Molecular Plant Breeding 2021, Vol.12, No.27, 1-16
http://genbreedpublisher.com/index.php/mpb

Research Article Open Access

Genome-wide Identification and Analysis of β-galactosidase (BGAL) Gene Family in Cotton

Cao Xiaocong [1,2], Zhang Chaojun [2], Gou Haoqi [2], Wang Xiaoyan [3], Qiao Kaikai [2], Ma Qifeng [2], Zhang Guiyin [1], Fan Shuli [1,2]

1 Hebei Agricultural University, Hebei Base of state Key Laboratory of Cotton Biology, Baoding, 071001, P.R. China

2 Institute of Cotton Research of the Chinese Academy of Agricultural Sciences, State Key Laboratory of Cotton Biology Cotton, Anyang; 455000, P.R. China

3 Anyang Institute of Technology, Anyang, 455000, 130033, P.R. China

Corresponding author email: zhangguiy63@126.com; fsl427@126.com

Molecular Plant Breeding, 2021, Vol.12, No.27 doi: 10.5376/mpb.2021.12.0027

Received: 15 Sep., 2021

Accepted: 25 Sep., 2021

Published: 21 Oct., 2021

Preferred citation for this article:

Cao X.C., Zhang C.J., Gou H.Q., Wang X.Y., Qiao K.K., Ma Q.F., Zhang G.Y., and Fan S.L., 2021, Genome-wide identification and analysis of β-galactosidase (BGAL) gene family in cotton, Molecular Plant Breeding, 12(27): 1-16 (doi: 10.5376/mpb.2021.12.0027)

Abstract Plant β-galactosidases (BGALs) are the important glycosidase that hydrolyses the non-reductive terminaβ-D-galactosidase residues from glycochains, glycolipids and glycoproteins. In order to reveal the regulation mechanism of β-galactosidases in cotton development, the whole genome analysis of *BGAL* gene family were carried out in this study, which laid a foundation for further understanding of the function of *BGAL* genes in cotton pollen. A total of 153 *BGAL* genes were respectively identified in *Gossypium hirsutum, G.barbadense, G.arboreum*, *G.raimondii*. Phylogenetic tree analysis showed that the *BGAL* genes are divided into eight subgroups, with the same number of exons and genetic structure in the same subgroup. The co-evolutionary analysis showed that there were multiple gene pairs between *G.hirsutum* and diploid cotton. The transcriptomic data showed that some genes in *G.hirsutum* were specifically expressed in different tissues. For example, *GhBGAL6* and *GhBGAL32*, are highly expressed in all tissues; *GhBGAL33*, *GhBGAL7*, *GhBGAL18* and *GhBGAL43* were highly expressed in stamens (anther and filament) and petals. Further qRT-PCR results showed that some genes (such as *GhBGAL6*, *GhBGAL7* and *GhBGAL17*) were highly expressed in stamens, which may have a certain regulatory effect on the development of stamens. This study explored the evolution and function of the *BGAL* gene family in the genomes of cotton, which can provide a theoretical basis for subsequent research on *BGAL* genes in cotton.

Keywords β-galactosidases; Bioinformatics; Cotton; Gene family

Glycoside hydrolases (GH, EC3.2.1) are glycoside bond enzymes capable of hydrolyzing between two or more carbohydrates or between carbohydrate and non-carbohydrate parts (Henrissat and Bairoch, 1993). Glycosidic hydrolases are found in almost all organisms and can hydrolyze the glycosidic bonds of various carbohydrate compounds (including monoglycosides, oligosaccharides, polysaccharides, saponins, glycoproteins, etc.) by endoscopically or exscopically to produce monosaccharides, oligosaccharides, or sugar complexes. Glycoside hydrolase has undergone structural changes in the course of evolution. According to the difference in protein structure of hydrolase, it can be divided into 135 families (Glycoside hydrolase family), namely GH1-GH135 (Henrissat and Bairoch, 1993). The structural similarity of members of each family is very high, and according to the structural characteristics of the structural domain of the catalytic active center, these families can be classified into 14 clans, namely GH-A-GH-N (Ahn et al., 2007). β-galactosidase (BGAL) only exists in the glucoside hydrolase family GH1, 2, 3, 35, 42, 50 and 59, belonging to GHA. Plant β-galactosidase (EC3.2.1.23) is found only in the GH35 family, *Arabidopsis thaliana* (Ahn et al., 2007), *Solanum Lycopersicum* (Smith and Gross, 2000), *Carica papaya* (Lazan et al., 2004), and *Oryza sativa* (Tanthanuch et al., 2008) all contain *BGAL* family genes. This suggests that *BGAL* gene diversity is prevalent in plants.

The degradation of lactose, proteoglycan, glycolipid, oligosaccharide, and polysaccharide widely present in plants, animals and microorganisms are mainly caused by the hydrolysis of the terminal non-reducing β-D-galactoside residues in β-D-galactoside (Lombard et al., 2014) by β-galactosidase (BGAL). BGALs in plants are classified into two categories: I and II. Class I is composed of β-(1→4) exigent galactanase, which can act specifically on

β-(1→4) galactanase in pectin to produce galactose residues. Class II specifically hydrolyzes β-(1→3)- and β-(1→6) galactose residues of arabinogalactan-proteins (AGPs) to produce monuronic acids, but has no activity for β-(1→4) galactose in pectin (Sørensen et al., 2000). Class I BGAL has a specific effect on β-(1→4) β-galactosidase residues in pectin and xyloglucan, so class I BGAL plays an important role in cell wall structure and intercellular adhesion. Studies have reported that BGAL is involved in pectin decomposition during fruit ripening, softening the cell wall of plant fruits in a variety of plants such as kiwifruit, persimmon, sweet cherry, mango and peach (Guo et al., 2018). Class II BGAL is involved in many stages of the development of other plant tissues, such as spinach leaves, mung bean seedlings, radish hypocotyls and young leaves, and the meristem region of root, cotyledon, vascular tissue, trichome and pollen of tobacco (Hrubá et al., 2005). In addition, the seeds of Tropueolum mujus L., Copaifera langsdorffii and Hymenaea courbaril were observed in cotyledon. β-galactosidase is involved in the degradation of xyloglucan (Ahn et al., 2007), and it has been reported that BGAL acts with α-xyloglucase, β-glucosidase and other enzymes to achieve the degradation of xyloglucan (Wang et al., 2018). Therefore, BGAL plays an important role in plant cell wall remodeling.

Plant pollen wall plays an important role in pollen development, and the development of pollen wall will directly affect the process of pollen development, thus affecting the fertility of plants (Tian et al., 2014). The main component of pollen wall is relatively stable sporopollenin. In the mononucleate stage of microspore development, sporopollenin derived from tapetum will cover pollen and form pollen wall structure, which can effectively protect pollen from external influence (Moctezuma et al., 2003). During pollen maturity, pollen wall components such as sporopollenin degrade normally and release pollen. The synthesis and normal degradation of sporopollenin are both complex biological processes, which are jointly regulated by many genes (Smith and Gross, 2000). β-galactosidase and other genes are important genes involved in sporopollenin degradation (Ban et al., 2018). β-galactosidase has been associated with pollen development in a variety of species, such as AtBGAL7 and AtBGAL15 genes in Arabidopsis thaliana (Hrubá et al., 2005) and OsBGAL5, OsBGAL12, OsBGAL14 and OsBGAL15 genes in Oryza sativa L. play a role in early microspore development and pollen development. In Tobacco (Rogers et al., 2001), northern hybridization showed that BGAL family genes were specifically expressed in anthers, mature pollen grains and late microspore development, indicating that they also played a role in the growth process of pollen tubes. In Chinese cabbage (Liu et al., 2013), many genes of BGAL family are specifically expressed in pollen.

Studies have shown that BGAL is not only an essential enzyme in plant growth and development, but also affects the development of pollen wall. Although the BGAL family has been studied in several species, it has not been reported in cotton. With the rapid development of sequencing technology, the genome sequences of *G.hirsutum*, *G.arbadense*, *G.Raimondii* have been sequenced and analyze. In this study, bioinformatics methods and bioinformation analysis tools were used to identify and analyze the BGAL family in cotton (Lu et al., 2018), and the expression pattern of *GhBGAL* genes were analyzed. The qRT-PCR results showed that *GhBGALs* was expressed in different tissues. Furthermore, the function of BGAL genes in *G.hirsutum* and the tissues where it plays a role were predicted, thus providing a theoretical basis for the subsequent in-depth study of the application of *BGAL* gene family in *G.hirsutum*.

