

SCIENTIFIC REPORTS



OPEN

Genome-wide identification of glutathione S-transferase gene family in pepper, its classification, and expression profiling under different anatomical and environmental conditions

Shiful Islam¹, Saikat Das Sajib¹, Zakya Sultana Jui², Shatil Arabia², Tahmina Islam² & Ajit Ghosh^{1,3}

Glutathione S-transferases (GSTs) compose a family of multifunctional enzymes involved in the numerous aspects of regulating plant growth, development, and stress response. An *in silico* genome-wide analysis of pepper (*Capsicum annuum* L.) was performed to identify eighty-five GST genes that were annotated according to their chromosomal location. Segmental duplication contributed more than tandem duplication for the expansion of GST gene family in pepper. All the identified members belong to ten different classes which are highly conserved among *Arabidopsis*, rice, tomato and potato counterparts indicating the pre-dicot-monocot split diversification of GST classes. Gene structure, protein domain, and motif organization were found to be notably conserved over the distinct phylogenetic groups, which demonstrated the evolutionary significant role of each class. Expression of most of the *CaGST* transcripts as well as the total pepper GST activity was found to be significantly up-regulated in response to cold, heat, drought, salinity and osmotic stress conditions. Presence of various hormone and stress-responsive cis-elements on most of the putative *CaGST* promoter regions could be directly correlated with the alteration of their transcripts. All these findings might provide opportunities for future functional validation of this important gene family in pepper.

Cellular detoxification is an elementary biological process in all living organisms including animals, plants, and microorganisms. It provides protection against different environmental noxious agents as well as reactive oxygen species (ROS) formed from different metabolic pathways, to ensure the optimum cellular condition for the growth and survival¹. After the conversion of toxic components into reactive electrophiles by cytochrome P450s, these reactive molecules are subsequently transformed by the exertion of phase II enzymes, namely Glutathione S-transferase (GST) for the final degradation process². GSTs are metabolic isozymes that form a complex family with versatile functions in plants³. Among its myriad functions, this family mainly works as a detoxification route in plants by conjugating glutathione (γ -Glu-Cys-Gly) to a diverse range of hydrophobic, electrophilic, xenobiotic compounds and redox buffering to form a soluble S-glutathionylated (R-SG) products⁴. This gene family possesses a high transcript abundance in most of the tissues and thus, act as a biomarker for the detection and monitoring of organ/tissue damage both in plants and animals^{1,5}.

GST genes are abundantly found in animals, plants and even in some prokaryotes⁶. The family consists of cytosolic, mitochondrial and microsomal GSTs. Among them, the mitochondrial and microsomal members

¹Department of Biochemistry and Molecular Biology, Shahjalal University of Science and Technology, Sylhet, 3114, Bangladesh. ²Plant Breeding and Biotechnology Laboratory, Department of Botany, University of Dhaka, Dhaka, 1000, Bangladesh. ³Max-Planck Institute for Plant Breeding Research, Carl-von-Linne-Weg 10, D-50829, Cologne, Germany. Shiful Islam and Saikat Das Sajib contributed equally. Correspondence and requests for materials should be addressed to T.I. (email: subarna.islam@gmail.com) or A.G. (email: ajitghoshbd@gmail.com)

showed significant differences in their sequence and biosynthesis process as compared with the cytosolic GSTs⁷. Classification of the plant soluble GST members is evaluated based on the sequence conservation, genomic organization, kinetic and physicochemical properties, and immunological cross reactivity⁸. There are fourteen classes of GSTs found so far in plants, including tau (U), phi (F), lambda (L), dehydroascorbate reductase (DHAR), theta (T), zeta (Z), eukaryotic translation elongation factor 1B- γ subunit (EF1B γ), tetra-chloro hydroquinone dehalogenase (TCHQD), microsomal prostaglandin E-synthase type 2 (mPGES-2), glutathionyl hydroquinone reductase (GHR), metaxin, Ure2p, hemerythrin (H) and iota (I)⁸. The first four of them (tau, phi, lambda, DHAR) and recently identified two new classes (hemerythrin and iota) are highly plant-specific⁸.

Tau and phi classes have drawn much interest as they are the most abundant classes in plants and perform a major role in the xenobiotics metabolism⁹. Theta-class of GST is mainly involved in the oxidative metabolism and zeta GSTs convert maleylacetoacetate to fumarylacetoacetate in a glutathione-dependent reaction¹⁰. The lambda and DHAR function as thiol transferases by replacing its serine residue to cysteine^{11,12}. The five other classes of GST, such as mPGES-2, GHR, metaxin, hemerythrin, and iota possess cysteine in their active site. Catalytic mechanism of Ure2p is mediated by the asparagine residue which plays a key role in the glutathione stabilization¹³. However, the catalytic nature of the EF1B γ class is not clear to date⁸.

Previous studies have emphasized the involvement of GSTs in various abiotic stress response^{14,15}. The tau class GSTs perform a potential role against the oxidative damages, chemical toxicity, and physical stress agents^{16,17}. Over-expression of *TaGSTU1B* and *TaGSTF6* genes in wheat enhanced tolerance against drought¹⁸. Ectopic expression of a rice tau glutathione S-transferase, *OsGSTU4* improves tolerance to salinity and oxidative stresses in *Arabidopsis* by up-regulating several stress responsive and cellular detoxifying genes¹⁹. Similarly, ectopic expression of stress-inducible *GmGSTU4*, *CsGSTU1* and *CsGSTU2* transcripts in transgenic tobacco plants enhanced tolerance against diphenyl ether herbicide, salinity, and drought stresses^{20,21}.

Genome-wide analysis of GST gene family had been carried out in various plant species, and identified 55 members in *Arabidopsis*¹⁵, 84 in barley²², 65 in *Brassica oleracea*²³, 49 in *G. arboreum* and 59 in *G. raimondii*²⁴, 27 in Japanese larch²⁵, 42 in maize²⁶, 62 in China-pear²⁷, 90 in potato²⁸, 37 in *Physcomitrella patens*²⁹, 81 in populus³⁰, 79 in rice¹⁴, 101 in soybean³¹, 14 in sunflower³², 23 in sweet orange³³, 73 in *Medicago*³⁴, 49 in *Capsella*³⁵ and 90 in tomato³⁶. Although detailed genome sequence information of *Capsicum annuum* is publicly available, genome-wide analysis of GST gene family was not performed yet. Pepper has a relatively large genome size of 3.48 Gb with a life cycle of around 95 days³⁷. In this study, a comprehensive genome-wide analysis of GSTs has been accomplished in *Capsicum annuum* and identified a total of 85 members. Analyses of chromosomal position, physicochemical characteristics, conserved motifs, and subcellular localization of these identified members showed a great variation among themselves. An evolutionary trajectory is drawn between pepper and one of its closest relative tomato, based on their class wise sequence analysis. Furthermore, expression patterns of all the identified genes and total GST activity were analyzed under different developmental and environmental conditions. This study will facilitate a door for the researchers to identify the specific gene/member of the family for crop improvement and stress management.

Results

Identification and nomenclature of GST genes in *Capsicum annuum*. Eighty-five full length genes encoding putative GST proteins were identified in *C. annuum* and classified into ten classes: tau, phi, theta, zeta, lambda, EF1B γ , DHAR, TCHQD, MGST, and GHR. The tau and phi classes are found to be the most abundant with 59 and 6 members, respectively (Table 1). The length of the *CaGST* transcripts ranged from 306 bp (*CaGSTU29*) to 14430 bp (*CaGHR1*), whereas the deduced proteins are 101 to 361 amino acids long. The molecular weight (MW) of the *CaGST* proteins vary from the lowest 11.62 kDa (*CaGSTU29*) to the highest of 74.94 kDa (*CaGSTF5*). However, the predicted pI values ranged from 5 to 9. The average length, MW, and pI of the *CaGST* proteins were found to be 230 aa, 26.3 kDa and 6.5, respectively (Table 1). Most of the *CaGST* proteins were predicted to be localized in the cytoplasm, followed by chloroplast, mitochondria, nucleus, extracellular space, and plasma membrane. Secondary structure analysis showed the presence of a higher percentage of α -helix in *CaGST* proteins as compared with β -sheets (Table S1). The percentage of the extended strand was fluctuating widely among the *CaGST* proteins, in a range of 10 to 28%. Glycosylation analysis showed that 40 *CaGST* proteins have a potential N-glycosylation site, whereas *CaGSTF5* and *CaGHR1* possess the highest number of predicted glycosylation site with seven and five, respectively (Table S2).

Genomic organization of the pepper GST gene family. All the *CaGST* genes are distributed randomly and unevenly in all the 13 chromosomes of pepper (Fig. S1). Chromosome 9 is the most densely populated with fifteen genes (17.6%), followed by ten genes each in chromosome 7 and 0, then nine genes each in chromosome 2 and 11 (Fig. S1). A total of sixteen gene clusters were observed on 12 different chromosomes. Among these clusters, eleven were formed among tau members, one each was for the theta and GHR cluster. Twenty-four sets of *CaGST* proteins appeared to be $\geq 80\%$ similar, that indicates the possible gene duplication events among these genes (Table S3). A maximum of nine duplicated GST genes were located in chromosome 7, followed by six in chromosome 9 and five in chromosome 11. Out of the twenty-four gene pairs, twelve pairs were found as tandemly duplicated, whereas rest twelve pairs appeared as segmentally duplicated. Furthermore, the substitution rate of non-synonymous (d_N) and synonymous (d_S) ratios were calculated to investigate the selective constraints on the duplicated *CaGST* gene pairs, where values >1 , <1 , and equal to 1 implies positive selection, purifying selection, and neutral selection, respectively. All the identified duplicated *CaGST* gene pairs showed the d_N/d_S value less than 0.7, signifying the influence of purifying selection behind the evolution of these gene pairs. Moreover, the estimated divergence time of the duplicated gene pairs varies from 1.18 Mya to 17.84 Mya (Table S3). Identification of the exon-intron organization of 85 *CaGST* genes revealed a group-specific exon/intron patterns within each GST class. A similar type of exon/intron number and length were notable in the

