Genome-wide analysis of sulfur-encoding biosynthetic genes in rice (Oryza sativa L.) with Arabidopsis as the sulfur-dependent model plant

Muhammad-Redha Abdullah-Zawawi1,2, Nisha Govender1*, Nor Azlan Nor Muhammad3, Norfarhan Mohd-Assaad3, Zamri Zainal1,3 & Zeti-Azura Mohamed-Hussein1,3

Sulfur is an essential element required for plant growth and development, physiological processes and stress responses. Sulfur-encoding biosynthetic genes are involved in the primary sulfur assimilation pathway, regulating various mechanisms at the gene, cellular and system levels, and in the biosynthesis of sulfur-containing compounds (SCCs). In this study, the SCC-encoding biosynthetic genes in rice were identified using a sulfur-dependent model plant, the Arabidopsis. A total of 139 AtSCC from Arabidopsis were used as reference sequences in search of putative rice SCCs. At similarity index > 30%, the similarity search against Arabidopsis SCC query sequences identified 665 putative OsSCC genes in rice. The gene synteny analysis showed a total of 477 syntenic gene pairs comprised of 89 AtSCC and 265 OsSCC biosynthetic genes in Arabidopsis and rice, respectively. Phylogenetic tree of the collated (AtSCCs and OsSCCs) SCC-encoding biosynthetic genes were divided into 11 different clades of various sizes comprised of branches of subclades. In clade 1, nearing equal representation of OsSCC and AtSCC biosynthetic genes imply the most ancestral lineage. A total of 25 candidate Arabidopsis SCC homologs were identified in rice. The gene ontology enrichment analysis showed that the rice-Arabidopsis SCC homologs were significantly enriched in the following terms at false discovery rate (FDR) < 0.05: (i) biological process; sulfur compound metabolic process and organic acid metabolic processes, (ii) molecular function; oxidoreductase activity, acting on paired donors with incorporation or reduction of molecular oxygen and (iii) KEGG pathway; metabolic pathways and biosynthesis of secondary metabolites. At less than five duplicated blocks of separation, no tandem duplications were observed among the SCC biosynthetic genes distributed in rice chromosomes. The comprehensive rice SCC gene description entailing syntenic events with Arabidopsis, motif distribution and chromosomal mapping of the present findings offer a foundation for rice SCC gene functional studies and advanced strategic rice breeding.

Sulfur (S) is an important macronutrient for plant growth and development, immunity, and stress mitigation. In sulfur-deficient soils, plants invoke stress resistance and xenobiotic detoxification1,2. Plant S assimilation is translated into sulfur-containing compounds (SCCs), a class of important secondary metabolites. Plants utilize freely available sulfate in the soil to synthesize SCCs for growth and functional metabolisms. The primary S assimilation pathway integrates carbon, nitrogen, and S for the synthesis of various SCCs such as glutathione, S-adenosylmethionine, S-methylmethionine, sulfoquinovosyldiaclylglycerol, ferredoxin and thiol-group containing plant defensins3. Both the thiolredoxins and glutathionones are redox modulators with detoxifying abilities. In...
view of the ecological perspective, various vital biological functions which include oxidative stress mitigation, heavy-metal detoxification \(^2\) and plant defense responses against biotic factors \(^3\) are regulated by SCCs.

Rice (Oryza sativa L.) is the second most preferred food crop consumed worldwide, after wheat. Cultivated in over 144 countries around the world, rice feeds half the world population (3 billion people) and warrants global food security \(^4\). It is predicted that rice production exceeding 800 million tonnes is required to meet the calorie demand of the expected world population in 2025 \(^4\). With climate change in the chart of global issues, abiotic stresses are strongly impacting rice productivity. Major limiting factors in the rice production system includes drought, heat, cold and salinity. In others, waterlogged paddy soils inherent toxic elements such as Cd, As and Fe. Rapid response to stressors regulates stress mitigation responses which include transmembrane transport, glutathione metabolism, signal transduction, and redox control \(^5\). In rice, S-associated genes, metabolites and proteins have shown involvement in abiotic stress responses and mitigation. For example, in Cd and As co-contaminated soils, the glutathione metabolism-related genes (Oso1g0537700 and Oso1g0539000) were significantly up-regulated relative to the control conditions. During rice drought stress response, the glutathione S-transferase activities were significant increased \(^6\). In another study, glutathione peroxidases and thiol-based antioxidant enzymes regulated the ABA-independent osmotic stress signalling in rice \(^7\). Although the role of SCC-encoding genes and SCCs in rice stress response have been documented by numerous studies, little is known about the SCC gene distribution and pattern, and putative functions at the rice genome scale. The SCC genome-level information is important to shed new information and knowledge in innovative rice breeding strategies.

Plant SCC distribution varies greatly with species. In the Brassicaceae family, more than two hundred different types of glucosinolates (GLSs) with potent roles in defense responses have been reported \(^8\), \(^9\). The GLS-myrosinase defense system gets activated during a pathogen attack to form unstable aglycone intermediates. Thereafter, a range of toxic volatile compounds (isothiocyanates, nitriles, and thiocyanates) is produced during hydrolysis for deterrence against the invading pathogen/pests \(^8\), \(^9\). In others, camalexin, an indole-type phytoalexin SCC is produced for adaptivity against abiotic stress and pathogen attack, alike \(^10\). Camalexin derived from tryptophan is converted to indole-3-acetaldoxime, which later switches into indole-3-acetonitrile upon dehydration \(^15\). Arabidopsis (Brassicaceae) and rice from the grass family (Poaceae) are S-dependent families. With about 10–30% of S expressed in the plant tissues, the first is ranked as the most S-dependent family, \(^11\), \(^16\)–\(^19\).

