Global dissection of the BAHD acyltransferase gene family in soybean: Expression profiling, metabolic functions, and evolution

Muhammad Zulfiqar Ahmad

Anhui Agricultural University <https://orcid.org/0000-0001-7201-8937>

Xiangsheng Zeng

Anhui Agricultural University

Qiang Dong

Huazhong Agriculture University

Sehrish Manan

Huazhong Agriculture University

Huanan Jin

Huazhong Agriculture University

Penghui Li

Anhui Agricultural University

Xiaobo Wang

Anhui Agricultural University

Vagner A. Benedito

West Virginia State University

Jian Zhao (✉ jianzhao@ahau.edu.cn)

Anhui Agricultural University <https://orcid.org/0000-0002-4416-7334>

Research article

Keywords: BAHD, Domestication, Evolution, Gene expression, Genome, Positive Selection, Soybean, Stress

Posted Date: March 31st, 2020

DOI: <https://doi.org/10.21203/rs.2.21482/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Members of the BAHD acyltransferase (ACT) family play important roles in plant defence against biotic and abiotic stresses. A better understanding of the functions that specific members of this family play in stress defence can lead to better breeding strategies for stress tolerance. Previous genome-wide studies explored other acyltransferase families, but so far not a single study has been published on genome-wide or positive selection analyses of the BAHD genes in *Glycine max*.

Results: A total of 103 genes of the BAHD family (GmACT genes) were identified from the soybean genome, which could be grouped into four phylogenetic clades (I-IV). Clade III was further divided into two sub-clades (IIIA and IIIB). In each clade, the gene structure and functional motifs were relatively well conserved. These ACT genes were unequally distributed on all 20 chromosomes, and 16 paralogous pairs were found within the family. Positive selection analysis revealed amino acids under strong positive selection, which suggests that the evolution of this gene family modulated soybean domestication. Some GmACT genes showed expression specific under specific conditions, while others showed constitutive expression in all soybean tissues or conditions analysed. Most of the expression of ACT genes in soybean was repressed with AI 3+ and fungal elicitor exposure, except for GmACT84, which expression increased in these conditions 2- and 3-fold, respectively. The promoter of GmACT84 contains the maximum number of stress-responsive elements among all GmACT genes and it is especially enriched in MYB-related elements.

Conclusions: This study provides a genome-wide analysis of the BAHD gene family and assessed their expression profiles in soybean. We found evidence of a strong positive selection of GmACT genes. Our findings will help efforts of functional characterisation of ACT genes in soybean in order to discover their involvement in growth, development, and defence mechanisms.

Background

A large number of secondary metabolites in plants play significant roles in their response to environmental challenges [1]. Specialised metabolites in a diverse range of plant species are synthesised through the alteration of basic structures by different reactions, such as acylation [2], methylation [3], and glycosylation [4]. For example, the modification of phenolic compounds through acylation enhances their functions *in planta*. Indeed, acylation to flavones, isoflavone or anthocyanins has a substantial impact on their solubilization, stabilization, transport, vacuolar uptake, and colour modulation [5-7]. The sinapoyl moiety conjugation to malate enhances the anti-UV property of flavonoids [8]. Avenacin acylation plays a demonstrated role in plant defence [9].

BAHD comprises of a large acyltransferase (ACT) gene family found in plants and prokaryotes. The family acronym originates from the first four acyltransferases identified [benzyl alcohol *O*-acetyltransferase BEAT from *Clarkia breweri* (order Myrtales) [10]; anthocyanin *O*-hydroxycinnamoyltransferase AHCT from *Gentiana triflora* (Gentianales) [11]; *N*-

hydroxycinnamoyl/benzoyltransferase HCBT from *Dianthus caryophyllus* (Caryophyllales) [12]; and deacetylindoline 4-*O*-acetyltransferase DAT from *Catharanthus roseus* (Gentianales) [13]. They use CoA conjugates (e.g., acetyl-, malonyl-, β -phenylalanine, tiglyl, anthraniloyl, and benzoyl- moieties) as donors to modify different acceptor compounds [2, 14]. The *O*- or *N*-acylation reaction is carried out by BAHD acyltransferases via the acceptor's hydroxyl (OH) or amine (NH₂) group with different CoA donor substrates [2]. The synthesis of a variety of polymers and secondary metabolites like lignin, volatiles, suberin, cutin, pigments, and defence-related compounds in plants involves these enzymes [14]. Furthermore, several plant BAHD genes that catalyse the formation of diverse metabolites have been characterised. For example, several plant BAHD enzymes, such as GmIMaT1, GmIMaT3 [15], GmMT7 [16], MtMaT1, 2, 3 [17], MaT4, 5, 6 [7] was shown in the heterologous system *E. coli* to synthesise malonylated flavones, isoflavones or anthocyanins. Moreover, HQT was demonstrated to synthesise the antioxidant chlorogenic acid [18], hydroxycinnamoyl-CoA:hydroxyphenyllactate transferases to make rosmarinic acid [19], the hydroxycinnamoyl-CoA:glycerol transferase to produce water-soluble antioxidant hydroxycinnamate glycerol esters [20], and hydroxycinnamoyl/benzoyl-CoA:anthranilate transferase (HCBT) to synthesise therapeutic benzoyl and hydroxycinnamoyl anthranilates [21]. The sequencing and annotation of many plant genomes revealed the presence of BAHD family genes, not only in the models *Arabidopsis thaliana* and *Oryza sativa* but also in orphan species, such as cardoon (*Cynara cardunculus*, Asteraceae). Their respective genomes contain 64, 119, and 69 BAHD genes [14, 22, 23]. So far, despite the centrality of these enzymes to plant specialised metabolism, there has been no genome-wide analysis of the BAHD family in the model legume crop, *Glycine max*.

To date, only two BAHD genes were reported in soybean: *GmIMaT1* and the two allelic forms of *GmIMaT3* [15] called *GmMT7* [16] and *GmIF7MaT* [24], which respectively malonylate the isoflavones daidzin, genistin, and glycitin. In this study, we have executed a global analysis of all BAHD genes in the soybean genome. The chromosomal distribution, phylogeny, gene structure, duplication events, positive selection, and transcriptional dynamics in response to different stresses in this gene family were analysed.

Results And Discussion

Identification of BAHD family members in the soybean genome

We searched genes in the soybean genome encoding proteins containing the specific BAHD catalytic motif (HXXXX, where X means any amino acid residue) and identified 103 genes (Fig. 2A; Additional files 1: File S3). The predicted BAHD proteins showed large disparity in length, from 274 (GmACT31) to 511 (GmMaT2) amino acid residues. However, most of them (70 out of the 103 members) ranged between 450-511 residues. The range of predicted molecular weight is from 30.93-56.70 kDs, whereas the majority of the genes code a predicted product ranging 45-52 kDs, which falls within the values of characterised BAHD proteins [14]. The predicted pI value of soybean BAHD proteins ranged from acidic 5.05 (Glyma.10G061900_GmACT53) to basic 9.27 (Glyma.02G264600_GmACT2) (Additional files 1: File S3). The predicted subcellular localization indicates that soybean BAHD proteins could be localized to several

organelles of the cell. However, GmIMaT1 and GmIMaT3 are primarily localised to the endoplasmic reticulum and cytosol, respectively [15]. Most BAHD proteins were predicted to be localized to the cytosol (37/103), followed by mitochondria (27/103), and the plasma membrane (20/103), whereas only 2 proteins are predicted to localize to the nucleus (Additional files 1: File S3). This result is expected, since many of the reactions of specialised metabolism involving BAHD enzymes occur in the cytosol and mitochondria prior to the products to be exported wither to an organelle for accumulation or outside the cell [7, 15, 16, 25, 26].

Phylogenetic relationships of the soybean BAHD members

The evolutionary relationship of the soybean BAHD genes was explored by constructing a phylogenetic tree to compare them with 43 functionally characterised BAHD genes from different plant species (Fig. 1). On the basis of phylogenetic relationships, which defines substrate donor and acceptor of the characterized proteins in the family, the 103 soybean BAHD genes were classified into four major clades (I-IV) and clade III was further subdivided into two subclades (IIIA and IIIB) (Fig. 1). Our results based on the phylogenetic relationships of the characterised soybean ACT proteins would predict their correct enzymatic functions.

Clade I contains 20 soybean BAHD genes along with 16 characterised genes from different plant species (Fig. 1). Most members of this clade code for benzoyltransferases, alkyl-hydroxycinnamate ester, or cell-wall acyltransferases. They use benzoyl as a donor substrate to form benzyl benzoates, such as CbBEBT, NtBEBT [27], or salicyl benzoate (PtSABT) and involved in plant defence mechanisms against herbivory [28]. Many characterised BAHD enzymes of clade I are responsible for the formation of benzenoid compounds [27-29]. Among them, PtSABT from *Populus trichocarpa* uses different CoA thioesters and alcohols as a donor to catalyse the production of salicyl benzoate for the synthesis of a variety of acetates, benzoates and cinnamates [28]. PtBEBT from *Populus trichocarpa* is more selective to produce benzyl benzoate [28] and takes part in the synthesis of salicinoids. CbBEBT produces benzyl benzoate in *Clarkia breweri* that composes its floral aroma. The *CbBEBT* gene is highly expressed in the flower, and especially in the stigma [27]. These BEBT genes code for products that have the capability to produce compounds that are important for plant defence or aroma [27-29]. Four BEBT genes from soybean (*GmACT3*, 4, 5, 20) have high similarity with the characterised BEBT genes. Similar to *CbBEBT*, *GmACT3* and 4 also showed high expression levels in the flower, leading us to hypothesize that it likely plays a function in catalyzing aroma compounds. These genes are also highly expressed in the stem and the leaf, alike *PtSABT* and *PtBEBT*, which produce defence compounds. Some members of clade I are also involved in the biosynthesis of cutin and suberin, which work as lipid barriers on epidermal surfaces of plant organs. These barriers regulate gas, water and solute fluxes [30] and play critical roles against pathogen attacks [2]. Alkylhydroxycinnamoyl ester BAHD genes, such as *AtASFT* [31], *StFHT* [32], *AtDCF* [33], and *AtFACT* [34] use feruloyl, *p*-coumaroyl, sinapoyl, and caffeoyl as a donor to form aromatic suberin, alkyl, or ω -hydroxy fatty acids, which are used for the synthesis of cutin polymers that in turn are important components of the cuticle and assist in the deposition of suberin into the inner side of the primary cell wall of secondary shoots and roots [2]. They also play a role in the synthesis of root waxes

that contain hydroxycinnamates in *A. thaliana* [34]. Ten genes from soybean fell in clade I and, therefore having the same predicted transferase function (Fig. 1; Additional files 1: File S3). Interestingly, we found in soybean only one orthologue of disinapoylspermidine (*AtSDT*; [35]) and another of coumaroylspermidine (*AtSCT*; [35]), that are annotated as *GmACT2* and *12*, respectively (Fig. 1). Meanwhile all the other genes in this clade have multiple copies or paralogous in the soybean genome. Monolignols, which participate in lignin formation, are synthesised through the phenylpropanoid pathway, primarily through the activity of hydroxycinnamoyltransferases (HCTs). The BAHD-ACTs also used hydroxycinnamic acid (HCA) as a donor substrate to esterify monolignols [36]. These cell wall-related BAHD-ACTs also belong to clade I, such as *OsPMT* [36], *BdPMT* [37], and *OsAT10* [38], which use *p*-coumaroyl-CoA as a donor substrate to form lignols and glucuronoarabinoxylan (GAX). These activities are responsible for producing the lignocellulosic biomass in grasses, which could have a major yield impact on feedstocks for making biofuels [37]. The soybean genes *GmACT1*, *GmACT13*, *GmACT14*, *GmACT16*, *GmACT17*, and *GmACT19* were also predicted to have cell wall-related functions, and they may play essential roles in the production of liquid biofuel compounds.