Sl no	Gene Name	Locus ID	CDS coordinate(5'-3')	Strand	Gene (bp)	Protein (aa)	MW (kDa)	pI	Localization
1	<i>CaGSTU1</i>	Capana00g001895	438859152-438859976	+	825	229	26.25	6.14	Cy ^{a,b}
2	<i>CaGSTU2</i>	Capana00g002164	455860084-455861271	+	1188	225	26.00	5.53	Cy ^{a,b}
3	<i>CaGSTU3</i>	Capana00g003105	524690584-524692284	-	1701	203	23.67	8.46	Cy ^{a,b}
4	<i>CaGSTU4</i>	Capana00g003106	524702849-524704044	-	1196	219	25.56	6.60	Cy ^{a,b}
5	<i>CaGSTU5</i>	Capana00g004478	647297751-647300086	-	2336	221	25.12	5.21	Cy ^a , Cp ^b
6	<i>CaGSTU6</i>	Capana00g004596	657125404-657128170	+	2767	222	25.73	5.59	Cy ^a , Cy ^b
7	<i>CaGSTU7</i>	Capana00g004598	657374139-657375389	-	1251	220	25.34	5.75	Cy ^{a,b}
8	<i>CaGSTU8</i>	Capana00g004621	659428697-659429026	+	330	109	12.84	6.97	Cy ^{a,b} , Mt ^a , Cp ^a
9	<i>CaGSTU9</i>	Capana00g004665	665401358-665402749	-	1392	224	25.87	5.57	Cy ^a , Cp ^b
10	<i>CaGSTU10</i>	Capana01g002551	166700570-166702635	+	2066	221	25.08	5.08	Cy ^a , Nu ^b
11	<i>CaGSTU11</i>	Capana01g003292	222778803-222779584	-	782	225	25.63	6.33	Cy ^{a,b}
12	<i>CaGSTU12</i>	Capana01g003296	222806252-222809195	-	2944	222	25.91	6.00	Cy ^{a,b}
13	<i>CaGSTU13</i>	Capana02g000947	103858121-103858901	-	781	153	17.40	5.57	Cy ^{a,b}
14	<i>CaGSTU14</i>	Capana02g000948	103899028-103899345	-	318	105	11.98	9.13	Cy ^a , Cp ^a , Nu ^b
15	<i>CaGSTU15</i>	Capana02g000950	104101699-104102578	+	880	225	26.05	5.36	Cy ^a , Cp ^b
16	<i>CaGSTU16</i>	Capana02g000952	104250329-104251764	+	1436	224	25.67	5.40	Cy ^{a,b}
17	<i>CaGSTU17</i>	Capana03g000768	11322594-11323373	-	780	212	24.43	5.44	Cy ^a , Cp ^b
18	<i>CaGSTU18</i>	Capana03g000769	11362575-11363976	-	1402	233	27.12	5.29	Cy ^{a,b}
19	<i>CaGSTU19</i>	Capana03g004402	253298052-253300358	+	2307	222	25.91	5.00	Cy ^a
20	<i>CaGSTU20</i>	Capana03g004562	257827721-257828950	-	1230	225	25.75	6.10	Cy ^{a,b}
21	<i>CaGSTU21</i>	Capana03g004565	258031244-258032469	+	1226	225	26.05	6.25	Cy ^{a,b}
22	<i>CaGSTU22</i>	Capana03g004566	258056794-258057997	+	1204	225	25.93	7.02	Cy ^{a,b}
23	<i>CaGSTU23</i>	Capana06g002861	210286274-210287439	-	1166	220	25.36	5.53	Cy ^{a,b}
24	<i>CaGSTU24</i>	Capana07g002003	210859245-210865929	+	6685	220	25.49	5.89	Cy ^{a,b}
25	<i>CaGSTU25</i>	Capana07g002004	210867066-210867991	+	926	222	25.60	5.79	Cy ^{a,b}
26	<i>CaGSTU26</i>	Capana07g002005	210902595-210903564	+	970	175	20.42	6.84	Cy ^{a,b} , Cp ^a
27	<i>CaGSTU27</i>	Capana07g002006	210941269-210948401	+	7133	282	32.96	8.35	Cy ^{a,b}
28	<i>CaGSTU28</i>	Capana07g002007	210981175-210982949	+	630	209	24.34	8.08	Cy ^{a,b}
29	<i>CaGSTU29</i>	Capana07g002008	211057582-211057887	+	306	101	11.62	5.14	Cy ^{a,b}
30	<i>CaGSTU30</i>	Capana07g002009	211066612-211068955	+	2344	238	27.61	6.25	Cy ^{a,b}
31	<i>CaGSTU31</i>	Capana07g002010	211152285-211153567	+	1223	220	25.54	5.78	Cy ^{a,b}
32	<i>CaGSTU32</i>	Capana07g002011	211155192-211156505	-	1314	219	25.32	6.84	Cy ^{a,b}
33	<i>CaGSTU33</i>	Capana07g002012	211160558-211161672	-	1115	203	23.75	7.00	Cy ^{a,b}
34	<i>CaGSTU34</i>	Capana08g001515	132057487-132058584	+	1098	226	25.76	5.38	Cy ^{a,b}
35	<i>CaGSTU35</i>	Capana08g001518	132082843-132083882	+	1040	202	22.88	5.36	Cy ^a , Cp ^a
36	<i>CaGSTU36</i>	Capana08g001520	132086808-132087750	+	943	181	20.61	5.11	Cy ^{a,b} , Cp ^a
37	<i>CaGSTU37</i>	Capana09g001740	199536703-199537891	+	1189	223	25.30	5.30	Cy ^a , Cp ^b
38	<i>CaGSTU38</i>	Capana09g001741	200130618-200131846	+	1229	220	25.04	5.48	Cy ^a , Cp ^b
39	<i>CaGSTU39</i>	Capana09g001742	200167723-200169649	+	1927	224	26.36	6.77	Cy ^a , Nu ^b
40	<i>CaGSTU40</i>	Capana09g001760	202017261-202018643	+	1383	217	25.05	5.28	Cy ^{a,b}
41	<i>CaGSTU41</i>	Capana09g001761	202086834-202088344	+	1511	220	24.85	5.41	Cy ^a , Cp ^b
42	<i>CaGSTU42</i>	Capana09g001762	202236152-202237088	+	937	220	25.45	5.12	Cy ^a , Cp ^b
43	<i>CaGSTU43</i>	Capana09g001763	202238222-202239079	+	858	154	17.67	5.12	Cy ^a , Mt ^a
44	<i>CaGSTU44</i>	Capana09g001764	202383049-202384214	+	1166	220	25.38	5.53	Cy ^{a,b}
45	<i>CaGSTU45</i>	Capana09g001858	213200912-213201949	+	1038	217	25.28	5.38	Cy ^a , Cp ^b
46	<i>CaGSTU46</i>	Capana09g001859	213228499-213229454	-	956	219	24.86	5.87	Cy ^{a,b}
47	<i>CaGSTU47</i>	Capana09g001860	213238117-213239430	-	1314	221	25.19	5.23	Cy ^a , Nu ^b
48	<i>CaGSTU48</i>	Capana09g001861	213240773-213241709	+	937	224	25.76	5.58	Cy ^a , Cp ^b
49	<i>CaGSTU49</i>	Capana09g001862	213303661-213304284	+	624	143	16.62	4.84	Cy ^a , Mt ^a
50	<i>CaGSTU50</i>	Capana09g002045	226590245-226590574	+	330	109	13.06	8.78	Cy ^{a,b} , Mt ^a
51	<i>CaGSTU51</i>	Capana11g001524	178109400-178111728	+	2329	231	25.76	6.18	Cy ^a , Cp ^{a,b}
52	<i>CaGSTU52</i>	Capana11g001525	178112748-178114176	+	1429	215	23.85	8.88	Cy ^a , Cp ^{a,b}
53	<i>CaGSTU53</i>	Capana11g001528	178187645-178188768	+	1124	220	24.84	6.75	Cy ^a , Cp ^{a,b}
54	<i>CaGSTU54</i>	Capana11g001532	178455711-178456697	+	987	229	26.20	6.76	Cy ^a , Cp ^b
55	<i>CaGSTU55</i>	Capana11g001533	178457971-178458582	+	612	203	23.42	6.31	Pm ^a , Ec ^b
56	<i>CaGSTU56</i>	Capana11g001535	178483866-178484900	+	1035	229	26.01	5.57	Cy ^a , Cp ^b

Continued

Sl no	Gene Name	Locus ID	CDS coordinate(5'-3')	Strand	Gene (bp)	Protein (aa)	MW (kDa)	pI	Localization
57	<i>CaGSTU57</i>	Capana11g001536	178558594–178560496	–	1903	229	26.25	5.45	Cy ^a , Cp ^b
58	<i>CaGSTU58</i>	Capana11g001537	178588610–178590596	+	1987	229	26.13	5.62	Cy ^a , Cp ^b
59	<i>CaGSTU59</i>	Capana12g000354	7186608–7188114	+	1507	219	25.57	5.47	Cy ^a , Nu ^b
60	<i>CaGSTF1</i>	Capana02g002285	142882377–142883807	+	1431	228	26.02	6.09	Cy ^a , Cp ^b
61	<i>CaGSTF2</i>	Capana03g003600	230428836–230430376	+	1541	213	24.19	5.69	Cy ^a , Cp ^b
62	<i>CaGSTF3</i>	Capana06g003058	218203021–218205366	+	2346	213	23.76	6.39	Cy ^a , Cp ^b
63	<i>CaGSTF4</i>	Capana06g001819	54683346–54684364	–	1019	251	29.29	8.68	Cy ^{a,b} , Mt ^a , Nu ^a
64	<i>CaGSTF5</i>	Capana12g000609	14792240–14802489	+	10250	654	74.94	7.13	Pm ^a , Cy ^b
65	<i>CaGSTF6</i>	Capana12g000612	14873894–14878274	+	4381	252	28.64	5.26	Cy ^a , Ec ^a , Cp ^b
66	<i>CaGSTT1</i>	Capana01g000218	3161948–3164992	+	3045	186	21.15	9.74	Mt ^a , Nu ^a , Cy ^b
67	<i>CaGSTT2</i>	Capana01g000219	3169497–3173899	+	4403	250	28.34	9.25	Cy ^a , Mt ^a , Nu ^a
68	<i>CaGSTT3</i>	Capana01g000221	3200545–3205818	+	5274	241	27.33	9.26	Cy ^{a,b} , Mt ^a
69	<i>CaGSTT4</i>	Capana12g001178	52439179–52442473	–	3295	235	26.46	6.45	Cy ^a , Cp ^b
70	<i>CaGSTZ1</i>	Capana01g001395	45340308–45347371	–	7064	283	32.48	8.34	Mt ^a , Cy ^a , Cp ^{b,c}
71	<i>CaGSTZ2</i>	Capana12g000896	32868383–32869334	+	952	107	12.47	5.01	Pm ^a , Cy ^b
72	<i>CaGSTL1</i>	Capana05g000544	14609977–14614464	+	4488	314	35.45	8.71	Cp ^{a,b,c}
73	<i>CaGSTL2</i>	Capana09g001864	213395287–213400333	+	5047	239	27.75	5.57	Cy ^{a,b}
74	<i>CaGSTL3</i>	Capana10g001792	183887466–183889739	–	2274	222	25.80	5.24	Cy ^{a,b}
75	<i>CaGSTL4</i>	Capana10g001806	186021678–186025506	–	3829	253	29.10	5.75	Cy ^{a,b} , Mt ^a
76	<i>CaEF1Bγ1</i>	Capana11g000124	3480968–3485480	–	4513	414	46.83	5.65	Cy ^a , Cp ^b
77	<i>CaEF1Bγ2</i>	Capana08g000330	43923121–43924032	–	912	204	23.47	8.67	Cy ^a , Nu ^a
78	<i>CaDHAR1</i>	Capana05g000799	31400365–31405449	+	5085	295	32.75	8.33	Mt ^a , Cp ^{a,b,c}
79	<i>CaDHAR2</i>	Capana05g002401	214455819–214460536	–	4718	212	23.57	5.88	Cy ^{a,b}
80	<i>CaTCHQD</i>	Capana00g004998	686182991–686184172	+	1182	268	31.64	9.06	Mt ^a , Cy ^a
81	<i>CaMGST1</i>	Capana02g002296	143033395–143035039	+	1645	144	16.38	9.15	Pm ^a
82	<i>CaMGST2</i>	Capana04g000166	2049728–2054459	–	4732	323	36.21	9.02	Cp ^{a,c}
83	<i>CaGHR1</i>	Capana02g001236	116162258–116176687	+	14430	361	41.43	6.80	Mt ^a , Nu ^a , Cp ^c
84	<i>CaGHR2</i>	Capana02g000926	102853919–102854627	–	709	123	13.73	4.37	Nu ^a , Cy ^a
85	<i>CaGHR3</i>	Capana06g000146	1927452–1929409	+	1958	373	41.82	6.61	Cp ^{a,c}