1 Results and Analysis

1.1 Identification of members of the BGAL genes in cotton

A total of 153 *BGAL* genes were identified and renamed according to the sequence of genes on chromosomes. Among them, there were 51 *G.hirsutum*, named *GhBGAL1~GhBGAL51*. There were 54 varieties of *G.barbadense*, named *GbBGAL1~GbBGAL54*; There were 24 varieties of *G.raimondii*, named *GrBGAL1~GrBGAL24*; There were 24 *G.arboreum*, named *GaBGAL1~GaBGAL24* (Table 1). According to the quantitative analysis of BGAL genes in different species, the number of BGAL family in four cotton species were higher than that in *Arabidopsis*. The number of tetraploid cotton genes were more than twice as many asdiploid cotton, which was consistent with the doubling of tetraploid cotton in BGAL genes. The number of *BGAL* gene family in the diploid cotton species *G.arboreum* and *G. raimondii* are the same, and the positions in chromosomes are roughly the same, indicating high homology. The physicochemical properties of the identified BGAL family showed that the length of the

BGAL family proteinsranged from 335 to 891 amino acid residues (aa). The lengths of *GbBGAL26*, *GaBGAL13* and *GaBGAL23* were all greater than 1 000 aa; The molecular weights of most proteins were in the range of 70~100 kDa, with the largest molecular weight of 170.819 kDa and the smallest molecular weight of 36.612 kDa. The isoelectric points ranged from 4.90 to 9.457, and the number of exons ranged from 9 to 20. According to the prediction of signal peptides, all 139 genes had signal peptides at the n-terminal, while the other 14 genes had no signal peptides. Due to the large number of BGAL gene family, the physical and chemical properties of different cotton species were different, and the physical and chemical properties such as protein length, relative molecular weight and isoelectric point of *G.arboreum* were higher than those of other cotton.

1.2 Phylogenetic analysis of BGAL protein in cotton

In order to understand the evolutionary relationship of *BGAL* gene family in cotton, the contiguous analysis of the amino acid sequences of BGAL family in four cotton species (*G.hirsutum*, *G.arboreum*, *G.barbadense*, and *G. raimondii*) and *Arabidopsis thaliana* was performed using MEGA 7.0 software, and the phylogenetic tree was constructed (Figure 1). According to our results the *BGAL* is divided into 4 large groups A, B, C and D. Group A was divided into five subgroups, A1 to A5, and group C was divided into C1 and C2. Group A has the largest number of family members, 88 in total. Among them, group A1 has the largest distribution, which contains 64 family members. There were only 7 *BGAL* gene members in both A2 and A4 groups, among which there was only one *G.hirsutum* gene *GhBGAL6* in A2 subgroup and two *G.hirsutum* genes in A4 subgroup, respectively *GhBGAL19* and *GhBGAL45*. Group B contains 31 members and has no subfamily classification. There are two subgroups in group C, consisting of 45 family members. Subgroup C1 contains only 3 members of the Arabidopsis *BGAL* genes. There were 40 members of the cotton *BGAL* genes in subgroup C2. In group D, there were 5 cotton *BGAL* family members and 1 *Arabidopsis BGAL* genes.

1.3 Chromosome distribution and collinearity analysis of *BGAL* genes in cotton

The *BGAL* genes were used for chromosome localization (Figure 2). According to our results, it was found that the distribution of genes was uneven, and some chromosomes or certain regions were closely arranged. Among them, 51 *GhBGALs* family genes were located on 20 chromosomes, except A04, A08, A09, D04, D08 and D09. Except for A04, A08, A09, A13, D08 and D09 chromosomes, the other 20 chromosomes of 53 *GbBGALs* genes were distributed in *G.barbadense*. In addition, *GbBGAL50* gene was located in scaffold (D13) fragment, and the location of BGAL family genes in *G.barbadense* and *G.hirsutum* was roughly similar. Suggesting evolutionary similarities between different species; Among the 24 *BGAL* genes in *G.arboreum*, 23 were distributed in Chr02, Chr03, Chr04, Chr05, Chr06, Chr07, Chr10, Chr11, Chr12, Chr13, and *GaBGAL24* existed in scaffold (tig00008658). A total of 24 BGAL members were distributed on the other 11 chromosomes except Chr04 and Chr06.

In order to understand the evolutionary relationship of BGAL family genes, collinear analysis was performed on the BGAL family genes of diploid *G.arboreum*, *G. raimondii* and tetraploid *G.hirsutum* (Figure 3). There were 19, 20 and 21 collinear gene pairs respectively in the subgenomes of *G.arboreum* and *G.hirsutum*, AD of *G.hirsutum*, and between *G.hirsutum* and *G.raimondii*. According to our research, there is a closer evolutionary relationship between *G.hirsutum* and *G.raimondii*. The above results indicated that genome rearrangement of BGAL family genes occurred in the process of polyploidy.

Table 1 Basic information of *BGAL* gene family in cotton

Species	Gene name	Gene ID	Protein length (aa)	Molecular weight (kDa)	Exon number	Signal peptide prediction
Gossypium hirsutum	GhBGAL1	GH_A01G1796	843	94.907	14	Y
	GhBGAL2	GH_A01G1797	843	94.859	14	Y
	GhBGAL3	GH_A02G1374	843	94.013	19	Y
	GhBGAL4	GH_A02G1949	463	52.057	12	Y
	GhBGAL5	GH_A03G2219	674	75.605	16	N
	GhBGAL6	GH_A05G0759	845	92.449	19	Y
	GhBGAL7	GH_A05G1366	871	97.011	19	Y
	GhBGAL8	GH_A05G1369	738	82.103	19	Y
	GhBGAL9	GH_A05G1896	733	82.097	18	Y
	GhBGAL10	GH_A05G2522	704	79.322	18	Y
	GhBGAL11	GH_A05G4025	716	81.228	18	Y
	GhBGAL12	GH_A06G0061	734	82.374	18	Y
	GhBGAL13	GH_A06G0504	818	91.745	19	Y
	GhBGAL14	GH_A06G0505	691	78.171	18	Y
	GhBGAL15	GH_A06G0780	842	93.682	18	Y
	GhBGAL16	GH_A06G0860	335	36.612	9	N
	GhBGAL17	GH_A07G0116	739	82.475	18	Y
	GhBGAL18	GH_A10G0601	846	93.691	19	Y
	GhBGAL19	GH_A10G1370	853	95.892	18	Y
	GhBGAL20	GH_A11G1880	854	95.261	19	Y
	GhBGAL21	GH_A11G3402	856	96.873	16	Y
	GhBGAL22	GH_A12G0009	813	90.951	20	Y
	GhBGAL23	GH_A12G0363	844	95.384	18	Y
	GhBGAL24	GH_A12G2231	746	83.217	17	N
	GhBGAL25	GH_A13G2127	341	39.151	9	N
	GhBGAL26	GH_A13G2188	749	84.666	16	N
	GhBGAL27	GH_D01G1914	843	94.94	14	Y
	GhBGAL28	GH_D01G1915	843	94.882	14	Y
	GhBGAL29	GH_D02G2389	732	81.984	17	Y
	GhBGAL30	GH_D03G0113	854	94.87	18	Y
	GhBGAL31	GH_D03G0592	843	94.199	19	Y
	GhBGAL32	GH_D04G0348	716	81.479	18	Y
	GhBGAL33	GH_D05G0756	845	92.462	19	Y
	GhBGAL34	GH_D05G1370	845	93.613	19	Y
	GhBGAL35	GH_D05G1376	738	82.226	19	Y
	GhBGAL36	GH_D05G1933	733	82.255	18	Y
	GhBGAL37	GH_D05G2544	704	79.216	18	Y
	GhBGAL38	GH_D06G0047	734	82.028	18	Y
	GhBGAL39	GH_D06G0474	697	78.68	18	Y
	GhBGAL40	GH_D06G0761	842	93.754	18	Y
	GhBGAL41	GH_D06G0841	406	45.018	10	Y
	GhBGAL42	GH_D07G0123	738	82.144	18	Y
	GhBGAL43	GH_D10G0226	838	94.084	19	Y
	GhBGAL44	GH_D10G0637	846	94.003	19	Y
	GhBGAL45	GH_D10G1521	890	99.71	19	Y
	GhBGAL46	GH_D12G0008	845	94.174	19	Y
	GhBGAL47	GH_D12G0342	844	95.38	18	Y
	GhBGAL48	GH_D12G2249	817	91.287	19	Y
	GhBGAL49	GH_D13G2108	774	87.331	18	N
	GhBGAL50	GH_D13G2171	856	96.839	16	Y
	GhBGAL51	GH_D13G2531	877	99.153	17	Y

