Comparative genomic study of ALDH gene superfamily in Gossypium: A focus on Gossypium hirsutum under salt stress

Yating Dong, Hui Liu, Yi Zhang, Jiahui Hu, Jiyu Feng, Cong Li, Jinhong Chen*, Shuijin Zhu*

Department of Agronomy, Zhejiang Key Laboratory of Crop Germplasm, Zhejiang University, Hangzhou, China

* shjzhu@zju.edu.cn (SZ); jinhongchen@zju.edu.cn (JC)

Abstract

Aldehyde dehydrogenases (ALDHs) are a superfamily of enzymes which play important role in the scavenging of active aldehydes molecules. In present work, a comprehensive whole-genomic study of ALDH gene superfamily was carried out for an allotetraploid cultivated cotton species, G. hirsutum, as well as in parallel relative to their diploid progenitors, G. arbor-eum and G. raimondii. Totally, 30 and 58 ALDH gene sequences belong to 10 families were identified from diploid and allotetraploid cotton species, respectively. The gene structures among the members from same families were highly conserved. Whole-genome duplication and segmental duplication might be the major driver for the expansion of ALDH gene superfamily in G. hirsutum. In addition, the expression patterns of GhALDH genes were diverse across tissues. Most GhALDH genes were induced or repressed by salt stress in upland cotton. Our observation shed lights on the molecular evolutionary properties of ALDH genes in diploid cottons and their alloallotetraploid derivatives. It may be useful to mine key genes for improvement of cotton response to salt stress.

Introduction

Endogenous aldehydes molecules are intermediates in common metabolic pathways [1]. However, excess aldehydes are toxic and deleterious to organism since the reaction of their carbonyl group with cellular nucleophiles. Aldehyde dehydrogenases (ALDHs; enzyme class EC: 1.2.1.3), considered as ‘aldehyde scavengers’, can metabolize a variety of aromatic and aliphatic aldehydes to their corresponding carboxylic acids by irreversible oxidation [2, 3]. ALDHs comprise a gene superfamily which are evolutionarily conserved and have been found in both prokaryotes and eukaryotes [4]. According to the criteria established from ALDH Gene Nomenclature Committee (AGNC), ALDHs can be divided into 24 families throughout all taxa [3]. Plant species contain 14 distinct families including ALDH2, ALDH3, ALDH5, ALDH6, ALDH7, ALDH10, ALDH11, ALDH12, ALDH18, ALDH19, ALDH21, ALDH22, ALDH23, and ALDH24. Among them, families ALDH10, ALDH12, ALDH19, ALDH21, ALDH22, ALDH23, and ALDH24 are unique in plant kingdom. To date, genome-wide analysis of ALDH gene superfamily were
performed in many plant species. There were 16 ALDH genes in Arabidopsis [5], 20 in rice [6], 18 in soybean [7], 26 in populus [8], and 23 in maize [9], etc.

Since the first identified plant ALDH gene rf2 was reported to function in male fertility of maize [10], previous studies have demonstrated that ALDH genes are involved in various metabolic and molecular detoxification pathways. Plant ALDH genes are induced under wide range of abiotic stresses such as drought, cold, high salinity, and heavy metals which indicated their potential role in improvement of plant stress tolerance. It has been proved that ALDH7A1 is a novel enzyme that involved in cellular defense against hyperosmotic stress [11]. Overexpression of ALDH311 in Arabidopsis could enhance the plant’s tolerance to many stresses [12]. Whereas, OsALDH11 and OsALDH22 were highly reduced by drought stress in rice [6]. What’s more, it was reported that transferring the TraeALDH7B1-5A of wheat into Arabidopsis conferred significant drought tolerance in transgenic plants [13]. In addition, ectopically expressing the soybean antiquitin-like ALDH7 gene in Arabidopsis and tobacco resulted in improvement of tolerance towards drought, salinity, and oxidative stress [14].

Though ALDH gene superfamily has been reported in G. raimondii [15], little is known about their detail information in other cotton species, especially the potential role under salt stress in upland cotton.

Cotton is one of the most important economic crop worldwide. There are approximately 50 cotton species in Gossypium genus, among which there are four cultivated species. They include two diploids, Gossypium arboreum (A1) and G. herbaceum (A0), and two natural allotetraploids, G. hirsutum (AD1) and G. barbadense (AD2). Compared with wild species, the cultivated ones are able to produce economically valuable fibers. It has been proved that allotriploid cottons were diversified from the same polyploidization events nearly 1–2 million years ago [16]. In addition, the genomes of G. arboreum and G. raimondii (D1) were considered to be the potential donors of A-subgenome and D-subgenome of the two allotetraploid cotton species, respectively. Recently, the four cotton species have been sequenced completely [17–22]. G. hirsutum accounts for over 90% of commercial cotton production globally and is an ideal model for polyploidy research. As a kind of pioneer crop in saline-alkali, the molecules and mechanisms related with salt stress response are still remain to be uncovered. As mentioned above, ALDHs are proposed to play an important role in plants under abiotic stress. The publications of genome sequences data of these four cotton species give us an access to investigate ALDH gene superfamily systematically in Gossypium, and mine key genes for improvement of plant salt tolerance.

In this study, comparative genomics approaches were applied to analyze ALDH gene superfamily in G. hirsutum and its diploid progenitors G. arboreum and G. raimondii. At the same time, the other cultivated allotetraploid cotton species G. barbadense was also adopted for systematic evolution investigation. The potential roles of ALDH gene superfamily in G. hirsutum response to salt stress were highlighted. The genetic structure and evolutionary relationship analyses were carried out, and the tissue-specific expression profile of ALDH gene superfamily in G. hirsutum was generated. Our results provided insights into the evolutionary processes of polyploidization with ALDH gene superfamily as an example, and associated the genomic sub-structures for the improvement of cotton tolerance to salt stress.

Materials and methods

Database search and sequence retrial for ALDH proteins

The completed genome sequences of four cotton species G. arboreum [21], G. raimondii [17], G. hirsutum acc. TM-1 [20] and G. barbadense cv. Xinhai21 [22] were downloaded from CGP (http://cgp.genomics.org.cn/), Phytozome (http://phytozome.jgi.doe.gov/pz/portal.html#)
The published ALDH proteins of Arabidopsis [5] and rice [6] were obtained from TAIR (http://www.Arabidopsis.org/) and MSU (http://rice.plantbiology.msu.edu/), respectively. Afterwards, the ALDH proteins from Arabidopsis and rice were used as queries to search against those cotton genome databases with BlastP and tBlastN program with a stringent E value cut-off (\(e^{-20}\)). Then, all hits were subjected to Pfam (http://pfam.sanger.ac.uk/) [23] and NCBI Conserved Domain Database (http://www.ncbi.nlm.nih.gov/cdd) [24] to confirm the presence of the conserved domain. Interproscan (http://www.ebi.ac.uk/Tools/pfa/iprscan/) program [25] was subsequently applied to determine each candidate member of ALDH protein superfamily. The retrieved sequences possessing motifs Pfam00171 (ALDH family), PS00687 (ALDH glutamic acid active site), PS00070 (ALDH cysteine active site), KOG2450, KOG2451, KOG2453, and KOG2456 (all aldehyde dehydrogenase) were retained for further analyses. To characterize the members of \(G. \) hirsutum ALDH superfamily, the pI and molecular weight of the full-length proteins were calculated by Compute pI/Mw tool from ExPASy (http://web.expasy.org/cgi-bin/compute_pi/pi_tool) [26]. And the CELLO v2.5 (http://cello.life.nctu.edu.tw/) [27] was applied to predict the subcellular localization.

