nature communications

Article

https://doi.org/10.1038/s41467-025-59428-4

Large-scale transcriptome mining enables macrocyclic diversification and improved bioactivity of the stephanotic acid scaffold

Received: 22 October 2024

Accepted: 23 April 2025

Published online: 06 May 2025

Check for updates

Xiaofeng Wang[1,3], Khadija Shafiq[1,3], Derrick A. Ousley[1,3], Desnor N. Chigumba[1], Dulciana Davis[1], Kali M. McDonough[1], Lisa S. Mydy[1], Jonathan Z. Sexton[1,2] & Roland D. Kersten[1]

Nearly 10,000 plant species are represented by RNA-seq datasets in the NCBI sequence read archive, which are difficult to search in unassembled format due to database size. Here, we optimize RNA-seq assembly to transform most of this public RNA-seq data to a searchable database for biosynthetic gene discovery. We test our transcriptome mining pipeline towards the diversification of moroidins, which are plant ribosomally-synthesized and posttranslationally-modified peptides (RiPPs) biosynthesized from copper-dependent peptide cyclases. Moroidins are bicyclic compounds with a conserved stephanotic acid scaffold, which becomes cytotoxic to non-small cell lung adenocarcinoma cells with an additional C-terminal macrocycle. We discover moroidin analogs with second ring structures diversified at the crosslink and the non-crosslinked residues including a moroidin analog from water chickweed, which exhibits higher cytotoxicity against lung adenocarcinoma cells than moroidin. Our study expands stephanotic acid-type peptides to grasses, Lowiaceae, mints, pinks, and spurges while demonstrating that large-scale transcriptome mining can broaden the medicinal chemistry toolbox for chemical and biological exploration of eukaryotic RiPP lead structures.

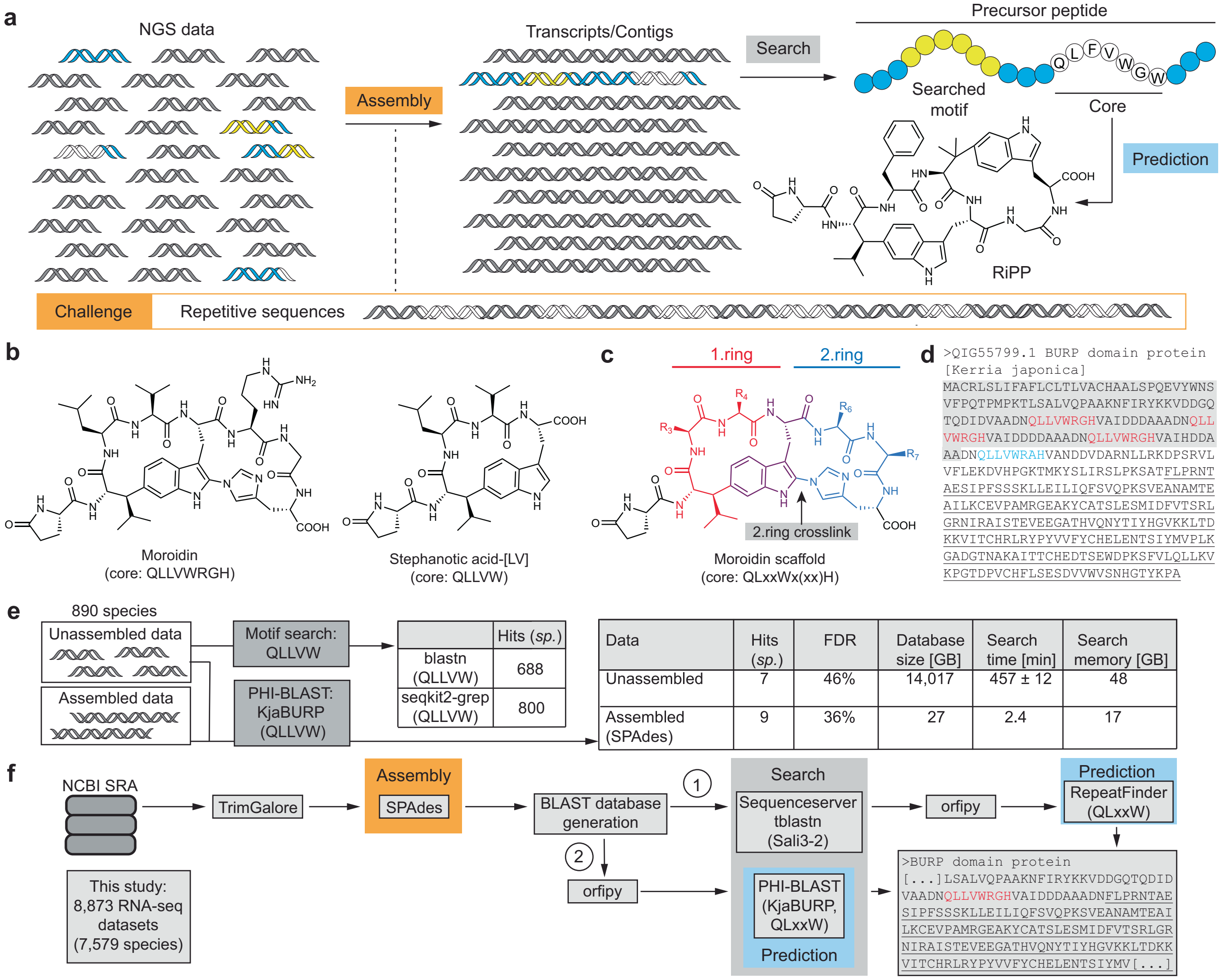
Plants produce a large diversity of ribosomally-synthesized and posttranslationally-modified peptides (RiPPs) with endogenous functions in development and abiotic and biotic stress responses[1–3]. These bioactivities have been translated into pesticides, immunosuppressants, and lead structures for chemotherapy drug candidates[4,5]. RiPPs are natural products biosynthesized by the ribosomal formation of a precursor peptide that is posttranslationally modified in a core peptide sequence and proteolytically cleaved to release the modified core as the mature RiPP[6,7]. In prokaryotes, the discovery of RiPPs has been driven over the past two decades by advances in sequencing technologies, the availability of large numbers of genome sequences, and the identification of genes and motifs involved in RiPP biosynthesis for bioinformatic prediction[8–12]. However, genome mining as a strategy for plant RiPP discovery is constrained by the limited number of high-quality plant genomes with annotated gene products in searchable repositories such as the NCBI non-redundant protein database (Supplementary Table 1)[13]. In addition, plant genome mining of natural product biosynthetic genes in the context of their metabolic pathways is complicated by eukaryotic gene structures and less biosynthetic gene clustering than in prokaryotes[14]. An alternative to plant genomics-guided discovery is the use of transcriptomes to predict peptide chemistry[15–17].

In the past decade, over 9000 phylodiverse plant species have had their transcriptomes sequenced[18] and deposited as RNA-seq data in the NCBI Sequence Read Archive (SRA)[19] (May 2024, Supplementary Table 1). However, only a small fraction of this data has been utilized for gene-guided discovery of plant natural products to date. Most transcriptomes are sequenced using next-generation sequencing

[1]Department of Medicinal Chemistry, University of Michigan, Ann Arbor, MI, USA. [2]Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA. [3]These authors contributed equally: Xiaofeng Wang, Khadija Shafiq, Derrick A. Ousley. e-mail: rkersten@umich.edu

(NGS) technologies that provide short-read sequence data[20,21], and a key step in the transcriptomics-guided discovery of RiPPs is the assembly of such short-read sequences into longer transcripts that represent full-length coding sequences of expressed genes. For non-model organisms, i.e., those lacking a reference genome, transcriptome assembly is performed de novo using specialized short-read NGS assemblers[22]. Three common RNA-seq assemblers are Trinity[23], SPAdes[24,25], and SOAPdenovo-trans[26], which have been described as efficient in multi-species benchmarking[22]. Trinity and SPAdes were applied in RiPP discovery studies from plant RNA-seq data[27–30] and SOAPdenovo-trans was originally employed to assemble transcriptomes of the 1000 plant transcriptomes initiative (1KP)[18]. Another common assembler, MEGAHIT[31], was developed for fast and memory-efficient assembly of metagenomic short-read NGS data, but has limited applications in RNA-seq data assembly to date[32]. De novo transcriptome mining of RiPPs requires assembly of a precursor peptide gene so that the core peptide sequence is connected to a searchable sequence motif for bioinformatic precursor gene identification (Fig. 1a).

Burpitides are a recent addition to the diversity of plant RiPPs (Supplementary Fig. 1)[33–35], which consist of 4-8 amino acids and feature crosslinks between aromatic amino acids, such as tryptophan or tyrosine, and other amino acids or the peptide backbone[15,29,30,36]. These crosslinks often involve functionalization of C(sp^3)-H-bonds by copper-dependent burpitide cyclases comprising a C-terminal BURP domain[37–41]. The stephanotic acid-type burpitide scaffold is derived from a QLxxW core peptide (where x is any proteinogenic amino acid), and defined by a C-C-bond between a Leu-Cβ and a tryptophan-indole-C6[42,43]. These peptides can be monocyclic such as stephanotic acid-[LV] from *Cercis canadensis*, or bicyclic such as moroidin from *Dendrocnide moroides* (Fig. 1b)[44,45]. In moroidins, the stephanotic acid scaffold forms the first ring and the second ring is formed by a C-N-crosslink between the tryptophan-indole-C2 and the imidazole-N1 of a C-terminal histidine (Fig. 1c). Reported moroidins are 7–8 amino acids-long with the second ring comprising 2–4 amino acids[15,44,45]. Moroidin and its analog celogentin C exhibit moderate in vitro cytotoxicity against non-small cell lung cancer whereas stephanotic acid-[LV] has no such detectable cytotoxicity, indicating that the



```
>QIG55799.1 BURP domain protein
[Kerria japonica]
MACRLSLIFAFLCLTLVACHAALSPQEVYWNS
VFPQTPMPKTLSALVQPAAKNFIRYKKVDDGQ
TQDIDVAADNQLLVWRGHVAIDDDAAADNQLL
VWRGHVAIDDDDAAADNQLLVWRGHVAIHDDA
AADNQLLVWRAHVANDDVDARNLLRKDPSRVL
VFLEKDVHPGKTMKYSLIRSLPKSATFLPRNT
AESIPFSSSKLLEIILIQFSVQPKSVEANAMTE
AILKCEVPAMRGEAKYCATSLESMIDFVTSRL
GRNIRAISTEVEEGATHVQNYTIYHGVKKLTD
KKVITCHRLRYPYVVFYCHELENTSIYMVPLK
GADGTNAKAITTCHEDTSEWDPKSFVLQLLKV
KPGTDPVCHFLSESDVVWVSNHGTYKPA
```

	Hits (*sp.*)
blastn (QLLVW)	688
seqkit2-grep (QLLVW)	800

Data	Hits (*sp.*)	FDR	Database size [GB]	Search time [min]	Search memory [GB]
Unassembled	7	46%	14,017	457 ± 12	48
Assembled (SPAdes)	9	36%	27	2.4	17

```
>BURP domain protein
[...]LSALVQPAAKNFIRYKKVDDGQTQDID
VAADNQLLVWRGHVAIDDDAAADNFLPRNTAE
SIPFSSSKLLEILIQFSVQPKSVEANAMTEAI
LKCEVPAMRGEAKYCATSLESMIDFVTSRLGR
NIRAISTEVEEGATHVQNYTIYHGVKKLTDKK
VITCHRLRYPYVVFYCHELENTSIYMV[...]
```

Fig. 1 | Transcriptome mining of plant RiPPs. a General workflow of transcriptomics-guided discovery of plant RiPPs including de novo assembly of NGS data as a challenging step addressed in this work. **b** Chemical structures of stephanotic acid-type burpitides. **c** The moroidin scaffold. **d** Moroidin precursor peptide KjaBURP from *Kerria japonica*. Moroidin core peptides are highlighted in red and blue color. Sequence highlighted in gray was not assembled with the BURP domain by SOAPdenovo-trans in 1kp database[18]. BURP-domain sequence is underlined. **e** Comparison of unassembled and assembled RNA-seq data search for discovery of stephanotic acid-type burpitides. **f** Transcriptome mining workflow of this study with either (1) Sequenceserver-based tblastn search or (2) commandline PHI-BLAST search. *sp.* species, FDR false discovery rate. Source data are provided as a Source Data file.

second ring of moroidin is required for its bioactivity[15,36]. The moroidin burpitide cyclase KjaBURP, discovered in *Kerria japonica* through transcriptome mining, contains four moroidin core peptide sequence repeats within a 380-amino-acid protein domain (Fig. 1d)[15]. Similarly, many plant RiPP precursors feature multiple core peptides embedded in sequence repeats[1,3,29,30,36,46,47], which pose a challenge for de novo assembly as it can result in incomplete assembly of the precursor-defining sequence motifs and its core peptide(s) (Fig. 1d, and Supplementary Fig. 1)[29]. An example is the misassembly of moroidin precursor gene *KjaBURP* from *Kerria japonica* transcriptome by SOAPdenovo-trans[26] within the 1kp database which misses the N-terminal repetitive core peptide domain with moroidin cores (QLLVWRGH) required for moroidin discovery in japanese kerria (Fig. 1d). Given the importance of de novo assembly of precursor gene transcripts from RNA-seq data, a systematic comparison of NGS assemblers in the context of scaled plant RiPP discovery from transcriptomes is needed.

Herein, we assess three de novo short-read assemblers to generate precursor peptides from RNA-seq datasets of RiPP-producing plants and apply the most effective assembler towards mining the transcriptomes of 7569 plant species for new bioactive stephanotic acid-type burpitides and their biosynthetic cyclases.

Results

RNA-seq assembly benchmarking for discovery of plant RiPP precursor peptides

Towards large-scale transcriptome mining of plant RiPPs such as stephanotic acid-type burpitides, we wanted to first test if sequencing data assembly is advantageous for peptide discovery from RNA-seq data. We searched unassembled and SPAdes-de novo assembled transcriptomes of 890 plant species for the presence of the stephanotic acid core peptide QLLVW (Fig. 1e). The target plant species had metabolomic data deposited in plantMASST[48], a repository of tandem mass spectrometry-based metabolomic data, and which can thus be tested for the presence of the stephanotic acid-QLLVW analyte. The unassembled RNA-seq data were first searched for the presence of the QLLVW motif as nucleotide or protein queries via blastn or seqkit2[49], respectively. These searches of the stephanotic acid motif yielded too many false positives (98–99%) for effective peptide discovery (Fig. 1e). Next, we searched homologs of moroidin cyclase KjaBURP with a QLLVW motif in both unassembled and assembled data via PHI-BLAST, a protein BLAST with target motif query[50]. Searching assembled data yielded 9 of 11 stephanotic acid-containing plants based on metabolomic data (Supplementary Fig. 2) compared to 7 hits in the unassembled data query. In addition, the assembled data search had a lower false discovery rate (36% versus 46%) compared to the unassembled data search. The BLAST database size of assembled transcriptomes was 0.2% of unassembled RNA-seq data and search time and memory was 0.5% and 33%, respectively, with the same computational resource for assembled RNA-seq data (Fig. 1e, and Supplementary Fig. 2). Given the smaller database size, lower search cost, and better search results, we focused on applying RNA-seq assembly for plant peptide discovery from short sequence motifs.

