ensemble of DE genes as the starting dataset to select tissue- and accession-specific genes. We defined tissue-specific genes as those that were only DE in the leaf (or any other tissue) of both accessions with respect to all other samples. Similarly, we defined the accession-specific genes as those that were exclusively DE in a single tissue of a specific accession with respect to all other remaining tissues from the same accession. Extracting the accession-specific genes, in particular, allowed us to look at the global transcriptional divergence of the individual tissues, an aspect that was not evident looking only at the pairwise comparisons. The total numbers of tissue- and accession-specific genes are reported in Figure 6(a), the heatmap representing the expression values of the various tissue- and accession-specific genes is instead presented in Figure 6(b). In this heatmap, the genes are ordered in nine modules (a-i) each representing the expression values in a specific tissue or tissue/accession combination (this module information is also available for individual genes in Table S4). As it would have been partially expected, the leaf is the most divergent tissue in terms of the number of differential transcripts (Figure 6a-c); the Gene Ontology (GO) categories in this module are enriched for those representing typical functions of an active photosynthetic tissue (e.g. carbon fixation, pyruvate metabolism, synthesis of isoprenoids; Table S5, Leaf). A large fraction of genes (3318; Figure 6b) is instead driving the large divergence of the AD Leaf-specific genes from all remaining tissues. When we compared the unique GO categories individually for the sets of AD leaf- and MW leaf-specific genes, we found that the AD leaf was characterized by the presence of categories related to tryptophan- and indole-related metabolism (Table S5, AD Leaf). The module containing the MW-leaf specific genes, despite the lower number of specific genes (1185; Figure 6c) was instead enriched for several gene categories, mainly related to primary metabolism (Table S5, MW Leaf).

Despite the larger metabolic differences driving the separation of the seedlings from leaves, pods and flowers (Figure S6), this tissue is instead relatively less divergent in terms of transcriptional variation, with respect to all other remaining tissues. Indeed only a minor fraction of genes are specifically DE in seedlings (720 genes; Figure 6d); Furthermore, the two sets of the AD seedling- and MW seedling-specific genes (Figure 6e,f) differ only in a few unique GO categories: among these, the AD seedlings were enriched for transcripts involved in developmental processes and (polar) primary and lipid metabolism (Table S5; AD Seedling); while the MW seedling-specific genes contained categories mainly related to metabolism

Table 1 Number of DE transcripts detected in each pairwise comparison (P-values Benjamini–Hochberg corrected < 0.05).

| Pairwise comparison | No. of DE transcripts |
|---------------------|-----------------------|
| Leaf/seedling       | 13 954 9959           |
| Leaf/pod            | 13 678 9164           |
| Pod/seedling        | 9202 9273            |

© 2018 The Authors. The Plant Journal published by John Wiley & Sons Ltd and Society for Experimental Biology., The Plant Journal, (2019), 97, 1132–1153
of sulfur-containing amino acids and regulation of cell growth (Table S5, MW seedling).

We then looked at the transcriptional differences detected in the pods. In general, similar to the seedlings, pods were characterized by a far lower number of specific genes with respect to the leaf (620; Figure 6g); the pod AD-specific genes (Figure 6h) were enriched, among others, for GO categories related to glyco-/glycerolipid metabolism (Table S5, AD Pod). Pod MW-specific genes (Figure 6i) differed for the presence of several GO categories all related to protein synthesis (Table S5, MW Pod).

**Phylogenetic analysis of candidate genes**

Given that the majority of predicted biochemical reactions described here involve glycosylations and hydroxylations of common backbones, we performed a phylogenetic analysis on the members of two gene families (UGT and CYP) known to be involved in these kinds of reactions. We thus retrieved from the reference genome sequence (AD accession, Schmutz et al., 2014), a set of amino acid sequences representing UDP-dependent glycosyltransferases (UGTs) and cytochrome P450 (CYPs). These families were selected using the following signatures obtained from InterProScan: PF00201 (for UGTs) and PF00067 (for CYPs). For each gene family we also included additional gene family members that have been characterized in other species producing similar compounds. The complete list of genes used is provided in Table S6.

