Novel cyclic peptides in seeds of *Annona muricata* are ribosomally synthesised

Mark F. Fisher[1,2], Jingjing Zhang[1,2], Oliver Berkowitz[3], James Whelan[3] and Joshua S. Mylne[1,2]*

The University of Western Australia, [1] School of Molecular Sciences & [2] The ARC Centre of Excellence in Plant Energy Biology, 35 Stirling Highway, Crawley, Perth 6009, WA, Australia

[3] Department of Animal, Plant and Soil Sciences, School of Life Sciences & ARC Centre of Excellence in Plant Energy Biology, AgriBio, The Centre for AgriBioscience, La Trobe University, Bundoora 3086, VIC, Australia

[*] Corresponding author: Joshua S. Mylne - joshua.mylne@uwa.edu.au

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 specialised endoproteases. Cyclic peptides from the *Annona* genus display 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.

Short Summary:

By assembling a transcriptome of *Annona muricata de novo*, we show that the known orbitide, annomuricatin A, is encoded by a short transcript that also encodes a novel second peptide, annomuricatin D. We discovered and sequenced eight additional annomuricatins encoded by five transcripts.

Introduction

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-characterised. These are the cyclotides (Saether *et al.* 1995; Craik *et al.* 1999), cyclic knottins (Hernandez *et al.* 2000) and PawS-derived peptides (PDPs) (Mylne *et al.* 2011; Elliott *et al.* 2014), all of which are stabilised by disulfide bridges and are cyclised by a cysteine protease called asparaginyl endopeptidase (AEP) (Saska & Craik 2008; Gillon *et al.* 2008; Mylne *et al.* 2012; Bernath-Levin *et al.* 2015). As such, these families of cyclic peptides are known as ribosomally-synthesised 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 5 to 16 amino acid residues (Arnison *et al.* 2013; Fisher *et al.* 2019). Orbitides are also considered to be RiPPs, but much less is known about their biosynthesis than for the other peptide families mentioned above. One family of orbitides, the PawL-derived peptides (PLPs), are closely related to the PDPs, and probably cyclised by the same mechanism (Jayasena *et al.* 2017; Fisher *et al.* 2018), but most other orbitides do not contain Asp or Asn residues, which are required for cyclisation by AEP (Arnison *et al.* 2013). 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.* (2011) demonstrated that the segetalins, a group of orbitides found in the seeds of *Vaccaria hispanica* (*Saponaria vaccaria* L.), are individually encoded by short genes that express a propeptide. The researchers also showed evidence of a genetic origin for orbitides in *Dianthus caryophyllus* and some *Citrus* species. Okinyo-Owiti *et al.* (2014) characterised three novel orbitides from flax (*Linum usitatissimum* L.), 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) revealed that the orbitides curcacyclines A and B are genetically encoded (Arnison *et al.* 2013). In all these cases the cyclic peptide is excised from a highly conserved N-terminal leader sequence and C-terminal follower sequence. The core peptide (the linear precursor of the final cyclic peptide) shows little conservation, except at its N- and C-termini.

There are many other orbitides whose biosynthesis is unknown. This includes the ~35 orbitides found in plants of the Annonaceae family, including some with cytotoxic (Wélé *et al.* 2004a; Wélé *et al.* 2006) and anti-inflammatory (Chuang *et al.* 2008; Yang *et al.* 2008) activities. Two of these

orbitides are found in the seeds of *Annona muricata*. Annomuricatin A (cyclo-GPFVSA) was characterized by Li *et al.* (1995) and annomuricatin B (cyclo-PNAWLGT) by the same group (Li *et al.* 1998). A third orbitide, annomuricatin C, was reported by Wélé *et al.* (2004b), but was later shown by structural studies to be identical to annomuricatin A (Wu *et al.* 2007).

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 synthesised. We also identified transcripts encoding a further eight novel annomuricatins (E-L) and sequenced them by mass spectrometry.

Results and Discussion

To characterise the biosynthesis of the annomuricatins we extracted total RNA from *A. muricata* seeds, performed RNA-seq and assembled a transcriptome using established methods (Jayasena *et al.* 2017). Using cyclic permutations of the two known peptides annomuricatin A and annomuricatin B, we searched the transcriptome for sequences encoding them; we were overwhelmed with hundreds of contigs that had the potential to encode annomuricatin A, but found none that could code for annomuricatin B.

