Genome-enabled discovery of anthraquinone biosynthesis in *Senna tora*

Sang-Ho Kang, Ramesh Prasad Pandey, Chang-Muk Lee, Joon-Soo Sim, Jin-Tae Jeong, Beom-Soon Choi, Myunghee Jung, Daniel Ginzburg, Kangmei Zhao, So Youn Won, Tae-Jin Oh, Yeisoo Yu, Nam-Hoon Kim, Ok Ran Lee, Tae-Ho Lee, Puspalata Bashyal, Tae-Su Kim, Woo-Haeng Lee, Charles Hawkins, Chang-Kug Kim, Jung Sun Kim, Byoung Ohg Ahn, Seung Yon Rhee & Jae Kyung Sohng

*Senna tora* is a widely used medicinal plant. Its health benefits have been attributed to the large quantity of anthraquinones, but how they are made in plants remains a mystery. To identify the genes responsible for plant anthraquinone biosynthesis, we reveal the genome sequence of *S. tora* at the chromosome level with 526 Mb (96%) assembled into 13 chromosomes. Comparison among related plant species shows that a chalcone synthase-like (CHS-L) gene family has lineage-specifically and rapidly expanded in *S. tora*. Combining genomics, transcriptomics, metabolomics, and biochemistry, we identify a CHS-L gene contributing to the biosynthesis of anthraquinones. The *S. tora* reference genome will accelerate the discovery of biologically active anthraquinone biosynthesis pathways in medicinal plants.
**Sen**na *tor*a (L.) Roxb., also known as *Cassia tor*a, is a favorite of ancient Chinese and Ayurvedic herbal medicine that is now widely used around the world and recorded as Model List of Essential Medicines by the World Health Organization. Recent studies point to *S. tor*a’s beneficial activities against microbial and parasitic infections, prevention or delay of the onset of neurodegenerative diseases, and diabetes. *S. tor*a’s positive health impact is attributed to the significant amount of anthraquinones in mature seeds and other parts of the plant. As an ancient medicine, anthraquinones from seeds of other *Sen*na species are commonly used for treating various diseases. Despite the extensive applications of *Sen*na plants in medicine and industry, molecular and genomic studies of this remarkable genus of plants have been limited. Elucidating the genes responsible for the biosynthesis of anthraquinones in *S. tor*a will aid molecular breeding and the development of tools for probing its biochemistry.

Anthraquinones are aromatic polyketides made by bacteria, fungi, insects, and plants. Besides their medicinal benefits, natural anthraquinones are garnering attention as alternatives to synthetic dyes that damage aquatic ecosystems. More than three decades ago, radiolabeled feeding experiments indicated that the A and B rings of anthraquinones were derived from shikimate and mevalonate/methyl-D-erythritol 4-phosphate pathways. More than three decades ago, radiolabeled feeding experiments indicated that the A and B rings of anthraquinones were derived from shikimate and α-ketoglutarate via O-succinylbenzoate and C ring from mevalonate pathway via isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP) or 2-C-methyl-D-erythritol 4-phosphate (MEP) pathway. Contrarily, recent studies speculated biosynthesis of anthraquinones in plants to occur via a polyketide pathway. Type III polyketide synthase (PKS) enzymes could actively catalyze seven successive decarboxylative condensations of malonyl-CoA to produce an octaketide chain. The linear polyketide chain undergoes cyclization and decarboxylation reactions to produce the core unit of polyketides such as atrochryson carboxylic acid followed by decarboxylation to atrochryson and dehydration to emodin anthrone. Type III polyketide synthase (PKS) enzymes could actively catalyze seven successive decarboxylative condensations of malonyl-CoA to produce an octaketide chain. The linear polyketide chain undergoes cyclization and decarboxylation reactions to produce the core unit of polyketides such as atrochryson carboxylic acid followed by decarboxylation to atrochryson and dehydration to emodin anthrone.

### Results and discussion

**S. tor*a genome assembly and annotation.** We generated one of the highest-quality genomes for medicinal plants. The *S. tor*a cultivar Myeongyungun genome was assembled with Pacific Biosciences long-read sequencing (146.2× coverage) by FALCON v0.4 (Supplementary Tables 1 and 2). To improve the quality of genome assembly, we performed error correction with Sequel data by Arrow v2.1.0 and further corrected it with 101.2× Illumina data using BWA and GATK. *S. tor*a has an estimated genome size of ~547 Mb based on k-mer analysis (Supplementary Fig. 2). Through chromosome conformation capture (Hi-C) mapping, we generated 13 chromosome-scale scaffolds (hereafter called chromosomes, Chr1–Chr13) totaling 502.6 Mb, 95.5% of the ~526.4 Mb of the assembled genome (Table 1 and Supplementary Figs. 3 and 4). We evaluated the quality of assembly using Benchmarking Universal Single Copy Orthologs (BUSCO) and 10 bacterial artificial chromosome (BAC) clones, and comparing to a linkage map (Supplementary Fig. 5). BUSCO estimates 94.3% completeness (Supplementary Table 3), suggesting that the assembly includes most of the *S. tor*a gene space. BAC sequence alignments showed high mapping rates (99.8%) with the assemblies (Supplementary Fig. 6 and Supplementary Table 4). We also built a genetic map of diploid *S. tor*a, to which 401.1 Mb of the assembled scaffolds were mapped (Supplementary Fig. 7). The 13 linkage groups matched well to the 13 chromosomes, indicating the high quality of *S. tor*a genome assembly (Supplementary Fig. 8).