We highlight below a few particularly interesting candidates that were selected on the basis of their phylogenetic proximity to previously characterized genes in other plant species. Moreover, we performed ortholog analysis with these candidates (Figure S7), including several legume and non-legume plant species as described in the Experimental procedures section. We start with candidate UGTs related to flavonoids and other phenylpropanoid genes (Figure 7). Flavonoid UGTs usually form distinct clusters based on the specificity for their sugar attachment position (Yonekura-Sakakibara et al., 2014). Flavonoid 3-O-glycosyltransferases (F3GT) are involved in a modification commonly found in plants and are essential for stable accumulation of flavonoids in Arabidopsis, namely the flavonoid-O-glycosylation at the 3-position (Tohge et al., 2005; Yonekura-Sakakibara and Hanada, 2011). Four common bean candidates (Phvul.002G214300, Phvul.002G214400, Phvul.006G201100 and Phvul.003G148300) were identified in the F3GT cluster. Results from orthologous analysis include five common bean orthologs of F3GT including our four candidates (Figure S7, F3GT). The fifth ortholog, Phvul.002G214500 was very lowly expressed in the tissues included in this work. Flavonoid 7-O-glycosyltransferases (F7GT) are usually found in dicots in regions including up to eight orthologs in close proximity (Tohge et al., 2013, 2017). Seven F7GT bean candidates were found (Phvul.001G182100, Phvul.001G182200, Phvul.001G182300, Phvul.001G182400, Phvul.001G182600, Phvul.001G182800, and Phvul.007G152800)
clustering together with previously characterized AtF7GlcT (Jones et al., 2003) and SbF7GlcT from Scutellaria baicalensis (Hirotani et al., 2000). Six of these candidates were found to be localized in the same region of chromosome 1 (spanning approximately 55 kbp), in a similar arrangement to that previously described for other species. Additionally, Phvul.011G151100 was identified in the same clade as AtF7RhaT from Arabidopsis (Yonekura-Sakakibara et al., 2007). From orthology analysis one large orthogroup comprising 35 common bean orthologs included all the candidates in chromosome 1 as well as Arabidopsis AtF7GlcT (Figure S7, F7GT). Phvul.011G151100 was included in another orthogroup together with AtF7RhaT and two other orthologs in common bean (Figure S7, F7RT). The two orthologs, Phvul.003G036900 and Phvul.011G151200 exhibited similar expression pattern with higher expression in AD leaves while Phvul.011G151100 was predominantly expressed in MW Pods. We also included several UGTs involved in glycosylation of the sugar moieties of flavonoid glycosides (F3GGT, Figure 7) (Yonekura-Sakakibara et al., 2014). A group of eight common bean candidates was identified within this clade (Phvul.007G135800, Phvul.011G128700, Phvul.005G093300, Phvul.005G093500, Phvul.005G093600, Phvul.011G138500, Phvul.011G138600 and Phvul.011G136700). All of these candidates were included within the same orthogroup, which included 11 orthologs in common beans (Figure S7, FGGT). From the three orthologs that were not included among our candidates only Phvul.011G136400 was expressed at particularly high levels in leaves. Four candidates were found in a cluster with glycosyltransferases involved in the glycosylation in the 7-position of flavones and isoflavones and the 4’-position of chalcones (Noguchi et al., 2007) (Phvul.004G103900, Phvul.004G104200, Phvul.005G005600 and Phvul.005G005400, IFTGT; Figure 7). These candidates were included in an orthologous group with two other orthologs in common bean from which Phvul.004G104100 is the only gene expressed in leaves and roots (Figure S7, UGT88E). The last family of flavonoid-related UGTs investigated was the anthocyanin-5-O-glycosyltransferases (ASGT), but no candidates were identified for Common bean in this study. Two other classes of phenylpropanoid-related genes were included in our phylogenetic analysis: monogalactosyl 4-O-glycosyltransferases (M4GT) (Lanot et al., 2006) and a sinapic acid:UDP-glucosyltransferase (SGT). Two common bean candidates were identified close to these M4GTs, Phvul.003G021900 and Phvul.002G026900, both belonging to the group of 17 orthologs of MGT in common bean (Figure S7, MGT). Finally a group of five candidates (Phvul.007G039300, Phvul.007G039400, Phvul.007G039500, Phvul.007G039600 and Phvul.003G239700) formed a cluster including Arabidopsis SGT (AT3G21560), a gene involved in the production of 1-O-sinapoylgucose for anthocyanin biosynthesis (Yonekura-Sakakibara et al., 2012). Interestingly, from these candidates, Phvul.007G039300 was the closest to the characterized SGT and the only one included in the same orthogroup as the Arabidopsis protein (Figure S7, SGT). All other candidates in this clade were in a different orthogroup (Figure S7, SGT-like) as with orthologs exclusively presenting in the brassica and legume species evaluated within this work. We speculate that some of these candidates could be involved in the production of a series of hydroxycinnamates identified in common bean, including phaseic acid.

Saponin-related UGTs (Figure 8) are characterized to a much lower extent than flavonoid UGTs. For our phylogenetic analysis we included characterized proteins from soyasaponin biosynthesis in soybean and Medicago, as well as other classes of triterpenoid saponins in Medicago and G. uralensis and steroidal alkaloids in Solanum species. All of these proteins, with the exception of soybean GmSGT3 included in the transfer of a rhamnosyl group from UDP-rhamnose to soyasaponin III, belong to the same family as flavonoid-related F7GTs (Shibuya et al., 2010). Indeed all saponin-related candidates identified belong to the same large orthogroup as the flavonoid-related F7GTs (Figure S7, F7GT), with the exception of the candidates in the clusters of GmSGT3 and MtUGT73K as highlighted below. From our phylogenetic reconstruction we can observe a clearly distinct clade formed by GmSGT3 and seven Phaseolus candidates (Phvul.006G208300, Phvul.006G208400, Phvul.006G208500, Phvul.010G036800, Phvul.010G036900, Phvul.010G037000 and Phvul.010G037600). A similar clade can be observed, including soybean GmSGT2 involved in transferring a galactosyl group from UDP-galactose to soyasapogenol B monoglucuronide (Shibuya et al., 2010) and five common bean candidates (Phvul.002G016400, Phvul.002G016500, Phvul.007G020600, Phvul.007G020700 and Phvul.007G020800). One common bean candidate (the candidate Phvul.010G101600) was identified in the cluster including MtUGT73K1, a UDP-glucosyltransferase shown to have specificity towards hederagenin and soyasapogenols B and E in Medicago (Achnine et al., 2005) and its putative counterpart in G. uralensis Glyur002597s00038050 (Mochida et al., 2017). Orthologs of these genes are only found in legumes, numbering between one and four copies in the species investigated (Figure S7, UGT73K). Common bean has four copies from which only Phvul.010G101800 and Phvul.010G101600 are expressed in the investigated tissues, at particularly higher levels in the seedlings. A cluster with soybean’s GmUGT73F2 and GmUGT73F4 catalyzing the addition of glucose and xylose, respectively, to an arabinose residue at the C-22 position of soyasaponins from the A class (Sayama et al., 2012) and Medicago’s UGT73F3 involved in the glucosylation of multiple sapogenins at position C-28 (Naoumkina et al., 2010) also included the bean candidate Phvul.003G097300. The last cluster of UGTs involved in
triterpenoid biosynthesis including candidates from beans was formed by *G. uralensis* GuUGAT involved in a two-step glucuronosylation yielding glycyrrhizin (Xu et al., 2016) and three candidates localized in tandem in the same region of chromosome 3 (Phvul.003G046900, Phvul.003G047000 and Phvul.003G047100). Finally, no candidates were identified close to glycoalkaloid-related proteins, and a group of three candidates (Phvul.008G090500, Phvul.008G090600 and Phvul.008G090900) was localized in between the clade of GmSGT2 and GuUGAT.