As a prerequisite for large-scale plant transcriptome assembly, we aimed to identify a de novo NGS assembler that balances accuracy, speed, and computational cost for generating transcripts of published plant RiPP precursor genes. We included non-burpitide precursor peptides to expand the benchmarking data to 27 RNA-seq datasets from RiPP-producing plants. These datasets encompassed 36 precursor genes of (a) burpitides[15,29,30,36], (b) cysteine-rich peptides[51–62], (c) cyclotides[3,63,64], (d) orbitides[46,47,65], (e) PawS-derived RiPPs[66], and (f) linear RiPPs[1,67,68] (Supplementary Fig. 1). The 36 corresponding precursor peptides encoded 138 total core peptides and 65 unique core peptide sequences. 120 of the total core peptides were localized in sequence repeats within 18 precursor peptides. These datasets allowed us to assess the de novo assembly of sequence repeats within precursor peptides for subsequent bioinformatic core peptide detection. Each RNA-seq dataset was trimmed and assembled in triplicate on a computing cluster using the de novo NGS assemblers Trinity[23], SPAdes[24,25], or MEGAHIT[31]. MEGAHIT was the fastest of all assemblers and used the lowest total memory (Table 1). SPAdes assembly of all datasets required on average 19% more time and 15% more memory than MEGAHIT, while Trinity assembly needed on average ~19x more time and ~15x more memory than MEGAHIT (Table 1). For the detection of core peptides for RiPP discovery, a core should be assembled with a searchable motif, such as a BURP domain, to enable discovery of its RiPP. Some RiPPs require the correct assembly of short, often repetitive core peptide domains in addition to assembly with a searchable motif for successful RiPP discovery. SPAdes assembled more total core peptides with searchable motifs (67%) than Trinity (47%) and MEGAHIT (49%). Furthermore, SPAdes assembled 74% unique core peptides correctly with searchable motifs for RiPP discovery, whereas Trinity (57%) and MEGAHIT (58%) assembled less unique cores correctly. To further assess the accuracy of precursor gene assembly, the sum of sequence similarity and sequence identity values of all assembled precursors were determined by comparison to the database precursor sequences. Herein, SPAdes had the highest similarity and identity scores (81.4% and 80.6%, respectively) (Table 1). Finally, SPAdes assembled the most repetitive cores (62%), followed by MEGAHIT (43%) and Trinity (37%). Based on replicates of all de novo benchmarking data assemblies, MEGAHIT and SPAdes assemblies were fully reproducible, whereas differences could be observed for individual Trinity assemblies. Read trimming before assembly resulted in more detected core peptides and more accurate precursor peptide assembly (Supplementary Table 2). Ultimately, our benchmarking identified SPAdes as the de novo assembler, which generated precursor peptide transcripts with most unique core peptides at low computational cost and high speed. This assessment was further realized by a scaled de novo transcriptome assembly of the 1kp higher plant RNA-seq data with SPAdes and comparison of moroidin precursor search results via PHI-BLAST with the published datasets assembled by de novo transcriptome assembler SOAPdenovo-trans (Supplementary Table 3, Supplementary Data 1)[18,26]. SPAdes-assembled 1kp data enabled identification of six fused moroidin cyclases with twelve different moroidin core peptides in five species, whereas published 1kp data contained only three fused moroidin cyclases with five different core peptides in three species (Supplementary Table 3, Supplementary Data 1), showing that de novo SPAdes assembly can yield more predicted peptides at scale than existing plant transcriptomes from other assemblers.

An alternative to de novo RNA-seq assembly is genome-guided transcriptome assembly which requires an annotated reference genome[69]. With increasing numbers of high-quality plant genomes complementing available RNA-seq data, we also compared genome-guided transcriptome assembly with StringTie[70] or Trinity to de novo assemblers regarding precursor peptide assembly (Table 2). Overall, genome-guided assembly via StringTie outperformed de novo SPAdes assembly in time, total detected cores, and assembly of cores in sequence repeats. However, SPAdes assemblies of precursor peptides were superior to genome-guided assembly by StringTie or Trinity in the number of detected unique cores and identity scores (Table 2). Based on lower number of high-quality plant genomes compared to plant RNA-seq datasets (Supplementary Table 1), we proceeded with a de novo assembly strategy with SPAdes.

Large-scale transcriptome mining of fused moroidin cyclases

Given that the second ring of moroidins is required for their lung adenocarcinoma cell cytotoxicity, we aimed to generate and search a large plant transcriptome database bioinformatically for new moroidin chemotypes and underlying burpitide cyclases for bicyclic diversification of the stephanotic acid scaffold for improved bioactivity.

Table 1 | Benchmarking results of de novo transcriptome assembly for plant RiPP precursor discovery

De novo assembler	Total time [min]	Total memory usage [GB]	Total detected cores (138 total cores)	Detected unique correct cores (65 total unique cores)	Detected repetitive cores (120 total repetitive cores)	Similarity [sum %] (total similarity score: 3600)	Identity [sum %] (total identity score: 3600)
MEGAHIT (v1.2.9)	1371 ± 66	210 ± 0	67 ± 0	38 ± 0	51 ± 0	2741 ± 0	2738 ± 0
SPAdes (v3.15.5)	1627 ± 42	241 ± 3	93 ± 0	48 ± 0	74 ± 0	3012 ± 0	2981 ± 0
Trinity (v2.15.1)	21833 ± 558	813 ± 23	64 ± 0	37 ± 0	44 ± 1	2618 ± 10	2597 ± 7

27 RNA-seq datasets were trimmed and assembled which include 36 precursor genes with 138 core peptides and 65 unique core peptides. Each RNA-seq dataset was assembled in triplicate (*n* = 3). Displayed values are means with one standard deviation shown as an error. Source data are provided as a Source Data file.

Table 2 | Benchmarking results of de novo and genome-guided transcriptome assembly for plant RiPP precursor discovery

Assembler	Approach	Total time [min]	Total memory usage [GB]	Total detected cores (98 total cores)	Detected unique correct cores (38 total unique cores)	Detected repetitive cores (87 total repetitive cores)	Similarity [sum %] (total similarity score: 2000)	Identity [sum %] (total identity score: 2000)
MEGAHIT (v1.2.9)	De novo	940 ± 44	123 ± 1	42 ± 0	24 ± 0	32 ± 0	1582 ± 0	1587 ± 0
SPAdes (v3.15.5)	De novo	1185 ± 39	180 ± 1	66 ± 0	32 ± 0	56 ± 0	1728 ± 0	1713 ± 0
Trinity (v2.15.1)	De novo	13683 ± 621	485 ± 6	41 ± 0	25 ± 0	30 ± 1	1527 ± 2	1521 ± 2
StringTie (v2.2.1)	Genome-guided	922 ± 621	1635 ± 2	77 ± 0	27 ± 0	65 ± 0	1565 ± 0	1545 ± 0
Trinity (v2.15.1)	Genome-guided	10545 ± 417	1787 ± 4	62 ± 0	28 ± 0	51 ± 0	1795 ± 1	1691 ± 1

16 RNA-seq datasets with corresponding annotated genomes were trimmed and assembled which include 20 precursor genes with 98 core peptides and 38 unique core peptides. Each RNA-seq dataset was assembled in triplicate (*n* = 3). Displayed values are means with one standard deviation shown as an error. Source data are provided as a Source Data file.

To create a large plant transcriptome database, 8893 plant transcriptomes were de novo assembled with SPAdes from RNA-seq data from the NCBI SRA representing 7579 plant species from 498 plant families. The assembly database was searched by tblastn[71,72] for the presence of BURP-domain-encoding genes as a starting point for the discovery of stephanotic acid-type burpitides. Burpitide cores can be predicted from BURP-domain protein sequences of fused burpitide peptide cyclases, i.e., their core peptide(s) are encoded in the N-terminal domain (RD22 BURP domains) or within the BURP domain (USP BURP domains)[33,37]. The assembled plant transcriptome database included 27,224 full-length non-redundant BURP-domain sequences, which were then analyzed via RepeatFinder[36] for BURP-domain sequences with one or more internal stephanotic acid core peptide motif(s). Among the BURP-domain proteins, candidate stephanotic acid-type burpitide cyclases were identified from 37 species, which represented 7 plant families without reported stephanotic acid chemotypes (Supplementary Table 4, Supplementary Data 2). In addition to Sequenceserver-tblastn search, we also applied PHI-BLAST search with moroidin cyclase KjaBURP and the stephanotic acid core QLxxW as a search motif. PHI-BLAST search yielded the same core peptide hits as Sequenceserver-tblastn search but PHI-BLAST resulted in 28 additional predicted stephanotic acid-type core peptides from transcripts without or with incomplete BURP-domain sequence (Supplementary Tables 4 and 5, Supplementary Data 2 and 3). Given the focus on fused burpitide cyclases in this study, we excluded these additional core peptides in subsequent RiPP discovery experiments.

Candidate cyclases from 17 species included a stephanotic acid core with a histidine at positions 7–9 from the N-terminal glutamine that identified them as putative moroidin cyclases capable of forming a second-ring via Trp-His-crosslinking (Supplementary Table 4, Supplementary Data 2). The predicted moroidin cyclases included 46 unique core peptides with most of the sequence diversity occurring in the second-ring, i.e., at core positions 6–9 (34 unique motifs, Supplementary Table 5, Supplementary Data 2). In addition, several predicted cyclases featured extended stephanotic acid cores with tyrosine or tryptophan at positions 7–9 instead of histidine indicating the possibility of new second-ring crosslinks (Supplementary Table 4, Supplementary Data 2).

Characterization of stephanotic acid-type burpitides with a Trp-Val-crosslinked second-ring

We first set out to search for a stephanotic acid-type burpitide with a second-ring-crosslink via a C-terminal tryptophan. One corresponding burpitide cyclase with a core including a C-terminal tryptophan (Supplementary Table 4, Supplementary Data 2) was encoded in a transcriptome of *Glechoma hederacea* (Lamiaceae, Fig. 2a), encoding the core peptides QLFVWGW and QLLVWGW (Supplementary Table 4, Supplementary Data 2). An analyte with a mass value of 914.419 Da, which matched the core peptide QLFVWGW with mass shifts corresponding to pyroglutamate (PyroGlu) formation and two macrocyclizations, was detected in the methanolic extract of *G. hederacea* leaves (Fig. 2b, and Supplementary Fig. 3). The candidate burpitide

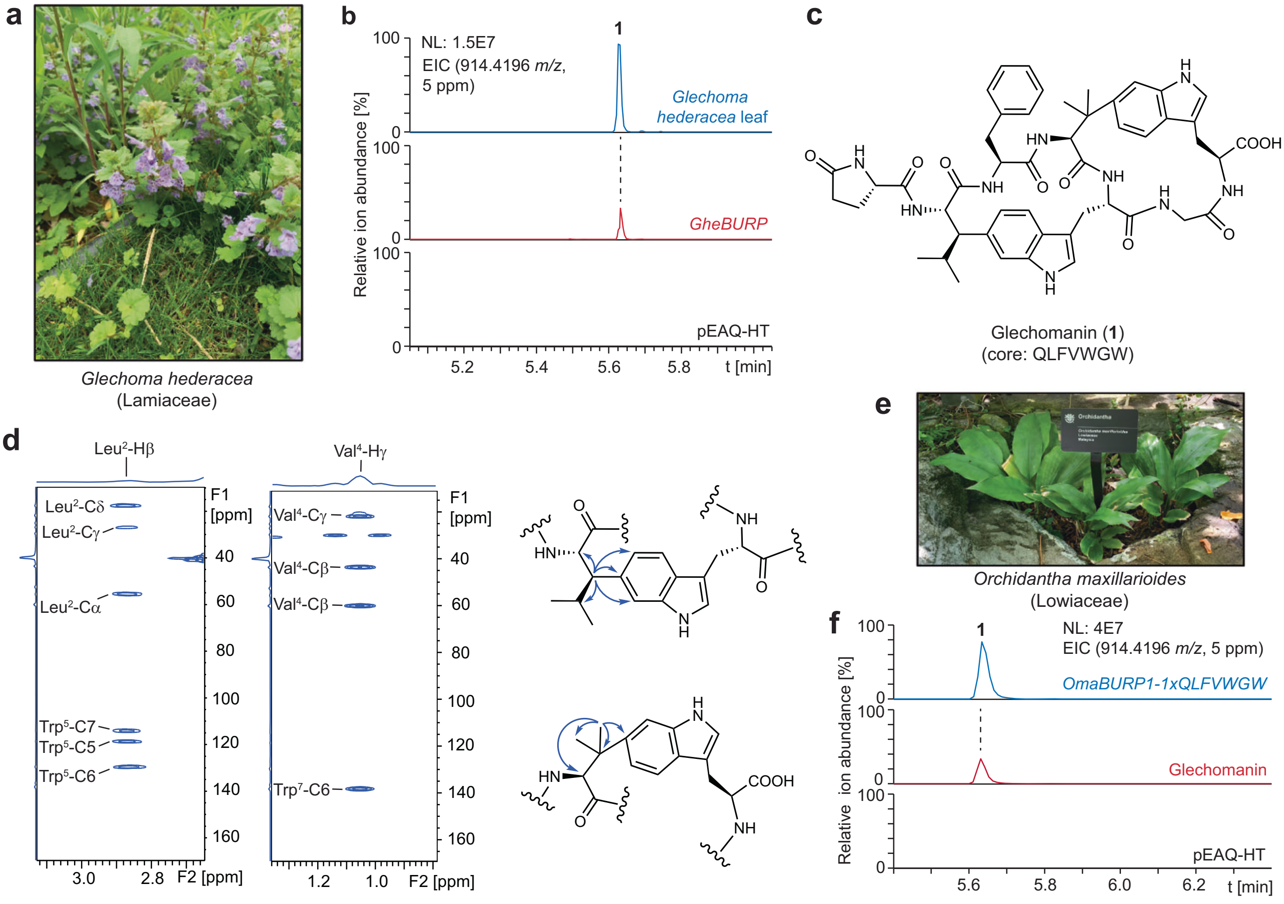


Fig. 2 | Characterization of stephanotic acid-burpitides with a Trp-Val-crosslinked second ring. **a** *Glechoma hederacea* in the University of Michigan Arboretum. **b** Glechomanin detection in methanolic leaf extracts of *G. hederacea* and of transgenic *N. benthamiana* 6 days after infiltration with *Agrobacterium tumefaciens* LBA4404 pEAQ-HT-*GheBURP*. **c** Glechomanin structure. **d** Key HMBC signals of glechomanin side-chain-crosslinks. **e** *Orchidantha maxillarioides* plant in Missouri botanical garden. **f** Glechomanin detection in methanolic leaf extract of transgenic *N. benthamiana* 6 days after infiltration with *A. tumefaciens* LBA4404 pEAQ-HT-*OmaBURP1-1xQLFVWGW*. EIC: extracted ion chromatogram. Source data are provided as a Source Data file.

cyclase gene, named *GheBURP*, of this analyte was cloned from the leaf transcriptome of *G. hederacea* and included four core peptides, three QLFVWGW and one QLLVWGW. Transient expression of GheBURP in *Nicotiana benthamiana* via the pEAQ-HT expression vector[73] through *Agrobacterium tumefaciens* LBA4404 infiltration yielded the same QLFVWGW analyte in methanolic leaf extract of transgenic tobacco six days after infiltration (Fig. 2b). To elucidate the structure of burpitide 1 (Fig. 2c), the analyte was isolated from *G. hederacea* collected from the University of Michigan Arboretum due to higher peptide yields in the source plant than transgenic *N. benthamiana* expressing GheBURP, and subsequently subjected to NMR and Marfey's analysis.

An N-terminal pyroglutamate was characterized by COSY, TOCSY and HMBC and the spin systems of all other amino acids of the predicted core peptide QLFVWGW were determined by COSY, TOCSY, HMBC and ROESY. The core peptide sequence QLFVWGW was confirmed by HMBC and ROESY (Supplementary Figs. 4–10, and Supplementary Tables 6 and 7). The two macrocyclizations were defined as Leu2-C*β* to Trp5-C6 such as in stephanotic acid-[LV] and Val4-C*β* to Trp7-C6. The Leu2-Trp5-crosslink was characterized by (a) HMBC of Leu-H*β* to Trp5-C6, Trp5-C5, Trp5-C7 (Fig. 2d) and ROESY from Leu-H*β* to Trp5-H7, from Leu-H*α* to Trp5-H5 and from Leu-H*δ*1 to Trp5-H7 (Supplementary Figs. 8, 9, 11 and 12) and (b) the detection of stephanotic acid-[LV] (core: QLLVW) after transient expression of GheBURP, which encodes a QLLVWGW core, in *N. benthamiana* (Supplementary Fig. 13). The Val4-Trp7-crosslink was characterized by (a) HMBC of Val4-H*γ*1 to Val4-C*β* and Trp7-C6 (Fig. 2d) and (b) ROESY of Val4-H*γ*1 with Trp7-H7 and of Val4-H*γ*2 with Trp7-H5 and Trp7-H7 and of Val4-H*α* with Trp7-H7 (Supplementary Fig. 8, 9, 11 and 12). In addition, no Val4-H*β* signal was detected. Furthermore, two indole-NH-signals were detected excluding a crosslink via an indole-nitrogen as seen in lyciumins[29]. Marfey's analysis of glechomanin revealed PyroGlu1 and Phe3 are both L-configuration (Supplementary Fig. 14). Moroidin features the same macrocyclic Leu-Trp-connection in the first ring as glechomanin and we observed that the chemical shift value and coupling constants of the β-position hydrogen in the Leu2 fragment of moroidin and glechomanin were comparable (δ_H 3.06 (dd, *J* = 3.0, 11.5 Hz)[15] and δ_H 2.87 (dd, *J* = 3.1, 11.6 Hz), respectively) (Supplementary Table 6). Therefore, we deduce that the β-carbon of Leu2 in glechomanin shares the same relative configuration as observed in moroidin. In addition, the stereochemistry of the first five glechomanin residues (QLFVW) was assigned as stephanotic acid-[LV] because transient expression of WT GheBURP, which has a QLLVWGW core, and GheBURP-1xQLLVW, which has a stephanotic acid-[LV] core peptide, yielded stephanotic acid-[LV] (Supplementary Fig. 13). The stereochemistry of the C-terminal tryptophan was determined as L-Trp7 (Trp7-*S*-C*α)* by comparison of glechomanin NOE data to two geometry-optimized structures of glechomanin with *R*- or *S*-configuration at Trp7-C*α* and the coupling constant of the hydrogen atoms at the *α* and *β1* positions ($J_{\alpha\beta1}$ = 11.2 Hz) (Supplementary Fig. 15). In addition to glechomanin from *G. hederacea*, we detected a glechomanin analyte with the core peptide QLLVWKW in *Orchidantha maxillarioides* (Lowiaceae) (Fig. 2e) based on a predicted glechomanin cyclase in its transcriptome (Supplementary Table 4, Supplementary Data 2). Cloning and heterologous expression of the corresponding burpitide cyclase gene *OmaBURP1* in *N. benthamiana* yielded the putative glechomanin-QLLVWKW analyte as observed in *O. maxillarioides* leaf extract (Supplementary Fig. 16). The structure of this analyte was confirmed as a glechomanin scaffold by expression of a synthetic OmaBURP1 with one core peptide of QLFVWGW, i.e., the glechomanin core, in transgenic *N. benthamiana* which resulted in the detection of glechomanin from ground ivy (Fig. 2f).