**S. tor*a’s genomic content is consistent with other sequenced plant genomes.** A total of 45,268 genes were annotated with the average gene length (3,157 bp), exon sequence length (217 bp with 4.37 exons per gene), and intron length (655 bp) that were similar to those of other legume species (Table 1 and Supplementary Fig. 9). Among the protein-coding genes, 31,010 (68.50%) showed homology to characterized genes based on BLAST searches and 25,453 (56.23%) and 17,450 (38.55%) were assigned to Gene Ontology (GO) terms and KEGG pathways, respectively (Supplementary Table 5). As expected, the genes were unevenly distributed with an increase in density toward the ends of the pseudomolecules (Fig. 1a). We also identified genes encoding for 839 tRNA, 752 rRNA, 3,278 IncRNA (Fig. 1a, Supplementary Table 6, and Supplementary Data 1), and 1,644 transcription factors (TFs) from 36 families that accounted for 3.63% of the protein-coding genes (Supplementary Table 7).

### Comparative genomics and gene family evolution analysis

To assess candidate gene families involved in anthraquinone biosynthesis, we compared the *S. tor*a genome with those from 15 related plant species. Reciprocal pairwise comparisons of the 16 species (15 legumes and grapevine) revealed that *S. tor*a has...
The most species-specific genes of all the 16 plants compared, with 7,231 (15.9%) genes that are specific to S. tora (Fig. 1b, Supplementary Fig. 10, and Supplementary Table 8). We compared gene family expansion and contraction across the species to identify gene families that were expanded or contracted in S. tora. Of the 36,597 gene families found among the sixteen species, 2,874 and 3,371 gene families were expanded and contracted in S. tora, respectively (Supplementary Fig. 11). The gene families that were expanded in S. tora were enriched for several Gene Ontology (GO) and KEGG terms, including those involved in specialized metabolism including “phenylpropanoid biosynthesis,” “isoflavanoid biosynthesis,” and “terpene biosynthesis,” likely reflecting the importance of genes for the biosynthesis of phenolics, isoflavonoids, and terpenoids in S. tora (Fig. 1c and Supplementary Data 2). To investigate S. tora metabolism further, we developed a genome-scale metabolic network database of S. tora named StoraCyc and identified enriched metabolic pathways. Expanded gene families in S. tora were enriched in phenolic and nitrogen-containing specialized metabolism, cofactor, carbohydrate, and hormone metabolism of StoraCyc (Supplementary Fig. 12 and Supplementary Table 9). We also examined enriched metabolic domains in families that are expanded only in S. tora, rapidly expanded in S. tora, and rapidly expanded only in S. tora. Phenolic specialized metabolism was the only domain of metabolism enriched in all these families (Supplementary Fig. 12 and Supplementary Data 3).

We next probed which of the lineage-specifically expanded families might be involved in anthraquinone biosynthesis. In plants, type III polyketide synthases such as chalcone synthases (CHSs) are involved in the biosynthesis of plant specialized metabolites, particularly acetate-pathway-derived flavonoids, stilbenes, and aromatic polyphenols. The S. tora CHS family contains twelve CHS (Supplementary Fig. 13) and sixteen CHS-L genes (Supplementary Fig. 13 and Supplementary Table 10). Interestingly, the CHS-L gene family specifically and rapidly expanded only in the S. tora genome (16 genes in S. tora,
To test the hypothesis that CHS-Ls might be involved in anthraquinone biosynthesis in S. tora, we turned to the tissue that is enriched in anthraquinones, the seed. We profiled anthranoids from seven developmental stages of the seed (Fig. 2a), using ten standard anthraquinones (Supplementary Table 12) as references for quantification. Anthraquinone accumulation varied in each stage (Fig. 2b). Importantly, the profile shifted toward modified derivatives such as glucoaurantio-obtusin, aurantio-obtusin, obtusifolin, and chryso-obtusin during late stages of seed development (Fig. 2b and Supplementary Table 13) essentially becoming major storage metabolites in dry seeds.

To identify genes involved in the biosynthesis of anthraquinones during seed development, we performed transcriptome and metabolome analysis from developing seeds. The majority (68%) of genes decreased in expression during seed maturation (Supplementary Fig. 14 and Supplementary Data 9). Similarly, metabolic gene expression decreased across all metabolic domains during seed maturation (Supplementary Fig. 15 and Supplementary Data 4), consistent with metabolite-profiling results, which showed that the majority of primary metabolites involved in central carbon metabolism were reduced after stage 4.

Fig. 2 Analysis of anthraquinone contents and CHS-L gene expression during S. tora seed development. a Developmental progression of S. tora seeds (Stage 1–Stage 7). b Concentrations of ten anthraquinones during the seven developmental stages of S. tora seeds (mean ± SD, n = 3). Dots represent individual values. c, d Scaled transcript expression profiles (in transcripts per million, TPM) of cluster 3 (554 genes) and cluster 6 (1,513 genes) during seed development and enriched metabolic domains within these two clusters. e Heatmap represents normalized transcripts per million (TPM) from two biological replicates. S1–S7 represents the seed-development stages of S. tora. The source data underlying Fig. 2b are provided as a Source Data file.

Metabolite profiling and transcriptomics of seed development.
To test the hypothesis that CHS-Ls might be involved in
mixtures containing 13C₃-malonyl-CoA containing CHS-L9 or heat-denatured CHS-L9 enzyme. The inset shows a zoomed region of the EIC chromatogram. (ii) EIC for the mass 317.12 Da in reaction mixtures containing CHS-L9, CHS, or heat-denatured CHS-L9. (iii) EIC for the mass 335.13 Da in reaction mixtures containing 13C₃-malonyl-CoA containing CHS-L9 or heat-denatured CHS-L9 enzyme. Dots in the structure represent 13C-labeled carbons. (iv) EIC for the mass 301.07 Da in reaction mixtures containing malonyl-CoA as substrate containing CHS-L9, CHS, or heat-denatured CHS-L9. The inset shows a zoomed region of the EIC chromatogram. (ii) EIC for the mass 317.12 Da in reaction mixtures containing 13C₃-malonyl-CoA containing CHS-L9 or heat-denatured CHS-L9 enzyme. e PKS-mediated biosynthetic pathway of anthraquinones. Two pathway intermediates, atrochrysone carboxylic acid and endocrocin anthrone, were produced in the CHS-L9-catalyzed reaction mixture.