In this study, the SCC-encoding biosynthetic genes in rice are identified and characterized using Arabidopsis as the reference genome model of an S-dependent plant family. The Arabidopsis genome is an excellent reference for the identification of S-encoding biosynthetic genes in rice. There is a burst of SCC-related functional experimental and databases \(^20\), \(^21\) extensively reported in Arabidopsis; low-affinity sulphate transporters; \(^22\) S dioxygenase activity in ETHE1 knockout mutant; \(^23\) S deficiency responsive genes; \(^24\) Arabidopsis S metabolome; \(^25\) S-containing secondary metabolites from Arabidopsis. The syntenity and similarity of the Arabidopsis-rice SCC homologous sequences are visualized and the enrichment analysis along a cross-comparison of the corresponding motif sequences is provided to gain information on the extent of similarities. The findings extent to compare and capture the Arabidopsis-rice evolutionary relationship, predict the ecological functions of SCC genes in rice and provide the genetic basis for stress mitigation and defense response enhancement in rice breeding.

Materials and methods

**Arabidopsis and rice genome sequences.** Arabidopsis thaliana and O. sativa genome sequences and genome annotations were obtained from the Phytozome v13.0 database (https://phytozome-next.jgi.doe.gov/) \(^26\), Arabidopsis Information Resource (TAIR) v10.0 (https://www.arabidopsis.org) \(^27\) and O. sativa Genome Annotation Project Database (RGAP) v7.0 (http://rice.uga.edu/) \(^28\). The Arabidopsis genome was set as reference sequence against the rice (query) sequences.

**Sulfur-containing compound (SCC)-encoding biosynthetic gene mining.** The SCC-encoding genes in Arabidopsis (AtSCC) were mined from AraCyc version 14.0 (https://pmn.plantcyc.org/) \(^29\) using the following keywords: (i) glucosinolate, and (ii) camalexin. The AtSCC biosynthetic protein sequences were designated as query for the identification of corresponding homologs (O. sativa SCC biosynthetic genes) in the rice genome using the BLAST program (http://blast.ncbi.nlm.nih.gov) \(^30\). Reciprocal searching was applied using BLASTP default parameters: e-value = 1e-10 and sequence similarity > 30%. The gene positions were determined by parsing the genome annotation file and the BLAST output. The genomic feature information (General Feature Format) file was concatenated as the input data for subsequent analysis.

**Syntenity analysis.** The Multiple Collinearity Scan Toolkit X software (MCScanX) was employed for the identification of collinear blocks of homologous sequences and multiple alignment of collinear blocks to the chromosomes. Input files were executed by the MCScan function and the expected number of occurrences (E) of the collinear blocks was calculated \(^31\). The following default parameters were applied: E-value cut-off = 1e-05 and match_size = 5. The collinear blocks of interspecies were labelled as AtSCC and OsSCC, denoting A. thaliana and O. sativa, respectively. All rice-Arabidopsis collinear blocks of gene pairs (two interspecies chromosomal positions) were identified and visualized using Rcircos software \(^32\).

**Multiple sequence alignment and phylogenetic analysis.** A multiple sequence alignment of the rice-Arabidopsis SCC-encoding biosynthetic genes was performed using the Multiple Sequence Comparison by Log-Expectation (MUSCLE) (https://www.ebi.ac.uk/Tools/msa/muscle/) with the following settings: gap open penalty = -2.9, gap extension = 0, and hydrophobicity multiplier = 1.2. Phylogenetic analysis was performed using Molecular Evolutionary Genetics Analysis (MEGA) v7.0 (http://megasoftware.net) \(^33\). The maximum-likelihood (ML) by Tamura-Nei substitution model and phylogeny test using 1000 replicates of the bootstrap
Motifs search distributions, gene structure analysis and chromosomal mapping. The ex- ontron architecture of AtSCC and OsSCC biosynthetic genes was visualized using the Gene Structure Display Server 2.0. Conserved motifs were identified using the Multiple Expectation Maximization (MEME) v4.11.3 (http://meme-suite.org) tool with the following parameters: the number of motifs = 10, motif site distributions mode = 0/1 occurrence per sequence (zoops)37. The consensus motif sequences were annotated using Database of protein domains, families and functional sites (PROSITE) (http://prosite.expasy.org)38, Pfam, database for protein families v35.0 (http://pfam.xfam.org/)39 and Conserved Domain Database v3.19 (CDD) (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml)40. The chromosoma gene loci were mapped using the Chromosome Map Tools available in TAIR (https://www.arabidopsis.orgjsp/ChromosomeMap/tool.jsp)41 and Oryzabase (http://viewer.shigen.info/oryzaw/maptool/MapTool.do)42 of A. thaliana and O. sativa genes, respectively. Genes separated by less than five genetic loci within 5 to 100 kb were scored as tandem duplications.

Gene ontology (GO) enrichment and pathway. Functional enrichment analysis of the SCC-encoding biosynthetic genes (AtSCC and OsSCC) was performed using ShinyGO v0.75 (http://bioinformatics.sdstate.edu/go75/) with p-value cut-off set at false discovery rate (FDR) = 0.05: (i) Gene ontology classification43 and (ii) KEGG pathway enrichment44. The A. thaliana and O. sativa Japonica genomes were set as reference datasets. The 20 top-most significantly enriched AtSCC and OsSCC genes were identified using the Venn webserver (https://bioinformatics.psbgent.be/webtools/Venn/).

Results

Identification of putative OsSCC biosynthetic genes using synteny analysis. A total of 139 AtSCC biosynthetic genes were obtained from a rapid search performed with the following descriptions: (i) glucosinolate activation (herbivore attack and intact plant cell) pathways, (ii) aliphatic glucosinolate (derived from homomethionine, dihomomethionine, trihomomethionine, hexahomomethionine, pentahomomethionine, and tetrahomomethionine), (iii) indolic glucosinolate (tryptophan derivative), (iv) aromatic glucosinolate (phenylalanine derivative) and (v) camalexin. The sequence homology search identified a total of 838 SCC biosynthetic genes in O. sativa. A total of 173 sequences were discarded due to low sequence similarity (< 30%) and the remaining 665 candidates were subjected to synteny analysis. Under various combinations, a total of 477 syntenic gene pairs with 89 AtSCC and 265 OsSCC biosynthetic genes were identified (Supplementary 1).

