Novel Cyclic Peptides in Seed of *Annona muricata* are Ribosomally Synthesized

Mark F. Fisher,[†] Jingjing Zhang,[†] Oliver Berkowitz,[‡] James Whelan,[‡] and Joshua S. Mylne[†]*

[†]School of Molecular Sciences, The University of Western Australia, 35 Stirling Highway, Crawley WA 6009, Australia

[‡]Department of Animal, Plant and Soil Sciences, School of Life Sciences & ARC Industrial Transformation Research Hub in Medicinal Agriculture, AgriBio building, La Trobe University, Bundoora 3086, VIC, Australia

ABSTRACT

Small, circular proteins are reported to have antimicrobial, cytotoxic and a host of other bioactivities. In bacteria and fungi they can be made by non-ribosomal peptide synthetases, but in plants they are exclusively ribosomal and the ligation reaction is performed by specialized endoproteases. Cyclic peptides from the *Annona* genus are said to possess cytotoxic and anti-inflammatory activities, but their biosynthesis is unknown. The medicinal soursop plant, *Annona muricata*, has been reported to contain annomuricatins A (cyclo-PGFVSA) and B (cyclo-PNAWLGT). Here, using *de novo* transcriptomics and tandem mass spectrometry, we identify a suite of short transcripts for precursor proteins for ten validated annomuricatins, nine of which are novel. In their precursors, annomuricatins are preceded by an absolutely conserved Glu and each peptide sequence has a conserved proto-terminal Pro, revealing parallels with the segetalin orbitides from the seeds of *Vaccaria hispanica,* which are processed through ligation by a prolyl oligopeptidase in a transpeptidation reaction.

Plants express a number of different types of cyclic peptides. There are several families of head-to-tail cyclic peptides in plants whose biosynthesis has been well-characterized. These are the cyclotides,[1, 2] cyclic knottins,[3] and PawS-derived peptides (PDPs),[4, 5] all of which are stabilized by disulfide bridges and are cyclized by a cysteine protease called asparaginyl endopeptidase (AEP).[4, 6-9] As such, these families of cyclic peptides are known as ribosomally-synthesized and post-translationally modified peptides (RiPPs).

There is another group of head-to-tail cyclic peptides in plants, known as orbitides. These similarly consist only of proteinogenic amino acid residues, but have no disulfide bonds and vary from five to 16 amino acid residues, although most are 7-9 residues.[10, 11] Orbitides are also considered to be RiPPs, but much less is known about their biosynthesis than the peptide families mentioned above. One family of orbitides, the PawL-derived peptides (PLPs), are closely related to the PDPs, and probably cyclized by the same mechanism,[12, 13] but most other orbitides do not contain Asp or Asn residues, which are required for cyclisation by AEP.[10] Thus most orbitides probably have a different mechanism of cyclisation to PLPs.

The genes encoding orbitides are known only in a small number of cases. Condie et al.[14] demonstrated that the segetalins, a group of orbitides found in the seeds of *Vaccaria hispanica* (*Saponaria vaccaria* L., Caryophyllaceae), are individually encoded by short genes that express a propeptide. The researchers also showed evidence of a genetic origin for orbitides in some *Citrus* species. Okinyo-Owiti et al.[15] characterized three novel orbitides from flax (*Linum usitatissimum* L., Linaceae), called cyclolinopeptides 17-19, and found that these cyclolinopeptides are derived from gene-encoded precursors, which for one gene had several copies of the same peptide. A search of expressed sequence tags from *Jatropha curcas* (Euphorbiaceae) suggested that curcacyclines A and B are genetically encoded.[10] In these species the cyclic peptide is excised from a highly conserved *N*-terminal leader sequence and *C*-terminal follower sequence (though the *Jatropha* peptides appear to have no follower sequence). The core peptides (the linear precursors of the final cyclic peptide) show little conservation, except at their *N*- and *C*-termini.

There are many other orbitides whose biosyntheses are not known. These include the ~35 orbitides found in plants of the Annonaceae family, including some said to possess cytotoxic[16, 17] and anti-inflammatory[18, 19] activities. Two of these orbitides are found in the seeds of *Annona muricata*. Annomuricatin A (cyclo-GPFVSA) was characterized by Li et al.[20] and annomuricatin B (cyclo-PNAWLGT) by the same group.[21] A third orbitide, annomuricatin C, was reported by Wélé et al.[22] but was later shown by structural studies to be identical to annomuricatin A.[23]

We were interested in the genetic origins of the annomuricatins, which we investigated by combining *de novo* transcriptomics with peptide tandem mass spectrometry. Having sequenced a novel orbitide, annomuricatin D, from tandem mass spectrometry data, we were able to identify a single transcript encoding both annomuricatin D and the previously-known annomuricatin A, thus demonstrating that annomuricatins are ribosomally synthesized. We were then able to identify transcripts encoding a further eight novel annomuricatins (E-L) and sequenced them by mass spectrometry.

RESULTS AND DISCUSSION

To characterize the biosynthesis of the annomuricatins, total RNA was extracted from *A. muricata* seeds, RNA-seq performed and a transcriptome assembled using established methods.[12] Using cyclic permutations of the two known peptides annomuricatin A and annomuricatin B, the transcriptome was searched for sequences encoding them; we were overwhelmed with hundreds of contigs that had the potential to encode annomuricatin A, but found none that could encode annomuricatin B.

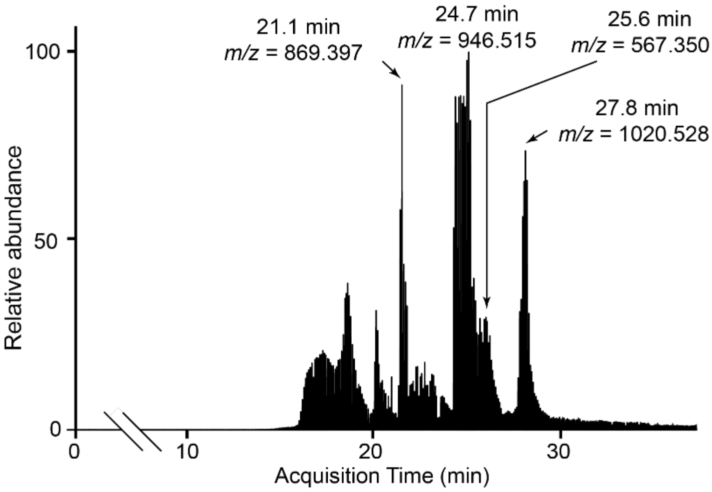


Figure 1. Total ion current chromatogram for extracts of *A. muricata* seeds with the *m/z* values and the acquisition times of the largest peaks labelled. All ions shown are $[M+H]^+$.

A liquid chromatography-mass spectrometry (LC-MS) analysis of seed peptide masses for *A. muricata* revealed a mass consistent with the presence of annomuricatin A, no mass for annomuricatin B, but critically it revealed a host of other seemingly abundant masses (Figure 1). One of these, eluting at 24.7 min, was especially abundant and the protonated molecule had a measured *m/z* of 946.515 $[M+H]^+$ (Figure 2A). Treatment of the sample with 1.2 M hydrochloric acid followed by liquid chromatography-tandem mass spectrometry (LC-MS/MS)[24] caused a new peak to appear, indicating a mass 18 Da heavier with its major components at *m/z* 482.767 $[M+2H]^{2+}$ and 964.524 $[M+H]^+$, although the original peak (now measured at *m/z* 946.514) remained (Figure 2B). If the 946.5 *m/z* ion was a novel annomuricatin, this is consistent with partial acid hydrolysis of its peptide backbone leading to an 18 Da mass increase.[24]

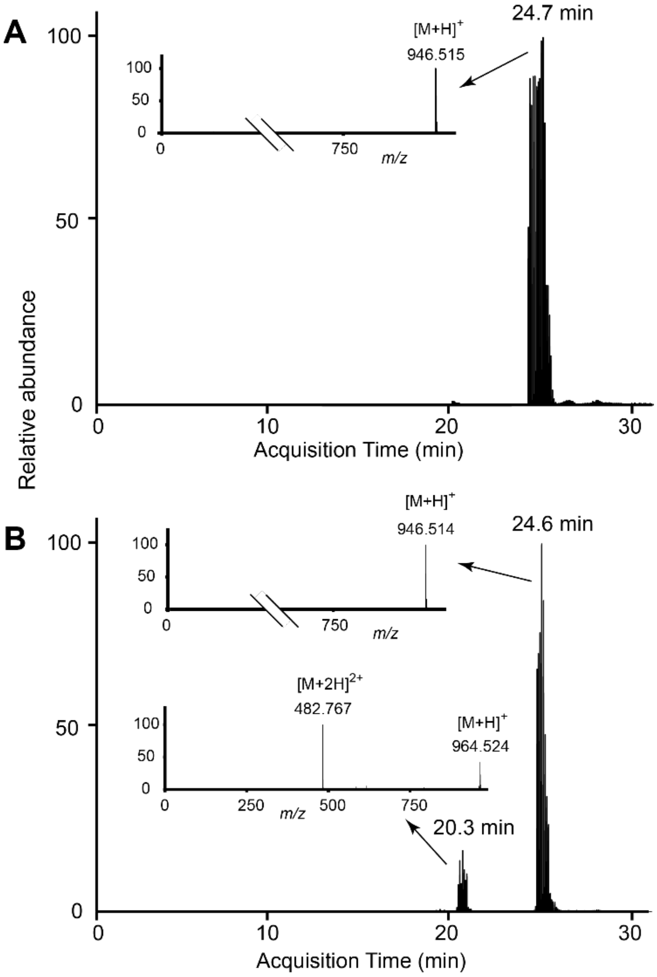


Figure 2. Combined extracted ion chromatograms at *m/z* 946.515 and *m/z* 964.526 for peptide extracts of *A. muricata* seeds. (A) Native extract; (B) After treatment with dilute hydrochloric acid caused the appearance of peaks at *m/z* 964.524 $[M+H]^+$ and 482.767 $[M+2H]^{2+}$ in the mass spectrum, consistent with backbone hydrolysis if the peak at *m/z* 946.5 $[M+H]^+$ was a cyclic peptide. Insets show the mass spectrum at each peak in the chromatogram.

Peptide evidence for annomuricatin D

Working on the hypothesis that the *m/z* 946.5 ion was a novel annomuricatin, we sequenced the protonated molecule of the hydrolyzed product (*m/z* 964.5) from MS/MS data to give the sequence SXFPPXPGH (where X represents either of the two isobaric residues Leu or Ile) (Figure 3), corresponding to an exact *m/z* of 964.526 for the acyclic peptide. The order of the Ser and Ile/Leu residues at the *N*-terminus could not be determined unambiguously by MS/MS, however. We tentatively named this novel cyclic peptide annomuricatin D.

