

Draft Genome of *Tanacetum Coccineum*: Genomic Comparison of Closely Related *Tanacetum*-Family Plants

Takanori Yamashiro

Dainihon Jochugiku Co., Ltd

Akira Shiraishi

Suntory Foundation for Life Sciences

Koji Nakayama

Dainihon Jochugiku Co., Ltd

Honoo Satake (✉ satake@sunbor.or.jp)

Suntory Foundation for Life Sciences

Research Article

Keywords:

Posted Date: January 13th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1133616/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

The plant *Tanacetum coccineum* (painted daisy) is closely related to *Tanacetum cinerariifolium* (pyrethrum daisy). However, *T. cinerariifolium* produces large amounts of pyrethrins, a class of natural insecticides, whereas *T. coccineum* produces much smaller amounts of these compounds. Thus, comparative genomic analysis is expected to contribute a great deal to investigating the difference in biological defense systems, including pyrethrin biosynthesis. Here, we elucidated the 9.4-Gb draft genome of *T. coccineum*, consisting of 2,836,647 scaffolds and 103,680 genes. Comparative analyses of the draft genome of *T. coccineum* and that of *T. cinerariifolium*, generated in our previous study, revealed distinct features of *T. coccineum* genes. While the *T. coccineum* genome contains more numerous ribosome-inactivating protein (RIP)-encoding genes, the number of higher-toxicity type-II RIP-encoding genes is larger in *T. cinerariifolium*. Furthermore, the number of histidine kinases encoded by the *T. coccineum* genome is smaller than that of *T. cinerariifolium*, suggesting a biological correlation with pyrethrin biosynthesis. Moreover, the flanking regions of pyrethrin biosynthesis-related genes are also distinct between these two plants. These results provide clues to elucidation of species-specific biodefense systems, including the regulatory mechanisms underlying pyrethrin production.

Introduction

Several plants of the genus *Tanacetum* biosynthesize plant-specific metabolites, called pyrethrins, that are neurotoxic to insects. Due to their highly selective toxicity to insects and ready decomposition in the presence of sunlight and oxygen, pyrethrins are widely used as commercial insecticides in products such as mosquito coils. Pyrethrins comprise six ester compounds: pyrethrin I and II, jasmolin I and II, and cinerin I and II. One species of *Tanacetum*, *T. cinerariifolium*, has been shown to produce pyrethrins exclusively in its ovaries, and a portion of the biosynthetic pathway for pyrethrins has been elucidated¹.

Previously, we elucidated the draft genome of *T. cinerariifolium*, the high pyrethrin-producing species². *T. cinerariifolium* is known to upregulate pyrethrin biosynthesis by emitting a specific combination of volatile organic compounds (VOCs) when the plant body is wounded³. Interestingly, our genome analysis demonstrated that *T. cinerariifolium* encodes multiple histidine kinases, including gas-induced-type histidine kinases related to ethylene receptors, suggesting a functional correlation between histidine kinases and pyrethrin biosynthesis². Furthermore, *T. cinerariifolium* possesses a greater number of ribosome-inactivating protein (RIP) -encoding genes than other plants, the apparent result of gene multiplication; in contrast, the number of genes encoding endochitinases, which are involved in defense against fungal pathogens, is smaller than those of other plants. These features of defense-related genes suggest that *T. cinerariifolium* has undergone specific evolutionary adaptation for protection against natural predators in the dry environment of this plant's area of origin, Dalmatia².

Another member of the same genus, *T. coccineum*, produces considerably lower amounts of pyrethrins, despite phylogenetically high relatedness to *T. cinerariifolium*⁴ (Fig. 1). Moreover, *T. coccineum* is native to a region, extending from Persia to the Caucasus, which is hot and humid. In contrast, *T. cinerariifolium* originated from Dalmatia, a region that is relatively dry. Such climatic differences suggest that these two plant species may have acquired distinct defense strategies against their natural enemies that vary according to the environment. Comparison between the genomes of these two species is expected to reveal the evolutionary processes and diversification of defense strategies, and to contribute to the development of technologies for more efficient production of pyrethrins.

In the present study, we elucidated the draft genome of *T. coccineum* and performed comparative genomic analyses between these two related species.

Results And Discussion

Sequencing of the *T. coccineum* genome

Paired-end (PE) libraries and mate-pair (MP) libraries (with 3-, 5- and 8-kb insert sizes; MP-3kb, MP-5kb, and MP-8kb, respectively) of the *T. coccineum* genome were generated and then sequenced using Illumina NovaSeq 6000 instruments. Miseq (MS) libraries were generated, then sequenced, using the Illumina MiSeq system. For long reads, a PacBio (PB) library was generated and sequenced using PacBio Sequel II system. The total base-counts of the sequence reads for PE, MP-3kb, MP-5kb, MP-8kb, MS, and PB amounted to 854 Gb, 99 Gb, 108 Gb, 109 Gb, 27 Gb, and 93 Gb, respectively (Table 1).

Table 1
Statistics of sequence reads

Library	Insert size (bp)	Read length (bases)	Number of reads	Total read length (bases)
PE	350	151	5,732,398,372	854,270,829,961
MP-3kb	3000	151	698,859,570	99,096,491,543
MP-5kb	5000	151	750,513,382	107,931,325,410
MP-8kb	8000	151	762,709,822	109,359,355,641
MS	550	301	97,731,712	26,503,089,921
PB		Ave. 10,738	8,670,092	93,100,193,428
Read lengths for Illumina NGS (PE, MP, MS) were indicated in max length, and PacBio read (PB) was indicated in average length. PE: paired-end; MP: mate-pair; MS: Miseq; PB: PacBio; NGS, next-generation sequencing.				

Size estimation of the *T. coccineum* genome

Prior to genomic DNA assembly, we estimated the 1C DNA content of *T. coccineum* by flow cytometry, using the *Chrysanthemum seticuspe* genome (cultivar Gojo-0, 3 pg/1C)⁵ as an internal standard. The estimated *T. coccineum* size was 9.4 pg/1C (Supplemental Fig. 1A), which corresponds to 9.4 Gb. This size is approximately 1.6 times larger than the 5.8 pg/1C obtained in a previous report in which the size was measured by Feulgen densitometry genome estimation⁶. To further validate the 9.4 Gb genome size, we performed a k-mer spectrogram analysis of the Illumina short reads of the *T. coccineum* genome (Supplemental Fig. 1B) using Jellyfish⁷. The k-mer spectrograms showed two major distributions: one with a maximum coverage of 1 and the other with a multimodal distribution with a maximum coverage of 44. Since the distribution with the maximum coverage=1 was considered to be caused by sequencing errors, we treated the data after coverage=11, which is the minimum value between the two distributions, as the k-mer derived from the correct genome. In accordance with previous studies⁸, the genome size estimated from the k-mer distribution with a maximum of 44 for coverage=11 and above was 9.8 Gb. These analyses led to the conclusion that the *T. coccineum* used in the present study has a genome size of approximately 9 Gb.

Sequence assembly and annotation of the *T. coccineum* genome

The reads obtained from next-generation sequencing (NGS) were subjected to contig assembly and scaffolding to estimate the genome sequence (Fig. 2). A collection of 6,500,576 contigs with a total length of 8.57 Gb (Table 2) was constructed by assembling PE reads and MS reads using SOAPdenovo⁹. The assembled contigs then were scaffolded with PB and MP reads using SSPACE^{10,11}, which concatenates contig sequences, and the Gapfiller¹² and TGS-Gapcloser programs¹³, which fill the inter-contig unknown bases ('N's) in scaffolds with 'A/T/G/C' (Fig. 2), as described in our previous report. Since the accuracy of gap-filled sequences depends on PB reads that exhibit lower sequence accuracy than PE reads, the scaffold sequences were polished using POLCA¹⁴. The total length of the resultant scaffolds was 9.46 Gb, which corresponded well with the flow cytometry-estimated genome size for *T. coccineum* (Supplemental Fig. 1). The N50 value of the scaffolds was 27.8 Kb, and the maximum length of the scaffolds was 331 Kb (Table 2). Subsequently, the draft genome was subjected to analysis by AUGUSTUS, resulting in the prediction of 1,582,136 putative genes.

Table 2
Statistics of genome assembly.

	Contigs	Scaffolds
Total number of sequence fragments	6,500,576	2,836,647
Total length (bp)	8,565,698,618	9,463,677,832
N50 (bp)	8,465	27,784
Length of longest contig (bp)	149,916	331,286
GC content (%)	34.9	35.1

Completeness of the draft genome sequences was evaluated using BUSCO¹⁶, which counts complete (C), fragmented (F), and missing (M) conserved genes in genome sequences. Sequence analysis of 1614 conserved core plant genes confirmed that 97.8% of conserved genes (92.7% as complete and 5.1% as fragmented) were present in the *T. coccineum* genome assembly (Table 3). These scores indicated quality as high as that obtained for *T. cinerariifolium*², and we therefore used the *T. coccineum* draft genome for subsequent analysis.

Table 3
Annotation statistics for draft genome.