The syntenic gene pairs were randomly distributed across the chromosomes with sizes, as annotated by the gene number (GN). In rice, the syntenic GN distribution were as following: OsChr1; syntenic GN = 45, OsChr2; syntenic GN = 32, OsChr3; syntenic GN = 28, OsChr11; syntenic GN = 23, OsChr4, OsChr 7 and OsChr 9; syntenic GN = 20, OsChr12; syntenic GN = 18, OsChr8; syntenic GN = 16, and OsChr5; GN = 13. In A. thaliana, the highest number of syntenic genes were distributed in AtChr1 (syntenic GN = 31), followed by AtChr5 (syntenic GN = 19), AtChr3 (syntenic GN = 15), AtChr2 (syntenic GN = 13) and AtChr4 (syntenic GN = 11) (Fig. 1).

The distribution of syntenic gene pairs (SPs) was higher in AtChr1 (SPs = 267) and AtChr5 (SPs = 113) in comparison to AtChr2 (SP = 37) and AtChr3 (SP = 33). Overall, a total of 41 AtSCC and 25 OsSCC biosynthetic genes were linked with at least four synteny blocks. Ten OsSCC biosynthetic genes from the cytochrome P450 gene family with at least five or more synteny blocks were identified as following: CYP89D1 (LOC_Os01g24800), CYP706C2 (LOC_Os01g50490), CYP73A35P (LOC_Os02g57760), CYP71A3 (LOC_Os01g72400), CYP71U3 (LOC_Os02g17760), CYP51H4 (LOC_Os02g21810), CYP73A40 (LOC_Os03g26770), CYP86E1 (LOC_Os03g38290), CYP81A6 (LOC_Os03g55240) and CYP735A4 (LOC_Os09g23820) (Supplementary 1).

Phylogenetic analysis of the SCC biosynthetic genes in A. thaliana and O. sativa. The phylogenetic tree comprised of 89 AtSCC and 265 OsSCC biosynthetic genes show 11 different clades of various sizes, as annotated by the gene number (GN). Clade 8 emerged as the largest group with GN = 65, followed by clade 7 (GN = 59), clade 2 (GN = 46), clade 6 (GN = 44), clade 9 (GN = 40), clade 11 (GN = 29), clade 4 (GN = 26), and clade 1 and clade 10 with GN = 20, each. Clade 5 and clade 2 were the smallest in size, with GN = 3 and GN = 2, respectively. There were 7 clades comprised of SCt and AtSCC biosynthetic genes in combination: clade 1, clade 4, clade 5, clade 6, clade 7, clade 8, clade 9, clade 10 and clade 11. Clade 1 showed nearing an equal number of OsSCC and AtSCC biosynthetic genes. In clade 1, AtNIT2 (At1g44300), AtNIT1 (At1g44310), AtNIT4 (At5g22300) and OsNRT2 (LOC_Os02g42330) were present together (Fig. 2).

In clade 4, synteny events between the AtBGLU34 (At1g47600) biosynthetic gene and OsbBLUG1 (LOC_Os06g25150), OsbBLUG9 (LOC_Os04g39814), OsbBLUG37 (LOC_Os11g08120), OsbBGLUG27 (LOC_Os08g39860) and OsbBGLUG29 (LOC_Os09g51410) biosynthetic genes were identified. Likewise, clade 8 showed a collinear relationship between the OsSOT (LOC_Os09g08190) biosynthetic gene and the AtSOT18 (At1g74090) biosynthetic gene. In clade 9, AtACO9 (At5g34440) was grouped together with Os2ODD5 (LOC_Os03g32470), OsFLS1 (LOC_Os09g18450), Os2ODD16 (LOC_Os01g24980), and Os2ODD26 (LOC_Os03g63900), whilst OsHIS1 (LOC_Os02g17940) was associated with AtACO4 (At1g03400) and AtACO8 (At3g61400). There were three syntenic pairs identified in clade 10: (i) OsCOMT4 (LOC_Os02g57760)-AtIGMT5 (At1g67920) biosynthetic genes, (ii) OsCOMT5 (LOC_Os04g09640)-AtIGMT1 (At1g21100) biosynthetic genes and, (iii) OsCOMT (LOC_Os08g06100)-AtIGMT1 (At1g21100) biosynthetic genes. In clade 11, both OsGT1 (LOC_Os11g04860) and OsIAGLU (LOC_Os09g11290) biosynthetic genes were identified as syntenic pairs of AtUGT74B1 (At1g24100). No syntenic evidence was present in clade 6 (Figs. 1 and 2).
Conserved motif analysis. A total of ten conserved motifs were identified from *A. thaliana* and *O. sativa* SCC biosynthetic genes in clade 1, clade 4, clade 6, clade 8, clade 9, clade 10 and clade 11. The detailed motif sequence information and annotations are provided in Supplementary 2. The motif distribution was similar within the clade level. All the SCC-encoding biosynthetic genes contained at least one motif, whereas a total of...
14 genes displayed all 10 motifs with mosaic patterning. No apparent pattern was observed among the motifs within the different species. The following motifs were annotated as isopropyl malate dehydrogenase (IPMDH): motifs 2, 3, 4, 7, 9 and 10. All the SCC-encoding biosynthetic genes in clade 4 displayed motif 6 (annotated as glucosidase) (Supplementary 2). Motif 1 and motif 5 contain the conserved sulfotransferase domain (Fig. 3). In clade 9, at least nine different motifs were consistently present in the member genes. Motifs 1, 2, 4 and 5 are annotated with the O-methyltransferase domain. The conserved motifs 1, 2, 4, 8 and 9 were also described as UDP-glycosyltransferase (Fig. 3) (Supplementary 2).