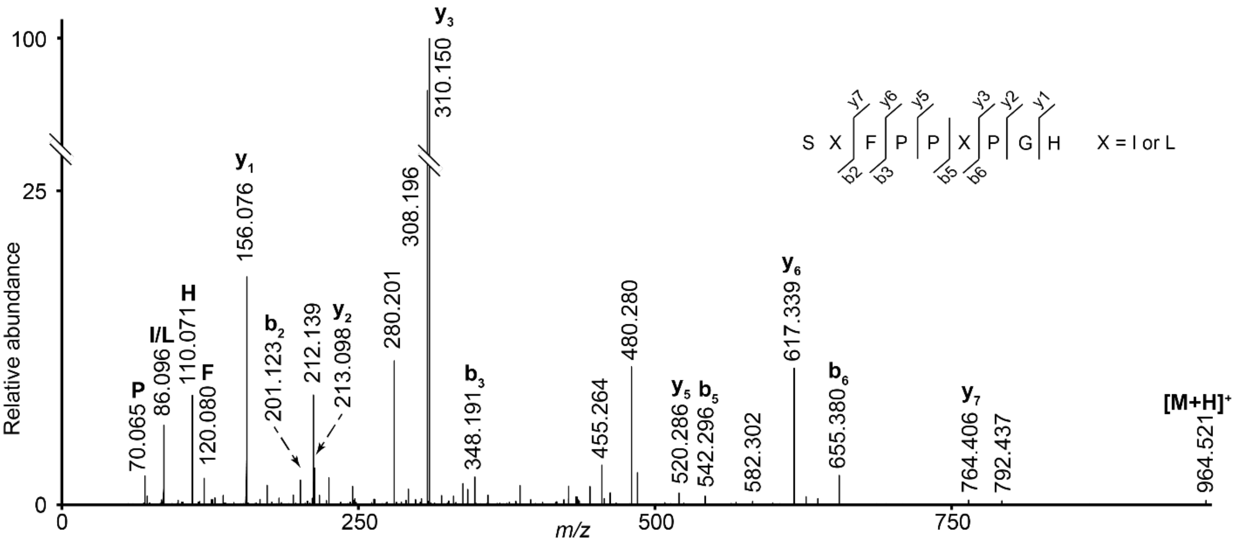


Figure 3. MS/MS confirmation of annomuricatin D. Partial hydrolysis using hydrochloric acid gave a sequence of b and y ions, with the ring breakage *C*-terminal to the His residue (inset shows the peptide sequence). Note that the order of the Ser and Ile/Leu residues could not be unambiguously determined.

Identification of novel transcripts encoding annomuricatins

Returning to search the *A. muricata* transcriptome using tBLASTn with all 72 possible annomuricatin D query sequences (i.e. all circular permutations of the two possible 9-residue sequences with four possible Ile/Leu combinations) produced just one contig with a perfect match, namely GHSIFPPIP. In the contig matching annomuricatin D, we observed that the complete ORF also encoded annomuricatin A (Figure 4).

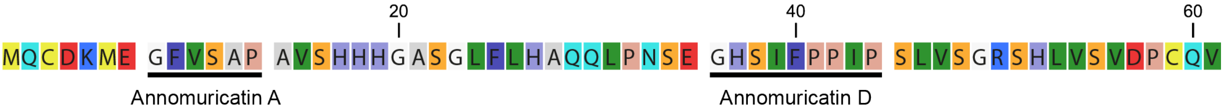

20
40
60
MQCDKME GFVSAP AVSHHHGASGLFLHAQQLPNSE GHSIFPPIP SLVSGRSHLVSVDPCQV
Annomuricatin A
Annomuricatin D

Figure 4. Searches of the *A. muricata* transcriptome by tBLASTn using annomuricatin D as the query sequence revealed a contig whose ORF encoded annomuricatin D as well as the previously published annomuricatin A.

Within this predicted precursor protein, the annomuricatin sequences both ended with Pro and were both preceded by Glu, suggesting their proteolytic maturation was conserved and involved two different enzymes – one targeting Glu that would release the *N*-terminus of the core peptide and a protease that targeted Pro that would probably perform a cleavage-coupled transpeptidation to the freed *N*-terminus to form a cyclic annomuricatin in a manner similar to other cyclic RiPPs. The sequence for annomuricatin A was *N*-terminal to annomuricatin D in the encoded sequence and so we named the transcript *<u>P</u>ro<u>a</u>nno<u>m</u>uricatinAD*, abbreviated to *PamAD*. Using the *PamAD* transcript as a tBLASTn query, five similar coding sequences with the potential to encode nine annomuricatins were discovered.

```
PamAD    MQCD----KME GFV---SAP AVSHHHGASGLFLHAQ-QLPNSE GHSIFPPIP SLVSGRSHLV--SVDPCQV*
         M--DSP--KSE GLVT---VP SVAYHHTASGLFLHTH-QLPNSE GLSA--VTP GHHSSPDLLI--SVHPCQA*
                 RSE-VPSLYLVMRGDSYSHH-SSGLFLHL--QLPNNE MIGD-YSYP AFVEDRRHLLSVSVDPYQV*
         MEME------E TLVPG-YSP SVSHHHGSSGLFLHAH-HLPNSE GFMHS-PVP -LVSCRGQLV--SVDPCQV*
         MEMDFPNARSE G-PPLYL-A MRVDSYSHDASLLFLPKQLPNIE RPFDF-LFP AFIRDRRHLLFVSVDPYQV*
                  ME-LVRVPDYSP-SASHHHGSSGFFLH--SQLPNNE GLGI---YP VFVGHRWHLL--SVYPYQV*
```

Figure 5. Manual alignment of the translated sequences of *Proannomuricatin* genes from the transcriptome of *A. muricata* seeds. The putative annomuricatin core peptide sequences are separated from the leader and follower sequences with spaces and have a green background. The absolutely conserved Glu in the P1 position is shown with a yellow background. Note the absolutely conserved Pro (pink background) which may be significant in processing of the peptide precursors, after cleavage at the *N*-terminus of each core peptide.

Consistent with the lack of peptide mass evidence, no evidence was found for a transcript matching the previously reported annomuricatin B sequence.[21] To confirm the sequences obtained by RNA-seq and assembly of contigs, primers were designed against the end of each contig to amplify a full length ORF and used genomic DNA as the template. We were able to amplify all six genes and found them to be intronless and a 100% match to the contigs assembled by RNA-seq.

Confirmation of annomuricatins E to L by tandem mass spectrometry

Knowing the sequence of potential peptides facilitates their confirmation by LC-MS/MS. Using the gene-encoded sequences, we predicted the expected mass for each peptide and could identify masses that matched predictions in all the putative *Pam* genes (Table 1). These putative masses named annomuricatin E to L were each sequenced by LC-MS/MS (Figures S1-S9). Annomuricatin I was found in two forms having either a reduced (Figure S5) or oxidized methionine (Figure S6). It is not possible to say whether this is a biologically relevant post-translational modification or occurred during peptide extraction and purification. We have therefore not classified the two forms as separate peptides.

Although peaks were found in the mass spectrum corresponding to a second putative annomuricatin in the *PamL* transcript, MS/MS data were not consistent with the predicted sequence from the transcript (Figure S10). We also searched for a possible second peptide in the *PamG* transcript, but none of the sequences searched was found in the mass spectrometry data.

Table 1. Summary of annomuricatin cyclic peptides, the genes that encode each and evidence for the peptides from mass spectrometry. N.D. = not detected.

Annomuricatin	Sequence	Gene	Peptide evidence
A (also C)	GFVSAP	*PamAD*	Li et al.[20]
B	NAWLGTP [a]	*N.D.*	N.D.
D	GHSIFPPIP	*PamAD*	Figure 2 Figure 3
E	GLVTVP	*PamEF*	Figure S1
F	GLSAVTP	*PamEF*	Figure S2
G	MIGDYSYP	*PamG*	Figure S3
H	TLVPGYSP	*PamHI*	Figure S4
I	GFMHSPVP [b]	*PamHI*	Figure S5 Figure S6[b]
J	GPPLYLA	*PamJK*	Figure S7
K	RPFDFLFP	*PamJK*	Figure S8
L	GLGIYP	*PamL*	Figure S9

[a] Annomuricatin B, which could not be detected at either the peptide or transcript level, was proposed by Li et al.[21] to have the sequence cyclo-PNAWLGT. Based on the conservation of Pro at the *C*-terminus of the core peptide we have written the sequence as NAWLGTP.

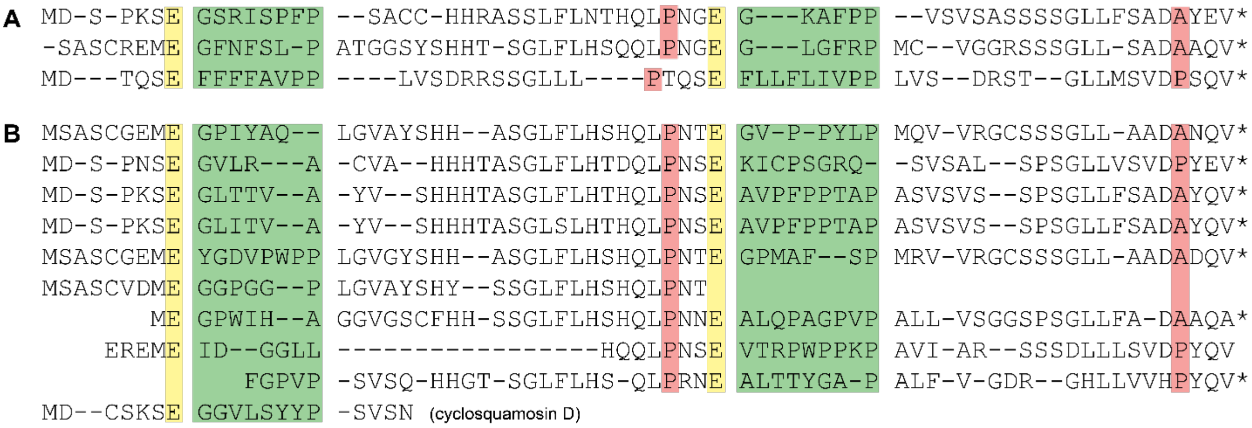
[b] Annomuricatin I (cyclo-GFMHSPVP) was found in two forms – one with standard Met (Figure S5) and the other with an oxidized-Met (Figure S6).

Like many other cyclic peptides, most of the core peptide sequence is not highly conserved, except for the *N*-terminal and *C*-terminal residues. In the case of the annomuricatins, the *N*-terminus is most often Gly, which tends to be the case in the majority of cyclic RiPPs, and the *C*-terminus residue is Pro or, in one case, Ala, both of which can be cleaved by a prolyl oligopeptidase.[25] In the leader sequence, Glu in the P1 position to the core peptide is absolutely conserved and the P2 residue varies, but is most often Ser. The only other very highly conserved residue is found at the fourth residue from the *C*-termini of the two propeptides formed after the *N*-terminus of each core peptide is cleaved. This residue is invariably Pro (Figure 5), and it is tempting to speculate that this is in some way required for the prolyl oligopeptidase to cyclize the core peptide.

The annomuricatins described here are typical in size for orbitides at between six and nine residues. Many annomuricatins contain mainly hydrophobic residues such as Val, Gly, Ala, Leu and Ile, which again is typical of orbitides. Annomuricatin K is unusual in that it has an Arg and an Asp residue. Both residues are rare in most orbitides except for the PLPs, which have an absolutely conserved Asp at the *C*-terminus of the core peptide, essential for their macrocyclization.