Identification of a candidate anthraquinone synthase family.

With these data in hand, we searched specifically for CHS-L genes that were induced in stage 4 when the primary metabolite levels decrease and anthraquinones start to accumulate. Among the 16 CHS-L genes, two genes (ST007G228250 (CHS-L9) and ST003G058250 (CHS-L6)) showed high expression levels at stage 4 (Fig. 2e and Supplementary Table 15). This was consistent with the previous observation that CHS-L9 activity increases during seed maturation (32% genes represented by 5 clusters, Supplementary Table 14). In addition, genes in clusters 3 and 6 showed similar expression patterns as anthraquinone biosynthetic enzymes. To identify genes that showed similar expression patterns as anthraquinone biosynthesis, we first identified all genes that were differentially expressed relative to stage 1 during seed development. Co-expression analysis of differentially expressed genes during seed development detected nine co-expression clusters (Supplementary Fig. 1). Among them, clusters 3 and 6 showed similar patterns to anthraquinone accumulation in which genes were highly induced starting stage 5 (Fig. 2c, d). Cluster 6 was statistically overrepresented with genes annotated as transferases, and oxidoreductases, which may reflect enzymes involved in the tailoring of anthraquinones to produce gluco-obtusifolin, glucoaurantio-obtusin, and other derivatives including aurantio-obtusin (Fig. 2b and Supplementary Table 14). In addition, genes in clusters 3 and 6 were enriched with specialized, fatty acid and lipid, cofactor, and carbohydrate metabolism in StoraCyc (Fig. 2c, d, and Supplementary Data 7).

Biochemical confirmation of an anthraquinone enzyme class.

To perform enzymatic assays, we expressed ST007G228250 (CHS-L9) and ST003G058250 (CHS) heterologously in E. coli and purified them to homogeneity (Supplementary Fig. 19). Enzyme assays were conducted in a phosphate buffer saline containing malonyl-CoA for successive condensation reactions to produce polyketides. ST007G228250 (CHS-L9)-catalyzed reaction mixture revealed the existence of two molecules with a molecular mass of 319.08 Da and 301.07 Da (Fig. 3). Neither of these metabolites was detected in reactions containing ST003G058250 (CHS-L9) control, indicating that these two masses are most likely the products of the PKS-catalyzed reaction (Fig. 3). The ESI-MS spectrum showed a compound with a distinct peak at m/z 319.0827 (retention time (tR) 4.45 min), which corresponds exactly to the mass of atrochrysone carboxylic acid (C₁₉H₁₁O₂ with 319.0818 Da in the proton-adduct mode). Furthermore, the theoretical isotope model for the same chemical formula corroborated perfectly to the observed isotope mass (Supplementary Fig. 20). Likewise, the ESI-MS spectrum of the latter metabolite (tR 4.62 min) m/z 301.0715 matched to the mass of endocrocin anthrone (C₂₀H₁₂O₆ with calculated exact mass of 301.0712 Da) for which the theoretical isotope mass model and observed mass isotope were perfectly aligned (Supplementary Fig. 20). To further verify these metabolites as PKS-derived products, a set of
reactions were conducted with STO07/G228250 (CHS-L9) and heat-denatured STO07/G228250 (dead CHS-L9) containing 1C6-
malonyl-CoA as substrate. The EIC for all carbon-based labo-
trony-hydroxy carboxylic acid (13C6H12O7, exact mass: 335.1355 Da) (Fig. 3 and Supplementary Fig. 20) and endocroc in anthrone (13C6H12O6, exact mass: 317.1249 Da) (Fig. 3 and Supplementary
Fig. 20) was confirmed to be present in only the CHS-L9 catalyzed reaction. The observed ESI-MS spectra aligned to the
theoretical isotope mass of the corresponding metabolites. Except for these two metabolites, none of the other octaketide met abolites, such as emodin anthrone, endocroc in, emodin, chrysophanol, or isidacin, was produced even when NADPH was added in the reaction mixtures.

Previous studies reported the production of deraliment products (SEK4 and SEK4b) by plant octaketide synthases such as HpPKS2 and ArOKS (Supplementary Fig. 1) 33,34. We
were unable to detect any of the octaketide metabolites, none of the other octaketides, and oxidation. It is possible that other CHS-L genes might also
oxidize products in the reaction mixture. The complete
theoretical isotope mass of the corresponding metabolites. Except
for these two metabolites, none of the other octaketide metabo-
lites, such as emodin anthrone, endocroc in, emodin, chrysophanol,
or isidacin, was produced even when NADPH was added in the reaction mixtures.

In summary, the reference genome of S. tora revealed the rapid
evolution of putative polyketide synthase genes. By combining met abolomics, transcriptomics, and biochemical characterization of a candidate polyketide synthase, we discovered the anthranoid- forming enzyme in plants. With these tools in hand, elucidation of genes involved in the rest of the anthraquinone biosynthesis pathway in S. tora and other species will be accelerated. These resources can also be used as a platform to develop a medicinally useful cultivar of S. tora with a high content of bioactive molecules.