Clade II of the soybean BAHD family contains 26 genes that cluster together with 9 characterised genes from other plants (Fig. 1). This clade includes many genes related to shikimate/quinic hydroxycinnamoyltransferases (HCT/HQT) and acyl-CoA *N*-acyltransferases, such as hydroxycinnamoyl-CoA, spermidinehydroxycinnamoyltransferase (SHT), hydroxycinnamoyl-CoA: *N*-hydroxycinnamoyl/benzoyltransferase (HCBT), and rosmarinic acid synthase, which is alternatively called hydroxycinnamoyl-CoA:hydroxyphenyllactate hydroxycinnamoyltransferase (RAS). These members use hydroxycinnamoyl, caffeoyl or *p*-coumaroyl as donor substrates to form chlorogenic acid, hydroxycinnamoylquinic, caffeoyl, or *p*-coumaroylquinic or malate, hydroxycinnamoyl glycerol, hydroxycinnamoylshikimate, hydroxyphenyllactates, anthranamides, and spermidines [33, 39-44]. Clade II mostly contains acyl-CoA *N*-ACTs that catalyze *N*-acylation instead of *O*-acylation. These reactions catalyze hydroxycinnamoyl and amine moieties that are the entry point of the phenolamide pathway [2]. Hydroxycinnamic acid (HCA) amides are classified as phytoalexins [45]. Most of the plant *N*-HCTs that have been characterised so far are from *A. thaliana* and *Nicotiana attenuata* and involved in defence against herbivores [46]. Meanwhile, other members of the clade II, such as HCT/HQT and their relatives, produce diverse types of compounds that play protective roles against UV photooxidation [39] and pathogenic resistance [42], such as lignin, chlorogenic acid (CGA) and quinicates. CGA has been implicated in cell wall development [47] and root hair formation [48]. Seventeen soybean BAHD genes are closely related to *TpHCT2*, which transfers hydroxycinnamoyl moieties (*p*-coumaroyl and caffeoyl) to malic acid (Fig. 1). The transferase activity of *TpHCT2* is involved in phaelic acid synthesis, particularly in the leaves of red clover (*Trifolium pratense*, Fabaceae) [43]. *TpHCT2* is related to *GmACT27*, *28*, and *45*, that have high and specific expression in leaves (Fig. 1) and may play a role in phaelic acid synthesis. Meanwhile, *GmACT21*, *22*, *32*, and *33* fall closest with *TpHCT1A/1B*, which use caffeoyl or *p*-coumaroyl as a donor to form shikimates or malates [43] and play a major role in monolignol synthesis [40, 43]. Lignin precursors are synthesised through *p*-coumaroyl derivatives rather than methoxylated analogues in response to pathogen attacks [49]. *TpHCT* activity was higher in the red clover stem and capable of

transferring *p*-coumaroyl moieties from the respective CoA to shikimic acid, which is critical for the biosynthesis of monolignol lignin precursors [40, 43]. *AtHCT* expression dramatically changed the lignin composition, which demonstrates the essential function of HCT in phenylpropanoid metabolism [40, 50]. HCT homologues have been identified in several vascular plant species [2]. In conformity with other members of this group, *GmACT21*, *22*, and *32* also presented high expression levels in the stem, and they may transfer *p*-coumaroyl moieties to respective acceptors with essential roles in lignin biosynthesis. Other four members of this clade (*GmACT35*, *36*, *38*, and *39*) are closely related to SHT from *Arabidopsis thaliana*, which uses hydroxycinnamoyl as a donor substrate to form hydroxycinnamoylspermidine [39]. This compound is abundant in pollen of many species, which indicates a role of these phenylpropanoids in plant sexual reproduction [39]. Indeed, *AtSHT* expression was high and specific to the tapetum cells of the anther during early flower development in order to supply the building blocks for cell wall development in the pollen [39]. The metabolic profiling of flower bud tissues in *Arabidopsis* revealed that *AtSHT* knockout or RNAi plants had a positive correlation between acyltransferase protein accumulation and the amounts of trihydroxyferuloyl and dihydroxyferuloyl sinapoyl spermidines [39]. *GmACT35* and *38* were also more highly expressed in the flower than any other tissues and are expected to have a function in the acylation of spermidine in the pollen.

Clade III is further divided into sub-clades IIIA and IIIB (Fig. 1). Sub-clade IIIA contains 25 genes along with 19 already characterised from other species. Most of the characterised members of clade IIIA acylate phenolic glucoside anthocyanins or other flavonoids (flavones or isoflavone). Malonylation of clade IIIA ACTs depends on the position of the acceptor substrate. Clade IIIA can be divided into aliphatic and aromatic ACTs. Fourteen soybean BAHD genes fall in the aliphatic ACTs, and they are functionally annotated as isoflavonoid malonyltransferases (Additional file 1: File S3) along with the characterised *At5MaT*, *At3At1*, *At3At2* from *Arabidopsis* and *MtMaT1*, *MtMaT2* and *MtMaT4* from *Medicago truncatula* (Fig. 1), which facilitate the transport or storage of metabolites [7, 35]. Among the 14 aliphatic ACTs in soybean, only two, *GmIMaT1* and *3* [15] have been functionally characterised and play a distinct role against abiotic and hormonal stresses [15]. Their transcription patterns correlated positively with the malonylisoflavonoid composition of tissues in different parts of the plant under different stress conditions [15]. Another aliphatic group in this clade IIIA contains 6 soybean ACTs along with *MtMaT5* and *MtMaT6* [7] from *Medicago* that malonylate different anthocyanins, such as cyanidin, delphinidin, and pelargonidin conjugates at 3-*O*-glucosides. These genes are very specific for anthocyanin modification at the C3 position, and they have no activity towards C5 or C7 of anthocyanin or other flavones or isoflavones [7]. The soybean genes that closely relate to *MtMaT5* and *MtMaT6* (Fig. 1) are predicted to function in anthocyanin modification at 3-*O*-glucoside (Additional file 1: File S3). This group of clade IIIA is most probably associated with the stability of the anthocyanin glucosides since anthocyanin acylation at 3-*O* sugar with the malonyl moiety has been suggested to increase its stability [14].

Clade IIIA also contains 7 soybean BAHD-ACT genes (*GmACT52*, *57*, *61*, *63*, *72*, *76* and *77*) closely associated with the aromatic ACTs from *Arabidopsis*, *At3At1* and *At3At2* (Fig. 1; [35]). *At3At1* and *At3At2* use *p*-coumaroyl, feruloyl and caffeoyl-CoA as donors to specifically modify the C6 hydroxyl of glucose at

three anthocyanin positions, but it does not have any affinity towards sinapoyl-CoA. We predict that these 7 aromatic ACTs also prefer coumaroyl-CoA to modify the anthocyanin at the C3 position just like their closely related *Arabidopsis* orthologues (Fig. 1; Additional file 1: File S3). We did not find in the soybean genome any aromatic gene that was closely related to a gene able to modify 5-*O*-glucosides, such as *Gt5AT* that acylates the C6 hydroxyl of 5-*O*-glucosyl residues of anthocyanin with either caffeoyl, *p*-coumaroyl [25] or aliphatic anthocyanin MaTs, such as *At5MaT* (Fig. 2; [51]) and *Ss5MaT1* from *Arabidopsis thaliana* (Brassicaceae) and *Salvia splendens* (Lamiaceae), respectively (Fig. 1; [52]).

Clade IIIB contains 9 GmACT genes from soybean with 7 characterised genes (Fig. 1). This clade contains malonyltransferases (MaTs) that catalyse flavonoids with malonyl-CoA to form malonates, acetyl CoA:deacetylindoline 4-*O*-acetyltransferases (DAT) which catalyse the last step in vindoline biosynthesis [13], acylsugar acyltransferases (ASAT), which is important in the acylsucrose biosynthetic pathway and uses sucrose and acyl-CoA ester substrate specificity to contribute in different types of accumulated acylsucroses [53]. Acylsugars are polyesters of short- to medium-length acyl chains of sucrose or glucose backbones that are produced in secretory glandular trichomes of many solanaceous plants, including tomato (*Solanum* spp.). These compounds act against different insects through a mechanism involving in a multitrophic defence interaction involving their metabolism to volatile fatty acids [54]. *Ss5MaT2* also fell in clade IIIB separated from other the MaTs [14]. These BAHD-ACTs share the anthocyanin-specific motif (NYFGNC) in addition to the BAHD motif (HXXXD). The characterised member of clade IIIB (*Ss5MaT2* from *Salvia splendens*) malonylates the anthocyanin glycosides bisdemalonylsalvianin and shisonin by transferring the malonyl moiety to the C4 position of the 5-*O*-glucosyl residue [55]. The soybean genes of clade IIIB also suggest forming salvianin, which has two malonyl moieties, and malonylshisonin, and plays an important role in flower colour [55]. Although the biochemical properties of *Ss5MaT2* are similar to that of clade IIIA members, it fell into the clade with ASAT and DAT.

Clade IV contains 21 soybean BAHD genes along with five characterised genes (Fig. 1; Additional file 1: File S3). The members of this clade have varied functions. Most are predicted to function as shikimate *O*-hydroxycinnamoyltransferase or transferase family, except for *GmACT88* and *GmACT97*, which are predicted to function as an anthranilate *N*-hydroxycinnamoyl (HCBT)-related transferase, as well as *GmACT79* and *81*, which are predicted to function as a shikimate *O*-hydroxycinnamoyltransferase/quinate *O*-hydroxycinnamoyltransferases. *Gt5At* from *Gentiana triflora*, an anthocyanin aromatic acyltransferase, catalyses the transfer of hydroxycinnamic acid moieties from their CoA esters to the glycosyl groups of anthocyanins [25]. HCBT produces the aromatic amide dianthramide in carnation [44]. Anthraniloyl-CoA: methanol AT (AMAT) uses anthraniloyl-CoA as a donor substrate to acylate methanol and produce a foxy odour compound (methylantranilate) in concord grapes [56]. Moreover, *GmACT93* and *96*, which are closely related to the gene *DEFECTIVE IN CUTICULAR RIDGES* (*DCR*) from *Arabidopsis thaliana*, *Medicago truncatula*, and *Populus trichocarpa* [26] and predicted to involved in cutin formation. In the flower, DCR converts monomers into polymeric cutin, which is composed of fatty acid, glycerol and aromatic monomers [57]. Changes in epidermal cell differentiation and post-genital organ fusion were observed due to defective cuticle in DCR-deficient plants. *DCR* was

also expressed in the root cap, lateral root primordia, and developing lateral root. This expression profile points to its involvement in suberin biosynthesis [26]. Downregulation of DCR in *dcr* mutants led to excessive root branching and changes of root architecture by influencing lateral root emergence and growth [26]. All in all, these analyses help us formulate educated hypotheses about the diversity and evolution of substrate specificity within the BAHD family in soybean.