Number of predicted genes	1,582,136
BUSCO v5	C: 92.7% (Single: 70.8%, Duplicated: 21.9%) F: 5.1% M: 2.2%
Number of predicted TEs	772,794
Number of predicted genes encoding products with known protein signatures	103,680
*C: percentage of full-length conserved genes in BUSCO notation; F: percentage of fragmented conserved genes in BUSCO notation; M: percentage of missing genes in BUSCO notation; TE: transposable element.	

Since a larger number of transposable elements (TEs) is detected than that of functional genes in the genomes of plants, including *T. cinerariifolium*², we first analyzed the TE component of the assembled *T. coccineum* genome. TEs were detected and annotated using hmmpfam against the Gypsy database (GyDB)¹⁷, revealing the presence of 772,794 TEs. The value of 82,212 TEs/Gb in the *T. coccineum* genome was slightly larger than the 73,957 TEs/Gb observed in the *T. cinerariifolium* genome². Furthermore, the predicted genes were subjected to InterProScan to provide high-confidence annotation, revealing the presence of 103,680 putative genes encoding products that exhibited known protein signatures. Thus, a high-quality 9.4-Gb *T. coccineum* draft genome was assembled and shown to include a total of 772,794 TEs and 103,680 plausible genes based on 854 Gb of PE reads, 316 Gb of MP reads, 26.5 Gb of MS reads, and 93.1 Gb of PB reads.

Inter-genus comparative analysis of TE classification

We analyzed the ratio of each TE clade against all TE regions in the genomes of *T. coccineum*, *T. cinerariifolium*, *C. seticuspe*, *Artemisia annua*, *Helianthus annuus*, *Nicotiana tabacum*, *Oryza sativa*, and *Arabidopsis thaliana*. *T. cinerariifolium*, *C. seticuspe*, *A. annua*, and *H. annuus* (which belong to the Asteraceae family), and *N. tabacum*, *O. sativa*, and *A. thaliana* (which are model organism) were used as described in the previous study². The top-5 clade ratios of each plant are shown in Table 4. In *T. coccineum*, members of the *sire*-clade TEs were the most abundant TE clade, which was also observed in the three Asteraceae plants (*T. cinerariifolium*, *C. seticuspe*, and *A. annua*). In the *T. cinerariifolium* genome, the second-largest ratio of clades was *athila*, followed (in order) by *del*, *oryco*, and *lentiviridae*; in the *T. coccineum* genome, the second-largest clade was *del*, followed (in order) by *athila*, *oryco*, and *tork*. These results suggested that the *del*- and *tork*-clade TEs multiplied after evolutionary divergence of *T. cinerariifolium* and *T. coccineum* from a common ancestor.

Table 4
The top-5 most-abundant TE clades in each species.

Rank	Tco	Tci	Cs	Aa	Ha	Nt	Os	At
1	sire (25.7)	sire (33.0)	sire (32.0)	sire (21.8)	del (37.7)	del (40.4)	tat (11.4)	athila (9.54)
2	del (15.3)	athila (17.0)	athila (10.9)	athila (19.6)	sire (9.85)	tat (20.5)	retroviridae (8.97)	retroviridae (4.89)
3	athila (12.5)	del (12.0)	oryco (5.11)	del (6.57)	lentiviridae (8.72)	athila (9.87)	del (8.39)	caulimovirus (4.15)
4	oryco (7.25)	oryco (6.34)	lentiviridae (5.06)	oryco (4.59)	tat (6.76)	sire (3.02)	tork (4.73)	badnavirus (4.05)
5	tork (4.70)	lentiviridae (4.92)	del (5.03)	tork (4.01)	athila (5.17)	tork (2.80)	alpharetroviridae (4.66)	tork (3.08)
*Parenthesized numbers indicate the ratio (%) of each clade against total TE regions in that species. Tco: <i>T. coccineum</i> ; Tci: <i>T. cinerariifolium</i> ; Cs: <i>C. seticuspe</i> ; Aa: <i>A. annua</i> ; Ha: <i>H. annuus</i> ; Nt: <i>N. tabacum</i> ; Os: <i>O. sativa</i> ; At: <i>A. thaliana</i> .								

To examine whether *del*- and *tork*-clade TEs had multiplied in a common ancestor of the Asteraceae or independently in individual genera, we estimated molecular phylogenetic trees of the reverse transcriptase (RT) domains encoded by the *del* and *tork* sequences and evaluated the number of co-clustered genes in single-genus clusters, as described in the previous study². In these molecular phylogenetic analyses, the TEs that multiplied in a common ancestor are positioned in orthologous clusters, while the TEs that multiplied after divergence from a common ancestor are positioned in clusters with the TEs from single plant species (multiplied clusters). The phylogenetic analysis revealed that 67%, 62%, 73%, 68%, and 86% of *del* TEs constituted multiplied clusters for *T. coccineum*, *T. cinerariifolium*, *C. seticuspe*, *A. annua*, and *H. annuus*, respectively (Fig. 3A). Likewise, 57%, 37%, 54%, 38%, and 71% of *tork* TEs were shown to be multiplied in the respective organisms (Fig. 3B). These results indicated that more than half of the *del* TEs and more than one-third of the *tork* TEs were multiplied in the individual genera, but the other TEs were conserved as common ancestor TEs within Asteraceae. Collectively, these results suggested that most of *del* and *tork* TEs were multiplied in the individual lineages of the respective *Asteraceae* genera, leading to the major TEs in *T. coccineum*.

Pyrethrin-related enzymes encoded in the *T. coccineum* genome

T. coccineum predicted proteins with high homology to 9 known *T. cinerariifolium* pyrethrin biosynthesis-related proteins (TciLOX1¹⁸, TciADH2¹⁹, TciALDH1¹⁹, TciCDS²⁰, TciGLIP²¹, TciJMH²², TciPYS²³, TciCCH²⁴, and TciCCMT²⁴.) were detected by BLASTP²⁵ (Table 5, Supplemental Fig. 2), indicating that a complete set of known pyrethrin-related enzymes is conserved in the *T. coccineum* genome.

Table 5

T. coccineum genome-encoded proteins corresponding to known pyrethrin biosynthesis-related proteins.

Known pyrethrin related enzymes	Corresponding proteins of <i>T. coccineum</i>	Protein sequence similarity
TciLOX1	Tco_0863779	Identities = 847/861 (98%), Positives = 853/861 (99%), Gaps = 0/861 (0%)
TciADH2	Tco_0487905	Identities = 340/378 (90%), Positives = 359/378 (95%), Gaps = 2/378 (1%)
TciALDH1	Tco_0682217	Identities = 448/499 (90%), Positives = 471/499 (94%), Gaps = 1/499 (0%)
TciCDS	Tco_1315810	Identities = 358/395 (91%), Positives = 374/395 (95%), Gaps = 0/395 (0%)
TciGLIP	Tco_1108878	Identities = 337/365 (92%), Positives = 348/365 (95%), Gaps = 0/365 (0%)
TciJMH	Tco_0572988	Identities = 450/512 (88%), Positives = 479/512 (94%), Gaps = 2/512 (0%)
TciPYS	Tco_1240348	Identities = 465/488 (95%), Positives = 475/488 (97%), Gaps = 0/488 (0%)
TciCCH	Tco_0360514	Identities = 470/498 (94%), Positives = 484/498 (97%), Gaps = 1/498 (0%)
TciCCMT	Tco_1190813	Identities = 358/374 (96%), Positives = 361/374 (97%), Gaps = 5/374 (1%)
The information of "Protein sequence similarity" column is introduced by BLASTP program with each known pyrethrin related enzymes as a query. Tci: <i>T. cinerariifolium</i> ; Tco: <i>T. coccineum</i> ; ADH2: alcohol dehydrogenase 2; ALDH1: aldehyde dehydrogenase 1; CCMT: 10-carboxychrysanthemol 10-methyltransferase; CDS: chrysanthemol diphosphate synthase; CHH: chrysanthemol 10-hydroxylase; GLIP: GDSL (Gly-Asp-Ser-Leu motif) lipase; JMH: jasmone hydroxylase; LOX1: 13-lipoxygenase; PYS: pyrethrolone synthase.		

Synteny analysis of genes encoding pyrethrin-related enzymes

The distribution of genes within the scaffolds that included loci encoding proteins corresponding to TciLOX1, TciADH2, TciALDH1, TciCDS, TciGLIP, TciJMH, TciPYS, TciCCH, and TciCCMT were analyzed using the GenomeJack software program. All genes were located on separate scaffolds, and TEs were located on the flanking regions of the genes encoding all pyrethrin-related enzymes, with the exception of the genes encoding TcoCCMT and TcoGLIP. The gene encoding the Tco_1190812 protein, which contains a Jacalin-like lectin domain, was located upstream of the gene encoding the TcoCCMT protein (Fig. 4A). A BLASTP search using the Tco_1190812 sequence as a query detected a predicted protein with sequence similarity (E-value of 3×10^{-93} ; 92.72% identity) to a segment of an *Artemisia annua* mannose-binding lectin (accession No. PWA73033.1). In the *T. cinerariifolium* genome, a gene encoding a corresponding Jacalin-like lectin (Accession No. GEW32189.1) also was found upstream of the locus encoding TciCCMT, suggesting that this synteny is conserved.