Methods

**DNA sequencing.** We sequenced a cultivated diploid Senna tora cv. Myeongyun (voucher number: IT987928) grown in Jeonju, Korea (N: 35° 49′; E: 127° 09′). The total DNA was extracted from young fresh leaves of S. tora cv. Myeongyun using the modified cetyltrimethylammonium bromide (CTAB) method 43. DNA purity and concentration were checked by electrophoresis analysis on 1.2% agarose gel and by DropSend®96 Spectrophotometer (Thienan, Belgium). A total of 34 single-
molecule real-time (SMRT) cars were run on the PacBio RS II system and 5 cars on the Sequel system using P6/C4 chemistry. We generated a total of 80.01 Gb of clean reads (Supplementary Table 1).

Illumina sequencing libraries were prepared according to the Illumina protocols. Briefly, 1 μg of genomic DNA was fragmented by Covaris. The fragmented DNA was repaired, and the base adenine was ligated to the 3′ end. Illumina adapters were then ligated to the fragments, and the proper samples were selected. The size-selected product was PCR-amplified, and the final product was validated. The Agilent Bioanalyzer 2100 was used to estimate the end (PE) and 3–20 kb mate-pair (MP) libraries and 500-bp PE using the HiSeq™ 2500 and MiSeq platforms (Illumina, San Diego, USA), respectively. Finally, we generated a total of 577.93 Gb of clean reads for the 200- and 500-bp PE and 3-, 5-, 10-, and 20-kb MP libraries (Supplementary Table 1).
The ten BAC clones were completely sequenced using 454 Life Sciences GS FLX System (GS FLX) and ABI 3730xl DNA Analyzer. Analyzed sequencing data were assembled using Newbler v2.8 (https://www.ncbi.nlm.nih.gov/assembly/GCA_000507345.1/) and used to create contigs or scaffolds. To fill the gap of sequences, we used primer walking. The primer walking method has been widely used in genome sequencing projects to determine the order of contigs and connect the remaining sequence gaps between the contigs. This way, a draft sequence for ten BAC clones was created. Finally, completed BAC clone sequences were checked by using the HISeq sequence data for sequence error correction. To validate the genome assembly obtained by FALCON, we performed an all-by-all alignment (-minIdentity = 80–99, -minScore = 100, -fastMap) of the 10 complete BACs and the assemblies using BLAT v3.2.4 (ref. 65).

Genotype-by-sequencing linkage analysis. Genomic DNA was extracted from the two parents by Myogenius (voucher number: IT99788) and S. tora (voucher number: IT104602)) and 153 F2 progeny using a Qiagen plant DNAeasy kit. Two genotype-by-sequencing (GBS) libraries were prepared using ApekI restriction enzyme as described in Elshire et al. 69. The GBS libraries (74 F2 individuals + two parents; 79 F2 individuals + two parents) were sequenced on an Illumina HiSeq2500 system. Low-quality bases and adapter sequences were trimmed using Trimmomatic v0.36 (ref. 66) and the trimmed reads from each sample were mapped to the S. tora draft assembly using BWA-MEM63. HaplotypeCaller in GATK39 was used to call single-nucleotide polymorphisms (SNPs) and generate a raw vcf file. High-quality biallelic SNPs were selected using VCFtools68 with the following parameters: --minMAPQ 30, --minMinor 0, --minMarg 0, -in hardref. The draft contig assemblies were used for genome assembly obtained by FALCON, we performed an all-by-all alignment (-minIdentity ≥ 20, and (3) missing genotype ≤ 30%. The SNP positions that showed polymorphic homologous SNPs between the parents were retained for linkage analysis. Linkage analysis was conducted using QTL IciMapping v4.1 (ref. 65) with the Kosambi function.

A total of 8,721,8 million raw PE reads were generated from two Apeki GBS libraries, and 372 million trimmed PE reads were used for subsequent linkage analysis. Of those, 89.8% reads were mapped to the S. tora reference assembly and 88.6% (329.5 million reads) were concordantly mapped, which was representing about 2.1 million properly mapped PE reads per sample. The GATK HaplotypeCaller called 2,287,678 and 4,778 million raw consensus and trimmed variants from libraries 1 and 2, respectively. After low-quality SNPs were filtered, 7,584 and 15,604 high-quality SNPs were obtained from libraries 1 and 2, respectively, and 5,071 markers were commonly represented in both. Three genetics maps independently constructed with three sets of SNP markers (from libraries 1 and 2) were evaluated. Two sets of SNPs (two parents) were used in a number of anchor contigs and genome representation, and the map that used common markers between libraries 1 and 2 was selected for further analysis. This map contained 2,654 nonredundant markers representing 3,587 cM within 12 linkage groups (LG1 contained only one marker).

With this linkage map, we tried to regrupo markers by increasing group number parameters from 13 to 25; however, the efforts were not successful to make the 13th linkage group. Finally, the linkage map was compared to pseudochromosomes constructed by Hi-C and we were able to split LG5 into two groups: one with three contigs (164 markers covering 34.4 Mb) and the other with 8 contigs (263 markers covering about 44 Mb). The final S. tora genetic map with Hi-C in LG5 had 13 linkage groups with 4,455 markers spanning 2,780 cM of genetic distance (Supplementary Data 8). It enabled to anchor 111 contigs (contig 3 split into two contigs: c31-1 and c31-2) to 13 linkage groups, which represented about 401 Mb of S. tora sequence assembly (Supplementary Table 17). Genetically, LG8 was the longest linkage group (347.5 cM) followed by LG13 (343 cM) and LG5 (331 cM). Whereas, 487 markers anchored about 45 Mb of sequences in LG5, which was the longest anchored chromosome. Physical distance per genetic distance was calculated as 144 kb/Mc on average across the whole genome (Supplementary Fig. 7).