Table 1. List of identified GST genes in pepper along with their detailed genomic and proteomic information. Abbreviations- CDS, coding DNA Sequence; MW, Molecular Weight; pI, Isoelectric point; bp, base pair; aa, amino acid; kDa, kilodalton; Cp, Chloroplast; Ec, Extracellular; Cy, Cytoplasm; Mt, Mitochondria; Nu, Nucleus; Pm, Plasma-membrane. ^aLocalization prediction by CELLO v.2.5 (<http://cello.life.nctu.edu.tw/>). ^bLocalization prediction by pSORT (<https://wolfsort.hgc.jp/>). ^cChloroplast localization signal confirmed by ChloroP (<http://www.cbs.dtu.dk/services/ChloroP/>).

phylogenetically related members (Fig. 1). The tau-class *CaGSTs* typically contained one/two exons, while phi *GST* genes contained two to five exons in its gene structure (Fig. 1). *CaGSTZ1*, *CaGSTL1*, and *CaGSTL4* contained the maximum number of nine introns and ten exons, followed by 9 exons in *CaGSTL2*, 8 exons in *CaGSTL3*, 7 exons in *CaGSTT2*, *CaGSTT3*, *CaGSTT4* and *CaDHAR1* (Fig. 1). An interesting pattern of intron distribution was observed among the putative paralogous members. Most of the paralogs showed same number and similar sized intron such as gene cluster of *CaGSTU13*, 46, 42 and *CaGSTU23*, 44, 45; while some of them showed intron gain/loss phenomenon as compared with their paralogs (*CaGSTU26* and 30, *CaGSTL3* and 4, *CaEF1Bγ1* and 2, *CaGHR1* and 2, and *CaDHAR1* and 2). Few of them showed partial deletion of exon such as *CaGSTU54* and 56 as compared with their two paralogs- *CaGSTU57* and 58 (Fig. S2A), while some others showed exon duplication such as *CaGSTU14* and *CaGSTU47* (Fig. S2B).

Domain architecture analysis showed that 60 out of 85 *CaGST* proteins contained two conserved GST domains, namely, N-terminal and C-terminal (Fig. S3). Only one N-terminal domain present in 22 *CaGSTs*, whereas only C-terminal domain present in 3 *CaGST* proteins (*CaGSTU13*, *CaGSTU43*, and *CaGHR2*). Additional distinct EF1Bγ domain (PF00736) was found to be present in the *CaEF1Bγ1* and *CaEF1Bγ2* proteins, while the MAPEG domain (PF01124) was present only in *CaMGST1* protein. Analysis of conserved motifs of *CaGST* proteins identified ten distinct motifs (Table S4, and Fig. S4). Among them, four motifs are located in the N-terminal GST domain, while the rest six reside in the C-terminal GST domain of the proteins. Motif3 is found to be present in the N-terminal of 65 proteins; followed by motif 1, motif 4, motif 2, motif 7, and motif 5 with 58, 57, 55, 53 and 51 sites, respectively. Interestingly, lambda class specific pattern was shown by motif 9, while motif 10 was present only in four tau class members (*CaGSTU1*, *CaGSTU20*, *CaGSTU21*, and *CaGSTU22*). Among others, motif 6 and 8 were present in 44 and 16 *CaGSTs*, respectively. The lowest number of one motif (either motif 4 or 8) is present in *CaGSTZ2*, *CaGSTL2*, *CaEF1Bγ2*, *CaMGST1*, *CaGHR2*, and *CaGHR3*.

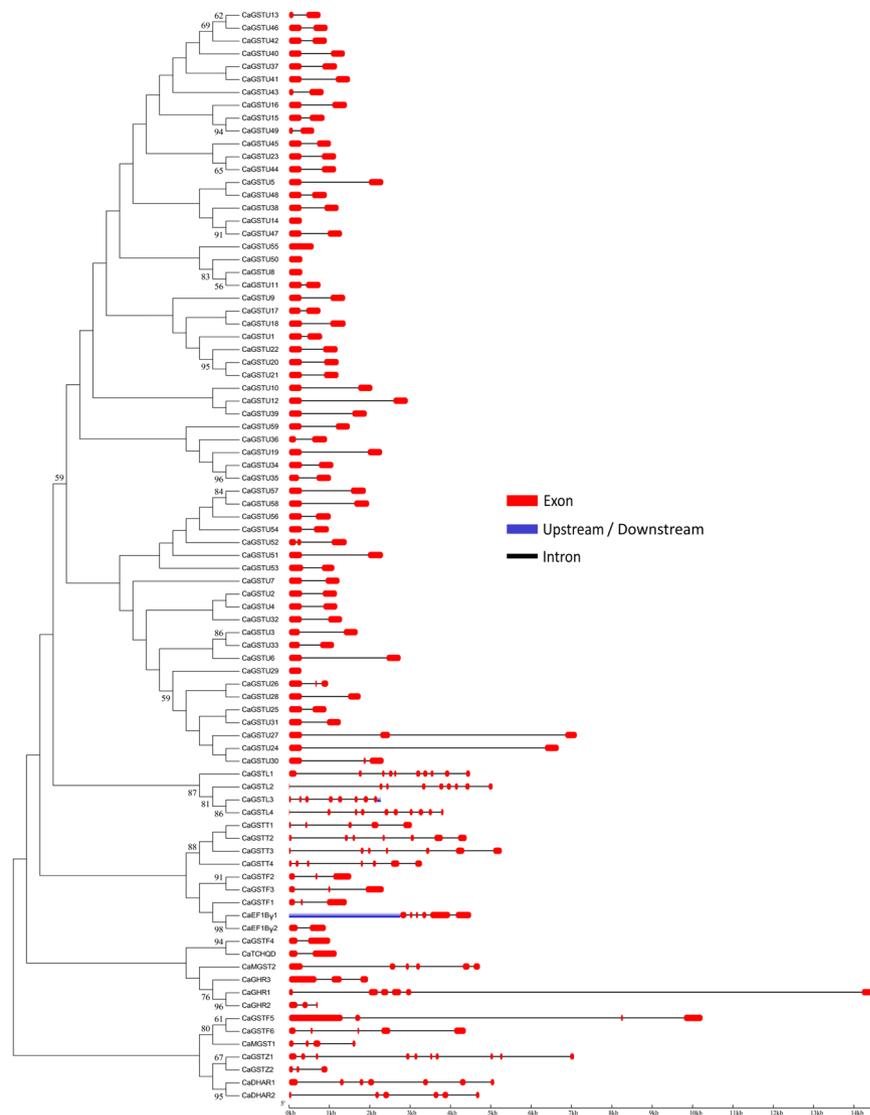


Figure 1. Phylogenetic and Structural analysis of *CaGST* genes. **(A)** Number at each node in the phylogenetic tree represents the bootstrap value higher than 50. The different class of *GST* genes forms separate clades. **(B)** The schematic diagram represents the gene structure of all 85 *CaGST* genes identified in this study. Exons are shown as red boxes; introns are shown as black lines, and the upstream/downstream regions are shown as blue boxes. The relative size of the full transcript, intron, exon, and upstream region could be inferred from the supplied scale in kilobase pair (kb).

Phylogenetic analysis of GST proteins. The evolutionary relationship of GST family members was predicted by comparing them to different plant species. A total number of 401 full length GST protein sequences from five different plant species- pepper, *Arabidopsis*, rice, tomato, and potato were aligned to create an unrooted maximum likelihood phylogenetic tree (Fig. 2). All these *CaGST* proteins were found to be closely associated with an individual class of GSTs. Most classes formed a monophyletic group with very few exceptions of *OsGSTU32*, *CaGSTU10*, and *SlGSTT1* (Fig. 2). The tree clearly suggested that the largest and most abundant class of GST is tau (brown circle) with 59, 52, 28, 57 and 66 members in pepper, rice, *Arabidopsis*, tomato, and potato; respectively. All these tau members are distributed in 6 different small clades under a large superclade, which indicates the presence of internal variation among these tau members. Similarly, the second largest clade is formed by the phi members with a green rectangle. The clustering of same classes of GST from five plant species demonstrates that the presence of all these GST individual classes during the divergence of plants followed a species-specific gene duplication. The major divergences that divided the family into 10 individual classes might occur in the common ancestor of all the investigated species.

Molecular evolution of *Capsicum* and tomato GST family. To investigate the lineage-specific expansion of *GST* genes in *Capsicum* and tomato genome, a phylogenetic analysis was performed using the GST protein sequences from *Capsicum* and tomato. A total of 85 *Capsicum* and 90 tomato GSTs fell into ten distinct classes. The divergence point between *Capsicum* and tomato were labeled with circles on certain specific nodes of the