TciGLIP (i.e., the *T. cinerariifolium* GDSL (Gly-Asp-Ser-Leu motif) lipase) is the key enzyme in the final esterification of pyrethrin biosynthesis²¹. Syntenic analysis showed that open reading frames (ORFs) encoding putative GLIPs are present in the regions downstream of both the TciGLIP- and TcoGLIP-encoding genes. However, the *T. cinerariifolium* glutathione S-transferase-encoding gene (accession No. GEU71427.1) positioned upstream of *TciGLIP* and the hypothetical protein-coding gene positioned downstream of tandem GLIP-encoding genes in *T. cinerariifolium* are replaced by TEs in the *T. coccineum* genome (Fig. 4B), suggesting that this tandem GLIP-encoding locus translocated after the divergence of *T. coccineum* and *T. cinerariifolium*. Although the transcriptional regulatory mechanism of the TciGLIP gene is yet to be determined, this difference in the flanking region of these GLIP-encoding genes provides a clue to investigating the mechanisms regulating the possible differential expression of the genes encoding TciGLIP and TcoGLIP.

Functional annotation of the *T. coccineum* genes and inter-genus comparative analysis

Next, we investigated multiplication ratios of protein superfamilies in *T. coccineum* compared with those in other species, including *T. cinerariifolium*, which was described in our previous study. In brief, the predicted protein data sets of *T. coccineum*, *T. cinerariifolium*, *C. seticuspe*, *A. annua*, *H. annuus*, *N. tabacum*, *O. sativa*, and *A. thaliana* were subjected to analysis using InterProScan, and multiplication odds scores were calculated for each superfamily. A positive value for the multiplication odds score indicates that a genus possesses a higher number of multiplied genes in a given superfamily than other genera.

The highest and lowest multiplication odds scores for the biodefense-, signaling-, and metabolism-related were compared with those of other plants, including *T. cinerariifolium* (Table 6 and 7, respectively). To further compare *T. coccineum* with *T. cinerariifolium*, the multiplication odds scores for the superfamilies that were identified in the previous study² and not listed in the highest or the lowest table (Table 6 and 7, respectively) are shown in Table 8.

Table 6
Superfamilies with highest multiplication odds scores in *T. coccineum*.

Category	IPR ID	Superfamily name	Tco	Tci	Cs	Aa	Ha	Nt	Os	At
Biodefense	IPR036041	Ribosome-inactivating protein	1.96 (159)	1.29 (98)	-1.81 (7)	-1.00 (16)	-3.07 (0)	-3.07 (0)	-0.94 (17)	-3.07 (0)
Metabolism	IPR005848	Urease, alpha subunit	2.36 (108)	-0.14 (15)	-1.87 (1)	-1.46 (3)	-1.87 (1)	-0.87 (7)	-2.14 (0)	-1.87 (1)
Metabolism	IPR036226	Lipoxygenase, C-terminal domain	1.86 (232)	0.48 (86)	-0.22 (51)	-0.82 (32)	-1.12 (25)	-0.67 (36)	-1.86 (13)	-2.44 (7)
Metabolism	IPR033966	RuBisCO	1.60 (42)	-0.25 (8)	0.05 (11)	-0.15 (9)	-1.37 (1)	-0.95 (3)	-0.25 (8)	-1.15 (2)
Metabolism	IPR032466	Metal-dependent hydrolase	1.48 (166)	0.69 (94)	-1.13 (23)	-0.89 (28)	-0.32 (44)	-0.05 (54)	-1.61 (15)	-0.98 (26)
Metabolism	IPR036849	Enolase-like, C-terminal domain	1.38 (71)	0.91 (50)	-0.62 (14)	-0.55 (15)	-0.78 (12)	-0.29 (19)	-1.41 (6)	-1.29 (7)
Metabolism	IPR036396	Cytochrome P450	0.90 (1220)	0.19 (745)	0.16 (732)	0.07 (688)	-0.20 (568)	-0.12 (600)	-1.05 (314)	-0.85 (361)
Signaling	IPR035983	HECT, E3 ligase catalytic domain	1.22 (95)	0.84 (72)	0.10 (41)	-0.52 (25)	-0.90 (18)	-0.21 (32)	-1.84 (7)	-1.25 (13)
Parenthesized numbers indicate the number of genes categorized in each superfamily.										
Tco: <i>T. coccineum</i> ; Tci: <i>T. cinerariifolium</i> ; Cs: <i>C. seticuspe</i> ; Aa: <i>A. annua</i> ; Ha: <i>H. annuus</i> ; Nt: <i>N. tabacum</i> ; Os: <i>O. sativa</i> ; At: <i>A. thaliana</i> .										

Table 7
Superfamilies with lowest multiplication odds scores in *T. coccineum*.

Category	IPR ID	Superfamily name	Tco	Tci	Cs	Aa	Ha	Nt	Os	At
Signaling	IPR039512	RCHY1, zinc-ribbon	-1.18 (3)	-0.48 (8)	-0.09 (12)	-0.09 (12)	-0.01 (13)	1.28 (39)	-0.72 (6)	-0.09 (12)
*Parenthesized numbers indicate the number of genes categorized in each superfamily.										
Tco: <i>T. coccineum</i> ; Tci: <i>T. cinerariifolium</i> ; Cs: <i>C. seticuspe</i> ; Aa: <i>A. annua</i> ; Ha: <i>H. annuus</i> ; Nt: <i>N. tabacum</i> ; Os: <i>O. sativa</i> ; At: <i>A. thaliana</i> .										

Table 8
Superfamilies with characteristic odds scores in *T. cinerariifolium* genome

Category	IPR ID	Superfamily name	Tco	Tci	Cs	Aa	Ha	Nt	Os	At
Biodefense	IPR035992	Ricin B-like lectins	0.81 (44)	1.41 (69)	-0.34 (17)	-0.05 (22)	-1.22 (7)	-0.80 (11)	-1.10 (8)	-1.48 (5)
Biodefense	IPR036861	Endochitinase-like	-0.13 (7)	-1.13 (1)	-0.39 (5)	-0.25 (6)	0.53 (14)	0.29 (11)	0.09 (9)	0.37 (12)
Signaling	IPR036097	Signal transduction histidine kinase, dimerization/phosphoacceptor domain	-0.11 (32)	1.41 (101)	-0.62 (21)	-0.37 (26)	-0.28 (28)	0.35 (46)	-1.74 (7)	-0.74 (19)
Signaling	IPR024792	Rho GDP-dissociation inhibitor domain	0.48 (18)	1.24 (34)	-0.14 (10)	-0.58 (6)	-0.34 (8)	-0.14 (10)	-1.04 (3)	-1.04 (3)
Metabolism	IPR012347	Ferritin-like	0.72 (22)	1.29 (35)	-0.03 (11)	-0.71 (5)	-0.57 (6)	-0.86 (4)	-1.23 (2)	-0.57 (6)
Metabolism	IPR036909	Cytochrome c-like domain	0.40 (21)	1.16 (39)	-0.50 (9)	-0.84 (6)	-0.22 (12)	-0.16 (17)	-0.82 (7)	-0.82 (7)
Metabolism	IPR037069	Acyl-CoA dehydrogenase/oxidase, N-terminal domain	0.45 (22)	1.05 (36)	-0.60 (8)	-0.30 (11)	-0.22 (12)	0.22 (18)	-0.98 (5)	-0.84 (6)
*Parenthesized numbers indicate the number of genes categorized in each superfamily.										
Tco: <i>T. coccineum</i> ; Tci: <i>T. cinerariifolium</i> ; Cs: <i>C. seticuspe</i> ; Aa: <i>A. annua</i> ; Ha: <i>H. annuus</i> ; Nt: <i>N. tabacum</i> ; Os: <i>O. sativa</i> ; At: <i>A. thaliana</i> .										