Figure 2. Phylogenetic analysis of collated sulfur-encoding biosynthetic genes in Arabidopsis thaliana and rice (Oryza sativa). The tree is constructed with MEGA software.
The exon–intron structure of the SCC biosynthetic genes in *A. thaliana* and *O. sativa*. Generally, the number of exons (EN) and introns (IN) in *Arabidopsis* and rice displayed no apparent trend by species. Nevertheless, similar exon–intron architecture was observed among the clades of collated AtSCC and OsSCC biosynthetic genes. The number of EN in AtSCC and OsSCC biosynthetic genes ranged from 1 to 13 (Fig. 4). The AtBGLU34 (At1g47600), OsBGLU24 (LOC_Os06g21570) and OsBGLU27 (LOC_Os08g39860) biosynthetic genes showed the highest exon and intron distribution with EN = 13 and IN = 12, respectively. There were eight

| Clade 1 | Clade 2 | Clade 3 | Clade 4 | Clade 5 | Clade 6 | Clade 7 | Clade 8 | Clade 9 | Clade 10 | Clade 11 |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| AT3G44300 | AT3G44310 | Os02g42330 | AT5G22300 | Os03g45320 | AT1G80560 | AT1G31180 | AT5G14200 | AT2G44490 | AT1G47600 | AT5G25980 |
| Os06g21570 | Os04g39814 | Os11g08120 | Os08g39860 | Os09g31410 | Os04g08824 | AT2G22330 | AT4G39950 | Os09g08190 | AT1G18590 | AT1G74090 |
| AT5G61290 | AT5G07800 | Os01g40570 | Os06g14390 | AT5G23010 | Os12g04440 | AT1G03400 | AT3G61400 | AT5G43440 | AT2G25450 | Os02g17940 |
| Os01g24980 | Os03g63900 | Os03g32470 | Os09g18450 | Os02g57760 | Os08g06100 | Os04g09604 | AT1G76790 | AT1G21100 | AT1G75530 | AT5G37170 |
| Os09g11290 | Os06g23800 | AT2G43100 | Os11g04860 | AT2G43840 | AT1G24100 | Os09g11290 | Os06g23800 | AT5G43100 | Os11g04860 | AT2G43840 |

**Figure 3.** Motif distribution structure of *Arabidopsis thaliana* and *Oryza sativa* sulfur-encoding biosynthetic genes grouped by clades. The *A. thaliana* (ATXXXXXXX) and *O. sativa* (OsXXXXXXX) gene IDs are written in black and red, respectively. Detailed information on the motif sequence information and annotation is available in Supplementary 2.
Two exon gains were observed in genes in established synteny and similarity against motifs distributions and exon–intron structure of 18 gained one exon, while their syntenic pairs At-exon losses in each biosynthesis genes (Table 1). At and (vi) At and Os remaining clades displayed a similar trend; the tryptophan signalling pathways. The highest number of genes were significantly enriched in the metabolic pathways and GO and pathway enrichment analysis of rice and Arabidopsis S-glycoside metabolic process, glycosinolate metabolite process and glucosinolate metabolic processes. In both Arabidopsis SCC biosynthetic genes, the following terms were significantly enriched: (i) oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen was the most significantly enriched (with more than 180 hits) term in both rice and Arabidopsis SCC biosynthetic genes, the following terms were commonly present.

The KEGG pathway enrichment showed involvement of the rice-Arabidopsis homologous genes in 10 different signalling pathways. The highest number of genes were significantly enriched in the metabolic pathways and biosynthesis of secondary metabolites with a total number of genes of 80 and 67, respectively. The tryptophan metabolism and 2-oxocarboxylic acid metabolism were fairly high at 25 and 21, respectively (Fig. 6).

Chromosomal distributions of the SCC biosynthetic genes in A. thaliana and O. sativa. Highly conserved SCC biosynthetic genes were physically mapped on the Arabidopsis and rice genomes. The SCC biosynthetic gene distribution in Arabidopsis and rice chromosomes are unequal (Fig. 5). In Arabidopsis, chromosome 1 showed the highest gene number (GN = 10), followed by chromosome 5 (GN = 8), chromosome 2, (GN = 5), chromosome 3 (GN = 3) and chromosome 4 (GN = 1). The rice SCC biosynthetic genes are distributed in all the 12 chromosomes except chromosomes 5 and 7. Chromosomes 1, 3, 4, 6, 8, 10, 11 and 12 contain one to three OsSCC biosynthetic genes, and the highest number of OsSCC biosynthetic genes (GN = 4) are distributed on chromosomes 2 and 9. No tandem duplications are observed among the SCC biosynthetic genes; no two gene loci are arranged in close proximity and genes are separated by more than five duplicated blocks (Fig. 5). The OsIPM1 and OsIPMS2 encoding proteins have the longest protein length (635 aa) in rice and AtTCP79B3 (543 aa) in Arabidopsis. OsIAGLU and AtIPM12 are the shortest protein-encoding gene in rice (113 aa) and Arabidopsis (256 aa), respectively. More than half of the genes encoded by the SCC biosynthetic genes are acidic, with a theoretical pl (isoelectric point) ranging from 4.63 to 6.24 (Arabidopsis) and 5.1 to 6.8 (rice). The average molecular weight (MW) of AtSCC biosynthetic genes is 45.94 kDa and 43.24 kDa for the OsSCC biosynthetic genes (Table 2).

Discussion
Sulfur (S) is a secondary macronutrient that regulates plant physiology, growth and developmental processes such as photosynthesis, biosynthesis of sulfur-containing compounds (SCCs) and hormone biosynthesis. It is the 4th major nutrient for crop production after nitrogen, phosphorus and potassium. In higher plants, the S acquisition and assimilation consumes high energy. The S element is taken up by plants as sulphate ions mainly via roots and a small amount can be absorbed through leaves. In rice, the S element, ‘S-containing genes and associated SCCs are critically involved in stress-responsive mechanisms’.