Among the CYPs (Figure 9), putative candidates for three classes of flavonoid-related genes were identified: the only two flavonoid 3′-hydroxylase orthologs in common bean (F3′H, Phvul.003G185500 and Phvul.L001623, Figure S7) involved in the hydroxylation of 3′-position of dihydrokaempferol or kaempferol yielding dihydroquercetin and quercetin (Schoenbohm et al., 2005); one flavonoid 3′,5′-hydroxylase (F3′5′H, Phvul.006G018800), a protein that catalyzes the 3′,5′-hydroxylation of dihydroflavonols leading to the production of delphinidin-based pigments (Takahashi et al., 2010) and the only ortholog of this gene expressed in the analyzed tissues (Figure S7, F3′5′H); and the only two legume-specific isoflavone synthases orthologs (IFS, Phvul.003G074000 and Phvul.003G051700, Figure S7) highly expressed in the seedlings involved in the first step committed with the production of isoflavonoids (Veitch, 2013). Additionally, we could also identify three candidates of the general phenylpropanoid metabolism: two cinnamic acid 4-hydroxylase (C4H, Phvul.003G027400 and Phvul.006G079700) and a ferulic acid 5-hydroxylase (F5H, Phvul.003G259200).

Saponin-related CYPs included previously characterized CYP93E9 (Phvul.008G128700.1) (Moses et al., 2014).
involved in β-amyrin hydroxylation and several other genes divided in four main clades comprising proteins from the CYP72A and CYP716A families, as well as two smaller clades formed with GuCYP88D6 and MtCYP87D16. The CYP716A family was recently shown to be a major contributor to triterpenoid diversity, being conserved in eudicots and involved in the biosynthesis of a variety of different aglycones (Miettinen et al., 2017). Even though there are no saponin-related CYP716A members characterized in soybean or common bean, we could identify three genes (Phvul.002G302100, Phvul.008G034400 and Phvul.008G034500) exhibiting high similarity to CYP716A12 and CYP716A179 from *Medicago* and *G. uralensis*, respectively. They indeed represent all the orthologs of the *Medicago* gene in common beans (Figure S7, CYP716A). Both these enzymes exhibit triterpene C-28 oxidase activity (Fukushima et al., 2011, 2013; Tamura et al., 2017). In the family CYP72A, one gene (Phvul.011G000200) exhibited high similarity to CYP72A61 from *Medicago*, which is involved in 24-hydroxy-β-amyrin-C-22β-hydroxylation (Fukushima et al., 2013). Four candidates (Phvul.005G063000, Phvul.011G161400, Phvul.011G161600 and Phvul.011G161700) were identified in a clade with CYP72A65v2 and CYP72A63 from *Medicago* as well as CYP72A154 from *G. uralensis*. Both CYP72A154 and CYP72A63 are shown to possess β-amyrin C-30 oxidase activity (Seki et al., 2011), while CYP72A65v2 activity is still elusive but it was shown to be highly co-expressed with β-amyrin synthase (Fukushima et al., 2013). Most of these candidates were all localized in a region with several genes with a similar intron–exon structure in tandem repeats. Moreover, several other orthologs of this family were present in common beans (18 in total; Figure S7, CYP72A) and were highly expressed. These could represent interesting targets to explain the great diversity of saponin structures identified in this work. Finally, four genes belonging to the same orthogroup were closely related to the soybean gene CYP72A69 (Glyma15g39090), which was recently
characterized as a monoxygenase involved in the hydroxylation of soyasapogenol B to soyasapogenol A. (Yano et al., 2017).

**Transcript-metabolite correlation analysis**

In order to further restrict the number of candidate genes involved in the biosynthesis of early phenylpropanoids, flavonoids and saponins, we analyzed the correlation between the expression of the *Phaseolus* genes selected from the phylogenetic analysis and the accumulation of specialized metabolites from selected subnetworks represented in Figures 3 and 5. The results are represented in Figure 10. As expected, we observed that highly correlated clusters of transcript–metabolite pairs (red areas in Figure 10) were strongly associated with the predominant tissue in which metabolites accumulate. Although none of the clusters of transcripts are associated with a single class of metabolites, nevertheless they show enrichment for phenylpropanoids, flavonoids or saponins. The high correlations observed for some transcripts with respect to multiple groups of compounds, in addition to reflecting shared tissue specificity, may also point at the general enzymatic promiscuity of these genes towards different substrates. This is a well documented phenomenon in specialized metabolism and there are a few described cases in legumes, for example in which UGTs active towards triterpenes can also glycosylate flavonoids (Achnine et al., 2005; Naoumkina et al., 2010). Certain specific CYP families display high catalytic promiscuity (Hamberger and Bak, 2013), particularly towards more similar substrates, as it is the case for CYP716A12 of *M. truncatula* capable of oxidizing β-amyrin, α-amyrin and lupeol (Fukushima et al., 2011, 2013). Such lack of specificity could explain the high correlation of these transcripts to compounds belonging to distinct groups and even to distinct classes.

Two clusters of transcripts showing particular evident correlation with groups of metabolites also showed consistent annotations for metabolite and transcript putative classes within them. The first one corresponds mainly to the seedling-specific saponins (Figure 6) in the biological subnetworks B2, B3 and B11 which show high correlation to the cluster of UGTs and CYPs highlighted in green (Figure 10). Looking into the levels of these transcripts and metabolites (Figure 5) we can observe that they are particularly intense in the Andean accession. The second cluster showing particularly intense and consistent transcript–metabolite correlations include metabolites in subnetworks B39, B12 and B22 (Figures 3 and 5 and Table S1) together with several flavonol and flavonoid-related
transcripts, as well as the phenylpropanoid-related SGTs and M4GT (highlighted in blue in Figure 10). This group of metabolites and transcripts are mainly associated with flowers (Figures 3 and 5). Unfortunately, flowering of the MW accession was very sensitive to the growth conditions used in this study and it was impossible to get enough material for the transcriptomic analysis. Accession specificity of this group of transcripts and metabolites was therefore not discussed further.

DISCUSSION

In this study we provided a comprehensive description of the metabolic variation present in several tissues and accessions of common bean. Despite the large economic importance of *P. vulgaris*, the extent of its natural metabolic variance has not been analysed thoroughly, with the exception of a few studies. These studies focused on changes in specific classes of metabolites such as anthocyanins and their relationship to seed color (Choung et al., 2003) and the response to various perturbations, including the effects of illumination upon common phenolic compounds (Aguilera et al., 2014), and phosphorous stress on primary metabolism (Hernández et al., 2007). The lack of an extensive survey of metabolic diversity is perhaps unsurprising given that a major obstacle in metabolomics is the annotation of unknown metabolites. This obstacle is particularly challenging due to the large diversity that specialized metabolites show across tissues and species, in particular in a large plant family such as Fabaceae (Wink, 2013). Although hydroxycinnamates, flavonoids and triterpene saponins (the three classes analyzed in this study) showed a wide distribution in all tribes of Fabaceae, as demonstrated in this study, they showed a marked chemical diversity even in a single, although highly polymorphic, species such as *P. vulgaris*.