GheBURP belongs to the RD22-type burpitide cyclases, which are autocatalytic and contain repetitive N-terminal core peptide domains[33,36]. To characterize GheBURP as a glechomanin cyclase, GheBURP was synthesized as an *E. coli*-codon-optimized gene construct with one core peptide of QLFVWGW and expressed heterologously as a His-tagged protein in *E. coli* BL21(DE3) (Supplementary Fig. 17). GheBURP-1xQLFVWGW was subsequently purified via denature-refolding from inclusion bodies and immobilized nickel-affinity chromatography (Supplementary Fig. 17)[74]. LC-MS/MS-based bottom-up proteomic analysis of GheBURP-1xQLFVWGW with trypsin yielded an analyte with the mass of 2398.134 Da matching the sequence QVTTYNAASDNGNQLFVWGWK and, thus, including the glechomanin core peptide (Supplementary Fig. 18 and 19, Supplementary Table 8). Incubation of GheBURP-1xQLFVWGW with 1 mM cupric chloride at pH 7 for 24 h and subsequent tryptic digest resulted in the detection of a tryptic core peptide species with a loss of four hydrogens in the peptide motif QLFVWGWK (Supplementary Fig. 18). The characterized MS/MS fragmentation of these tryptic core peptides indicated the formation of two macrocyclization bonds in the QLFVWGW core peptide (Supplementary Fig. 20, Supplementary Table 9). To confirm the mass loss of four hydrogens in the core peptide as the crosslinks of the glechomanin structure, the tryptic core peptide species was purified by HPLC from a scaled GheBURP-1xQLFVWGW enzyme assay with cupric chloride followed by trypsin proteolysis. The purified modified GheBURP peptide was then incubated with an aminopeptidase and a carboxypeptidase for 24 h. LC-MS analysis of the exopeptidase assay detected an analyte matching glechomanin (Supplementary Fig. 18) which shows that the mass loss observed after GheBURP enzyme assay represents the bicyclic bonds of glechomanin. Our chemotyping and biochemical analysis revealed that stephanotic acid-type burpitide pathways yielding bicyclic scaffolds via a C-terminal tryptophan-crosslink exist in plants.

Characterization of stephanotic acid-type burpitides with a Tyr-Phe-crosslinked second-ring

Next, we aimed to identify a stephanotic acid pathway yielding a bicyclic scaffold with a C-terminal tyrosine from candidate cyclases with corresponding core peptides in our transcriptome database (Supplementary Table 4, Supplementary Data 2). A transcriptome of annual mercury (*Mercurialis annua*, Euphorbiaceae) encoded RD22-type BURP-domain proteins with putative stephanotic acid-type cores QLFFWRY and QLFFWKY with C-terminal tyrosines. The full-length sequence of one such BURP-domain protein, named ManBURP, was obtained from the recently released annual mercury genome (GenBank ID XP_050232080.1)[75]. Transient expression of a full-length ManBURP in *N. benthamiana* yielded an analyte with the mass 1038.483 Da, which matched the core peptide sequence QLFFWRY after loss of four hydrogens corresponding to two putative macrocyclizations and loss of ammonia due to N-terminal pyroglutamate formation (Supplementary Fig. 21). The same analyte was detectable in an herbarium sample of *M. annua* (Fig. 3a, b) but we could not obtain this source plant in sufficient quantities for peptide isolation. Therefore, we transiently expressed a truncated ManBURP-1xQLFFWRY construct via pEAQ-HT expression system in *N. benthamiana* at scale and purified 5 mg of the target peptide from methanolic extracts of 1.3 kg of transgenic tobacco leaves.

The pure peptide, named mercurialin, was characterized by analysis of 1D and 2D NMR spectroscopic data (Fig. 3c, Supplementary Table 10, Supplementary Figs. 22–29). All its amino acids, including an N-terminal pyroglutamate, were determined by COSY, TOCSY, HMBC and NOESY. The core peptide sequence QLFFWKY was confirmed by HMBC and ROESY. Mercurialin has the characteristic stephanotic acid crosslink between Leu2-C*β* and Trp5-C6, evidenced by HMBC correlations of Leu-H*β* with Trp5-C5, C6 and C7 as well as NOESY correlations from Leu2-H*β* to Trp5-H7, from Leu2-H*α* to Trp5-H5, and from Leu2-H*δ1* to Trp5-H7. The crosslinking position of the second identified ring is established between Phe4-C*β* and Tyr7-O, as unambiguously supported by HMBC correlations of Phe4-H*β* with Tyr7-C4 and

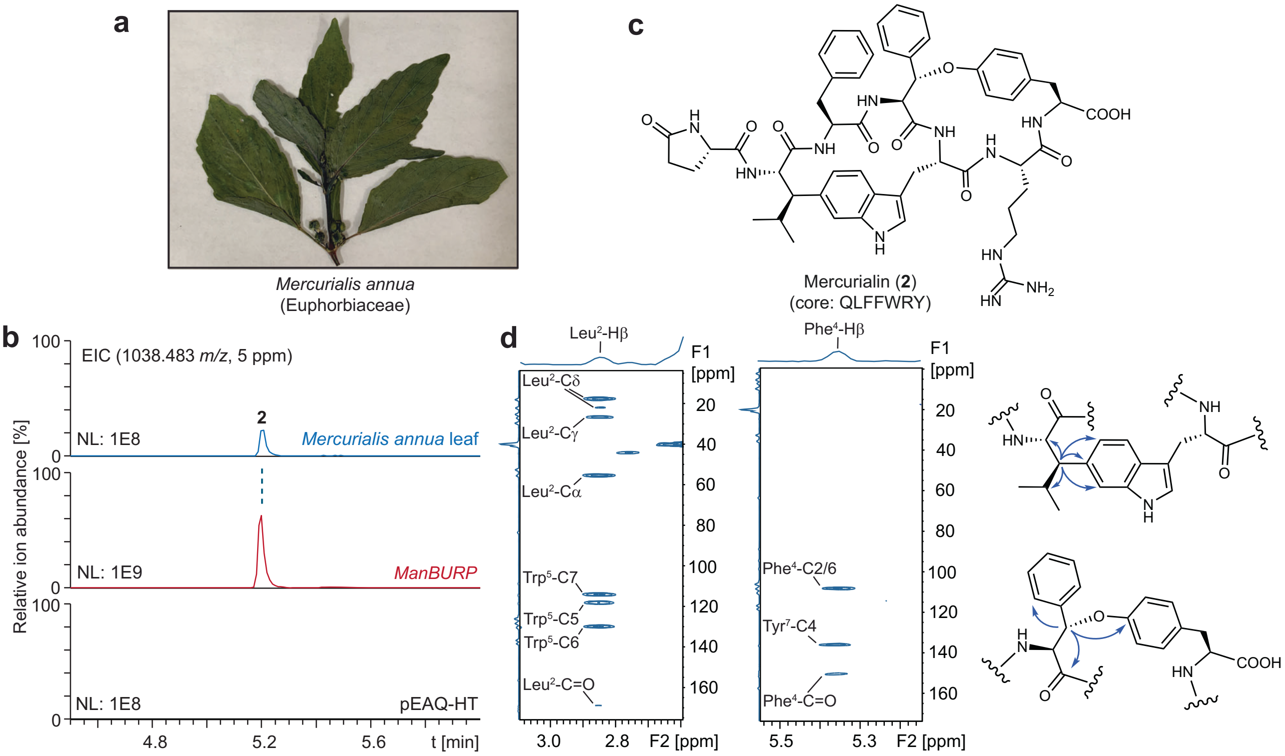


Fig. 3 | Characterization of stephanotic acid-burpitides with a Tyr-Phe-crosslinked second ring. a *Mercurialis annua* from University of Michigan herbarium. **b** Mercurialin detection in methanolic leaf extracts of *M. annua* and of transgenic *N. benthamiana* 6 days after infiltration with *A. tumefaciens* LBA4404 pEAQ-HT-*ManBURP*. **c** Mercurialin structure. **d** Key HMBC correlations of mercurialin side-chain-crosslinks. EIC: extracted ion chromatogram. Source data are provided as a Source Data file.

NOESY correlations from Phe4-Hβ to Tyr7-H3 and Tyr7-H5 (Fig. 3d, and Supplementary Table 10, and Supplementary Figs. 26 and 27). To determine the stereochemistry of mercurialin, we performed Marfey's analysis, which indicates that PyroGlu, Phe, and Arg are all present in the L-configuration. The stereochemistry of the stephanotic acid ring (QLFFW) was inferred as the stephanotic acid stereochemistry based on (a) the similar coupling constants of Leu-Hβ in mercurialin and glechomanin (Supplementary Tables 6 and 10) and (b) the generation of stephanotic acid-[LV] after transient expression of ManBURP-1xQLLVWRY in *N. benthamiana* (Supplementary Fig. 30). Given that the amino acid backbone synthesized by the ribosome exhibits an L-configuration, the stereochemistry of the C-terminal tyrosine is assigned as L-Tyr. The second macrocyclization site at the Cβ of phenylalanine in core residue 4 was determined to be *S* by comparing NOE correlations of mercurialin with two geometry-optimized structures featuring *R*- or *S*-configuration at Phe4-Cβ (Supplementary Fig. 31).

A truncated construct of mercurialin cyclase ManBURP with one core of QLFFWRY was expressed and purified like glechomanin cyclase construct GheBURP-1xQLFVWGW (Supplementary Fig. 32). Bottom-up proteomic analysis of the ManBURP-1xQLFFWRY with LysC digestion identified an analyte with a mass of 4123.068 Da, corresponding to the sequence IPQNYNFQPEPNQLFFWRYQNRGNDIIRLPSMK, which includes the mercurialin core peptide (Supplementary Fig. 33 and 34, Supplementary Table 11). Incubation of ManBURP-1xQLFFWRY with 1 mM cupric chloride at pH 7 for 24 h, followed by LysC digestion, resulted in a loss of four hydrogens in the detected LysC-proteolytic core peptide (Supplementary Fig. 33). MS/MS fragmentation confirmed two macrocyclic bonds in the QLFFWRY core peptide (Supplementary Fig. 35, and Supplementary Table 12). To verify the 4H loss as crosslinks in the mercurialin structure, the LysC-proteolytic core peptide was purified by HPLC and incubated with aminopeptidase and carboxypeptidase for 24 h. LC-MS detected a product consistent with mercurialin (Supplementary Fig. 33), confirming the 4H mass loss reflects the two macrocyclic crosslinks of mercurialin. Mercurialin and its cyclase from *M. annua* show that the stephanotic acid pathway has also evolved bicyclic scaffolds with a C-terminal tyrosine-crosslink in plants.

Chemotaxonomic expansion and cancer cell cytotoxic improvement of moroidins

Bioinformatic analysis of moroidin core peptides within our predicted moroidin cyclases showed that the core QLLVWRGH representing moroidin (Fig. 1b) only appeared twice in nineteen candidate moroidin cyclases whereas the core peptide QLLVWRNH was present in nine of the candidate moroidin cyclases from three plant families (Fig. 4a, Supplementary Table 4, Supplementary Data 2). Motivated by a rationale that evolution of a single core in multiple plant families might suggest potent bioactivity[76], we targeted moroidin-QLLVWRNH for isolation and biological evaluation. A candidate moroidin-QLLVWRNH burpitide cyclase was identified in the transcriptome of water chickweed (*Stellaria aquatica*, Caryophyllaceae) (Fig. 4b). The cloned candidate cyclase from *S. aquatica* contained 18 QLLVWRNH cores and a corresponding moroidin-QLLVWRNH analyte was detected in the methanolic extract of water chickweed from a park close to Lake Erie, Michigan (Fig. 4c). Moroidin-QLLVWRNH was purified (32 mg) and characterized by LC-MS/MS, NMR and transient tobacco expression of *Stellaria aquatica* burpitide cyclase SaqBURP and reported moroidin cyclase KjaBURP with an updated core peptide (KjaBURP-1xQLLVWRNH), as a moroidin natural product (Fig. 4c, d, and Supplementary Table 13, Supplementary Figs. 36–44). High-content screening of moroidin-QLLVWRNH against H1437, H1299, and U2OS cancer cell lines revealed significant cytotoxicity in H1437 non-small cell lung adenocarcinoma cells ($IC_{50} = 1.4\ \mu M$, Fig. 5a), exhibiting a five-fold lower IC_{50} than celogentin C, the most potent previously reported

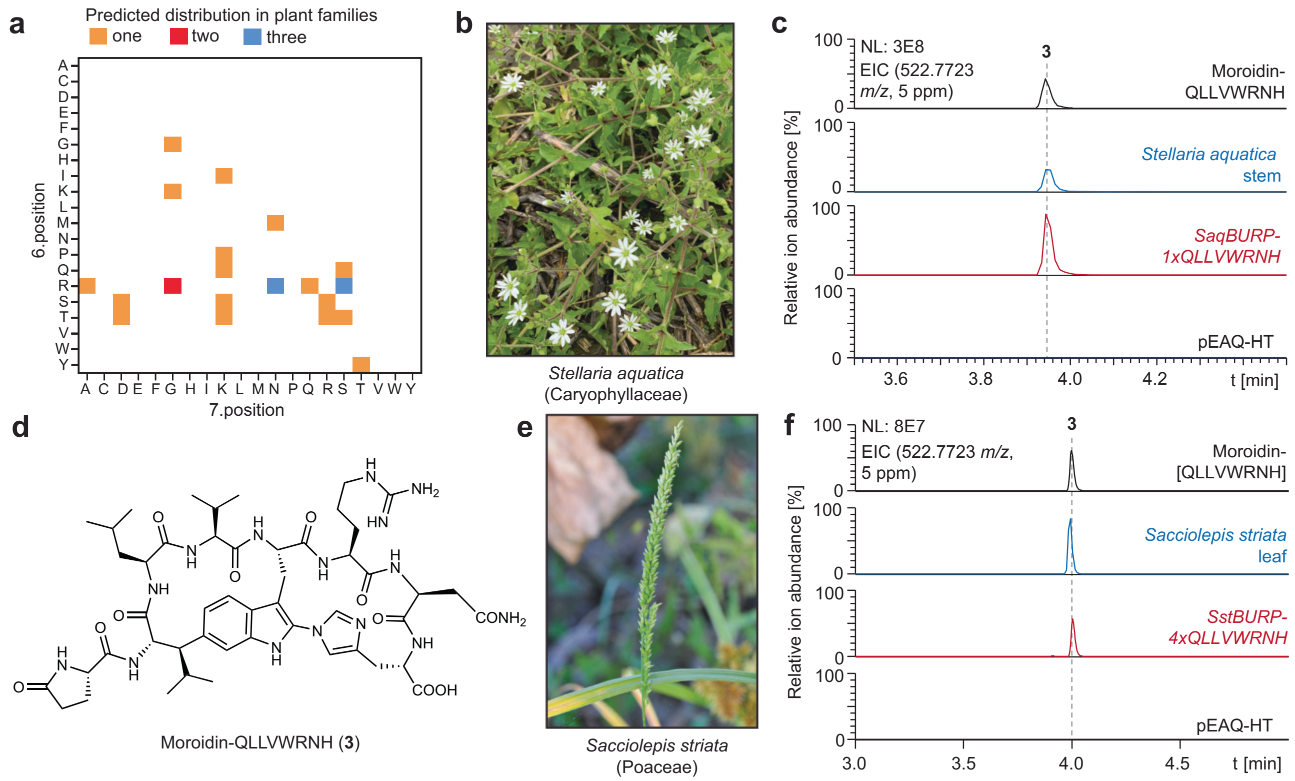


Fig. 4 | Characterization of moroidin burpitides in pinks and grasses. a Second-ring sequence diversity of moroidin cores in transcriptomics-derived fused burpitide cyclases. **b** *Stellaria aquatica*. **c** LC-MS characterization of moroidin-QLLVWRNH in *S. aquatica* and transgenic *N. benthamiana* after 6 days of expression of *SaqBURP-1xQLLVWRNH*. **d** Moroidin-QLLVWRNH structure. **e** *Sacciolepis striata*. Photo credit: Larry Allain, U.S. Geological Survey. **f** LC-MS characterization of moroidin-QLLVWRNH in *S. striata* and in transgenic *N. benthamiana* after 6 days of expression of *SstBURP-4xQLLVWRNH*. EIC: extracted ion chromatogram. Source data are provided as a Source Data file.