Continued Table 1

Species	Gene name	Gene ID	Protein length (aa)	Molecular weight (kDa)	Exon number	Signal peptide prediction
Gossypium barbadense	GbBGAL1	GB_A01G1909	843	94.907	14	Y
	GbBGAL2	GB_A01G1910	843	94.859	14	Y
	GbBGAL3	GB_A02G1396	843	94.019	19	Y
	GbBGAL4	GB_A02G1974	854	94.902	18	Y
	GbBGAL5	GB_A03G2302	694	77.729	17	Y
	GbBGAL6	GB_A05G0763	845	92.461	19	Y
	GbBGAL7	GB_A05G1374	845	93.863	19	Y
	GbBGAL8	GB_A05G1923	733	82.097	18	Y
	GbBGAL9	GB_A05G2555	704	79.394	18	Y
	GbBGAL10	GB_A05G4115	716	81.3	18	Y
	GbBGAL11	GB_A06G0065	734	82.416	18	Y
	GbBGAL12	GB_A06G0512	818	91.745	19	Y
	GbBGAL13	GB_A06G0513	701	79.058	19	Y
	GbBGAL14	GB_A06G0815	842	93.541	18	Y
	GbBGAL15	GB_A06G0889	335	36.612	9	N
	GbBGAL16	GB_A07G0107	739	82.413	18	Y
	GbBGAL17	GB_A10G0209	838	93.891	19	Y
	GbBGAL18	GB_A10G0606	846	93.691	19	Y
	GbBGAL19	GB_A10G1442	853	95.908	18	Y
	GbBGAL20	GB_A11G1891	854	95.215	19	Y
	GbBGAL21	GB_A11G3483	624	70.724	13	N
	GbBGAL22	GB_A12G0007	845	94.135	19	Y
	GbBGAL23	GB_A12G0370	844	95.428	18	Y
	GbBGAL24	GB_A12G2322	776	86.535	18	N
	GbBGAL25	GB_A13G2262	801	90.599	18	N
	GbBGAL26	GB_A13G2325	1 352	152.026	27	N
	GbBGAL27	GB_D01G2004	843	94.907	14	Y
	GbBGAL28	GB_D01G2005	843	94.891	14	Y
	GbBGAL29	GB_D02G2446	732	81.964	17	Y
	GbBGAL30	GB_D03G0111	854	94.898	18	Y
	GbBGAL31	GB_D03G0592	843	94.227	19	Y
	GbBGAL32	GB_D04G0356	716	81.479	18	Y
	GbBGAL33	GB_D05G0751	845	92.462	19	Y
	GbBGAL34	GB_D05G1378	822	91.044	19	Y
	GbBGAL35	GB_D05G1385	738	82.226	19	Y
	GbBGAL36	GB_D05G1950	733	82.229	18	Y
	GbBGAL37	GB_D05G2555	701	79.179	18	Y
	GbBGAL38	GB_D06G0067	734	82.055	18	Y
	GbBGAL39	GB_D06G0504	818	91.794	19	Y
	GbBGAL40	GB_D06G0505	696	78.665	18	Y
	GbBGAL41	GB_D06G0803	842	93.764	18	Y
	GbBGAL42	GB_D06G0888	406	45.029	10	Y
	GbBGAL43	GB_D07G0116	737	82.028	18	Y
	GbBGAL44	GB_D10G0215	838	94.084	19	Y
	GbBGAL45	GB_D10G0632	846	93.86	19	Y
	GbBGAL46	GB_D10G1505	890	99.756	19	Y
	GbBGAL47	GB_D11G1938	854	95.113	19	Y
	GbBGAL48	GB_D12G0009	845	94.222	19	Y
	GbBGAL49	GB_D12G0347	844	95.38	18	Y
	GbBGAL50	GB_D12G2325	817	91.229	19	Y
	GbBGAL51	GB_D13G2210	891	100.484	19	Y
	GbBGAL52	GB_D13G2276	856	96.769	16	Y
	GbBGAL53	GB_D13G2624	877	99.169	17	Y
	GbBGAL54	GB_D13G2752	877	99.303	17	Y

Continued Table 1

Species	Gene name	Gene ID	Protein length (aa)	Molecular weight (kDa)	Exon number	Signal peptide prediction
Gossypium arboretum	GaBGAL1	Ga02G0116	854	94.789	18	Y
	GaBGAL2	Ga02G0696	843	94.069	19	Y
	GaBGAL3	Ga02G1107	843	94.905	14	Y
	GaBGAL4	Ga02G1108	843	94.905	14	Y
	GaBGAL5	Ga03G2554	717	80.305	17	Y
	GaBGAL6	Ga04G1771	716	81.386	18	Y
	GaBGAL7	Ga05G0795	845	92.515	19	Y
	GaBGAL8	Ga05G1432	859	95.624	19	Y
	GaBGAL9	Ga05G1436	738	82.132	19	Y
	GaBGAL10	Ga05G2010	733	82.07	18	Y
	GaBGAL11	Ga05G2676	665	75.104	19	Y
	GaBGAL12	Ga06G0075	734	82.416	18	Y
	GaBGAL13	Ga06G0454	1 442	162.587	36	Y
	GaBGAL14	Ga06G0745	819	91.171	19	Y
	GaBGAL15	Ga07G0121	738	82.293	18	Y
	GaBGAL16	Ga10G1511	890	99.971	19	Y
	GaBGAL17	Ga10G2452	846	93.691	19	Y
	GaBGAL18	Ga11G0419	838	95.085	16	Y
	GaBGAL19	Ga11G2096	842	93.936	19	Y
	GaBGAL20	Ga12G0769	781	87.485	20	Y
	GaBGAL21	Ga12G2694	853	96.562	20	N
	GaBGAL22	Ga13G2300	795	90.071	19	N
	GaBGAL23	Ga13G2368	1 512	170.819	33	N
	GaBGAL24	Ga14G0924	829	93.606	17	Y
Gossypium raimondii	GrBGAL1	Gorai.001G011600	738	82.291	18	Y
	GrBGAL2	Gorai.002G197100	754	85.441	14	Y
	GrBGAL3	Gorai.002G197200	843	94.905	14	Y
	GrBGAL4	Gorai.003G011900	854	94.746	18	Y
	GrBGAL5	Gorai.005G246600	732	81.963	17	Y
	GrBGAL6	Gorai.007G198900	859	95.819	19	Y
	GrBGAL7	Gorai.008G000600	845	94.389	20	Y
	GrBGAL8	Gorai.008G036000	844	95.309	18	Y
	GrBGAL9	Gorai.008G221000	797	89.279	19	Y
	GrBGAL10	Gorai.009G078400	845	92.564	19	Y
	GrBGAL11	Gorai.009G140800	845	93.659	19	Y
	GrBGAL12	Gorai.009G141300	738	82.197	19	Y
	GrBGAL13	Gorai.009G198100	733	82.194	18	Y
	GrBGAL14	Gorai.009G265200	699	78.656	18	Y
	GrBGAL15	Gorai.010G007300	734	82.196	18	Y
	GrBGAL16	Gorai.010G052200	818	91.819	19	Y
	GrBGAL17	Gorai.010G052300	591	66.492	16	Y
	GrBGAL18	Gorai.010G082900	843	93.962	18	Y
	GrBGAL19	Gorai.011G021800	838	94.072	19	Y
	GrBGAL20	Gorai.011G065300	846	93.858	19	Y
	GrBGAL21	Gorai.011G155500	890	99.933	19	Y
	GrBGAL22	Gorai.012G037700	716	81.437	18	Y
	GrBGAL23	Gorai.013G216400	829	93.61	17	Y
	GrBGAL24	Gorai.013G222800	794	89.91	16	Y