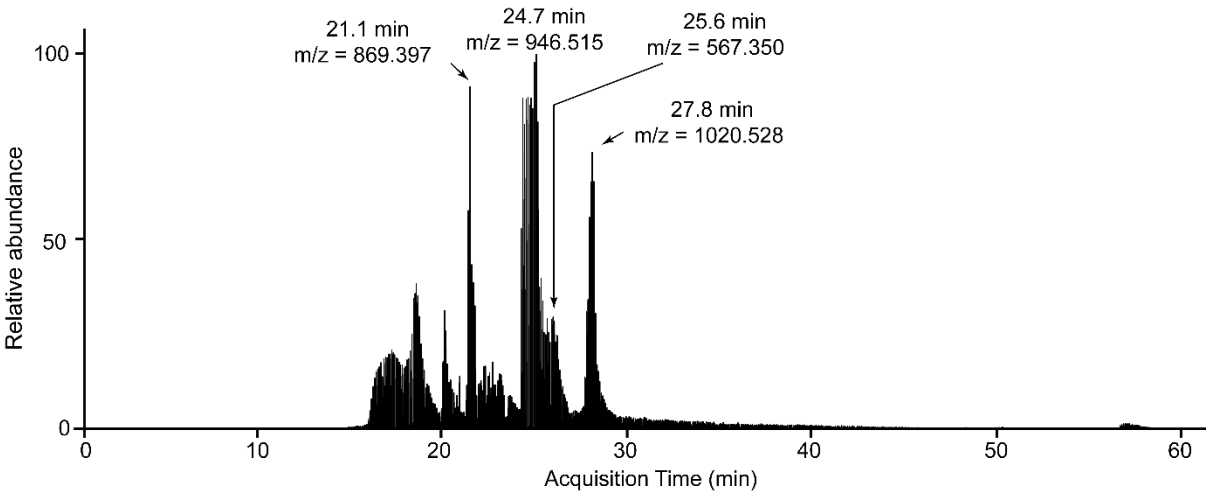


Figure 1. Total ion current chromatogram for extracts of *Annona muricata* seeds with the m/z values and the acquisition times of the largest peaks labelled. All ions shown are $[M+H]^+$.

An 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 (**Fig. 1**). One of these, eluted at 24.7 min, was especially abundant and the protonated molecule had a measured m/z of 946.515 $[M+H]^+$ (**Fig. 2A**). Treatment of the sample

with 1.2 M hydrochloric acid followed by LC-MS/MS 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 (**Fig. 2B**). If the 946 m/z ion was a novel annomuricatin, this could be interpreted as partial acid hydrolysis of its peptide backbone leading to an 18 Da mass increase.