Motif analysis and gene structure

Variation in length and sequence was found in the conserved region (PF02458/IPR003480) of soybean BAHD genes (Additional files 1: File S4). Among the ten representative BAHD motifs searched and annotated through Pfam and SMART tools, three motifs (1, 4 and 6) were found related to BAHD transferase family (Additional files 1: File S4). The biological importance of these recurring motifs required an in-depth investigation. All members in clades I and II contained all 10 conserved motifs whereas clade III and IV members missed some of them (Fig. 2B). We found a typical motif pattern in most of the paralogous pairs. However, some presented differences in motif composition and arrangement, such as pairs GmACT94-99, 93-96, 31-46, 21-22, 52-72, 7-9 and 15-18 (Fig. 2B).

The architecture of soybean BAHD genes was analysed to establish general features within each clade. The exon-intron structure of each gene is shown along with the phylogenetic tree (Fig. 2C). Genes within the same clade contain almost the same length and number of exons and introns, with a few exceptions (Fig. 2C). Introns of clade I members are usually much larger than in other clades. Seventy percent of clade I members (14 out of 20) contain two exons while 15% (3/20) contain one and three exons (Fig. 2C; Additional files 1: File S3). *GmACT12* contains the longest intron (13,062 bp) in the whole set of the BAHD gene family (Fig. 2C). In clade II, 14 genes contain only one, 8 genes contain two, and 3 genes contain three exons. *GmACT21*, *22*, *32* and *33* have long introns compared to the other clade II members. Clade III genes varied in exon number, but 77.78% (28 out of 36) contains just a single exon. Two clade III members (GmACT51 and 72) contain the maximum exon number (4) in the whole soybean BAHD family. Most members of clade IV (76%) contain only one exon while the remaining ones contain two exons. Among the 16 paralogous pairs, 10 did not show remarkable differences except that for GmACT21-GmACT22 from clade II which GmACT22 has an untranslated region (UTR) longer (4982bp) while GmACT21 had (1212bp) at 5' end. The other six paralogous pairs showed small variations in exon number or intron length (Fig. 2C).

Chromosomal localization and duplication of soybean BAHD genes

A chromosome ideogram was constructed on the basis of the physical position of the soybean BAHD-ACT genes (Additional file 1: File S3). All the 20 soybean chromosomes contain BAHD-ACT genes, although the number on each chromosome varies between 1 and 15 genes per chromosome (Fig. 3). Chromosome 18 has the highest number of genes (15%) followed by chromosomes 8 (11%) and 19 (10%) (Fig. 3; Additional file 1: File S3). On the other hand, chromosomes 1, 9 and 20 contain only a single gene each (Fig. 3; Additional file 1: File S3). The positions of BAHD genes on the soybean chromosome,

often away from the centromere (Fig. 3), which explained their high recombination rates and active translation [58].

Tandem and segmental duplications expanded BAHD genes into a large family in different plant genomes [59]. Gene redundancies can be predicted by the identification of closely related paralogues in the genome whereas the accurate identification of true orthologues between species can lead to the creation of strong hypothesis of common gene functions. The soybean genome contains a high number of BAHD genes (103) as compared to 69 in *Cynara cardunculus* [23], 64 in *A. thaliana* [22], but less the 119 BAHD genes in the *Oryza sativa* genome [14]. This large gene family might be the result of the two rounds of whole-genome duplication (WGD) events in soybean that occurred 58-60 (the Papilionoidea allotetraploidization event) and 13 Mya (recent soybean allotetraploidization) [60]. The presence of paralogous genes in a plant genome is more often related to their functional category than the genetic proximity between species [61]. We found sixteen paralogous pairs in the soybean BAHD family (Additional file 2: Table S1). These paralogues are present in all four phylogenetic clades (Additional file 2: Table S1): Clade I contains 7 members; clade II has 3, clade III contains 2, and clade IV has 4 paralogous pairs (Additional file 2: Table S1). The presence of clear paralogues is evidence of a recent common ancestry and helps to elaborate strong hypotheses about the functional conservation of BAHD genes in soybean along with those already characterised in different species. The selection history of coding sequences can be assessed through the analysis of base substitutions of the protein code called the K_a/K_s ratio. To investigate the deviation of duplicated BAHD acyltransferases, we determined the non-synonymous substitutions per site (K_a), the synonymous substitutions per site (K_s), and the respective K_a/K_s ratios for each paralogous pair. The partition of each pair was estimated through the K_s values. All paralogous pairs had a K_s value from 0.09 to 0.28, which is consistent with the soybean duplication events (Additional file 2: Table S1). The selection pressure can be determined through the K_a/K_s ratio. Most of the paralogous pairs showed negative K_a/K_s ratios (Additional file 2: Table S1) suggesting that the BAHD gene family in soybean experienced strong purifying selection pressure and is bound by strong evolutionary constraints to maintain its stability [62]. Divergence time analysis showed that these paralogous genes evolved between 7.38 to 22.95 million years ago (Mya), which is consistent with the recent allotetraploidization of the species that occurred around 13 Mya (Additional file 2: Table S1).

Positive and purifying selection of BAHD genes in soybean

The K_a/K_s ratio is used as a measure of the direction and magnitude of gene selection [63]. The K_a and K_s substitution rates in this study were estimated on the basis of simple probability theory. Positive selection was considered significant with K_a/K_s ratio > 1 [64]. Comparisons of nucleotide substitution per site in the 103 BAHD genes were carried out on the basis of the K_a/K_s ratio in each individual branch of the phylogenetic tree. Maximum Likelihood computations of K_a and K_s were conducted using the HyPhy software package [65]. The analysis involved 26 nucleotide sequences from clade I, 20 from clade II, 36 from clade III and 21 from clade IV. All positions containing gaps or missing data were eliminated. The neutrality analysis based on maximum likelihood computation of K_a/K_s values of all soybean BAHD

genes revealed 16 coding sites under positive selection between codons 69- 342 in clade I, 6 coding sites between codons 33-143 in clade II, 19 coding sites in clade III and 15 sites in clade IV under positive selection, all of which showed K_a/K_s ratios > 1 (Additional file 2: Table S2). The changes in codons, as well as sequences for pairwise comparison, were estimated at the K_a and K_s basis. The variability between K_a and K_s polymorphisms at the allelic distribution level presents quite an explicit proof for positive selection [66]. The allelic distribution in various coding sites can be revealed by the strapping test of neutrality, which is applicable to genomic data containing a large number of polymorphisms, and that along with the homogeneity test, allows for comparing the frequency distribution of synonymous sites [67, 68]. The changes in the coding region during the evolutionary time under analysis showed the cumulative presence of synonymous, non-synonymous and ambiguous (indels) codons (Additional file 3: Figure S1). Synonymous and non-synonymous mutations showed low cumulative behaviour at the start site and then gradually increased during the evolutionary time analysed, whereas the opposite behaviour was observed for ambiguous codons at the start point and then acquired a static behaviour at the end of the code. The very same behaviour of synonymous, non-synonymous and ambiguous codons was found in all four clades of the BAHD family (Additional file 3: Figure S1).

We estimated the Likelihood Rate Test (LRT) through the empirical Bayes method at specific branching points and identified various diversifying selection sites. The mixed-effect model of evolution (MEME) was used to identify the diversifying selection of BAHD genes also on the basis of empirical Bayes procedure [69]. MEME identified 7, 17, 16, and 4 episodic diversifying coding sites ($p < 0.05$) in clades I-IV, respectively (Additional file 2: Table S3). The synonymous (α) and non-synonymous (β) substitution rates were calculated, and coding sites with $\beta > \alpha$ were considered as significant to determine sites under diversifying selection. The Maximum-Likelihood estimation of $\beta > \alpha$ in MEME was obtained for nucleotide position 60 and 434 in clade I, 189 and 463 in clade II, 277 and 643 in clade III, and 299 in clade IV (Additional file 2: Table S3). The prevalence of purifying or natural selection masked the episodic natural selection with the transient period of adaptive evolution. Thus, we could not identify sites under positive selection due to the sensitivity of tests used [70] and the stringent parameters imposed, such as $p < 0.05$ and the empirical Bayes factor threshold of analysis for selection of sites (> 15), so that the best outcomes were obtained when the selected branches were placed in the conserved lineage.

Ambiguities in posterior gene- and site-specific distributions were calculated through FUBAR using the MCMC approach [71, 72]. This analysis revealed the following number of sites in BAHD genes under pervasive diversifying selection: one site in clade I, four in clades II and III each, and two in clade IV (Additional file 2: Table S4). The diversifying evolutionary sites were detected through the random effect model using the rate distribution with a large number of parameters at the significance level at 95% confidence interval and $\Pr [\beta > \alpha]$ values (Additional file 2: Table S4). The class rate weight at each coding site under positive selection was calculated through a bivariate general discrete distribution. The posterior mean values ranged between 0.80-0.92, which are closer to the value of the potential scale reduction factor (Additional file 2: Table S4) and indicates MCMC convergence. Coding sites with $\beta > \alpha$ and empirical Bayes factor (EBF) > 15 were considered to be under diversifying selection. The EBF values for

each coding site under positive selection were calculated with a net effective sample size (Additional file 2: Table S4). PSRF values for each site under positive selection were <1.03, whereas the effective population size was >100. The detected selection across many coding sites was potentially improved through deducing the allocation of gene-specific selection parameters. Altogether, the coding sites under positive selection unveiled in this analysis provides strong evidence of diversifying selection in BADH genes under selection in the soybean lineage.

The evolutionary fingerprint of soybean BAHD genes

Nucleotide and amino acid substitutions in the codon model were used to identify synonymous and nonsynonymous substitutions [71, 73]. The genetic algorithm in the codon model evolution was used to identify the evolutionary fingerprint in coding sites of BAHD genes. The 9,052 models generated were used in codon model selection on the basis of likelihood log and modified Bayesian Information Criterion (mBIC). Selective effects with an exchangeable preference for particular amino acid were found in BAHD genes through a model that used the combined empirical codon and transition/transversion-related physiochemical parameters [74]. The model with log (L) value -16,360.3 was considered optimum for the amino acid substitution analysis. We observed mBIC values ranging from 33,701.1 for clade II to a maximum value of 81,506.1 for clade IV (Additional file 2: Table S5). All clades showed three-class rates for the distribution of amino acids in different classes (Fig. 5) with the estimation of a single K_a/K_s substitution rate (Additional file 2: Table S5). A genetic multi-rate model algorithm was used to analyse the class rates in order to calculate substitution rates at the amino acid level during the evolutionary time scale (Fig. 5). The substitution rate in each class was calculated through genetic models using the Stanfel class parameters (Fig. 5; [75]). The substitution rate distributed the amino acids into the three classes through evolutionary rate cluster, and the substitution pair ACGILMPSTV revealed 90% substitution, whereas DENQ had 50% substitution, and FWY and HKR presented < 50% substitution rates. The best-fit model with mBIC values defined the clustering rate, and the numerical values of corresponding rate class substitution are inferred with Maximum-Likelihood estimation, which were computed evolutionary evidence ratio for the gene [71].