Among biodefense-related superfamilies, genes encoding proteins with the “Ribosome-inactivating protein (RIP)” (IPR036041) domain showed multiplication in the *T. coccineum* genome, exhibiting a multiplication score of 1.96 (Table 6). In the previous study, this superfamily was also multiplied in *T. cinerariifolium*. Although the odds score of “Ribosome-inactivating protein (RIP)” (IPR036041) in the *T. coccineum* genome is 1.5 times higher than that in *T. cinerariifolium* (Table 6), the odds score of “Ricin B-like lectins” (IPR035992) in the *T. coccineum* genome is less than that of *T. cinerariifolium* (Table 8). RIPs, including ricin, show high toxicity to a wide range of species, including insects, bacteria, and viruses, serving as biodefense molecules for the producing plant²⁶. RIPs are categorized into type I and type II due to the absence or presence (respectively) of the ricin B lectin domain²⁷. The ricin B lectin domain is involved in internalization via binding of target cell glycans; therefore, type-II RIPs have higher toxicity than type-I RIPs. Taken together, these results demonstrated that genes encoding higher-toxicity type-II RIPs are multiplied (i.e., more abundant) in the *T. cinerariifolium* genome compared with the *T. coccineum* genome, suggesting that *T. coccineum* may have been subjected to, or may be more sensitive to, natural enemies under wild conditions, compared with *T. cinerariifolium*. In the previous study, a gene encoding a putative insecticidal type-II RIP Tci_399175 (accession No. GEY27201.1) that showed sequence similarity to the *Sambucus nigra* insecticidal RIP SNA-I (*S. nigra* agglutinin-I, accession No. O22415.1)²⁸, was found in the *T. cinerariifolium* genome². A BLASTP search of the *T. coccineum* genome with the SNA-I sequence returned Tco_1336120. An alignment of the RICIN domain, which is important for identifying target cells, confirmed that this putative insecticidal RIP is also encoded in the *T. coccineum* genome (Supplemental Fig. 3). These results indicated that a SNA-I-like insecticidal RIP is conserved in both *Tanacetum* species. In combination, these comparative analyses of the RIP genes verified that type-I and type-II RIPs are abundant in the *T. coccineum* and *T. cinerariifolium* genomes, respectively, suggesting distinct RIP-associated defense strategies between these two plants.

“Endochitinase-like superfamily” (IPR036861), which plays a pivotal role in defense against fungal pathogens, is present in the *T. coccineum* genome at levels similar to those seen in other genera, and more abundant than that seen in *T. cinerariifolium* (Table 8). While *T. cinerariifolium* is native to a region with a dry environment, *T. coccineum* is native to a region with a humid environment,

indicating that these plants have been exposed to distinct natural enemies. These observations suggested that distinct defense strategies may have evolved in the different lineages leading to *T. cinerariifolium* and *T. coccineum*, reflecting the differing areas of origin of the two species.

The metabolism-related superfamily showed the highest gene multiplication values in *T. coccineum*, with genes encoding “Urease, alpha subunit” (IPR005848), “Lipoxygenase, C-terminal domain” (IPR036226), “RuBisCo” (IPR033966), “Metal-dependent hydrolase” (IPR032466), “Enolase-like, C-terminal domain” (IPR036849), and “Cytochrome P450” (IPR036396) exhibiting multiplication scores of 2.36, 1.86, 1.60, 1.48, 1.38 and 0.90 respectively (Table 6). In particular, genes encoding metalloproteins such as lipoxygenases, metal-dependent hydrolases, and cytochrome P450 were multiplied, as was observed for the *T. cinerariifolium* genome². Since some pyrethrin-related proteins belong to the superfamily of cytochrome P450s or lipoxygenases, the corresponding genes may have multiplied in the common lineage shared by *T. coccineum* and *T. cinerariifolium*, given that both species have the ability to synthesize pyrethrins. The cytochrome P450 superfamily-encoding genes were 1.6 times more numerous in the *T. coccineum* genome than in the *T. cinerariifolium* genome. Molecular phylogenetic analysis showed that 57% of the *T. coccineum* cytochrome P450s were not included in orthologous gene clusters but constituted single-genus clusters (Fig. 5), suggesting that some of orthologous cytochrome P450s were multiplied in the *T. coccineum*-specific lineage. These results support the view that the cytochrome P450s of *T. cinerariifolium* and *T. coccineum* might have multiplied during the respective evolutionary processes and then acquired an ability to produce species-specific plant specialized metabolites, including pyrethrins. Moreover, these results suggested that species-specific secondary metabolites may be more abundant in *T. coccineum* than *T. cinerariifolium*. the Further investigation of *T. coccineum* secondary metabolites is needed.

While the high multiplication of proteins harboring the “HECT, E3 ubiquitin ligase catalytic domain” (IPR035983) was observed in *T. coccineum* (Table 6) (as is the case in *T. cinerariifolium*), low multiplication was seen for proteins containing “RCHY, zinc-ribbon” (IPR039512), a domain that is contained in the RING-finger-type E3 ubiquitin ligases (Table 7). These results suggested that genes encoding the HECT-type E3 ubiquitin ligases are multiplied in the *T. coccineum* genome, while genes encoding the RING-finger-type E3 ubiquitin ligases are not. Investigation of the biological significance of this apparent imbalance in the amplification of E3 ubiquitin ligase-encoding genes awaits further study.

Similarly, genes encoding proteins containing the “Signal transduction histidine kinase, dimerization/ phosphoacceptor domain” (IPR036097) are multiplied in *T. cinerariifolium*, but not in the *T. coccineum* genome (Table 8). *In planta*, histidine kinases are involved in responding to environmental stimuli, including sunlight, plant hormones, and ethylene. *A. thaliana* ethylene receptor 1 (AtETR1) is a typical gas-induced histidine kinase that possesses both a HATPase (histidine kinase-like ATPase) domain and a REC (phosphoacceptor receiver) domain²⁹. We surveyed the histidine kinase-encoding genes of *T. coccineum* and *T. cinerariifolium* for the presence of HATPase and REC domains; data for the number of genes containing each domain are presented as Venn diagrams in Fig. 6A. The number of genes encoding both HATPase and REC domains were 13 and 38 for *T. coccineum* and *T. cinerariifolium*, respectively. A molecular phylogenetic tree of the predicted histidine kinase proteins is presented in Fig. 6B; the AtETR1 protein is indicated, as are clusters for the 5 and 4 paralogs found in *T. cinerariifolium* and *T. coccineum*, respectively. Comparative analysis detected not only orthologous clusters but also the apparent multiplication of a *T. cinerariifolium*-specific single-genus cluster (Fig. 6B, green).

Our previous study implicated a gas-induced histidine kinase in the VOC-mediated regulation of pyrethrin production in *T. cinerariifolium*². The present study also suggested a correlation between the extent of pyrethrin production and the number of histidine kinase proteins in *T. cinerariifolium* and *T. coccineum*. These results suggested that *T. cinerariifolium* have acquired a gas (i.e., VOC) -induced pyrethrin production system via species-specific multiplication of histidine kinase-encoding genes. Investigation of the functional relationship between histidine kinases and VOC-induced pyrethrin production is underway.

Conclusion

In this study, we elucidated the draft genome of *T. coccineum*. Comparative genomic analyses between *T. coccineum* and a closely related species, *T. cinerariifolium*, revealed characteristic features of *T. coccineum* genes. First, while *T. coccineum* harbors a greater number of RIP-encoding genes than *T. cinerariifolium*, fewer higher-toxicity type-II RIP-encoding genes are present in the *T. coccineum* genome, suggesting that *T. coccineum* is subjected or sensitive to natural enemies in its original growth environment.

Second, *T. coccineum* shares typical pyrethrin biosynthesis-related genes with *T. cinerariifolium*. Third, the number of histidine kinase-encoding genes and the sequences of regions flanking pyrethrin biosynthesis-related genes differ between these two closely related species. Such species-specific genomic features presumably are involved in the differential regulation of pyrethrin production. Consequently, the present study provides clues to the elucidation of the relationship between species-specific specialized plant metabolites and biodefense systems, and to the development of metabolic engineering fundamentals for effective and sustainable pyrethrin production.

Materials And Methods

Plant materials and genome sequencing

The seeds of *Tanacetum coccineum* (cultivar: Robinson mix) were obtained commercially from Sakata Seed Co., Ltd., Japan (production number, 906435). The use of *Tanacetum coccineum* seeds has been approved by Sakata Seed Co. Ltd. All the experimental work on plant material described in this study complies with the relevant institutional, national, and international guidelines and legislation.

Genomic DNA was extracted from seeds using a DNeasy Plant Mini Kit (Qiagen) according to the manufacturer's instructions. PE and MP libraries with three different insert sizes (3, 5, and 8 kb) of the extracted DNA were constructed using a TruSeq DNA PCR-Free kit (Illumina) and a Nextera Mate-Pair Sample Prep Kit (Illumina), respectively. The PE and MP libraries then were subjected to 151 × 2 cycles of paired-end sequencing, using NovaSeq 6000 Illumina instruments. MS libraries of the extracted DNA were constructed using a TrueSeq DNA PCR-Free kit (Illumina). The MS library was subjected to 301 × 2 cycles of paired-end sequencing, using an Illumina Miseq System. PB libraries of the extracted DNA were constructed using a SMRTbell Express Template Prep Kit (PacBio). The PB libraries were subjected to sequencing using a PacBio Sequel II system.

Genome size estimation using flow cytometry

T. coccineum were grown from seed under field conditions for 6 months, and a leaf was obtained on of the resulting plants. A 5-mm-square section was excised from the leaf using a razor and treated with Quantum Stain UV and PI for DNA (Quantum Analysis) according to the manufacturer's protocol. The genome size of *T. coccineum* was estimated using a CyFlow SL flow cytometer (Sysmex Partec). Segments from leaves of *T. cinerariifolium* (wild type, 7.1 Gb) and *Chrysanthemum seticuspe* (3 Gb, cultivar Gojo-0)⁵ also were assayed as size references. The *C. seticuspe* Gojo-0 strain used in this study and related information are available from the National BioResource Project (<https://shigen.nig.ac.jp/chrysanthemum/top.jsp>).