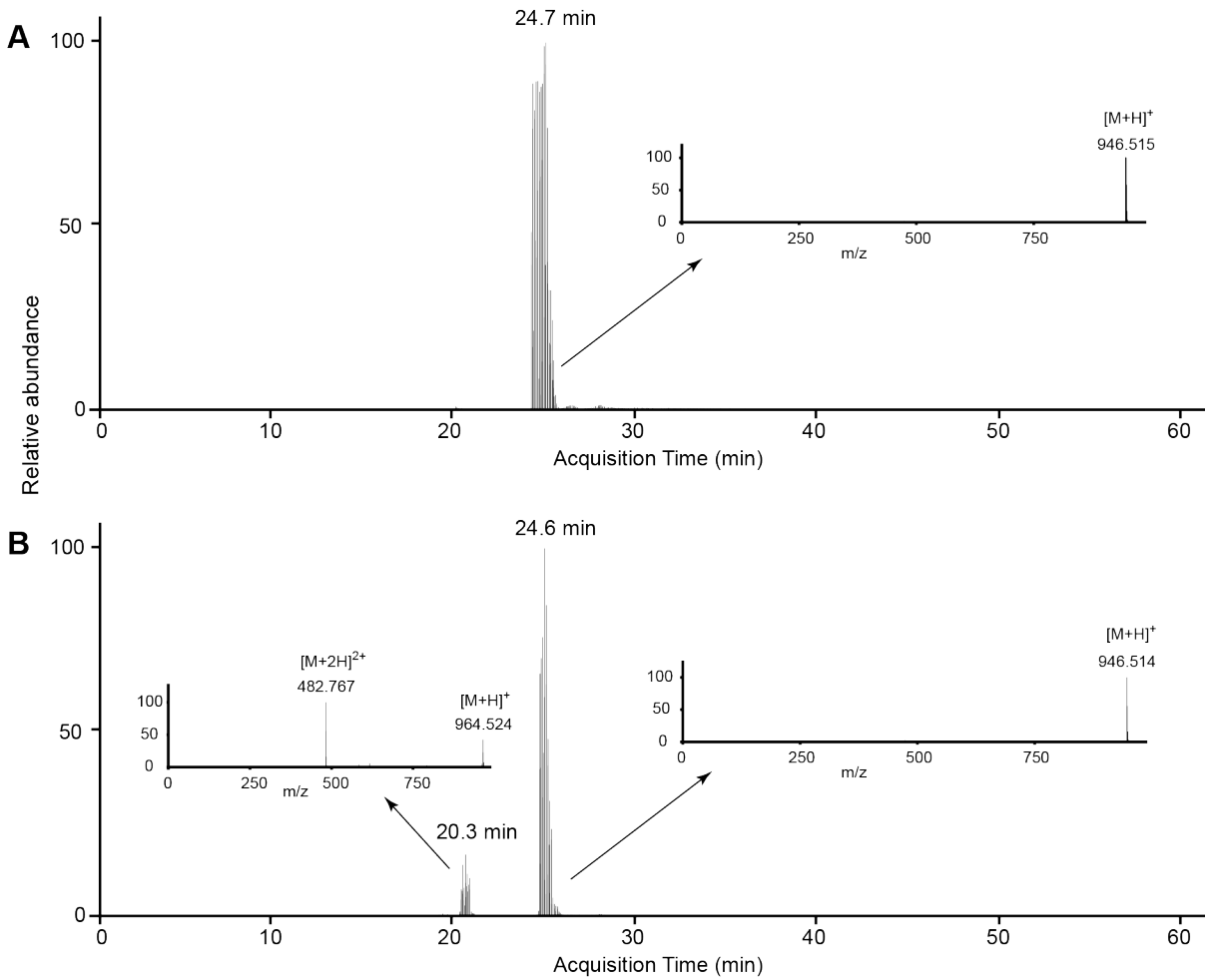


Figure 2. Combined extracted ion chromatograms at m/z 946.515 and m/z 964.526 for peptide extracts of *Annona muricata* seeds. **A**: Native extract; **B**: After treatment with dilute hydrochloric acid caused the appearance of peaks at m/z 964.524 $[M+H]^{+}$ and 482.767 $[M+2H]^{2+}$ in the mass spectrum, 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 hydrolysed 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) (**Fig. 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.

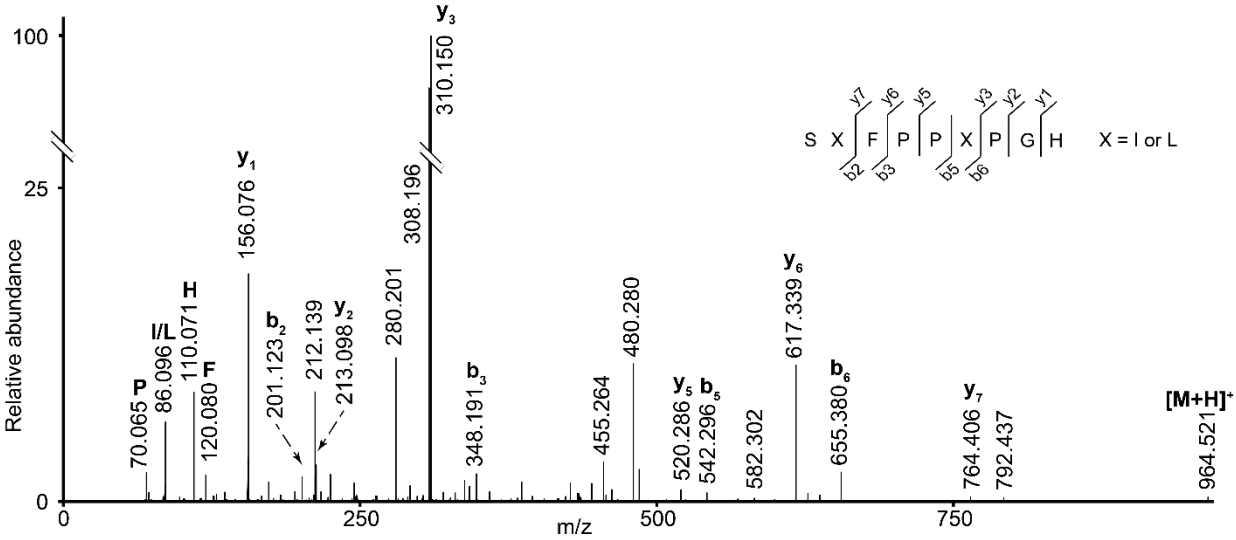


Figure 3. MS/MS confirmation of annomuricatin D. Partial hydrolysis using hydrochloric acid gave a sequence of b and y ions, with the ring breakage C-terminal to the His residue. (inset shows the peptide sequence). Note that the order of the Ser and Ile/Leu residues could not be unambiguously determined.

Identification of novel transcripts encoding annomuricatins

Returning to search the *Annona 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 (**Fig. 4**).