Analysis of *cis*-regulatory elements in promoter regions of BAHD genes in soybean

Extensive studies on ACTs indicate their roles in plant growth, development, and stress tolerance. Many *cis*-acting regions are found in the upstream region of *GmACTs* genes according to our *in silico* analyses (Additional file 1: File S5). The 103 *GmACT* promoter regions contain *cis*-elements responsive to light (43%), hormones (29%), environmental stresses (20%), and plant growth (8%) (Fig. 4A). These regions contained transcription factor binding sites involved in response to plant hormones (Fig. 4B), biotic stresses and pathogen defence, MYB binding sites, and response to heat, low temperature, and drought (Fig. 4C). The pattern of *cis*-acting regions differed among the soybean BAHD gene promoters (Additional file 1: File S5). *GmACT3* (clade I) contains the maximum number of *cis*-elements (29), whereas *GmACT82* (clade IV) contains 26 hormone-responsive elements (Additional file 1: File S5). *GmACT34* contains the maximum number of ABA-responsive elements (8), and *GmACT45* contains 9 ethylene-responsive motifs

(Additional file 1: File S5). *GmACT84* and *96* (clade IV), along with *GmACT66* (clade III) contain the maximum number of environmental stress-response elements (18, 17, and 15, respectively) in their promoter regions. Members of clades III and IV contain the maximum number of MYB-related *cis*-elements (Additional file 1: File S5). *GmACT84* showed the most (12) MYB related elements, and also showed high expression values, up to seven-fold higher in response to Al^{3+} and two-fold higher to fungal elicitor than the control (Additional file 1: File S1c, S1d). The promoter regions of paralogous pairs were also compared. Both members of each pair showed variation in *cis*-elements related to stress or hormonal responses or other signals in their promoter regions. Some hormonal, stress or growth-related elements were often present only in one member of the pair (Additional file 1: File S5). Although the *GmACT* paralogous pairs have little variation in protein length, the sequences of their promoter regions were highly different, which suggests divergent specific roles of each member. We concluded from this analysis that *GmACT*s likely play vital roles in plant stress perception, signalling, or biotic and abiotic stress defence.

Expression patterns of soybean BAHD genes in different tissues

Expression data for the 103 soybean BAHD genes in different soybean tissues were extracted from Phytozome (Fig. 6; Additional file 1: File S1a). Most of the genes showed tissue expression patterns that corresponds to their putative functions as predicted by their phylogenetic positions. The benzenoid-related genes (*GmACT3* and *4*) had maximum expression in above-ground organs, whereas *GmACT5* was highly expressed only in the root (Fig. 6; Additional file 1: File S1a). Benzenoid compounds are produced mostly in flowers and known to be involved in the aroma of numerous fruits and contribute to improving flavour quality [29]. Indeed, already characterised genes from this clade, *PhBPBT* from *Petunia x hybrida* and *CbBEBT* from *Clarkia breweri* are highly expressed in floral parts and consistently involved with benzylbenzoate compounds [27, 29]. Monolignol synthetic genes (e.g., *GmACT1* and *16*) showed high expression in root and pod. Likewise, *OsPMT* and *BdPMT*, which are specific to monolignols with *p*-coumarate and ubiquitously expressed but particularly high in spikelets. When overexpressed, these genes led to a three-fold increase in pCA lignin compared to the control [37]. Monolignol-pCA conjugates integrate with lignin biosynthesis through oxidative radical coupling in order to generate the pCA appendages [36, 37]. Polyamine conjugates are abundant in seeds and serve as nitrogen reserves during germination [45]. The Arabidopsis *SMT* gene is highly expressed in seeds and produce disinapoyl spermidine and its glucoside while being mainly expressed in roots to produce coumaroylated spermidines [41]. Soybean also contains *SMT* and *SCT* related genes, *GmACT2* and *12*, respectively, which are expressed constitutively, even though *GmACT12* expression is barely detectable in the stem, leaves and roots but spermidine still accumulates at high levels in seeds and roots [41].

The clade II members, *GmACT29* and *30*, showed high expression only in the seeds. *GmACT40* is expressed in roots and nodules, while *GmACT38* transcriptional activity is especially high in nodules and flowers (Fig. 6; Additional file 1: File S1a). On the other hand, *GmACT21*, *22*, *32*, *35*, *41*, and *45* showed expression in all tissues and at high levels (Fig. 6; Additional file 1: File S1a) suggesting putative roles in cell wall development and plant defence by producing CGA or quinate-like compounds, provided their

close relatedness to *TpHCT1A* and *TpHCT1B* (Fig. 2; [43]). Phenolamide, an important compound produced by clade II BAHD enzymes, is often associated with floral induction and development. *GmACT21*, *22*, and *32* showed particularly high transcription in the stem and flowers, similarly to *TpHCT1*, which was shown to have 4 to 5-fold higher expression in these organs than the leaf and plays an important role in stem lignification by participating in monolignol biosynthesis [40, 43]. *GmACT27*, *28* and *45* are very similar at the amino acid sequence and also showed the same expression pattern as reported for *TpHCT2* [43]. A reduction in the biosynthesis of phenolamides or their conjugates leads to male sterility in plants. Tri-substituted hydroxycinnamoyl spermidines were synthesised in the pollen coat, and this conjugation is catalysed by *SHT* [76]. *GmACT22*, *32* and *35* also had their highest expression in the flower and, given their phylogenetic position and expression patterns, these genes might as well be involved in the synthesis or conjugation of spermidines in the pollen coat of soybean.

The clade III genes, *GmACT67* and *72* are highly and specifically expressed in the flower, *GmACT55* in the root, *GmACT49*, *54*, and *65* in the nodule, whereas *GmACT66* is expressed in the leaf, *GmACT47* in stem and flower, *GmACT57* in root and flower, and *GmACT71* in seed and the pod (Fig. 6; Additional file 1: File S1a). The putative isoflavonoid malonyltransferase genes, such as *GmMaT1*, *2*, *3*, *4*, *GmACT68*, *69*, and *70*, showed high expression levels in almost all tissues. Other related members presented a more specific expression pattern, such as *GmACT71* in seeds and pods, and *GmACT50* in the leaf (Fig. 6; Additional file 1: File S1a). The different expression patterns of these genes suggest putative roles of malonylated flavonoids in storage, symbiosis signalling, flower colour, and resistance against biotic and abiotic stresses. Most of clade IIIB members had high expression in leaves and the stem, similarly to the characterised members of this group. High and specific expression of *CrDAT* in leaves correlates with the synthesis of vindoline from tabersonine in *Catharanthus roseus* (Apocynaceae) [13]. Likewise, the expression of ASAT genes was higher in the trichomes of *S. lycopersicum* stem and *S. pennellii* leaf, which correlates with acylsugars accumulation in these organs [53].

DCR genes are specifically expressed in the epidermis of vegetative and reproductive organs. Additionally, they also showed expression in the lateral root primordia and developing lateral roots [26]. The cuticle of different organs is the primary barrier against several biotic and abiotic stresses [77]. High expression of DCR genes in young leaves and stem is predictive of their role in the formation of cutin polymers in vegetative organs [26]. During leaf expansion, the cutin polymeric matrix is formed, and DCR plays a vital role to incorporate the hydroxyl fatty acid chains into the cutin polymer [1]. It is also critical to convert monomer units into its polymeric form of cutin in the flower. DCR develops the lipid polyester for the plant epidermal cell differentiation and patterning, trichome development, and protection from undesired post-genital fusions in aboveground organs [26]. We noticed that almost all clade IV genes showed high expression levels in seeds, roots, and nodules. Some genes, such as *GmACT82* and *95*, presented particularly high expression levels in roots and flowers, *GmACT94* and *91* in seeds, *GmACT99* in roots and leaves (Fig. 6; Additional file 1: File S1a).

Expression profiles of BAHD genes during nodule development in soybean

From our analysis of RNA-Seq data derived from soybean nodules at different developing stages [78], BAHD genes were differentially expressed at different developmental stages of the nodule, since its development until senescence (Fig. 7; Additional file 1: File S1b). Most of clade II and clade IV genes showed higher expression in roots than nodules, while 50 and 80% genes in clades I and III showed the opposite pattern. The expression levels of *GmACT39, 40, 41* (clade II), *27* and *33* (clade IV) were 10 to 40-fold higher in the initial stages of nodule development and reached up to >80-fold in later in the development (stage 5, pre-senescence) than roots. Even though most of clade I and III genes were expressed more highly in nodules than roots, their expression levels were not much higher than the genes in clades II and IV, except for *GmACT12* and *68*, which showed more than 2 and 70 fold higher expression than roots, respectively (Fig. 7; Additional file 1: File S1b). The expression patterns of *GmACT35, 39, 86, 87*, and *GmMaT1* gradually increased as the nodule developed, peaking at stage 5 and then dropping at senescence (stage 6), which imply roles during nodule development. Although the expression levels of *GmMaT1* and *GmACT86* were higher in roots than nodules, their expression increased as the nodule developed. Therefore, they may play a symbiotic role in the relationship between root and nodule development by modifying secondary metabolites in roots for secretion into the rhizosphere in order to attract rhizobia, such as the isoflavonoids that act as a signalling molecule for inducing *nod* factor biosynthesis in rhizobia. The transcription level of *GmMaT3* was the highest after *GmACT87* than any other BAHD genes during the different nodule stages analysed, and its expression remained constant throughout nodule development and almost similar to that of the root (Fig. 7; Additional file 1: File S1b). The expression levels of *GmACT15, 18, 40, 41, 75*, and *GmMaT2* were higher in the initial stages and gradually decreased as the nodule matured, reaching the lowest expression levels at senescence (Fig. 7; Additional file 1: File S1b). Therefore, these genes may function in rhizobial interaction or nodule initiation and development.

Expression patterns of soybean BAHD genes in response to different stresses

The RNA-Seq data available allowed us to analyse the expression profile of BAHD family genes in plants subjected to acidic condition (low pH 4.0), aluminium treatment 50 μM Al^{3+} (pH 4.0), and fungal elicitor after 2 and 24 h. Most of the BAHD genes expression decreased in response to low pH and Al^{3+} exposure, except for *GmACT13* (clade I) and *36* (clade II). Their expression levels doubled at low pH but decreased or remained the same upon Al^{3+} stress (Fig. 8; Additional file 1: File S1c). The expression of *GmACT49*, *GmMaT1* and *3* (clade III) as well as *GmACT84* and *92* (clade IV) increased dramatically in response to both stresses (Fig. 8; Additional file 1: File S1c). The expression of *GmACT49, 84* and *92* increased about 11-, 7-, 40-fold in response to low pH, and 6-, 7-, 70-fold in response to Al^{3+} stress, respectively. Meanwhile, the expression of *GmMaT1* and *3* significantly increased under these conditions as compared to the control (Fig. 8; Additional file 1: File S1c). Therefore, the BAHD genes that showed elevated transcription levels after low pH and Al^{3+} treatments are good candidate for studying Al^{3+} toxicity tolerance mechanisms in soybean. On the other hand, the expression of *GmACT12* (clade I), *39* (clade II), *62, 68, 75, GmMaT2* and *4*, (clade III) as well as *GmACT86* (clade IV) decreased between 5 and 70-fold, revealing their sensitivity to these stresses.