Identification of IncRNA. A pipeline for IncRNA identification was designed according to a previous study10. In brief, among the total transcripts obtained from reference-guided assembly of transcriptome data, transcripts with open-reading frame (ORF) for ≥100 amino acids and ≥200 nucleotides were removed. We also removed sequences with homology to protein sequences based on BLAST search against SwissProt67 and Panther databases68. The remaining sequences were used in Coding Potential Calculator (CPC)30 and transcripts with CPC score ≥ 1.0 were removed as CPC scores between 1 and 1 are “weak noncoding” or “weak coding”. From the remaining transcripts, fluoride-eating RNAs (URNA, RNAs, mRNA transcripts) were removed by comparing with RNAcentral database91 sequences (cutoff E value of 1e–10), and those completely matching with S. tora reference protein-coding gene sequences were also removed. Finally, we only retained the longest isoform for each gene to obtain the final set of 3,278 IncRNAs (Supplementary Data 1).

Phylogenetic tree construction and evolution-rate estimation. To understand the evolutionary patterns of the S. tora genome and gene families, we performed comparative analysis using genome assembly sequences of S. tora (v2.09) and seven different stages of seeds, suggesting that the assembly includes most of the S. tora gene space (Supplementary Table 20).

Phylogenetic tree construction and evolution-rate estimation. To understand the evolutionary patterns of the S. tora genome and gene families, we performed comparative analysis using genome assembly sequences of S. tora (v2.09) and seven different stages of seeds, suggesting that the assembly includes most of the S. tora gene space (Supplementary Table 20).

Identification of IncRNA. A pipeline for IncRNA identification was designed according to a previous study10. In brief, among the total transcripts obtained from reference-guided assembly of transcriptome data, transcripts with open-reading frame (ORF) for ≥100 amino acids and ≥200 nucleotides were removed. We also removed sequences with homology to protein sequences based on BLAST search against SwissProt67 and Panther databases68. The remaining sequences were used in Coding Potential Calculator (CPC)30 and transcripts with CPC score ≥ 1.0 were removed as CPC scores between 1 and 1 are “weak noncoding” or “weak coding”. From the remaining transcripts, fluoride-eating RNAs (URNA, RNAs, mRNA transcripts) were removed by comparing with RNAcentral database91 sequences (cutoff E value of 1e–10), and those completely matching with S. tora reference protein-coding gene sequences were also removed. Finally, we only retained the longest isoform for each gene to obtain the final set of 3,278 IncRNAs (Supplementary Data 1).
Markov graph clustering (MGL v1.4–137) algorithm, through all-vs-all sequence similarity search by BLASTP v2.2.29–39 with an E-value cutoff of 1e-3. The orthologous genes in each homomorphic cluster are used to map genes onto multiple-sequence alignment with MAFFT v7.305b39 and the alignments were corrected with Gblocks v0.91b.40 The phylogenetic tree was reconstructed using IQ-Tree v1.6.0.2–68, using a maximum likelihood method with 1,000 bootstrap iterations. Here, the longest protein in each genome was selected among the proteins in each orthologous cluster. From the trees, the gene changes such as contraction and expansion were observed among the genomes using CAFE v3.1 method.41 Rapid expansion/contraction is indicated by statistically significant non-random expansion/contraction at P < 0.01, as described in CAFE.42 The evolutionary divergence timescale of the species was obtained from the clock and Yule model with the JTT substitution model (the gamma category count set to 4), which was implemented in BEAST2 method.43 The calibration priors were set as 58–70 MYA for the common ancestor of S. tora and Mimosa pudica, do not have the soybean-specific WGD event (Supplementary Fig. 27).44 Homology analysis with 6,310 orthologous genes shared by S. tora and 15 other green plant species was used to construct a phylogenetic tree based on a concatenated sequence alignment using MAFFT.45 In this phylogenetic tree, S. tora was, as expected, clustered with other legume crops, although the evolutionary distance from S. tora to Papilionoideae such as soybean, Medicago truncatula, and chickpea was relatively large (Supplementary Fig. 11). The phylogenetic tree confirmed the grouping of Caesalpinioideae species such as S. tora and M. pudica. The first divergence time between Caesalpinioideae and Papilionoideae was estimated to be approximately 81.9–93.6 MYA (Supplementary Fig. 11). Furthermore, Senna and Chamaecrista genera diverged from the Mimosoid clade (Fairstherbia albida and Mimosa pudica) –59.4–66.5 MYA. (Supplementary Fig. 11).

Enzyme prediction. S. tora enzymes and metabolic pathways were predicted using the Plant Metabolic Network (PMN)’s pipelines.46 This process starts by annotating amino acid sequences with the Ensemble Enzyme Prediction Pipeline (E2P) version 2.0.47 Protein domains and assignments are then aligned to a database of known enzymes called Reference Protein Sequence Database (RPSD) version 4.2.48 Metabolic pathways were predicted using PathLogic software, which is part of the Pathway Tools v23.5 package from SRI.49 The predicted pathways were further refined using PMN’s Semi-Automated Validation Infrastructure (SAVI) pipeline.50 Deep learning and machine learning algorithms were then applied on manual curations. In total, 6,159 enzymes and 442 metabolic pathways were predicted in S. tora into a database called StoraCyc, available online at https://plantcyc.org.

To identify enriched metabolic pathways among expanded or rapidly expanded gene families, we created the following four datasets: Dataset 1 included 15,921 genes from 2,874 families that were expanded or rapidly expanded in S. tora. Dataset 2 included 10,571 genes from 1,775 families that were exclusively expanded or rapidly expanded in S. tora. Dataset 3 included 7,382 genes from 411 families that were exclusively expanded or rapidly expanded in S. tora, as expected. The background used in this enrichment analysis was all genes in S. tora. P values of the enrichment analysis were calculated with a hypergeometric test using the phyper() function followed by multiple test correction using false-discovery rate (FDR) via Padjust(), both functions from the stats package version 3.6.2 R version 3.6.3. Significant enrichment was defined as an adjusted P-value < 0.01.