For example, the glutathione S-transferase (GST), a detoxification enzyme ubiquitously present in vertebrates and invertebrates plays an important role in xenobiotic compound detoxification. GST activity is associated with oxidative stress protection as it acts as a mediating substrate in various biochemical reactions, interacts with phytohormones and redox metabolites, and coordinates stress-induced signalling events. Glutathione (GSH) mediates abiotic and biotic stress resistance using the ROS-scavenging mechanism of the first defense line system.
in crop plants. Extensive studies have evident GSH-mediated tolerance mechanisms against salinity, drought, heavy metal toxicity, chilling and herbicides in rice, wheat, barley, soybean and canola. The effect of S
amendment on plant defense response had contributed to similar evidence. As such, the soil amendment of S-containing fertilizer on wheat varieties increased resistance against brown rust and improved the overall productivity\(^9\).

Rice yield-impeding factors include pest and pathogen, climate, weather, soil infertility, heavy metal contamination and others. Presently, rice yield enhancement strategies are vigorously carried out by tapping into various aspects of rice biology. Genetic studies, molecular breeding, genetic engineering, heterosis breeding and population improvement are amongst the most sought-after tools utilized in modern rice breeding\(^{49-51}\). Since a large number of studies on rice S and SCCs have been linked to stress mechanisms and defense responses, a comprehensive annotation of SCC-encoding genes in the rice genome is important to necessitate enhanced manipulation strategies in breeding approaches\(^3-6\).

In this study, a total of 665 OsSCC biosynthetic genes were identified as the homologs of AtSCC query sequences. A total of 477 syntenic gene pairs (Arabidopsis-rice) and 25 rice SCC biosynthetic genes (AtSCC homologs) were obtained using a comprehensive analysis entailing synteny, phylogenetic, conserved motif distribution and gene structure. The synteny analysis identified the gene order and compared the genomic structural distribution and gene structure. The synteny analysis identified the gene order and compared the genomic structural changes of the target genes. Shared synteny assumes a common ancestor/evolutionary origin and a syntenic fragment shares a similar function\(^{57,58}\). A small number of genes identified as Arabidopsis-rice syntenies, suggests the early Angiosperm divergence of monophyletic monocot from its eudicot relatives\(^59\). The monocot rice genome with 5 chromosomes typically diverged from the eudicot Arabidopsis genome (7 chromosomes) of a higher

| OsSCC ID | Criteria | 1 | 2 | 3 | 4 |
|----------|----------|---|---|---|---|
| LOC_Os02g42330 | OsNit2 | 7.00E-158 | 1 | 3/5 | 5/5 |
| LOC_Os03g45320 | OsIPMDH | 0.00E-000 | 1 | 10/10 | 11/8 |
| LOC_Os04g35320 | OsIPMDH | 0.00E+00 | 1 | 10/10 | 11/9 |
| LOC_Os09g31410 | OsBGLU29 | 2.00E-149 | 4 | 7/9 | 6/13 |
| LOC_Os08g39860 | OsBGLU27 | 3.00E-158 | 4 | 9/9 | 13/13 |
| LOC_Os06g21570 | OsBGLU24 | 1.00E-156 | 4 | 9/9 | 13/13 |
| LOC_Os14g39814 | OsBGLU9 | 1.00E-82 | 4 | 7/9 | 7/13 |
| LOC_Os11g8120 | OsBGLU35 | 2.00E-41 | 4 | 5/9 | 7/13 |
| LOC_Os09g81890 | OsSOT | 8.00E-66 | 8 | 4/4 | 1/1 |
| LOC_Os06g14390 | OsACO4 | 1.00E-84 | 9 | 7/9 | 2/3 |
| LOC_Os06g14390 | OsACO4 | 5.00E-75 | 9 | 7/9 | 2/3 |
| LOC_Os09g18450 | OsFLS | 3.00E-111 | 9 | 9/9 | 3/3 |
| LOC_Os11g24980 | Os2ODD16 | 7.00E-50 | 9 | 9/9 | 4/3 |
| LOC_Os03g6900 | Os2ODD26 | 4.00E-46 | 9 | 9/9 | 1/3 |
| LOC_Os03g32470 | Os2ODD25 | 1.00E-35 | 9 | 9/9 | 3/3 |
| LOC_Os11g04670 | OsIPMS1 | 7.00E-173 | 9 | 1/1 | 12/10 |
| LOC_Os12g44440 | OsIPMS2 | 5.00E-173 | 9 | 1/1 | 12/10 |
| LOC_Os02g17940 | Os2ODDD12 | 1.00E-34 | 9 | 9/9 | 4/3 |
| LOC_Os01g24980 | Os2ODD16 | 3.00E-39 | 9 | 9/9 | 4/3 |
| LOC_Os03g32470 | Os2ODD25 | 8.00E-23 | 9 | 9/9 | 3/3 |
| LOC_Os02g57760 | OsCOMT4 | 3.00E-40 | 10 | 7/8 | 4/3 |
| LOC_Os08g06100 | OsROMT9 | 6.00E-92 | 10 | 8/8 | 2/3 |
| LOC_Os04g96064 | OsCOMT5 | 3.00E-74 | 10 | 8/8 | 4/3 |
| LOC_Os02g43830 | OsSat2 | 2.00E-067 | 11 | 1/1 | 1/1 |
| LOC_Os11g04860 | OsUGT75E1 | 4.00E-058 | 11 | 7/8 | 1/2 |
| LOC_Os09g11290 | OsIAGLU | 2.00E-12 | 11 | 2/8 | 1/2 |
| LOC_Os04g08824 | OsCP79A10 | N/A | 6 | 9/10 | 3/3 |
| LOC_Os04g08824 | OsCP79A10 | N/A | 6 | 9/10 | 3/3 |
| LOC_Os10g40570 | OsFMOGS-ox-like5 | N/A | 8 | 8/8 | 7/7 |
| LOC_Os10g40570 | OsFMOGS-ox-like5 | N/A | 8 | 8/8 | 7/7 |
| LOC_Os06g23800 | OsFMOGS-ox | N/A | 11 | 1/1 | 6/2 |