Transcriptomic analyses in response to fungal elicitor (FE) showed that the expression of 29% of the BAHD genes (30 out of the 103) changed significantly, with 10 genes increasing and 20 decreasing their expression (Fig. 9; Additional file 1: File S1d). The expression of *GmACT35*, *36* (clade II) *GmMaT1* and *2*, *GmACT50*, *57*, *58* (clade III), *84*, *85* and *92* (clade IV) increased while the other genes decreased their transcription. An increase in expression was observed for *GmACT58* (40-fold), *57* (20-fold), *36* (12-fold) as well as *84*, *92*, and *GmMaT1* (2-fold each) (Fig. 9; Additional file 1: File S1d). This expression pattern suggests a role in resistance to fungal attack by the synthesis and secretion of modified metabolites. Indeed, fungal and bacterial infections, or elicitor treatment in Solanaceae plants and *Arabidopsis*, increase the synthesis of coumaroyl or feruloyl tyramines [79, 80]. The anti-microbial activity was demonstrated for dicoumaroyl-caffeoyl spermidine, which inhibits mycelial growth of *Pyrenophora* and reduces powdery mildew infection in barley seedlings [44]. Likewise, constitutive accumulation of tyramine derivatives in transgenic tomato promotes resistance to *Pseudomonas syringae* [81]. *MtMaT3* expression increased in response to fungal infection and decreased in response to copper treatment [17]. The observed modulation of BAHD family genes in soybean in response to fungal elicitor may correlate with an increased acylation of active compounds against fungal attacks, which warrants further analysis.

Expression of BAHD genes in seed coat and cotyledons of different soybean genotypes

Soybean RNA-Seq data of coat and cotyledon tissues dissected from developing seeds (stage 5, 380-450 mg in weight) of three varieties with distinct seed coat colours were downloaded [82] and analysed. Interestingly, some BAHD family genes were highly expressed in the seed coat while low levels were found in the cotyledons, and vice versa. *GmACT5* (clade I), *GmACT25* (clade II), *GmMaT2* and *GmIMaT3* (clade III), and *GmACT84* (clade IV) showed higher expression in the cotyledon compared to the seed coat (Additional file 1: File S1e). This result suggests a function in the accumulation of modified metabolites, potentially related to resistance against seed borers. The expression levels of *GmACT4*, *GmACT15* (clade I), *GmACT21*, *GmACT22*, and *GmACT35* (clade II), *GmACT57*, *GmMaT4*, *GmACT70* (clade III), *GmACT79*, and *GmACT81* (clade IV) were higher in brown than black seed coats and almost absent in yellow seed coats (Additional file 1: File S1e), which suggests a role in anthocyanin modification.

Expression modulation of BAHD genes by MYB transcription factors, GmLEC2a, and GmWRi1b in the hairy root system

MYB transcription factors (TF) tightly regulate the biosynthesis of flavonoids and anthocyanins [83]. Many MYB factors have been reported to either positively or negatively regulate isoflavonoid accumulation in soybean roots and seeds as well as the production of anthocyanins and proanthocyanidins (PAs) in seeds [83]. Two soybean MYB TFs, a CPC-like R3-MYB repressor (GmMYB3) and an R2R3-MYB activator (GmMYB7), were overexpressed in soybean hairy roots to perform a transcriptome analysis on BAHD genes. The expression of *GmACT17* (clade I), *GmACT23*, *GmACT40*, *GmACT41*, *GmACT43* (clade II), *GmACT75* (clade III) and *GmACT98* (clade IV) was upregulated by *GmMYB3* (Additional file 1: File S1f; Additional file 3: Fig. S2) while the genes *GmACT9*, *GmACT12* (clade I), *GmACT21*, *GmACT22*, *GmACT28* (clade II) were induced by *GmMYB7* (Additional file 1: File S1f;

Additional file 3: Fig. S2). The transcription levels of *GmACT24* (clade II) and *GmMaT2* (clade III) were upregulated 2 to 5-fold by both, *GmMYB3* and *GmMYB7*, as compared to the control. Most of the induced genes belong to clades I and II, which are involved in the modification of diverse compounds by adding ferulyl, caffeoyl, sinapoyl, *p*-coumaroyl, and hydroxycinnamoyl moieties to a substrate. On the other hand, the upregulated genes of clade III may be involved in the synthesis of flavonoids or anthocyanins since we observed overexpression of *MtPAR*, which a previous report showed to lead to a 10-fold increase in flavonoid contents [83]. *GmMYB3* and *GmMYB7* negatively regulated most of the clade III genes, but not *GmACT44* (clade II). The expression levels of *GmACT54*, *GmACT57*, *GmACT62*, *GmACT64*, and *GmACT65* were suppressed in both *GmMYB* overexpressed soybean hairy roots (Additional file 1: File S1f; Additional file 3: Fig. S2).

Higher isoflavonoid levels were reported in hairy roots than the control [84]. An active complex is formed via a MYB activator to prompt isoflavone biosynthesis. When the *Medicago truncatula* proanthocyanidin (PA) biosynthesis regulator (*MtPAR*) was heterologously expressed in soybean hairy roots decreased contents of isoflavonoids, such as daidzin, glycitin, genistin, malonyldaidzin, and daidzein up to 3, 2, 4, 2.5, and 8-fold, respectively, as compared to control through repressing the expression of the *isoflavone synthase* (*IFS*) gene [83]. In fact, the *Medicago* MYB TF (*MtPAR*) suppressed the expression of not only *IFS* but also other isoflavonoid biosynthesis genes. Indeed, *MtPAR* has a documented negative impact on isoflavone production through expression suppression of the biosynthesis genes [83].

The expression of some BAHD acyltransferases were also significantly modulated in *GmLEC2a*-overexpressing hairy roots [85]. While some genes specifically expressed in leaves (*GmACT26*) and roots (*GmACT84*, *GmACT85*) should not be native *GmLEC2a* targets of due to their distinct expression patterns, genes that are highly expressed in seeds (*GmACT15*, *GmACT34*), pods (*GmACT71*, *GmACT99*) or open flowers (*GmACT75*, *GmACT97*) could potentially be natural targets. Many of these genes were repressed in the *GmLEC2a*-overexpressing system (Additional file 1: File S1g). This result suggests that *LEC2a* may mainly upregulate the primary metabolism while repressing genes of the specialised metabolism, including those coding for enzymes that modify metabolic structures, such as BAHD genes.

Some of the specialised metabolic genes were also regulated by the soybean WRINKLED1 (*GmWRI1b*), an AP2/EREBP family TF that regulates fatty acid metabolism in seeds, as shown by their capability of modifying gene expression significantly in hairy roots [86]. The expression of 63% (65 out of 103 genes) of the BAHD genes was modulated significantly in *GmWRI1b*-OE hairy roots, with 51% of the genes upregulated while were 49% downregulated (Additional file 1: File S1h). Interestingly, the two genes (*GmACT2* and *GmACT4*) that have higher expression in above ground soybean parts i.e., stem, leaf, flower, were upregulated in *GmWRI1b*-OE hairy roots. BAHD genes that are expressed in specific organs, such as leaves (*GmACT28*), stems (*GmACT47*), and flowers (*GmACT47*, *GmACT57*, *GmACT82*) were also upregulated by *GmWRI1b* in hairy roots. Some genes expressed in root nodules (*GmACT9*, *GmACT44*, *GmACT49*) were upregulated, whereas others (*GmACT54*, *GmACT87* and *GmACT89*) were downregulated. On the other hand, most of the genes specifically expressed in roots (*GmACT18*, *GmACT21*, *GmACT22*, *GmACT23*, *GmACT36*, *GmACT40*, *GmACT41*, *GmACT43*, *GmACT86*) and seeds

(*GmACT15*, *GmACT94*) were downregulated in *GmWRi1b*-OE hairy roots. The genes *GmACT56* and *GmACT62* that were naturally highly expressed in roots and other organs were upregulated, but those root-expressed genes that were also expressed in nodules (e.g., *GmACT92*, *GmACT98*) were rather down-regulated. We also noticed that some BAHD genes with very low counts in all organ analysed were significantly upregulated (*GmACT33*) or downregulated (*GmACT17*) while most of the clade IV genes that targeted acyl-lipids for wax or cuticle synthesis were down-regulated in the *GmWRi1b*-overexpressing hairy roots. (Additional file 1: File S1h).

Validation of BAHD gene expression in developing nodules and in response to stresses via qRT-PCR

In order to validate the *in silico* expression analysis obtained from public RNA-Seq data, and further assess the involvement of BAHD genes in physiological processes, we performed quantitative reverse-transcription PCR (qRT-PCR). We analysed the expression profiles during nodule development and root responses to Al^{3+} and fungal elicitor exposures. To select BAHD genes for expression analyses in soybean, we considered the following criteria: (i) high expression in different tissues; (ii) representation of genes in all phylogenetic clades; (iii) inclusion of some paralogous pairs. The expression profiles revealed by qRT-PCR were generally consistent with the RNA-seq data. The expression of *GmACT2* was higher in nodules than roots but remained static in all nodule developing stages. In contrast, the expression levels of *GmACT21*, *GmACT22*, *GmACT55*, *GmACT68*, *GmACT79*, *GmACT81*, *GmACT84* and *GmACT86* were higher in roots while lower but static during nodule development. *GmACT12*, *GmACT15*, *GmACT41* expression was higher during the initial stages of nodule development but gradually decreased and remained low during nodule senescence. *GmACT44* expression was initially low and steadily increased and was found high expression during nodule senescence (Fig. 7B). *GmACT84* expression was high in response to Al^{3+} (Fig. 8B) and fungal elicitor (Fig. 9B) exposures compared to the control. At the same time, most of the other genes tested significantly decreased their expression or remained virtually unchanged.

Conclusions

The findings described in this study will help to reveal the roles that BAHD genes play in soybean tissues as well as in other plant species. We identified 103 BAHD genes in the soybean genome spread over all of its 20 chromosomes. The distribution of genes within the genome, their organization and structure suggest a complex evolutionary history of the BAHD family in this legume species. Several cases of segmentally duplicated genes were found, which could have originated during its last genome duplication event. The expression dynamics of BAHD genes in different tissues of the mature plant, during nodule development, and in roots responding to different stressors indicates that this family is broadly involved in organogenesis and stress responses. Our findings warrant further investigation on the molecular and physiological functions of members of this family.

Materials And Methods

Identification of BAHD family genes in the *Glycine max* genome

BLASTP against the *G. max* proteome dataset obtained from the genome annotation (<https://phytozome.jgi.doe.gov/pz/portal.html>) was used to search for BAHD genes using two characterised protein sequences from legumes as query: GmMT7 from soybean [16], and MaTs from *Medicago truncatula* [7, 17]. The HXXXD motif was searched manually in each retrieved sequence to confirm it as a candidate BAHD protein. The annotation of soybean BAHD genes was based on their fall in the clade and position on the chromosome (*GmACT1–99*) along with the already characterised genes Glyma.18G268200 (*GmIMaT1*), Glyma.13G056100 (*GmIMaT3*) [15], Glyma.18G258000 (*GmMaT2*) and Glyma.18G268400 (*GmMaT4*) (Fig. 1A). Features of BAHD proteins, such as the number of amino acid residues, along with the start-to-end position of their respective gene in the genome, and chromosome location were acquired from the Phytozome database. Physical protein parameters of each putative gene product, such as molecular weight (Mw) and isoelectric point (pI) were calculated with the pI/Mw tool at ExPASy (<http://www.expasy.org/tools/>) using the default parameters. Additional files 1: File S3 lists details of BAHD gene products.

Chromosomal location, phylogeny, and gene structure analysis of the BAHD genes

PhenoGram Plot (<http://visualization.ritchielab.psu.edu/phenograms/plot>) was used to create the image of chromosomal locations of BAHD family genes using the chromosomal position information available at Phytozome. A phylogenetic analysis with functionally characterised proteins from different species and all BAHD proteins encoded in the soybean genome was conducted to explore the evolutionary relationships of this family. An unrooted phylogenetic tree was constructed following the Neighbour-Joining method in MEGA 6.0 [87]. The tool Gene Structures Display Server (<http://gsds.cbi.pku.edu.cn>) was used to determine the exon/intron structures of individual BAHD genes.