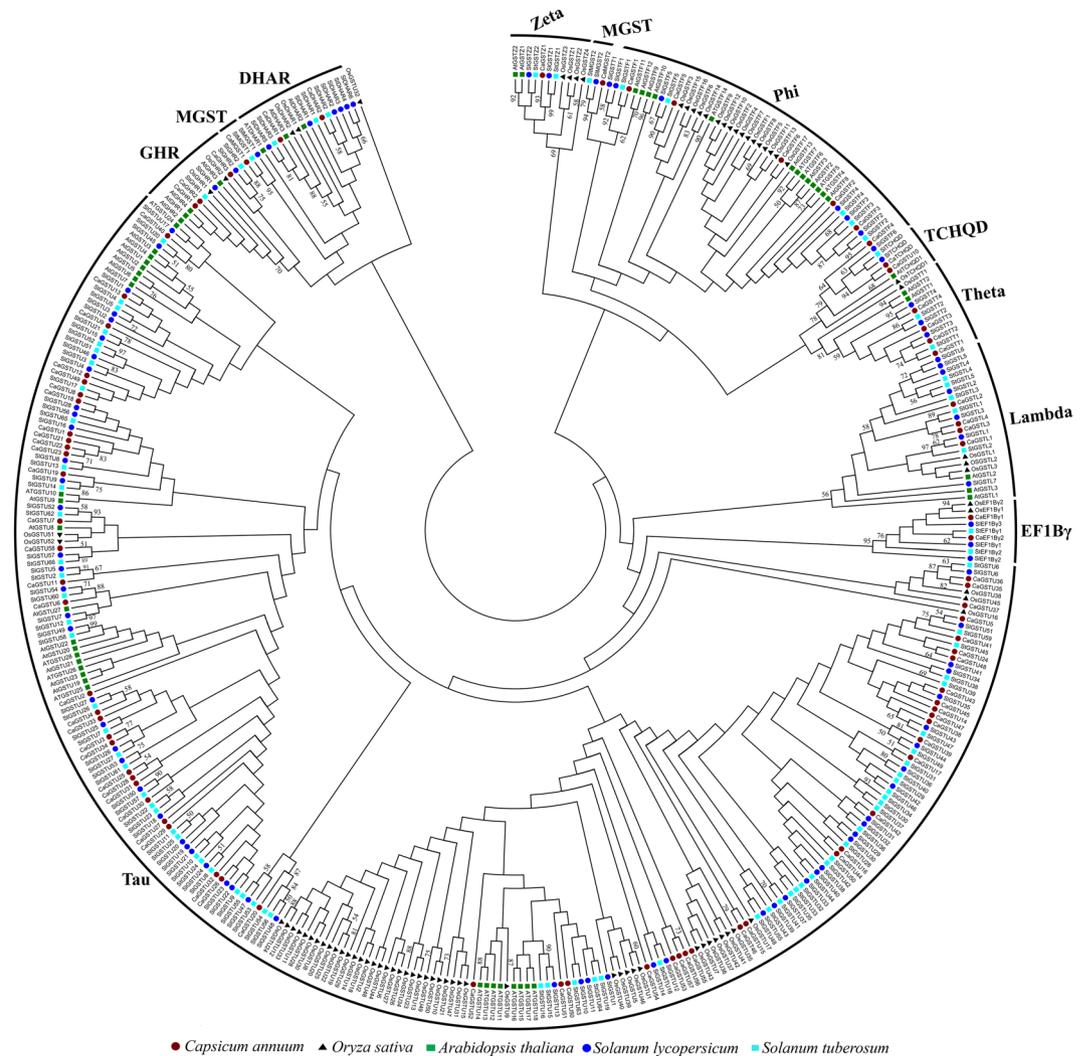


Figure 2. Phylogenetic analysis of GST proteins. GST proteins from five species- pepper, tomato, potato, *Arabidopsis*, and rice were used to construct an unrooted phylogenetic tree using MEGA Maximum-likelihood method with 1000 bootstraps. GST members from tau, lambda, zeta, DHAR, theta, GHR, TCHQD, phi, MGST, EF1B γ classes were marked with red, green, cyan, blue, violet, gray, coral, yellow, pink, royal blue; respectively. Members of each class formed different clades with few exceptions.

phylogenetic tree that represents the most recent common ancestral (MRCA) genes before the split (Figs S5–S10). Most of the MRCA have a member from both the species, while some of them have representation from either *Capsicum* or tomato, indicating a subsequent loss in tomato and *Capsicum* genome, respectively. There were at least 55 MRCA for tau GST genes between *Capsicum* and tomato out of a total of 59 and 57 members, respectively (Fig. 3). After the split, *Capsicum* gained 17 genes and lost 13 genes, resulting in the 59 tau GST genes; while tomato gained 13 genes and lost 11 genes, resulting in 57 tau genes. For phi GST, there are at least 7 MRCA between *Capsicum* and tomato (Fig. 3). After the split, both of them lost one member without any gain, resulting 6 phi members in each species (Fig. 3). Similarly, two genes have been lost from 6 MRCA of theta GST to result in 4 existing members in each species.

For zeta, lambda, EF1B γ , and GHR GSTs, there are at least three ancestral genes in the MRCA analysis of *Capsicum* and tomato (Fig. 3). After the split, both *Capsicum* and tomato lost one zeta member without any gain; no loss with one gain for *Capsicum* and three gain for tomato lambda GST; one loss without gain for *Capsicum* and two loss with two gain for tomato EF1B γ GST; and no loss/gain for *Capsicum* and one loss without any gain for tomato GHR GST genes. For DHAR GST, the MRCA of *Capsicum* and tomato had at least two ancestral members (Fig. 3). After the split, there was no gain/loss event in *Capsicum*, while four genes have been gained by tomato. For TCHQD GST and MGST classes, there were no identifiable gene gain or loss events after the split of these two species (Fig. 3).

Expression analysis of *CaGST* transcripts in different tissues. To investigate the putative roles of *CaGST* genes in *C. annuum* growth and development, expression of all the identified *CaGST* genes were analyzed in the 57 different tissues and organs based on the RNA-seq data. All these tissues could be broadly represented

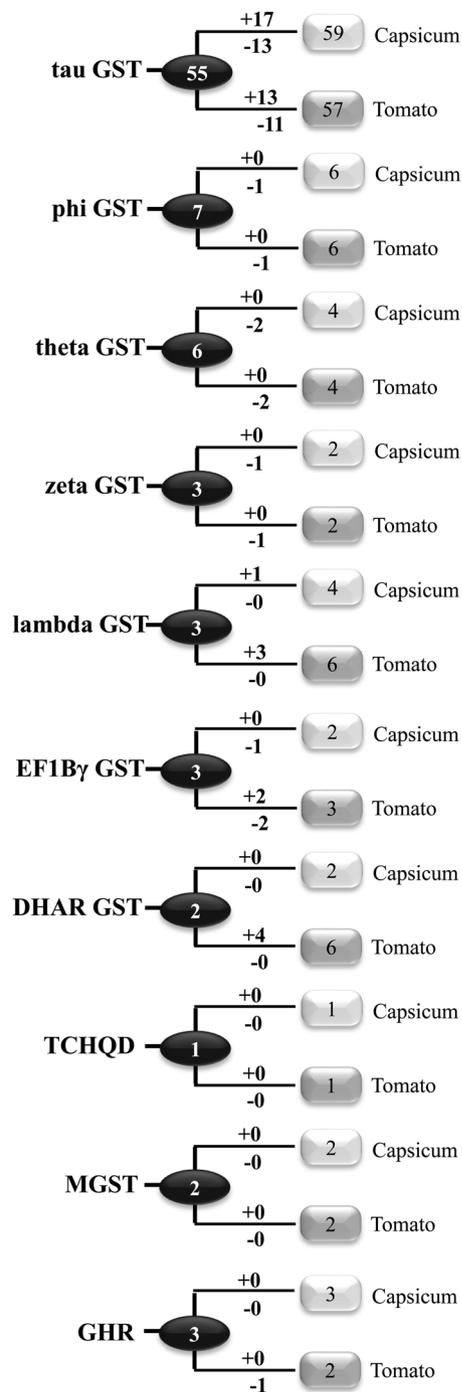


Figure 3. Copy number variation between *Capsicum* and tomato *GST* genes. The numbers presented inside the circles and rectangles represent the numbers of *GST* genes as common ancestral and species-specific, respectively. Numbers on the branches with plus and minus symbols represent the numbers of gene gains and losses, respectively throughout the evolution between these two species. The light gray boxes represent *Capsicum*, while dark gray boxes symbolize tomato.

into nine major organs/stages: seedlings, flower, petal, ovary, anther, fruit, pericarp, seed, and placenta. The analysis revealed a differential pattern of expression for different *CaGST* transcripts depending on the type of tissues and organs (Fig. 4). Based on the differential expression patterns, all these genes could be classified into three groups: a) Some *CaGSTs* showed extremely low levels of expression in almost all tissues and organs, b) Some *CaGSTs* exhibited low to medium levels of expression among different organs/tissues, and c) Some were highly expressive across all the tissues of the entire life cycle. Among all, *CaDHAR2* and *CaEF1B γ 1* showed the maximum level of expression in all the tissues, while other members of this clade such as *CaGSTF3*, *CaGSTU31*, *CaGSTU24*, *CaGSTU3*, *CaGSTT3*, *CaGSTL2*, *CaGSTU12*, *CaGSTL3*, *CaGSTU32*, *CaGSTT4*, *CaMGST1*, and *CaGHR1* exhibited a high level of expression in most of the tissues (Fig. 4). Notably, two clusters of *GST* genes

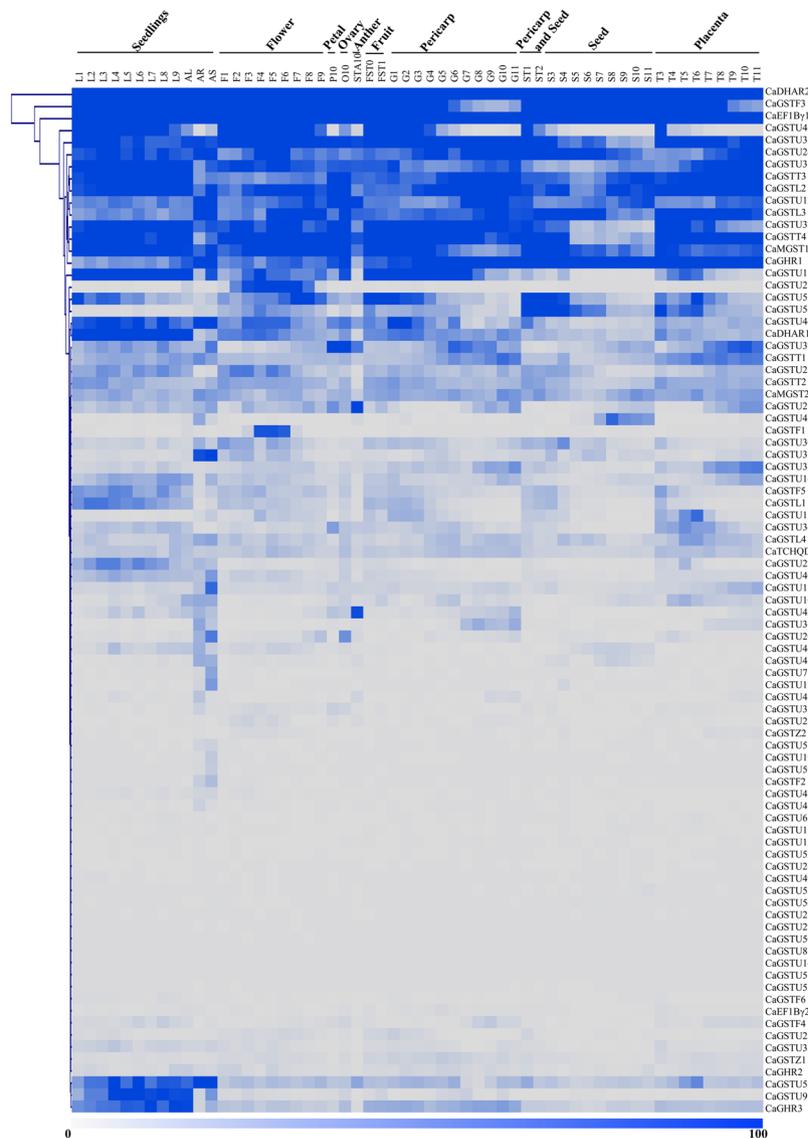


Figure 4. Expression profiling of *CaGST* genes at different anatomical tissues. Expression of all the identified *CaGST* transcripts was analyzed in 57 developmentally diverse tissues using the RNA-seq data. All these tissues could be categorized into nine major stages, such as seedlings, flower, petal, ovary, anther, fruit, pericarp, seed, and placenta. Heatmap with hierarchical clustering was performed using the expression values in MeV software package with Manhattan correlation. The highest level of expression is represented by dark blue (100%), while the low level is presented as white (0%). Thus, the intensity of the color in the heatmap is directly proportioned to the transcript abundance of each member.

from *CaGSTU51* to *CaDHAR1* and from *CaGSTU5* to *CaGHR3* (Fig. 4) maintained a medium to high levels of expression in all the analyzed tissues. Interestingly, some of the members showed a highly tissue-specific expression, such as *CaGSTU41* showed a low level of expression in pericarp, seed and placental otherwise the level is high in other tissues; *CaGSTU28* and *CaGSTF1* possessed only flower-specific expression; expression of *CaGSTU56* raised significantly in seed and placenta; and the cluster of *CaGSTU5* to *CaGHR3* showed a high level of seedling specific expression (Fig. 4). Thus, different *GST* transcripts might have different developmental and tissue-specific regulation to maintain their specific localized or ubiquitous function.