Genome size estimation using k-mer depth information

Separately, genome size was estimated using k-mer depth analysis, as described previously⁸. Briefly, the occurrence of each *k*-mer was counted using Jellyfish software and the homozygous *k*-mer depth peak (*C*) was determined from a histogram of the *k*-mer depths. The genome size then was calculated using the formula

$$Genomesize = \frac{n * (L - k + 1)}{C}$$

where *n*, *L*, and *k* indicate the total number of reads, average read length, and *k*-mer size (respectively)

De novo assembly of genome sequences

The obtained read sequences were cleaned for assembly as described in our previous study². Adapters derived from the Truseq or Nextera mate-pair Sample Prep Kit, low-quality reads, and short reads (<36 bp in length) were trimmed using Trimmomatic version 0.36³⁰.

Contig sequences were generated from the PE reads by a four-step process ("pre-assembly", "SOAPdenovo assembly", "clean-up", and "merging with other assembly result"). First, the cleaned PE reads, including overlaps, were detected and pre-assembled with PANDAsseq³¹. Second, these pre-assembled reads and remaining reads were subjected to SOAPdenovo v2.04-r240⁹ with multi *k*-mers from 80 to 127 to generate contig sequences. Third, the SOAPdenovo-generated contigs were cleaned by trimming and merging the contigs showing more than 95% sequence identity with other contig sequences.

In contrast to the previous study, a “hybrid assembling” process was incorporated to generate pre-scaffolds using SSPACE-longread v1-1¹⁰ with PacBio reads as guides, including setting minimum overlap length to 20 bp, minimum number of links to 3, and maximum link ratio to 0.3.

Scaffold sequences were generated by a three-step process (“scaffolding”, “gapfilling”, and “polishing”). First, PE and MP reads were mapped to contig sequences using bowtie version 2.3.4.3³² with the –local option; read pairs mapped discordantly were subjected to the following steps, and the selected reads and contigs were scaffolded using SSPACE-STANDARD version 3.0¹¹ (BaseClear), including setting the minimum number of links to 3. Second, the positions for which the bases were unknown (e.g., ‘N’) in the scaffolds were filled by GapFiller v1-10¹² (BaseClear), for which the minimum number of overlaps was set to 30 and the number of reads to trim off was set to 10, and then filled by using TGS-Gapcloser¹³ with the default parameter. Additionally, the scaffolds were polished using the POLCA program¹⁴ to complete the draft genome. Third, the completeness of the draft genome was evaluated using BUSCO-v5³³ with the embryophyta_odb9 protein set.

Gene prediction and annotation

The assembled genome was subjected to analysis with AUGUSTUS 3.3.1¹⁵ trained with the training data generated by BUSCO using the –long option. TEs in predicted genes were identified by hmmpfam in HMMER 2.3.1³⁴ and comparison to GyDB 2.0¹⁷m using an E-value cutoff of 1.0. To identify high-confidence genes, the genes encoding proteins with known protein signatures were detected using InterProScan 5.33-72.0³⁵ and annotated using Blast2GO³⁶.

Comparative analysis of TE content versus that in other plants

The TE content also was estimated for six other genome-elucidated plants, including *A. thaliana* (TAIR10³⁷), *N. tabacum* (Ntab-TN90³⁸), *O. sativa* (assembly Build 4.0³⁹), *H. annuus* (HA412HO_v1.1⁴⁰), *A. annua* (ASM311234v1⁴¹), and *C. seticuspe* (CSE_r1.0⁴²). As described for TE detection in *T. cinerariifolium*, the coding regions of these genomes were estimated using AUGUSTUS 3.3.1¹⁵ with the *Arabidopsis* model set and the default parameters, and the TEs in these predicted transcriptomes were extracted using hmmpfam and comparison to GyDB. According to GyDB¹⁷ classifications, the percentage of genomic regions occupied by each clade of TEs was calculated as an accumulation score.

Molecular phylogenetic trees for *deh*- and *tork*-clade TEs also were estimated using the ORTHOSCOPE method⁴³, as described in our previous study⁴⁴. The amino acid sequences of hmmpfam-extracted reverse-transcriptase domains encoded by the TEs in *T. cinerariifolium*, *H. annuus*, *A. annua*, and *C. seticuspe* were aligned using CLUSTAL W-mpi 0.13⁴⁵, and maximum-likelihood phylogenetic trees based on a JTT matrix-based model⁴⁶ with 100 bootstraps were created using Fast Tree 2.1.10 (JTT model, CAT approximation)⁴⁷.

Comparative analysis of protein superfamily content versus that in other plants

We also detected protein signatures using InterProScan³⁵ analysis in six genome-elucidated plants in the previous study and *T. cinerariifolium*, using methods as described above for the InterProScan analysis of *T. coccineum*. Having determined the number of genes possessing each superfamily signature, the multiplication odds score for each combination of InterProScan-detected superfamily signature (Sig) and plant genus (Genus) were calculated as follows:

$$\text{Multiplicationoddscore}(\text{Genus}, \text{Sig}) = \log_2 \frac{N(\text{Genus}, \text{Sig}) + PS}{N(\text{Sig}) + PS}$$

where N (Genus, Sig) represents the number of the plant-genus genes with an InterProScan-detected superfamily signature and PS represents the pseudo-count constant, which was set to 5.00.

In a further analysis for functional proteins, we subjected the predicted proteins of the RIP-related superfamily to a BLASTP search²⁵ (version 2.7.1). *S. nigra* agglutinin I (SNA-I, accession No. O22415.1)²⁸ was used as the query for RIPs. The detected proteins (SNA-I, Tci_399175² and Tco_1336120) were subjected to sequence alignment using CLUSTAL W-MPI 0.13⁴⁵, and maximum-likelihood phylogenetic trees based on a JTT matrix-based model⁴⁶ with 500 bootstraps were created using MEGA software⁴⁸.

The predicted histidine kinases of *T. coccineum* and *T. cinerariifolium* were subjected to searches using the Conserved Domain Database (CDD v.3.19)⁴⁹ to detect HATPase (histidine kinase-like ATPase) and REC (phosphoacceptor receiver) domains. Molecular phylogenetic trees were further estimated for histidine kinases. AtETR1 and histidine kinases of *T. coccineum* and *T. cinerariifolium* were aligned using CLUSTAL W-MPI 0.13⁴⁵, and maximum-likelihood phylogenetic trees based on a JTT matrix-based model⁴⁶ with 500 bootstraps were created using MEGA software⁴⁸.

Abbreviations

ADH2, alcohol dehydrogenase 2; ALDH1, aldehyde dehydrogenase 1; CCMT, 10-carboxychrysanthemic acid 10-methyltransferase; CDS, chrysanthemyl diphosphate synthase; CHH, chrysanthemol 10-hydroxylase; ETR1, ethylene receptor 1; GLIP, GDSL (Gly-Asp-Ser-Leu motif) lipase; HATPase, histidine kinase-like ATPase; JMH, jasmone hydroxylase; LOX1, 13-lipoxygenase; MP, mate-pair; MS, Miseq; PB, PacBio; PE, paired-end; PYS, pyrethrolone synthase; REC, phosphoacceptor receiver; RIP, ribosome-inactivating protein; RT, reverse transcriptase; SNA-I, *Sambucus nigra* agglutinin I; TE, transposable element; VOC, volatile organic compound

Declarations

Author contributions: H. S. and K. N. designed the experiments. T. Y., A. S., H. S., and K. N. performed the experiments, analyzed the data, and wrote the paper.

TT matrix-based model⁴⁶ with 500 bootstraps were created using MEGA software⁴⁸.

Data availability

The draft genome sequences and annotations, along with the raw reads for PE, MP, MS, and PB, have been deposited in the DNA Data Bank of Japan (DDBJ) under the BioProject accession Code PSUB016075.

Acknowledgments

We are grateful to Forte, Ltd., for English language editing prior to submission; to the Macrogen Japan Corp. (www.macrogen-japan.co.jp) for Illumina platform sequencing; and to the Cytotechs Corp. (www.cytotechs.com) for flow cytometric estimation of the genome size.

References

1. Lybrand, D. B., Xu, H., Last, R. L. & Pichersky, E. How Plants Synthesize Pyrethrins: Safe and Biodegradable Insecticides. *Trends Plant Sci* **25**, 1240–1251, doi:10.1016/j.tplants.2020.06.012 (2020).
2. Yamashiro, T., Shiraishi, A., Satake, H. & Nakayama, K. Draft genome of *Tanacetum cinerariifolium*, the natural source of mosquito coil. *Sci Rep* **9**, 18249, doi:10.1038/s41598-019-54815-6 (2019).
3. Kikuta, Y. *et al.* Specific regulation of pyrethrin biosynthesis in *Chrysanthemum cinerariaefolium* by a blend of volatiles emitted from artificially damaged conspecific plants. *Plant Cell Physiol* **52**, 588–596, doi:10.1093/pcp/pcr017 (2011).
4. Katsuda, Y. Progress and future of pyrethroids. *Top Curr Chem* **314**, 1–30, doi:10.1007/128_2011_252 (2012).
5. Nakano, M. *et al.* A pure line derived from a self-compatible *Chrysanthemum seticuspe* mutant as a model strain in the genus *Chrysanthemum*. *Plant Sci* **287**, 110174, doi:10.1016/j.plantsci.2019.110174 (2019).
6. Bennett, M. D., Smith, J. B. & Heslop-Harrison, J. S. Nuclear DNA amounts in angiosperms. *Proc R Soc Lond B* **216**, doi:10.1098/rspb.1982.0069 (1982).
7. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770, doi:10.1093/bioinformatics/btr011 (2011).