**Phylogenetic analysis and genomic organization prediction**

For phylogenetic analysis of all the putative ALDH proteins, multiple sequence alignments were created using ClustalX 2.0 [28] with default option, followed by adjustment and refinement with BioEdit V7.2.5 [29]. Then, phylogenetic trees were constructed with MEGA 5.2 software [30] using a neighbor-joining (NJ) method. The parameters were as follows: poisson correction model, pairwise deletion and bootstraps test with 1000 replicates for statistical reliability. Furthermore, maximum likelihood (ML) analysis with PhyML software [31] was applied in the tree construction to test the reliability of NJ method.

The structures of ALDH genes were parsed from respective genome files, and portrayed graphically using the online program Gene Structure Display Server (http://gsds.cbi.pku.edu.cn/) [32].

**Chromosomal location and gene duplication**

To map the location of ALDH genes in \(G. \) hirsutum, the chromosomal distribution of ALDH genes were illustrated by Circos software [33] according to their positional information provided in the genome files. Two types of ALDH gene duplication events were identified within the \(G. \) hirsutum genome. Only the length coverage covered > 80% of the longer one between aligned gene sequences and the similarity of the aligned regions was > 80% can be defined as duplication events [34–36]. Referring to different chromosomal location, they can be designated as tandem duplication or segment duplication. PAL2NAL v14 [37] was then run on these full-length ALDH gene pairs to calculate the nonsynonymous substitutions rate (dN) and synonymous substitution rate (dS) of evolution. The ratio of dN to dS (dN/dS) were then assessed to determine the selective pressure of duplicate genes [38,39].

**Plant materials, growth condition, and salt treatments**

One-week-old seedlings of the upland cotton genetic standard line, \(G. \) hirsutum acc. TM-1, were transplanted into polypots (10 cm in diameter) with full-strength Murashige and Skoog (MS) medium and transferred to growth chamber with temperature of 28°C, relative humidity of 60%, and photoperiod of 16 hours light and 8 hours dark. At the appearance of the true leaf, the seedlings were subjected to salt treatment by transferring them to a MS medium with additional 0, 100, 150 and 200 mM NaCl, which represented the control condition, slight stress,
moderate stress, and severe stress, respectively. Three biological replicates were conducted for each sample. After treatments for two weeks, the root, stem, cotyledon and leaf were harvested from each individual, immediately frozen with liquid nitrogen and then stored at -80˚C for RNA isolation.

RNA isolation and expression analysis of ALDH genes

Fifty-eight pairs of ALDH gene specific primers from *G. hirsutum* were used to study the expression profile of ALDH gene superfamily by qRT-PCR. Total RNAs of all the collected samples were isolated using EASYspin Plus RNAprep Kit (Aidlab, Beijing, China). A NanoDrop 2000 Spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) was used to detect the quantity and quality of total RNAs. Approximately 500 ng RNA was reverse transcribed using the PrimerScript 1st Strand cDNA synthesis kit (TaKaRa, Dalian, China) to synthesis cDNA. All the protocols were followed the manufacturer’s instructions. qPCR was performed with Lightcycler 96 system (Roche, Mannheim, Germany) using SYBR the premix Ex taq (TakaRa, Dalian, China) in 20 μL volume according to the supplier’s protocols. The specific primers used in current research are listed in S1 Table. *G. hirsutum UBQ7* was used as internal control to normalize all data. Each gene was run in triplicate from three biological replicates. $2^{-\Delta\Delta Ct}$ method was carried out to calculate the relative expression levels [40]. And the heatmap for expression profiles were generated with the Mev 4.0 software [41].

Results

Characterization of upland cotton ALDH gene superfamily

The completed genome sequencing of cotton species, *G. arboresum* (A$_2$), *G. raimondii* (D$_3$), *G. hirsutum* (A$_1$D$_1$), and *G. barbadense* (A$_2$D$_2$) resulted in the whole-genome exploration of ALDH gene superfamily in *Gossypium*. In this study, 30, 30, 58, and 58 non-redundant ALDHs encoding members of 10 ALDH gene families (ALDH2, ALDH3, ALDH5, ALDH6, ALDH7, ALDH10, ALDH11, ALDH12, ALDH18, ALDH22) were identified respectively in the aforementioned four *Gossypium* (S2 Table). The nomenclature and description of ALDH genes in the four *Gossypium* species were referred as the criteria established by the ALDH Gene Nomenclature Committee (AGNC). According to the AGNC criteria, deduced cotton ALDH sequences with greater than 40% identical to other previously identified ALDH sequences composed a family, sequences with less than 40% identical would form a new ALDH protein family. For sequences that were more than 60% identical, they were grouped as a protein subfamily. To classify each protein family based on AGNC, all the ALDH proteins from *G. arboresum, G. raimondii, G. hirsutum,* and *G. barbadense* were designated as GaALDH, GrALDH, GhALDH, and GbALDH, respectively. The proteins belonged to different families were followed by the family designation number (1, 2, 3, 4, etc.), and subsequently by a subfamily designation letter (A, B, C, D, etc.). Finally, an individual gene number was added according to chromosomal order within each subfamily. Moreover, A and D were assigned to distinguish genes from A- and D-subgenome of allotetraploid cotton species.

As illustrated in Table 1, family 2 was the largest one with 15 ALDH genes in allotetraploid cottons and eight in diploid cottons, respectively. Families 5, 7, 12 and 22 were the smallest, with only one representative in the diploid progenitors. Compared to other well characterized plant ALDHs, *G. hirsutum* and *G. barbadense* ALDH gene superfamilies were the most expanded ones with 58 members. In *G. hirsutum*, these candidate ALDH genes encoded proteins ranging from 33 kDa (GhALDH2C3D) to 124 kDa (GhALDH6B3D). And the other detailed information of *G. hirsutum* ALDH proteins such as the length, isoelectric points (pI), and the predicted subcellular localization were listed in Table 2.
Phylogenetic and structural analyses of upland cotton ALDH gene superfamily

To assess the functional relevance of members of upland cotton ALDH gene superfamily, phylogenetic relationships among G. hirsutum ALDHs and other plant species was established. An unrooted phylogenetic tree derived from the ALDH amino acid sequences of G. hirsutum, Arabidopsis and rice was illustrated in Fig 1. The phylogenetic tree can be classified into 10 major groups which represented the 10 distinct ALDH protein families of G. hirsutum. In consistent with other previous studies, families 2, 5 and 10 grouped together, and families 3 and 22 were connected by a node, which belongs to the plant ALDH core families. Family 18 was the most phylogenetically distantly related ALDH family from the view of this topology. Meanwhile, an ML tree reconstructed with PhyML was almost consistent with the NJ tree except for minor differences at some branches (S1 Fig).