Metabolite diversity across species or accessions is usually driven by the presence of genetic polymorphisms at the level of structural or regulatory genes. These genes may determine, for example, the emergence of metabolic novelty in terms of specific decorations of the basic backbones (e.g. acylations and sugar decorations of flavonoids) (Tohge et al., 2016). Conversely, metabolic specialization across developmental stages or different tissues of a single plant may be the result of epigenetic changes and subsequent transcript differential expression, as during tomato fruit development (Zhong et al., 2013).

While multiple approaches do exist for mining this metabolite diversity across species or within tissues of a single plant, they usually require analysis of tandem mass spectrometry data, spectroscopic techniques (e.g. NMR) for structural elucidation, confirmation of metabolite annotation with authentic standards and computational approaches (Li et al., 2015, 2016; Perez de Souza et al., 2017). In this study, we described a simpler approach and provided an example of an annotation pipeline that can be applied to metabolic datasets obtained from the profiling of multiple tissues from large collections of species/accessions. The approach described is based on the combination of first-order mass spectra, tandem MS, chromatographic behavior and on the integration of these datasets to extract groups of metabolites that could be related based on the existence of a potential biochemical conversion between them (Morreel et al., 2014; Schwahn et al., 2014). This simple approach was not only capable of significantly improving the annotation of unknown mass signals, but also provided a network of metabolite–metabolite relationships that proved extremely helpful for data interpretation.

The pipeline described here used metabolic annotation in combination with additional approaches based on stable isotopic labeling, transcriptome sequencing and phylogenetic analysis of members of different gene families of specialized metabolism. The integration of these approaches allowed possible metabolic pathways for three classes of specialized metabolites of *P. vulgaris* to be proposed.

For hydroxycinnamates and flavonoids, the proposed metabolic pathways could be further confirmed looking at the incorporation of labeled phenylalanine. Stable isotopic labeling with $^{13}$C$_6$-phenylalanine was applied not only as a mean to measure metabolic fluxes (Antonio et al., 2013; Fernie and Morgan, 2013), but also with the aim of supporting the annotations of metabolites based on the network approach (Giavalisco et al., 2009). Preferential enrichment of the label towards hydroxycinnamates, and a minor incorporation of $^{13}$C$_6$-phenylalanine into flavonoids was shown. This result could be interpreted as a significantly higher flux towards the latter, and could be anticipated considering the relationship between hydroxycinnamates and lignin biosynthesis, one of the largest carbon sinks in most land plants (Wang et al., 2018). The lack of incorporation of the labeled phenylalanine into saponins is probably due to the different biosynthetic origin of these molecules. Triterpene saponins are derived from the mevalonate isoprenoid pathway, and their diversity is achieved through subsequent modifications of aglycone downstream of 2,3-oxidosqualene cyclization (Seki et al., 2015). Further support to our proposed pathways for saponin biosynthesis in *Phaseolus* would therefore require the use of a different labeled precursor (e.g., labeled mevalonate; Trojanowska et al., 2001).

Our study showed that hydroxycinnamates, flavonoids and triterpene saponins have highly distinct accumulation profiles across the four sampled tissues. A wide variation in these three classes was also observed across the accessions. Although two of the accessions analyzed here (MD and MW) were selected to maximize genetic diversity.
(from a larger panel described in Bellucci et al., 2014), we anticipate that a much wider metabolic diversity in specialized metabolites would emerge from profiling larger collections of accessions (ongoing in the Bean Adapt project; www.beanadapt.org/site). In a recent study, the profiles of over 300 mass signals from trifoliate leaves of 29 accessions of *P. vulgaris* (a set including domesticated and wild accessions from the Mesoamerican, Andean and Inca gene pool) efficiently discriminated different *Phaseolus* gene pools and clustered the wild forms away from the domesticated accessions. The metabolic data were highly consistent with the genomic differentiation of the accessions. In particular, the combination of the metabolic profiles with the genetic data showed that the production of flavonoids (kaempferol and luteolin) was associated with the intraspecific differentiation and to the species radiation across the whole geographical distribution of *Phaseolus* (Rendón-Anaya et al., 2017). Diversity of flavonoids in *P. vulgaris* could have a functional significance and possibly also reflects the role that these metabolites have in the interaction with rhizobia and establishment of nodulation (Ng et al., 2015).

With the recent developments in next generation sequencing and popularization of RNA-seq, it became feasible to perform large experiments integrating metabolomics and transcriptomics data. This combination is a powerful means of associating the phenotypic metabolic data with changes in gene expression which has been extensively used to investigate metabolic natural variation in a wide range of species and conditions (Alseekh et al., 2018).

**Figure 10.** Heatmap of metabolite–transcript correlations. Pearson correlations were calculated between the expression level of a subset of putative biosynthetic genes (selected from the phylogenetic analyses, x-axis) with the accumulation of flavonoids, phenylpropanoids and saponins on selected subnetworks (y-axis). High positive correlations are shown in red, while high negative (inverse) correlations are depicted in blue. Columns have been clustered according to Euclidean distance. The colors in the legend represent metabolites belonging to the different subnetworks assigned in Table S1 including putative flavonoids (subnetwork B12, B22 and B7), putative hydroxycinnamates (subnetworks B39 and MS30) and putative saponins (subnetworks B2, B3 and B11).
The Plant Journal

Secondary metabolism of common beans

© 2018 The Authors.
The Plant Journal published by John Wiley & Sons Ltd and Society for Experimental Biology.,
The Plant Journal, (2019), 97, 1132–1153