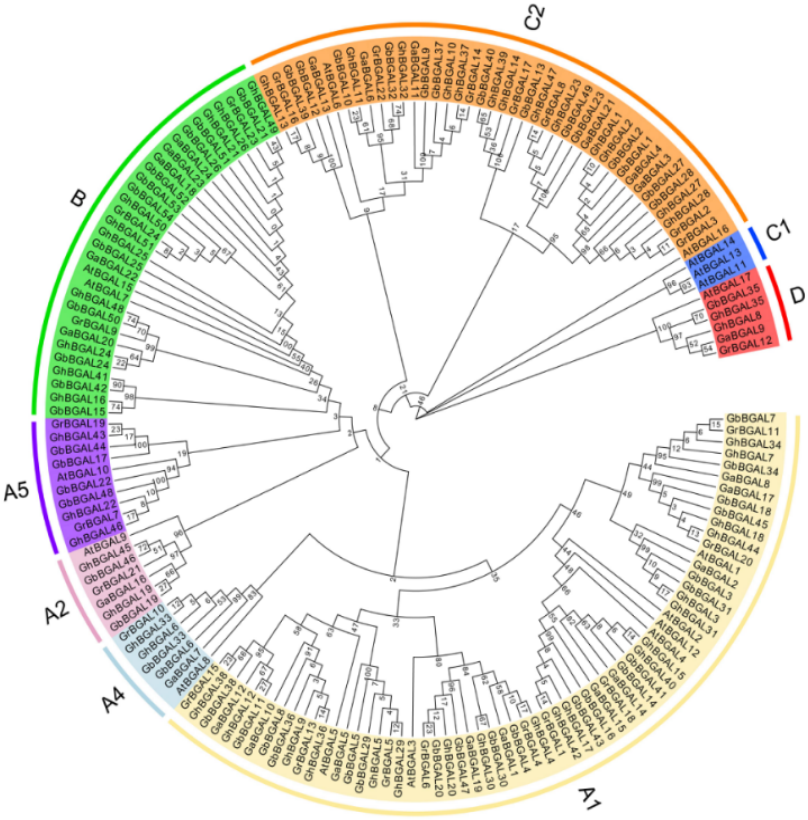

Figure 1 phylogenetic tree of *BGAL* gene members

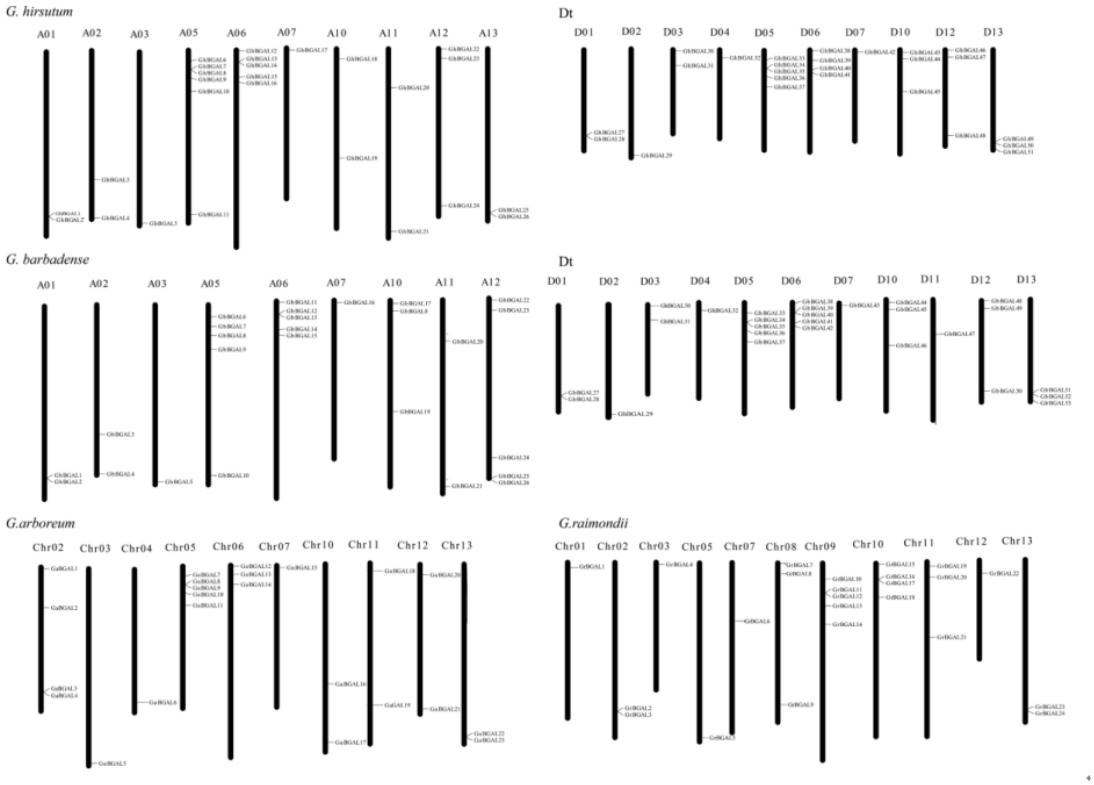

Figure 2 Chromosomal distribution of *BGAL* genes in cotton

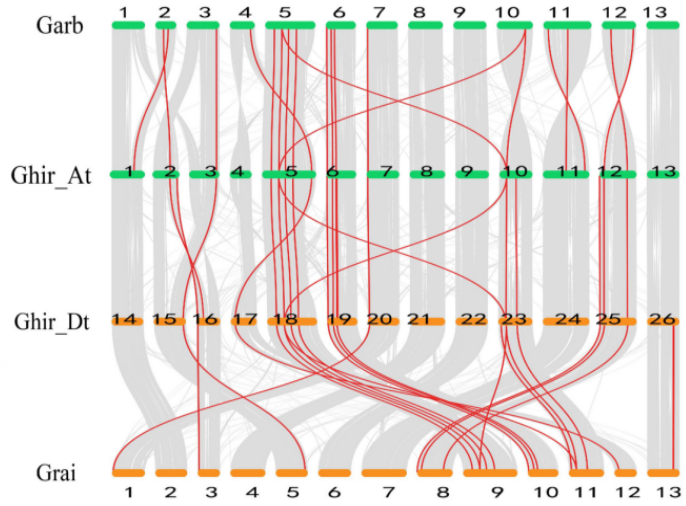

Figure 3 The collinearity of *BGAL* genes in the A genome of *G.barbadense*, the AD subgenome of *G.hirsutum* and the D genome of *G.arboreum*

Note: The gray lines represent collinear relationships within different genomes, and the red lines represent collinear gene pairs in the *BGAL* genes

1.4 Sequence alignment and structure analysis of *GhBGALs* genes

In order to further understand the evolution of the *BGAL* family of *G.hirsutum*, the exon-intron structure of 51 *BGAL* genes of *G.hirsutum* was studied (Figure 4), and it was found that the number of exons in BGAL genes of *G.hirsutum* was high. Except for 9 exons in *GhBGAL16* and *GhBGAL25* genes, all other genes have 10 to 20 exons (including 4 exons in 16 exons, 3 exons in 17 exons, 19 exons in 18 exons, 16 exons in 19 exons, and 1 exon in 20 exons). The number of introns in the *BGAL* genes is large, and the distribution of introns in most genes is very dense. Introns of A2, A5, and some of the C2 subfamily genes are scattered. Analysis results of the conserved motifs of *G.hirsutum BGALs* showed that motif 3, motif 4, motif 5 were found in all 7 subfamilies of *G.hirsutum*, indicating that these three motifs are the most conserved motifs in the *BGALs* genes, but the function of these conserved motifs remains to be studied. In addition, among the seven subgroups, subgroup C and subgroup A have the most motif species, but subgroup A has one more motif 6 than subgroup C, and the two subgroups A and C may be the most similar in function. The results showed that all *GhBGALs* genes in the same subfamily had similar gene structure and conserved motif, which strongly supported the reliability of phylogenetic and evolutionary classification (Figure 4).

1.5 Analysis of promoter elements of *GhBGALs*

In order to understand the transcriptional regulation and potential function of *GhBGALs*, it is important to study the cis-acting elements in its promoter region. *GhBGALs* has many cis-acting elements, which can be roughly divided into three categories: A, B and C (Figure 5). A is the plant hormone response element, including auxin response element, gibberellin response element, salicylic acid response element, abscisic acid response element, methyl jasmonate response element and flavonoid response element. B is stress response element, including stress response element, drought response element, damage response element and low temperature response element; C is other responsive elements, including those related to the development of palisade mesophyll tissue, photoperiod regulatory elements, seed development, meristem development, and endosperm development. Among all *GhBGALs* cis-acting elements (Figure 6), the number of plant hormone-responsive elements was the largest, and almost every *GhBGAL* genes had one or two plant hormone-responsive elements, among which abiotic acid response elements and methyl jasmonate response elements were the largest. Among the stress response elements, drought response elements and low temperature response elements were the most numerous, and one third of *GhBGALs* promoters contained these two stress response elements. The number of other response elements is relatively small, generally only exist in a few genes, such as *GhBGAL1*, *GhBGAL2*, *GhBGAL5*, *GhBGAL13*,

GhBGAL115 genes photoperiod regulatory elements; Seven *GhBGALs* promoter sequences contained expression response elements in endosperm.