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 *ProannomuricatinAD*, abbreviated to *PamAD*. Using the *PamAD* transcript as a tBLASTn query, we discovered five similar coding sequences with the potential to encode nine annomuricatins.

```
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 *Annona 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 a lack of peptide mass evidence, we found no evidence of a transcript matching the previously reported annomuricatin B sequence (Li *et al.* 1998). To confirm the sequences obtained by RNA-seq and assembly of contigs, we designed primers 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 tandem mass spectrometry. 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 (**Supp. Fig. 1-9**). Annomuricatin I was found in two forms having either a reduced or oxidised methionine (**Supp. Fig. 6**). 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 we found peaks 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 (**Sup. Fig. 10**). 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.* 1995)
B	NAWLGTP [a]	*N.D.*	N.D.
D	GHSIFPPIP	*PamAD*	**Fig. 2**, **Fig. 3**
E	GLVTVP	*PamEF*	**Fig. S1**
F	GLSAVTP	*PamEF*	**Fig. S2**
G	MIGDYSYP	*PamG*	**Fig. S3**
H	TLVPGYSP	*PamHI*	**Fig. S4**
I	GFMHSPVP [b]	*PamHI*	**Fig. S5**, **Fig. S6** [b]
J	GPPLYLA	*PamJK*	**Fig. S7**
K	RPFDFLFP	*PamJK*	**Fig. S8**
L	GLGIYP	*PamL*	**Fig. S9**

[a] Annomuricatin B, which we could not detect at either the peptide or transcript level, was proposed by Li *et al.* (1998) 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 (**Fig. S5**) and the other with an oxidized-Met (**Fig. S6**).

Like many other cyclic peptides, most of the core peptide sequence is not highly conserved, except for the N-terminus and C-terminus 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 (Moriyama,

Nakanishi & Sasaki 1988). 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 (**Fig. 5**), and it is tempting to speculate that this is in some way required for the prolyl oligopeptidase to cyclise 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 macrocyclisation.

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, we downloaded RNA-seq data from the NCBI Sequence Read Archive. These data were generated from RNA isolated from the mature flowers of *Annona squamosa* (Liu *et al.* 2016) and leaves of *Annona muricata* (www.onekp.com). 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. We identified three candidate transcripts in *A. muricata* leaves, which were different from those in the *A. muricata* seeds, encoding a total of six putative peptides (**Fig. 6A**). We also found ten candidate transcripts from *A. squamosa*, mostly encoding two possible peptides, though several varied considerably from the canonical sequence, with the putative core peptide lacking a C-terminal Pro or Ala (**Fig. 