2015; Li et al., 2015, 2016; Schulz et al., 2015). One important limitation of correlation-based integration methodologies, such as the one used in this work is that, while easy to establish, they usually detect a great deal of false positives (Redestig and Costa, 2011). Hence the importance of including as many constraints as possible, restricting data to relevant classes of enzymes, for example, and analyzing their correlation to specific groups of metabolites. We have used here, as an additional constraint, the criteria of phylogenetic proximity to well characterized genes of specialized metabolism of several legume species (Medicago, soybean, Lotus, Glycyrrhiza, peanut, chickpea, lentil and pea) in order to restrict the list of putative UGT and CYP genes involved in the reconstructed pathways of bean phenylpropanoids, flavonoids and triterpenoid sapogenins. Additionally, several hypothesis can be derived from the phylogentic relationships of candidates and characterized genes. In Arabidopsis, it has been suggested that sugar donor specificity was gained after earlier diversification based on the sugar attachment position: the F3GT cluster contains genes Fd3GlcT (At5g17050), Fd3RhaT (At1g30530) and Fd3AraT (At5g17030), which transfer glucosyl, rhamnosyl and arabinosyl units, respectively (Yonekura-Sakakibara and Hanada, 2011). Multiple candidates found in common bean, Phvul.006G201100 for example, may fulfill this prediction as it was found to be closer to VmF3GalT from Vigna mungo and could similarly represent a flavonoid 3-O-galactosyltransferase (Mato et al., 1998). It is also interesting to note that the clade containing GmSGT3 and its orthologs represents a sister clade of the one containing flavonoid-related F3GGTs, suggesting that these two classes may have evolved from closely related ancestral genes. Indeed, the orthogroup including GmSGT3 and all its orthologs among our common bean candidates also includes orthologs in several other species known not to accumulate triterpenoid sapogenins (Figure S7, SGT3). Furthermore, from the known phylogenic structure of UGT, two sister clades including Phaseolus candidates in chromosome 6 and chromosome 10 may be predicted to be involved in the transfer of different sugar moieties to the same acceptor. A very similar structure can be also observed for the clade including soybean GmSGT2 and involved in transfer of a galactosyl group from UDP-galactose to soyasapogenol B monoglucuronide (Shibuya et al., 2010). In this clade, two candidates localized in the same region of chromosome two (Phvul.002G016400 and Phvul.002G016500) cluster together with GmSGT2, while three other candidates in a sister clade are co-localized on chromosome 7 (Phvul.007G020600, Phvul.007G020700 and Phvul.007G020800).

In conclusion, the data provided here represents an example of how the integration of metabolic profiling, isotope labeling, expression data and phylogenetic inference can lead to the reconstruction of specialized metabolic pathways, therefore helping to elucidate part of the tissue- and accession-specific metabolic diversity in common bean. They furthermore act as a powerful resource foundation that will greatly enable the study of these and indeed other metabolic pathways in this crop species.

EXPERIMENTAL PROCEDURES

Plant growth

Seeds from three accessions of common bean (P. vulgaris L.) were grown under controlled greenhouse conditions (relative humidity 70%, average day/night temperature 25°C, 12/12 h day/night cycle). Different tissues were harvested, and immediately snap-frozen in liquid nitrogen. Harvested tissues included the first trifoliate fully expanded leaf, flowers in the R2 stage, 10-day-old pods counting from the beginning of R3 stage, roots of plants grown on sand until bolting of the first trifoliate leaf and seedlings harvested 5 d after germination. The frozen samples were subsequently powdered in a Mixer Mill (Retsch, MM 300, https://www.retsch.com) using 5 mm id steel balls (Th. Geyer Berlin, https://www.thgeyer.com/en/) for 1 min at 25 Hz, and stored at −80°C.

Metabolite profiling

Metabolite extraction for non-targeted metabolite profiling was performed as described by Tohge and Fernie (2010). High-resolution mass spectrometry was performed on a Thermo Exactive (https://www.thermofisher.com/de/de/home.html) machine following chromatographic separation in an Waters (http://www.waters.com/waters/home.htm?locale=e) Acquity ultraformance liquid chromatography (UPLC) machine using a Waters column HSS T3 C18 (100 mm l × 2.1 mm i.d. × 1.8 μm particle size), with column temperature control at 40°C. A 20 min elution gradient at 0.4 mL min⁻¹ using UPLC MS-grade water + 0.1% formic acid (A) and acetonitrile (UPLC MS-grade) + 0.1% formic acid (B) was applied as follows: 1 min at isocratic condition with 1% (B), followed by a 10 min gradient from 1 to 40% (B), 2 min gradient from 40 to 70% (B), 2 min gradient from 70 to 100% (B), 1 min at isocratic condition with 100% (B), 1 min gradient from 100% to 1% (B) and 3 min at isocratic condition with 1% (B). Recorded mass range was set from m/z 100 to m/z 1500. Ionization was performed using an electrospray ionization (ESI) source with capillary conditions set to 3 kV and 200°C, drying gas at 350°C, sheet gas flow and auxiliary gas flow at 60 and 20 U, respectively, and skimmer and tube lens voltages at 25 and 130 V, respectively. Data-independent MSMS spectra were obtained from pooled samples combining all three genotypes for each tissue at three different collision energies: 10, 20 and 40 V.

13C labeling and data analysis

Leaves of the three accessions were used for 13C labeling experiment using 13C⁰-phenylalanine (Cambridge Isotope Laboratories, https://www.isotope.com/). Leaf discs were harvested using a leaf disk cutter with 19 mm of internal diameter, previously selected for producing approximately 50 mg samples. Leaf discs were cut directly inside of an incubation media consisting of 10 mM MES-KOH (pH 6.5) containing 10 mM of one of the following substrates: 13C⁰ (ring) phenylalanine or unlabeled phenylalanine. Samples were kept in the incubation medium for 4 h after which they were immediately snap-frozen and prepared for metabolite profiling as
described above. The experiment was repeated in triplicate and data evaluation was carried out by directly calculating the ratios between the peak areas of each m/z previously described and the expected enriched m/z, considering that the incorporation of each unit of phenylalanine corresponds to an m/z shift of +6.