Primary metabolite profiling. Metabolome analysis was performed with 21 samples of frozen seed powders (~50 mg each) collected from seven seed developmental stages using the Capillary Electrophoresis Time of Flight Mass Spectrometry (CE-TOF-MS). CE-TOF-MS was run in two modes for cationic and anionic metabolites. Cationic metabolite levels were analyzed using a commercial fused silica capillary column (H3304-1002, HMT, Inc.; 150 μm x 80 cm) with a commercial cationic electrophoresis buffer (H3301-1001, HMT), or anionic electrophoresis column (H3301-1020, HMT). A commercial sheath liquid (H3301-1020, HMT) was delivered at a rate of 10 μL/min. Approximately 10 mL of sample solution was injected at a pressure of 50 mbar for 10 s, and applied capillary voltages were set at 27 kV (cathode mode) and 30 kV (anode mode), respectively. For both cationic and anionic separations, the electrophoretic voltage was held constant (500 V) and resuspended in 50 μL of ultrapure water immediately before the measurement. Anion trapping was achieved using a commercial fused silica capillary column (H3304-1002, HMT, 150 μm x 80 cm) with a commercial anionic electrophoresis buffer (H3301-1001, HMT) or anionic electrophoresis column (H3301-1020, HMT) as described in CAFE.51 Peaks detected in CE-TOF-MS were extracted using an automated integration software (MasterHands ver. 2.16.0.15 developed at Keio University) in order to obtain peak information including m/z, migration time (MT), and peak area. The peak detection limit was set at the signal-noise ratio (S/N) of 3. Signal peaks corresponding to isotopomers, adduct ions, and other product ions of known metabolites were excluded, and the remaining peaks were annotated with putative metabolites from the MasterHands database based on their MTs and m/z values.

Anthraxione extraction and analysis. S. tora seeds were collected and sorted into seven different ripening stages (Stage 1–Stage 7) depending on their size, color, and hardness. Classified seeds were ground with a mortar and pestle using liquid nitrogen to a fine powder and freeze-dried. Powdered samples (20 mg) were extracted with 1 mL of methanol using sonication for 30 min at 60 °C. After extraction, samples were centrifuged at 500 × g for 3 min at 25 °C, and the supernatant was filtered through a 0.2 μm Acrodisc MS syringe filters with PTFE membrane filters (Pall Corporation, Port Washington, NY, USA). Each sample was completely dried by EvaT-0200 Total Concentration System equipped with EvaS-3600 N2 generator (Gooyong engineering, Seoul, Korea), mixed with methanol, and filtered again with Acrodisc 0.2 μm MS syringe filter for liquid chromatography–mass spectrometry (LC–MS) analysis.

Quantitative analysis of anthraquinones was performed by a 3200 QTRAP mass spectrometer with a Turbo V ion source (AB Sciex, Ontario, CA, USA) coupled with a VANQUISH UHPLC system (Thermo Fisher Scientific, CA, USA) equipped with binary solvent manager, sample manager, column heater, and photodiode array detector. HPLC was performed on a ZORBAX Eclipse Plus column (150 × 4.6 mm, 5 μm) at room temperature (5 °C). The mobile phase consisted of 5% m/m ammonium acetate in water (reagent A) and 100% acetonitrile (reagent B). The gradient conditions were as follows: 0–1 min, 10% B; 1–4.5 min, 10–30% B; 4.5–8 min, 30–50% B; 8–11 min, 50–100% B; 11–14 min, 100% B. The flow rate was 0.5 mL/min and two microliters of samples were injected. For detecting peaks from test samples, MS parameter in ESI-negative mode was used as follows: nebulizing gas, 50 psi; heating gas, 50 psi; curtain gas, 20 psi; desolvation temperature, 550 °C; ion-spray voltage floating, 4.5 kV. The data obtained from MRM mode were quantitated using MultiQuant 3.0.2 software (AB SCIEX).

RNA sequencing and analysis. Total RNA was isolated from seven developmental stages of seeds (Stage 1–Stage 7) (Supplementary Table 2). RNA extraction and RNA-Seq library preparations were performed, and RNA-Seq libraries were sequenced on the Illumina NovaSeq 6000 (Illumina, San Diego, CA). RNA-Seq data with high-quality bases (PHRED score (Q) < 20) and adaptor contamination were removed by Trimmomatic v0.36 using the parameters "ILLUMINACLIP:TruSeq3-SE:2:30:10 LEADING:3 SLIDINGWINDOW:4/15 MINLEN:36". After checking for quality scores and read lengths, RNA-Seq reads were mapped to S. tora genome using STAR-2.5.0a with default parameters (ARCs, 0.5; Expectation Maximization, 7; RSEM-1.3.1).109 method was used to obtain the expression value for each gene in the genome (Supplementary Data 9). The read counts estimated by RSEM were subset to read*g < 0.005. After mapping, these reads were classified into different expression classes along with the statistical significance based on the FDR. Furthermore, we applied the standard filters, i.e., genes per million (TPM) ≥ 0.3, read counts ≥ 5, and log2-fold changes ≥ 1 or ≤ –1 to derive the final list of differentially expressed genes.110 Finally, the expressed genes (i.e., TPM ≥ 0.3 and read count ≥ 5) were included to show the different expression patterns during seed development. An in-house R script was used to generate the heatmap.

To analyze gene expression patterns of metabolic genes during seed development, all expressed (TPM > 0) S. tora genes predicted to catalyze small-molecule metabolism in StoraCyc were mapped to StoraCyc’s metabolic domains
Homogenates were centrifuged at 13,475 × g. Cells were harvested by centrifugation, washed twice with 100 mM phosphate buffer saline (pH 7.5). An identical reaction mixture containing the same amount of nucleotide triphosphate was used for protein production. Cultures were induced at 13,475 × g for 3 min.