Other annomuricatin-like peptides

The number of similar annonomuricatins suggests the genes encoding them have duplicated and diverged. To investigate whether this type of gene is more widespread among the Annonaceae, RNA-seq data were downloaded from the NCBI Sequence Read Archive. These data were generated from RNA isolated from the mature flowers of *Annona squamosa*[26] and leaves of *A. muricata*.[27] We assembled transcriptomes and searched them with tBLASTn using the *pamAD* transcript as the query sequence. This approach identified several contigs with strong sequence similarity to *Pam* genes. Three candidate transcripts were identified in *A. muricata* leaves, which were different from those in the *A. muricata* seeds, encoding a total of six putative peptides (Figure 6A). Ten candidate transcripts from *A. squamosa* were also found most encoding two possible peptides, though several varied considerably from the canonical sequence, with the putative core peptide lacking a *C*-terminal Pro or Ala (Figure 6B).



```
A  MD-S-PKSE GSRISPFP --SACC-HHRASSLFLNTHQLPNGE G---KAFPP --VSVSASSSSGLLFSADAYEV*
   -SASCREME GFNFSL-P ATGGSYSHHT-SGLFLHSQQLPNGE G---LGFRP MC--VGGRSSSGLL-SADAAQV*
   MD---TQSE FFFFAVPP ----LVSDRRSSGLLL----PTQSE FLLFLIVPP LVS--DRST--GLLMSVDPSQV*

B  MSASCGEME GPIYAQ-- LGVAYSHH--ASGLFLHSHQLPNTE GV-P-PYLP MQV-VRGCSSSGLL-AADANQV*
   MD-S-PNSE GVLR---A -CVA--HHHTASGLFLHTDQLPNSE KICPSGRQ- -SVSAL--SPSGLLVSVDPYEV*
   MD-S-PKSE GLTTV--A -YV--SHHHTASGLFLHTHQLPNSE AVPFPPTAP ASVSVS--SPSGLLFSADAYQV*
   MD-S-PKSE GLITV--A -YV--SHHHTASGLSLHTHQLPNSE AVPFPPTAP ASVSVS--SPSGLLFSADAYQV*
   MSASCGEME YGDVPWPP LGVGYSHH--ASGLFLHSHQLPNTE GPMAF--SP MRV-VRGCSSSGLL-AADADQV*
   MSASCVDME GGPGG--P LGVAYSHY--SSGLFLHSHQLPNT
          ME GPWIH--A GGVGSCFHH-SSGLFLHSHQLPNNE ALQPAGPVP ALL-VSGGSPSGLLFA-DAAQA*
       EREME ID--GGLL -----------------HQQLPNSE VTRPWPPKP AVI-AR--SSSDLLLSVDPYQV
                FGPVP -SVSQ-HHGT-SGLFLHS-QLPRNE ALTTYGA-P ALF-V-GDR--GHLLVVHPYQV*
   MD--CSKSE GGVLSYYP -SVSN  (cyclosquamosin D)
```

Figure 6. Alignments of transcripts from (A) *A. muricata* leaves and (B) *A. squamosa* flowers. Putative core peptides are shown with a green background, the absolutely conserved Glu with a yellow background, and the highly conserved Pro or Ala in pink. Asterisks are stop codons, shown where the ORF is complete. The last sequence shown corresponds to the known cyclic peptide, cyclosquamosin D[28], and appears to be only a fragment of the full ORF.

The cases where the putative core peptides do not have Pro or Ala at the *C*-terminus may represent genes where one of the two peptides encoded has degenerated into a non-functional sequence, as appears to have occurred in the *PamG* and *PamL* transcripts from *A. muricata* seeds. Again, the Glu in the P1 position at the *N*-terminus of the core peptide is absolutely conserved. The highly conserved Pro in the follower sequence is again prominent, though sometimes replaced by Ala. In one case it was in the fifth position from the *C*-terminus rather than the fourth. Without a tissue sample on which to perform MS/MS analysis, it was impossible to say whether these differences from the *A. muricata* seed sequences prevent production of the cyclic peptide.

What can be seen in these homologs of the *Pam* genes of *A. muricata* seeds is that similar transcripts are present in more than one Annonaceae species. The putative propeptides encoded by the transcripts have a high degree of sequence similarity and most of them appear to encode two cyclic peptides. We therefore suggest this type of orbitide-encoding gene could occur across the Annonaceae and would be an interesting topic for further research. It is also noteworthy that the putative peptides from the leaves of *A. muricata* are quite different to the seed peptides, suggesting the annomuricatins could have organ-specific functions.

Biosynthesis of annomuricatins

The *Proannomuricatin* transcripts are short (~200 nt) and the leader and follower sequences around the core peptide are highly conserved (Figure 5). The core peptide sequence is highly variable, but the *N*-terminus has a highly conserved Gly, preceded by an absolutely conserved Glu in the P1 position. The *C*-terminal Pro of the core peptide is also absolutely conserved, except for the presence of Ala in one instance.

Much work has been done on the biosynthesis of plant cyclic peptides that rely on AEP for their maturation and cyclisation.[4-8, 29] A similar depth of understanding exists for the orbitide segetalin A and its relatives, whose cyclization of a conserved *C*-terminal Ala residue is performed by the prolyl oligopeptidase PCY1.[14, 30] Prior to cyclization, the segetalins are cleaved at the *N*-terminus of the core peptide by an as-yet-uncharacterized enzyme, OLP1,[30] which presumably is able to recognize the conserved Gln or Glu residue at the P1 position. Another example of peptides cyclized by a prolyl oligopeptidase are the amanitins from the fungus *Galerina marginata;* these are cyclized by GmPOPB. Unlike the segetalins, pro-amanitin is cleaved by POPB at the *N*-terminus of the core peptide as well as at the *C*-terminus due to the conserved Pro at the P1 position to the *N*-terminus.[31] Based on the conserved residues in *Pam* genes and parallels with other cyclic peptide biosyntheses, annomuricatin precursors are likely to be cleaved first at the core peptide *N*-terminus by a Glu-targeting protease and then cleaved at the *C*-terminus and ligated by a POP. This may parallel the action of OLP1 and PCY1 in *V. hispanica*, since the former appears to target Glu or Gln at the proto-*N*-terminus, and the latter cleaves at Ala or Pro.

Here we have shown that one known and nine novel annomuricatins in the seeds of *A. muricata* are encoded by six very similar short genes; four of them encode two cyclic peptides, and the other two encode one peptide each. Similar genes are also found in the leaves of *A. muricata* and flowers of *A.*

squamosa, indicating that such cyclic peptides may be present in other Annonaceae species. Comparison with the segetalins of *V. hispanica* indicates that a prolyl oligopeptidase is the likely cyclisation agent. Further study is required to identify this POP and to characterize its structure and mechanism of action.

EXPERIMENTAL SECTION

Plant material

Seeds of *A. muricata* were purchased from B & T World Seeds (Paguignan, France). While under quarantine, seeds were treated with Gaucho 600 insecticide (Bayer CropScience) to comply with Western Australian regulations. Soon after, seeds were frozen in liquid nitrogen, ground to a fine powder and stored at -80 °C until required.

Seed peptide extraction

Seed peptides were extracted as previously described.[12] Briefly, peptides were extracted in 50% MeOH / 50% $CHCl_2$. Phases were separated by the alternate addition of $CHCl_3$ and 0.05% trifluoroacetic acid. The upper, aqueous phase was dried overnight in a vacuum centrifuge (Labconco) prior to purification.

Purification of seed extracts

Seed peptide extracts were purified according to the method previously described.[13] Briefly, the crude extract was purified by solid-phase extraction using a 30 mg Strata-X polymeric reversed-phase column (Phenomenex). The extract was applied to the column as an aqueous solution of 5% MeCN (v/v) / 0.1% formic acid (v/v), then purified peptides were eluted with 85% MeCN (v/v) / 0.1% formic acid (v/v). The extract was dried in a vacuum centrifuge and redissolved in 5% MeCN (v/v) / 0.1% formic acid (v/v) for LC-MS analysis. HPLC-grade solvents were used throughout (Honeywell).

LC-MS/MS for peptide sequencing

Samples (2 µL) were injected onto an EASY-Spray PepMap C18 column (75 µm x 150 mm, 3 µm particle size, 10 nm pores; Thermo Fisher Scientific) using a Dionex UltiMate 3000 nano UHPLC system (Thermo Fisher Scientific) at flow rate of 200 nL/min by the “µL pick-up” method. A gradient elution was run from 5% solvent B to 95% solvent B over 40 minutes. Solvent A was 0.1% formic acid in water and solvent B was 0.1% formic acid in MeCN (Fisher Scientific). The resulting electrospray (source voltage 1,800 V) was analyzed by an Orbitrap Fusion mass spectrometer (Thermo Fisher Scientific) running in positive ionization, data-dependent, “top speed” MS/MS mode, employing the Orbitrap mass analyzer for both MS and MS/MS measurements at a resolution of 120,000 for MS and 60,000 for MS/MS. Parameters were set as follows: HCD fragmentation alternating between 14% ± 3% and 23% ± 3% energy, MS scan range from 400 to 1600 *m/z*, minimum MS/MS *m/z* 50, isolation window 1.2, ACG 400,000 (MS) and 500,000 (MS/MS), maximum injection time 200 ms (MS) and 250 ms (MS/MS) and 2 microscans for MS/MS. Only ions with an intensity > 100,000 were fragmented for MS/MS.

Peptide sequencing

Peptides were sequenced by visual examination of MS/MS spectra, aided by fragment predictions from the program mMass[32] for cyclic peptides and MS Product on the ProteinProspector website for acyclic peptides.[33] The parameters used for MS-Product selected a, b, y and immonium ions, plus internal fragments. Neutral losses were set to water (when S, T, E or D present) or ammonia (when R, K, Q or N present). Other parameters were left at their default values. Similar options were chosen for mMass except that y ions were not selected; these do not appear in the mass spectra of cyclic peptides because such peptides lack a carboxyl-terminus.

***Annona muricata* RNA-seq and transcriptome assembly**

Seeds of *A. muricata* (100 mg) were ground to a powder using a mortar and pestle cooled with liquid nitrogen. Total RNA was isolated using the Spectrum Plant Total RNA kit (Sigma Aldrich) and quality validated on a TapeStation 2200 system (Agilent). RNA-seq libraries were generated using the TruSeq Stranded Total RNA with Ribo-Zero Plant kit according to the manufacturer's instructions (Illumina) and sequenced on a NextSeq 550 system (Illumina) as paired-end reads with a length of150 bp and an average quality score (Q30) of above 90%. The raw reads were deposited in the NCBI Sequence Read Archive under accession number SRR8959862.

The *A. muricata* transcriptome was assembled using CLC Genomics Workbench 10.0.1 (QIAGEN Aarhus A/S). The raw reads were trimmed to a quality threshold of Q30 and minimum length 50, and the assembly was performed with word size 64 and minimum contig length 200. Other parameters remained at their default values.

Other *Annona* transcriptome assemblies

RNA-seq paired-read data were downloaded from the Sequence Read Archive of the National Institutes of Health for *A. squamosa* mature flowers (run SRR3478571) and *A. muricata* leaves (run ERR2040135). Using CLC Genomics Workbench 11.0, transcriptomes were assembled for each of these datasets. For *A. squamosa*, the raw paired-end reads were trimmed to a quality threshold of 22 and minimum length 50, and the assembly was performed with word size 64, minimum contig length 50 and bubble size 100. For *A. muricata* leaves, the raw paired-end reads were trimmed to a quality threshold of 20 and minimum length 50, and the assembly was performed with word size 50 and minimum contig length 50. In both cases, all other parameters remained at their default values.

Cloning of *Proannomuricatin* genes

Genomic DNA was extracted from 2 g of frozen *A. muricata* seed powder with the DNEasy Mericon Food Kit (QIAGEN) according to the manufacturer's instructions. DNA was quantified using a NanoDrop 2000 (Thermo Fisher Scientific).

The genomic DNA template was amplified by the polymerase chain reaction (PCR) using *Pfu* Ultra High-Fidelity DNA polymerase (Agilent Technologies). Each 50 µL reaction consisted of genomic DNA (~12 ng), 5 µL *Pfu* Ultra DNA polymerase reaction buffer (10x), 400 µM mixed dNTPs, 0.5 µL *Pfu* Ultra DNA polymerase and 0.4 µM of each of the appropriate forward and reverse primers (Table S1). PCR amplification was performed in a Veriti 96-well thermocycler (Applied Biosystems) programmed as follows: 95 °C for 2 min followed by 5 cycles of 95 °C for 30 s; 65 °C for 30 s; 72 °C for 30 s then 30 cycles of 95 °C for 30 s; 60 °C for 30 s; 72 °C for 30 s; and finally 72 °C for 10 min (*PamAD, PamEF, PamG*), or 95 °C for 2 min followed by 35 cycles of 95 °C for 30 s; 60 °C for 30 s; 72 °C for 30 s; and finally 72 °C for 10 min (*PamHI, PamJK, PamLM*).