RNA sequencing

Total RNA was extracted with TRIzol (Invitrogen, https://www.thermofisher.com/de/de/home.html) and RNA purity checked with a BioAnalyzer. The cDNA libraries were constructed following illumina standard protocols and paired-end sequenced on a illumina HiSeq 3000 machine by the Max Planck Genome Centre in Cologne, Germany. Samples included the same material utilized for metabolomics analysis of AD accession flowers, seedlings, leaves and pods, as well as Mesoamerican seedlings, leaves and pods. As for the metabolomics, all samples were analyzed in triplicate. RNA-seq reads were aligned to the P. vulgaris genome v2.1 downloaded from Phytozome v12.1 (http://phytozome.jgi.doe.gov/). Data processing and analysis were performed using the LSTrAP workflow (Proost et al., 2017), which included all steps further described. Adapter sequences were removed from fastq files by Trimomatic (Bolger et al., 2014), and aligned to the genome using Bowtie 2 (Langmead and Salzberg, 2012) and TopHat 2 (Kim et al., 2013). Read counts aligned to each annotated gene were computed with HTSeq (Anders et al., 2015). The results were passed through LSTrAP quality control and TPM normalized. The final data were used to calculate coexpression network and coexpression clusters based on Pearson’s correlation and MCL clustering, respectively. Additionally, LSTrAP submitted sequences to InterProScan (Jones et al., 2014) for functional analysis of protein sequences. For differential gene expression, read counts from HTSeq were analyzed using the R package DESeq 2 (Love et al., 2014). Genes were considered differentially expressed based on a P-value corrected by the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) below 0.05.

Phylogenetic analysis

Amino acid sequences for the genes of interest were obtained from the P. vulgaris genome v2.1 available in the web platform Phytozone v12. These data were further combined with sequences from NCBI and Plaza databases (Proost et al., 2015) of previously characterized genes involved in targeted pathways of different legume species and Arabidopsis. Evolutionary analyses were conducted in MEGA v6.06 (Tamura et al., 2013). Protein sequences were aligned by the algorithm MUSCLE (Edgar, 2004) and the alignments were then used to construct phylogenetic trees. The evolutionary history was inferred by using the maximum likelihood method based on the Jones–Taylor–Thornton (JTT) matrix-based model (Jones et al., 1992). Initial trees for the heuristic search were obtained automatically by applying neighbor-joining and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with the superior log likelihood value. The trees were drawn to scale, with branch lengths measured in the number of substitutions per site. All positions containing gaps and missing data were eliminated from the alignments and the bootstrap was set to 1000 replicates.

Orthologous analysis

Orthologous analysis was performed using the software OrthoFinder (Emms and Kelly, 2015) within LSTrAP pipeline. The following species were included for the comparison and its genomes were downloaded from Phytozone v12.1: Arabidopsis thaliana, Brassica oleracea, Brassica rapa, Capsicum annuum, Cicer arietinum, Fragaria vesca, Glycine max, Linum usitatissimum, Lotus japonicas, Lupinus angustifolius, Medicago truncatula, Oryza sativa, Phaseolus vulgaris, Solanum lycopersicum, Sorghum bicolor, Trifolium pretense, Triticum aestivum, Vigna angularis, Vigna radiata, Vigna unguiculata, Zea mays.

ACCESSION NUMBERS

Germplasm used included accessions number G22837 (MW), PI300668 (MD) and G19833 (AD, reference genome).

ACKNOWLEDGEMENTS

LPS gratefully acknowledges the scholarship granted by the Brazilian National Council for Scientific and Technological Development (CNPq) and support from the Max Planck Society and IMPRS-PMPG program. Work on common bean in the Fernie laboratory is sponsored by grants from the ERA-NET for Coordinating Action in Plant Sciences-2nd ERA-CAPScall, BEAN_ADAPTProject. We are grateful to Dr Saleh Alseekh and Valerio Di Vittori for their help in obtaining orthologous analysis data.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

SUPPORTING INFORMATION

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

Figure S1. PCA plot of the complete metabolite datasets.

Figure S2. PCA plots of metabolites for Andean domesticated accession.

Figure S3. PCA plots of metabolites for Mesoamerican domesticated accession.

Figure S4. PCA plots of metabolites for Mesoamerican wild accession.

Figure S5. Transcripts PCA of Andean domesticated and Mesoamerican wild samples.

Figure S6. PCA plot of the metabolite datasets excluding roots.

Figure S7. Gene count across multiple species for each orthogroup including common bean candidate genes.

Table S1. Metabolite annotation and dataset from MS-DIAL and MS-FINDER.

Table S2. Summary of common reactions of specialized metabolism.

Table S3. Read counts from RNA-seq data.

Table S4. TPM normalized RNA-seq data.

Table S5. Results of Gene Ontology enrichment analysis.

Table S6. List of genes used for phylogenetic analyses.

REFERENCES

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

Afendi, F.M., Okada, T., Yamazaki, M., Hira-Morita, A., Nakamura, Y., Nakamura, K., Ikeda, S., Takahashi, H., Altaf-Ul-Amin, M. and Darusman, L.K. (2018) KNAPsAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. Plant Cell Physiol. 59, e1-e1.

Aguilera, Y., Liebana, R., Hernández, T., Rebollo-Hernandez, M., Sanchez Puelles, C., Benitez, V. and Martin-Cabrejas, M.A. (2014) Effect of light illumination on the content of melatonin, phenolic compounds, and antioxidant activity during germination of lentils (Lens culinaris L.) and kidney beans (Phaseolus vulgaris L.). J. Agri. Food Chem. 62, 10736–10743.
Secondary metabolism of common beans

1151

Dong, M., He, X. and Liu, R.H.

Cavill, R., Jennen, D., Kleinjans, J. and Bried

Dixon, R.A., Dey, P.M. and Whitehead, I.M.

Bitocchi, E., Rau, D., Bellucci, E., Rodriguez, M., Murgia, M.L., Gioia, T.,

The Plant Journal

©

Alseekh, S., Tohge, T., Wendenberg, R.

Bitocchi, E., Bellucci, E., Giardini, A.

J. Agric. Food Chem.

©

polyploidy and hybridization for crop improvement.

L.) is revealed by sequence data.

Hiratani, M., Kuroda, R., Suzuki, H. and Yoshikawa, T.

Proc. Natl Acad. Sci. USA, 109, E788–E796.