Transcript analysis of *CaGSTs* in response to various abiotic stresses. To identify the abiotic stress-responsiveness, expression profiling of all *CaGST* transcripts were further analyzed in response to five different abiotic stresses viz. cold, heat, drought, salinity and osmotic using the Illumina RNA-Seq data (Fig. 5). Transcriptome profiling was performed in the leaf and root tissue of hot pepper at six different time points such as 1 h, 1.5 h, 3 h, 6 h, 12 h, and 24 h. Most of the genes showed a different level of upregulation with few downregulation events. As compared with the leaf samples (Fig. 5A), the genes showed more upregulation in the root samples (Fig. 5B) as root is one of the first organs to perceive the adverse conditions. A large cluster of genes (*CaGSTU5* to *CaEF1Bγ2*) showed a minimum alteration in response to all these stresses in both the leaf and root tissues.

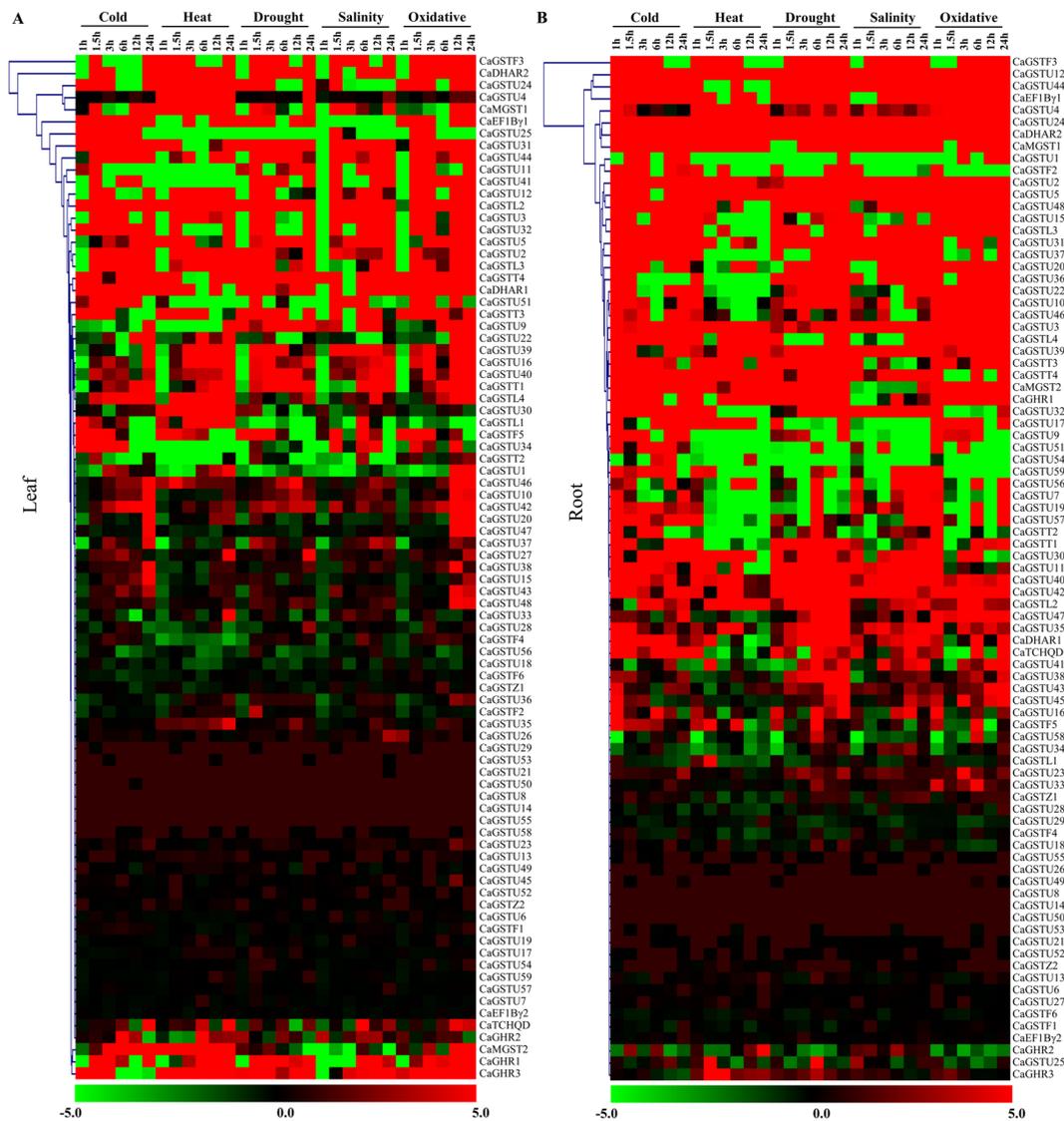


Figure 5. Alteration of *CaGST* transcripts in various abiotic stresses. Expression of all 85 *CaGST* transcripts was analyzed from the leaf (A) and root (B) samples treated with five major abiotic stresses, such as cold, heat, drought, salinity and oxidative. Relative fold change of transcript abundance in seven different time points of each stress (1 h, 1.5 h, 3 h, 6 h, 12 h, 24 h, and 48 h) was plotted with MeV software package. The color scale, depicted at the down of each heatmap, represents the intensity of alterations where green color indicates downregulation and red indicates upregulation.

Two clusters of genes; one from *CaGSTF3* to *CaMGST1*, and another from *CaGSTU2* to *CaGSTU48*, showed the maximum upregulation in the root against all these stresses (Fig. 5B), while there is no such unique pattern of upregulation in the leaf (Fig. 5A). Moreover, there are a few stress-specific transcript alteration events in the root (Fig. 5B) such as *CaGSTU1* showed cold and oxidative stress-specific upregulation; a cluster of *CaGSTU15* to *CaGSTU46* showed heat-specific downregulation; and *CaGSTF2*, *CaGSTU51*, and *CaGSTU54* exhibited cold-specific upregulation. Likewise, a cold specific upregulation for *CaGSTU25* and *CaGSTU34*; and drought and salinity specific upregulation for *CaGSTU11* and *CaGSTU44* was observed in the leaf samples (Fig. 5A).

Stress-responsive alteration of total GST activity. In order to validate the abiotic stress specific up-regulation of the majority of *CaGST* transcripts, the total GST activity was measured in response to the same abiotic stress conditions- cold, heat, drought, salinity, and oxidative; and compared with the respective untreated control conditions (Fig. 6). As shown in Fig. 6, a strong positive induction of GST activity was observed in response to all these stress treatments. GST enzyme activity enhanced significantly in response to heat, cold, salt, and oxidative stresses as compared with their untreated (0 h) sample (Fig. 6). A gradual enhancement of GST activity with time was observed for heat, cold and salt stresses, while total GST activity under oxidative stress reached a maximum level within 12 h of stress imposition, and maintained the level until 24 h observation period (Fig. 6F). Among all these stresses, drought showed minimum induction with a slow rate over the time

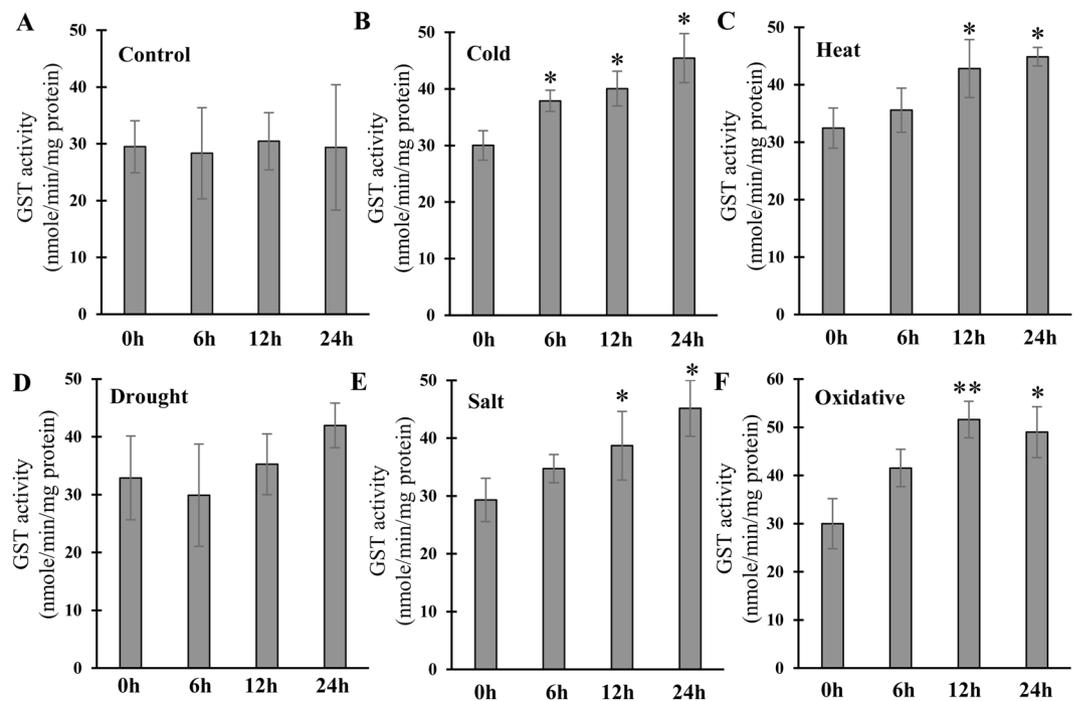


Figure 6. Measurement of total GST activity in response to various abiotic stresses. Total GST enzyme activity was measured in response to various abiotic stresses such as cold, heat, drought, salinity, and oxidative at four different time points of stress exposure. The activity was represented as nmoles/min/mg protein. All the experiments were repeated thrice and represent as the average \pm standard deviation ($n = 3$). The significance level of the paired student's two-tailed t-test is represented as * and ** with a p-value less than 0.05 and 0.01; respectively.

of 24 h (Fig. 6D). However, the level of total GST activity in the untreated control sample remains almost similar within the 24 h experimental period (Fig. 6A).