PCR products were purified using the QIAquick PCR Purification Kit (QIAGEN) according to the manufacturer's instructions. DNA was eluted in 30 µL of water, quantified on a NanoDrop 2000 and sent for dideoxy sequencing using the forward PCR primers mentioned above (Garvan Institute, Darlinghurst NSW, Australia).

The six *Proannomuricatin* (*Pam*) genes from this study were deposited in GenBank under accession numbers MK836460-MK836465.

AUTHOR INFORMATION

Corresponding author

*Joshua S. Mylne, School of Molecular Sciences, The University of Western Australia, 35 Stirling Highway, Crawley, Perth 6009, Australia. E-mail: joshua.mylne@uwa.edu.au

Author Contributions

M.F.F. and J.S.M conceived the study; O.B. and J.W. performed RNA-seq; J.Z. and M.F.F. assembled the *A. muricata* transcriptome; M.F.F performed all other experiments and analyzed data; M.F.F and J.S.M. wrote the manuscript with help from all other authors.

Funding Sources

M.F.F. was supported by the Australian Research Training Program and a Bruce and Betty Green Postgraduate Research Scholarship. J.Z. was supported by an International Postgraduate Research Scholarship and a University Postgraduate Award from The University of Western Australia. J.S.M. was supported in part by an ARC Future Fellowship (FT120100013). This work was supported by ARC grant DP190102058 to J.S.M. and CE140100008 to J.W.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Nicolas L. Taylor of the School of Molecular Sciences at the University of Western Australia for providing advice and assistance in mass spectrometry aspects of this project.

The authors acknowledge the facilities, and the scientific and technical assistance of the Australian Microscopy & Microanalysis Research Facility at the Centre for Microscopy, Characterisation & Analysis, The University of Western Australia, a facility funded by the University, State and Commonwealth governments.

REFERENCES

(1) Saether, O.; Craik, D. J.; Campbell, I. D.; Sletten, K.; Juul, J.; Norman, D. G. *Biochemistry* **1995,** 34, 4147-4158.

(2) Craik, D. J.; Daly, N. L.; Bond, T.; Waine, C. *J. Mol. Biol.* **1999,** 294, 1327-1336.

(3) Hernandez, J. F.; Gagnon, J.; Chiche, L.; Nguyen, T. M.; Andrieu, J. P.; Heitz, A.; Trinh Hong, T.; Pham, T. T.; Le Nguyen, D. *Biochemistry* **2000,** 39, 5722-5730.

(4) Mylne, J. S.; Colgrave, M. L.; Daly, N. L.; Chanson, A. H.; Elliott, A. G.; McCallum, E. J.; Jones, A.; Craik, D. J. *Nat. Chem. Biol.* **2011,** 7, 257-259.

(5) Elliott, A. G.; Delay, C.; Liu, H.; Phua, Z.; Rosengren, K. J.; Benfield, A. H.; Panero, J. L.; Colgrave, M. L.; Jayasena, A. S.; Dunse, K. M.; Anderson, M. A.; Schilling, E. E.; Ortiz-Barrientos, D.; Craik, D. J.; Mylne, J. S. *Plant Cell* **2014,** 26, 981-995.

(6) Gillon, A. D.; Saska, I.; Jennings, C. V.; Guarino, R. F.; Craik, D. J.; Anderson, M. A. *Plant J.* **2008,** 53, 505-515.

(7) Mylne, J. S.; Chan, L. Y.; Chanson, A. H.; Daly, N. L.; Schaefer, H.; Bailey, T. L.; Nguyencong, P.; Cascales, L.; Craik, D. J. *Plant Cell* **2012,** 24, 2765-2778.

(8) Bernath-Levin, K.; Nelson, C.; Elliott, A. G.; Jayasena, A. S.; Millar, A. H.; Craik, D. J.; Mylne, J. S. *Chem. Biol.* **2015,** 22, 571-582.

(9) Saska, I.; Gillon, A. D.; Hatsugai, N.; Dietzgen, R. G.; Hara-Nishimura, I.; Anderson, M. A.; Craik, D. J. *J. Biol. Chem.* **2007,** 282, 29721-29728.

(10) Arnison, P. G.; Bibb, M. J.; Bierbaum, G.; Bowers, A. A.; Bugni, T. S.; Bulaj, G.; Camarero, J. A.; Campopiano, D. J.; Challis, G. L.; Clardy, J.; Cotter, P. D.; Craik, D. J.; Dawson, M.; Dittmann, E.; Donadio, S.; Dorrestein, P. C.; Entian, K.-D.; Fischbach, M. A.; Garavelli, J. S.; Goransson, U.; Gruber, C. W.; Haft, D. H.; Hemscheidt, T. K.; Hertweck, C.; Hill, C.; Horswill, A. R.; Jaspars, M.; Kelly, W. L.; Klinman, J. P.; Kuipers, O. P.; Link, A. J.; Liu, W.; Marahiel, M. A.; Mitchell, D. A.; Moll, G. N.; Moore, B. S.; Muller, R.; Nair, S. K.; Nes, I. F.; Norris, G. E.; Olivera, B. M.; Onaka, H.; Patchett, M. L.; Piel, J.; Reaney, M. J. T.; Rebuffat, S.; Ross, R. P.; Sahl, H.-G.; Schmidt, E. W.; Selsted, M. E.; Severinov, K.; Shen, B.; Sivonen, K.; Smith, L.; Stein, T.; Sussmuth, R. D.; Tagg, J. R.; Tang, G.-L.; Truman, A. W.; Vederas, J. C.; Walsh, C. T.; Walton, J. D.; Wenzel, S. C.; Willey, J. M.; van der Donk, W. A. *Nat. Prod. Rep.* **2013,** 30, 108-160.

(11) Fisher, M. F.; Payne, C. D.; Rosengren, K. J.; Mylne, J. S. *J. Nat. Prod.* **2019,** 82, 2152-2158.

(12) Jayasena, A. S.; Fisher, M. F.; Panero, J. L.; Secco, D.; Bernath-Levin, K.; Berkowitz, O.; Taylor, N. L.; Schilling, E. E.; Whelan, J.; Mylne, J. S. *Mol. Biol. Evol.* **2017,** 34, 1505-1516.

(13) Fisher, M. F.; Zhang, J.; Taylor, N. L.; Howard, M. J.; Berkowitz, O.; Debowski, A. W.; Behsaz, B.; Whelan, J.; Pevzner, P. A.; Mylne, J. S. *Plant Direct* **2018,** 2, e00042.

(14) Condie, J. A.; Nowak, G.; Reed, D. W.; John Balsevich, J.; Reaney, M. J. T.; Arnison, P. G.; Covello, P. S. *Plant J.* **2011,** 67, 682–690.

(15) Okinyo-Owiti, D. P.; Young, L.; Burnett, P.-G. G.; Reaney, M. J. T. *Peptide Science* **2014,** 102, 168-175.

(16) Wélé, A.; Landon, C.; Labbe, H.; Vovelle, F.; Zhang, Y.; Bodo, B. *Tetrahedron* **2004,** 60, 405-414.

(17) Wélé, A.; Zhang, Y.; Dubost, L.; Pousset, J.-L.; Bodo, B. *Chem. Pharm. Bull. (Tokyo)* **2006,** 54, 690-692.

(18) Chuang, P. H.; Hsieh, P. W.; Yang, Y. L.; Hua, K. F.; Chang, F. R.; Shiea, J.; Wu, S. H.; Wu, Y. C. *J. Nat. Prod.* **2008,** 71, 1365-70.

(19) Yang, Y. L.; Hua, K. F.; Chuang, P. H.; Wu, S. H.; Wu, K. Y.; Chang, F. R.; Wu, Y. C. *J. Agric. Food Chem.* **2008,** 56, 386-92.

(20) Li, C.; Tan, N.; Lu, Y.; Liang, H.; Mu, Q.; Zheng, H.; Hao, X.; Zhou, J. *Acta Botanica Yunnanica* **1995,** 17, 459-462.

(21) Li, C.; Tan, N.; Zheng, H.; Mu, Q.; Hao, X.; He, Y.; Zou, J. *Phytochemistry* **1998,** 48, 555-556.

(22) Wélé, A.; Zhang, Y.; Caux, C.; Brouard, J.-P.; Pousset, J.-L.; Bodo, B. *C. R. Chim.* **2004,** 7, 981-988.

(23) Wu, L.; Lu, Y.; Zheng, Q.-T.; Tan, N.-H.; Li, C.-M.; Zhou, J. *J. Mol. Struct.* **2007,** 827, 145-148.

(24) Fisher, M. F.; Mylne, J. S. *J. Proteome Res.* **2019,** 18, 4065-4071.

(25) Moriyama, A.; Nakanishi, M.; Sasaki, M. *J. Biochem.* **1988,** 104, 112-117.

(26) Liu, K.; Feng, S.; Pan, Y.; Zhong, J.; Chen, Y.; Yuan, C.; Li, H. *Front. Plant Sci.* **2016,** 7, 1695.

(27) Leebens-Mack, J. H.; Barker, M. S.; Carpenter, E. J.; Deyholos, M. K.; Gitzendanner, M. A.; Graham, S. W.; Grosse, I.; Li, Z.; Melkonian, M.; Mirarab, S.; Porsch, M.; Quint, M.; Rensing, S. A.; Soltis, D. E.; Soltis, P. S.; Stevenson, D. W.; Ullrich, K. K.; Wickett, N. J.; DeGironimo, L.; Edger, P. P.; Jordon-Thaden, I. E.; Joya, S.; Liu, T.; Melkonian, B.; Miles, N. W.; Pokorny, L.; Quigley, C.; Thomas, P.; Villarreal, J. C.; Augustin, M. M.; Barrett, M. D.; Baucom, R. S.; Beerling, D. J.; Benstein, R. M.; Biffin, E.; Brockington, S. F.; Burge, D. O.; Burris, J. N.; Burris, K. P.; Burtet-Sarramegna, V.; Caicedo, A. L.; Cannon, S. B.; Çebi, Z.; Chang, Y.; Chater, C.; Cheeseman, J. M.; Chen, T.; Clarke, N. D.; Clayton, H.; Covshoff, S.; Crandall-Stotler, B. J.; Cross, H.; dePamphilis, C. W.; Der, J. P.; Determann, R.; Dickson, R. C.; Di Stilio, V. S.; Ellis, S.; Fast, E.; Feja, N.; Field, K. J.; Filatov, D. A.; Finnegan, P. M.; Floyd, S. K.; Fogliani, B.; García, N.; Gâteblé, G.; Godden, G. T.; Goh, F.; Greiner, S.; Harkess, A.; Heaney, J. M.; Helliwell, K. E.; Heyduk, K.; Hibberd, J. M.; Hodel, R. G. J.; Hollingsworth, P. M.; Johnson, M. T. J.; Jost, R.; Joyce, B.; Kapralov, M. V.; Kazamia, E.; Kellogg, E. A.; Koch, M. A.; Von Konrat, M.; Könyves, K.; Kutchan, T. M.; Lam, V.; Larsson, A.; Leitch, A. R.; Lentz, R.; Li, F.-W.; Lowe, A. J.; Ludwig, M.; Manos, P. S.; Mavrodiev, E.; McCormick, M. K.; McKain, M.; McLellan, T.; McNeal, J. R.; Miller, R. E.; Nelson, M. N.; Peng, Y.; Ralph, P.; Real, D.; Riggins, C. W.; Ruhsam, M.; Sage, R. F.; Sakai, A. K.; Scascitella, M.; Schilling, E. E.; Schlösser, E.-M.; Sederoff, H.; Servick, S.; Sessa, E. B.; Shaw, A. J.; Shaw, S. W.; Sigel, E. M.; Skema, C.; Smith, A. G.; Smithson, A.; Stewart, C. N.; Stinchcombe, J. R.; Szövényi, P.; Tate, J. A.; Tiebel, H.; Trapnell, D.; Villegente, M.; Wang, C.-N.; Weller, S. G.; Wenzel, M.; Weststrand, S.; Westwood, J. H.; Whigham, D. F.; Wu, S.; Wulff, A. S.; Yang, Y.; Zhu, D.; Zhuang, C.; Zuidof, J.; Chase, M. W.; Pires, J. C.; Rothfels, C. J.; Yu, J.; Chen, C.; Chen, L.; Cheng, S.; Li, J.; Li, R.; Li, X.; Lu, H.; Ou, Y.; Sun, X.; Tan, X.; Tang, J.; Tian, Z.; Wang, F.; Wang, J.; Wei, X.; Xu, X.; Yan, Z.; Yang, F.; Zhong, X.; Zhou, F.; Zhu, Y.; Zhang, Y.; Ayyampalayam, S.; Barkman, T. J.; Nguyen, N.-p.; Matasci, N.; Nelson, D. R.; Sayyari, E.; Wafula, E. K.; Walls, R. L.; Warnow, T.; An, H.; Arrigo, N.; Baniaga, A. E.; Galuska, S.; Jorgensen, S. A.; Kidder, T. I.; Kong, H.; Lu-Irving, P.; Marx, H. E.; Qi, X.; Reardon, C. R.; Sutherland, B. L.; Tiley, G. P.; Welles, S. R.; Yu, R.; Zhan, S.; Gramzow, L.; Theißen, G.; Wong, G. K.-S.; One Thousand Plant Transcriptomes, I. *Nature* **2019,** 574, 679-685.