8. Alm Rosenblad, M. *et al.* Genomic Characterization of the Barnacle *Balanus improvisus* Reveals Extreme Nucleotide Diversity in Coding Regions. *Mar Biotechnol (NY)* **23**, 402–416, doi:10.1007/s10126-021-10033-8 (2021).
9. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18, doi:10.1186/2047-217X-1-18 (2012).
10. Boetzer, M. & Pirovano, W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15**, 211, doi:10.1186/1471-2105-15-211 (2014).
11. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579, doi:10.1093/bioinformatics/btq683 (2011).
12. Nadalin, F., Vezzi, F. & Policriti, A. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* **13 Suppl 14**, S8, doi:10.1186/1471-2105-13-S14-S8 (2012).
13. Xu, M. *et al.* TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *Gigascience* **9**, doi:10.1093/gigascience/giaa094 (2020).
14. Zimin, A. V. & Salzberg, S. L. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Comput Biol* **16**, e1007981, doi:10.1371/journal.pcbi.1007981 (2020).
15. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644, doi:10.1093/bioinformatics/btn013 (2008).
16. Waterhouse, R. M. *et al.* BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol Biol Evol* **35**, 543–548, doi:10.1093/molbev/msx319 (2018).
17. Llorens, C. *et al.* The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res* **39**, D70-74, doi:10.1093/nar/gkq1061 (2011).
18. Ramirez, A. M., Yang, T., Bouwmeester, H. J. & Jongsma, M. A. A trichome-specific linoleate lipxygenase expressed during pyrethrin biosynthesis in pyrethrum. *Lipids* **48**, 1005–1015, doi:10.1007/s11745-013-3815-1 (2013).
19. Xu, H. *et al.* Coexpression Analysis Identifies Two Oxidoreductases Involved in the Biosynthesis of the Monoterpene Acid Moiety of Natural Pyrethrin Insecticides in *Tanacetum cinerariifolium*. *Plant Physiol* **176**, 524–537, doi:10.1104/pp.17.01330 (2018).
20. Rivera, S. B. *et al.* Chrysanthemyl diphosphate synthase: isolation of the gene and characterization of the recombinant non-head-to-tail monoterpene synthase from *Chrysanthemum cinerariaefolium*. *Proc Natl Acad Sci U S A* **98**, 4373–4378, doi:10.1073/pnas.071543598 (2001).
21. Kikuta, Y. *et al.* Identification and characterization of a GDSL lipase-like protein that catalyzes the ester-forming reaction for pyrethrin biosynthesis in *Tanacetum cinerariifolium*- a new target for plant protection. *Plant J* **71**, 183–193, doi:10.1111/j.1365-313X.2012.04980.x (2012).
22. Li, W., Zhou, F. & Pichersky, E. Jasmonate Hydroxylase, a Key Enzyme in the Synthesis of the Alcohol Moiety of Pyrethrin Insecticides. *Plant Physiol* **177**, 1498–1509, doi:10.1104/pp.18.00748 (2018).
23. Li, W., Lybrand, D. B., Zhou, F., Last, R. L. & Pichersky, E. Pyrethrin Biosynthesis: The Cytochrome P450 Oxidoreductase CYP82Q3 Converts Jasmonolone To Pyrethrolone. *Plant Physiol* **181**, 934–944, doi:10.1104/pp.19.00499 (2019).
24. Xu, H. *et al.* Pyrethric acid of natural pyrethrin insecticide: complete pathway elucidation and reconstitution in *Nicotiana benthamiana*. *New Phytol* **223**, 751–765, doi:10.1111/nph.15821 (2019).
25. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402, doi:10.1093/nar/25.17.3389 (1997).
26. Bolognesi, A., Bortolotti, M., Maiello, S., Battelli, M. G. & Polito, L. Ribosome-Inactivating Proteins from Plants: A Historical Overview. *Molecules* **21**, doi:10.3390/molecules21121627 (2016).
27. Zhu, F., Zhou, Y. K., Ji, Z. L. & Chen, X. R. The Plant Ribosome-Inactivating Proteins Play Important Roles in Defense against Pathogens and Insect Pest Attacks. *Front Plant Sci* **9**, 146, doi:10.3389/fpls.2018.00146 (2018).
28. Shahidi-Noghabi, S., Van Damme, E. J. & Smagghe, G. Carbohydrate-binding activity of the type-2 ribosome-inactivating protein SNA-I from elderberry (*Sambucus nigra*) is a determining factor for its insecticidal activity. *Phytochemistry* **69**, 2972–2978, doi:10.1016/j.phytochem.2008.09.012 (2008).
29. Shakeel, S. N., Wang, X., Binder, B. M. & Schaller, G. E. Mechanisms of signal transduction by ethylene: overlapping and non-overlapping signalling roles in a receptor family. *AoB Plants* **5**, plt010, doi:10.1093/aobpla/plt010 (2013).

30. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120, doi:10.1093/bioinformatics/btu170 (2014).
31. Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G. & Neufeld, J. D. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* **13**, 31, doi:10.1186/1471-2105-13-31 (2012).
32. Langdon, W. B. Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Min* **8**, 1, doi:10.1186/s13040-014-0034-0 (2015).
33. Manni, M., Berkeley, M. R., Seppey, M., Simao, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol* **38**, 4647–4654, doi:10.1093/molbev/msab199 (2021).
34. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763, doi:10.1093/bioinformatics/14.9.755 (1998).
35. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240, doi:10.1093/bioinformatics/btu031 (2014).
36. Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676, doi:10.1093/bioinformatics/bti610 (2005).
37. Lamesch, P. *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* **40**, D1202–D1210, doi:10.1093/nar/gkr1090 (2012).
38. Sierro, N. *et al.* The tobacco genome sequence and its comparison with those of tomato and potato. *Nat Commun* **5**, 3833, doi:10.1038/ncomms4833 (2014).
39. Rice Annotation, P. *et al.* The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res* **36**, D1028–D1033, doi:10.1093/nar/gkm978 (2008).
40. Badouin, H. *et al.* The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **546**, 148–152, doi:10.1038/nature22380 (2017).
41. Shen, Q. *et al.* The Genome of *Artemisia annua* Provides Insight into the Evolution of Asteraceae Family and Artemisinin Biosynthesis. *Mol Plant* **11**, 776–788, doi:10.1016/j.molp.2018.03.015 (2018).
42. Hirakawa, H. *et al.* De novo whole-genome assembly in *Chrysanthemum seticuspe*, a model species of Chrysanthemums, and its application to genetic and gene discovery analysis. *DNA Res* **26**, 195–203, doi:10.1093/dnares/dsy048 (2019).
43. Inoue, J. & Satoh, N. ORTHOSCOPE: An Automatic Web Tool for Phylogenetically Inferring Bilaterian Orthogroups with User-Selected Taxa. *Mol Biol Evol* **36**, 621–631, doi:10.1093/molbev/msy226 (2019).
44. Shiraishi, A. *et al.* Repertoires of G protein-coupled receptors for Ciona-specific neuropeptides. *Proc Natl Acad Sci USA* **116**, 7847–7856, doi:10.1073/pnas.1816640116 (2019).
45. Li, K. B. ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics* **19**, 1585–1586, doi:10.1093/bioinformatics/btg192 (2003).
46. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275–282, doi:10.1093/bioinformatics/8.3.275 (1992).
47. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490, doi:10.1371/journal.pone.0009490 (2010).
48. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547–1549, doi:10.1093/molbev/msy096 (2018).
49. Lu, S. *et al.* CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res* **48**, D265–D268, doi:10.1093/nar/gkz991 (2020).

Figures



Tanacetum cinerariifolium



Tanacetum coccineum

Figure 1

Flowers of *Tanacetum cinerariifolium* and *Tanacetum coccineum*.