Presence of Cis-elements in the *CaGST* promoters. To understand the tissue-specificity or stress-responsive transcriptional regulation of *CaGST* genes, the 1000 bp 5' upstream region from the transcriptional start site (ATG) was retrieved and scanned through PlantCARE database for the identification of important cis-regulatory elements. The promoters possessed several cis-elements that confer hormone and stress response. We have identified the presence of seven hormone-related such as abscisic acid-responsive (ABRE), auxin-responsive (AUXRR-core), ethylene-responsive (ERE), gibberellin-responsive (GARE and P-box), salicylic acid-responsive elements (TCA-element) and MeJA-responsive element (TGACG-motif); seven defense and stress-responsive elements such as fungal elicitor-responsive (Box-W1, W-Box), heat stress-responsive (HSE), low-temperature-responsive (LTR) elements, MYB binding site involved in drought-inducibility (MBS), stress responsiveness (TC-rich repeats), and wound-responsive element (WUN-motif); and one transcriptional enhancer (pyrimidine-rich motif) (Fig. 7). The upstream region of most *CaGST* members contained at least one hormone-related and one stress-related element. The top four abundant elements involved in the hormone and stress responses are - MBS element with 85 instances, 83 instances of the defense and stress-responsive element (TC-rich repeats), 64 instances of HSE element and 63 instances of TCA element (Table S5). The promoter of *CaTCHQD* contains the highest number of 15 cis-element, followed by the promoter of *CaGSTU26* with 13 members. Whereas the minimum number of one cis-regulatory element was positioned in the promoter of *CaGSTU51* and *CaMGST1*, followed by the promoter of *CaGSTU7* and *CaGSTU8* with 2 motifs (Fig. 7). Presence of such diverse type of hormone and stress-inducible cis-elements could be directly correlated with the positive alteration of the *CaGST* transcripts (Fig. 5) and total GST activity (Fig. 6) under various abiotic stress conditions.

Discussion

Hot Pepper (*C. annuum*) is one of the most economically important, nutritious and world-wide grown spicy vegetable³⁷. Various environmental stresses including salt, wind, cold, temperature, drought, humidity and osmotic stress cause a serious threat and thus, constrain the total production of pepper³⁸. Water deficit notably diminishes the final fruit production in *Capsicum chinense* by interfering with the process of flowering and fruit development³⁹. Much focus has been given in recent years to understand the stress responses and adaptation mechanisms of pepper by identifying many of the master regulatory genes^{40–42}. GSTs are a promiscuous enzyme family that plays a vital role in the growth and development as well as stress management. A comprehensive genome-wide exploration identified a total of eighty-five *CaGST* gene members (Table 1) with at least one conserved GST domain (Fig. S3). The number of identified *GST* genes are higher in pepper as compared to most other species;

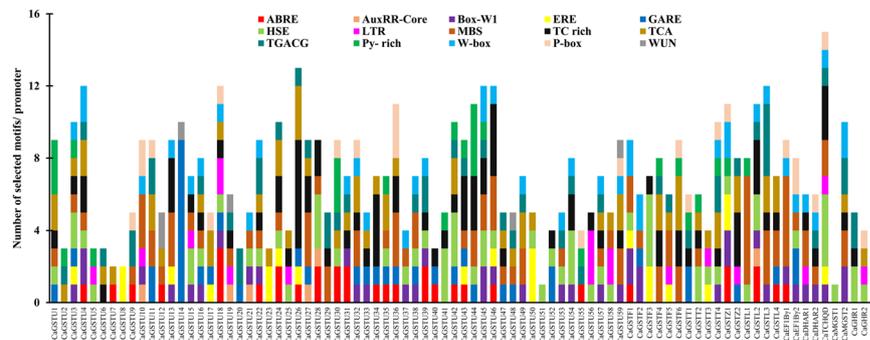


Figure 7. Analysis of the putative promoter of *CaGST* genes. One kb 5' upstream sequences of all the identified *CaGST* genes were retrieved from the genome database and analyzed through PlantCARE to identify the presence and number of various cis-acting regulatory elements. Different hormone-responsive and stress-related elements were identified and plotted against a bar diagram. The abundance of different regulatory elements on each of the promoter was represented with different colors.

but fewer than barley, tomato, potato and soybean^{14,15,22–25,28–33,36,43,44}. Among the ten identified classes of *CaGST*, tau is the most abundant class, followed by phi (Table 1). The almost similar pattern of tau and phi dominance was observed in other plant species; and thus, termed as plant-specific GSTs^{15,45}. In spite of the ubiquitous abundance of the tau class in tracheophytes (25–62 copies), they are completely missing in *Physcomitrella patens* and green algae³⁵. The possible reason behind the tau and phi class-specific expansion might be their substantial functional influence on the xenobiotics metabolism and stress tolerance against both pathogens and environmental factors^{35,46–48}.

The expansion of a gene family occurs mainly through three evolutionary mechanisms such as tandem duplication, segmental duplication, and transposition⁴¹. Due to the polyploidy nature of most plants, segmental duplication occurs frequently as compared with the other two means. The members of *CaGST* gene family arise largely due to several rounds of tandem and segmental gene duplications. Among the 59 tau GSTs in *Capsicum*, sixteen (27%) were created by segmental duplication and fourteen (23%) by the tandem duplication event, indicating that segmental duplication has the major contribution for the rapid expansion of the tau GSTs in *Capsicum* (Table S3). Apart from that, gene clustering also played an important role in the family expansion where 52 out of 59 *CaGSTUs* (88%) were presented in the eleven clusters on eight chromosomes (0, 1, 2, 3, 7, 8, 9 and 11). Similarly, 3 out of 4 theta GST (75%), 3 out of 3 DHAR GST (100%), 2 out of 4 lambda GST (50%), 2 out of 6 phi GST (33%) formed one cluster each (Fig. S1). This indicates a major contribution of tandem clustering towards the expansion of gene members in each class. Interestingly, different GST proteins from the same genomic cluster showed distinct variation in their enzymatic activity, catalytic efficiency, substrate affinity, and specificities^{34,35}. The reasons behind these extensive tandem duplication event in *GST* gene family with the diverse kinetic property are still unknown.

Most *CaGST* genes shared a similar exon-intron structure within the same phylogenetic group (Fig. 1), indicating that the evolution of GST domains may be closely related to the diversification of gene structure. Gene structure analysis showed variation in the presence of exon number in various *CaGST* genes. The number of exons varied from one to a maximum of ten in *CaGST* genes, where the majority of them were single exonic (Fig. 1). Likewise *CaGSTs*, the presence of nine introns was reported previously for tomato and potato *GST* genes^{28,36}, while a maximum number of 16 and 14 were reported in *Vigna radiate* and *Chinese cabbage*, respectively^{49,50}. Previous reports suggested that introns influence and enhance the expression of a gene in a eukaryotic organism, which has been experimentally validated with heterologous gene expression in *Arabidopsis thaliana*^{51,52}. The clustering of most intronless or intron-containing genes into the same group (Fig. 1), suggested that this may be a unique feature of the evolution of pepper *GST* gene family. Due to the less selection pressure, intronic sequences possess a higher rate of gain and loss as compared with exonic sequences⁵³. In two of the five segmentally duplicated rice gene pairs showed intron gain event⁵³. An interesting study of 612 pairs of paralogs from seven representative gene families and 300 pairs of orthologs from different species, concluded that orthologs are more conserved with significantly fewer structural changes as compared with paralogs of similar evolutionary time⁵⁴.

Phylogenetic analysis revealed that *CaGSTs* were closely allied to the same class of GSTs of four other plant species- tomato, potato, *Arabidopsis*, and rice (Fig. 2). This suggested that the evolution and divergence of each *GST* classes have happened before the split of monocot and dicot. However, *CaGST* members are more closely related to the tomato and potato as compared with *Arabidopsis* and rice counterparts, reflecting the fact that pepper, tomato, and potato belong to the same superfamily of eudicots and diverged more recently from a common ancestor³⁷. Orthologous gene-based phylogenetic analysis of grape, papaya, pepper, tomato, potato, and *Arabidopsis* genome concluded that pepper has been separated from tomato and potato ~36 Mya ago⁵⁵. Thus, pepper GSTs were compared with that of tomato to elucidate the lineage-specific expansion and genome diversity among these two species (Fig. 3). Thirteen clades contained only *CaGST* and seven clades contained only tomato *GST* (Figs S3–S10), indicating that gene loss might have occurred in these clades. The number of clades indicated that there were at least 85 ancestral *GST* genes before the *Capsicum*-tomato split (Fig. 3).

Extensive expression analysis revealed the developmental stage and tissue specific transcript alteration of *CaGST* genes (Fig. 4). Similarly, nine genes were expressed ubiquitously, ten showed root specificity and two

expressed in leaves out of 37 tau *GST* of *Sorghum bicolor*⁵⁶. Six sunflower *GST* genes were mainly expressed in leaves, four in seeds and two each in flowers and roots out of a total of 14 identified members³². In addition to the developmental alteration, expression of *GST* genes also showed significant variation in response to adverse environmental conditions (Fig. 5). Plants exposure with various abiotic stresses such as cold, heat, drought and salt resulted in the enhancement ROS level and thus, caused oxidative stress⁵⁷. Tau *GSTs* could protect the cell by enhancing the detoxification of herbicides such as atrazine, metolachlor, flurodifen, and thiocarbamates⁵⁸ and maintaining higher *GST* activity under salinity and oxidative stress¹⁹. *JrGSTT1* enhanced chilling stress tolerance of *Juglans regia* by protecting oxidative enzymes, scavenging ROS, and elevating the expression of several stress-related genes⁵⁹. Upregulation of most of the *SlGST* transcripts in response to multiple abiotic stresses could be directly harmonized with the enhancement of total *GST* activity under similar conditions³⁶. Similarly, upregulation of 6 *HaGSTs* (*HaGSTU1*, *U2*, *U5*, *U6*, *F2*, and *Z1*) expression showed a significant positive correlation with the changes in their respective *GST* activity³². The present study also found the positive relationship between the *GST* transcript upregulation and enzyme activity in pepper against four abiotic stresses and oxidative injury (Fig. 6). Heterologous expression of one of the sweet orange tau *GST* or *JrGSTT1* in tobacco enhanced tolerance against herbicide, salinity, chilling and drought stresses^{56,59}. Several stress-responsive motifs were identified in the putative promoter region of *CaGST* genes (Fig. 7). Cis-acting elements played important to control/regulate the expression of genes and thus, modulating plant response against stress and developmental changes⁶⁰. Two commonly present abiotic stress-inducible cis-acting elements, dehydration-responsive element (DRE) and ABRE are found to be interdependent in the ABA-responsive expression of *Atrd29A* gene^{60,61}. Similarly, the presence of two putative low-temperature responsive cis-elements in the 5'-proximal region of *BN115* gene was found to be indispensable for its cold-induced expression in *Brassica napus*⁶¹. The highly abiotic stress responsive genes of the present study such as *CaGSTF3*, *CaGSTL2*, *CaGSTL3*, *CaEF1Bγ1*, and *CaGSTU44*, showed the presence of a variable amount of HSE, MBS, TC-rich repeat, and LTR motifs (Figs 6, 7). All these common motifs might work synergistically depending on the type of stress to induce the expression of a maximum number of stress-responsive genes.