(28) Morita, H.; Sato, Y.; Kobayashi, J. i. *Tetrahedron* **1999,** 55, 7509-7518.

(29) Saska, I.; Craik, D. J. *Trends Biochem. Sci.* **2008,** 33, 363-368.

(30) Barber, C. J. S.; Pujara, P. T.; Reed, D. W.; Chiwocha, S.; Zhang, H.; Covello, P. S. *J. Biol. Chem.* **2013,** 288, 12500-12510.

(31) Luo, H.; Hong, S.-Y.; Sgambelluri, R. M.; Angelos, E.; Li, X.; Walton, Jonathan D. *Chem. Biol.* **2014,** 21, 1610-1617.

(32) Niedermeyer, T. H. J.; Strohalm, M. *PLoS One* **2012,** 7, e44913.

(33) Baker, P. R.; Clauser, K. R. ProteinProspector. http://prospector.ucsf.edu/ (5th December 2019),

SUPPORTING INFORMATION

Supplementary Tables

Table S1. PCR primers used for the amplification and sequencing of the putative *Proannomuricatin* genes identified in transcriptome data.

Gene	Forward Primer	Reverse Primer
ProannomuricatinAD	5′-TGGGAGACAG GAGATACAGA GAGAGA-3′	5′-ATCGGTAGAA AGAGAGAGAG ACAAGACAG-3′
ProannomuricatinEF	5′-AGTCGCAGCT CGCAGGTACA-3′	5′-AGCAGCAAAG AAACACAGCA GAAA-3′
ProannomuricatinG	5′-AGGAGCGAAG TGCCTTCGTT GTA-3′	5′-TGCACACACA TGCGGCAAGG-3′
ProannomuricatinHI	5′-AGCTAAGCAA TCGCAGTTGG CA-3′	5′-GGAAGAGAAG ACAGGACGGC CA-3′
ProannomuricatinJK	5′-TGCCCAATAT CGTGAGGGAC GC-3′	5′-AGATGGCCAC CTCGCTCCAT-3′
ProannomuricatinL	5′-TGGACAGACA TGGAGTTGGT GC-3′	5′-TGACAAGGAC TTCCATCGAG GC-3′

Supplementary Figures

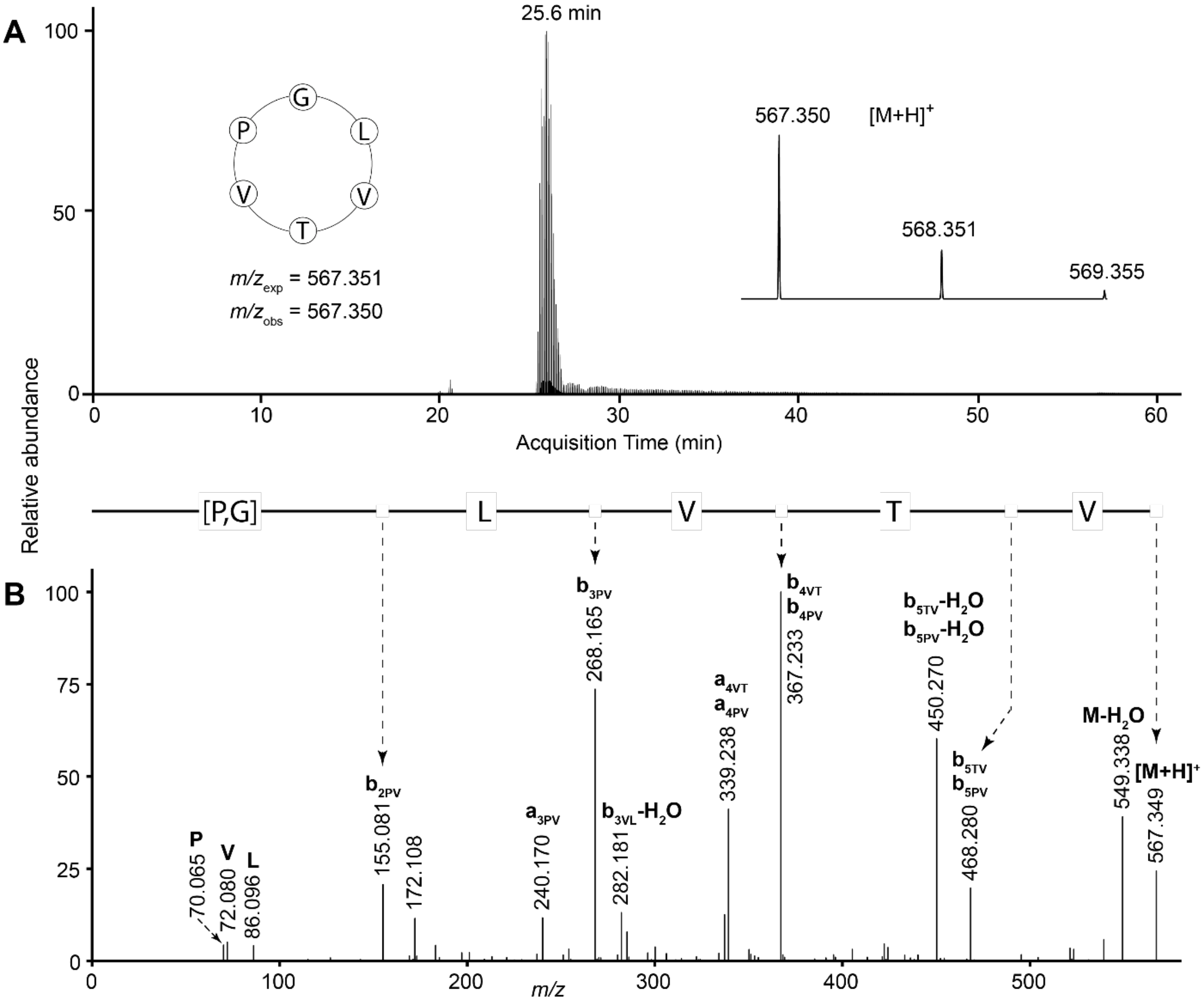


Figure S1. LC-MS data for annomuricatin E with sequence cyclo-GLVTVP. (A) Extracted ion chromatogram showing acquisition time of the peptide, with (inset left) peptide sequence with expected and observed mass-to-charge ratios (*m/z*) and (inset right) peptide mass spectrum. (B) Tandem mass spectrum of the fragmented precursor ion. Immonium ions are denoted by the one-letter code of the residue they represent. The derived amino acid sequence is shown above the spectrum; residues in square brackets could not be assigned an order.

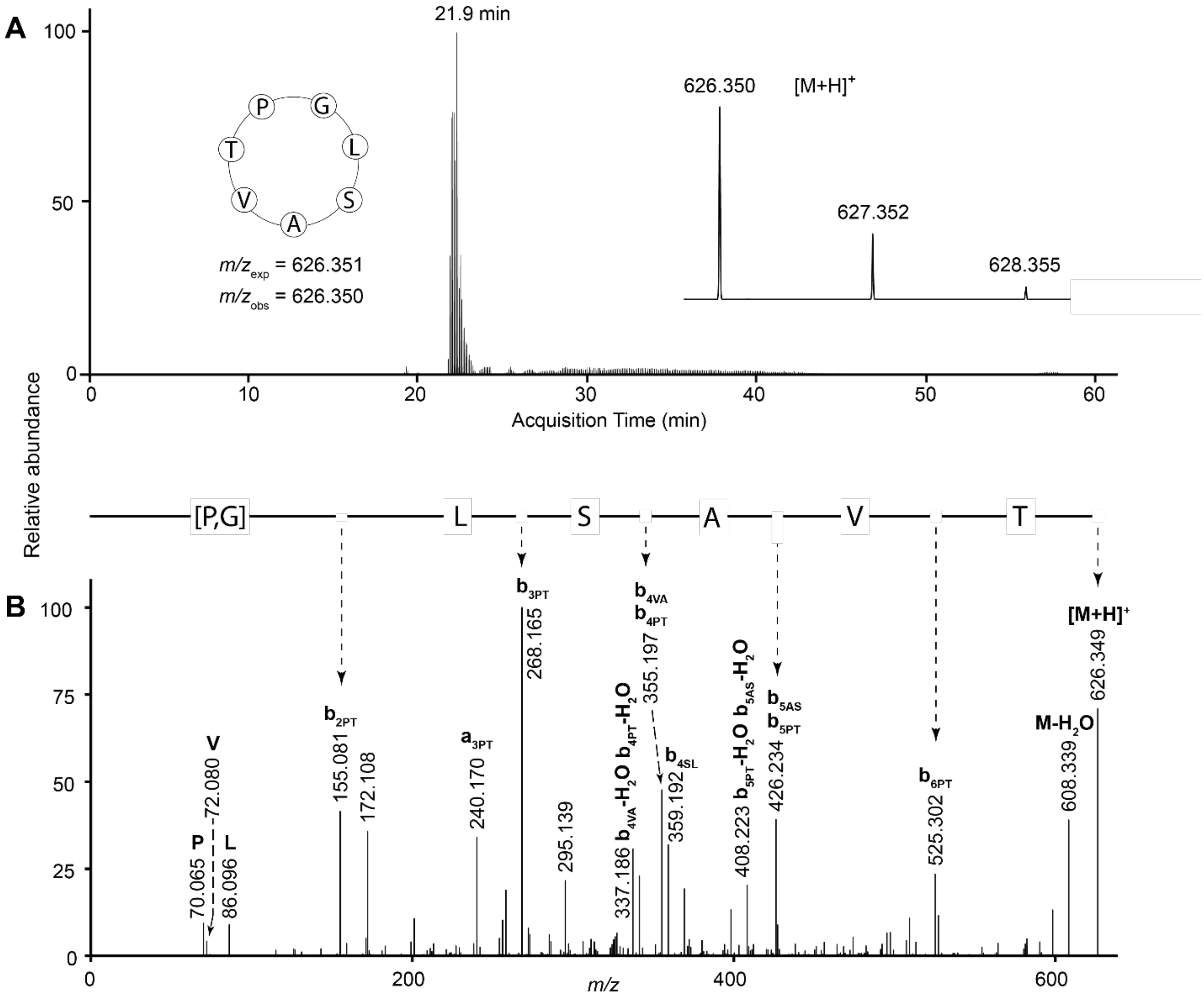


Figure S2. LC-MS data for annomuricatin F with sequence cyclo-GLSAVTP. (A) Extracted ion chromatogram showing acquisition time of the peptide, with (inset left) peptide sequence with expected and observed mass-to-charge ratios (*m/z*) and (inset right) peptide mass spectrum. (B) Tandem mass spectrum of the fragmented precursor ion. Immonium ions are denoted by the one-letter code of the residue they represent. The derived amino acid sequence is shown above the spectrum; residues in square brackets could not be assigned an order.