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) *Annona muricata* leaves and (B) *Annona 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. The last sequence shown corresponds to the known cyclic peptide, cyclosquamosin D (Morita, Sato & Kobayashi 1999), 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 homologues 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 tissue-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 (**Fig. 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 (Saska & Craik 2008; Gillon *et al.* 2008; Mylne *et al.* 2012; Bernath-Levin *et al.* 2015); a similar depth of understanding exists for the orbitide segetalin A and its relatives, whose cyclisation of a conserved C-terminal Ala residue is performed by the prolyl oligopeptidase PCY1 (Barber *et al.* 2013; Condie *et al.* 2011). Prior to cyclisation, the segetalins are cleaved at the N-terminus of the core peptide by an as-yet-uncharacterised enzyme, OLP1 (Barber *et al.* 2013), which presumably is able to recognise the conserved Gln or Glu residue at the P1 position. Another example of peptides cyclised 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 (Luo *et al.* 2014). 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 *Vaccaria 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 *Vaccaria hispanica* indicates that a prolyl oligopeptidase is the likely cyclisation agent. Further study is required to identify this POP and to characterise its structure and mechanism of action.

Materials and Methods

Plant material

Seeds of *Annona 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 described by Jayasena *et al.* (2017). Briefly, peptides were extracted in 50% methanol, 50% dichloromethane. Phases were separated by the alternate addition of chloroform and 0.1% trifluoroacetic acid. The upper, aqueous phase was dried overnight in a vacuum centrifuge prior to purification.

Purification of seed extracts

Seed peptide extracts were purified according to the method previously described by Fisher *et al.* (2018). 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% acetonitrile (v/v) and 0.1% formic acid (v/v), then purified peptides were eluted with 85% acetonitrile (v/v) and 0.1% formic acid (v/v). The extract was dried in a vacuum centrifuge (Labconco) and redissolved in 5% acetonitrile (v/v) and 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 acetonitrile (Fisher Scientific). The resulting electrospray (source voltage 1,800 V) was analysed by an Orbitrap Fusion mass spectrometer (Thermo Fisher Scientific) running in positive ionisation, data-dependent, "top speed" MS/MS mode, employing the Orbitrap mass analyser 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 (Niedermeyer & Strohalm 2012) for cyclic peptides and MS Product on the ProteinProspector website for acyclic peptides (http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msproduct).