Table 1. Mining for Oryza sativa sulfur-encoding biosynthetic genes (OsSCC) with Arabidopsis sulfur-encoding biosynthetic gene (AtSCC) input data. Selection criteria are described as following: (1) synteny events; (2) phylogenetic clade; (3) motif composition (Os/At); and (4) number of exon (EN) with AtSCC biosynthetic genes (Os/At).
The synteny analysis of Arabidopsis-rice SCC biosynthetic genes implies the ancient existence of SCC biosynthetic genes, even before the divergence of the Arabidopsis-rice (eudicot-monocot).

The SCC biosynthetic gene distribution pattern suggests the occurrence of an expansion event during evolution which could have possibly gone through gene co-localization or inter-chromosomal translocation. The phylogenetic and gene structure pattern of the SCC-encoding biosynthetic genes suggest exon loss and gain events during Arabidopsis-rice (eudicot-monocot) evolution. The exon–intron arrangement pattern in 25 AtSCC and 18 OsSCC suggests that the species-specific genome features are conserved. The mosaic patterning of the SCC gene exon–intron regions could be associated with evolutionary forces that shaped the SCC biosynthetic gene structure dynamics.

Motifs are frequently occurring (conserved) regions within a DNA sequence. Found within the regulatory regions such as promoters and 3' UTRs, the 4–10 base pair motifs carry significant genome regulatory functions. Two species are likely to be close relatives if they share a high content of common motifs. During speciation, mutations lead to either an accumulation or loss of motifs (motif turnover) and thus, a motif content analysis is often regarded as more advantageous than the counterpart sequence similarity search analysis. Our results showed that at least 10 different motifs identified in the Arabidopsis and rice SCC-encoding biosynthetic genes have similar distribution patterns by clades.
As shown in Table 1, five -glucosidase genes from clade 4 showed syntenies with glucosidase genes from clade 10 were characterized as O-methyltransferase, a key gene in response. In clade 1, SCC biosynthetic genes identified in this study showed potential functional roles in plant defense whereas, in Clade 10, there are 4 motifs corresponding to O-methyltransferase domain (Supplementary Table S1). The DNA and protein sequences. Likewise in Clade 4, about 7 different motifs are annotated as glycosyl hydrolase.

Table 2. Sulfur-encoding biosynthetic gene, chromosomal and protein level description in Arabidopsis and rice. Each gene is characterized according to its chromosome number, chromosomal loci, open reading frame (ORF) and physical characteristics of the encoding protein.

| Gene ID     | Gene name     | Chr | Location       | ORF length (bp) | Protein Length | PI  | MW (kDa) |
|-------------|---------------|-----|----------------|-----------------|----------------|-----|----------|
| AT1G03400   | AAtCO4        | 1   | 842,747–844,190| 1056            | 351            | 6.15| 39.13    |
| AT1G21100   | AAtGMT1       | 1   | 7,386,839–7,388,428| 1122           | 373            | 5.01| 40.869   |
| AT1G24100   | AAtUGT74B1    | 1   | 8,525,435–8,527,087| 1383           | 460            | 4.63| 51.002   |
| AT1G31180   | AAtM3         | 1   | 11,142,714–11,144,633| 1215           | 404            | 5.55| 43.847   |
| AT1G47600   | AAtBGU14      | 1   | 17,491,732–17,494,759| 1536           | 511            | 8.21| 57.542   |
| AT1G74900   | AAtSOT18      | 1   | 27,862,909–27,864,193| 1053           | 350            | 5.5  | 40.456   |
| AT1G76790   | AAtRGT5       | 1   | 28,822,186–28,823,673| 1104           | 567            | 4.76| 40.222   |
| AT2G22330   | AAtCYP79B3    | 2   | 9,488,554–9,491,187| 1632           | 543            | 8.17| 61.457   |
| AT2G25450   | AAtGLS-OH     | 2   | 10,829,916–10,831,655| 1080           | 359            | 6.24| 40.351   |
| AT2G34100   | AAtPM12       | 2   | 17,920,660–17,921,689| 771            | 256            | 6.01| 27.043   |
| AT2G43300   | AAtNT2        | 3   | 15,983,311–15,985,535| 1020           | 339            | 5.24| 37.153   |
| AT3G61400   | AAtAC08       | 3   | 22,718,956–22,720,397| 1113           | 370            | 5.64| 41.601   |
| AT4G39950   | AAtCYP79B2    | 4   | 18,525,246–18,527,579| 1626           | 541            | 8.73| 63.347   |
| AT5G08700   | AAtFMOGS-OX-like9| 5   | 2,486,576–2,489,296| 1383           | 460            | 6.21| 52.337   |
| AT5G14200   | AAtM1D1       | 5   | 4,576,202–4,578,402| 1230           | 409            | 5.81| 44.161   |
| AT5G23010   | AAtM1A1       | 5   | 7,703,092–7,706,896| 1521           | 506            | 7.28| 55.125   |
| AT5G34440   | AAtAC09       | 5   | 17,455,233–17,456,657| 1098           | 565            | 6.18| 40.86    |
| AT5G61290   | AAtFMOGS-OX-like8| 5   | 24,648,558–24,650,815| 1386           | 461            | 4.9  | 52.406   |
| LOC_Os01g24980| Os2ODD16     | 1   | 14,077,629–14,080,716| 1035           | 344            | 5.62| 38.731   |
| LOC_Os01g40570| OsFMOGS-OX-like5| 10 | 21,724,416–21,727,181| 1449           | 482            | 5.69| 53.726   |
| LOC_Os11g04670| OsPMS1       | 11  | 1,989,201–1,995,087| 1908           | 635            | 6.46| 68.448   |
| LOC_Os11g04860| OsUGT75E1    | 11  | 2,067,727–2,069,430| 1449           | 482            | 5.38| 54.068   |
| LOC_Os11g08120| OsUGL35      | 11  | 4,262,908–4,265,304| 579            | 197            | 9.81| 22.062   |
| LOC_Os12g04440| OsPMS2       | 12  | 1,888,943–1,894,920| 1908           | 635            | 6.46| 68.461   |