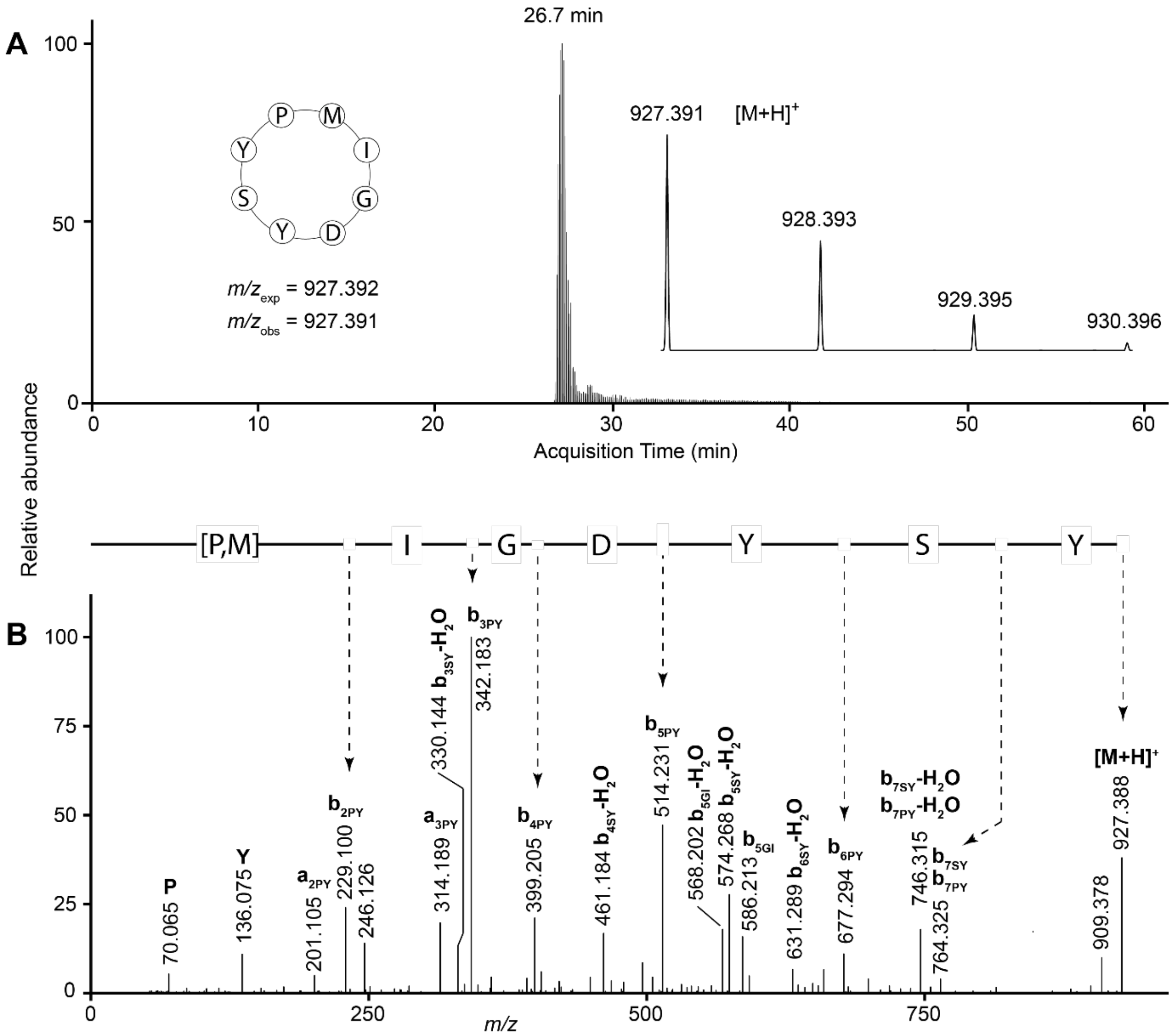


Figure S3. LC-MS data for annomuricatin G with sequence cyclo-MIGDYSYP. (A) Extracted ion chromatogram showing acquisition time of the peptide, with (inset left) peptide sequence with expected and observed mass-to-charge ratios (*m/z*) and (inset right) peptide mass spectrum. (B) Tandem mass spectrum of the fragmented precursor ion. Immonium ions are denoted by the one-letter code of the residue they represent. The derived amino acid sequence is shown above the spectrum; residues in square brackets could not be assigned an order.

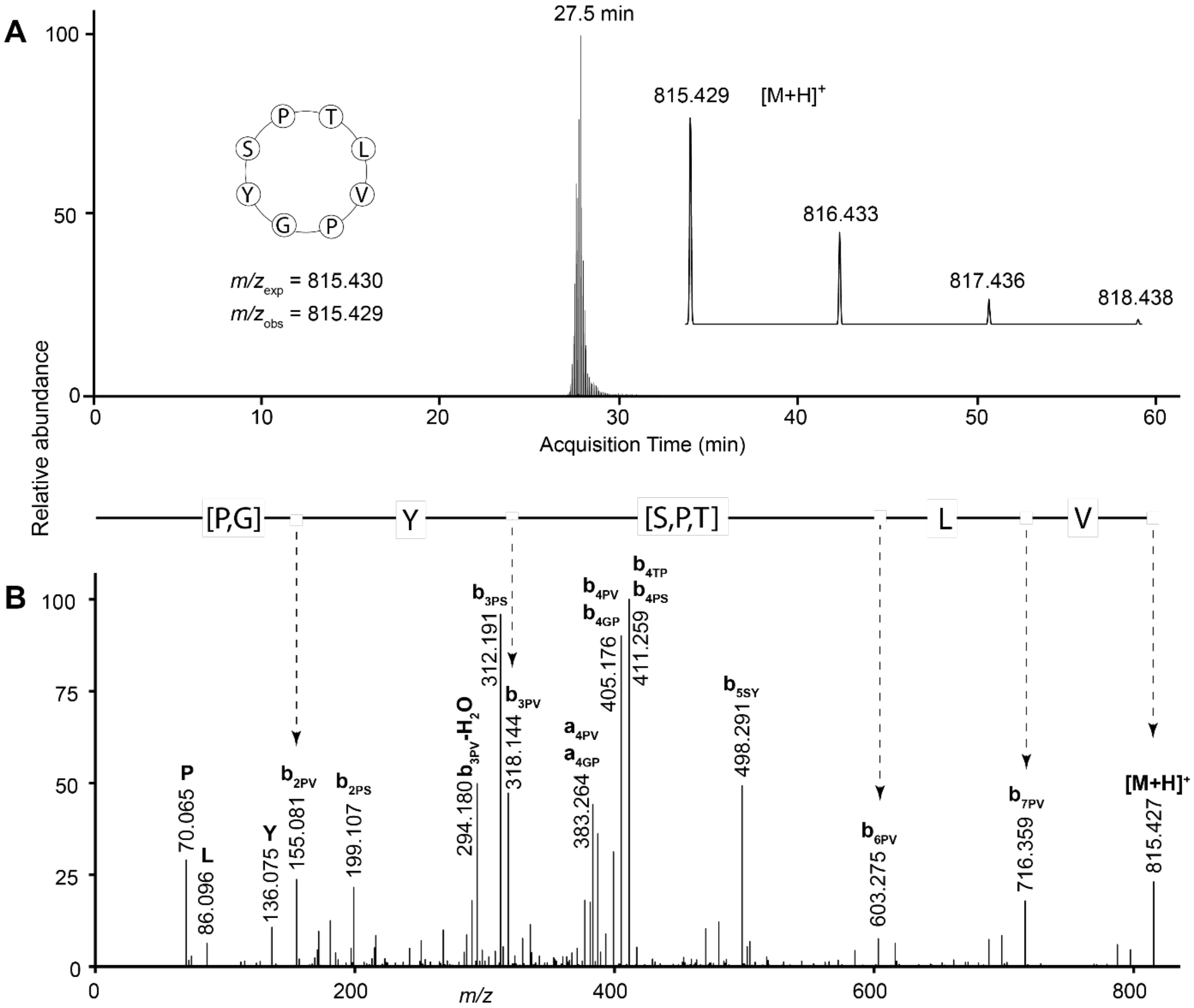


Figure S4. LC-MS data for annomuricatin H with sequence cyclo-TLVPGYSP. (A) Extracted ion chromatogram showing acquisition time of the peptide, with (inset left) peptide sequence with expected and observed mass-to-charge ratios (*m/z*) and (inset right) peptide mass spectrum. (B) Tandem mass spectrum of the fragmented precursor ion. Immonium ions are denoted by the one-letter code of the residue they represent. The derived amino acid sequence is shown above the spectrum; residues in square brackets could not be assigned an order.