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). We left other parameters at their default values. We chose similar options 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 *Annona muricata* (100 mg) were ground to a powder using a mortar and pestle. 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 SRA accession PRJNA539840.

The *Annona 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 30 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

We downloaded RNA-seq paired-read data from the Sequence Read Archive of the National Institutes of Health for *Annona squamosa* mature flowers (run SRR3478571) and *Annona muricata* leaves (run ERR2040135). Using CLC Genomics Workbench 11.0, we assembled transcriptomes 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 *Annona 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 (**Supp. Table 1**).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.

Funding Information

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.

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 *Annona muricata* transcriptome; M.F.F performed all other experiments and analysed data; M.F.F and J.S.M. wrote the manuscript with help from all other authors.

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

Arnison, PG, Bibb, MJ, Bierbaum, G, Bowers, AA, Bugni, TS, Bulaj, G, Camarero, JA, Campopiano, DJ, Challis, GL*, et al.* 2013, 'Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature', *Nat Prod Rep*, vol. 30, no. 1, pp. 108-160.

Barber, CJS, Pujara, PT, Reed, DW, Chiwocha, S, Zhang, H & Covello, PS 2013, 'The two-step biosynthesis of cyclic peptides from linear precursors in a member of the plant family Caryophyllaceae involves cyclization by a serine protease-like enzyme', *J Biol Chem*, vol. 288, no. 18, pp. 12500-12510.

Bernath-Levin, K, Nelson, C, Elliott, AG, Jayasena, AS, Millar, AH, Craik, DJ & Mylne, JS 2015, 'Peptide macrocyclization by a bifunctional endoprotease', *Chem Biol*, vol. 22, no. 5, pp. 571-582.

Chuang, PH, Hsieh, PW, Yang, YL, Hua, KF, Chang, FR, Shiea, J, Wu, SH & Wu, YC 2008, 'Cyclopeptides with anti-inflammatory activity from seeds of *Annona montana*', *J Nat Prod*, vol. 71, no. 8, pp. 1365-1370.

Condie, JA, Nowak, G, Reed, DW, John Balsevich, J, Reaney, MJT, Arnison, PG & Covello, PS 2011, 'The biosynthesis of Caryophyllaceae-like cyclic peptides in *Saponaria vaccaria* L. from DNA-encoded precursors', *Plant J*, vol. 67, pp. 682–690.

Craik, DJ, Daly, NL, Bond, T & Waine, C 1999, 'Plant cyclotides: A unique family of cyclic and knotted proteins that defines the cyclic cystine knot structural motif', *J Mol Biol*, vol. 294, no. 5, pp. 1327-1336.