Taken together, our results provide a comprehensive analysis of the *GST* gene family in pepper. Sequence and phylogenetic analysis of *GST* from five different plant species revealed the evolutionary conservation of each class of *GST* proteins. A close relationship between the expression and activity of *GST* with plant stress tolerance established *GST* as a major stress biomarker for the plant.

Materials and Methods

Sequence retrieval, analysis, and annotation. To retrieve all *GST* members in pepper, previously reported rice and *Arabidopsis* *GST* protein sequences from each class was taken as a query in the blastp search with the default parameters (e-value 10^{-10}) against the Pepper Genome Protein Database (release 2.0) (<http://peppersequence.genomics.cn>; <http://public.genomics.org.cn/BGI/pepper/>) and Sol Genomics Network (SGN) (<https://www.solgenomics.net/>). Sequences were then analyzed through the NCBI conserved domain database (<https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>) to identify the individual class of each identified members. Detailed information about the locus name, CDS coordinate (5'-3'), length of the transcript and peptide were collected from the Pepper Genome Database (<http://pepperhub.hzau.edu.cn/pegnm/>) and Sol Genomics Network (SGN). The ProtParam tool (<http://www.expasy.org/tools/protparam.html>) was used to calculate various physiochemical properties like molecular weight and theoretical isoelectric point (pI) of the identified proteins. The secondary structure of *GST* proteins was predicted using the SOPMA (Self-Optimized Prediction Method with Alignment, https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html). Moreover, N-glycosylation sites were identified using NetNGlyc 1.0 server (<http://www.cbs.dtu.dk/services/NetNGlyc/>). Furthermore, subcellular localization was predicted using the CELLO version 2.5 (<http://cello.life.nctu.edu.tw/>), pSORT (<http://www.genscript.com/wolf-psort.html>) and ChloroP server (<http://www.cbs.dtu.dk/services/ChloroP/>). Pfam (<http://pfam.xfam.org/>) was used to assess the conserved *GST_N* and *GST_C* domains in all the identified members. Domains were graphically depicted using the software 'Illustrator for Biological Sequences' (version 1.0.2). The conserved motifs of *CaGST* proteins were identified using the online MEME server (<http://meme-suite.org/tools/meme>).

Chromosome localization, gene structure and duplications. The physical location of *CaGST* genes was collected from the Pepper Genome Database and the positions of these *CaGST* genes were plotted to thirteen *C. annuum* chromosomes. Exon-intron structure of *CaGST* genes was obtained by gene structure display server (<http://gsds.cbi.pku.edu.cn/>) using the corresponding genomic and CDS sequence. Duplication events were predicted by blastp search (e-value 10^{-10}) with $\geq 80\%$ sequence identity in the Pepper Genome Database⁶². Two or more homologous genes within 100 Kb region on the same chromosome were considered as tandemly duplicated (TD)⁶², while those located beyond 100 kb region were designated as segmental duplication (SD). Synonymous rate (d_s), non-synonymous rate (d_n), and evolutionary constraint (d_n/d_s) between the duplicated *CaGST* gene pairs were analyzed using the PAL2NAL online tool (<http://www.bork.embl.de/pal2nal/>). Divergence time (T) of each duplicated gene pairs was calculated using the formula: $(T = d_s/2\lambda)$, where λ is considered as a fixed rate of 1.5×10^{-8} substitutions per site per year for dicotyledonous plants⁶³.

Phylogenetic analysis. To analyze the evolutionary relationship, *GST* protein sequences from five different species- *Capsicum annuum*, *Arabidopsis thaliana*, *Oryza sativa*, *Solanum lycopersicum* and *Solanum tuberosum* were obtained from respective genome database and class information was gathered from the previously published literatures^{14,15,28,36}. The phylogenetic tree was constructed using the default parameters of the maximum likelihood method in MEGA 7 software with 1000 bootstrap replicates. The James Taylor Thornton (JTT) substitution model was set with the site coverage cutoff of 95%. To investigate the lineage-specific expansion of *GST*

genes in *Capsicum annuum* and *Solanum lycopersicum*, ten class-specific phylogenetic trees were constructed using MEGA 7 software according to the above mentioned procedure.

Pepper RNA-seq data analysis. Illumina RNA sequenced gene expression profiling data of each *CaGST* genes at different tissues and developmental stages and in response to five abiotic stresses was obtained from the pepper hub transcriptome database (<http://pepperhub.hzau.edu.cn/petdb/>). For abiotic stress treatments, datasets were obtained from the database for 0, 1, 1.5, 3, 6, 12 and 24 h of cold, heat, drought, salinity and osmotic stress treatments for leaf and root tissue, and the relative fold change of expression was calculated based on their control value (0 h). Heat maps with hierarchical clustering were performed using the default parameters of TIGR Multiple Experiment Viewer (MEV) 4.9 software package with the Manhattan correlation⁶⁴.

Plant materials, stress treatment, and total GST activity. *C. annuum* (BARI Morich-3) were germinated and grown as described previously by Guo *et al.*⁴⁰. Fifteen days old plants were soaked with normal water (for experimental control), or 200 mM NaCl (for salinity), or 5% mannitol (for drought stress), or 5 mM H₂O₂ (for oxidative stress), or normal water at 40°C (for heat stress) or 4°C (for cold stress). Shoot samples were collected at 0, 6, 12, and 24 hours post-stress induction. Total protein was extracted using the ice-cold extraction buffer containing 100 mM potassium phosphate buffer, pH 7.0, 50% glycerol, 16 mM MgSO₄ and 1 mM PMSF⁵⁷. After quantification of the protein by Bradford method⁶⁵, GST activity was measured by its ability to conjugate the reduced glutathione and 1-chloro-2,4-dinitrobenzene at 344 nm³⁶. Activity was calculated using the extinction coefficient of the product formed (9.6 mM⁻¹ cm⁻¹) and was expressed as nmoles of CDNB conjugated/minute/mg of total protein.

Analysis of putative promoter sequence. The 1000 bp 5' upstream sequences from the transcription start site of all the *CaGST* genes were acquired from the pepper hub (<http://pepperhub.hzau.edu.cn/pegnm/>) and then analyzed individually on the PlantCARE program (<http://bioinformatics.psb.ugent.be/webtools/plantcare/html/>) with default parameters for the identification of the important stress and hormone responsive cis-regulatory elements⁶⁶.

Data Availability

The authors declare that all the data and plant materials will be available without restrictions.

References

- Hayes, J. D. & McLellan, L. I. Glutathione and glutathione-dependent enzymes represent a co-ordinately regulated defence against oxidative stress. *Free Radic Res* **31**, 273–300 (1999).
- Edwards, R., Dixon, D. P., Cummins, I., Brazier-Hicks, M. & Skipsey, M. In *Organic xenobiotics and plants* 125–148 (Springer, 2011).
- Prova, C. The plant glutathione transferase gene family: genomic structure, functions, expression and evolution. *Physiologia plantarum* **119**, 469–479 (2003).
- Marrs, K. A. The functions and regulation of glutathione S-transferases in plants. *Annual review of plant biology* **47**, 127–158 (1996).
- Loguercio, C. *et al.* Alpha-glutathione transferases in HCV-related chronic hepatitis: a new predictive index of response to interferon therapy? *Journal of hepatology* **28**, 390–395 (1998).
- Basantani, M. & Srivastava, A. J. B. Plant glutathione transferases—a decade falls short. **85**, 443–456 (2007).
- Mohsenzadeh, S. *et al.* Plant glutathione S-transferase classification, structure and evolution. *African Journal of Biotechnology* **10**, 8160–8165 (2011).
- Lallement, P.-A., Brouwer, B., Keech, O., Hecker, A. & Rouhier, N. The still mysterious roles of cysteine-containing glutathione transferases in plants. *Frontiers in pharmacology* **5**, 192 (2014).
- Kumar, S. & Trivedi, P. Glutathione S-transferases: An Update for the Role in Arsenic Detoxification Mechanism. *Frontiers in Plant Science* **9**, 751 (2018).
- Sheehan, D., Meade, G. & Foley, V. M. Structure, function and evolution of glutathione transferases: implications for classification of non-mammalian members of an ancient enzyme superfamily. *Biochemical Journal* **360**, 1–16 (2001).
- Dixon, D. P., Davis, B. G. & Edwards, R. Functional divergence in the glutathione transferase superfamily in plants identification of two classes with putative functions in redox homeostasis in *Arabidopsis thaliana*. *Journal of Biological Chemistry* **277**, 30859–30869 (2002).
- Moons, A. Regulatory and functional interactions of plant growth regulators and plant glutathione S-transferases (GSTs). *Vitamins & Hormones* **72**, 155–202 (2005).
- Thuillier, A. *et al.* Atypical features of a Ure2p glutathione transferase from *Phanerochaete chrysosporium*. *FEBS letters* **587**, 2125–2130 (2013).
- Jain, M., Ghanashyam, C. & Bhattacharjee, A. Comprehensive expression analysis suggests overlapping and specific roles of rice glutathione S-transferase genes during development and stress responses. *BMC genomics* **11**, 73 (2010).
- Sappl, P. G. *et al.* The *Arabidopsis* glutathione transferase gene family displays complex stress regulation and co-silencing multiple genes results in altered metabolic sensitivity to oxidative stress. *The Plant Journal* **58**, 53–68 (2009).
- Skopelitou, K. *et al.* Characterization and functional analysis of a recombinant tau class glutathione transferase GmGSTU2-2 from *Glycine max*. *International journal of biological macromolecules* **94**, 802–812 (2017).
- Nutricati, E., Miceli, A., Blando, F. & De Bellis, L. Characterization of two *Arabidopsis thaliana* glutathione S-transferases. *Plant cell reports* **25**, 997–1005 (2006).
- Gallé, Á. *et al.* Glutathione transferase activity and expression patterns during grain filling in flag leaves of wheat genotypes differing in drought tolerance: response to water deficit. *Journal of plant physiology* **166**, 1878–1891 (2009).
- Sharma, R., Sahoo, A., Devendran, R. & Jain, M. Over-expression of a rice tau class glutathione s-transferase gene improves tolerance to salinity and oxidative stresses in *Arabidopsis*. *PLoS one* **9**, e92900 (2014).
- Kissoudis, C. *et al.* Stress-inducible GmGSTU4 shapes transgenic tobacco plants metabolome towards increased salinity tolerance. *Acta physiologiae plantarum* **37**, 102 (2015).
- Cicero, L. L., Madesis, P., Tsafaris, A. & Piero, A. R. L. Tobacco plants over-expressing the sweet orange tau glutathione transferases (CsGSTUs) acquire tolerance to the diphenyl ether herbicide fluorodifen and to salt and drought stresses. *Phytochemistry* **116**, 69–77 (2015).
- Rezaei, M. K., Shobbar, Z.-S., Shahbazi, M., Abedini, R. & Zare, S. Glutathione S-transferase (GST) family in barley: identification of members, enzyme activity, and gene expression pattern. *Journal of plant physiology* **170**, 1277–1284 (2013).