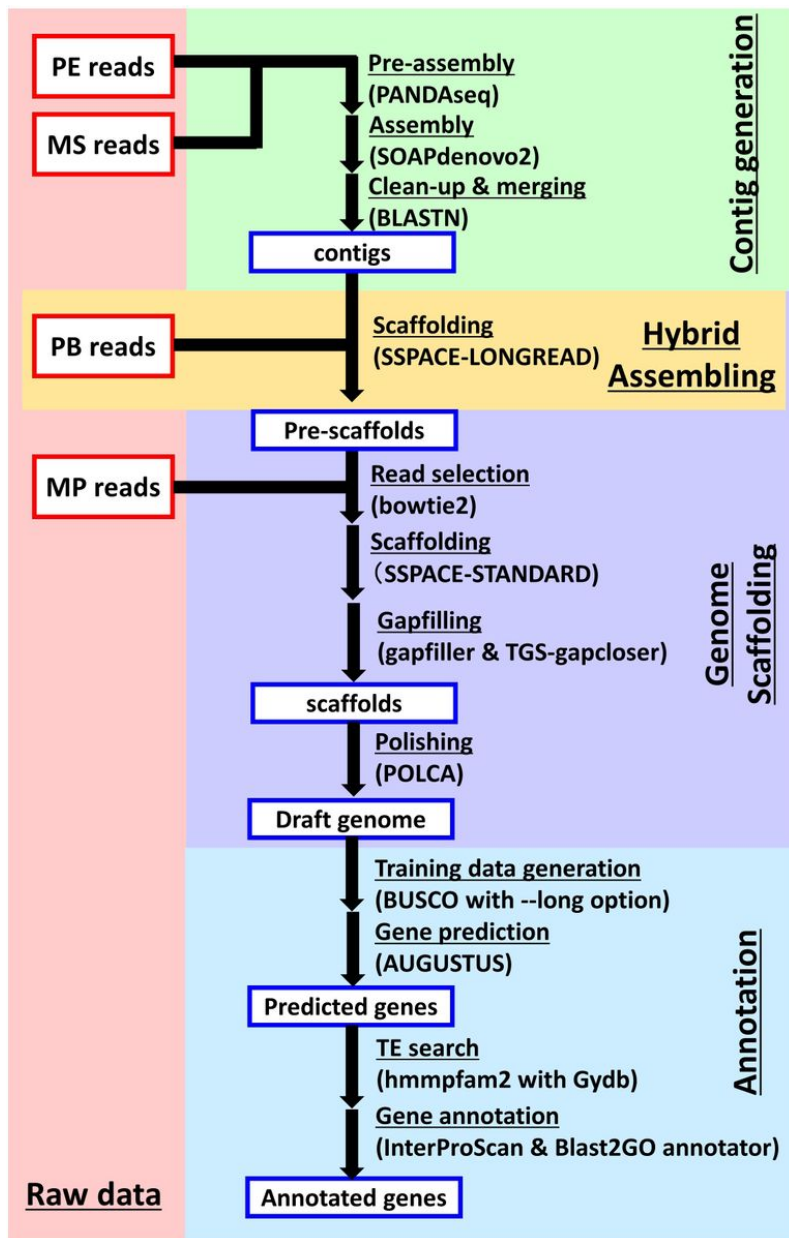


Figure 2

Flowchart of genome assembly and gene prediction.

The paired-end (PE) reads and Miseq (MS) reads were subjected to contig assembly by a three-step process, including “pre-assembly” using PANDAsseq, “contig assembly” using SOAPdenovo, and “clean-up & merging” using BLASTN. The PacBio (PB) reads were subjected to hybrid assembly using SSPACE-LONGREAD, generating pre-scaffolds. Then, the mate-pair (MP) reads were subjected to scaffold generation by a four-step process, including “read selection” using bowtie2, “scaffolding” using SSPACE-STANDARD, “gapfilling” using Gapfiller and TGS-Gapcloser, and “polishing” using POLCA. These processes yielded the complete draft genome. The coding sequences in the draft genome then were annotated by a four-step process, including “Training data generation” using BUSCO, “Gene prediction” using AUGUSTUS, “TE search” (TE, transposable element) using hmmpfam, and “Gene annotation” using InterProScan and Blast2GO.

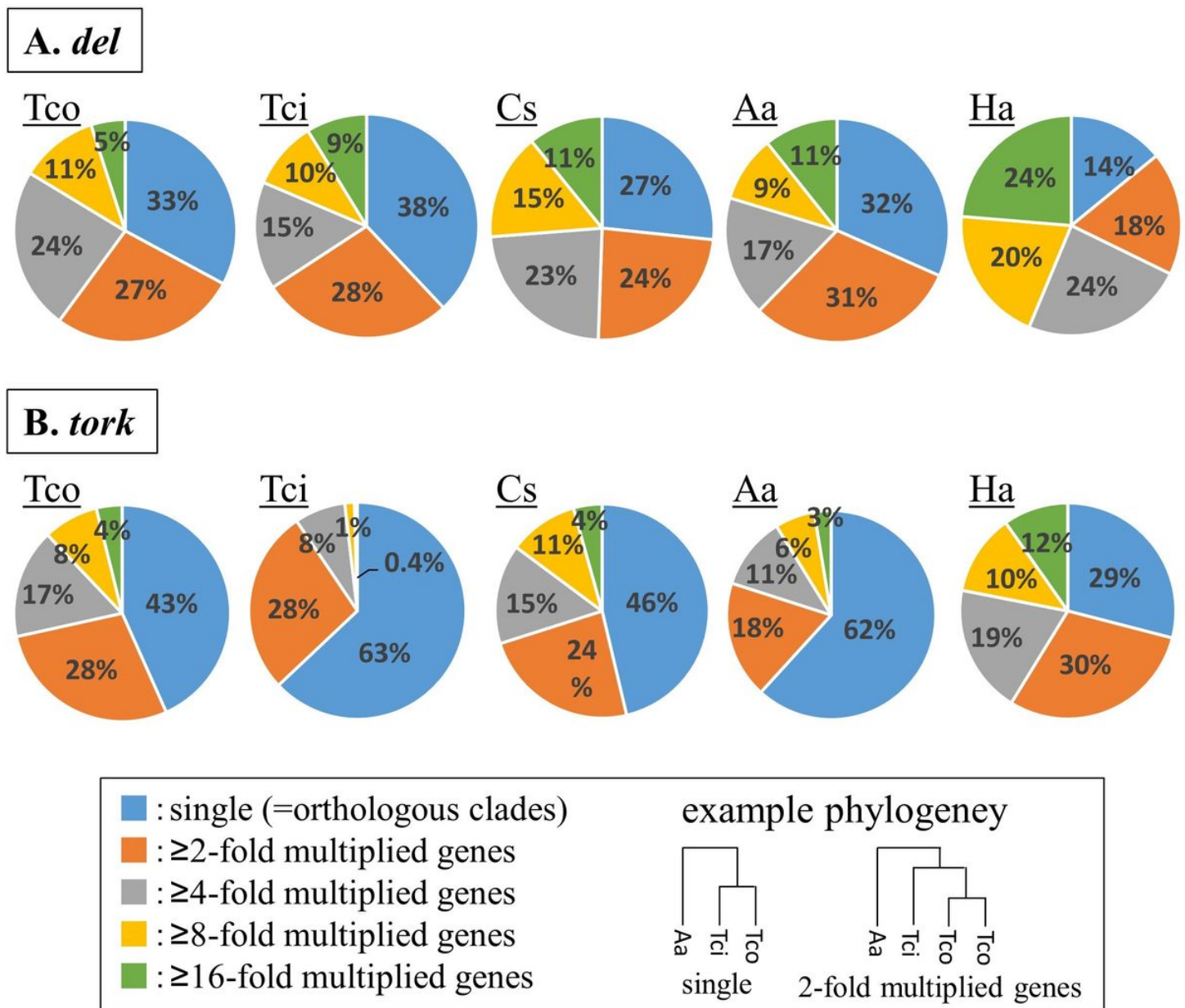


Figure 3

Multiplication analysis of (A) *del*-clade and (B) *tork*-clade transposable elements (TEs). Based on molecular phylogenetic trees for *del* TEs and *tork* TEs, the number of co-clustered genes in single-genus clusters, which reflects the inferred number of genus-specific duplication events, were counted for each genus. Tco: *T. coccineum*; Tci: *T. cinerariifolium*; Cs: *C. seticuspe*; Aa: *A. annua*; Ha: *H. annuus*