Calculation of K_a and K_s ratios

The Plant Genome Duplication Database (PGDD, <http://chibba.agtec.uga.edu/duplication/>) was used to obtain the K_a and K_s values of BAHD paralogous pairs. Divergence time (T) was calculated using the formula $T = K_s/2\lambda$, where $\lambda = 6.161029 \times 10^{-9}$ for *G. max* [88]. Selection pressures on duplicated genes were estimated using the K_a/K_s ratio for each paralogous gene pair. Assumptions of negative, neutral and positive evolution were considered for the values <1 , $=1$, and >1 , respectively.

Diversifying evolutionary selection analysis

A phylogenetic analysis was carried out by the HyPhy program on CDS sequences using the Nei-Gojobori method and applied in MEGA 6.0 to calculate K_a/K_s ratios for neutrality analysis of BAHD genes within each clade. Maximum-likelihood computation based on the Kimura-2 parameter model was used to infer the evolutionary history of the genes within each clade. The detection of episodic diversification effect of individual coding sites was performed through different approaches. The episodic diversifying and pervasive positive selections were identified at the individual branch site level through the mixed-effect

model of evolution (MEME). Markov Chain Monte Carlo (MCMC) was used in fast, unconstrained Bayesian approximation (FUBAR) to ensure the strength over predefined sites through the approximate Bayesian method [70]. The parameter $\omega = \beta/\alpha$ was used in MEME to fit the data in the GTR nucleotide model as initial values. The β^- and β^+ parameters were used to measure the selection pressure, whereas β^- , β^+ and α were used to estimate the variability of site-to-site rate substitution. LRT based on χ^2 asymptotic distribution was considered significant at p -value < 0.05. The codon model evolution used a phylogenetic Markov model that includes substitution rates, character frequencies, amino acid substitution rate clustering, and branch lengths estimated through the maximum-likelihood method.

Identification of conserved protein motifs and promoter *cis* elements

The identification of conserved regions within BAHD proteins was executed through MEME (<http://meme-suite.org/>), and their genomic assemblies were screened through the Pfam database (<http://pfam.sanger.ac.uk>). Regulatory elements and their analytical data were obtained using the Plant *cis*-acting Regulatory DNA Elements (PlantCARE) program (<http://bioinformatics.psb.ugent.be/webtools/plantcare>) to analyse the 1.5kb promoter region upstream of the start codon for each BAHD gene.

In silico gene expression analysis

Gene expression data derived from microarrays of identified BAHD genes in seven tissues (seed, root, nodule, stem, leaf, flower, and pod; Additional file 1: File S1a) was obtained at Phytozome v.12.0 (<https://phytozome.jgi.doe.gov/pz/portal.html>).

Plant growth, RNA-Seq data analysis, and gene expression confirmation with qRT-PCR

Soybean seeds cv. Williams 82 were germinated in vermiculite in a light chamber at $25 \pm 2^\circ\text{C}$. Seedlings were transferred into a hydroponic culture system. The soybean seedlings were grown in an incubator at 28°C with a 16/8 h photoperiod for 24 h. Samples at different nodulation stages (N-1 to N-6) were harvest from root systems inoculated at different days, at about 10, 16, 22, 26, 32, 38 days after inoculation, which are named as N1, N2, N3, N4, N5, and N6 correspondingly. These nodules at different developmental stages can be marked by average dry weights, from N1 to N6 were about 0.0005, 0.001, 0.004, 0.01, 0.03, 0.07 gram. [85, 78]. The seedlings were hydroponically subjected to acidic treatment (pH 4.0) and 50 mM Al^{3+} stress (under pH 4.0) [15, 89] for 10 days before harvesting the roots for expression analysis. For elicitor (FE) treatment, wild-type seedlings were transferred to hydroponic medium and grown under controlled condition for 14 days. A *Fusarium* spp. oligosaccharide elicitor was prepared. The phenol/sulphuric acid method was used to determine the concentration of the elicitor. Elicitors (30 mg/mL fungal elicitor or 50 μM MeJA) or inhibitors were added to the solution for each treatment. Roots were collected at 2 and 24h after treatment for gene expression analysis. In all cases, the samples were collected, snap frozen in liquid N_2 , and stored at -80°C for RNA extraction.

Illumina HiSeq2500 sequencing libraries were synthesised and further sequenced. The data were utilised to quantify the expression of soybean BAHD family genes (i.e. the number of sequence reads/million reads aligned) in nodule developing stages (Additional files 1: File S1b), response in Al³⁺ stressed roots (Additional files 1: File S1c) and fungal elicitor treated roots (Additional files 1: File S1d), genotypes with different seed coat colours (Additional files 1: File S1e), as well as hairy roots overexpressing the following soybean genes: MYB transcription factors (*MYB3* and *MYB7*) (Additional files 1: File S1f), *GmLEC2a* (Additional files 1: File S1g), and *GmWRi1b* (Additional files 1: File S1h).

Fourteen BAHD genes were selected for validation through qRT-PCR on the basis of their expression level in different soybean tissues during nodule developing and response to Al³⁺ and fungal elicitor (gene-specific primers are shown in Additional files 1: File S2). An iQ5 Real-Time PCR System (Bio-Rad) with 96-well plates in 20-µL reactions, each consisting of 2.5 µL Power SYBR Master Mix (Applied Biosystems), 1 µL primer mix (0.4 µL of each F/R primer + 0.2 µL H₂O) and 2 µL cDNA. Gene expression was normalised with the housekeeping gene *GmACTIN*.

Abbreviations

ACT: Acyltransferase

UV: Ultraviolet

HCBT: Hydroxycinnamoyl/benzoyl-CoA:anthranilate transferase

HCT: Hydroxycinnamoyl transferase

SHT: Spermidine hydroxycinnamoyl transferase

HCA: Hydroxycinnamic acid

MaT: Malonyltransferase

MW: Molecular weight

pI: Isoelectric point

cDNA: Complementary deoxynucleic acid

PlantCARE: Plant *cis*-acting Regulatory DNA Elements

MEME: Mixed-effect model evolution

MCMC: Markov Chain Monte Carlo

FUBAR: Fast unconstrained Bayesian approximation

kDs: Kilodaltons

WGD: Whole-genome duplication

MYA: Million years ago

EBF: Empirical Bayes factor

mBIC: Modified Bayesian Information Criterion

Declarations

Ethics approval and consent to participate

Not Applicable

Consent for publication

Not Applicable

Availability of data and materials

All data generated or analysed during this study are included in this published article or as supplementary information.

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported by the National Science Foundation of China (grant 31670294), the Ministry of Science and Technology of China (grant 2016YFD0100504; 2016YFD0100504), and the funding from Anhui Agricultural University and the State Key Laboratory of Tea Plant Biology and Utilization.

Authors' contributions

JZ planned and designed the research. MZ, XZ, QD, SM, HJ, PL, XW performed experiments and bioinformatics analyses. JZ and MZ wrote the manuscript. VAB proofread and edited the manuscript. All authors have read and approved the manuscript.

Acknowledgements

The authors thank lab members in Prof. Zhao's lab for all assistances in experiments and data analyses.

References

1. Reina JJ, Guerrero C, Heredia A. Isolation, characterization, and localization of AgaSGNH cDNA: a new SGNH-motif plant hydrolase specific to *Agave americana* L. leaf epidermis. *Journal of experimental botany*. 2007;58(11):2717-2731.
2. Bontpart T, Cheynier V, Ageorges A, Terrier N. BAHD or SCPL acyltransferase? What a dilemma for acylation in the world of plant phenolic compounds. *New Phytologist*. 2015;208(3):695-707.
3. Lam KC, Ibrahim RK, Behdad B, Dayanandan S. Structure, function, and evolution of plant O-methyltransferases. *Genome*. 2007;50(11):1001-1013.
4. Bowles D, Isayenkova J, Lim E-K, Poppenberger B. Glycosyltransferases: managers of small molecules. *Current Opinion in Plant Biology*. 2005;8(3):254-263.
5. Nakayama T, Suzuki H, Nishino T. Anthocyanin acyltransferases: specificities, mechanism, phylogenetics, and applications. *Journal of Molecular Catalysis B: Enzymatic*. 2003;23(2-6):117-132.
6. Zhao J. Flavonoid transport mechanisms: how to go, and with whom. *Trends in Plant Science*. 2015;20(9):576-585.
7. Zhao J, Huhman D, Shadle G, He X-Z, Sumner LW, Tang Y, Dixon RA. MATE2 mediates vacuolar sequestration of flavonoid glycosides and glycoside malonates in *Medicago truncatula*. *The Plant Cell*. 2011;23(4):1536-1555.
8. Landry LG, Chapple CC, Last RL. Arabidopsis mutants lacking phenolic sunscreens exhibit enhanced ultraviolet-B injury and oxidative damage. *Plant Physiology*. 1995;109(4):1159-1166.
9. Mugford ST, Qi X, Bakht S, Hill L, Wegel E, Hughes RK, Papadopoulou K, Melton R, Philo M, Sainsbury F. A serine carboxypeptidase-like acyltransferase is required for synthesis of antimicrobial compounds and disease resistance in oats. *The Plant Cell*. 2009;21(8):2473-2484.
10. Dudareva N, D'auria JC, Nam KH, Raguso RA, Pichersky E. Acetyl-CoA: benzylalcohol acetyltransferase—an enzyme involved in floral scent production in *Clarkia breweri*. *The Plant Journal*. 1998;14(3):297-304.
11. Fujiwara H, Tanaka Y, Fukui Y, Nakao M, Ashikari T, Kusumi T. Anthocyanin 5-aromatic acyltransferase from *Gentiana triflora*: purification, characterization and its role in anthocyanin biosynthesis. *European Journal of Biochemistry*. 1997;249(1):45-51.
12. Yang Q, Reinhard K, Schiltz E, Matern U. Characterization and heterologous expression of hydroxycinnamoyl/benzoyl-CoA: anthranilate N-hydroxycinnamoyl/benzoyltransferase from elicited cell cultures of carnation, *Dianthus caryophyllus* L. *Plant Molecular Biology*. 1997;35(6):777-789.
13. St-Pierre B, Laflamme P, Alarco AM, Luca E. The terminal O-acetyltransferase involved in vindoline biosynthesis defines a new class of proteins responsible for coenzyme A-dependent acyl transfer. *The Plant Journal*. 1998;14(6):703-713.
14. D'Auria JC. Acyltransferases in plants: a good time to be BAHD. *Current Opinion in Plant Biology*. 2006;9(3):331-340.
15. Ahmad MZ, Li P, Wang J, Rehman NU, Zhao J. Isoflavone malonyltransferases GmIMaT1 and GmIMaT3 differently modify isoflavone glucosides in soybean (*Glycine max*) under various stresses. *Frontiers in plant science*. 2017;8(735).