To identify enriched StoraCyc metabolic domains in clusters 3 and 6, we computed the metabolic domain annotations of genes annotated to those clusters to those annotated to all genes in S. tora. P values of the enrichment analysis were calculated with a hypergeometric test using phyper() followed by multiple test correction using False Discovery Rate via p.adjust(), both functions from the stats package version 3.6.2 in R version 3.6.3. Significant enrichment was defined as an adjusted P value ≤ 0.01. To avoid biases introduced by datasets containing a small number of genes, we defined a minimum threshold of at least ten genes present in each metabolic domain per cluster in order for the domain to be considered significantly enriched. The bubble plot was generated using ggplot2 version 3.3.2 in R version 3.6.3.

**References**

1. World Health Organization. Model List of Essential Medicines, 21st List 2019 (WHO, Geneva, 2019).
2. Manoharan, R. K., Lee, J.-H., Kim, Y.-G. & Lee, J. Alizarin and chrysazin inhibit biofilm and hypophilic adhesion by Candida albicans. Front. Cell Infect. Microbiol. 7, 447–447 (2017).
3. Parvez, M. K. et al. The anti-hepatitis B virus therapeutic potential of anthraquinones derived from Aloe vera. Phytother. Res. 33, 2960–2970 (2019).
4. Roa-Linares, V. C. et al. Anti-herpetic, anti-dengue and antineoplastic activities of simple and heterocycle-fused derivatives of terpenyl-1,4-naphthoquinone and 1,4-anthraquinone. Molecules 24, 1279 (2019).
5. Wang, Q.-W. et al. Anti-influenza A virus activity of rhein through regulating oxidative stress, TLR4, Akt, MAPK, and NF-κB signal pathways. PLoS ONE 13, e0191793 (2018).
6. Dhananjayan, M. R., Milev, Y. P., Kron, M. A. & Nair, M. G. Synthesis and activity of substituted anthraquinones against human filarial parasite, Brugia malayi. J. Medicinal Chem. 48, 2822–2830 (2005).
7. Li, Y. & Jiang, J.-G. Health functions and structure-activity relationships of natural anthraquinones from plants. Food Funct. 9, 6063–6080 (2018).
8. Ravi, S. K., Narasingappa, R. B., Prasad, M., Javagal, M. R. & Vincent, B. Antioxidant potential of Aloe vera L. on type 2 diabetes mellitus. Food Funct. 9, 439–444 (2018).
9. Chaube, M. & Kapoor, V. P. Structure of a galactomannan from the seeds of Cassia alata L. J. Medicinal Chem. 30, 439–439 (2018).
10. Choudhri, P. et al. De novo sequencing, assembly and characterisation of Aloe vera transcriptome and expression of genes related to saponin and anthraquinone metabolism. BMC Genomics 19, 427 (2018).
11. Mahesh, V. K., Sharma, R., Singh, R. S. & Upadhya, S. K. Anthraquinones and kaempferol from Cassia species section fistulae. J. Nat. Products 47, 733–733 (1984).
12. Boy, H. I. A. et al. Recommended medicinal plants as source of natural products: a review. Digital Chin. Med. 1, 131–142 (2018).
13. Aloe vera bioactivity and expression of genes related to saponin and anthraquinone metabolism. BMC Genomics 19, 427 (2018).
14. Kang, S.-H. et al. De novo transcriptome sequence of Senna tora provides insights into anthraquinone biosynthesis. PLoS ONE 15, e0225564 (2020).
16. Mehta, R. H., Ponnuchamy, M., Kumar, J. & Reddy, N. R. R. Exploring drought stress-regulated genes in senna (Cassia angustifolia Vahl.): a transcriptomic approach. Funct. Integr. Genomics 17, 1–25 (2017).

17. Deng, Y. et al. Full-length transcriptome survey and expression analysis of Cassia obtusifolia to discover putative genes related to aurantio-obtusin biosynthesis, seed formation and development, and stress response. Int. J. Mol. Sci. 19, 2476 (2018).

18. Chiang, Y.-M. et al. Characterization of the Aspergillus nidulans monodictyophene gene cluster. Appl. Environ. Microbiol. 76, 2067–2074 (2010).

19. Shamim, G., Ranjan, K. S., Pandey, M. D. & Ramani, R. Biochemistry and biosynthesis of insect pigments. Eur. J. Entomol. 111, 149–164 (2014).

20. Duval, J., Pecher, V., Poujol, M. & Leselier, E. Research advances for the extraction, analysis and uses of anthraquinones: a review. Ind. Crops Products 94, 812–833 (2016).

21. Javid, R. & Qazi, U. Y. Catalytic oxidation process for the degradation of synthetic dyes: an overview. Int. J. Environ. Res. Public Health 16, 2066 (2019).

22. Tkaczyk, A., Mitrowska, K. & Posyniak, A. Synthetic organic dyes as synthetic dyes: an overview. Tetrahedron Lett. 9, 1395–1396 (2018).

23. Bauch, H. J. & Leistner, E. Aromatic metabolites in cell suspension cultures of Galium mollugo. Planta Med. 33, 105–123 (1978).

24. Leistner, E. Isolation, identification, and biosynthesis of anthraquinones in cell suspension cultures of Morinda citrifolia (author’s thesis). Planta Med. 28, 214–224 (1975).

25. Leistner, E. in Medicinal and Aromatic Plants VIII. Biotechnology in Agriculture and Forestry Vol. 33 (ed. Bajaj, Y. P. S.) 296–307 (Springer, 1995).