To obtain further insight into the evolutionary relationship among G. hirsutum ALDH gene superfamily and other three surveyed cotton species, all the putative ALDH proteins from G. arboreum, G. raimondii, G. hirsutum and G. barbadense were also aligned to construct an unrooted phylogenetic tree. As expected, the topology was similar to that generated with ALDH proteins from G. hirsutum, Arabidopsis and rice. As Fig 2A displayed, the core ALDH families 2, 5 and 10 clustered tightly, while family 18 was still the most phylogenetically distant. Form the view of evolution, one member of ALDH genes in diploid cottons would be correspondent with two homeologs from the A and D subgenomes of allotetraploid cotton species. In this study, most ALDH genes from G. hirsutum shown a one-to-one correspondence with those from its diploid progenitors, and the same phenomena was found in the other one cultivated allotetraploid cotton species G. barbadense. The inconsistencies including each a member loss of subfamilies 2B, 6B, and 11A in G. barbadense, and subfamilies 2C, 6B and 7B in G. hirsutum. Furthermore, one more putative ALDH gene was present in the ALDH10A subfamily of G. barbadense and ALDH3H subfamily of G. hirsutum respectively. In particular, the homologous genes were almost in the terminal branches of the phylogenetic tree with high bootstrap values. And those genes within the same subfamilies from the same subgenome of allotetraploids tended to cluster together, suggesting close relationship between them. Surprisingly, ALDH genes from the A subgenome and D subgenome of G. hirsutum shown a bias to those from G. arboreum and G. raimondii. Meanwhile, the phylogenetic tree reconstructed with ML method was almost consistent with the one of NJ method except for minor differences at some branches (S2 Fig), which validated the reliability of our results.

Table 1. The number of ALDH gene superfamily members identified in Gossypium.

| Family | G. arboreum | G. raimondii | G. hirsutum | G. Barbadense |
|--------|-------------|--------------|-------------|--------------|
| Family 2 | 8           | 8            | 15          | 15           |
| Family 3 | 5           | 6            | 12          | 11           |
| Family 5 | 1           | 1            | 2           | 2            |
| Family 6 | 4           | 3            | 6           | 6            |
| Family 7 | 1           | 1            | 1           | 2            |
| Family 10 | 2           | 2            | 4           | 5            |
| Family 11 | 3           | 3            | 6           | 5            |
| Family 12 | 1           | 1            | 2           | 2            |
| Family 18 | 4           | 4            | 8           | 8            |
| Family 22 | 1           | 1            | 2           | 2            |
| Total   | 30          | 30           | 58          | 58           |

https://doi.org/10.1371/journal.pone.0176733.t001
| Family | Gene Name | Gene identifier | Chromosome | Genomics position | CDS Size (AA) | Mw(kDa) | pI | Subcellular Localization | Strand |
|--------|------------|-----------------|-------------|-------------------|---------------|---------|----|-------------------------|--------|
| Family2 | GhALDH2B1A | Gh_A03G1229     | A03         | 86766784–86770208 | 1179 392      | 42.64381 | 5.70 | Chloroplast minus       |        |
|        | GhALDH2B2A | Gh_A07G2011     | A07         | 76376925–76380509 | 1596 531      | 57.63800 | 7.59 | Mitochondrial plus      |        |
|        | GhALDH2B3A | Gh_A12G1314     | A12         | 69558172–69561824 | 1596 531      | 57.69817 | 6.81 | Mitochondrial plus      |        |
|        | GhALDH2B4A | Gh_A12G2471     | A12         | 87183896–87187146 | 1623 540      | 58.81244 | 7.18 | Mitochondrial plus      |        |
|        | GhALDH2B1D | Gh_D02G1665     | D02         | 57172439–57176663 | 1530 509      | 55.34062 | 6.80 | Chloroplast minus       |        |
|        | GhALDH2B2D | Gh_D07G2232     | D07         | 53375582–53379207 | 1593 530      | 57.64695 | 8.03 | Mitochondrial plus      |        |
|        | GhALDH2B3D | Gh_D12G1438     | D12         | 44221515–44225115 | 1596 531      | 57.74128 | 6.81 | Mitochondrial plus      |        |
|        | GhALDH2B4D | Gh_D12G2599     | D12         | 58840207–58843432 | 1626 541      | 58.92050 | 7.18 | Chloroplast plus        |        |
|        | GhALDH2C1A | Gh_A05G0157     | A05         | 1670874–1678965   | 1515 504      | 54.79314 | 8.33 | Cytoplasmic minus       |        |
|        | GhALDH2C2A | Gh_A06G0526     | A06         | 10799343–10802205 | 1494 497      | 53.99529 | 7.11 | Cytoplasmic minus       |        |
|        | GhALDH2C3A | Gh_A07G0063     | A07         | 761623–777944     | 2739 912      | 98.77655 | 5.97 | Cytoplasmic plus        |        |
|        | GhALDH2C1D | Gh_D05G0221     | D05         | 2045263–2049124   | 1215 404      | 43.71334 | 7.05 | Cytoplasmic minus       |        |
|        | GhALDH2C2D | Gh_D06G0580     | D06         | 9239274–9242178   | 1494 497      | 53.95119 | 7.11 | Cytoplasmic minus       |        |
|        | GhALDH2C3D | Gh_D07G0046     | D07         | 498249–502320     | 933 310       | 33.44073 | 7.05 | Cytoplasmic minus       |        |
|        | GhALDH2C4D | Gh_D07G0047     | D07         | 506235–511937     | 1548 515      | 55.99429 | 5.85 | Cytoplasmic minus       |        |
| Family3 | GhALDH3F1A | Gh_A02G1616     | A02         | 82646608–82651284 | 1440 479      | 53.27214 | 8.62 | PlasmaMembrane minus    |        |
|        | GhALDH3F2A | Gh_A05G0568     | A05         | 6073916–6077932   | 1470 489      | 54.74277 | 7.62 | PlasmaMembrane minus    |        |
|        | GhALDH3F1D | Gh_D03G1006     | D03         | 792938–797752     | 1440 479      | 53.29011 | 8.66 | PlasmaMembrane plus     |        |
|        | GhALDH3F2D | Gh_D05G0679     | D05         | 5668003–5672363   | 1473 490      | 54.81081 | 8.06 | PlasmaMembrane minus    |        |
|        | GhALDH3H1A | Gh_A02G0751     | A02         | 14030633–14036873 | 1473 490      | 53.89183 | 8.27 | Mitochondrial minus     |        |
|        | GhALDH3H2A | Gh_A05G3716     | A06         | 27711–31024      | 1182 393      | 42.36826 | 8.81 | PlasmaMembrane minus    |        |
|        | GhALDH3H3A | Gh_A06G0384     | A06         | 6429365–6435437   | 1473 490      | 53.62758 | 8.61 | Mitochondrial plus      |        |
|        | GhALDH3H1D | Gh_D02G0793     | D02         | 12945983–12952298 | 1473 490      | 53.92185 | 8.27 | Mitochondrial minus     |        |
|        | GhALDH3H2D | Gh_D05G2245     | D05         | 21555587–21560927 | 1482 493      | 54.20499 | 8.79 | Chloroplast plus        |        |
|        | GhALDH3H3D | Gh_D06G0414     | D06         | 5912130–5918225   | 1473 490      | 53.69162 | 8.61 | Mitochondrial plus      |        |
|        | GhALDH3I1A | Gh_A05G3311     | A05         | 86650719–86655533 | 1473 490      | 54.23894 | 6.15 | Chloroplast minus       |        |
|        | GhALDH3I1D | Gh_D04G0292     | D04         | 4378806–4385236   | 1650 549      | 61.09512 | 8.65 | Chloroplast plus        |        |
| Family5 | GhALDH5F1A | Gh_A01G0799     | A01         | 17510464–17544827 | 1530 509      | 54.67401 | 7.58 | Chloroplast minus       |        |