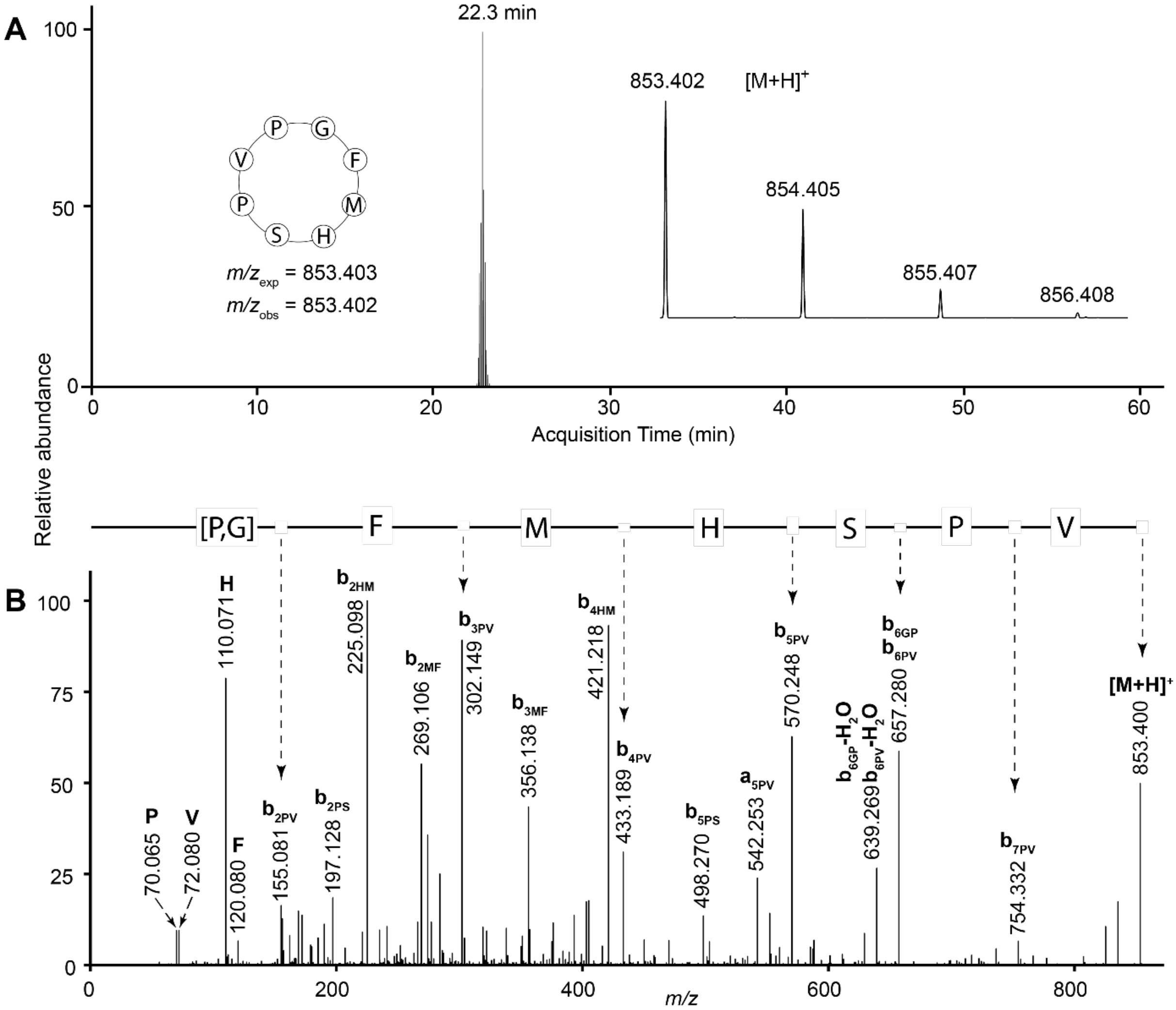


Figure S5. LC-MS data for annomuricatin I with sequence cyclo-GFMHSPVP. (A) Extracted ion chromatogram showing acquisition time of the peptide, with (inset left) peptide sequence with expected and observed mass-to-charge ratios (*m/z*) and (inset right) peptide mass spectrum. (B) Tandem mass spectrum of the fragmented precursor ion. Immonium ions are denoted by the one-letter code of the residue they represent. The derived amino acid sequence is shown above the spectrum; residues in square brackets could not be assigned an order.

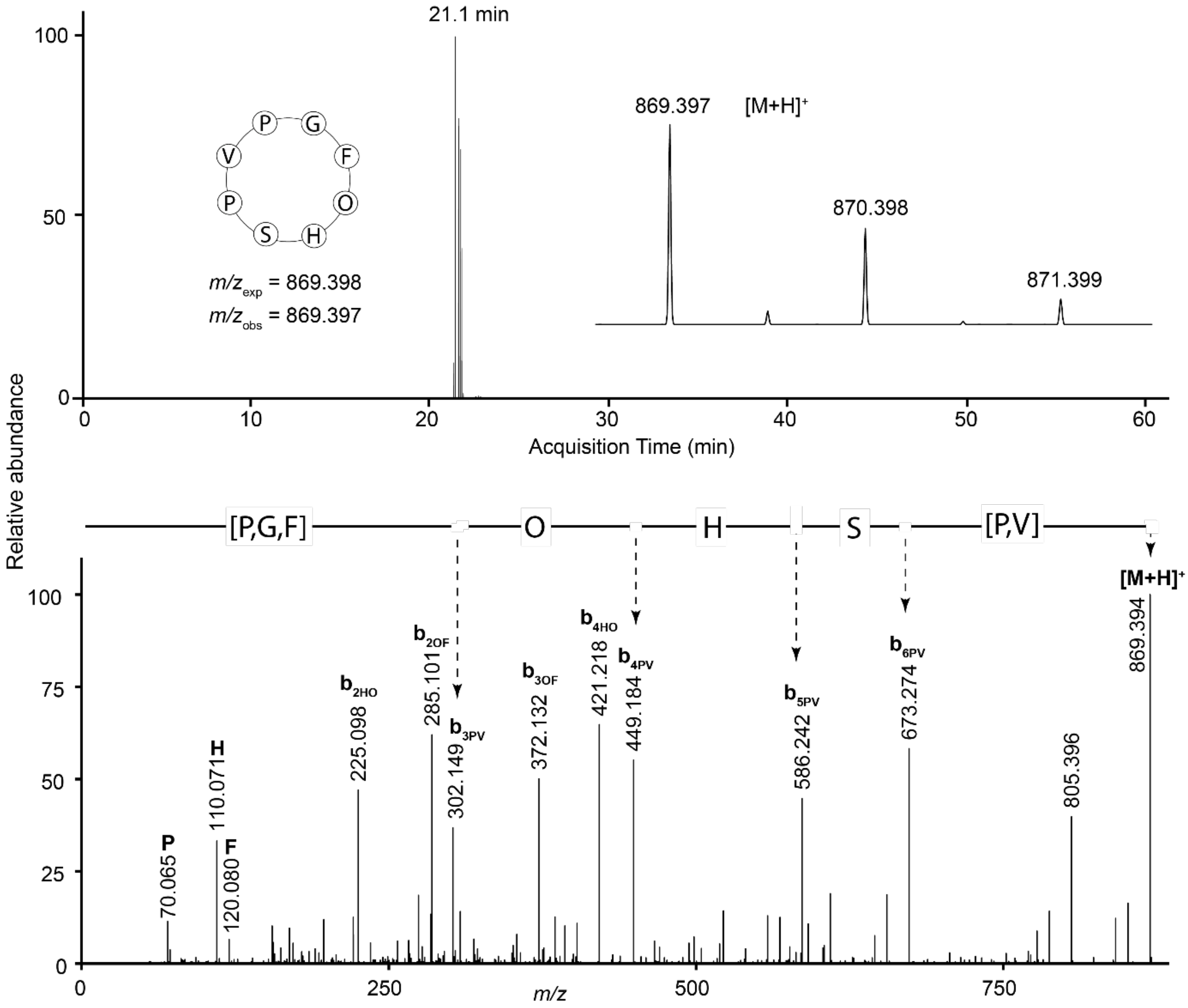


Figure S6. LC-MS data for annomuricatin I with sequence cyclo-GFMHSPVP, although the Met was oxidized (hence O in figure).(A) Extracted ion chromatogram showing acquisition time of the peptide, with (inset left) peptide sequence with expected and observed mass-to-charge ratios (*m/z*) and (inset right) peptide mass spectrum.(B) Tandem mass spectrum of the fragmented precursor ion. Immonium ions are denoted by the one-letter code of the residue they represent. The derived amino acid sequence is shown above the spectrum; residues in square brackets could not be assigned an order.

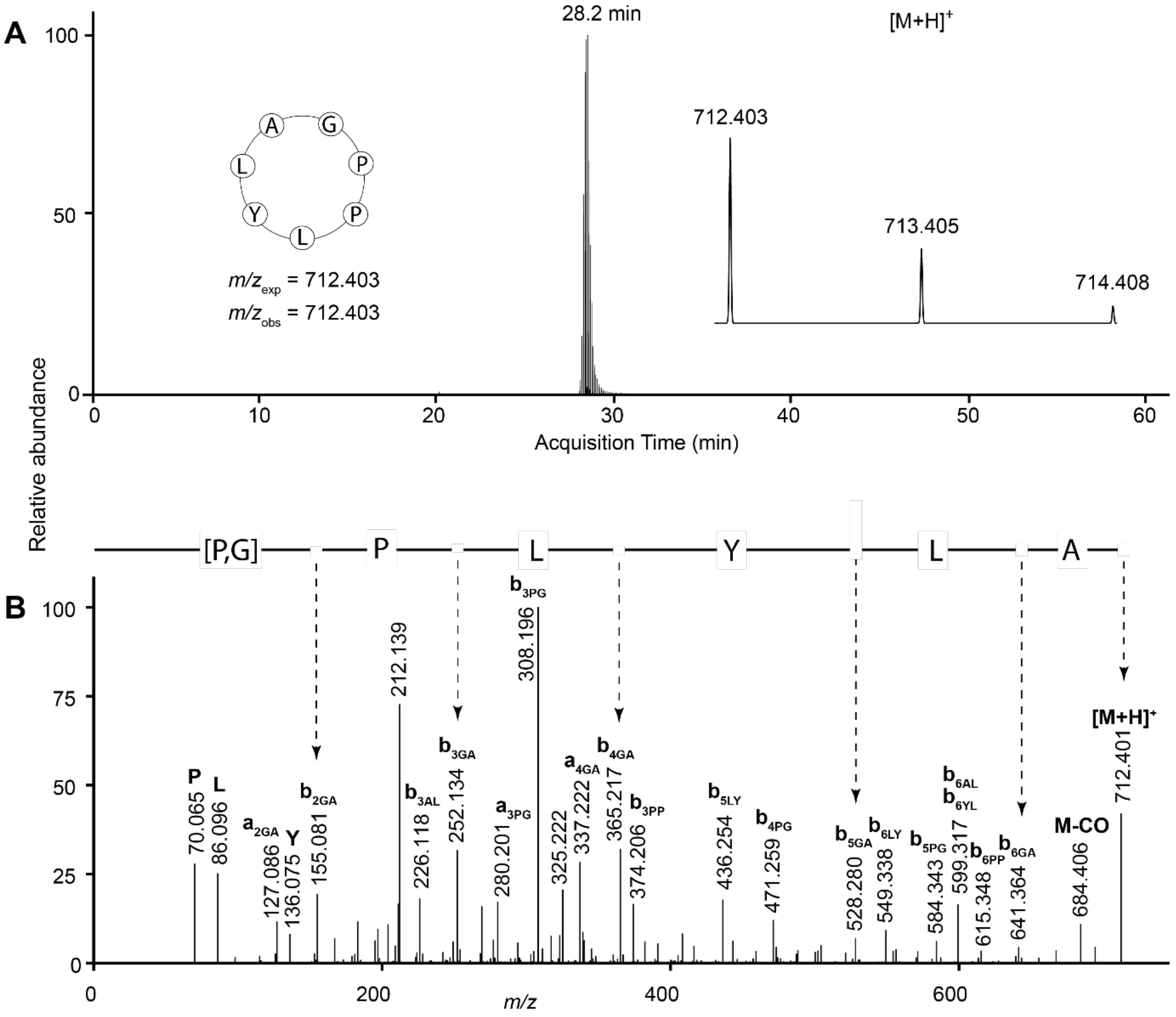


Figure S7. LC-MS data for annomuricatin J with sequence cyclo-GPPLYLA. (A) Extracted ion chromatogram showing acquisition time of the peptide, with (inset left) peptide sequence with expected and observed mass-to-charge ratios (*m/z*) and (inset right) peptide mass spectrum. (B) Tandem mass spectrum of the fragmented precursor ion. Immonium ions are denoted by the one-letter code of the residue they represent. The derived amino acid sequence is shown above the spectrum; residues in square brackets could not be assigned an order.