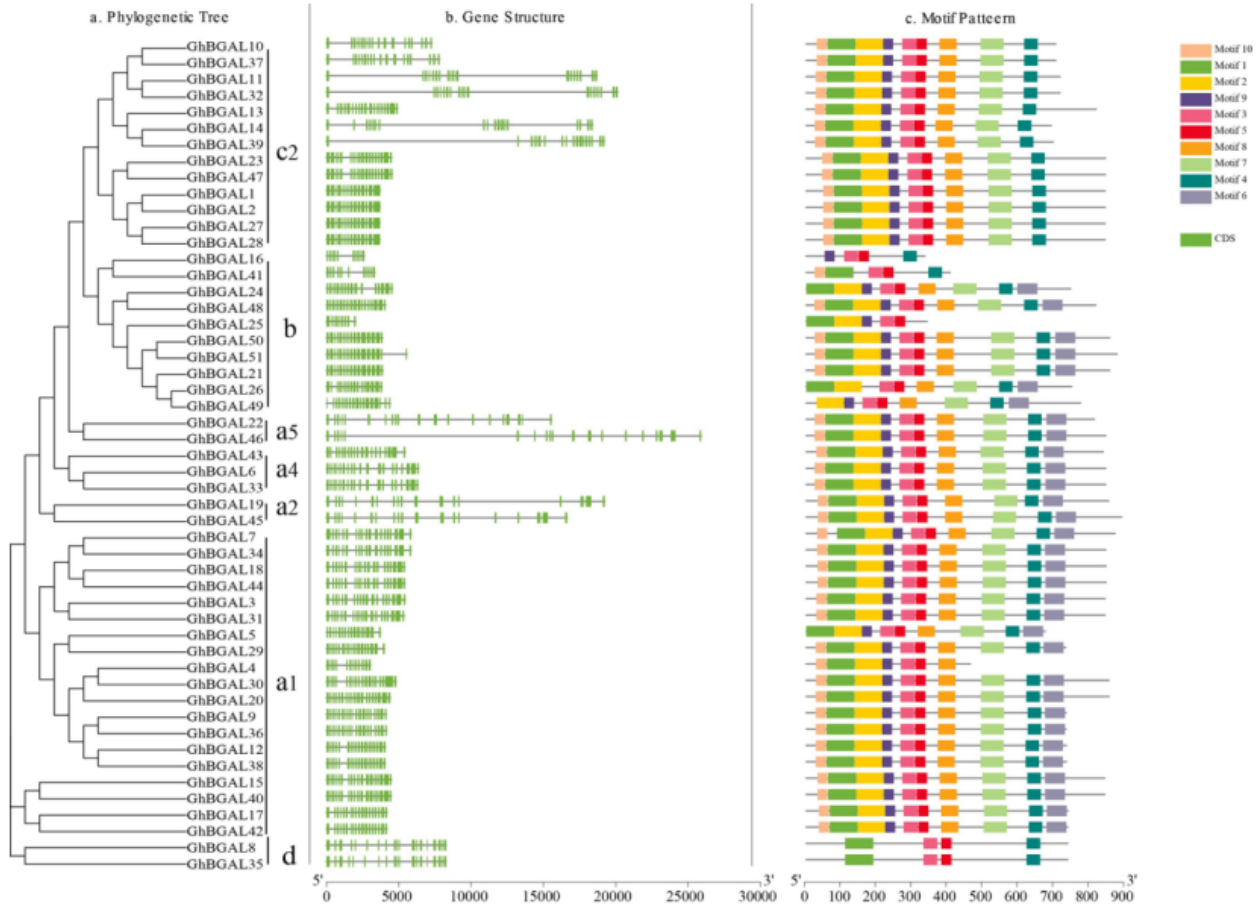

Figure 4 phylogenetic tree, conserved motif and gene structure of *BGAL* proteins in *G.hirsutum*

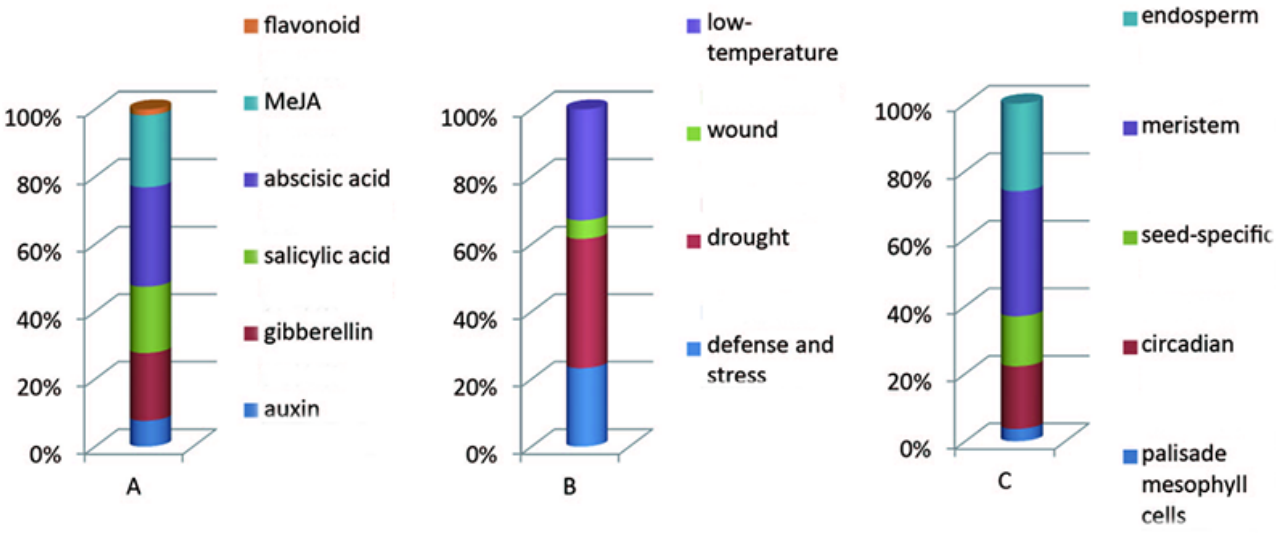

Figure 5 *GhBGALs* promoter cis-acting element

Note: A: Hormone response element; B: Stress response element; C: Other response elements

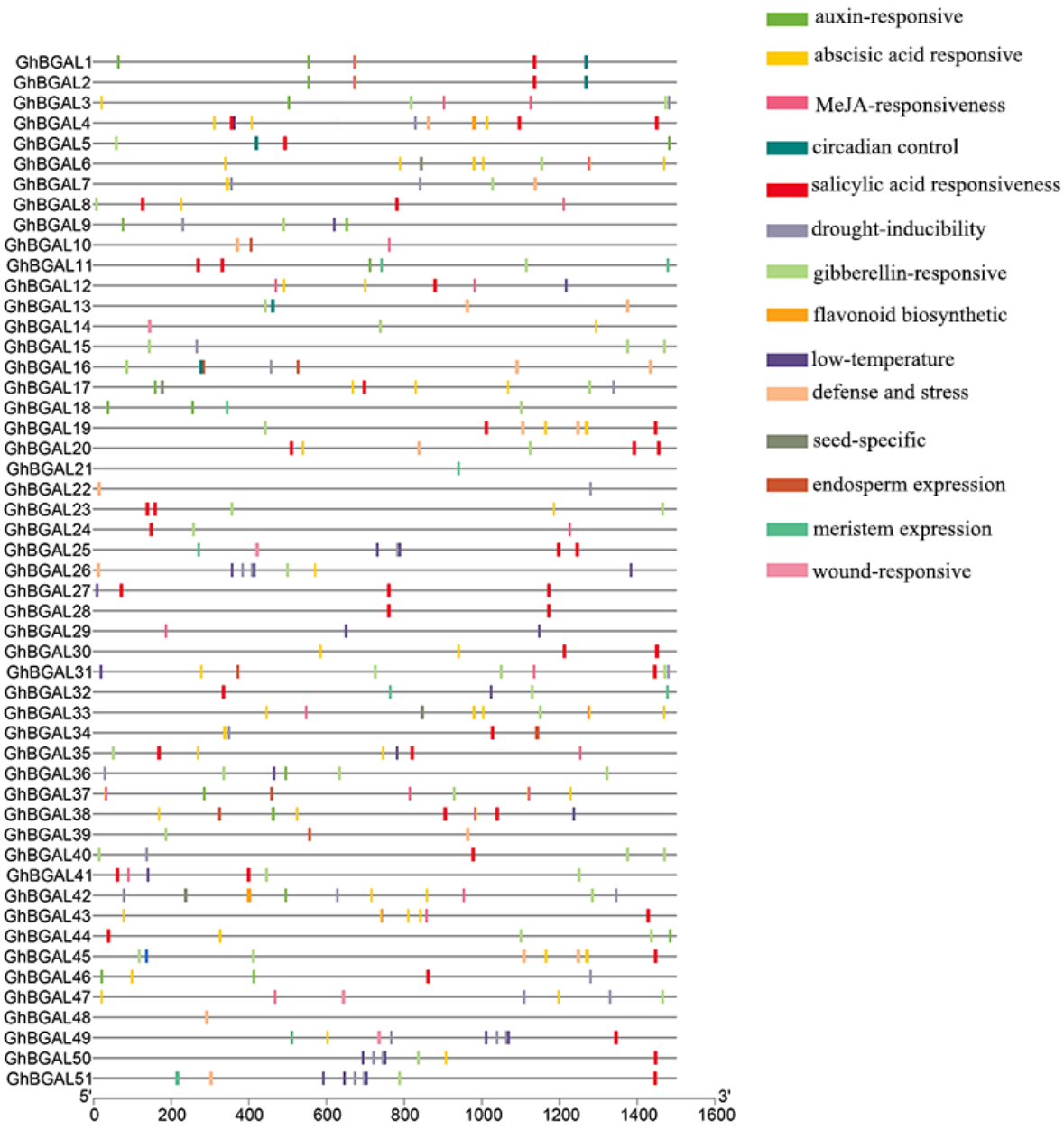

Figure 6 *GhBGALs* promoter cis-acting element

1.6 Tissue specific expression pattern and qRT-PCR analysis of *GhBGALs*

In order to understand the expression of BGAL genes in *G.hirsutum*, 51 *BGAL* genes in different tissues (anther, filaments, pistil, bracts, sepals, petals, torus, roots, leaves, and stems) in *G.hirsutum* were analyzed based on tissue transcriptome data (Figure 7). According to the results, different genes (such as *GhBGAL43* and *GhBGAL38*) were expressed significantly differently in the same tissues (such as anther, filaments and petals), and the same genes (such as *GhBGAL42*) were expressed significantly differently in different tissues (such as stamen and pistil). The expression pattern graph was clustered according to rows, and the overall analysis of genes expression differences could be divided into three parts: high expression, no expression and low expression. The expressions of 22 genes from *GhBGAL46* to *GhBGAL2* in all tissues showed no significant difference and were all in the state of low expression or even no expression. From *GhBGAL42* to *GhBGAL43* at the top of the pattern diagram, these 9 genes were highly specifically expressed in stamens (anther and filament), petals and sepals, and were less expressed in other tissues. The 20 genes from *GhBGAL38* to *GhBGAL37*, shown at the bottom of the pattern diagram, were generally expressed at low levels in all tissues. *GhBGAL6* and *GhBGAL32* were highly expressed in all tissues.