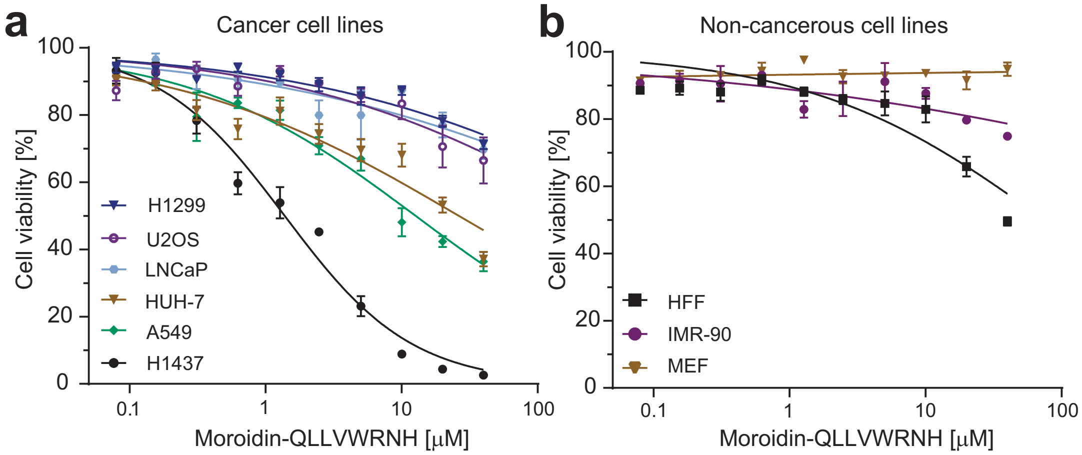
moroidin analog[18]. To evaluate selectivity and bioactivity, the peptide was orthogonally screened in three additional cancer cell lines (A549, LNCaP, and HUH-7) and three non-cancerous fibroblast lines (IMR-90, HFF, and MEF, Fig. 5a, b). Moroidin-QLLVWRNH was found inactive in H1299, U2OS, and LNCaP cells, but exhibited moderate dose-responsive cytotoxicity in A549 ($IC_{50} = 12.7\,\mu M$) and HUH-7 ($IC_{50} = 26.3\,\mu M$) cells. Notably, no significant dose-responsive cytotoxicity was observed in any non-cancerous fibroblast lines, underscoring its promising cancer cell specificity-an essential factor for clinical translation.

We identified additional moroidin analogs in *Sacciolepis striata* (Poaceae) and *Orchidantha maxillarioides* (Lowiaceae), two plant families without reported moroidin chemotypes. A fused burpitide cyclase with core peptides for moroidin-QLLVWRNH was predicted in the *S. striata* transcriptome (Fig. 4e) and moroidin-QLLVWRNH was confirmed in the methanolic extract from a herbarium *S. striata* specimen, matching the chemotype of moroidin-QLLVWRNH from *S. aquatica* and KjaBURP-1xQLLVWRNH (Fig. 4f). Similarly, four moroidin core peptides predicted in a transcript from the *O. maxillarioides* transcriptome and sequenced from cloned burpitide cyclase gene *OmaBURP2* were verified as analytes in methanolic leaf extracts of *O. maxillarioides* by LC-MS/MS (Supplementary Figs. 45–48). Two of these analytes were confirmed as moroidins through LC-MS comparison with methanolic extracts of transgenic *N. benthamiana* expressing OmaBURP2 or KjaBURP with one core peptide of QLLVWRSH or QLLVWPKH (Supplementary Figs. 45–48).

Discussion

Here we applied de novo plant transcriptome assembly and transcriptome mining for large-scale bioinformatic discovery of bioactive stephanotic acid-type burpitides. Based on our benchmarking analysis, SPAdes emerged as the most effective assembler for scaled transcriptome assembly due to a balance of high speed, high accuracy and low computational cost. MEGAHIT is faster, but not as accurate in de novo assembly of precursor transcripts as SPAdes, likely due to its design to maximize contig length during metagenomic assembly. SPAdes was best in assembly of repetitive plant RiPP precursor genes given their high number of detected core peptides in sequence repeats compared to other tested NGS assemblers. SPAdes has been applied in transcriptome mining of burpitides from fused and split precursor peptides but at one order of magnitude smaller data scale and with smaller outcome in terms of characterized burpitide chemistry (Supplementary Table 14).

Our transcriptome mining of fused burpitide cyclases utilized most plant species with RNA-seq data in the NCBI SRA and revealed a large diversity of second ring analogs of the stephanotic acid scaffold with the core peptide QLLVWRNH being found in three plant families and therefore representing an interesting candidate for biological evaluation. Moroidin-QLLVWRNH demonstrated selective dose-responsive cytotoxicity against H1437 lung cancer cells and exhibited limited toxicity in non-cancerous fibroblasts. These findings propose moroidin-QLLVWRNH as a lead structure with a specific mechanism-of-action in non-small cell lung cancer, offering a putative target for rational drug design based on the stephanotic acid scaffold. Furthermore, our study showed that other stephanotic acid-type burpitides exist in nature which utilize a tryptophan or tyrosine as the C-terminal residue for crosslinking in the second ring. Although glechomanin and mercurialin did not exhibit in vitro dose-responsive cytotoxicity (Supplementary Fig. 49), their characterized cyclases could increase the number of biosynthetically feasible second-ring analogs of moroidin in the future. The minimal cytotoxicity of glechomanin and mercurialin against H1437 lung adenocarcinoma cells



Cell line	IC_{50} [μM]
H1299	>40
U2OS	>40
LNCaP	>40
HUH-7	26.3
A549	12.7
H1437	1.4
HFF	>40
IMR-90	>40
MEF	>40

Fig. 5 | Bioactivity of moroidin-QLLVWRNH. a In vitro anticancer assays of moroidin-QLLVWRNH. Cell viability in six human cancer cell lines is plotted for the peptide across ten concentrations: H1299 (non-small cell lung cancer), U2OS (osteosarcoma), LNCaP (prostate cancer), HUH-7 (hepatoma), A549 (lung carcinoma), and H1437 (non-small cell lung adenocarcinoma). The data points represent the mean of triplicate samples ($n = 3$) and error bars represent one standard error of the mean. **b** In vitro counterscreen of moroidin-QLLVWRNH in three non-cancerous cell lines is plotted for the peptide across ten concentrations: HFF (human foreskin fibroblast), IMR-90 (human lung fibroblast), and MEF (mouse embryonic fibroblast). The data points represent the mean of triplicate samples ($n = 3$) and error bars represent one standard error of the mean. Corresponding IC_{50} values of (**a** and **b**) are listed. Source data are provided as a Source Data file.

could be due to the deviation from the QLLVW sequence of the moroidin first ring, a 7 aa-core peptide, or no second ring crosslink between the C-terminal residue and the central tryptophan-5-indole like in moroidins. Screening additional stephanotic acid-type cyclases with 8 aa-core peptides and C-terminal tyrosine or tryptophan residues could enable access to these stephanotic acid scaffolds.

Our study expands stephanotic acid-type burpitide chemistry by introduction of two bicyclic scaffolds and we uncover a broader chemotaxonomic distribution of stephanotic acid-type burpitides (Fig. 6, Supplementary Table 15) such as stephanotic acids in pinks, mints, spurges, grasses and Lowiaceae (Supplementary Figs. 13, 30 and 50); moroidins in pinks, grasses and Lowiaceae (Fig. 4c, f, and Supplementary Figs. 45–48); glechomanins in mints and Lowiaceae (Fig. 2b, f), and mercurialins in spurges (Fig. 3b). The Val-Trp- and Phe-Tyr-crosslinks of glechomanin and mercurialin represent new post-translational modifications in RiPP biosynthesis. C-C-crosslinks between amino acid side chains, such as the Leu-Trp-bond in the stephanotic acid scaffold and the Val-Trp-bond in glechomanin, are a common occurrence in cyclic peptide natural products. Glechomanin highlights that burpitide C-C-crosslinks usually include an unactivated carbon[15,30,36] which is similar to C-C-macrocyclic bonds in bacterial radical SAM-derived RiPPs[77–81]. In contrast, bacterial CYP450-peptide cyclases generate peptide macrocycles largely by C(sp^2)-C(sp^2)-bonds[82–89], if C-C-crosslinks are formed, with few C(sp^3)-C(sp^2)-exceptions[89,90] (Supplementary Fig. 51). Analogous to moroidin[15], a biosynthetic proposal for glechomanin and mercurialin peptides in *G. hederacea* and *M. annua* could be established based on transient expression studies of GheBURP and ManBURP in *N. benthamiana*, as well as in vitro reconstitution of GheBURP and ManBURP (Supplementary Fig. 52). Initially, GheBURP is translated by the ribosome, yielding the precursor peptide with an N-terminal domain containing three repeated core peptides for glechomanin. Subsequently, the BURP domain catalyzes the bicyclization of the core peptides within the N-terminal domain, serving as a substrate, in the presence of copper ions. The modified N-terminus will then undergo proteolytic cleavage by endopeptidases, resulting in the formation of a glechomanin analog with an N-terminal glutamine. This residue can cyclize either spontaneously or through a glutamine cyclotransferase to pyroglutamate[29]. Finally, exopeptidase cleavage facilitates the maturation of the C-terminus of glechomanin. Mercurialin biosynthesis occurs most likely through a similar mechanism from ManBURP in *M. annua* which features the formation of a cyclopeptide alkaloid crosslink, i.e., a tyrosine-phenol ether bond, after formation of the stephanotic acid scaffold.

In summary, our study informs the selection of current de novo NGS assemblers to efficiently address a major bottleneck in gene-guided discovery of plant natural products from RNA-seq data, i.e., the generation of large searchable genetic data. Since plants are eukaryotic organisms spanning a wide range of genome sizes from 61 Mbp to 160 Gbp[91,92] and plant RiPP precursor genes are structurally diverse[33], our results of plant transcriptome assembly should be applicable to other eukaryotic RNA-seq data for transcriptomics-guided peptide discovery[93]. Furthermore, our study can inspire next steps in development of stephanotic acid-type burpitide leads through discovery of a bioactive moroidin analog for mechanism-of-action studies and moroidin cyclases for further second-ring diversification.

Methods

Materials

All chemicals and reagents were from Thermo Fisher Scientific Inc. if not otherwise noted. All LC-MS solvents were Optima LC-MS-grade solvents (Fisher Scientific Inc.). All synthetic gene fragments were from Twist Biosciences Inc. Deoxynucleotide primers were from Integrated DNA Technologies Inc. LC-MS analysis was performed on a Thermo QExactive Orbitrap mass spectrometer coupled to a Vanquish Ultra high performance liquid chromatography instrument. LC-MS data was analyzed with QualBrowser from Thermo Xcalibur software package (v4.3.73.11, Thermo Scientific). Preparative and semi-preparative high performance liquid chromatography (HPLC) was performed on a Shimadzu HPLC with two binary LC20-AP pumps, a DGU-403 Degasser, an SPD-20A ultraviolet-visible light detector, and a FRC-10A fraction collector. NMR analysis was done on a Bruker Ascend 800 MHz NMR spectrometer with a 5 mm Triple Resonance Inverse Detection TCI CryoProbe. NMR data was analyzed with MestReNova (v14.0.0, Mestrelab Research).

Plant cultivation

Nicotiana benthamiana and *Glechoma hederacea* were grown from seeds in Premier Horticulture Pro Mix HP High Porosity with Mycorrhizae for plant growth under plant growth lights with a 16 h light/8 h dark cycle for 4–12 weeks at room temperature. *N. benthamiana* seedlings were transferred once to larger pots with the same soil and

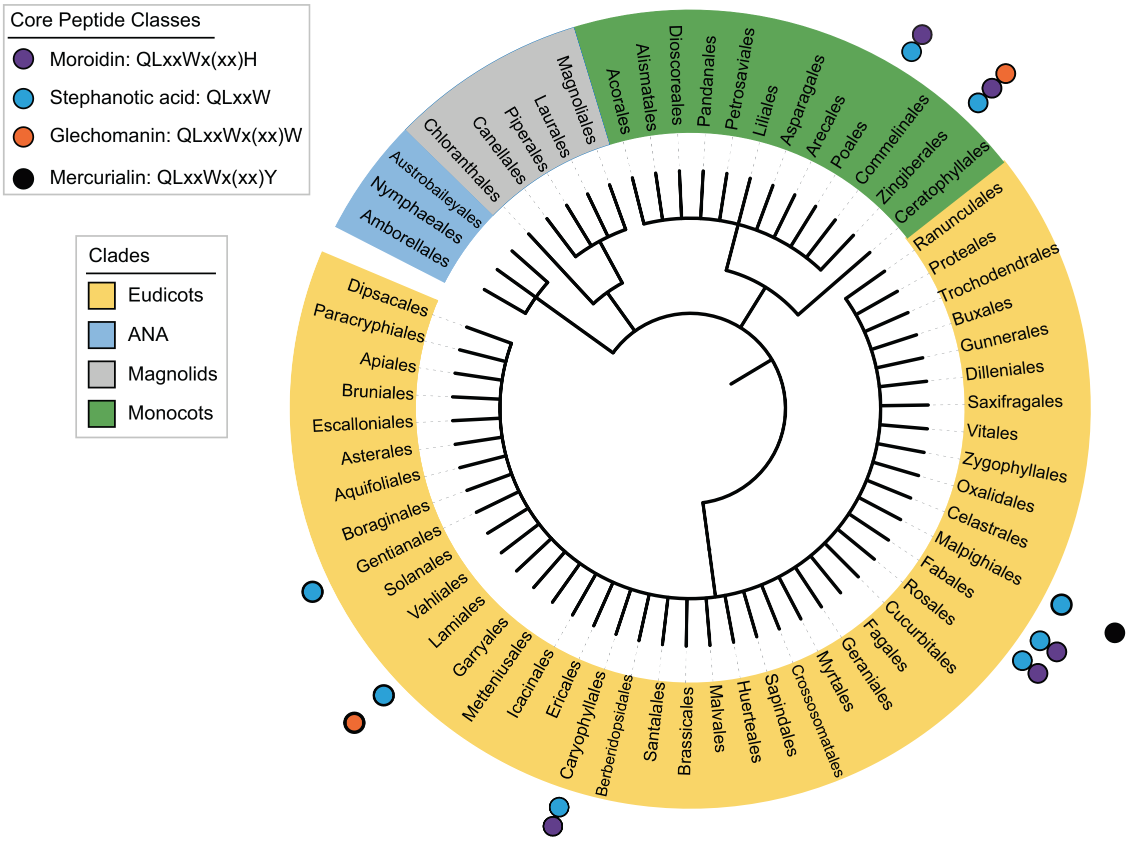


Fig. 6 | Phylogenetic distribution of characterized stephanotic acid-type burpitides in flowering plants. Displayed phylogenetic resolution are plant orders. See Supplementary Table 15 for species information. ANA: Amborellales, Nymphaeales, and Austrobaileyales clade; x: any proteinogenic amino acid.

added Osmocote fertilizer. *G. hederacea* seeds were obtained from the Seed Corner Inc. (Cat #: GROUNDIVYS100). *Sacciolepis striata* was collected from a sample in the University of Michigan Herbarium (Cat # 1423220). *Glechoma hederacea* for glechomanin isolation was collected from the University of Michigan Arboretum (42°16'32.7"N 83°43'27.2"W, June 2022/2023). *Stellaria aquatica* for moroidin-QLLVWRNH isolation and RNA isolation was collected from Pointe Mouillee State Game Area (41°59'54.8"N 83°11'47.7"W) with Department of Natural Resources permit 24–25 (June 2024). *Mercurialis annua* was collected from a sample in the University of Michigan Herbarium (Cat # 1642848). *Orchidantha maxillarioides* (accession 1989-4177-1) for peptide chemotyping and RNA isolation was collected from the Missouri Botanical Garden (June 2024). *Bauhinia forticata* for stephanotic acid-QLLVW chemotyping was purchased from Plant Delights Nursery Inc. *Kerria japonica* for stephanotic acid-QLLVW chemotyping was purchased as a mature plant from Green Promise Farms^{TM}.