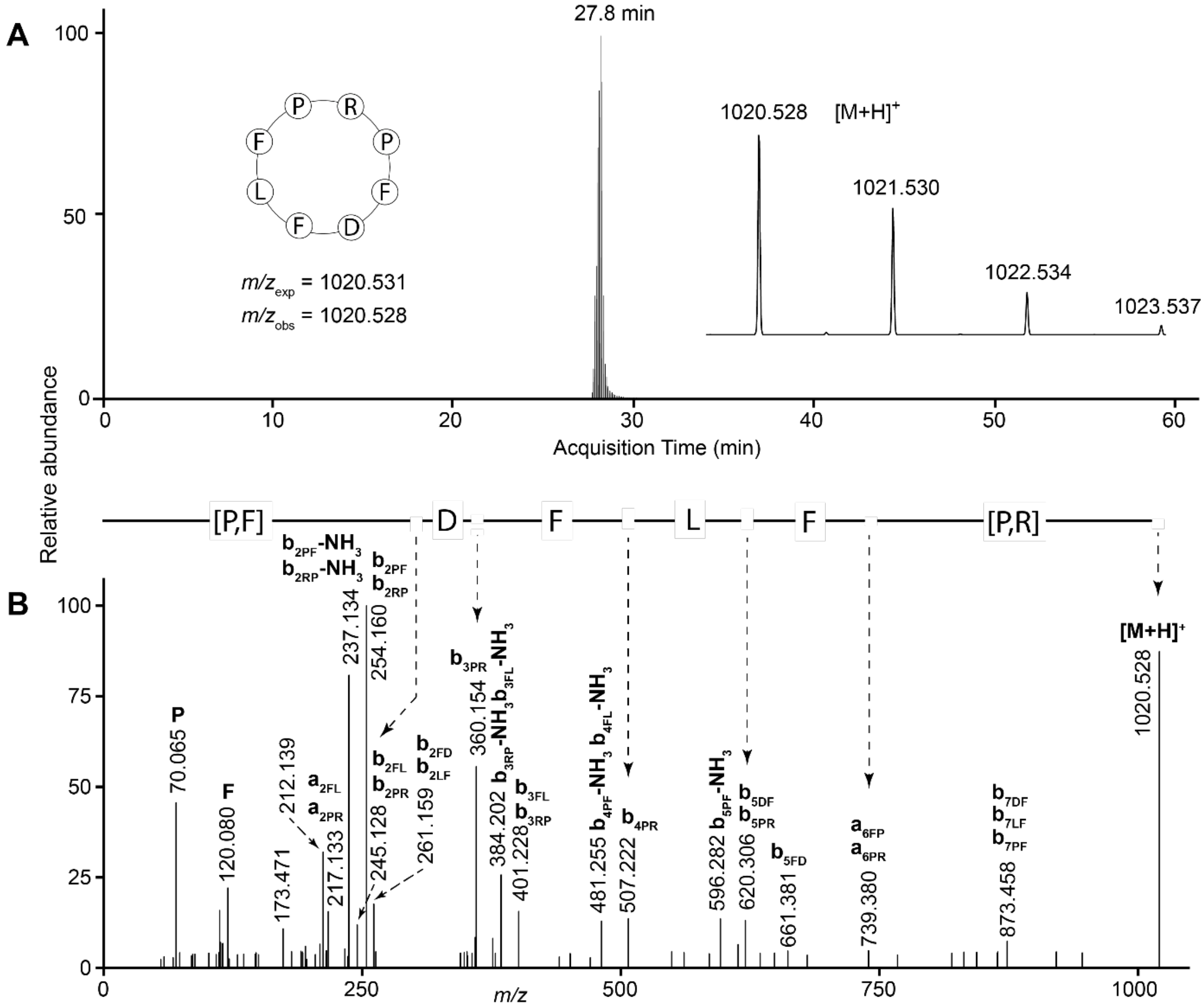


Figure S8. LC-MS data for annomuricatin K with sequence cyclo-RPFDFLFP. (A) Extracted ion chromatogram showing acquisition time of the peptide, with (inset left) peptide sequence with expected and observed mass-to-charge ratios (*m/z*) and (inset right) peptide mass spectrum. (B) Tandem mass spectrum of the fragmented precursor ion. Immonium ions are denoted by the one-letter code of the residue they represent. The derived amino acid sequence is shown above the spectrum; residues in square brackets could not be assigned an order.

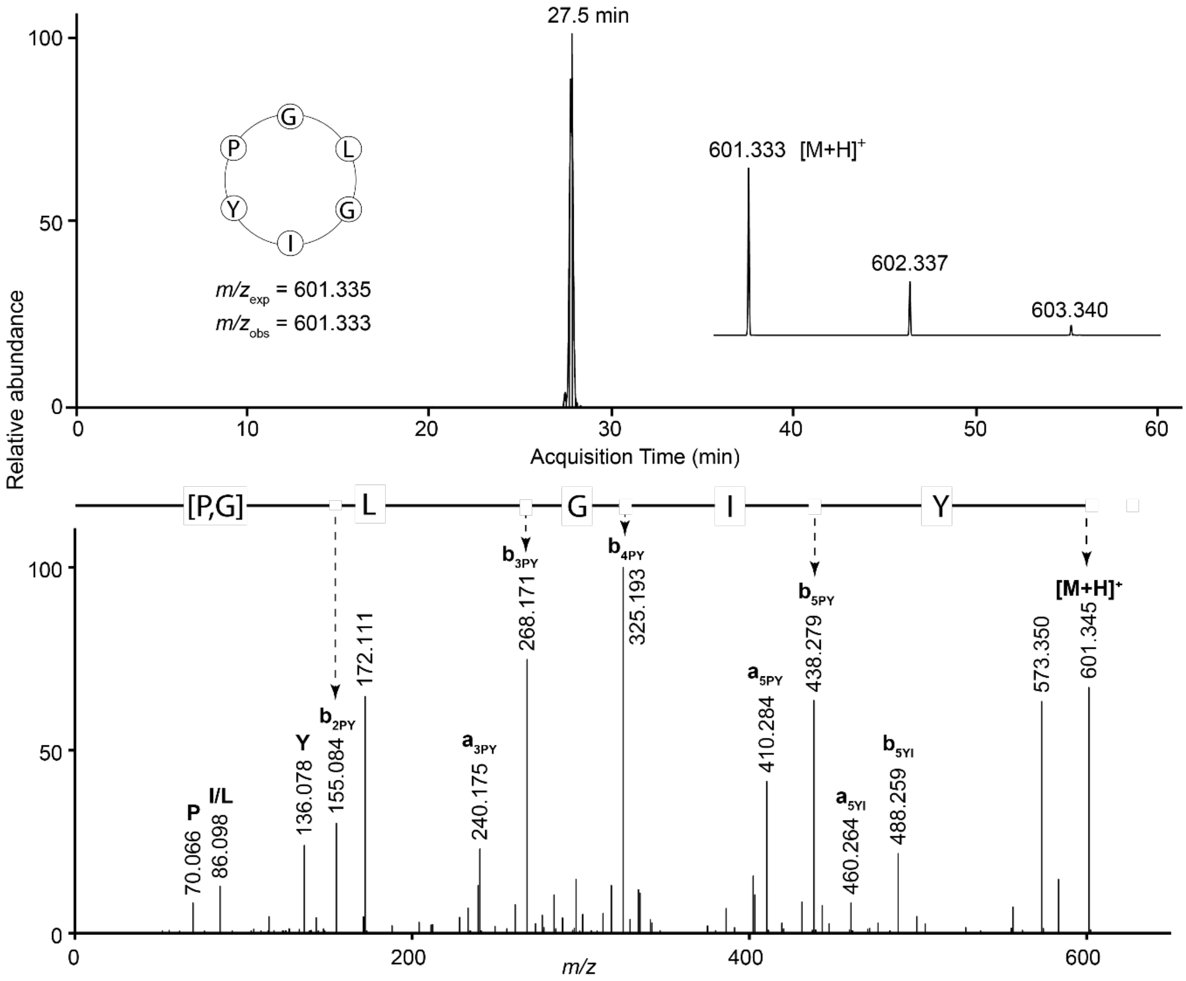


Figure S9. LC-MS data for annomuricatin L with sequence cyclo-GLGIYP. (A) Extracted ion chromatogram showing acquisition time of the peptide, with (inset left) peptide sequence with expected and observed mass-to-charge ratios (*m/z*) and (inset right) peptide mass spectrum. (B) Tandem mass spectrum of the fragmented precursor ion. Immonium ions are denoted by the one-letter code of the residue they represent. The derived amino acid sequence is shown above the spectrum; residues in square brackets could not be assigned an order.

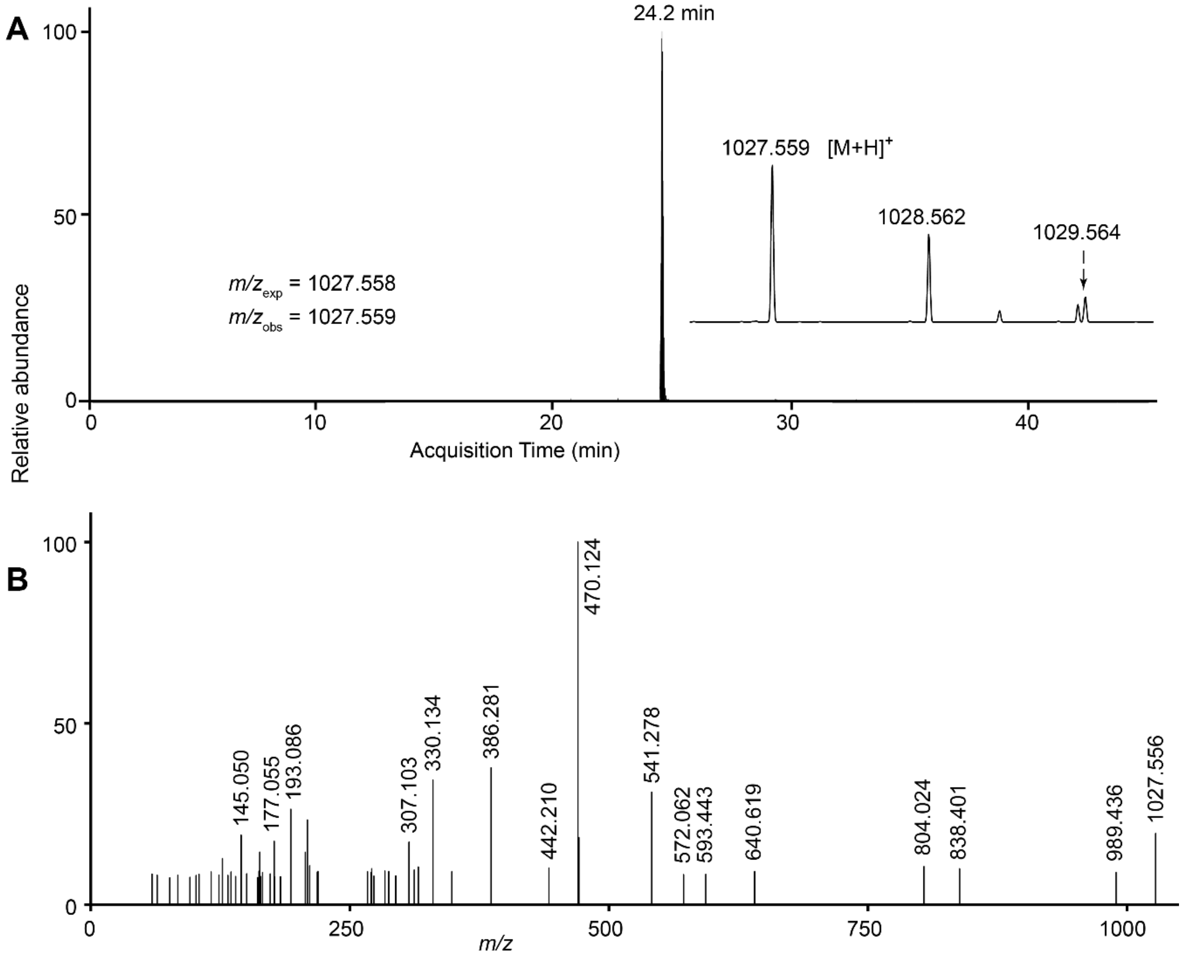


Figure S10. LC-MS data for putative peptide at *m/z* 1027.558. (A) Extracted ion chromatogram showing acquisition time of the peak, with (inset left) expected and observed mass-to-charge ratios (*m/z*) and (inset right) mass spectrum. **(B)** Tandem mass spectrum of the fragmented precursor ion.

TABLE OF CONTENTS/ABSTRACT GRAPHIC