GhBGAL12 and *GhBGAL37* were specifically expressed only in receptacle and pistil. The two genes, GhBGAL6 and GhBGAL32, were not expressed only in stems, and their expressions were low in other tissues. *GhBGAL33*, *GhBGAL7*, *GhBGAL18* and *GhBGAL43* were highly expressed in stamens (anthers and filaments) and petals. *GhBGAL22* was highly expressed in the pistil. *GhBGAL20* was highly expressed in roots and stems, but low in other tissues. *GhBGAL29* was highly specifically expressed in leaves. *GhBGAL38* was specifically expressed in torus, root and stem, but hardly expressed in other tissues.

In order to further understand the effect of *BGAL* y genes on plant tissues of *G.hirsutum*, 12 highly expressed *BGAL* genes were randomly selected in anther, filaments, pistil, bracts and sepals for qRT-PCR analysis (Figure 8). The gene *GhBGAL7* is specifically expressed only in the stamens, and the expressions of *GhBGAL6*, *GhBGAL7*, *GhBGAL17*, *GhBGAL18*, *GhBGAL33*, *GhBGAL41*, *GhBGAL42* and *GhBGAL43* in the stamens are significantly higher than those in the pistils, bracts and sepals. These genes may regulate the development of stamens. The primers used in this study are as follows (Table 2).

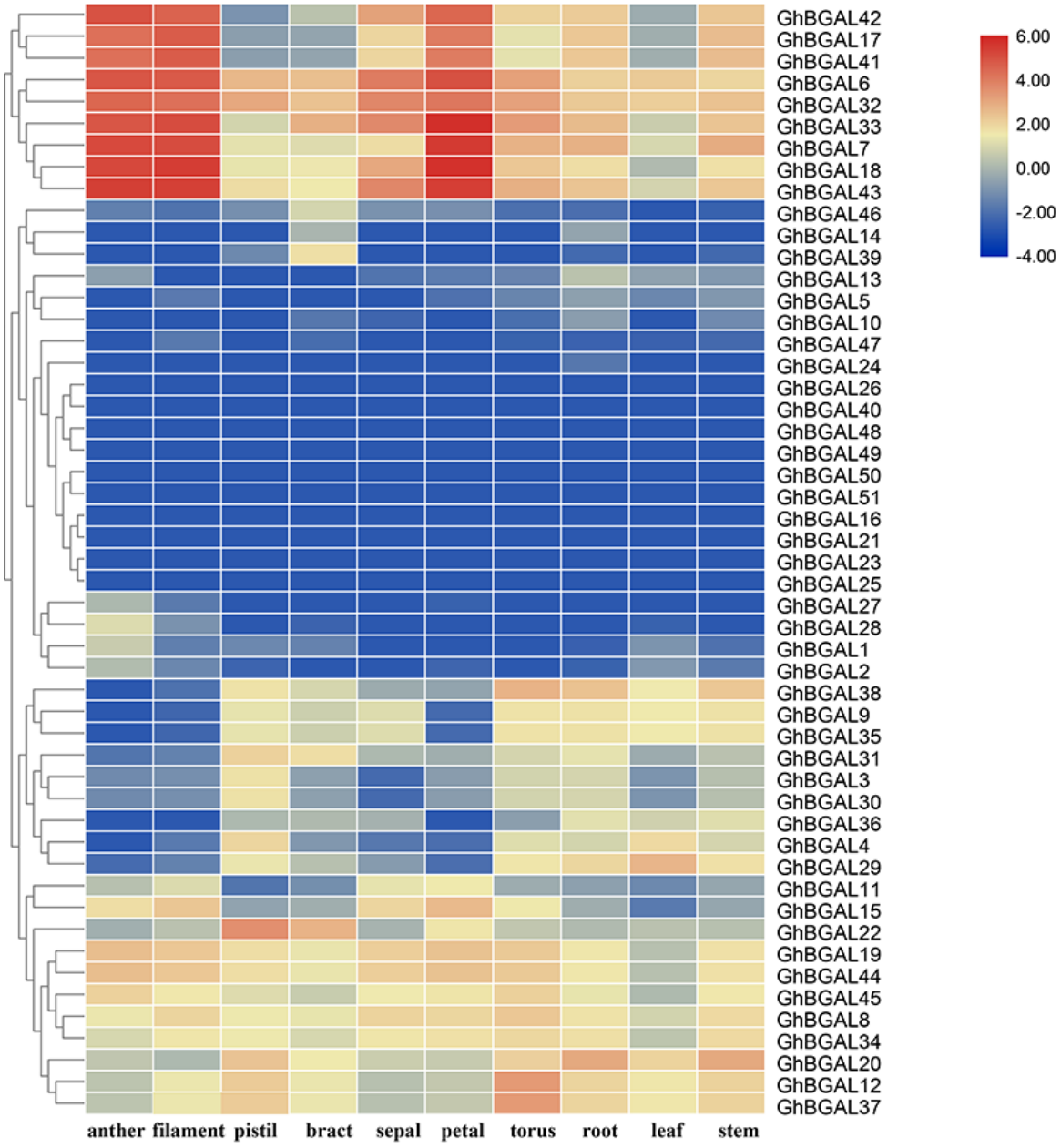

Figure 7 *GhBGALs* cluster expression pattern in different tissues and organs of *G.hirsutum*