RNA-seq data download and trimming

Paired-end RNA-seq data (Supplementary Table 4, Supplementary Data 2) were downloaded from NCBI SRA via fasterq-dump (--split-files) of the SRA Toolkit (v2.10.9)[19] to the Great Lakes High Performance Computing (HPC) Cluster at the University of Michigan, Ann Arbor. The datasets were trimmed via TrimGalore (v0.6.7) with default settings (Phred cutoff: 20, Default pair-cutoff: 20 bp).

Unassembled RNA-seq data search for stephanotic acid core peptides

Unassembled RNA-seq datasets of 890 plant species (Supplementary Table 3, Supplementary Data 1) were either searched for nucleotide sequence CAACTTCTTGTTTGG encoding stephanotic acid core peptide QLLVW (from the *Kerria japonica* moroidin cyclase KjaBURP, GenBank ID QIG55799.1) via blastn on the NCBI BLAST webserver (https://blast.ncbi.nlm.nih.gov/Blast.cgi, default parameters: matrix BLOSUM62, gap-open 11, gap-extended 1, Filter L, e-value 0.05) or for protein sequence of core peptide QLLVW via seqkit2 (v2.3.0)[49] on the Great Lakes HPC cluster. For webserver search, each SRA dataset was searched individually for the target nucleotide sequence with a maximum search hit number of 100. For seqkit2 search, the 890 species datasets were downloaded and trimmed in batches of 100 datasets. Duplicate reads were then removed from each dataset via seqkit2 rmdup command and the resulting fastq datasets were converted to fasta datasets via seqkit2 fq2fa command. The output fasta files were combined into one fasta-file for each 100-transcriptomes-batch and they were 6-frame translated via orfipy (v0.0.4, parameters: --pep --min 90 --between-stops)[94]. The orfipy output pep files were reduced in fasta-header size via awk command and the formatted pep file was searched via seqkit2 grep command for the presence of translated reads with the core peptide QLLVW. The identified reads were extracted from the seqkit2 grep output file for their respective SRA numbers with awk command. Unassembled data was converted to BLAST databases by makeblastdb and searched by commandline PHI-BLAST (BLAST+ v2.16.0, default settings, E-value 10)[50] for the presence of BURP-domain proteins including stephanotic acid core peptide QLLVW with KjaBURP GenBank accession QIG55799.1) as a query sequence. For comparison to assembled data, the same datasets were assembled de novo by SPAdes (v3.15.5, rnaSPAdes), combined and 6-frame translated by orfipy (v0.0.4, parameters: --pep --min 450 --between-stops), converted to BLAST databases by makeblastdb and searched by commandline PHI-BLAST (BLAST+ v2.16.0, default settings, E-value 10) for the presence of BURP-domain proteins including stephanotic acid core peptide QLLVW with KjaBURP. The ORF length of 90 bp was chosen for unassembled data 6-frame translation due to recommended PHI-BLAST database length minimum of 30 amino acids and a common RNA-seq read length of 90 bp. For benchmarking of stephanotic acid prediction via unassembled or assembled RNA-seq data searches, a deposited stephanotic acid-[LV] tandem mass spectrum (GNPS: CCMSLIB00006709935) was searched via plantMASST[48] (v2025.3.25, parameters: PM Tolerance (Da) 0.05, Fragment Tolerance (Da) 0.05, Cosine Threshold 0.7, Minimum Matched Peaks 3) in a metabolomic database including all 890 plant species searched in their RNA-seq data. False discovery rate was calculated as [(number of false positives)/((number of false positives) + (number of true positives))].

Transcriptome assembly

For de novo assembly benchmarking, datasets (Table 1) were trimmed via TrimGalore (v0.6.7) with default settings (Phred cutoff: 20, Default pair-cutoff: 20 bp) and assembled de novo with Trinity (v2.15.1)[23], SPAdes (v3.15.5)[24,25], or MEGAHIT (v1.2.9)[31] with the parameters in scripts specified on a GitHub repository (https://github.com/UM-KerstenLab4009/Moroidin-Transcriptomics). The working memory for all benchmarking datasets was 48 GB except for Trinity assembly of SRR7440026 data (80 GB memory), which failed at 48 GB memory. For large-scale de novo assembly, the datasets were assembled with SPAdes on the Great Lakes HPC cluster with the same parameters as for benchmarking. Assemblies which failed at 48 GB memory were assembled at 180 GB memory (Supplementary Table 4, Supplementary Data 2). For genome-guided assembly benchmarking, datasets (Table 2) were trimmed via TrimGalore (v0.6.7) with default settings. For each RNA-seq dataset, reads were aligned and mapped to the genome assembly fasta file with STAR (v2.7.11a) using the genome-specific gtf/gff3 annotation file. STAR alignment of a genome was done with the parameter –sjdbOverhang RNA-seq read-length minus 1. The resulting sam file was converted to bam file via samtools (v1.21). The transcriptome was assembled from the bam file either by StringTie (v2.2.1) or Trinity (v2.15.1). For genome-guided transcriptome assembly with Trinity, the parameter genome_guided_max_intron 50,000 was used. For StringTie transcriptome assembly, the final gtf file was converted with gffread (v0.12.7) to a fasta file.

Transcriptome analysis

For BLAST library preparation, fasta headers of all de novo assemblies and genome-guided assemblies were added with their SRA number and reduced in length below 51 letters. Precursor assembly benchmarking was done via Sequenceserver-based BLAST search: De novo assembled transcriptomes were searched for their corresponding precursor peptides via tblastn (v2.16.0 +)[71] on Sequenceserver (v3.1.0)[72] with the following BLAST parameters: e-value 1e-5, matrix BLOSUM62, gap-open 11, gap-extend 1, filter L. The top hit of each search was translated into protein sequence and aligned by Needle[95] pairwise alignment with the target precursor peptide. The resulting similarity and identity scores of all 36 precursor peptides were recorded (Table 1), added and compared to the maximum score (3600 = 36*100%) (Table 1). Large-scale search of BURP-domain transcripts was performed either via Sequenceserver BLAST and subsequent RepeatFinder analysis or via commandline PHI-BLAST and subsequent RepeatFinder analysis. For Sequenceserver-based BLAST search, all SPAdes de novo assembled transcriptomes (Supplementary Table 4, Supplementary Data 2) were combined into 100 transcriptome-databases and searched via tblastn (BLAST+ v2.16.0 +) for homologs of Sali3-2 as a BURP-domain-containing protein query (GenBank ID AAB66369) on Sequenceserver (v3.1.0) with the following BLAST parameters: E-value 1e-05, matrix BLOSUM62, gap-open 11, gap-extend 1, filter L, max, max_target_seqs 5000. tblastn hits from all databases were combined into a fasta-file and translated into protein sequences via orfipy (v0.0.4)[94] with the following parameters: --pep --min 450 --between-stops. The resulting protein sequence pep file was converted to a fasta file and analyzed by standalone RepeatFinder[36] on the Great Lakes HPC Cluster at the University of Michigan, Ann Arbor (see GitHub repository for RepeatFinder settings). Candidate stephanotic acid-type burpitide cyclases were characterized in RepeatFinder output as core QLxxW, where x is any proteinogenic amino acid, and presence in an open reading frame, which encoded a BURP-domain motif. For commandline BLAST search, all SPAdes de novo assemblies (Table 1) were combined into a fasta file and then translated into protein sequences via orfipy (v0.0.4)[94] with the following parameters: --pep --min 450 --between-stops. The resulting pep file was searched via PHI-BLAST (BLAST+ v2.16.0 +) with phi_pattern QQL-(x)2-W and KjaBURP (GenBank ID QIG55799.1) on default settings (E-value 10). The resulting hits were analyzed by RepeatFinder. In our searches, a BURP-domain motif was defined by the presence of at least half the BURP-domain-conserved residues (i.e., FF/Fx-C-C-CH of FF/Fx-C-C-CH-CH-CH-CH)[37]. Hits with a single QLxxW core peptide at the first N-terminal α-helix of BURP-domain proteins[74] were removed. PHI-BLAST results (Supplementary Table 4, Supplementary Data 2) also included predicted stephanotic acid-type core peptides from transcripts without or incomplete BURP-domain sequence. Reduction of the PHI-BLAST E-value 10 to 1e-5 had no effect on search results from fused burpitide cyclases but it reduced hits from incomplete or non-BURP-domain transcripts (Supplementary Table 5, Supplementary Data 3). In order to search for other core peptide motifs within BURP-domain proteins, the phi_pattern should be changed accordingly and a query BURP-domain protein sequence should be selected which includes the target motif, e.g., for a moroidin (core: QLLVWRGH) PHI-BLAST search, use the phi_pattern QQLLVWRGH and KjaBURP (GenBank ID QIG55799.1) which includes a QLLVWRGH motif in its sequence.

Peptide chemotyping

Leaf and stem plant material (0.2 g fresh weight) was collected from aerial tissues, frozen and ground with a MP Biomedicals FastPrep-24-5G Tissuelyser in 2-mL MP Biomedicals tubes with 2.3-mm Zirconia beads for 60 s at 6 m/s. 1 mL of 80% methanol was added to ground tissue and ground for 20 s at 6 m/s. The methanolic extract was incubated for 10 min in a 60 °C water bath, centrifuged for 5 min at 16,000 × *g* at room temperature, and filtered through Whatman 0.2 μm syringeless LC-MS filters (UN503NPEORG) before LC-MS analysis. Filtered extracts were transferred to LC-MS vials and analyzed by LC-MS/MS with the following settings: LC-Phenomenex Kinetex® 2.6 μm C18 reverse phase 100 Å 150 × 3 mm LC column; LC gradient, solvent A, 0.1% formic acid; solvent B, acetonitrile (0.1% formic acid); 0 min, 10% B; 5 min, 60% B; 5.1 min, 95% B; 6 min, 95% B; 6.1 min, 10% B; 9.9 min, 10% B; 0.5 mL/min, 30 °C; MS, positive ion mode; full MS, resolution 70,000; mass range 400–1200 *m/z*; dd-MS^2 (data-dependent MS/MS), resolution 17,500; AGC target 1×10^5, loop count 5, isolation width 1.0 *m/z*, collision energy 25 eV and dynamic exclusion 0.5 s. Sample numbers were $n = 3$ (biological replicates) for each source plant extraction and each transient gene expression experiment in *N. benthamiana* and $n = 1$ for *Mercurialis annua* chemotyping from Herbarium specimen. The negative control of transient gene expression experiment in *N. benthamiana* was an empty pEAQ-HT vector infiltration experiment.

Glechomanin isolation

Mature *Glechoma hederacea* (root, leaves, stem, flowers, 10 kg total) was collected and frozen at −80 °C and ground with dry ice in a food processor to a fine powder. The plant powder was extracted with 1 L methanol (HPLC grade) per 500 g plant powder for 16 h at 37 °C at 160 rpm. The crude extract was vacuum-filtered through a silica filter and then dried *in vacuo*. The dried methanol extract was resuspended in 2 L of water. The resuspension was partitioned 1:1 (v/v) with hexanes (3x) and ethyl acetate (2x) and then extracted with n-butanol (1:1, v/v, 2x). The n-butanol extract was dried *in vacuo*, resuspended in 20% methanol and applied to a Sephadex LH20 (Sigma-Aldrich, LH20100) liquid chromatography (LC) solid phase column (60 × 8 cm) with a mobile phase gradient of 20%, 40%, 60%, 80% and 100% methanol (500 mL each). LH20 LC fractions were analyzed for the presence of the target analyte with the following LC-MS settings: injection volume 2 µL; LC, Phenomenex Kinetex 2.6 µm C18 reverse phase 100 Å 50 × 3 mm LC column; LC gradient. solvent A, 0.1% formic acid; solvent B, acetonitrile (0.1% formic acid); 0 min, 5% B; 2.5 min, 95% B; 3.0 min, 95% B; 3.1 min, 5% B; 5.0 min, 5% B, 0.5 mL/min, 30 °C; MS, positive ion mode; full MS, resolution 35,000, mass range 400–1200 *m/z*; dd-MS^2, resolution 17,500; loop count 5, collision energy 25 eV and dynamic exclusion 0.5 s. LH20 LC fractions with glechomanin were combined and dried *in vacuo*.

Glechomanin extracts were separated by preparative HPLC with the following LC conditions in two rounds: 1. separation: Phenomenex Kinetex® 5 µm C18 100 Å LC Column 150 × 21.2 mm, LC gradients: solvent A - 0.1% TFA, solvent B - acetonitrile (0.1% TFA), 7.5 mL/min, 1. separation: 0 min: 10% B, 1 min: 10% B, 36 min: 50% B, 39 min: 95% B, 42 min: 95% B, 42.5 min: 10% B, 60.1 min: 10% B. 2. separation: Phenomenex Kinetex® 5 µm C18 100 Å LC Column 150 × 21.2 mm, 7.5 mL/min, 0 min: 15% B, 1 min: 15% B, 36 min: 35% B, 39 min: 95% B, 42 min: 95% B, 42.5 min: 15% B, 60.1 min: 15% B. Preparative HPLC fractions were analyzed for glechomanin described above. Subsequently, glechomanin fractions were combined, dried *in vacuo* and dissolved in 20% acetonitrile. The glechomanin fraction was separated by semipreparative HPLC with the following conditions: 1. separation: Kinetex® 5 µm C18 100 Å 250 × 10 mm column, solvent A - 0.1% TFA, solvent B - acetonitrile (0.1% TFA), 1.5 mL/min, 0 min - 20% B, 1 min - 20% B, 21 min - 30% B, 22 min - 95% B, 25 min - 95% B, 25.1 min - 20% B, 45 min - 20% B. 2. separation: solvent A - 0.1% TFA, solvent B - methanol (0.1% TFA), 1.5 mL/min, 0 min - 65% B, 1 min - 65% B, 30 min - 85% B, 31 min - 100% B, 36 min - 100% B, 37 min - 65% B, 47 min - 65% B. Subsequently, glechomanin fractions were combined and dried *in vacuo*. Glechomanin (2 mg) was a white powder.