(Continued)
| Family | Gene Name     | Gene identifier | Chromosome | Genomics position | CDS Size (AA) | Mw(kDa)  | pI       | Subcellular Localization | Strand |
|--------|---------------|----------------|------------|-------------------|--------------|---------|---------|--------------------------|---------|
| 6      | GhALDH6B1A    | Gh_A0 3G1560   | A03        | 96700062–96712653 | 2478 825     | 90.58171 | 8.17    | Mitochondrial minus      |         |
| 6      | GhALDH6B2A    | Gh_A0 3G1561   | A03        | 96700062–96712653 | 3219 1072    | 117.91685| 8.46    | Mitochondrial minus      |         |
| 6      | GhALDH6B3A    | Gh_A10G0793    | A10        | 16217188–16235086 | 1620 539     | 57.91762 | 8.61    | Mitochondrial minus      |         |
| 6      | GhALDH6B1D    | Gh_D02G2009    | D02        | 64300173–64305816 | 2118 705     | 77.40568 | 8.47    | Cytoplasmic minus        |         |
| 6      | GhALDH6B2D    | Gh_D02G2010    | D02        | 64307155–64315478 | 3405 1134    | 124.99664| 8.94    | Nuclear plus              |         |
| 7      | GhALDH7B1D    | Gh_D06G1578    | D06        | 52817685–52822968 | 1524 507     | 54.33663 | 6.42    | Chloroplast minus        |         |
| 10     | GhALDH10A1A   | Gh_A07G0563    | A07        | 7809148–7814113   | 1512 503     | 54.69899 | 5.34    | Chloroplast minus        |         |
| 10     | GhALDH10A2A   | Gh_A11G0380    | A11        | 3483181–3488821   | 1509 502     | 54.75215 | 5.47    | Cytoplasmic minus        |         |
| 10     | GhALDH10A1D   | Gh_D07G0629    | D07        | 7306529–7311487   | 1512 503     | 54.72795 | 5.34    | Chloroplast minus        |         |
| 10     | GhALDH10A2D   | Gh_D11G0441    | D11        | 3702955–3708589   | 1509 502     | 54.71415 | 5.79    | Cytoplasmic minus        |         |
| 11     | GhALDH11A1A   | Gh_A05G0415    | A05        | 4681423–4683777   | 1491 496     | 53.16555 | 7.50    | Cytoplasmic plus         |         |
| 11     | GhALDH11A2A   | Gh_A05G0479    | A05        | 5198467–5201069   | 1494 497     | 53.28571 | 6.80    | Cytoplasmic plus         |         |
| 11     | GhALDH11A3A   | Gh_A07G0209    | A07        | 2502115–2504488   | 1500 499     | 53.48300 | 6.80    | Cytoplasmic plus         |         |
| 11     | GhALDH11A1D   | Gh_D05G0533    | D05        | 4326082–4328437   | 1491 496     | 53.17862 | 7.88    | Cytoplasmic plus         |         |
| 11     | GhALDH11A2D   | Gh_D05G0594    | D05        | 4784042–4786694   | 1494 497     | 53.17963 | 7.11    | Cytoplasmic plus         |         |
| 11     | GhALDH11A3D   | Gh_D07G0263    | D07        | 2751699–2754071   | 1500 499     | 53.44692 | 7.13    | Cytoplasmic plus         |         |
| 12     | GhALDH12A1A   | Gh_A03G0575    | A03        | 15033080–15037682 | 1665 554     | 61.51080 | 6.98    | Mitochondrial minus      |         |
| 12     | GhALDH12A1D   | Gh_D03G0856    | D03        | 29709339–29713965 | 1665 554     | 61.49283 | 7.27    | Mitochondrial minus      |         |
| 18     | GhALDH18B1A   | Gh_A01G1899    | A01        | 98932667–98937902 | 2196 731     | 79.25592 | 6.23    | Mitochondrial plus       |         |
| 18     | GhALDH18B2A   | Gh_A04G1396    | scaffold1000_A04 | 9715–18472   | 2193 730     | 79.16256 | 6.24    | Mitochondrial plus       |         |
| 18     | GhALDH18B3A   | Gh_A10G2010    | A10        | 97863671–97869528 | 2133 710     | 76.84524 | 6.40    | Mitochondrial minus      |         |
| 18     | GhALDH18B4A   | Gh_A11G2925    | A11        | 93106493–93111298 | 2211 736     | 78.73630 | 6.55    | Mitochondrial minus      |         |
| 18     | GhALDH18B1D   | Gh_D01G2158    | D01        | 60548378–60553524 | 2208 735     | 79.78740 | 6.53    | Mitochondrial plus       |         |
| 18     | GhALDH18B2D   | Gh_D04G1204    | D04        | 39407171–39416330 | 2193 730     | 79.17467 | 6.31    | Mitochondrial minus      |         |
The genomic structures were vital to reveal the evolutionary history within ALDH gene families. We compared the ALDH gene structures and found that genes from the same subfamily usually possessed a highly conserved exon-intron organization within and across the four surveyed cotton species (Fig 2B). In G. hirsutum, the numbers of exons of ALDH genes varied from six to 22. Some ALDH genes even shared identical number and length of exon such as genes from subfamilies 2B, 2C, 10A, 11A, and 12A. Such conserved gene structures within each subfamily indicated that cotton ALDH genes have underwent duplication events during evolution. Compared with the ALDH genes of intra-species, the ALDH genes from the A and D subgenomes of the two allotetraploid cottons were more similar to those from their ancestor species respectively. Even so, exon gains or losses still occurred during evolution with subfamily 22A as an example. GaALDH22A1, GrALDH22A1, GhALDH22D1, and GbALDH22D1 each contained 14 exons, while GhALDH22A1A possessed 13 exons and GbALDH22A1A had 15 exons.

Chromosomal distribution and expansion patterns of upland cotton ALDH gene superfamily

The mapping of the gene loci shown that ALDH genes were distributed unevenly on 19 of 26 G. hirsutum chromosomes. As illustrated in Fig 3, Chr A05, D02, D05 and D07 contained five ALDH genes each, followed by Chr A03 and A07 on which four ALDH genes were located. Additionally, GhALDH3H2A and GhALDH18B2A were distributed on the scaffolds related to Chr A05 and A04, respectively. The remaining genes were dispersed on other chromosomes.