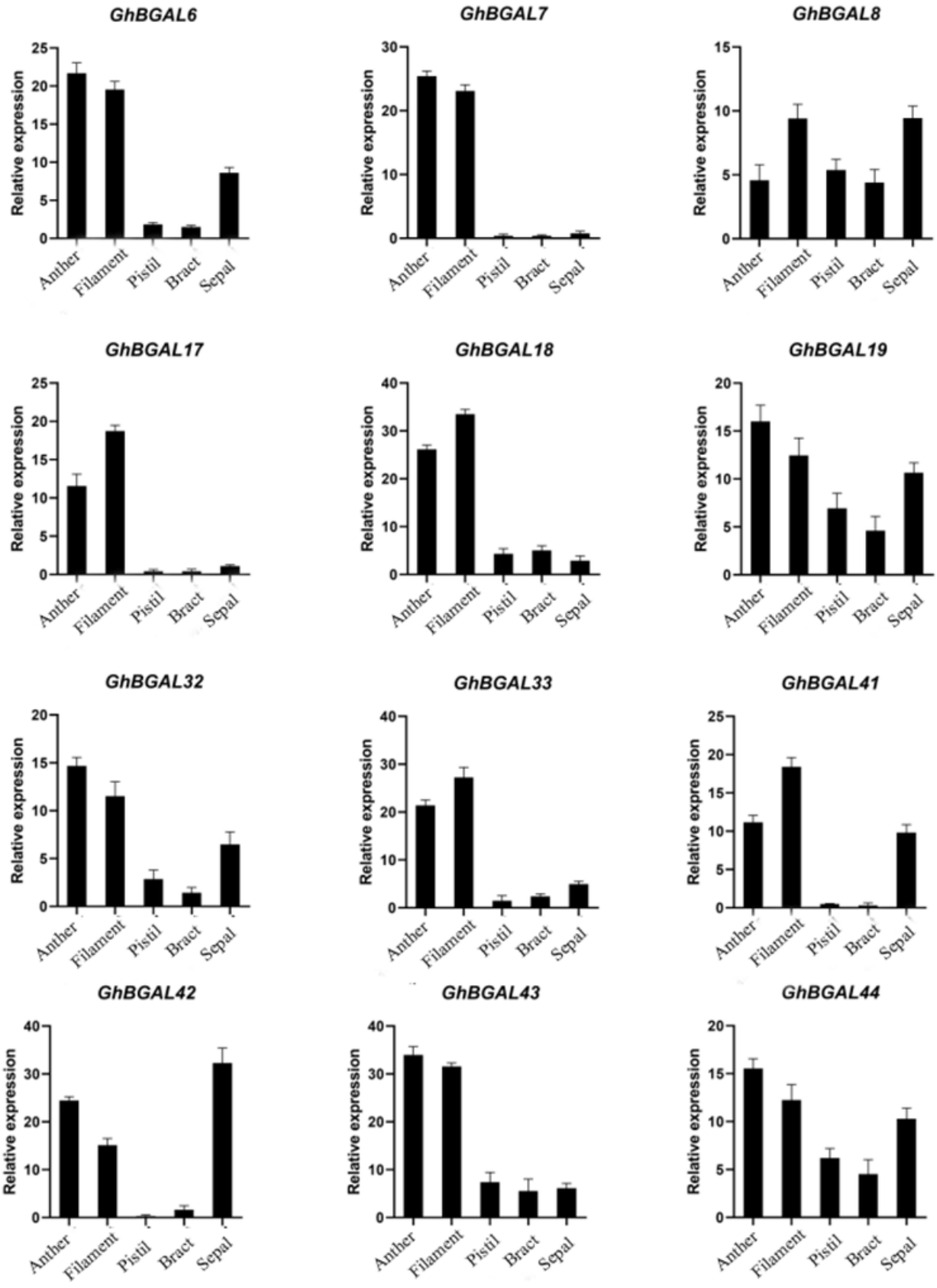

Figure 8 The expression level of *GhBGALs* in different tissues

Table 2 Primers for qRT-PCR

Gene name	Forward primers (5'-3')	Reverse primers (5'-3')'
GhBGAL6	CCAGAACTGGGGTGGAACATGC	AGCGCTAGAAAGGTCATCACGC
GhBGAL7	CAAGTACGACCACCACACTCCC	TGCGTTGGCGTTAGAAGAGGAG
GhBGAL8	ATGAATCGAAGTAAAAGCCTAG	TCAAAAAGTACGCAGCCCTAGT
GhBGAL17	TACCTGCTGTGACTGTCCCTGT	TTCCGATGTCCGGTTTCGTTCC
GhBGAL18	AAATGGTGGTTTTTGAACAAAG	TTAGAACTGGGAGACTCCAATA
GhBGAL19	ATGAAGGTGAGGGAGATGGTAT	TCACTGAGGGAGATGATGGTAT
GhBGAL32	ACCTGATTCTGAAGGGACCGACT	AAGCAGGTAGGGGACTCCTGTC
GhBGAL33	AAGTGGTGATTGCGAGCTCTGG	CCTTAACTGAGTGGCCGTCTCG
GhBGAL41	CACGCTATCCATCAGAGCGACC	GGACTTTGGCCATCTGGCAACT
GhBGAL42	TCGGTGGATCAGACATTGCGTT	CTTCCTTCATGATACCTCTCTCGGT
GhBGAL43	GGGCCATCTTTCCCCATGACAA	GCCCCAGGGATGGCATACTTTT
GhBGAL44	GGGCCATCTTTCCCCATGACAA	GCCCCAGGGATGGCATACTTTT

2 Discussion

β-galactosidase can hydrolyze pectin, and has the function of softening fruit and promoting cell wall metabolism during fruit ripening. In this study, 153 *BGAL* genes were identified from *G.hirsutum*, *G.barbadense*, *G. raimondii* and *G.hirsutum* by bioinformatics, including 51 in *G.hirsutum* and 17 in *Arabidopsis thaliana*. The *BGAL* genesy has its typical sequence structure, and its sequence is GGP (LIVW)2-X(2)-Q-X-E-N-E. Multiple sequence comparison and analysis of the protein sequence of *BGAL* genes in cotton showed that the typical sequence of 5 genes was incomplete and part of Cys residues were missing. The other 148 genes in the cotton family all had relatively complete typical sequences, which were the sites of β-galactosidase specific binding substrates (Ahn et al., 2007). The protein sequence alignings supported the identification results of genes members.

In the phylogenetic analysis, most of the identified *BGAL* genes in cotton appear in pairs with high homology, and these genes may also have certain similarities in function. The 153 *BGAL* genes of cotton were divided into 8 subgroups, consistent with the grouping of *Arabidopsis thaliana*, indicating that *BGAL* genes of cotton have homology with *Arabidopsis thaliana* genes (Figure 2), but some protein sequences show differences in the evolutionary process, which may be related to subfunctionalization and natural selection of species. There is no subgroup A3 in the phylogenetic tree group because it is a bryophyte specific cluster (Ahn et al., 2007). There are no *BGAL* genes in the C1 subfamily, but only *AtBGALs* genes, which may be due to the lack of conserved protein sequences in this subfamily, and the loss of *BGAL* genes in the evolution of cotton. The number of genes in subgroup D is at least six, but contains *BGAL* genes of each species, which may be related to the conservation of genes in subgroup D. Subgroup A1, the most intensively studied among other species, encodes β-galactosidase, hydrolyzing β-(1,3)-and β-(1,4)-lactose oligosaccharide in cell walls.

Chromosome location analysis showed that the BGAL family members of *G.hirsutum*, *G.arboreum* and *G.barbadense* were randomly distributed on different chromosomes, while the 26 BGAL members of *G.raimondii* were mainly distributed on chromosomes 8, 9, 10 and 11. The number of genes in the tetraploid cotton species *G.hirsutum* and *G.barbadense* was twice that of the diploid cotton species *G. raimondii* and *G.arboreum*. It is possible that the *BGAL* y genes were not lost in the evolution of cross doubling in the tetraploid cotton species *G.hirsutum* and *G.barbadense*. In order to better understand the relationship between cotton species evolution, collinearity analysis was performed between different cotton species and AD subgenomes of *G.hirsutum*. It was found that the phenomenon of chromosome dislocation and even inversion existed in cotton, and there were coevolutionary gene pairs at some chromosomal sites, which may also have similar functions.

Analysis of conserved domain results showed that except for subgroup D, all *GhBGALs* genes contained motif 1, and the structures of the three subgroups A, B, and C were almost all similar, indicating that they may have similar functions. The gene expression pattern diagram could be systematically analyzed according to subgroup classification. Analysis of cis-acting elements in *GhBGALs* promoters reveals that almost every gene has a plant hormone responsive element. Plant hormones such as auxin and methyl jasmonate play an important role in the

regulation of plant growth and development. The regulation of MeJA in the process of flower development mainly involves filament elongation, pistil development, anther development and anther dehiscing and other physiological processes. MeJA plays an important role in regulating anther dehysis in many plants, such as wheat, rice and rapeseed. Therefore, it is speculated that the response elements of plant hormones in the promoter sequence of *GhBGALs* may regulate the anther development of cotton.

Expression pattern analysis, *GhBGAL14* and *GhBGAL39* in cotton organizations almost all is in not express state, in the analysis of sequence alignment, the two gene sequences of BGAL typical incomplete, lead to can't specific to identify the role of β-galactosidase glucoside enzyme loci, cannot provide the energy for plant growth and development, not express in various organizations. β-galactosidase glucoside enzyme catalyzed by the cell walls of large and complex side chains on the metabolism of galactose and metabolism in the cell wall, the family of *BGALs* specific roles in plant glucan of pectin in plant cell walls, wood or Arab galactose protein degradation, in the process of mature plants, *BGAL*sexpansion of cell wall conduction, degradation and signal molecules play an important role. BGAL family in mature fruit and seed development process of more, such as tomato and *arabidopsis thaliana*, a new study suggests that the yeast expression tomato TBG4 can hydrolysis of plant cell wall substrates, alkali soluble in tomato fruits during ripening stage and chelating activity of pectin mol, highest homologous genes *AtBGAL4* as hydrolysis enzyme involved in the degradation of pectin, *GhBGAL17* and *GhBGAL42*, which have the highest homology with *GhBGAL17* and *GhBGAL42* in *G.hirsutum*, are highly expressed in the reproductive organs of cotton, and these two genes may affect the development of flower organs by specifically hydrolyzing the pectin components in the cell wall or pollen wall of flowers.

Studies in recent years have found that the BGAL family influences pollen development and fertility of plants by regulating the development of pollen wall, and genes such as β-galactosidase participate in the degradation of pollen wall during the anabolic process of pollen wall. *AtBGAL16* in Arabidopsis is expressed specifically in mature pollen grains, and *GhBGAL32* in *G.hirsutum* is expressed specifically in stamens (filaments and anthers), suggesting that this gene may regulate anther development to some extent.*AtBGAL7* and *AtBGAL15* belong to group B genes in *Arabidopsis thaliana*, which, like *OsBGAL5*, *OsBGAL12*, *OsBGAL14* and *OsBGAL15* genes of group B in rice, play an important role in the early microspore stage of pollen development and pollen development. *OsBGAL15* is highly expressed in flower organs. Studies on BGAL family transcription in fertile and sterile Chinese cabbage (Liu et al., 2013) found that *BcBGAL7* and *BcBGAL7* were specifically expressed in the anthers of fertile plants. In the expression pattern of *G.hirsutum*, *GhBGAL41* in the B subgroup of *GhBGALs* is highly expressed in the anther (anther and filament) and petal, which is consistent with the expression in *Arabidopsis thaliana*, rice and Cabbage. It is suggested that *GhBGAL41* may affect the development of pollen by regulating the development of pollen wall of cotton, which provides a theoretical basis for the study of male sterility of cotton.

3 Materials and Methods

3.1 Identification of BGAL family members and analysis of physical and chemical properties

Download the genomic, CDS and protein sequences of the four cotton species (*Gossypium hirsutum*, ZJU; *Gossypium arboretum*, JGI; *Gossypium barbadense*, ZJU; *Gossypium raimondii*, CRI) from Cotton Functional Genomics Database (CottonFGD) (https://cottonfgd.org/). Download biological information on other species such as Arabidopsis from the JGI database (http://www.phytozome.net). Download the configuration file of the GH35 conservative domain (PF01301) HIDDEN Markov Model (HMM) from the Pfam database (http://pfam.xfam.org). HMMER 3.0 and BLASTP were used to search for *BGAL* genes in the genomes of cotton and other species. Redundant genes were removed from the HMM and BLASTP results. The remaining genes were further identified by SMART (http://smart.embl-heidelberg.de/). The physicochemical properties of proteins such as amino acid length and isoelectric point (pI) of all BGAL family members of *G.hirsutum*, Sea *G.arboreum*, *G.arboreum* and *G. raimondii* were retrieved from the Cotton Functional Genomics Database (Cotton FGD) (https://cottonfgd.org/). Using cNLS Mapper (http://nls-mapper.iab.keio.ac.jp/cgi-bin/NLS_Mapper_form.cgi) and NetNES1.1 server (http://www.cbs.dtu.dk/services/NetNES/) to check and ratify a signal BGAL gene (NLS) and nuclear export signal (NES).