16. Dhaubhadel S, Farhangkhoei M, Chapman R. Identification and characterization of isoflavonoid specific glycosyltransferase and malonyltransferase from soybean seeds. *Journal of experimental botany*. 2008;59(4):981-994.
17. Yu XH, Chen MH, Liu CJ. Nucleocytoplasmic-localized acyltransferases catalyze the malonylation of 7-O-glycosidic (iso) flavones in *Medicago truncatula*. *The Plant Journal*. 2008;55(3):382-396.
18. Kim B-G, Jung WD, Mok H, Ahn J-H. Production of hydroxycinnamoyl-shikimates and chlorogenic acid in *Escherichia coli*: production of hydroxycinnamic acid conjugates. *Microbial cell factories*. 2013;12(1):15.
19. Bloch SE, Schmidt-Dannert C. Construction of a chimeric biosynthetic pathway for the de novo biosynthesis of rosmarinic acid in *Escherichia coli*. *ChemBioChem*. 2014;15(16):2393-2401.
20. Kim IA, Kim B-G, Kim M, Ahn J-H. Characterization of hydroxycinnamoyltransferase from rice and its application for biological synthesis of hydroxycinnamoyl glycerols. *Phytochemistry*. 2012;76(25-31).
21. Eudes A, Juminaga D, Baidoo EE, Collins FW, Keasling JD, Loqué D. Production of hydroxycinnamoyl anthranilates from glucose in *Escherichia coli*. *Microbial cell factories*. 2013;12(1):62.
22. D'Auria JC, Gershenzon J. The secondary metabolism of *Arabidopsis thaliana*: growing like a weed. *Current Opinion in Plant Biology*. 2005;8(3):308-316.
23. Moglia A, Acquadro A, Eljounaidi K, Milani AM, Cagliero C, Rubiolo P, Genre A, Cankar K, Beekwilder J, Comino C. Genome-wide identification of BAHD acyltransferases and in vivo characterization of HQT-like enzymes involved in caffeoylquinic acid synthesis in *Globe artichoke*. *Frontiers in plant science*. 2016;7(1424).
24. Suzuki H, Nishino T, Nakayama T. cDNA cloning of a BAHD acyltransferase from soybean (*Glycine max*): isoflavone 7-O-glucoside-6"-O-malonyltransferase. *Phytochemistry*. 2007;68(15):2035-2042.
25. Fujiwara H, Tanaka Y, Yonekura-Sakakibara K, Fukuchi-Mizutani M, Nakao M, Fukui Y, Yamaguchi M, Ashikari T, Kusumi T. cDNA cloning, gene expression and subcellular localization of anthocyanin 5-aromatic acyltransferase from *Gentiana triflora*. *The Plant Journal*. 1998;16(4):421-431.
26. Panikashvili D, Shi JX, Schreiber L, Aharoni A. The Arabidopsis DCR encoding a soluble BAHD acyltransferase is required for cutin polyester formation and seed hydration properties. *Plant Physiology*. 2009;151(4):1773-1789.
27. D'Auria JC, Chen F, Pichersky E. Characterization of an acyltransferase capable of synthesizing benzylbenzoate and other volatile esters in flowers and damaged leaves of *Clarkia breweri*. *Plant Physiology*. 2002;130(1):466-476.
28. Chedgy RJ, Köllner TG, Constabel CP. Functional characterization of two acyltransferases from *Populus trichocarpa* capable of synthesizing benzyl benzoate and salicyl benzoate, potential intermediates in salicinoid phenolic glycoside biosynthesis. *Phytochemistry*. 2015;113(149-159).
29. Boatright J, Negre F, Chen X, Kish CM, Wood B, Peel G, Orlova I, Gang D, Rhodes D, Dudareva N. Understanding in vivo benzenoid metabolism in petunia petal tissue. *Plant Physiology*. 2004;135(4):1993-2011.

30. Pollard M, Beisson F, Li Y, Ohlrogge JB. Building lipid barriers: biosynthesis of cutin and suberin. *Trends in Plant Science*. 2008;13(5):236-246.
31. Molina I, Li-Beisson Y, Beisson F, Ohlrogge JB, Pollard M. Identification of an Arabidopsis feruloyl-coenzyme A transferase required for suberin synthesis. *Plant Physiology*. 2009;151(3):1317-1328.
32. Serra O, Hohn C, Franke R, Prat S, Molinas M, Figueras M. A feruloyl transferase involved in the biosynthesis of suberin and suberin-associated wax is required for maturation and sealing properties of potato periderm. *The Plant Journal*. 2010;62(2):277-290.
33. Rautengarten C, Ebert B, Ouellet M, Nafisi M, Baidoo EE, Benke P, Stranne M, Mukhopadhyay A, Keasling JD, Sakuragi Y. Arabidopsis Deficient in Cutin Ferulate encodes a transferase required for feruloylation of ω -hydroxy fatty acids in cutin polyester. *Plant Physiology*. 2012;158(2):654-665.
34. Kosma DK, Molina I, Ohlrogge JB, Pollard M. Identification of an Arabidopsis fatty alcohol: caffeoyl-coenzyme A acyltransferase required for the synthesis of alkyl hydroxycinnamates in root waxes. *Plant Physiology*. 2012;160(1):237-248.
35. Luo J, Nishiyama Y, Fuell C, Taguchi G, Elliott K, Hill L, Tanaka Y, Kitayama M, Yamazaki M, Bailey P. Convergent evolution in the BAHD family of acyl transferases: identification and characterization of anthocyanin acyl transferases from *Arabidopsis thaliana*. *The Plant Journal*. 2007;50(4):678-695.
36. Withers S, Lu F, Kim H, Zhu Y, Ralph J, Wilkerson CG. Identification of grass-specific enzyme that acylates monolignols with p-coumarate. *Journal of Biological Chemistry*. 2012;287(11):8347-8355.
37. Petrik DL, Karlen SD, Cass CL, Padmakshan D, Lu F, Liu S, Le Bris P, Antelme S, Santoro N, Wilkerson CG. p-Coumaroyl-CoA: monolignol transferase (PMT) acts specifically in the lignin biosynthetic pathway in *Brachypodium distachyon*. *The Plant Journal*. 2014;77(5):713-726.
38. Bartley LE, Peck ML, Kim S-R, Ebert B, Manisseri C, Chiniquy DM, Sykes R, Gao L, Rautengarten C, Vega-Sánchez ME. Overexpression of a BAHD acyltransferase, *OsAt10*, alters rice cell wall hydroxycinnamic acid content and saccharification. *Plant Physiology*. 2013;161(4):1615-1633.
39. Grienemberger E, Besseau S, Geoffroy P, Debayle D, Heintz D, Lapierre C, Pollet B, Heitz T, Legrand M. A BAHD acyltransferase is expressed in the tapetum of Arabidopsis anthers and is involved in the synthesis of hydroxycinnamoyl spermidines. *The Plant Journal*. 2009;58(2):246-259.
40. Hoffmann L, Besseau S, Geoffroy P, Ritzenthaler C, Meyer D, Lapierre C, Pollet B, Legrand M. Silencing of hydroxycinnamoyl-coenzyme A shikimate/quinate hydroxycinnamoyltransferase affects phenylpropanoid biosynthesis. *The Plant Cell*. 2004;16(6):1446-1465.
41. Luo J, Fuell C, Parr A, Hill L, Bailey P, Elliott K, Fairhurst SA, Martin C, Michael AJ. A novel polyamine acyltransferase responsible for the accumulation of spermidine conjugates in Arabidopsis seed. *The Plant Cell*. 2009;21(1):318-333.
42. Niggeweg R, Michael AJ, Martin C. Engineering plants with increased levels of the antioxidant chlorogenic acid. *Nature Biotechnology*. 2004;22(6):746-754.
43. Sullivan M. A novel red clover hydroxycinnamoyl transferase has enzymatic activities consistent with a role in phaelic acid biosynthesis. *Plant Physiology*. 2009;150(4):1866-1879.

44. Walters D, Meurer-Grimes B, Rovira I. Antifungal activity of three spermidine conjugates. *FEMS Microbiology Letters*. 2001;201(2):255-258.
45. Facchini PJ, Hagel J, Zulak KG. Hydroxycinnamic acid amide metabolism: physiology and biochemistry. *Canadian Journal of Botany*. 2002;80(6):577-589.
46. Onkokesung N, Gaquerel E, Kotkar H, Kaur H, Baldwin IT, Galis I. MYB8 controls inducible phenolamide levels by activating three novel hydroxycinnamoyl-coenzyme A: polyamine transferases in *Nicotiana attenuata*. *Plant Physiology*. 2012;158(1):389-407.
47. Mondolot L, La Fisca P, Buatois B, Talansier E, De Kochko A, Campa C. Evolution in caffeoylquinic acid content and histolocalization during *Coffea canephora* leaf development. *Annals of Botany*. 2006;98(1):33-40.
48. Narukawa M, Kanbara K, Tominaga Y, Aitani Y, Fukuda K, Kodama T, Murayama N, Nara Y, Arai T, Konno M. Chlorogenic acid facilitates root hair formation in lettuce seedlings. *Plant and cell physiology*. 2009;50(3):504-514.
49. Varbanova M, Porter K, Lu F, Ralph J, Hammerschmidt R, Jones AD, Day B. Molecular and biochemical basis for stress-induced accumulation of free and bound p-coumaraldehyde in cucumber. *Plant Physiology*. 2011;157(3):1056-1066.
50. Besseau S, Hoffmann L, Geoffroy P, Lapierre C, Pollet B, Legrand M. Flavonoid accumulation in *Arabidopsis* repressed in lignin synthesis affects auxin transport and plant growth. *The Plant Cell*. 2007;19(1):148-162.
51. D'Auria JC, Reichelt M, Luck K, Svatoš A, Gershenzon J. Identification and characterization of the BAHD acyltransferase malonyl CoA: Anthocyanidin 5-O-glucoside-6"-O-malonyltransferase (At5MAT) in *Arabidopsis thaliana*. *FEBS Letters*. 2007;581(5):872-878.
52. Suzuki H, Nakayama T, Yonekura-Sakakibara K, Fukui Y, Nakamura N, Nakao M, Tanaka Y, Yamaguchi M-a, Kusumi T, Nishino T. Malonyl-CoA: Anthocyanin 5-O-Glucoside-6'''-O-Malonyltransferase from Scarlet Sage (*Salvia splendens*) Flowers ENZYME PURIFICATION, GENE CLONING, EXPRESSION, AND CHARACTERIZATION. *Journal of Biological Chemistry*. 2001;276(52):49013-49019.
53. Fan P, Miller AM, Schilmiller AL, Liu X, Ofner I, Jones AD, Zamir D, Last RL. In vitro reconstruction and analysis of evolutionary variation of the tomato acylsucrose metabolic network. *Proceedings of the National Academy of Sciences*. 2016;113(2):E239-E248.
54. Kim J, Kang K, Gonzales-Vigil E, Shi F, Jones AD, Barry CS, Last RL. Striking natural diversity in glandular trichome acylsugar composition is shaped by variation at the Acyltransferase2 locus in the wild tomato *Solanum habrochaites*. *Plant Physiology*. 2012;160(4):1854-1870.
55. Suzuki H, Nakayama T, Nagae S, Yamaguchi M-A, Iwashita T, Fukui Y, Nishino T. cDNA cloning and functional characterization of flavonol 3-O-glucoside-6"-O-malonyltransferases from flowers of *Verbena hybrida* and *Lamium purpureum*. *Journal of Molecular Catalysis B: Enzymatic*. 2004;28(2-3):87-93.