Elliott, AG, Delay, C, Liu, H, Phua, Z, Rosengren, KJ, Benfield, AH, Panero, JL, Colgrave, ML, Jayasena, AS*, et al.* 2014, 'Evolutionary origins of a bioactive peptide buried within preproalbumin', *Plant Cell*, vol. 26, no. 3, pp. 981-995.

Fisher, MF, Zhang, J, Berkowitz, O, Whelan, J & Mylne, JS 2019, *An orbitide from Ratibida columnifera seeds containing 16 amino acid residues*.

Fisher, MF, Zhang, J, Taylor, NL, Howard, MJ, Berkowitz, O, Debowski, AW, Behsaz, B, Whelan, J, Pevzner, PA*, et al.* 2018, 'A family of small, cyclic peptides buried in preproalbumin since the Eocene epoch', *Plant Direct*, vol. 2, no. 2, p. e00042.

Gillon, AD, Saska, I, Jennings, CV, Guarino, RF, Craik, DJ & Anderson, MA 2008, 'Biosynthesis of circular proteins in plants', *Plant J*, vol. 53, no. 3, pp. 505-515.

Hernandez, JF, Gagnon, J, Chiche, L, Nguyen, TM, Andrieu, JP, Heitz, A, Trinh Hong, T, Pham, TT & Le Nguyen, D 2000, 'Squash trypsin inhibitors from *Momordica cochinchinensis* exhibit an atypical macrocyclic structure', *Biochemistry*, vol. 39, no. 19, pp. 5722-5730.

Jayasena, AS, Fisher, MF, Panero, JL, Secco, D, Bernath-Levin, K, Berkowitz, O, Taylor, NL, Schilling, EE, Whelan, J*, et al.* 2017, 'Stepwise evolution of a buried inhibitor peptide over 45 My', *Mol Biol Evol*, vol. 34, no. 6, pp. 1505-1516.

Li, C, Tan, N, Lu, Y, Liang, H, Mu, Q, Zheng, H, Hao, X & Zhou, J 1995, 'Annomuricatin A, a new cyclopeptide from the seeds of Annona muricata', *Acta Bot Yunnanica*, vol. 17, no. 4, pp. 459-462.

Li, C, Tan, N, Zheng, H, Mu, Q, Hao, X, He, Y & Zou, J 1998, 'Cyclopeptide from the seeds of Annona muricata', *Phytochemistry*, vol. 48, no. 3, pp. 555-556.

Liu, K, Feng, S, Pan, Y, Zhong, J, Chen, Y, Yuan, C & Li, H 2016, 'Transcriptome Analysis and Identification of Genes Associated with Floral Transition and Flower Development in Sugar Apple (Annona squamosa L.)', *Front Plant Sci*, vol. 7, p. 1695.

Luo, H, Hong, S-Y, Sgambelluri, RM, Angelos, E, Li, X & Walton, Jonathan D 2014, 'Peptide macrocyclization catalyzed by a prolyl oligopeptidase involved in α-amanitin biosynthesis', *Chem Biol*, vol. 21, no. 12, pp. 1610-1617.

Morita, H, Sato, Y & Kobayashi, Ji 1999, 'Cyclosquamosins A–G, cyclic peptides from the seeds of Annona squamosa', *Tetrahedron.*, vol. 55, no. 24, pp. 7509-7518.

Moriyama, A, Nakanishi, M & Sasaki, M 1988, 'Porcine Muscle Prolyl Endopeptidase and Its Endogenous Substrates', *The Journal of Biochemistry*, vol. 104, no. 1, pp. 112-117.

Mylne, JS, Chan, LY, Chanson, AH, Daly, NL, Schaefer, H, Bailey, TL, Nguyencong, P, Cascales, L & Craik, DJ 2012, 'Cyclic peptides arising by evolutionary parallelism via asparaginyl-endopeptidase-mediated biosynthesis', *Plant Cell*, vol. 24, pp. 2765-2778.