Moroidin-QLLVWRNH isolation

Stellaria aquatica aerial tissue (1.5 kg) was frozen at −80 °C and ground in a food processor with dry ice and extracted in 6 L of methanol for 16 h at 37 °C with agitation (160 rpm). The methanol extract was filtered over silica and dried under reduced pressure. The dried extract was resuspended in 2 L of water and partitioned sequentially with 2 L of hexanes (3x), ethyl acetate (2x), and n-butanol (2x). The n-butanol fraction was dried, resuspended in 10% methanol, and subjected to Sephadex LH20 chromatography (60 × 8 cm) using a methanol gradient (10%, 20%, 30%, 40%, 50%; 500 mL each). Fractions were analyzed by LC-MS with the following parameters: injection volume, 2 µL; column, Phenomenex Kinetex® 2.6 µm C18 100 Å (50 × 3 mm); solvent A, 0.1% formic acid in water; solvent B, acetonitrile (0.1% formic acid); 0.5 mL/min, 30 °C; gradient: 0 min, 5% B; 2.5 min, 95% B; 3.0 min, 95% B; 3.1 min, 5% B; 5.0 min, 5% B. MS settings: positive ion mode; full MS resolution, 35,000; mass range, 400–1200 *m/z*; dd-MS^2 resolution, 17,500; collision energy, 25 eV; dynamic exclusion, 0.5 s. Moroidin-QLLVWRNH-containing fractions were combined, dried *in vacuo*, and then separated by preparative HPLC with the following LC conditions in two rounds: Phenomenex Kinetex® 5 µm C18 100 Å LC column 150 × 21.2 mm; 1. separation: LC gradients: solvent A - 0.1% TFA in water, solvent B - acetonitrile (0.1% TFA), 7.5 mL/min, 1. separation: 0 min: 10% B, 1 min: 10% B, 36 min: 50% B, 39 min: 95% B, 42 min: 95% B, 42.5 min: 10% B, 60.1 min: 10% B. 2. separation: 7.5 mL/min, 0 min: 15% B, 1 min: 15% B, 36 min: 35% B, 39 min: 95% B, 42 min: 95% B, 42.5 min: 15% B, 60.1 min: 15% B. Fractions were analyzed for moroidin-QLLVWRNH by LC-MS as for LH20 LC and combined. Final purification was achieved by semipreparative HPLC (Kinetex® 5 µm C18 100 Å 250 × 10 mm, solvent A: 0.1% TFA in water, solvent B: acetonitrile, 0.1% TFA, 1.5 mL/min: 1. Separation: 0 min - 20% B, 1 min - 23% B, 45 min - 23% B), 2. Separation: 0 min - 23% B, 30 min - 23% B. Moroidin-QLLVWRNH (32 mg) was obtained as a white powder.

Mercurialin isolation

Transgenic tobacco leaves expressing ManBURP-1xQLFFWRY (1.3 kg) were frozen at −80 °C and ground in a food processor with dry ice and extracted in 8 L of methanol for 16 h at 37 °C with agitation (160 rpm). The methanol extract was filtered over silica and dried *in vacuo*. The dried extract was resuspended in 2 L of water and partitioned sequentially with 2 L of hexane (3x), ethyl acetate (2x), and n-butanol (2x). The n-butanol fraction was dried, resuspended in 10% methanol, and subjected to Sephadex LH20 chromatography (60 × 8 cm) using a methanol gradient (20%, 40%, 60%, 80%, 100%; 500 mL each). Fractions were analyzed by LC-MS with the following parameters: injection volume, 2 µL; column, Phenomenex Kinetex® 2.6 µm C18 100 Å (50 × 3 mm); solvent A, 0.1% formic acid in water; solvent B, acetonitrile (0.1% formic acid); 0.5 mL/min, 30 °C; gradient: 0 min, 5% B; 2.5 min, 95% B; 3.0 min, 95% B; 3.1 min, 5% B; 5.0 min, 5% B. MS settings: positive ion mode; full MS resolution, 35,000; mass range, 400–1200 *m/z*; dd-MS^2 resolution, 17,500; collision energy, 25 eV; dynamic exclusion, 0.5 s. Mercurialin-containing fractions were combined, dried *in vacuo*, and then separated by preparative HPLC with the following LC conditions in two rounds: Phenomenex Kinetex® 5 µm C18 100 Å LC Column 150 × 21.2 mm, solvent A - 0.1% TFA in water, solvent B - acetonitrile (0.1% TFA); 1. separation: LC gradients:, 7.5 mL/min, 1. separation: 0 min: 10% B, 1 min: 10% B, 36 min: 50% B, 39 min: 95% B, 42 min: 95% B, 42.5 min: 10% B, 60.1 min: 10% B. 2. separation: 7.5 mL/min, 0 min: 20% B, 1 min: 20% B, 36 min: 40% B, 39 min: 95% B, 42 min: 95% B, 42.5 min: 20% B, 60.1 min: 20% B. Fractions were analyzed for mercurialin by LC-MS and combined. Final purification was achieved by 2 rounds of semi-preparative HPLC (Kinetex® 5 µm C18 100 Å 250 x 10 mm column, solvent A: 0.1% TFA in water, solvent B: acetonitrile, 0.1% TFA, 1.5 mL/min; 1. and 2. separation: 0 min - 38% B, 35 min - 38% B. Mercurialin (5 mg) was obtained as a white powder.

Structure elucidation of glechomanin and mercurialin

For 1D and 2D NMR analysis, glechomanin was dissolved in DMSO-d_6 or methanol-d_4 (Cambridge Isotope Laboratories, Inc.), whereas mercurialin and moroidin-QLLVWRNH were prepared in DMSO-d_6. All data were acquired using corresponding Shigemi NMR tubes (Wilson Glass). To compare ROESY/NOESY data with geometry-optimized structures of glechomanin and mercurialin, geometry optimizations were performed for the Trp7-Cα stereoisomers of glechomanin and the Phe-Cβ stereoisomers of mercurialin. Initially, the four stereoisomer structures were energy-minimized using MM2 followed by MMFF94 calculations in Chem3D (v.22.2.0, Revvity Signals, maximum iterations set to 10,000). The resulting conformers were subsequently geometry-optimized using the B3LYP-D3/6-31 + G(d,p) method in Gaussian 16 (revision A.03)[96] on the Great Lakes HPC cluster and analyzed in Chem3D.

For Marfey's analysis, ~400 µg of purified glechomanin or mercurialin were hydrolyzed in 600 µL of 6 M hydrochloric acid (HCl) in a sealed thick-walled reaction vessel. The sample was heated at 110 °C and stirred overnight for 12 h. Subsequently, the hydrolysate was concentrated to dryness under nitrogen gas and redissolved in 100 µL water. Then, 100 µL of 1 M sodium bicarbonate ($NaHCO_3$) and 100 µL of a 1% acetone solution of Marfey's reagent, 1-fluoro-2,4-dinitrophenyl-5-L-alanine amide (L-FDAA), were added to the solution. The reaction mixture was incubated at 40 °C for 1 h and then quenched by adding 100 µL of 1 M HCl. For the preparation of L-FDAA-amino acid standard derivatives, 50 mM of each amino acid (D/L-Phe, D/L-Glu, D/L-Leu, D/L-Trp, D/L-Val, D/L-Phe, D/L-Arg and D/L-Tyr) dissolved in water (50 µL) was treated with 1 M $NaHCO_3$ (20 µL) and 1% L-FDAA (100 µL) at 40 °C for 1 h, respectively. After the reaction, the solution was quenched with 1 M HCl (20 µL) and diluted with acetonitrile (810 µL) for LC-MS analysis. LC-MS/MS analysis parameters were as follows: injection volume 2 µL; LC, Phenomenex Kinetex 2.6 µm C18 reverse phase 100 Å 150 × 3 mm LC column; LC gradient: solvent A, 0.1% formic acid; solvent B, acetonitrile (0.1% formic acid); 0 min, 10% B; 40 min, 60% B; 41 min, 95% B; 44 min, 95% B; 45 min, 10% B, 50 min, 10% B, 0.5 mL/min, 30 °C; MS, positive ion mode; full MS, resolution 70,000, mass range 100–1000 *m/z*. For the determination of D/L-Arg, the following LC-MS/MS parameters were employed: injection volume 5 µL; LC, Higgins Analytical PROTO300 C4 5 µm 250 × 4.6 mm LC column; LC gradient: solvent A, 0.1% formic acid; solvent B, acetonitrile (0.1% formic acid); 0 min, 10% B; 38 min, 20% B; 39 min, 95% B; 45 min, 95% B; 46 min, 10% B, 56 min, 10% B, 0.5 mL/min, 30 °C; MS, positive ion mode; full MS, resolution 70,000, mass range 100–1000 *m/z*. The configurations of the amino acid residues were determined by comparing their retention times and elution orders with those of FDAA derivatives of amino acid standards.

BURP-domain gene cloning

Leaves of 10-week-old *Glechoma hederacea*, stem tissue of *Stellaria aquatica*, and leaves of *Orchidantha maxillarioides* were collected and total RNA was extracted from each plant with QIAGEN RNeasy Plant Mini Kit. Complementary DNA was prepared from *G. hederacea* leaf total RNA, *S. aquatica* stem total RNA, and *O. maxillarioides* leaf total RNA with Thermo Scientific™ Maxima H Minus First Strand cDNA Synthesis Kit. Cloning primers for *GheBURP*, *SaqBURP, and OmaBURP1/2* were designed from candidate *GheBURP* transcripts, *SaqBURP* transcripts, and *OmaBURP1/2* transcripts, respectively (Supplementary Table 16). *GheBURP* and *OmaBURP1/2* were amplified with Phusion High-Fidelity DNA polymerase (New England Biolabs) and cloned by Gibson cloning assembly[97] (New England Biolabs) into pEAQ-HT[73] linearized by restriction enzymes AgeI and XhoI. Cloned *GheBURP* were sequenced by Sanger sequencing. Similarly, *SaqBURP* was cloned into pHis8 (pET28a with N-terminal 8xHis-tag) linearized by restriction enzyme NcoI. Cloned *SaqBURP* was sequenced by Nanopore sequencing (Plasmidsaurus Inc).

Transient gene expression in *Nicotiana benthamiana*

SstBURP-1xQLLVWRNH, *SstBURP-4xQLLVWRNH*, *SaqBURP-1xQLLVWRNH*, *OmaBURP1-1xQLFVWGW*, *ManBURP-1xQLFFWRY*, *ManBURP-1xQLLVWRY*, *GheBURP-1xQLLVW*, *KjaBURP-1xQLLVWRGH*, *KjaBURP-1xQLLVWRNH*, *KjaBURP-1xQLLVWRSH*, and *KjaBURP-1xQLLVWPKH* (Supplementary Data 4) were synthesized with N-terminal overhang TGCCCAAATTCG CGACCGGT and C-terminal overhang CTCGAGGCCTTTAACTCTGG for Gibson cloning into pEAQ-HT[73] linearized with AgeI and XhoI. *Agrobacterium tumefaciens* LBA4404 was transformed with pEAQ-HT-*GheBURP* wildtype or synthetic constructs (*GheBURP-1xQLLVW*), pEAQ-HT-*KjaBURP* synthetic constructs (*KjaBURP-1xQLLVWRNH, KjaBURP-1xQLLVWRSH, KjaBURP-1xQLLVWPKH*), pEAQ-HT-*SstBURP* synthetic constructs (*SstBURP-1xQLLVWRNH, SstBURP-4xQLLVWRNH*), pEAQ-HT-*SaqBURP* synthetic construct (*SaqBURP-1xQLLVWRNH*), pEAQ-HT-*OmaBURP1* wild-type and synthetic constructs (*OmaBURP1-1xQLFVWGW*), pEAQ-HT-*OmaBURP2* wildtype construct and pEAQ-HT-*ManBURP* wildtype and synthetic constructs (*ManBURP-1xQLFFWRY*, *ManBURP-1xQLLVWRY*) by electroporation (2.5 kV) and then plated on YM agar plates (0.4 g of yeast extract, 10 g of D-mannitol, 0.1 g of sodium chloride (NaCl), 0.2 g of magnesium sulfate (heptahydrate), 0.5 g of potassium phosphate (dibasic, trihydrate), 15 g of agar, and 1 L of deionized water, adjusted to pH 7) with 100 µg/mL rifampicin, 50 µg/mL kanamycin, and 100 µg/mL streptomycin, and incubated for 2 days at 30 °C. A 5 mL starter culture of the YM medium with 100 µg/mL rifampicin, 50 µg/mL kanamycin, and 100 µg/mL streptomycin was inoculated with a clone of *A. tumefaciens* LBA4404 pEAQ-HT-*GheBURP* or other precursor gene constructs and incubated for 24–36 h at 30 °C in a shaking incubator at 220 rpm. Subsequently, the starter culture was used to inoculate a 25 mL culture of the YM medium with 100 µg/mL rifampicin, 50 µg/mL kanamycin, and 100 µg/mL streptomycin, which was incubated for 24 h at 30 °C in a shaking incubator at 220 rpm. The cells from the 25 mL culture were centrifuged for 30 min at 3000 × *g*, the YM medium was discarded, and cells were resuspended in MMA medium (10 mM MES KOH buffer (pH 5.6), 10 mM magnesium chloride ($MgCl_2$), 100 µM acetosyringone) to give a final optical density (OD_{600}) of 0.8. The *Agrobacterium* suspension was infiltrated into the bottom of the leaves of *N. benthamiana* plants (6-8 weeks old) using 1 mL plastic syringes. *N. benthamiana* plants were placed in the shade 2 h before syringe infiltration. After infiltration, *N. benthamiana* plants were grown as described above for 6 days. Subsequently, infiltrated leaves were collected and subjected to peptide chemotyping. For peptide purification, infiltrated leaves were collected after 7 days and stored at -80 °C until purification.

Protein purification

E. coli-codon-optimized *GheBURP-1xQLFVWGW* and *ManBURP-1xQLFFWRY* (Supplementary Data 4) were cloned via Gibson cloning into NcoI-linearized pHis8 expression vector (pET28a with an N-terminal 8xHis-tag) with 5′-overhang GAAAACTTGTACTTCCAGGC CCATGGC and 3′-overhang CCATGGCGGATCCGAATTCGAG. They were transformed into *E. coli* BL21(DE3) respectively. Each colony of *E. coli* BL21(DE3) was inoculated to a 10 mL Luria-Bertani medium with 50 µg/mL kanamycin for 16 h incubation at 37 °C and 200 rpm. The starter culture was used to inoculate 1 L Terrific Broth liquid medium with 50 µg/mL kanamycin and the culture was incubated at 37 °C and 200 rpm until OD_{600} reached 0.7-1. The culture was then added with Isopropyl-β-D-1-thiogalactopyranoside to a final concentration of 50 µg/mL and cooled to 18 °C for incubation at 200 rpm for 16 h.

Cells were centrifugated at 3500 × *g* and 4 °C. The cell pellet was resuspended in solution buffer (50 mM Tris-hydrochloride (Tris-HCl) (pH 8), 1 mM ethylenediaminetetraacetic acid (EDTA, pH 8), 1 mM dithiothreitol (DTT), 25% sucrose (w/v)) with 5 mL per g wet cell pellet. Resuspended cells were sonicated on ice to break cells open for 2 min 57 sec at 80% amplitude (59 s on, 59 s off). Subsequently, 1 mg lysozyme and 1 mg DNAse I is added per mL lysate and 25 µL divalent cation solution (400 mM $MgCl_2$, 400 mM calcium chloride ($CaCl_2$)) is added per mL lysate. The lysate was then stirred at room temperature for 20 min. Next, the same volume of lysis buffer (50 mM Tris-HCl (pH 8), 1 mM EDTA (pH 8), 1 mM DTT, 200 mM NaCl, 1.0 % Triton X-100 was added to the lysate and mixed for 1 h at 4 °C. The lysate was sonicated again on ice for 2 min 57 sec and 60% amplitude (59 s on, 59 s off) and then added with EDTA solution (pH 8, final concentration 20 mM). The inclusion bodies were subsequently checked for full resuspension. The resulting lysate was centrifugated for 30 min at 6000 × *g* and 4 °C. The supernatant was discarded and the inclusion bodies were resuspended in the same volume as the discarded supernatant of Wash Buffer with Triton X-100 (50 mM Tris-HCl (pH 8), 1 mM EDTA (pH 8), 1 mM DTT, 100 mM NaCl, 0.5% Triton X-100). The inclusion body resuspension was sonicated on ice for 2 min 57 sec and 60% amplitude (59 s on, 59 s off). The inclusion bodies are checked for full resuspension and centrifugated for 30 min at 6000 × *g* and 4 °C. The inclusion bodies were washed one more time with Wash Buffer with Triton X-100, before the inclusion bodies were washed twice with Wash Buffer without Triton X-100 (50 mM Tris-HCl (pH 8), 1 mM EDTA (pH 8), 1 mM DTT, 100 mM NaCl). After the final centrifugation, the supernatant was discarded and the inclusion body pellet was resuspended in TE buffer (10 mM Tris-HCl (pH 8), 1 mM EDTA (pH 8)) at 5 mL per g wet inclusion bodies and then aliquoted as 1 mL inclusion body resuspension for storage at -80 °C before further protein purification.