56. Wang J, Luca VD. The biosynthesis and regulation of biosynthesis of Concord grape fruit esters, including 'foxy'methylantranilate. *The Plant Journal*. 2005;44(4):606-619.
57. Kannangara R, Branigan C, Liu Y, Penfield T, Rao V, Mouille G, Höfte H, Pauly M, Riechmann JL, Broun P. The transcription factor WIN1/SHN1 regulates cutin biosynthesis in *Arabidopsis thaliana*. *The Plant Cell*. 2007;19(4):1278-1294.
58. See DR, Brooks S, Nelson JC, Brown-Guedira G, Friebe B, Gill BS. Gene evolution at the ends of wheat chromosomes. *Proceedings of the National Academy of Sciences*. 2006;103(11):4162-4167.
59. Cannon SB, Mitra A, Baumgarten A, Young ND, May G. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC plant biology*. 2004;4(1):10.
60. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J. Genome sequence of the palaeopolyploid soybean. *Nature*. 2010;463(7278):178-183.
61. dos Santos AL, Chaves-Silva S, Yang L, Maia LGS, Chalfun-Júnior A, Sinharoy S, Zhao J, Benedito VA. Global analysis of the MATE gene family of metabolite transporters in tomato. *BMC plant biology*. 2017;17(1):185.
62. Juretic N, Hoen DR, Huynh ML, Harrison PM, Bureau TE. The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. *Genome Research*. 2005;15(9):1292-1297.
63. Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS. *PLoS genetics*. 2008;4(12):e1000304.
64. Oleksyk TK, Smith MW, O'Brien SJ. Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2010;365(1537):185-205.
65. Pond SLK, Muse SV: HyPhy: hypothesis testing using phylogenies. In: *Statistical Methods in Molecular Evolution*. Springer; 2005: 125-181.
66. Hughes AL, Friedman R. Codon-based tests of positive selection, branch lengths, and the evolution of mammalian immune system genes. *Immunogenetics*. 2008;60(9):495.
67. Ahmad HI, Liu G, Jiang X, Liu C, Chong Y, Huarong H. Adaptive molecular evolution of *MC 1R* gene reveals the evidence for positive diversifying selection in indigenous goat populations. *Ecology and evolution*. 2017a;7(14):5170-5180.
68. Manel S, Perrier C, Pratlong M, Abi-Rached L, Paganini J, Pontarotti P, Aurelle D. Genomic resources and their influence on the detection of the signal of positive selection in genome scans. *Molecular ecology*. 2016;25(1):170-184.
69. Yang Z, Dos Reis M. Statistical properties of the branch-site test of positive selection. *Molecular biology and evolution*. 2010;28(3):1217-1228.
70. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK. Detecting individual sites subject to episodic diversifying selection. *PLoS genetics*. 2012;8(7):e1002764.
71. Ahmad HI, Liu G, Jiang X, Liu C, Fangzheng X, Chong Y, Ijaz N, Huarong H. Adaptive selection at agouti gene inferred breed specific selection signature within the indigenous goat populations. *Asian-Australasian journal of animal sciences*. 2017b;00(1-7. DOI.org/10.5713/ajas.5716.0994.

72. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Molecular biology and evolution*. 2013;30(5):1196-1205.
73. Miyazawa S. Advantages of a mechanistic codon substitution model for evolutionary analysis of protein-coding sequences. *PloS one*. 2011;6(12):e28892.
74. Kosiol C, Holmes I, Goldman N. An empirical codon model for protein sequence evolution. *Molecular biology and evolution*. 2007;24(7):1464-1479.
75. Stanfel LE. A new approach to clustering the amino acid. *Journal of Theoretical Biology*. 1996;183(2):195-205.
76. Elejalde-Palmett C, de Bernonville TD, Glevarec G, Pichon O, Papon N, Courdavault V, St-Pierre B, Giglioli-Guivarc'h N, Lanoue A, Besseau S. Characterization of a spermidine hydroxycinnamoyltransferase in *Malus domestica* highlights the evolutionary conservation of trihydroxycinnamoyl spermidines in pollen coat of core Eudicotyledons. *Journal of experimental botany*. 2015;66(22):7271-7285.
77. Shepherd T, Wynne Griffiths D. The effects of stress on plant cuticular waxes. *New Phytologist*. 2006;171(3):469-499.
78. Ahmad MZ, Rehman Nu, Yu S, Zhou Y, Haq Bu, Wang J, Li P, Zeng Z, Zhao J. GmMAX2–D14 and–KAI interaction-mediated SL and KAR signaling play essential roles in soybean root nodulation. *The Plant Journal*. 2020;101(2):334-351.
79. Muroi A, Ishihara A, Tanaka C, Ishizuka A, Takabayashi J, Miyoshi H, Nishioka T. Accumulation of hydroxycinnamic acid amides induced by pathogen infection and identification of agmatine coumaroyltransferase in *Arabidopsis thaliana*. *Planta*. 2009;230(3):517.
80. Zacarés L, López-Gresa MP, Fayos J, Primo J, Bellés JM, Conejero V. Induction of p-coumaroyldopamine and feruloyldopamine, two novel metabolites, in tomato by the bacterial pathogen *Pseudomonas syringae*. *Molecular Plant-Microbe Interactions*. 2007;20(11):1439-1448.
81. Campos L, Lisón P, López-Gresa MP, Rodrigo I, Zacarés L, Conejero V, Bellés JM. Transgenic tomato plants overexpressing tyramine N-hydroxycinnamoyltransferase exhibit elevated hydroxycinnamic acid amide levels and enhanced resistance to *Pseudomonas syringae*. *Molecular Plant-Microbe Interactions*. 2014;27(10):1159-1169.
82. Chen B, Wang J, Zhang G, Liu J, Manan S, Hu H, Zhao J. Two types of soybean diacylglycerol acyltransferases are differentially involved in triacylglycerol biosynthesis and response to environmental stresses and hormones. *Scientific reports*. 2016;6(28541).
83. Li P, Dong Q, Ge S, He X, Verdier J, Li D, Zhao J. Metabolic engineering of proanthocyanidin production by repressing the isoflavone pathways and redirecting anthocyanidin precursor flux in legume. *Plant Biotechnology Journal*. 2016;14(7):1604-1618.
84. Gutierrez-Gonzalez JJ, Guttikonda SK, Tran L-SP, Aldrich DL, Zhong R, Yu O, Nguyen HT, Sleper DA. Differential expression of isoflavone biosynthetic genes in soybean during water deficits. *Plant and cell physiology*. 2010;51(6):936-948.

85. Manan S, Ahmad MZ, Zhang G, Chen B, Haq BU, Yang J, Zhao J. Soybean LEC2 regulates subsets of genes involved in controlling the biosynthesis and catabolism of seed storage substances and seed development. *Frontiers in plant science*. 2017;8(1604).
86. Chen B, Zhang G, Li P, Yang J, Guo L, Benning C, Wang X, Zhao J. Multiple GmWRI1s are redundantly involved in seed filling and nodulation by regulating plastidic glycolysis, lipid biosynthesis and hormone signalling in soybean (*Glycine max*). *Plant Biotechnology Journal*. 2020;18(1):155-171.
87. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular biology and evolution*. 2013;30(12):2725-2729.
88. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. *Science*. 2000;290(5494):1151-1155.
89. Wang J, Hou Q, Li P, Yang L, Sun X, Benedito VA, Wen J, Chen B, Mysore KS, Zhao J. Diverse functions of multidrug and toxin extrusion (MATE) transporters in citric acid efflux and metal homeostasis in *Medicago truncatula*. *The Plant Journal*. 2017;90(1):79-95.

Additional Files

Additional file S1: File S1a-h. Expression levels of *BAHD* genes in different soybean tissues, during nodule developing, root exposure to Al^{3+} , low pH conditions, fungal elicitors, and in seed tissues of genotypes with different coat colours, and hairy roots overexpressing *GmMYB3*, *GmMYB7*, *GmLEC2a* and *GmWRI1b*.

Additional file S1: File S2. List of primers of soybean *BAHD* genes used for qRT-PCR validation.

Additional file S1: File S3. Physical parameters of proteins coded by the soybean *BAHD* genes.

Additional file S1: File S4. Description of protein motifs in soybean *BAHD* sequences.

Additional file S1: File S5. *Cis*-acting promoter region analysis of soybean *BAHD* genes.

Additional file S2: Table S1. Duplicated *BAHD* genes in soybean and the estimated dates of the duplication events.

Additional file S2: Table S2. Maximum Likelihood analysis of *BAHD* proteins for codon by codon positive selection.

Additional file S2: Table S3. Mixed-effect model evolution (MEME)-based episodic diversifying selection *BAHD* genes.

Additional file S2: Table S4. Fast, unconstrained Bayesian approximation inferring pervasive diversifying selection of *BAHD* genes.

Phylogenetic tree of soybean BAHD family genes with characterised BAHD genes from different plant species. The phylogenetic tree was constructed through MEGA 6.0 using the Maximum Likelihood (ML) method. Bootstrap values in percentage (1000 replicates) are indicated on the nodes. Different clades are shaded in different colours: Clade I (pink), Clade II (olive), Clade III (blue), and Clade IV (maroon).

Figure 2

Phylogenetic analyses, conserved motif and gene structure analysis of the soybean BAHD family. (A) Phylogenetic tree and denomination of the 103 soybean BAHD genes identified. The genes are listed according to their position in the clade and location on the chromosome. Clades are shaded with different colours. (cf. Figure 1). (B) Soybean BAHD (GmACT) proteins motifs are shown as coloured boxes. The grey lines represent the full protein size (in amino acid residues). (C) Intron-exon structures of GmACT genes are shown. Introns are presented by black lines, blue boxes represent UTRs, and exonic coding regions are represented by maroon boxes.

Figure 3

Chromosomal locations and duplications of soybean BAHD genes. The chromosome number is indicated at the top of each structure. Circles with different coloured represent genes belonging to each clade. Each pair of segmental duplication is indicated by a line coloured according to their respective clade.

Figure 4

Cis-elements identified in all BAHD gene promoters in soybean. PlantCare were used to analyse the 1,500-bp upstream regions. (A) Percentage of cis-elements responsive to specific stimuli. (B) cis-elements responsive to different hormones and (C) environmental stresses.

Figure 5

Evolutionary rate clustering in structured genetic algorithm models. The models were inferred from the BADH gene alignment of different clades. Each cluster was labelled with the Maximum Likelihood estimate of its inferred rate. The nodes (residues) are annotated by their biochemical properties and Stanfel class, and the rates (edges) are labelled with the model-averaged rate estimates.

Figure 6

Expression patterns of soybean BAHD genes. The expression data of the BAHD genes in different tissues were retrieved from the Phytozome database and represented in the heatmap using TBTools.

Figure 7

Expression patterns of soybean BAHD genes during nodule development. (A) The data was retrieved from RNA-Seq public database, and the heatmap was designed using TBTools. (B) qRT-PCR analyses of selected BAHD genes during nodule development: N-1 to N-6: nodule development stages 1 to 6.

Figure 8

Expression patterns of soybean BAHD genes in response to low pH and Al³⁺ stress. (A) The root expression dataset was generated through RNA-Seq sequencing, and the heatmap was designed using TBTools. (B) qRT-PCR analyses of selected BAHD genes were done in the hairy roots exposed to acidic condition (pH 4.0) treatment and 50 mM Al³⁺ stresses (under pH 4.0), hydroponically cultivated seedlings were transferred to these conditions for 10 days before harvesting roots for gene expression.

Figure 9

Expression patterns of soybean BAHD genes in response to fungal elicitor. (A) The root expression dataset was generated through RNA-Seq sequencing, and the heatmap was designed using TBTools. (B) qRT-PCR analyses of selected BAHD genes exposed to stress for 0, 2 and 24 hours.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile31.ppt](#)
- [Additionalfile12.xls](#)
- [Additionalfile21.doc](#)