Bitocchi, E., Bellucci, E., Giardini, A. et al. (2013) Molecular analysis of the parallel domestication of the common bean (Phaseolus vulgaris) in Mesaoamerican and the Andes. New Phytol. 197, 300–313.

Biotic stress in common bean: root transcript and metabolic responses.

He, J., Benedito, V.A., Wang, M., Murray, J.D., Zhao, P.X., Tang, Y. and Udvardi, M.K. (2009) The Medicago truncatula gene expression atlas web server. BMC Bioinformatics, 10, 441–441.

Hamberger, B. and Bak, S. (2013) Plant P450s as versatile drivers for evolution of species-specific chemical diversity. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 368, 20120428.

He, J., Benedito, V.A., Wang, M., Murray, J.D., Zhao, P.X., Tang, Y. and Udvardi, M.K. (2009) The Medicago truncatula gene expression atlas web server. BMC Bioinformatics, 10, 441–441.

Hernández, G., Ramírez, M., Valdés-López, O. et al. (2007) Phosphorus stress in common bean: root transcript and metabolic responses. Plant Physiol. 144, 752–767.

Hirohata, A., Hodge, D., Jackson, R.G., George, G.L., Elias, L., Lim, E.K., Vaistij, P.E. and Bowles, D.J. (2008) The glucosyltransferase UGT72E2 is responsible for monoglucosyl d-glucosylnaturosyltransferase from hairy root cultures of Scutellaria baicalensis. Planta, 210, 1006–1013.

Horai, H., Arita, M., Kanaya, S. et al. (2010) MassBank: a public repository for sharing mass spectral data for life sciences. J. Mass Spectrom. 45, 703–714.

Hu, J., Chen, G., Zhang, Y., Cui, B., Yin, W., Yu, X., Zhu, Z. and Hu, Z. (2015) Anthocyanin composition and expression analysis of anthocyanin biosynthetic genes in kidney bean pod. Plant Physiol. Biochem. 97, 304–312.

Hungria, M., Joseph, C.M. and Phillips, D.A. (1991) Anthocyanidins and flavonols, major nod gene inducers from seeds of a black-seeded common bean (Phaseolus vulgaris L.). Plant Physiol. 97, 751–758.

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) A new approach to protein fold recognition. Nature, 358, 86–99.

Jones, P., Messner, B., Nakajima, J.-I., Schaffner, A.R. and Saito, K. (2003) UGT73C9 and UGT78D1, glycosyltransferases involved in flavonol glycoside biosynthesis in Arabidopsis thaliana. J. Biol. Chem. 278, 43910–43918.

Jones, P., Binns, D., Chang, H.-Y. et al. (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics, 30, 1236–1240.

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

Lai, Z., Tsgawa, H., Wohlgemuth, G. et al. (2017) Identifying metabolites by integrating metabolome databases with mass spectrometry chemoinformatics. Nat. Methods, 15, 53.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. Nat. Methods, 9, 357–359.

Lanot, A., Hodge, D., Jackson, R.G., George, G.L., Elias, L., Lim, E.K., Vaistij, P.E. and Bowles, D.J. (2008) The glucosyltransferase UGT72E2 is responsible for monoglucosyl d-glucosylnaturosyltransferase from hairy root cultures of Scutellaria baicalensis. Planta, 210, 1006–1013.

Horai, H., Arita, M., Kanaya, S. et al. (2010) MassBank: a public repository for sharing mass spectral data for life sciences. J. Mass Spectrom. 45, 703–714.
Perez del Canto, G., Garcia-Calderon, M., Sanchez, D.H., Udvardi, M.K., Kopka, J., Marquez, A.J. and Bettii, M. (2013) Transcriptomic and metabolic changes associated with photorespiratory ammonium accumulation in the model legume Lotus japonicus. Plant Physiol. 162, 1834-1844.

Plaschke, T., Castillo, S., Villarin-Briones, A. and Ortesi, M. (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics, 11, 395.

Pollier, J., Morreel, K., Geelen, D. and Goossens, A. (2011) Metabolite profiling of triterpene saponins in Medicago truncatula hairy roots by liquid chromatography fourier transform ion cyclotron resonance mass spectrometry. J. Nat. Prod. 74, 1462-1476.

Proost, S., Van Bel, M., Vaneechoute, D., Van de Peer, Y., Inze, D., Mueller-Roever, B. and Vandepoele, K. (2015) PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Res. 43, D974-D981.

Proost, S., Krawczyz, A. and Mutvili, M. (2017) LSTRAP: efficiently combining RNA sequencing data into co-expression networks. BMC Bioinformatics, 18, 444.

Redestig, H. and Costa, I.G. (2011) Detection and interpretation of met abolite transcript coresponses using combined profiling data. Bioinformatics, 27, 1357-1365.

Rendon-Anaya, M., Montero-Vargas, J.M., Saburido-Avila, S. et al. (2017) Gening and diversification of the origin and domestication of common bean unveils its closest sister species. Genome Biol. 18, 60.

Ryder, T.B., Cramer, C.L., Bell, J.N., Robbins, M.P., Dixon, R.A. and Lamb, C.J. (1984) Elicitor rapidly induces chalcone synthase mRNA in Phaseolus vulgaris cells at the onset of the phytoalexin defense response. Proc. Natl Acad. Sci. USA, 81, 5724-5728.

Ryder, T.B., Hedrick, S.A., Bell, J.N., Liang, X., Clouse, S.D. and Lamb, C.J. (1987) Organization and differential activation of a gene family encoding the plant defense enzyme chalcone synthase in Phaseolus vulgaris. Mol. Gen. Genet. 210, 219-233.

Saito, K., Yonekura-Sakakibara, K., Nakabayashi, R., Higasi, Y., Yamazaki, M., Tohge, T. and Fernie, A.R. (2013) The flavonoid biosynthetic pathway in Arabidopsis: structural and genetic diversity. Plant Physiol. Biochem. 72, 21-34.

Sanchez, D.H., Lippold, F., Redestig, H., Hannah, M.A., Erban, A., Kramer, U., Kopka, J. and Udvardi, M.K. (2008) Integrative functional genomics of salt acclimatization in the model legume Lotus japonicus. Plant J. 53, 973-984.