23. Vijayakumar, H. *et al.* Glutathione transferases superfamily: cold-inducible expression of distinct GST genes in Brassica oleracea. *International journal of molecular sciences* **17**, 1211 (2016).
24. Dong, Y. *et al.* Glutathione S-transferase gene family in Gossypium raimondii and G. arboreum: comparative genomic study and their expression under salt stress. *Frontiers in plant science* **7**, 139 (2016).
25. Yang, Q., Liu, Y.-J. & Zeng, Q.-Y. Biochemical functions of the glutathione transferase supergene family of Larix kaempferi. *Plant physiology and biochemistry* **77**, 99–107 (2014).
26. McGonigle, B., Keeler, S. J., Lau, S.-M. C., Koeppe, M. K. & O'Keefe, D. P. A genomics approach to the comprehensive analysis of the glutathione S-transferase gene family in soybean and maize. *Plant Physiology* **124**, 1105–1120 (2000).
27. Wang, L., Qian, M., Wang, R., Wang, L. & Zhang, S. J. P. G. R. Characterization of the glutathione S-transferase (GST) gene family in Pyrus bretschneideri and their expression pattern upon superficial scald development. **86**, 211–222, <https://doi.org/10.1007/s10725-018-0422-4> (2018).
28. Islam, M. S., Choudhury, M., Majlish, A.-N. K., Islam, T. & Ghosh, A. Comprehensive genome-wide analysis of Glutathione S-transferase gene family in potato (Solanum tuberosum L.) and their expression profiling in various anatomical tissues and perturbation conditions. *Gene* **639**, 149–162 (2018).
29. Liu, Y.-J., Han, X.-M., Ren, L.-L., Yang, H.-L. & Zeng, Q.-Y. Functional divergence of the glutathione S-transferase supergene family in Physcomitrella patens reveals complex patterns of large gene family evolution in land plants. *Plant physiology* **161**, 773–786, <https://doi.org/10.1104/pp.112.205815> (2013).
30. Lan, T. *et al.* Extensive functional diversification of the Populus glutathione S-transferase supergene family. *The Plant Cell* **21**, 3749–3766 (2009).
31. Liu, H.-J. *et al.* Divergence in enzymatic activities in the soybean GST supergene family provides new insight into the evolutionary dynamics of whole-genome duplicates. *Molecular biology and evolution* **32**, 2844–2859 (2015).
32. Ma, L. *et al.* Molecular cloning, identification of GSTs family in sunflower and their regulatory roles in biotic and abiotic stress. *World Journal of Microbiology and Biotechnology* **34**, 109, <https://doi.org/10.1007/s11274-018-2481-0> (2018).
33. Licciardello, C. *et al.* Characterization of the glutathione S-transferase gene family through ESTs and expression analyses within common and pigmented cultivars of Citrus sinensis (L.) Osbeck. *BMC plant biology* **14**, 39 (2014).
34. Han, X. M., Yang, Z. L., Liu, Y. J., Yang, H. L. & Zeng, Q. Y. Genome-wide profiling of expression and biochemical functions of the Medicago glutathione S-transferase gene family. *Plant Physiol Biochem* **126**, 126–133, <https://doi.org/10.1016/j.plaphy.2018.03.004> (2018).
35. He, G. *et al.* Genome-Wide Analysis of the Glutathione S-Transferase Gene Family in Capsella rubella: Identification, Expression, and Biochemical Functions. *Front Plant Sci* **7**, 1325, <https://doi.org/10.3389/fpls.2016.01325> (2016).
36. Islam, S., Rahman, I. A., Islam, T. & Ghosh, A. Genome-wide identification and expression analysis of glutathione S-transferase gene family in tomato: Gaining an insight to their physiological and stress-specific roles. *PLoS One* **12**, e0187504, <https://doi.org/10.1371/journal.pone.0187504> (2017).
37. Kim, S. *et al.* Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nature genetics* **46**, 270 (2014).
38. Goldberg, N. P. Chile pepper disorders caused by environmental stress. (Cooperative Extension Service, College of Agriculture and Home Economics ..., 2004).
39. Jaimez, R., Vielma, O., Rada, F. & García-Núñez, C. Effects of water deficit on the dynamics of flowering and fruit production in Capsicum chinense Jacq in a tropical semi-arid region of Venezuela. *Journal of Agronomy and Crop Science* **185**, 113–119 (2000).
40. Guo, M. *et al.* Genome-wide analysis of the CaHsp20 gene family in pepper: comprehensive sequence and expression profile analysis under heat stress. *Front Plant Sci* **6**, 806, <https://doi.org/10.3389/fpls.2015.00806> (2015).
41. Khan, A. *et al.* Genome-wide analysis of dirigent gene family in pepper (Capsicum annuum L.) and characterization of CaDIR7 in biotic and abiotic stresses. *Scientific reports* **8**, 5500 (2018).
42. Zhang, H.-X. *et al.* Genome-wide identification and analysis of the SBP-box family genes under Phytophthora capsici stress in pepper (Capsicum annuum L.). *Frontiers in plant science* **7**, 504 (2016).
43. Liu, Y.-J., Han, X.-M., Ren, L.-L., Yang, H.-L. & Zeng, Q.-Y. Functional divergence of the GST supergene family in Physcomitrella patens reveals complex patterns of large gene family evolution in land plants. *Plant physiology*, pp. 112.205815 (2012).
44. Wang, L., Qian, M., Wang, R., Wang, L. & Zhang, S. Characterization of the glutathione S-transferase (GST) gene family in Pyrus bretschneideri and their expression pattern upon superficial scald development. *Plant Growth Regulation*, 1–12 (2018).
45. Kumar, S. *et al.* Differential expression of rice lambda class GST gene family members during plant growth, development, and in response to stress conditions. *Plant molecular biology reporter* **31**, 569–580 (2013).
46. Loyall, L., Uchida, K., Braun, S., Furuya, M. & Frohnmeyer, H. Glutathione and a UV light-induced glutathione S-transferase are involved in signaling to chalcone synthase in cell cultures. *The Plant Cell* **12**, 1939–1950 (2000).
47. Benekos, K. *et al.* Overexpression of a specific soybean GmGSTU4 isoenzyme improves diphenyl ether and chloroacetanilide herbicide tolerance of transgenic tobacco plants. *Journal of biotechnology* **150**, 195–201 (2010).
48. Cummins, I. *et al.* Key role for a glutathione transferase in multiple-herbicide resistance in grass weeds. *Proceedings of the National Academy of Sciences*, 201221179 (2013).
49. Vaish, S. *et al.* In silico genome-wide identification and characterization of the glutathione S-transferase gene family in Vigna radiata. *Genome* **61**, 311–322 (2018).
50. Du, J. *et al.* Genome-wide identification and expression analysis of the glutathione S-transferase (GST) family under different developmental tissues and abiotic stresses in Chinese cabbage (Brassica rapa ssp. pekinensis). *PeerJ PrePrints* (2018).
51. Emami, S., Arumainayagam, D., Korf, I. & Rose, A. B. The effects of a stimulating intron on the expression of heterologous genes in Arabidopsis thaliana. *Plant biotechnology journal* **11**, 555–563 (2013).
52. Shaul, O. How introns enhance gene expression. *The international journal of biochemistry & cell biology* **91**, 145–155 (2017).
53. Lin, H., Zhu, W., Silva, J. C., Gu, X. & Buell, C. R. Intron gain and loss in segmentally duplicated genes in rice. *Genome Biol* **7**, R41, <https://doi.org/10.1186/gb-2006-7-5-r41> (2006).
54. Xu, G., Guo, C., Shan, H. & Kong, H. Divergence of duplicate genes in exon-intron structure. *Proc Natl Acad Sci USA* **109**, 1187–1192, <https://doi.org/10.1073/pnas.1109047109> (2012).
55. Qin, C. *et al.* Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. *Proceedings of the National Academy of Sciences* **111**, 5135–5140 (2014).
56. Chi, Y. *et al.* Expansion mechanisms and functional divergence of the glutathione S-transferase family in sorghum and other higher plants. *DNA research* **18**, 1–16 (2010).
57. Ghosh, A., Pareek, A., Sopory, S. K. & Singla-Pareek, S. L. A glutathione responsive rice glyoxalase II, OsGLYII-2, functions in salinity adaptation by maintaining better photosynthesis efficiency and anti-oxidant pool. *Plant J* **80**, 93–105, <https://doi.org/10.1111/tpj.12621> (2014).
58. Kawahigashi, H. Transgenic plants for phytoremediation of herbicides. *Curr. Opin. Biotechnol.* **20**, 225–230 (2009).
59. Yang, G. *et al.* In planta characterization of a tau class glutathione S-transferase gene from Juglans regia (JrGSTTau1) involved in chilling tolerance. *Plant Cell Rep* **35**, 681–692, <https://doi.org/10.1007/s00299-015-1912-8> (2016).
60. Narusaka, Y. *et al.* Interaction between two cis-acting elements, ABRE and DRE, in ABA-dependent expression of Arabidopsis rd29A gene in response to dehydration and high-salinity stresses. *Plant J* **34**, 137–148 (2003).

61. Jiang, C., Iu, B. & Singh, J. Requirement of a CCGAC cis-acting element for cold induction of the BN115 gene from winter Brassica napus. *Plant Mol Biol* **30**, 679–684 (1996).
62. Kayum, M. A. *et al.* Genome-wide characterization and expression profiling of PDI family gene reveals function as abiotic and biotic stress tolerance in Chinese cabbage (*Brassica rapa* ssp. *pekinensis*). *BMC genomics* **18**, 885 (2017).
63. Koch, M. A., Haubold, B. & Mitchell-Olds, T. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Molecular biology and evolution* **17**, 1483–1498 (2000).
64. Johnson, S. C. Hierarchical clustering schemes. *Psychometrika* **32**, 241–254 (1967).
65. Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* **72**, 248–254 (1976).
66. Lescot, M. *et al.* PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res* **30**, 325–327 (2002).

Acknowledgements

SI, SDS, and AG acknowledge the logistic support and laboratory facilities of the Department of Biochemistry and Molecular Biology, Shahjalal University of Science and Technology, Sylhet, Bangladesh. ZSJ, SA, and TI acknowledge Plant Breeding and Biotechnology Laboratory, Department of Botany, University of Dhaka for providing laboratory facilities. AG acknowledges the Humboldt Research Fellowship for Postdoctoral Researchers.

Author Contributions

A.G. conceived the idea. A.G. and T.I. designed the experiments. S.I., S.D.S., Z.S.J., and S.A. performed the experiments. A.G., T.I., and S.I. analyzed data. S.I. and S.D.S. wrote the initial draft of the manuscript. All authors read the manuscript and approved the final version.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-019-45320-x>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019