Unless otherwise specified, all subsequent procedures were conducted at room temperature. The denaturing solution was prepared in a 15 mL conical tube by mixing solid urea with 4 mL of 50 mM Tris-HCl (pH 8) and 100 mM NaCl, achieving a final concentration of 7.5–8.0 M, followed by the addition of 8 mM β-mercaptoethanol. Thawed inclusion bodies (1 mL aliquot) were then added to the denaturing solution and gently mixed. The mixture was inverted or pipetted every 10 min for 60 min until complete dissolution was achieved. Subsequently, the solution was filtered through a 0.45 μm syringe filter into a conical tube. The dissolved and filtered GheBURP-1xQLFVWGW/ManBURP-1xQLFFWRY was diluted into 500 mL of 50 mM Tris-HCl (pH 8), 500 mM NaCl, and 2.5 M urea to prevent aggregation during refolding. Before use, a 4 L stock of refolding buffer (50 mM Tris-HCl pH 8, 500 mM NaCl, 0.5 mM tris(2-carboxyethyl)phosphine (GoldBio), 10% glycerol) was cooled to 4 °C. The diluted GheBURP-1xQLFVWGW/ManBURP-1xQLFFWRY (each 500 mL) was then introduced into 3.5 kDa molecular weight cut-off (MWCO) dialysis tubing and dialyzed overnight at 4 °C in 4 L of refolding buffer. Following dialysis, the protein was removed from the tubing and centrifuged at 4000 × *g* at 4 °C for 30 min to eliminate aggregates prior to purification. Subsequently, 2.5 mL of Ni-NTA resin was added to a 20 mL column, washed with deionized water, and equilibrated with 10 column volumes (CV) of refolding buffer containing 30 mM imidazole at pH 8. The refolded protein was then applied onto the resin, and the flow-through was discarded. The resin was washed with 10 CV of refolding buffer containing 30 mM imidazole at pH 8, followed by elution of the protein with 10 CV of refolding buffer containing 600 mM imidazole at pH 8. A mixture containing 1 mg of TEV protease for every 10 mg of eluted protein was subjected to overnight dialysis in 3.5 kDa MWCO dialysis tubing at 4 °C with stirring in 4 L of TEV protease buffer (50 mM Tris-HCl (pH 8), 100 mM NaCl, 3 mM reduced glutathione (GSH), 0.3 mM oxidized glutathione (GSSG)). Subsequently, GheBURP-1xQLFVWGW/ManBURP-1xQLFFWRY was separated from uncleaved His$_8$-GheBURP-1xQLFVWGW/His$_8$-ManBURP-1xQLFFWRY, respectively, and the His-tagged TEV protease using a second 2.0 mL Ni-NTA column, which had been pre-equilibrated with 10 CV of TEV protease buffer in a disposable 20 mL column. The flow-through was then concentrated using Amicon Ultra-15mL 10 kDa MWCO centrifugal concentrators, subjected to SDS-PAGE analysis, and utilized for enzyme assays.

Enzyme assays

For core peptide analysis, enzyme assays were performed using 50 μL volumes containing either copper(II) chloride ($CuCl_2$) from JT Baker Chemical Co. or water. Assays for BURP-domain proteins comprised 300 μM GheBURP-1xQLFVWGW/100 μM ManBURP-1xQLFFWRY, 1 mM $CuCl_2$, and a citrate:phosphate buffer (pH 7) with final concentrations of 18 mM citrate and 165 mM Na_2HPO_4. The enzyme reactions were incubated for 24 h at 20 °C. Subsequently, 1 μg of trypsin (Sigma-Aldrich, T7575-1KT) in 20 μL of 1 M Tris-HCl (pH 8)/1 μg of LysC (NEB, P8109S) in 10 μL of 10 mM Tris-HCl (pH 8) buffer was added to GheBURP/ManBURP assay, respectively, and further incubated at 37 °C for 24 h. Each digest was centrifuged, added to Chromacol™ LC-MS vial, and subjected to LC–MS/MS analysis. LC-MS/MS settings: Injection volume 15 μL, LC-Higgins Analytical PROTO300 C4 5 μm 250 × 4.6 mm, solvent A-0.1% formic acid, solvent B-acetonitrile (0.1% formic acid), 0.7 mL/min, 30 °C, LC gradient: 0 min: 10% B, 10 min: 50% B, 11 min: 95% B, 13 min: 95% B, 13.1 min: 10% B, 20 min: 10% B, MS - Positive ion mode, Full MS: Resolution 35,000, mass range: 700-1250 *m/z* for GheBURP digest and 700-1450 *m/z* for ManBURP digest, dd-MS^2: resolution 17,500, AGC target 1e5, loop count 5, isolation window 0.4 *m/z*, collision energy 25 eV, dynamic exclusion 2 s. Scaled enzyme assays for exoproteolytic cyclic peptide formation were conducted on a 2 mL scale using purified GheBURP-1xQLFVWGW/ManBURP-1xQLFFWRY obtained from a 4 L/2 L expression and purification process. These assays included 130–350 μM of the BURP domain cyclase and 1 mM $CuCl_2$ in a citrate:phosphate buffer at pH 7. The assays were incubated for 24 h at 20 °C. Following incubation, trypsin/LysC was introduced into the GheBURP/ManBURP reaction mixtures at a protease-substrate molar ratio of 1:250/1:100, with the pH adjusted to 8 using a 1 M Tris-HCl (pH 8) buffer and incubated at 37 °C for 24 h. Two digests were subjected to preparative HPLC, respectively (LC settings, Phenomenex Kinetex 5 μm C18 100 Å LC Column 150 x 21.2 mm, 7.5 mL/min; solvent A, 0.1% TFA; solvent B, acetonitrile (0.1% TFA); LC gradient, 0 min, 10% B; 1 min, 10% B; 36 min, 50% B; 39 min, 95% B; 42 min, 95% B; 42.5 min, 10% B; 60.1 min, 10% B). HPLC fractions were analyzed by LC–MS analysis for target masses of modified core peptides with the following LC–MS parameters: injection volume 2 μL; LC, Phenomenex Kinetex 2.6 μm C18 reverse phase 100 Å 50 × 3 mm LC column; LC gradient. solvent A, 0.1% formic acid; solvent B, acetonitrile (0.1% formic acid); 0 min, 5% B; 2.5 min, 95% B; 3.0 min, 95% B; 3.1 min, 5% B; 5.0 min, 5% B, 0.5 mL/min, 30 °C; MS, positive ion mode; full MS, resolution 35,000, mass range 400–1200 *m/z*; dd-MS^2, resolution 17,500; loop count 5, collision energy 25 eV and dynamic exclusion 0.5 s. LC fractions containing target peptides were dried and resuspended in 30 μL DMSO for further analysis. Exopeptidase assays were conducted with ~25–50 μg of peptides (with 4% DMSO), carboxypeptidase Y (from *S. cerevisiae*, Sigma-Aldrich), and aminopeptidase N (from Rat, Sigma Aldrich) at a peptidase:peptides ratio of 1:10 (w/w). These assays were performed in 100 mM Tris-HCl buffer at pH 7 and 37 °C for 24 h. Following incubation, the assays were centrifuged and analyzed by LC-MS/MS as follows: Injection volume 15 μL, LC-Higgins Analytical PROTO300 C4 5 μm 250 × 4.6 mm, all gradients: solvent A-0.1% formic acid, solvent B-acetonitrile (0.1% formic acid), 0.7 mL/min, 30 °C, LC gradient: 0 min: 15% B, 60 min: 40% B, 61 min: 95% B, 63 min: 95% B, 63.1 min: 15% B, 80 min: 15% B, MS - Positive ion mode, Full MS: Resolution 35,000, mass range: 800–1300 *m/z* for GheBURP and 700–1450 *m/z* for ManBURP, dd-MS^2: resolution 17,500, AGC target 1e5, loop count 5, isolation width 0.4 *m/z*, collision energy 25 eV, dynamic exclusion 0.6 s. The analysis was compared to purified glechomanin or mercurialin standards.

Bioassay

Cancer cell lines H1437 (ATCC® CRL-5872), H1299 (ATCC® CRL-5803), A549 (ATCC® CCL-185), U2OS (ATCC® HTB-96), LNCaP (ATCC® CRL-1740), HUH-7 (JCRB® JCRB0403), and noncancerous cell lines IMR-90 (ATCC® CCL-186), HFF (ATCC® CRL-2088), and MEFs were maintained at 37 °C and 5% CO_2 atmosphere. Standard isolation and culture of primary mouse embryonic fibroblasts from pregnant female mice at the University of Michigan established the MEF cell line[98]. H1437, H1299 and LNCaP were grown in Roswell Park Memorial Institute 1640 Medium (Fisher Scientific, 11875119), U2OS were grown in McCoy's 5 A Medium (Thermo Fisher, 16600082), and the remaining lines grown in Dulbecco's Modified Eagle's medium (Fisher Scientific, 11995065). A549 cell media were supplemented with 15% Fetal Bovine Serum (FBS; Thermo Fisher, A5670701), while the remaining media were supplemented with 10% FBS and 1x penicillin-streptomycin (Gibco, 15140122).

384-well plates (Perkin Elmer, 6057300) were seeded with H1437, H1299, A549, U2OS, LNCaP, HUH-7, IMR-90, HFF, and MEF at 5000 cells per well in 50 μL of their respective complete media. Cell seeding densities were optimized such that the cells would be 90–100% confluent in DMSO vehicle control wells after 72 h. Cells were allowed to attach to the wells for 24 h at 37 °C and 5% CO_2. Compounds (10 mM) were added using the HPD300e digital compound dispenser in randomized, ten-point, two-fold dose response (n = 3) from a starting concentration of 40 μM. Compounds were normalized to the highest class DMSO volume of 1%, then incubated at 37 °C and 5% CO_2 for 48 h. Each cell line had $n = 16$ positive controls (10 μM taxol) and vehicle (DMSO). Cells were fixed with 4% paraformaldehyde for 20 min, permeabilized with 0.03% Triton X-100 for 20 min and rinsed twice with PBS between each step. Cells were then stained for 30 min with a 1:1000 PBS dilution of Hoechst 33340 (Thermo Scientific, H1399) for nuclei staining and 1:10,000 PBS dilution of CellMask Orange (Thermo Fisher, H32713) for cell delineation. Fixed and stained plates were washed twice with PBS, then left in a final volume of 25 μL per well for imaging using a Thermo Fisher CX5 high content imaging microscope with LED excitation (386/23 nm, 560/25 nm) at 10x magnification. Cells were imaged at a single z-plane as optimized through image-based autofocus for each channel, and the exposure times were selected to maximize the signal-to-background ratio. Cellpose[99] was used to segment and count cell nuclei for each cell line, a percent viability score for compounds was generated by normalizing the cell count to the average well-level cell counts observed in the negative control (DMSO), and dose-response curves were fit using the GraphPad Prism software (v10.1.0, Dotmatics).

Phylogenetic tree construction

The tree was constructed in the web server Interactive Tree of Life[100] displaying angiosperm orders distinguished as ANA grade, Magnolids, Monocots and Eudicots[101].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Gene sequences of BURP-domain precursor peptides were deposited in GenBank (*GheBURP*-PP316100 [https://www.ncbi.nlm.nih.gov/nuccore/PP316100.1/], *SstBURP*-PP316102 [https://www.ncbi.nlm.nih.gov/nuccore/PP316102], *SaqBURP*-PQ435901 [https://www.ncbi.nlm.nih.gov/nuccore/PQ435901], *OmaBURP1*-PQ435900 [https://www.ncbi.nlm.nih.gov/nuccore/PQ435900], *OmaBURP2*-PQ463684 [https://www.ncbi.nlm.nih.gov/nuccore/PQ463684]). LC-MS/MS datasets (*Glechoma hederacea* leaf extract, *Sacciolepis striata* leaf extract, *Mercurialis annua* stem extract, *Stellaria aquatica* stem extract, *Orchidantha maxillarioides* stem and leaf extract) were deposited in GNPS-MassIVE[102] under accession MSV000096083 [https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=20959bf587854448850f0ae82d834343]. Proteomic datasets of burpitide cyclase characterization were deposited in GNPS-MassIVE[102] under accession MSV000097169 [https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=811455e7894e47659396e322f5a0e5f4]. Transcriptome assemblies were deposited in Deep Blue Data repository of the University of Michigan under accession 9306t0217 [https://doi.org/10.7302/z14x-5m94][103]. NMR data was deposited in NP-MRD[104] (glechomanin-NP0350903 [https://np-mrd.org/natural_products/NP0350903], mercurialin-NP0350904 [https://np-mrd.org/natural_products/NP0350904], moroidin-QLLVWRNH-NP0350905 [https://np-mrd.org/natural_products/NP0350905]). Source data are provided with this paper.

Code availability

All scripts from this study were deposited in GitHub repository [https://github.com/UM-KerstenLab4009/Moroidin-Transcriptomics/].

References

1. Pearce, G., Moura, D. S., Stratmann, J. & Ryan, C. A. Production of multiple plant hormones from a single polyprotein precursor. *Nature* **411**, 817–820 (2001).
2. Pearce, G., Moura, D. S., Stratmann, J. & Ryan, C. A. Jr RALF, a 5-kDa ubiquitous polypeptide in plants, arrests root growth and development. *Proc. Natl Acad. Sci. Usa.* **98**, 12843–12847 (2001).
3. Jennings, C., West, J., Waine, C., Craik, D. & Anderson, M. Biosynthesis and insecticidal properties of plant cyclotides: the cyclic knotted proteins from Oldenlandia affinis. *Proc. Natl Acad. Sci. Usa.* **98**, 10614–10619 (2001).
4. Gründemann, C., Stenberg, K. G. & Gruber, C. W. T20K: an immunomodulatory cyclotide on its way to the clinic. *Int. J. Pept. Res. Ther.* **25**, 9–13 (2019).
5. Oguis, G. K., Gilding, E. K., Jackson, M. A. & Craik, D. J. Butterfly pea (Clitoria ternatea), a cyclotide-bearing plant with applications in agriculture and medicine. *Front. Plant Sci.* **10**, 645 (2019).
6. Arnison, P. G. et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.* **30**, 108–160 (2013).
7. Montalbán-López, M. et al. New developments in RiPP discovery, enzymology and engineering. *Nat. Prod. Rep.* **38**, 130–239 (2021).
8. Chen, I.-M. A. et al. IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res.* **45**, D507–D516 (2017).
9. Merwin, N. J. et al. DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products. *Proc. Natl Acad. Sci. Usa.* **117**, 371–380 (2020).
10. Mohimani, H. et al. Automated genome mining of ribosomal peptide natural products. *ACS Chem. Biol.* **9**, 1545–1551 (2014).
11. Tietz, J. I. et al. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat. Chem. Biol.* **13**, 470–478 (2017).
12. Kloosterman, A. M., Shelton, K. E., van Wezel, G. P., Medema, M. H. & Mitchell, D. A. RRE-Finder: a genome-mining tool for class-independent RiPP discovery. *mSystems* **5**, e00267–20 (2020).
13. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
14. Jacobowitz, J. R. & Weng, J.-K. Exploring uncharted territories of plant specialized metabolism in the postgenomic era. *Annu. Rev. Plant Biol.* **71**, 631–658 (2020).
15. Kersten, R. D. et al. Gene-guided discovery and ribosomal biosynthesis of Moroidin Peptides. *J. Am. Chem. Soc.* **144**, 7686–7692 (2022).
16. Kriger, D., Pasquale, M. A., Ampolini, B. G. & Chekan, J. R. Mining raw plant transcriptomic data for new cyclopeptide alkaloids. *Beilstein J. Org. Chem.* **20**, 1548–1559 (2024).
17. Fisher, M. F. et al. The genetic origin of evolidine, the first cyclopeptide discovered in plants, and related orbitides. *J. Biol. Chem.* **295**, 14510–14521 (2020).
18. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
19. Kodama, Y., Shumway, M., Leinonen, R. & International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.* **40**, D54–D56 (2012).
20. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
21. Hu, T., Chitnis, N., Monos, D. & Dinh, A. Next-generation sequencing technologies: an overview. *Hum. Immunol.* **82**, 801–811 (2021).
22. Hölzer, M. & Marz, M. De novo transcriptome assembly: a comprehensive cross-species comparison of short-read RNA-Seq assemblers. *Gigascience* **8**, giz039 (2019).
23. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
24. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
25. Bushmanova, E., Antipov, D., Lapidus, A. & Prjibelski, A. D. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *Gigascience* **8**, giz100 (2019).
26. Xie, Y. et al. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**, 1660–1666 (2014).
27. Park, S. et al. Cyclotide evolution: insights from the analyses of their precursor sequences, structures and distribution in violets (viola). *Front. Plant Sci.* **8**, 2058 (2017).