26. Leistner, E. & Zenk, M. H. Mevalonic acid a precursor of the substituted naphthoquinone and its related anthraquinone in a biosynthetic pathway. Anemone: animal or bacterial origin? Mitochondrial DNA Part B 1728 (2007).

27. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).

28. Anguraj Vadivel, A. K., Krysiak, K., Tian, G. & Dhaubhadel, S. Genome-wide identification and localization of chalcone synthase family in soybean (Glycine max [L.] Merr). BMC Plant Biol. 18, 325–325 (2018).

29. Pandith, S. A. et al. Cloning and functional expression of three type II PKS gene clusters from Streptomyces bottropensis. ChemBioChem 13, 224–230 (2012).

30. Zhang, C. et al. Biosynthetic Baeyer–Villiger chemistry enables access to two anthracene scaffolds from a single gene cluster in deep-sea-derived Streptomyces olivaceus SC50 T05. J. Nat. Products 81, 1570–1577 (2018).

31. Leistner, E. & Zenk, M. H. A comprehensive map of the anthraquinone biosynthesis of Photorhabdus luminescens. ChemBioChem 8, 1721–1728 (2007).

32. Zhou, Q. et al. Molecular mechanism of polyketide shortening in anthraquinone biosynthesis of Photorhabdus luminescens. Chem. Sci. 10, 3941–3949 (2019).

33. Allen, G. C., Flores-Vergara, M. A., Krasynska, S., Kumar, S. & Thompson, W. F. A modified protocol for rapid DNA isolation from plant tissues using cetrimethylammonium bromide. Nat. Protoc. 1, 2320–2325 (2006).

34. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimming tool for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014).

35. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. An efficient trimmer for de novo assembly based on chromatin interactions. Gigascience 3, 8 (2014).

36. Muziuchi, Y. et al. Novel type III polyketide synthases from Aloe barbescens. J. Agric. Food Chem. 57, 2437–2441 (2009).

37. Abe, I., Oguro, S., Tsutsui, Y., Sano, Y. & Noguchi, H. Engineered biosynthesis of plant polyketides: chain length control in an octaketide-producing plant type III polyketide synthase. J. Am. Chem. Soc. 127, 12709–12716 (2005).

38. Pillai, P. P. & Nair, A. R. Hypericin biosynthesis in Hypericum hookerianum Wight and Arn: investigation on biochemical pathways using metabolite inhibitors and suppression subtractive hybridization. C. R. Biol. 337, 571–580 (2014).

39. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212 (2015).

40. Aitshul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).

41. Anguraj Vadivel, A. K., Krysiak, K., Tian, G. & Dhaubhadel, S. Genome-wide identification and localization of chalcone synthase family in soybean (Glycine max [L.] Merr). BMC Plant Biol. 18, 325–325 (2018).

42. Pandith, S. A. et al. Functional promiscuity of two divergent paralogs of type III plant polyketide synthases. Plant Physiol. 171, 2599–2619 (2016).

43. Anderssen-Ranberg, J. et al. Synthesis of C-glucosylated octaketide anthraquinones in Nicotiana benthamiana by using a multispecies-based biosynthetic pathway. ChemBioChem 18, 1893–1897 (2017).
77. Varshney, R. K. et al. Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. Nat. Biotechnol. 30, 83–89 (2012).

78. Kang, Y. J. et al. Genome sequence of mungbean and insights into evolution within Vigna species. Nat. Commun. 5, 5443 (2014).

79. Varshney, R. K. et al. Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. Nat. Biotechnol. 31, 240–246 (2013).

80. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7, 562–578 (2012).

81. Haas, B. J. et al. Improving the Arabidopsis genome annotation using maximal exact matches data. Nucleic Acids Res. 37, 5654–5666 (2009).

82. Stankiewicz, M., Schoffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalised hidden Markov model that uses hints from external sources. BMC Bioinformatics 7, 62 (2006).

83. Blanco, E., Parra, G. & Guigó, R. Using geneid to identify genes. Curr. Protoc. Bioinformatics 4, 4.3 (2007).

84. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6, 31 (2005).

85. Götz, S. et al. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36, 3420–3438 (2008).

86. Jones, P. et al. InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240 (2014).

87. Jayakodi, H. et al. Genome-wide characterization of long intergenic non-coding RNAs (lincRNAs) provides new insight into viral diseases in honey bees Apis cerana and Apis mellifera. BMC Genomics 16, 680 (2015).

88. Consortium, T. U. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 47, D506–D515 (2018).

89. El-Gebali, S. et al. The Pfam protein families database in 2019. Nucleic Acids Res. 47, D277–D342 (2018).

90. Kong, L. et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. 35, W345–W349 (2007).

91. The RNAcentral Consortium. RNACentral: a hub of information for non-coding RNA sequences. Nucleic Acids Res. 47, D1250–D1251 (2019).

92. Li, L., Stoeckert, C. J. Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178–2189 (2003).

93. Kato, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780 (2013).

94. Cestresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552 (2000).

95. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274 (2015).

96. Han, M. V., Thomas, G. W. C., Lugo-Martinez, J. & Hahn, M. W. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFe 3. Mol. Biol. Evol. 30, 1987–1997 (2013).

97. Bouckaert, R. et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 15, e1006650 (2019).

98. Zhang, Z. et al. ParaAT: a parallel tool for constructing multiple protein-coding ORF (STO07G228250) discovered in this study. The remaining authors declare no competing interests.

99. Additional information Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-19681-1. Correspondence and requests for materials should be addressed to S.-H.K., S.Y.W. or T.J.O.

100. Peer review information Nature Communications thanks Joe Chappell, Jerome Gouzy, and Takayuki Tohe for their contribution to the peer review of this work. Peer reviewer reports are available.

101. Reprints and permission information is available at http://www.nature.com/reprints

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

© The Author(s) 2020, corrected publication 2021