For instance, in clade 1, six motifs were annotated as 3-isopropylmalate dehydrogenase despite differences in the DNA and protein sequences. Likewise in Clade 4, about 7 different motifs are annotated as glycosyl hydrolase family 1 whereas, in Clade 10, there are 4 motifs corresponding to O-methyltransferase domain (Supplementary 2). The OsSCC biosynthetic genes identified in this study showed potential functional roles in plant defense response. In clade 1, LOC_Os02g42330 (nitrilase 1), the syntenic pair of At1g44300 (nitrilase 2) was reported to participate in the tryptophan-dependent pathway of auxin biosynthesis in rice. Three OsSCC biosynthetic genes from clade 10 were characterized as O-methyltransferase, a key gene in Arabidopsis indolic glucosinolate modification. As shown in Table 1, five -glucosidase genes from clade 4 showed syntenies with glucosidase 34.
(AtBGLU34). AtBGLU34 plays a major role in response to salt stress\(^6\) and indolic glucosinolate biosynthesis\(^6\) in *Arabidopsis*.

The SCC biosynthetic genes distributed among the unique phylogenetic clades, carrying similar motif pattern are possibly sharing a similar function. The unique motifs in each clade could be associated with specific functional roles of the SCC biosynthetic genes. The current findings shed insights on the potential functional roles of SCC biosynthetic genes in rice as more than half of the genes were putatively involved in the biosynthesis of aliphatic glucosinolate and indolic glucosinolate. Based on the gene ontology and pathway enrichment analysis, the *Arabidopsis*-rice homologous SCC-encoding genes were significantly enriched in the sulfur compound metabolic process (BP), oxidoreductase activity, acting on paired donors with incorporation or reduction of molecular oxygen (MF) and biosynthesis of secondary metabolites (KEGG pathway) (Fig. 6). This may suggest the role of

---

**Figure 6.** Gene ontology (GO) and pathway enrichment analysis. The bubble plot represents the top 20 significantly enriched terms of the *Arabidopsis*-rice homologous SCC-encoding genes. The GO terms are presented in (i-ii) biological process and (iii-iv) molecular functions whereas the KEGG pathways are presented in (v-vii). Red arrows represent the terms shared among the *Arabidopsis*-rice orthologous genes. The results are visualized at P < 0.05 using ShinyGO v0.75 (http://bioinformatics.sdstate.edu/go75/).
the SCC-encoding genes in S assimilation, whereby the reduction of sulphate ion to sulphide and subsequent S-containing amino acids (methionine and cysteine) via the adenosine phosphosulphate pyrophosphate (APS) and phosphoadenosine phosphosulphate (PAPS) is catalyzed by the participating enzyme activities.

In plant breeding strategies, exploiting the naturally occurring genetic variation is of utmost fundamental in controlling genes of agroscopic importance. Physiological maps of rice SCC biosynthetic genes provided in this study could be harnessed for chromosomal region manipulated breeding techniques such as the target chromosome-segment substitution and hotspot chromosomal regional positioning of desirable candidate genes. The findings enable the selection of desirable target rice genes which are tightly linked to S and SCC-encoding genes with a putative functional role in stress response mechanisms.

Conclusions

Rice SCCs biosynthetic genes show syntenic associations with Arabidopsis homologs (AtSCCs). The high degree of conservation between the AtSCC and OsSCC genes suggests long conservation history which could be implicated in SCC gene functions in plant defense response. The present findings not only identified the rice SCC-encoding genes (OsSCC) but also stretch further to include chromosomal level-mapping to better inform new directions in rice functional research and breeding manipulation strategies.

Data availability

All open-source genomic datasets analysed in this study are available in the Phytozome v13.0 database (https://phytozome-next.jgi.doe.gov/), Arabidopsis Information Resource v10.0 (TAIR) (https://www.arabidopsis.org) and O. sativa Genome Annotation Project Database v7.0 (RGAP) (http://rice.uga.edu/).

Received: 28 January 2022; Accepted: 4 August 2022
Published online: 15 August 2022