To examine the driving force for gene evolution, the nonsynonymous and synonymous substitution (dN and dS) of duplicated genes were calculated using the full-length sequences. A dN/dS ratio of 1 was set as a cut-off value for identify genes under negative selection. As demonstrated in Table 3, almost all the duplicated gene pairs were likely under purifying selection pressure with the dN/dS ratio < 1, except for GhALDH3F2A/GhALDH3F2D, suggesting that the two genes had experienced positive selection.

Expression profiles of upland cotton ALDH gene superfamily under salt stress

A comprehensive qRT-PCR analysis was performed to obtain the expression patterns of ALDH gene superfamily in G. hirsutum. As displayed in Fig 4, most ALDH-encoding genes showed predominant expression in roots and stems compared with cotyledons and leaves. In most cases, the genes from the same family with conserved structure didn’t cluster together, suggesting a function divergence during evolution. Most GhALDH genes shown a tissue-specific expression pattern with the exception of GhALDH3H2A/GhALDH3H2D, GhALDH3H3A/
GhALDH3H3D, and GhALDH10A1A which exhibited abundant in all the tissues detected. Notably, for GhALDH10A1D, GhALDH11A1A/GhALDH11A1D, and GhALDH11A3D genes, high level accumulation existed in stem, cotyledon, and leaf, but not in root. On the contrary, the expression level of GhALDH2C1A/GhALDH2C1D, GhALDH18B4A/GhALDH18B4D, and GhALDH2C3A couldn’t be detected almost in all the four tissues surveyed.

Fig 1. Phylogenetic analysis of the ALDH proteins from G. hirsutum, Arabidopsis and rice. The unrooted phylogenetic tree was constructed using MEGA 5.2 by Neighbor-Joining method. Numbers on branches were bootstrap portions from 1000 replicates. Percentage bootstrap scores of <50% were hidden. The specific color indicated different families.

https://doi.org/10.1371/journal.pone.0176733.g001
Fig 2. Phylogenetic relationships and gene structures of ALDHs from *G. arboreum*, *G. raimondii*, *G. hirsutum*, and *G. barbadense*. (A) The unrooted phylogenetic tree was constructed using MEGA 5.2 by
Researches have shown that the plant ALDH genes were involved in a wide range of stress response pathways. Therefore, we particularly aim at the expression pattern changes of ALDH gene superfamily under salt stress in upland cotton. The heatmap of *G. hirsutum* ALDH gene superfamily expression profile under salt stress was presented in Fig 5. The relative expression levels of *GhALDHs* under salt treatment differed among each subfamilies. A majority of ALDH genes shown altered expression patterns of either induction or suppression associated with at least one salt treatment. In roots and stems, nearly no ALDH genes were induced under salt stress except for *GhALDH2C3A/GhALDH2C3D* and *GhALDH18B1A/GhALDH18B1D*. Transcripts of *GhALDH18B4A* was initially increased under a slight salinity conditions and then dropped under severe salinity conditions in roots. Fifty-four *GhALDH* genes shown an up-regulated expression trend in leaves in response to seriously salt treatments. By contrast, the number of up-regulated ALDH genes in cotyledons were less. In leaves, *GhALDH6B2A*, *GhaLDH12A1A*, *GhALDH2B2A*, and *GhALDH7B1D* presented a continuous increase of

![Fig 3. Chromosome distribution and gene duplication of GhALDH gene superfamily.](https://doi.org/10.1371/journal.pone.0176733.g003)
transcript accumulation under the salt treatments. And GhALDH2B3A, GhALDH2C1D, and GhALDH3H3A were down-regulated under stress condition.

### Discussion

The phylogenetic relationship of ALDH gene superfamily were highly related among two allotetraploidal cotton sepcies and their diploid progenitors

The releases of genome assembly for four *Gossypium* species makes it easy to analyze the stress response related gene families through comparative genomics method [17–22]. In this study, putative ALDH sequences belonging to 10 ALDH families were individually identified in the genomes of *G. arboreum*, *G. raimondii*, *G. hirsutum*, and *G. barbadense*. Compared with other known plant ALDH superfamilies such as *Arabidopsis* [5], rice [6], and *populus* [8], *Gossypium* possessed the most expanded ones, with 30 members in both diploid species and 58 in both allotetraploid species, respectively. A previous studies identified 30 *G. raimondii* ALDH genes based on the same genome data we used. We checked the result by BLASTP and tBlastN and found to be consistent. From the theory of evolution, one single ALDH gene in diploid cottons should be corresponding to two homeologs in their allotetraploid derivatives. However, the

| Duplicated gene 1 | Duplicated gene 2 | dN  | dS  | dN/dS |
|-------------------|-------------------|-----|-----|-------|
| GhALDH2B1A        | GhALDH2B1D        | 0.0298 | 0.0551 | 0.5407 |
| GhALDH2B2A        | GhALDH2B2D        | 0.0063 | 0.0404 | 0.1554 |
| GhALDH2B3A        | GhALDH2B3D        | 0.0088 | 0.0319 | 0.2749 |
| GhALDH2B4A        | GhALDH2B4D        | 0.0051 | 0.0465 | 0.1107 |
| GhALDH2C1A        | GhALDH2C1D        | 0.0137 | 0.0673 | 0.2040 |
| GhALDH2C2A        | GhALDH2C2D        | 0.0056 | 0.0153 | 0.3692 |
| GhALDH2C3A        | GhALDH2C3D        | 0.0181 | 0.0508 | 0.3560 |
| GhALDH2C3D        | GhALDH2C4D        | 0.0564 | 0.226  | 0.2494 |
| GhALDH3F1A        | GhALDH3F1D        | 0.0139 | 0.0496 | 0.2803 |
| GhALDH3F2A        | GhALDH3F2D        | 0.0145 | 0.0061 | 2.3895 |
| GhALDH3H1A        | GhALDH3H1D        | 0.0010 | 0.0431 | 0.0224 |
| GhALDH3H2A        | GhALDH3H2D        | 0.0091 | 0.0182 | 0.5000 |
| GhALDH3H3A        | GhALDH3H3D        | 0.0056 | 0.0393 | 0.1416 |
| GhALDH3H31A       | GhALDH3H31D       | 0.0106 | 0.0263 | 0.4025 |
| GhALDH5F1A        | GhALDH5F1D        | 0.0341 | 0.0649 | 0.5254 |
| GhALDH6B1A        | GhALDH6B1D        | 0.0085 | 0.0369 | 0.2314 |
| GhALDH6B2A        | GhALDH6B2D        | 0.0100 | 0.0388 | 0.258 |
| GhALDH6B3A        | GhALDH6B3D        | 0.0196 | 0.0443 | 0.4422 |
| GhALDH10A1A       | GhALDH10A1D       | 0.0039 | 0.0187 | 0.2062 |
| GhALDH10A2A       | GhALDH10A2D       | 0.0068 | 0.0209 | 0.3258 |
| GhALDH11A1A       | GhALDH11A1D       | 0.0047 | 0.0155 | 0.3015 |
| GhALDH11A3A       | GhALDH11A3D       | 0.0055 | 0.0357 | 0.1526 |
| GhALDH12A1A       | GhALDH12A1D       | 0.0049 | 0.0403 | 0.1222 |
| GhALDH18B1A       | GhALDH18B1D       | 0.0195 | 0.0621 | 0.3149 |
| GhALDH18B2A       | GhALDH18B2D       | 0.0036 | 0.0406 | 0.0883 |
| GhALDH18B3A       | GhALDH18B4A       | 0.1149 | 1.8984 | 0.0605 |
| GhALDH18B3D       | GhALDH18B4D       | 0.1159 | 2.0883 | 0.0555 |
| GhALDH22A1A       | GhALDH22A1D       | 0.0054 | 0.0407 | 0.1333 |
numbers of ALDH genes in *G. hirsutum* and *G. barbadense* were less than the total sum of those from *G. arboreum* and *G. raimondii*, not twofold theoretically. This could be explained by that gene loss during the evolution of allotetraploid cottons after speciation. In addition, the size of ALDH gene superfamily is the same in two diploid cottons, albeit the genome of *G.*
 arboreum is approximately twice larger than that of G. raimondii. This may be associated with the long terminal repeat (LTR) retrotransposons insertion along each chromosome in G. arboreum [17–19]. However, the relatively same size of ALDH gene superfamily among the four surveyed cotton species reflects the high conservation of ALDH genes during evolution.