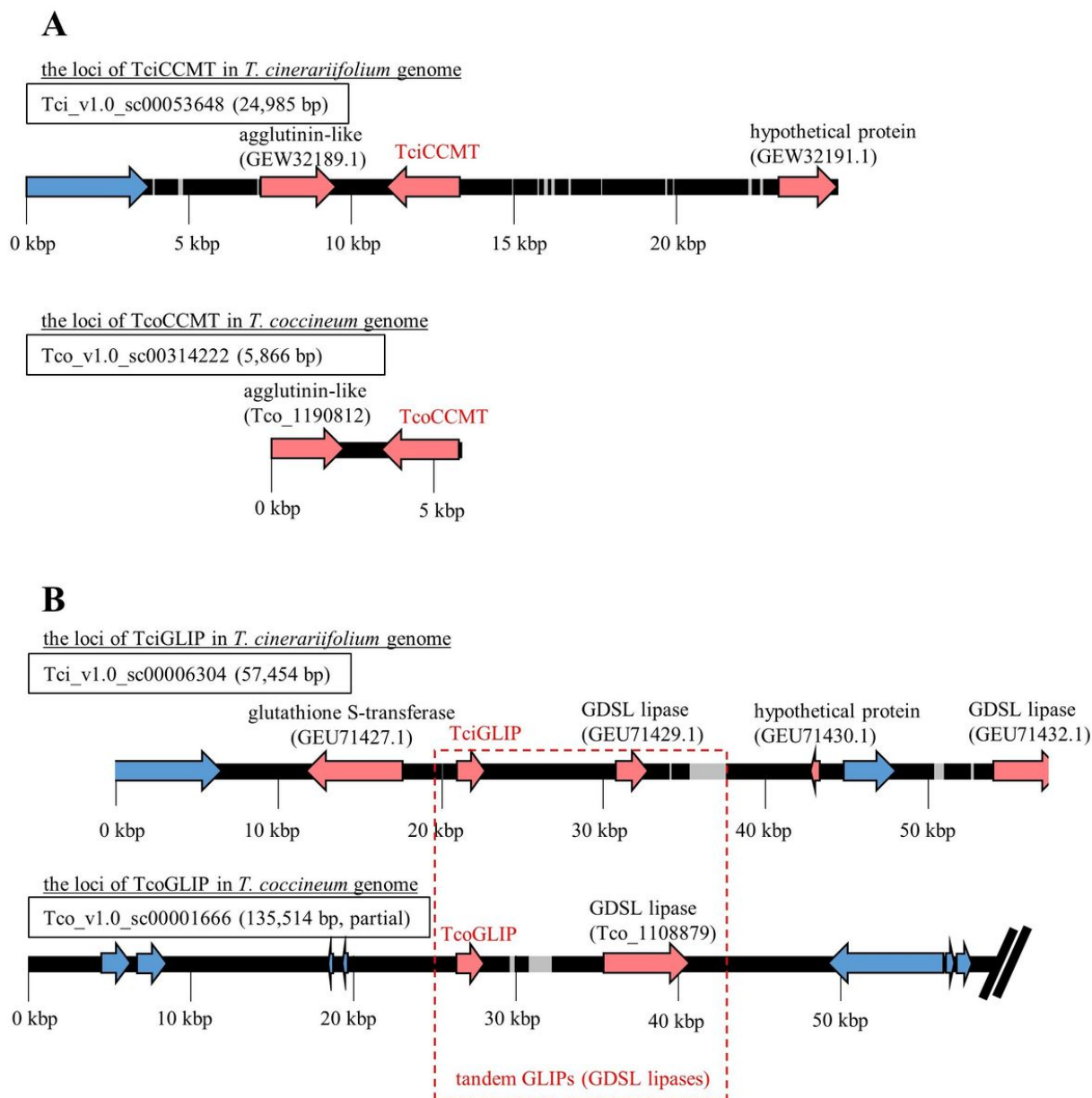


Figure 4

Scaffolds including genes encoding the pyrethrin-related enzymes Tci(o)CCMT (A) and Tci(o)GLIP (B). Gray region: gap; red arrows: protein-coding gene; blue arrow: transposable element (TE) gene.

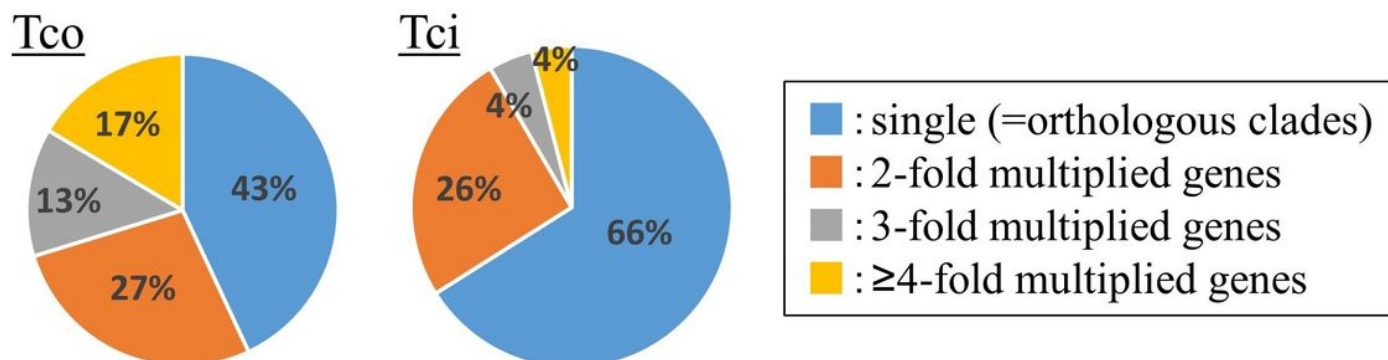


Figure 5

Multiplication analysis of genes encoding proteins with the superfamily “cytochrome P450” (IPR036396) domain in the *T. coccineum* and *T. cinerariifolium* genomes. Tco: *T. coccineum*; Tci: *T. cinerariifolium*.

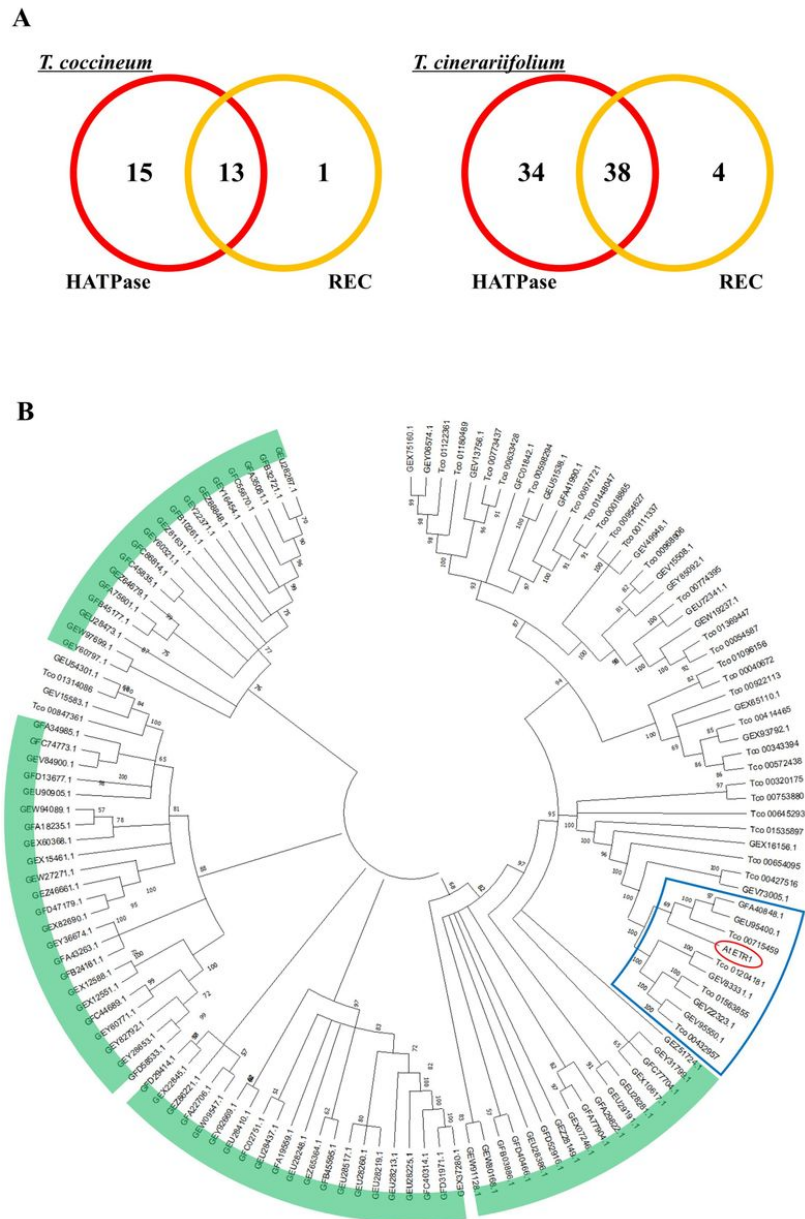


Figure 6

The number of gene encoding proteins containing histidine kinase domains (A), and phylogenetic analysis (B) of histidine kinase superfamily genes. (A) Venn diagram of numbers of *T. coccineum* and *T. cinerariifolium* histidine kinase superfamily genes encoding proteins with histidine kinase-like ATPase (HATPase) or phosphoacceptor receiver (REC) domains. (B) A phylogenetic tree of *T. coccineum* and *T. cinerariifolium* genes encoding proteins belonging to the signal transduction histidine kinase superfamily. *T. coccineum* genes and *T. cinerariifolium* genes are showed as the name with prefix "Tco" and accession No., respectively. The *A. thaliana* ethylene receptor 1 protein (AtETR1) is circled in red; the AtETR1 cluster which includes 5 paralogs found in *T. cinerariifolium* and 4 paralogs found in *T. coccineum* is surrounded by a blue line; *T. cinerariifolium*-specific clusters are indicated by green arcs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Yamashiroetal2021SupplementaryInformationfinal.docx](#)