3.2 Phylogenetic tree construction of *BGAL* genes

The identified BGAL gene members were searched and extracted for the protein sequences of cotton and Arabidopsis by HMMER 3.0 and BLASTP, and multiple sequence alignings were performed in MEGA 7.0 software. Use online software Evolview (https://www.omicsclass.com/article/671), beautify the evolutionary tree.

3.3 Chromosome localization and coevolutionary analysis of the *BGAL* genes

Biological information such as location and structure of BGAL family members were extracted from the gff3 annotation files of the genome of four cotton species. MapChart2.2 software was used to analyze and map the position information of *BGAL* genes on chromosomes in cotton. In order to reveal the coevolutionary relationship between *BGAL* families among cotton species, Circos was used to construct a coevolutionary analysis diagram. According to the sequence length of the alignment covering more than 70%, the similarity of the alignment region more than 70%, the gene duplication was determined. Beautify with Adobe Illustrator CC 2019.

3.4 Gene structure and conserved motif analysis of *GhBGALs*

Use of MEME (http://meme.sdsc.edu/meme/) for *GhBGALs* conservative motif is analyzed. Use online software GSDS (http://gsds.cbi.pku.edu.cn) gene exon-contains substructure analysis *GhBGALs*. The analyzed sequence alignment result files, exon-intron structure files and conservative domain files were combined and visualized using TBtools.

3.5 Analysis of promoter elements of *GhBGALs*

In order to study the relationship between BGAL family and hormones and stress, the sequence of 1 500 bp upstream of initiation codon (ATG) in BGAL family was selected from the genome sequence of *G.hirsutum*. Using PlantCARE database (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/) to identify and analyze gene upstream region of cis elements, after screening using TBtools to visualize it. In order to facilitate statistical analysis, all components are classified and sorted, and histogram is made.

3.6 Tissue expression pattern analysis of *GhBGALs*

To understand the expression of *GhBGALs* in different tissues. From NCBI SRA (Sequence read archive) database (http://www.ncbi.nlm.nih.gov/sra) download *G.hirsutum* in different organizations (PRJNA248163) transcriptome sequencing data. Based on the characteristics of *GhBGALs* family, the expression levels of representative tissues such as roots, stems, leaves, anthers, filaments, pistils, bracts, sepals, petals, and torus were selected to analyze the original data. The expression levels of genes were calculated by FPKM, and the data were visualization by TBtools.

3.7 RNA isolation and quantitative reverse transcription-polymerase chain reaction (qRT-PCR) of *GhBGALs*

G.hirsutum 'cv CCRI24' provided by Institute of Cotton Research of the Chinese Academy of Agricultural Sciences was planted in Anyang greenhouse, sampled at full flowering stage and stored at -80℃. Total RNA was extracted using the RNAprep Pure Plant Plus kit (Tiangen, Beijing, China) according to the instructions. 2 μg total RNA was reverse transcribed into cDNA using PrimeScript first strand cDNA synthesis Kit (TaKaRa, Dalian, China), and diluted for later use. Real-time PCR was performed using SYBR premixed Ex Taq (TaKaRa, Dalian, China) in ABI 7500 system (Applied Biosystems, Foster City, CA, USA). The system consisted of 20 μL, 10 μL SYBR Green PCR mix, 0.5 μL upstream and downstream primers, 2 μL diluted cDNA, and 7 μL ddH_2O. The reaction procedure of qRT-PCR was: 94℃ for 30 s; Cycle stage: 94℃ 5 s, 55℃ 15 s, 72℃ 10 s, 45 cycles; Dissolution curve stage: 94℃ 15 s, 60℃ 15 s, 95℃ 15 s, 4℃ storage. $2^{-\Delta\Delta CT}$ method was used to calculate the relative expression level of genes using Actin as the internal reference. Each tissue and each gene expression response had 3 biological replicates and 3 technical replicates.

Authors' contributions

CXC is the experimental design and experimental research executor of this study; CXC completed the data analysis and wrote the first draft of the paper; ZCJ, GHQ, WXY, MQF and QKK participated in experimental design and experimental result analysis; ZGY

and FSL were the architects and principals of the project, guiding experimental design, data analysis, paper writing and modification. All authors read and approved the final manuscript.

Acknowledgments

This research was supported by the National Natural Science Foundation of China for Young Scholars (31701474).

Reference

Ban Q.Y.., Han Y., He Y.H..,Jin M.J., Han S.K., Suo J.T., Rao J.P., 2018, Functional characterization of persimmon β-galactosidase gene *DkGAL1* in tomato reveals cell wall modification related to fruit ripening and radicle elongation, Plant Science, 274: 109-120
https://doi.org/10.1016/j.plantsci.2018.05.014

Smith D.L., and Gross K.C., 2000, A family of at least seven β-Galactosidase genes is expressed during tomato fruit development, Plant Physiology, 123(3): 1173-1183
https://doi.org/10.1104/pp.123.3.1173

Guo S.L., Song J., Zhang B.B. Jiang H., Ma R.J., Yu M.L. 2018, Genome-wide identification and expression analysis of beta-galactosidase family members during fruit softening of peach [*Prunus persica* (L.) Batsch], Postharvest Biology and Technology, 136: 111-123
https://doi.org/10.1016/j.postharvbio.2017.10.005

Henrissat B., and Bairoch A., 1993, New families in the classification of glycosyl hydrolases based on amino acid sequence similarities, Biochemical Journal, 293: 781-788
https://doi.org/10.1042/bj2930781

Hrubá P., Honys D., Twell D., Capková V., and Tupy J., 2005, Expression of *β*-galactosidase and *β*-xylosidase genes during microspore and pollen development, Planta, 220(6): 931-940
https://doi.org/10.1007/s00425-004-1409-0

Lazan H., Ng S.Y., Goh L.Y., and Ali Z.M., 2004, Papaya beta-galactosidase/galactanase isoforms in differential cell wall hydrolysis and fruit softening during ripening, Plant Physiology and Biochemistry, 42(11): 847-853
https://doi.org/10.1016/j.plaphy.2004.10.007

Liu J.L., Gao M.H., Lü M.L., and Cao J.S., 2013, Structure, evolution, and expression of the β-galactosidase gene family in *Brassica campestris* ssp. *chinensis*, 31(6): 1249-1260
https://doi.org/10.1007/s11105-013-0585-0

Lombard V., Golaconda R.H., Drula E., Coutinho P.M., and Henrissat B., 2014, The carbohydrate-active enzymes database (CAZy) in 2013, Nucleic Acids Research, 42: D490-495
https://doi.org/10.1093/nar/gkt1178

Lu Q., Shao F.J., and Qiu D.Y., 2018, Genome-wide analysis of gene family of lateral organ boundaries domain in *Populus trichocarpa*, Jiyinzuxue yu Yingyong Shengwuxue (Genomics and Applied Biology), 37(1): 313-325

Moctezuma E., Smith D.L., and Gross K.C., 2003, Antisense suppression of a β-galactosidase gene (TBG6) in tomato increases fruit cracking, J. Exp. Bot., 54(390): 2025-2033
https://doi.org/10.1093/jxb/erg214

Sørensen S.O., Pauly M., Bush M., Skjet M., Maureen M.C., Bernhardt B., Ulvskov P., 2000, Pectin engineering: modification of potato pectin by *in vivo* expression of an endo-1,4-*β*-D-galactanase, 97(13): 7639-7644
https://doi.org/10.1073/pnas.130568297

Rogers H.J., Bate N., Combe J., Sullivan J., Sweetman J., Swan C., Lonsdale D.M., and Twell D., 2001, Functional analysis of *cis*-regulatory elements within the promoter of the tobacco late pollen gene *g10*, Plant Mol. Biol., 45(5): 577-585
https://doi.org/10.1023/A:1010695226241

Tanthanuch W., Chantarangsee M., Maneesan J., Ketudat-Cairns JJJJJ., 2008, Genomic and expression analysis of glycosyl hydrolase family 35 genes from rice (*Oryza sativa* L.), BMC Plant Biology, 8(1): 84-80
https://doi.org/10.1186/1471-2229-8-84

Tian A.M., Liu J.L., and Cao J.S., 2014, Beta galactosidase in plants, Zhongguo Xibao Shengwuxue Xuebao (Chinese Journal of Cell Biology), 36(5): 703-707

Wang P., Li H., Jian Y., Lv Y.Z., and Gai J.T., 2018, Gene mining and evolutionary analysis of xylosyltransferase gene family in solanaceae, Jiyinzuxue yu Yingyong Shengwuxue (Genomics and Applied Biology), 37(1): 332-338

Ahn Y.O., Zheng M.Y., Bevan D.R., Esen A., Shiu S.H., Benson J., Peng H.P., Miller J.T., Cheng C.L., Poulton J.E., Shih M.C., 2007, Functional genomic analysis of *Arabidopsis thaliana* glycoside hydrolase family 35, Science Direct, 68(11): 1510-1520
https://doi.org/10.1016/j.phytochem.2007.03.021