In order to reveal the homologous relationships of ALDH gene superfamily among different taxa, a phylogenetic tree was generated with full-length ALDH proteins from G. hirsutum, Arabidopsis, and rice. As Fig 1 illustrated, GhALDHs were more closely related with AtALDHs than OsALDHs, which was consistent with the evolutionary relationships among the three species. The topology of the other phylogenetic tree constructed with full-length ALDH proteins from the four surveyed cottons was similar to that mentioned above. The two phylogenetic trees indicated that the plant core ALDH families 2, 5, and 10 were grouped together. And family 18 was the most distantly related one, which was similar to that from other plant species such as populus [8], grape [42], and P. trichocarpa [43]. Meanwhile, our results have complemented the earlier study of ALDH superfamily in G. raimondii [15] by the comparative genomics approach. Furthermore, it’s worth noting that families 5, 7, 12, 22 were represented by only one gene number in all the surveyed diploid species, and one or two counterparts in allotetraploids cottons. It is speculated that these families may act as ‘house-keeping genes’ to participate in the fundamental metabolism and physiological pathways of plants to keep balance of aldehyde concentration. In contrast, family 2 and family 3 are the two most expanded groups in the six plant species we investigated. Studies shown that the ALDH2 gene family can degrade the acetaldehyde generated through ethanolic fermentation [44,45]. In Arabidopsis,
ALDH3I1 expression only can be detected in leaves and induced by stress treatments such as ABA exposure, salinity, dehydration, heavy metals, oxidants and pesticides [46,47]. The expansion of ALDH2 and ALDH3 gene families compared with other families suggested that these ALDH genes may be essential for plants to cope with environmental stresses. Additionally, the ALDH gene members from the subgenomes of two allotetraploid cottons were more phylogenetic closely to their diploid genome ancestors. It reflected that ALDH superfamily evolved before the formation of allotetraploid cotton species by a polyploidization event.

The ALDH gene superfamily were greatly conserved in four Gossypium species

To explore evolutionary characters of ALDH genes among diploids and allotetraploids, we directly compared the gene structures of different species. A high level of structural identity was observed among the ALDH genes from the same subfamily. The conservation of gene structures correlated well with the phylogetic clades. Such phenomena indicated that cotton ALDH gene superfamily have underwent duplication events during evolution. Also, the ALDH gene members from the A-subgenome and D-subgenome of allotetraploid cottons were structurally more similar to those of its A- and D-genome progenitor, separately. It further supported the topology of our phylogenetic tree. Meanwhile, this might be a result from genome duplication occurrence earlier than segmental duplication. Gene duplication events, including tandem duplication, segmental duplication, transposition events, and whole-genome duplication, are the major reason for the amplification of gene family [48,49]. In G. hirsutum, whole-genome duplicaton mainly contributed to the expansion of ALDH gene superfamily. And an intriguing finding was that purifying selection predominated across the duplicated genes. A likely reason for this observation is that, for a new duplicate gene, deleterious loss-of-function mutations were tended to happen. However, purifying selection could eliminate it, thus fixed the retention in a genome and function of both duplicate loci [50–52].

ALDH gene superfamily shown a functional diversity in response to salt stress in G. hirsutum

The gene expression patterns can provide important clues for gene function. Our qRT-PCR results demonstrated the different expression patterns of ALDH gene superfamily across tissues of upland cotton. The conservation of ALDH gene superfamily in plants implied their significance in fundamental processes. There must exist strong selective pressure to maintain the gene function. Functional analyses of most ALDH genes shown that they shared the same stress response pattern among various plants [5]. In the study, a majority of GhALDH genes were up-regulated in leaves under severe salt stress, although roots are the tissues directly exposing to environmental stresses. This may be associated with the facts that these two tissues by themselves were distinct in structure and functions [53].

Arabidopsis ALDH gene AtALDH10A9 was reported to be weakly induced by different abiotic stressors [54]. Transgenic Arabidopsis plants overexpressing AtALDH7B4 were more tolerant to salt stress and show reduced accumulation of malondialdehyde (MDA) in comparison to the wild-type ones [55]. Analogously, the expression of orthologous gene GhALDH10A2D and GhALDH7B1D were induced significantly in leaves under salt stress in our study. In addition, most of the duplicated gene pairs demonstrated a high degree of functional divergence in response to salt treatment. In leaves, GhALDH2B3A shown a high level of accumulation in response to salt stress, while the closely-related gene GhALDH2B3D shown down-regulated. It could be explained by the assertion that the duplicated genes always underwent massive silencing and elimination after whole-genome duplication. This has long been recognized as a
pervasive force in plant evolution [56]. Nonfunctionalization, subfunctionalization, and neo-
functionalization are the three evolutionary fates of duplicate genes. And the functional diver-
gence among duplicate genes can increase their chance to be retained in a genome [57].
What’s more, the expression level dominance exhibited by \( G. \) hirsutum under salt stress was unbalanced toward the D-subgenome, more ALDH genes (27 of 30 ALDH genes) were induced than that of A-subgenome (23 of 28 ALDH genes) in leaves. This is consistent with the nature of its diploid ancestors, i.e., in the genomic group of A-genome progenitor, long fiber first involved; while in the D-genome parent, the feature of adaptation to adverse envi-
ronmental stresses was dominant.

Conclusions

A comparative genomics approach was carried out to investigate ALDH superfamily in upland cotton. The phylogenetic relationships and gene structure were evaluated in the four cotton species, \( G. \) arboreum, \( G. \) raimondii, \( G. \) hirsutum, and \( G. \) barbadense. The tissue-specific expression profiles of \( GhALDH \) gene superfamily were detected. Future work will reveal the physio-
logical role of different ALDH genes in dealing with abiotic stress in \( Gossypium \) species. Our
findings may provide a framework to understand the evolution of ALDH gene superfamily in plants and help in the identification of key genes which can be used in the improvement of salt tolerance for cotton.