28. Fisher, M. F. et al. A family of small, cyclic peptides buried in preproalbumin since the Eocene epoch. *Plant Direct* **2**, e00042 (2018).
29. Kersten, R. D. & Weng, J.-K. Gene-guided discovery and engineering of branched cyclic peptides in plants. *Proc. Natl Acad. Sci. Usa.* **115**, E10961–E10969 (2018).
30. Lima, S. T. et al. A widely distributed biosynthetic cassette is responsible for diverse plant side chain cross-linked cyclopeptides. *Angew. Chem. Int. Ed Engl.* **62**, e202218082 (2023).
31. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
32. Hempel, C. A. et al. Metagenomics versus total RNA sequencing: most accurate data-processing tools, microbial identification accuracy and perspectives for ecological assessments. *Nucleic Acids Res.* **50**, 9279–9293 (2022).
33. Chekan, J. R., Mydy, L. S., Pasquale, M. A. & Kersten, R. D. Plant peptides - redefining an area of ribosomally synthesized and post-translationally modified peptides. *Nat. Prod. Rep.* **41**, 1020–1059 (2024).
34. Kessler, S. C. & Chooi, Y.-H. Out for a RiPP: challenges and advances in genome mining of ribosomal peptides from fungi. *Nat. Prod. Rep.* **39**, 222–230 (2022).
35. Tan, N.-H. & Zhou, J. Plant cyclopeptides. *Chem. Rev.* **106**, 840–895 (2006).
36. Chigumba, D. N. et al. Discovery and biosynthesis of cyclic plant peptides via autocatalytic cyclases. *Nat. Chem. Biol.* **18**, 18–28 (2022).
37. Hattori, J., Boutilier, K. A., van Lookeren Campagne, M. M. & Miki, B. L. A conserved BURP domain defines a novel group of plant proteins with unusual primary structures. *Mol. Gen. Genet.* **259**, 424–428 (1998).
38. Treacy, B. K. et al. Bnm1, a Brassica pollen-specific gene. *Plant Mol. Biol.* **34**, 603–611 (1997).
39. Bassüner, R. et al. Abundant embryonic mRNA in field bean (Vicia faba L.) codes for a new class of seed proteins: cDNA cloning and characterization of the primary translation product. *Plant Mol. Biol.* **11**, 321–334 (1988).
40. Yamaguchi-Shinozaki, K. & Shinozaki, K. The plant hormone abscisic acid mediates the drought-induced expression but not the seed-specific expression of rd22, a gene responsive to dehydration stress in Arabidopsis thaliana. *Mol. Gen. Genet* **238**, 17–25 (1993).
41. Zheng, L., Heupel, R. C. & DellaPenna, D. The beta subunit of tomato fruit polygalacturonase isoenzyme 1: isolation, characterization, and identification of unique structural features. *Plant Cell* **4**, 1147–1156 (1992).
42. Christina Leung, T.-W., Williams, D. H., Barna, J. C. J., Foti, S. & Oelrichs, P. B. Structural studies on the peptide moroidin from laportea moroides. *Tetrahedron* **42**, 3333–3348 (1986).
43. Yoshikawa, K., Tao, S. & Arihara, S. Stephanotic Acid, a Novel Cyclic Pentapeptide from the Stem of Stephanotis floribunda. *J. Nat. Prod.* **63**, 540–542 (2000).
44. Morita, H., Shimbo, K., Shigemori, H. & Kobayashi, J. Antimitotic activity of moroidin, a bicyclic peptide from the seeds of Celosia argentea. *Bioorg. Med. Chem. Lett.* **10**, 469–471 (2000).
45. Kobayashi, J., Suzuki, H., Shimbo, K., Takeya, K. & Morita, H. Celogentins A-C, new antimitotic bicyclic peptides from the seeds of Celosia argentea. *J. Org. Chem.* **66**, 6626–6633 (2001).
46. Jayasena, A. S. et al. Stepwise evolution of a buried inhibitor peptide over 45 my. *Mol. Biol. Evol.* **34**, 1505–1516 (2017).
47. Song, Z., Burbridge, C., Schneider, D. J., Sharbel, T. F. & Reaney, M. J. T. The flax genome reveals orbitide diversity. *BMC Genomics* **23**, 534 (2022).
48. Gomes, P. W. P. et al. plantMASST-Community-driven chemotaxonomic digitization of plants. *bioRxivorg* https://doi.org/10.1101/2024.05.13.593988 (2024).
49. Shen, W., Sipos, B. & Zhao, L. SeqKit2: A Swiss army knife for sequence and alignment processing. *Imeta* **3**, e191 (2024).
50. Zhang, Z. et al. Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.* **26**, 3986–3990 (1998).
51. Shen, Y. et al. Potentides: New cysteine-rich peptides with unusual disulfide connectivity from *Potentilla anserina*. *Chembiochem* **20**, 1995–2004 (2019).
52. Villanueva, J. et al. Characterization of the wound-induced metallocarboxypeptidase inhibitor from potato1. *FEBS Lett.* **440**, 175–182 (1998).
53. Colilla, F. J., Rocher, A. & Mendez, E. gamma-Purothionins: amino acid sequence of two polypeptides of a new family of thionins from wheat endosperm. *FEBS Lett.* **270**, 191–194 (1990).
54. Broekaert, I., Lee, H. I., Kush, A., Chua, N. H. & Raikhel, N. Wound-induced accumulation of mRNA containing a hevein sequence in laticifers of rubber tree (Hevea brasiliensis). *Proc. Natl Acad. Sci. Usa.* **87**, 7633–7637 (1990).
55. Huang, R.-H. et al. Two novel antifungal peptides distinct with a five-disulfide motif from the bark of *Eucommia ulmoides* Oliv. *FEBS Lett.* **521**, 87–90 (2002).
56. Kumari, G. et al. Molecular diversity and function of jasmintides from Jasminum sambac. *BMC Plant Biol.* **18**, 144 (2018).
57. Kumari, G. et al. Cysteine-rich peptide family with unusual disulfide connectivity from *Jasminum sambac*. *J. Nat. Prod.* **78**, 2791–2799 (2015).
58. Tam, J. P. et al. Ginsentides: cysteine and glycine-rich peptides from the ginseng family with unusual disulfide connectivity. *Sci. Rep.* **8**, 16201 (2018).
59. Tailor, R. H. et al. A novel family of small cysteine-rich antimicrobial peptides from seed of impatiens balsaminals derived from a single precursor protein. *J. Biol. Chem.* **272**, 24480–24487 (1997).
60. Andreev, Y. A. et al. Genes encoding hevein-like defense peptides in wheat: distribution, evolution, and role in stress response. *Biochimie* **94**, 1009–1016 (2012).
61. Ghag, S. B., Shekhawat, U. K. S. & Ganapathi, T. R. Petunia floral defensins with unique prodomains as novel candidates for development of Fusarium wilt resistance in transgenic banana plants. *PLoS One* **7**, e39557 (2012).
62. Schrader-Fischer, G. & Apel, K. Organ-specific expression of highly divergent thionin variants that are distinct from the seed-specific crambin in the crucifer Crambe abyssinica. *Mol. Gen. Genet.* **245**, 380–389 (1994).
63. Mylne, J. S. et al. Cyclic peptides arising by evolutionary parallelism via asparaginyl-endopeptidase-mediated biosynthesis. *Plant Cell* **24**, 2765–2778 (2012).
64. Poth, A. G., Colgrave, M. L., Lyons, R. E., Daly, N. L. & Craik, D. J. Discovery of an unusual biosynthetic origin for circular proteins in legumes. *Proc. Natl Acad. Sci. Usa.* **108**, 10127–10132 (2011).
65. Condie, J. A. et al. The biosynthesis of Caryophyllaceae-like cyclic peptides in Saponaria vaccaria L. from DNA-encoded precursors. *Plant J.* **67**, 682–690 (2011).
66. Mylne, J. S. et al. Albumins and their processing machinery are hijacked for cyclic peptides in sunflower. *Nat. Chem. Biol.* **7**, 257–259 (2011).
67. Matsubayashi, Y. & Sakagami, Y. Phytosulfokine, sulfated peptides that induce the proliferation of single mesophyll cells of Asparagus officinalis L. *Proc. Natl Acad. Sci. Usa.* **93**, 7623–7627 (1996).
68. Fletcher, J. C., Brand, U., Running, M. P., Simon, R. & Meyerowitz, E. M. Signaling of cell fate decisions by CLAVATA3 in Arabidopsis shoot meristems. *Science* **283**, 1911–1914 (1999).

69. Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. *Nat. Rev. Genet.* **12**, 671–682 (2011).
70. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
71. 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).
72. Priyam, A. et al. Sequenceserver: a modern graphical user interface for custom BLAST databases. *Mol. Biol. Evol.* **36**, 2922–2924 (2019).
73. Sainsbury, F., Thuenemann, E. C. & Lomonossoff, G. P. pEAQ: versatile expression vectors for easy and quick transient expression of heterologous proteins in plants. *Plant Biotechnol. J.* **7**, 682–693 (2009).
74. Mydy, L. S. et al. An intramolecular macrocyclase in plant ribosomal peptide biosynthesis. *Nat. Chem. Biol.* **20**, 530–540 (2024).
75. Christenhusz, M. J. M. et al. The genome sequence of the annual mercury, Mercurialis annua L., 1753 (Euphorbiaceae). *Wellcome Open Res.* **9**, 102 (2024).
76. Negin, B. & Jander, G. Convergent and divergent evolution of plant chemical defenses. *Curr. Opin. Plant Biol.* **73**, 102368 (2023).
77. Schramma, K. R., Bushin, L. B. & Seyedsayamdost, M. R. Structure and biosynthesis of a macrocyclic peptide containing an unprecedented lysine-to-tryptophan crosslink. *Nat. Chem.* **7**, 431–437 (2015).
78. Nguyen, H. et al. Characterization of a radical SAM oxygenase for the ether crosslinking in darobactin biosynthesis. *J. Am. Chem. Soc.* **144**, 18876–18886 (2022).
79. Caruso, A., Martinie, R. J., Bushin, L. B. & Seyedsayamdost, M. R. Macrocyclization via an Arginine-Tyrosine Crosslink Broadens the Reaction Scope of Radical S-Adenosylmethionine Enzymes. *J. Am. Chem. Soc.* **141**, 16610–16614 (2019).
80. Nguyen, T. Q. N. et al. Post-translational formation of strained cyclophanes in bacteria. *Nat. Chem.* **12**, 1042–1053 (2020).
81. Ma, S. et al. Post-translational formation of aminomalonate by a promiscuous peptide-modifying radical SAM enzyme. *Angew. Chem. Int. Ed. Engl.* **60**, 19957–19964 (2021).
82. Forneris, C. C. & Seyedsayamdost, M. R. In vitro reconstitution of OxyC activity enables total chemoenzymatic syntheses of vancomycin aglycone variants. *Angew. Chem. Int. Ed. Engl.* **57**, 8048–8052 (2018).
83. Zhao, B. et al. Binding of two flaviolin substrate molecules, oxidative coupling, and crystal structure of Streptomyces coelicolor A3(2) cytochrome P450 158A2. *J. Biol. Chem.* **280**, 11599–11607 (2005).
84. Hug, J. J. et al. Biosynthesis of cittilins, unusual ribosomally synthesized and post-translationally modified peptides from *Myxococcus xanthus*. *ACS Chem. Biol.* **15**, 2221–2231 (2020).
85. Belin, P. et al. Identification and structural basis of the reaction catalyzed by CYP121, an essential cytochrome P450 in *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. Usa.* **106**, 7426–7431 (2009).
86. Zdouc, M. M. et al. A biaryl-linked tripeptide from Planomonospora reveals a widespread class of minimal RiPP gene clusters. *Cell Chem. Biol.* **28**, 733–739.e4 (2021).
87. Nam, H. et al. Exploring the diverse landscape of biaryl-containing peptides generated by cytochrome P450 macrocyclases. *J. Am. Chem. Soc.* **145**, 22047–22057 (2023).
88. Hu, Y. L. et al. P450-modified ribosomally synthesized peptides with aromatic cross-links. *J. Am. Chem. Soc.* **145**, 27325–27335 (2023).
89. He, B.-B. et al. Bacterial cytochrome P450 catalyzed post-translational macrocyclization of ribosomal peptides. *Angew. Chem. Int. Ed. Engl.* **62**, e202311533 (2023).
90. Nanudorn, P. et al. Atropopeptides are a novel family of ribosomally synthesized and posttranslationally modified peptides with a complex molecular shape. *Angew. Chem. Int. Ed Engl.* **61**, e202208361 (2022).
91. Fleischmann, A. et al. Evolution of genome size and chromosome number in the carnivorous plant genus Genlisea (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Ann. Bot.* **114**, 1651–1663 (2014).
92. Fernández, P. et al. A 160 Gbp fork fern genome shatters size record for eukaryotes. *iScience* **27**, 109889 (2024).
93. Lin, Z., Agarwal, V., Cong, Y., Pomponi, S. A. & Schmidt, E. W. Short macrocyclic peptides in sponge genomes. *Proc. Natl Acad. Sci. Usa.* **121**, e2314383121 (2024).
94. Singh, U. & Wurtele, E. S. orfipy: a fast and flexible tool for extracting ORFs. *Bioinformatics* **37**, 3019–3020 (2021).
95. Madeira, F. et al. The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024. *Nucleic Acids Res.* **52**, W521–W525 (2024).
96. Frisch, M. J. et al. Gaussian 16 Revision A.03. Gaussian Inc. Wallingford CT. https://gaussian.com/citation_a03/ (2016).
97. Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
98. Durkin, M. E., Qian, X., Popescu, N. C. & Lowy, D. R. Isolation of mouse embryo fibroblasts. *Bio Protoc.* **3**, e908 (2013).
99. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).
100. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
101. Cole, T. C. H., Hilger, H. H. & Stevens, P. Angiosperm phylogeny poster (APP) - flowering plant systematics, 2019. *PeerJ* https://doi.org/10.7287/peerj.preprints.2320v6 (2019).
102. Wang, M. et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
103. Kersten, R. D. et al. Large-scale transcriptome mining enables macrocyclic diversification and improved bioactivity of the stephanotic acid scaffold. *Large-scale transcriptome mining enables macrocyclic diversification and improved bioactivity of the stephanotic acid scaffold* https://doi.org/10.7302/z14x-5m94 (2025).
104. Wishart, D. S. et al. The Natural Products Magnetic Resonance Database (NP-MRD) for 2025. *Nucleic Acids Res.* **53**, D700–D708 (2025).

Acknowledgements

This study was supported by NIGMS (grant R35GM146934 to R.D.K., grant R01GM152417 to J.Z.S., grant F32GM146395 to. L.S.M., grant F31GM155959 to K.S., Pharmacological Sciences Training Program predoctoral fellowship T32GM140223 to D.O.), NIH OD (S10OD034245 to J.Z.S.), the Hermann Frasch Foundation (R.D.K.), a Rackham Merit Fellowship (D.O.), the PhRMA foundation (predoctoral fellowship, D.N.C.), a Rackham Predoctoral Fellowship (D.N.C.), and the National Science Foundation (Research Experience for Undergraduates, grant 1851985 to D.D.). We are grateful to Dr. Brad Ruhfehl for providing a sample of *Sacciolepis striata* and *Mercurialis annua* from the University of Michigan Herbarium. We thank Dr. Tim Cernak for collecting and providing *Glechoma hederacea*. We also thank Dr. George Lomonosoff (John Innes Center, UK) for sharing the pEAQ-HT vector. This research was supported in part through computational resources and services provided by Advanced Research Computing at the University of Michigan, Ann Arbor. We thank Meghan Forde and the Missouri Botanical Garden for access to *Orchidantha maxillarioides* plants.

Author contributions

R.D.K. conceived the idea. X.W., K.S., and D.A.O. contributed equally to this manuscript. D.A.O., K.S., L.S.M., D.N.C., and R.D.K. benchmarked de novo transcriptome assemblers and performed large-scale transcriptome assembly and analysis. X.W., K.S., K.M.M., D.D., and R.D.K. purified glechomanin. X.W. and R.D.K. characterized the glechomanin, mercurialin, and moroidin structures and peptide cyclases *in planta* and in vitro. K.S. and J.Z.S. performed and analyzed bioactivity assays. X.W., K.S., D.A.O., L.S.M., and R.D.K. wrote the manuscript. All authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-59428-4.

Correspondence and requests for materials should be addressed to Roland D. Kersten.

Peer review information *Nature Communications* thanks Christian Gruber, Chunshuai Huang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions 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.