References

1. Zhang, N. et al. Sulfur deficiency exacerbates phytotoxicity and residues of imidacloprid through suppression of thiol-dependent detoxification in lettuce seedlings. Environ. Pollut. 291, 118221 (2021).
2. Bednařek, P. Sulfur-containing secondary metabolites from arabidopsis thaliana and other Brassicaceae with function in plant immunity. ChemBioChem 13(13), 1846–1859 (2012).
3. Aarabi, F., Naake, T., Fernie, A. R. & Hoefgen, R. Coordinating sulfur pools under sulfate deprivation. Trends Plant Sci. 25, 1–10 (2020).
4. Yadav, P. & Srivastava, S. Effect of thiourea application on root, old leaf and young leaf of two contrasting rice varieties (Oryza sativa L.) grown in arsenic contaminated soil. Environ. Technol. Innov. 21, 101368 (2021).
5. Kopriva, S., Calderwood, A., Weckopp, S. C. & Koprivova, A. Plant sulfur and big data. Plant Sci. 241, 1–10 (2015).
6. Capaldi, F. R. et al. Sulfur metabolism and stress defense responses in plants. Trop. Plant Biol. 8, 60–73 (2015).
7. Chen, D. et al. N6-methyladenosine methylation analysis reveals transcriptome-wide expression response to salt stress in rice roots. Environ. Exp. Bot. 201, 104945 (2022).
8. Zhang, H., Li, Y. Y. & Zha, J. K. Developing naturally stress-resistant crops for a sustainable agriculture. Nat. Plants 4(12), 989–996 (2018).
9. Huang, Y. et al. Transcriptomic (RNA-seq) analysis of genes responsive to both cadmium and arsenic stress in rice root. Sci. Total Environ. 666, 445–460 (2019).
10. Hasanuzzaman, M. et al. Exogenous proline and glycine betaine mediated upregulation of antioxidant defense and glyoxalase systems provides better protection against salt-induced oxidative stress in two rice (Oryza sativa L.) varieties. Biomed. Res. Int. 2014, 757219 (2014).
11. Zhou, H. et al. Rice glutathione peroxidase1-mediated oxidation of bZIP68 positively regulates ABA-independent osmotic stress. Mol. Plant 15, 651–670 (2022).
12. Agerbirk, N. & Olsen, C. E. Glucosinolate structures in evolution. Phytochemistry 77, 16–45 (2012).
13. Kunstler, A., Gullner, G., Adam, A. L., Nagy, J. K. & Király, L. The versatile roles of sulfur-containing biomolecules in plant defense-a road to disease resistance. Plants 9, 1705 (2020).
14. Teutor, S. & Gershenson, J. Herbivore induction of the glucosinolate-myrosinase defense system: Major trends, biochemical bases and ecological significance. Phytochem. Rev. 8, 149–170 (2009).
15. Pedras, M. S. C., Yaya, E. E., Glawischnig, E. & Links, D. A. The phytoalexins from cultivated and wild crucifers: Chemistry and physiology. J. Med. Chem. 54(12), 6381–6405 (2011).
16. Bell, L. et al. Taste and flavor perceptions of glucosinolates, isothiocyanates, and related compounds. Mol. Nutr. Food Res. 62(18), e1700990 (2018).
17. Beck, T. K., Jensen, S., Bjoern, G. K. & Kidmose, U. The masking effect of sucrose on perception of bitter compounds in brassica vegetables. J. Sens. Stud. 29(3), 190–200 (2014).
18. Hawkesford, M. et al. Functions of macronutrients. In Marschner’s Mineral Nutrition of Higher Plants, 3rd edn (ed. Marschner, P.) 135–189 (Academic Press, 2012).
19. Burow, M., Wittstock, U. & Gershenson, J. Sulfur-containing secondary metabolites and their role in plant defense. In Sulfur Metabolism in Phototrophic Organisms (eds Hell, R. et al.) 201–222 (Springer, 2008).
20. Harun, S., Abdullah-Zawawi, M. R., Goh, H. H. & Mohamed-Hussein, Z. A. A comprehensive gene inventory for aliphatic glucosinolate biosynthetic pathway in Arabidopsis thaliana. J. Agric. Food Chem. 68(28), 7281–7297 (2020).
21. Ashari, K. S., Abdullah-Zawawi, M. R. & Harun, S. Rebuilding the transcriptional regulatory network in arabisobisp thaliana aliphatic glucosinolate biosynthetic pathway. Sains Malaysia. 47(12), 2993–3002 (2018).
22. Zuber, H. et al. Sultr4 mutant seeds of Arabidopsis have an enhanced sulphate content and modified proteome suggesting metabolic adaptations to altered sulphate compartmentalization. BMC Plant Biol. 10, 78 (2010).
23. Meghan, M. et al. Arabidopsis ETHE1 encodes a sulfur dioxygenase that is essential for embryo and endosperm development. Plant Physiol. 160(1), 226–236 (2012).
24. Wawrzynska, A. & Sirko, A. To control and to be controlled: Understanding the Arabidopsis SLIM1 function in sulfur deficiency through comprehensive investigation of the EIL protein family. Front Plant Sci. 5, 575 (2014).
25. Glaser, K. et al. Exploring the Arabidopsis sulfur metabolism. Plant Physiol. 177(1), 31–45 (2015).
26. Goodstein, D. M. et al. Phytozome: A comparative platform for green plant genomics. Nucleic Acids Res. 40, D1178–D1186 (2012).
27. Huala, E. et al. The Arabidopsis information resource (TAIR): A comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucleic Acids Res. 29(1), 102–105 (2001).
Author contributions
The first author (MRAZ) is funded by Skim Zamalah Universiti Penyelidikan, Universiti Kebangsaan Malaysia. This work was supported by ERGS/1/2013/STG07/UKM/02/3 awarded to ZAMH by the Malaysian Ministry of Higher Education and Ministry of Science, Technology and Innovation. The Ph.D. scholarship obtained by the first author (MRAZ) is funded by Skim Zamalah Universiti Penyelidikan, Universiti Kebangsaan Malaysia.

Acknowledgements
This work was supported by ERGS/1/2013/STG07/UKM/02/3 awarded to ZAMH by the Malaysian Ministry of Higher Education and Ministry of Science, Technology and Innovation. The Ph.D. scholarship obtained by the first author (MRAZ) is funded by Skim Zamalah Universiti Penyelidikan, Universiti Kebangsaan Malaysia.

Author contributions
Conceptualization, M.R.A.Z. and Z.A.M.H.; methodology, investigation, writing—original draft preparation, M.R.A.Z.; validation, writing—review and editing, N.G., N.A.N.M., N.M.A., Z.Z. and Z.A.M.H.; visualization,
supervision, writing-revision and improved the draft manuscript, Z.A.M.H.; supervision and funding acquisi-
tion. All authors reviewed the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/
10.1038/s41598-022-18068-0.

Correspondence and requests for materials should be addressed to N.G.

Reprints and permissions information is available at www.nature.com/reprints.

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 licence, and indicate if changes were made. The images or other third party material in this
article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the
material. If material is not included in the article’s Creative Commons licence 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 licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022