Supporting information

S1 Fig. Phylogenetic relationships of ALDH proteins from \( G. \) hirsutum, \( Arabidopsis \) and \( rice \). The unrooted phylogenetic tree was constructed using PhyML software by Maximum Likelihood method with LG model. The bootstrap test was performed with 1,000 replicates. Different ALDH families were represented by specific colors.
(TIF)

S2 Fig. Phylogenetic relationships of ALDH proteins from \( G. \) arboreum, \( G. \) raimondii, \( G. \) hirsutum, and \( G. \) barbadense. The unrooted phylogenetic tree was constructed using PhyML software by Maximum Likelihood method with LG model. The bootstrap test was performed with 1,000 replicates. Different ALDH families were represented by specific colors.
(TIF)

S1 Table. Gene-specific primers for q-RT-PCR used in this study.
(XLSX)

S2 Table. The information of ALDH genes in \( Gossypium \).
(XLSX)

Acknowledgments

We are grateful to Li Chen, Rubing Zhao, Tianlun Zhao, Fan Zhang, and Jieqiong Huang (Zhejiang University, China) for their support in this study. We also appreciate valuable com-
ments for manuscript improvement from the editor and reviewers.

Author Contributions

Conceptualization: YD JC SZ.
Data curation: YD JC SZ.
Formal analysis: YD HL YZ.

Funding acquisition: SZ.

Investigation: YD JC.

Methodology: YD YZ JH Cong Li Cheng Li.

Project administration: YD SZ.

Resources: YD JF.

Software: HL JF.

Supervision: JC SZ.

Validation: HL YZ JH.

Visualization: YD JC SZ.

Writing – original draft: YD SZ.

Writing – review & editing: YD JC SZ.

References
1. Vasiliou V, Pappa A, Petersen DR. Role of aldehyde dehydrogenases in endogenous and xenobiotic metabolism. Chem-Biol Interact. 2000; 129: 1–19. PMID: 11154732
2. Lindahl R. Aldehyde dehydrogenases and their role in carcinogenesis. Crit Rev Biochem Mol. 1992; 27: 283–335.
3. Vasiliou V, Bairroch A, Tipton KF, Nebert DW. Eukaryotic aldehyde dehydrogenase (ALDH) genes: human polymorphisms, and recommended nomenclature based on divergent evolution and chromosomal mapping. Pharmacogenet Genom. 1999; 9: 421–434.
4. Sophos NA, Pappa A, Ziegler TL, Vasiliou V. Aldehyde dehydrogenase gene superfamily: the 2000 update. Chem-Biol Interact. 2003; 143: 5–22. PMID: 12904184
5. Kirch HH, Bartels D, Wei YL, Schnable PS, Wood AJ. The ALDH gene superfamily of Arabidopsis. Trends Plant Sci. 2004; 9: 371–377. https://doi.org/10.1016/j.tplants.2004.06.004 PMID: 15358267
6. Gao C, Han B. Evolutionary and expression study of the aldehyde dehydrogenase (ALDH) gene superfamily in rice (Oryza sativa). Gene. 2009; 431: 86–94. https://doi.org/10.1016/j.gene.2008.11.010 PMID: 19071198
7. Kotchoni SO, Jimenez-Lopez JC, Kayode APP, Gachorno EW, Baba-Moussa L. The soybean aldehyde dehydrogenase (ALDH) protein superfamily. Gene. 2012; 495: 128–133. https://doi.org/10.1016/j.gene.2011.12.035 PMID: 22226812
8. Tian F, Zhang J, Wang T, Xie Y, Zhang J, Hu J. Aldehyde dehydrogenase gene superfamily in Populus: organization and expression divergence between paralogous gene pairs. Plos One. 2015; 10: e0124669. https://doi.org/10.1371/journal.pone.0124669 PMID: 25909656
9. Jimenez-Lopez JC, Gachorno EW, Seufferheld MJ, Kotchoni SO. The maize ALDH protein superfamily: linking structural features to functional specificities. BMC Struct Biol. 2010; 10: 43. https://doi.org/10.1186/1472-6807-10-43 PMID: 21190582
10. Skribbe DS, Liu F, Wen T, Yandieu MD, Cui X, Cao J, et al. Characterization of the aldehyde dehydrogenase gene families of Zea mays and Arabidopsis. Plant Mol Biol. 2002; 48: 751–764. PMID: 11999848
11. Brocker C, Lassen N, Estey T, Pappa A, Cantore M, Orilova VV, et al. Aldehyde dehydrogenase 7A1 (ALDH7A1) is a novel enzyme involved in cellular defense against hyperosmotic stress. J Biol Chem. 2010; 285: 18452–18463. https://doi.org/10.1074/jbc.M109.077925 PMID: 20207735
12. Sunkar R, Bartels D, Kirch HH. Overexpression of a stress-inducible aldehyde dehydrogenase gene from Arabidopsis thaliana in transgenic plants improves stress tolerance. Plant J. 2003; 35: 452–464. PMID: 12904208
13. Chen J, Wei B, Li G, Fan R, Zhong Y, Wang X, et al. TraeALDH7B1-5A, encoding aldehyde dehydrogenase 7 in wheat, confers improved drought tolerance in Arabidopsis. Planta. 2015; 242: 137–151. https://doi.org/10.1007/s00425-015-2290-8 PMID: 25893867
14. Rodrigues SM, Andrade MO, Gomes AP, Damatta MC, Baracat-Pereira MC, Fontes EPB. Arabidopsis and tobacco plants ectopically expressing the soybean antiquitin-like ALDH7 gene display enhanced tolerance to drought, salinity, and oxidative stress. J Exp Bot. 2006; 57: 1909–1918. https://doi.org/10.1093/jxb/erj132 PMID: 16595581

15. He D, Lei Z, Xing H, Tang B. Genome-wide identification and analysis of the aldehyde dehydrogenase (ALDH) gene superfamily of Gossypium raimondii. Gene. 2014; 549: 123–133. https://doi.org/10.1016/j.gene.2014.07.054 PMID: 25058695

16. Wendel JF, Cronn RC. Polyploidy and the evolutionary history of cotton. Adv Agron. 2003; 78: 139–186.

17. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin DC, et al. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature. 2012; 492: 423–427. https://doi.org/10.1038/nature11798 PMID: 23257886

18. Wang K, Wang Z, Li F, Ye W, Wang J, Song G, et al. The draft genome of a diploid cotton Gossypium raimondii. Nat Genet. 2012; 44: 1098–1103. https://doi.org/10.1038/ng.2271 PMID: 22922876

19. Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, et al. Genome sequence of the cultivated cotton Gossypium arboreum. Nat Genet. 2014; 46: 567–572. https://doi.org/10.1038/ng.2987 PMID: 24836287

20. Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, et al. Sequencing of allotetraploid cotton (Gossypium hirsutum L. acc. TM-1) provides a resource for fiber improvement. Nat Biotechnol. 2015; 33: 531–537. https://doi.org/10.1038/nbt.3207 PMID: 25893781

21. Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, et al. Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. Nat Biotechnol. 2015; 33: 524–530. https://doi.org/10.1038/nbt.3208 PMID: 25893780

22. Liu X, Zhao B, Zheng H, Hu Y, Lu G, Yang C, et al. Gossypium barbadense genome sequence provides insight into the evolution of extra-long staple fiber and specialized metabolites. Sci Rep. 2015; 5: 14139. https://doi.org/10.1038/srep14139 PMID: 26420475

23. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014; 42: D222–D230. https://doi.org/10.1093/nar/gkt1223 PMID: 24288371

24. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu SN, Chitsaz F, Geer LY, et al. CDD: NCBI's conserved domain database. Nucleic Acids Res. 2015; 43: D222–D226. https://doi.org/10.1093/nar/gku1221 PMID: 25414356

25. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: protein domains identifier. Nucleic Acids Res. 2015; 43: D222–D226. https://doi.org/10.1093/nar/gku1221 PMID: 25414356

26. Bjellqvist B, Basse B, Olsen E, Celis JE. Reference points for comparisons of 2-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. Electrophoresis. 1994; 15: 529–539. PMID: 8055880

27. Yu CS, Lin CJ, Hwang JK. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci. 2004; 13: 1402–1406. https://doi.org/10.1110/ps.03479604 PMID: 15096640

28. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007; 23: 2947–2948. https://doi.org/10.1093/bioinformatics/btm404 PMID: 17846036

29. Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. Nucl Acids Symp Ser. 1999; 41: 95–98.

30. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGAS: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011; 28: 2731–2739. https://doi.org/10.1093/molbev/msr121 PMID: 21546353

31. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003; 52: 696–704. PMID: 14530136

32. Hu B, Jin J, Guo AY, Zhang H, Luo J, Gao G. GSDDS 2.0: an upgraded gene feature visualization server. Bioinformatics. 2015; 31: 1296–1297. https://doi.org/10.1093/bioinformatics/btu717 PMID: 25504850

33. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009; 19: 1639–1645. https://doi.org/10.1101/gr.092759.109 PMID: 19541911

34. Maher C, Stein L, Ware D. Evolution of Arabidopsis microRNA families through duplication events. Genome Res. 2006; 16: 510–519. https://doi.org/10.1101/gr.4680506 PMID: 16520461
35. Dong Y, Li C, Zhang Y, He Q, Daud MK, Chen J, et al. Glutathione S-transferase gene family in *Gossypium raimondii* and *G. arboreum*: Comparative genomic study and their expression under salt stress. Front Plant Sci. 2016; 7: 139. https://doi.org/10.3389/fpls.2016.00139 PMID: 26904090

36. Ouyang Y, Chen J, Xie W, Wang L, Zhang Q. Comprehensive sequence and expression profile analysis of Hsp20 gene family in rice. Plant Mol Biol. 2009; 70: 341–357. https://doi.org/10.1007/s11103-009-9477-y PMID: 19277876

37. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006; 34: W609–W612. https://doi.org/10.1093/nar/gkl315 PMID: 16845082

38. Goldman N, Yang ZH. Codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 1994; 11: 725–736. PMID: 7968486

39. Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. Perspective-Synteny and collinearity in plant genomes. Science. 2008; 320: 486–488. https://doi.org/10.1126/science.1153917 PMID: 18436778

40. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods. 2001; 25: 402–408. https://doi.org/10.1006/meth.2001.1262 PMID: 11846609

41. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, et al. TM4: A free, open-source system for microarray data management and analysis. Biotechniques. 2003; 34: 374–378. PMID: 12613259

42. Zhang YC, Mao LY, Wang H, Brocker C, Yin XJ, Vasiliou V, et al. Genome-wide identification and analysis of grape aldehyde dehydrogenase (ALDH) gene superfamily. Plos One. 2012; 7: e32153. https://doi.org/10.1371/journal.pone.0032153 PMID: 22355416

43. Zhang J. Evolution by gene duplication: an update. Trends Ecol Evol. 2003; 18: 292–298.

44. Kirch HH, Schlingensiepen S, Kotchoni S, Sunkar R, Bartels D. Detailed expression analysis of selected genes of the aldehyde dehydrogenase (ALDH) gene superfamily in *Arabidopsis thaliana*: biochemical requirements, metabolic pathways, and functional analysis. Frontiers Plant Sci. 2011; 2: 65.

45. Wei YL, Lin M, Oliver DJ, Schnable PS. The roles of aldehyde dehydrogenases (ALDHs) in the PDH bypass of *Arabidopsis*. BMC Biochem. 2009; 10: 7. https://doi.org/10.1186/1471-2091-10-7 PMID: 19320993

46. Kirch HH, Schlingensiepen S, Kotchoni S, Sunkar R, Bartels D. Detailed expression analysis of selected genes of the aldehyde dehydrogenase (ALDH) gene superfamily in *Arabidopsis thaliana*. Plant Mol Biol. 2005; 57: 315–332. https://doi.org/10.1007/s11103-004-7796-6 PMID: 15830124

47. Stiti N, Missihiou TD, Kotchoni SO, Kirch HH, Bartels D. Aldehyde dehydrogenases in *Arabidopsis thaliana*: biochemical requirements, metabolic pathways, and functional analysis. Frontiers Plant Sci. 2011; 2: 65.

48. Blanc G, Wolfe KH. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell. 2004; 16: 1667–1678. https://doi.org/10.1105/tpc.021345 PMID: 15208399

49. Flagel LE, Wendel JF. Gene duplication and evolutionary novelty in plants. New Phytol. 2009; 183: 557–564. https://doi.org/10.1111/j.1469-8137.2009.02923.x PMID: 19554355

50. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementarity, degenerative mutations. Genetics. 1999; 151: 1531–1545. PMID: 10101715

51. Lan T, Yang ZL, Yang X, Liu YJ, Wang XR, Zeng QY. Extensive functional diversification of the *Hsp20* gene family in rice. Plant Mol Biol. 2009; 70: 341–357. https://doi.org/10.1007/s11103-009-9577-5 PMID: 19963777

52. Tanaka KM, Takahasi KR, Takano-Shimizu T. Enhanced fixation and preservation of a newly arisen duplicate gene by masking deleterious loss-of-function mutations. Genet Res (Camb). 2011; 95: 557–564. https://doi.org/10.1017/S0016672311000147 PMID: 21935642

53. Guo Y, Haltier U, Ishitani M, Zhu JK. Molecular characterization of functional domains in the protein kinase SOS2 that is required for plant salt tolerance. Plant Cell. 2001; 13: 1383–1399. PMID: 11402167

54. Missihiou TD, Schmitz J, Klug R, Kirch HH, Bartels D. Betaine aldehyde dehydrogenase genes from *Arabidopsis* with different sub-cellular localization affect stress responses. Planta. 2011; 233: 369–382. https://doi.org/10.1007/s00425-010-1297-4 PMID: 21053011

55. Sunkar R, Bartels D & Kirch H. H. Overexpression of a stress-inducible aldehyde dehydrogenase gene from *Arabidopsis thaliana* in transgenic plants improves stress tolerance. Plant J 35, 452–464 (2003). PMID: 12904208

56. Jiao Y, Wickett NJ, Ayyampalayam S, Chandler AS, Landherr L, Ralph PE, et al. Ancestral polyploidy in seed plants and angiosperms. *Nature* 2011; 473: 97–100. https://doi.org/10.1038/nature09916 PMID: 21478875

57. Zhang J. Evolution by gene duplication: an update. Trends Ecol Evol. 2003; 18: 292–298.