Sanchez, D.H., Pleckenstain, F.L., Escaray, F., Erban, A., Kramer, U.T.E., Udvardi, M.K. and Kopka, J. (2011) Comparative ionomics and metabolomics in extremophile and glycolytic Lotus species under salt stress challenge the metabolic pre-adaptation hypothesis. Plant Cell Environ. 34, 655-671.

Sawaya, M., Ono, E., Takagi, K. et al. (2012) The Sp-1 glycosyltransferase locus regulates structural diversity of triterpenoid saponins of soybean. Plant Cell 24, 2123-2138.

Schmutz, J., McClean, P.E., Mamidi, S. et al. (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat. Genet. 46, 707-713.

Schoenbohm, C., Martens, S., Eder, C., Forkmann, G. and Weisshaar, B. (2006) Identification of the Arabidopsis thaliana Flavonoid 3’-Hydroxylase Gene and Functional Expression of the Encoded P450 Enzyme. Biol. Chem. 381(8), 749-753.

Schwahn, K., de Souza, L.P., Fernie, A.R. and Tohge, T. (2014) Metabolomic-assisted refinement of the pathways of steroidal glycoalkaloid biosynthesis in the tomato clade. J. Integr. Plant Biol. 56, 864-875.

Schulz, E., Tohge, T., Zuther, E., Fernie, A. and Hincha, D. (2015) Natural variation in flavonol and anthocyanin metabolism during cold acclimation in Arabidopsis thaliana accessions. Plant Cell Environ. 38, 1658-1672.

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

Seki, H., Sawai, S., Ohyama, K. et al. (2011) Triterpenes: plant functional genomics in licorice for identification of CYP72A154 involved in the biosynthesis of glycyrrhizin. Plant Cell 23, 4112-4123.
Secondary metabolism of common beans 1153

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

Shaiyulal, N.F.M., Muhlemann, J.K., Woodworth, B.D., Van Meerkercke, A., Verdonk, J.C., Ramirez, A.A., Haring, M.A., Dudareva, N. and Schuurink, R.C. (2016) COCOMOTDow-regulation activates anthocyanin biosynthesis in Petunia. Plant Physiol. 170, 717–731.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504.

Shibuya, M., Hoshino, M., Katsube, Y., Hayashi, H., Kushiro, T. and Ebizuka, Y. (2006) Identification of β-aminor and sophoradiol 24-hydroxylase by expressed sequence tag mining and functional expression assay. FEBS J. 273, 948–959.

Shibuya, M., Nishimura, K., Yasuyama, N. and Ebizuka, Y. (2010) Identification and characterization of glycosyltransferases involved in the biosynthesis of soyasaponin I in Glycine max. FEBS Lett. 584, 2258–2264.

Smith, C.A., O’Maille, G., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R. and Siuzdak, G. (2005) METLIN: a metabolite mass spectral database. Ther. Drug Monit. 27, 747–751.

Takahashi, R., Dubouzet, J.G., Matsumura, H., Yasuda, K. and Iwashina, T. (2010) A new allele of flower color gene W1 encoding flavonoid 3′-hydroxylase is responsible for light purple flowers in wild soybean Glycine soja. BMC Plant Biol. 10, 155.

Tamara, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30, 2725–2729.

Tamura, K., Seki, H., Suzuki, H., Kojoma, M., Saito, K. and Muranaka, T. (2017) CYP716A179 functions as a triterpene C-28 oxidase in tissue-cultured stolons of Glycyrrhiza uralensis. FEBS Lett. 584, 2258–2264.

Tohge, T. and Fernie, A.R. (2010) Combining genetic diversity, informatics and metabolomics to facilitate annotation of plant gene function. Nat. Rev. Genet. 11, 897–907.

Wang, P., Gun, L., Jaini, R., Klempien, A., McCoy, R.M., Morgan, J.A., Dudareva, N. and Chapple, C. (2018) A 13C isotope labeling method for the measurement of lignin metabolic flux in Arabidopsis stems. Plant Methods 14, 51.

Wink, M. (2013) Evolution of secondary metabolites in legumes (Fabaceae). S. Afr. J. Bot. 89, 164–175.

Xu, G., Cai, W., Gao, W. and Liu, C. (2016) A novel glucuronosyltransferase has an unprecedented ability to catalyse continuous two-step glucuronosylation of glycyrrhetinic acid to yield glycyrrhizin. New Phytol. 212, 123–135.

Yano, R., Takagi, K., Takada, Y. et al. (2017) Metabolic switching of astringent and beneficial triterpenoid saponins in soybean is achieved by a loss-of-function mutation in cytochrome P450 72A69. Plant J. 89, 527–539.

Yonekura-Sakakibara, K. and Hanada, K. (2011) An evolutionary view of functional diversity in family 1 glycosyltransferases. Plant J. 66, 182–193.

Yonekura-Sakakibara, K., Tohge, T., Niida, R. and Saito, K. (2007) Identification of a flavonol 7-O-hamnosyltransferase gene determining flavonoid pattern in Arabidopsis by transcriptome coexpression analysis and reverse genetics. J. Biol. Chem. 282, 14932–14941.

Yonekura-Sakakibara, K., Fukushima, A., Nakabayashi, R. et al. (2012) Two glycosyltransferases involved in anthocyanin modification delineated by transcriptome independent component analysis in Arabidopsis thaliana. Plant J. 69, 154–167.

Yonekura-Sakakibara, K., Nakabayashi, R., Sugawara, S., Tohge, T., Ito, T., Koyanagi, M., Kitajima, M., Takayama, H. and Saito, K. (2014) A flavonoid 3-O-glucoside-2″-O-glucosyltransferase responsible for terminal modification of pollen-specific flavonoids in Arabidopsis thaliana. Plant J. 79, 769–782.

Zhang, J.-Y., Cruz De Carvalho, M.H., Torres-Jerez, I., Kang, Y.U.N., Allen, S.N., Huhman, D.V., Tang, Y., Murray, J., Sumner, L.W. and Udvardi, M.K. (2014) Global reprogramming of transcription and metabolism in Medicago truncatula during progressive drought and after rewatering. Plant Cell Environ. 37, 2553–2576.

Zhong, S., Fei, Z., Chen, Y.-R. et al. (2013) Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. Nat. Biotechnol. 31, 154.