Mylne, JS, Colgrave, ML, Daly, NL, Chanson, AH, Elliott, AG, McCallum, EJ, Jones, A & Craik, DJ 2011, 'Albumins and their processing machinery are hijacked for cyclic peptides in sunflower', *Nat Chem Biol*, vol. 7, no. 5, pp. 257-259.

Niedermeyer, THJ & Strohalm, M 2012, 'mMass as a software tool for the annotation of cyclic peptide tandem mass spectra', *PLoS One*, vol. 7, no. 9, p. e44913.

Okinyo-Owiti, DP, Young, L, Burnett, P-GG & Reaney, MJT 2014, 'New flaxseed orbitides: Detection, sequencing, and 15N incorporation', *Peptide Science*, vol. 102, no. 2, pp. 168-175.

Saether, O, Craik, DJ, Campbell, ID, Sletten, K, Juul, J & Norman, DG 1995, 'Elucidation of the primary and three-dimensional structure of the uterotonic polypeptide kalata B1', *Biochemistry*, vol. 34, pp. 4147-4158.

Saska, I & Craik, DJ 2008, 'Protease-catalysed protein splicing: a new post-translational modification?', *Trends Biochem Sci*, vol. 33, no. 8, pp. 363-368.

Wélé, A, Landon, C, Labbe, H, Vovelle, F, Zhang, Y & Bodo, B 2004a, 'Sequence and solution structure of cherimolacyclopeptides A and B, novel cyclooctapeptides from the seeds of Annona cherimola', *Tetrahedron.*, vol. 60, no. 2, pp. 405-414.

Wélé, A, Zhang, Y, Caux, C, Brouard, J-P, Pousset, J-L & Bodo, B 2004b, 'Annomuricatin C, a novel cyclohexapeptide from the seeds of *Annona muricata*', *C R Chim*, vol. 7, no. 10–11, pp. 981-988.

Wélé, A, Zhang, Y, Dubost, L, Pousset, J-L & Bodo, B 2006, 'Cyclic peptides from the seeds of Annona glauca and A. cherimola', *Chem Pharm Bull*, vol. 54, no. 5, pp. 690-692.

Wu, L, Lu, Y, Zheng, Q-T, Tan, N-H, Li, C-M & Zhou, J 2007, 'Study on the spatial structure of annomuricatin A, a cyclohexapeptide from the seeds of *Annona muricata*', *J Mol Struct*, vol. 827, no. 1, pp. 145-148.

Yang, YL, Hua, KF, Chuang, PH, Wu, SH, Wu, KY, Chang, FR & Wu, YC 2008, 'New cyclic peptides from the seeds of Annona squamosa L. and their anti-inflammatory activities', *J Agric Food Chem*, vol. 56, no. 2, pp. 386-392.

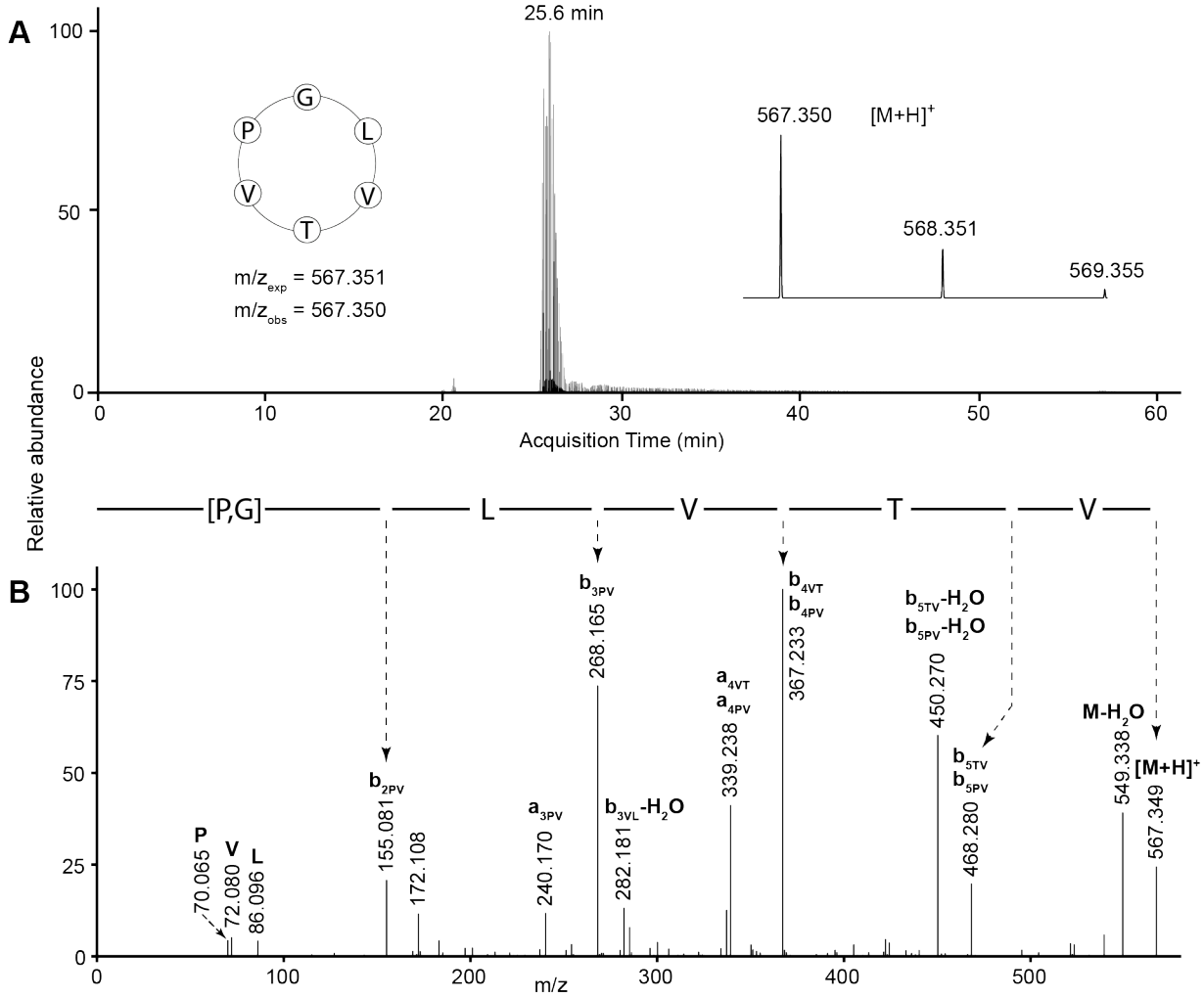
Supplementary Information

Supplementary Tables

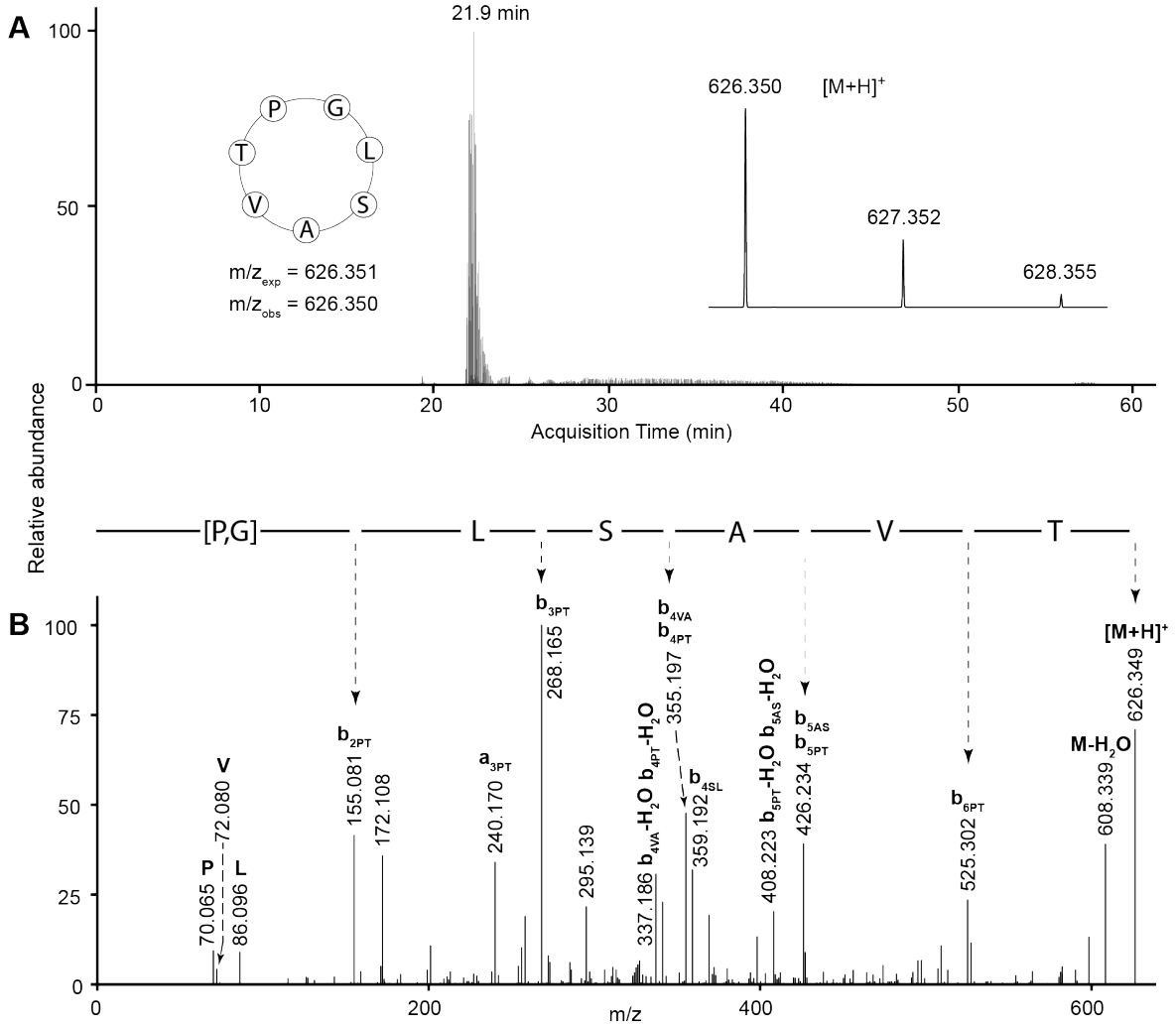
Supplementary Table 1. PCR primers used for the amplification and sequencing of the putative *Proannomuricatin* genes identified in transcriptome data.

Gene	Forward Primer	Reverse Primer
ProannomuricatinAD	5′-TGG GAG ACA GGA GAT ACA GAG AGA GA-3′	5′-ATC GGT AGA AAG AGA GAG AGA CAA GAC AG-3′
ProannomuricatinEF	5′-AGT CGC AGC TCG CAG GTA CA-3′	5′-AGC AGC AAA GAA ACA CAG CAG AAA-3′
ProannomuricatinG	5′-AGG AGC GAA GTG CCT TCG TTG TA-3′	5′-TGC ACA CAC ATG CGG CAA GG-3′
ProannomuricatinHI	5′-AGC TAA GCA ATC GCA GTT GGC A-3′	5′-GGA AGA GAA GAC AGG ACG GCC A-3′
ProannomuricatinJK	5′-TGC CCA ATA TCG TGA GGG ACG C-3′	5′-AGA TGG CCA CCT CGC TCC AT-3′
ProannomuricatinL	5′-TGG ACA GAC ATG GAG TTG GTG C-3′	5′-TGA CAA GGA CTT CCA TCG AGG C-3′

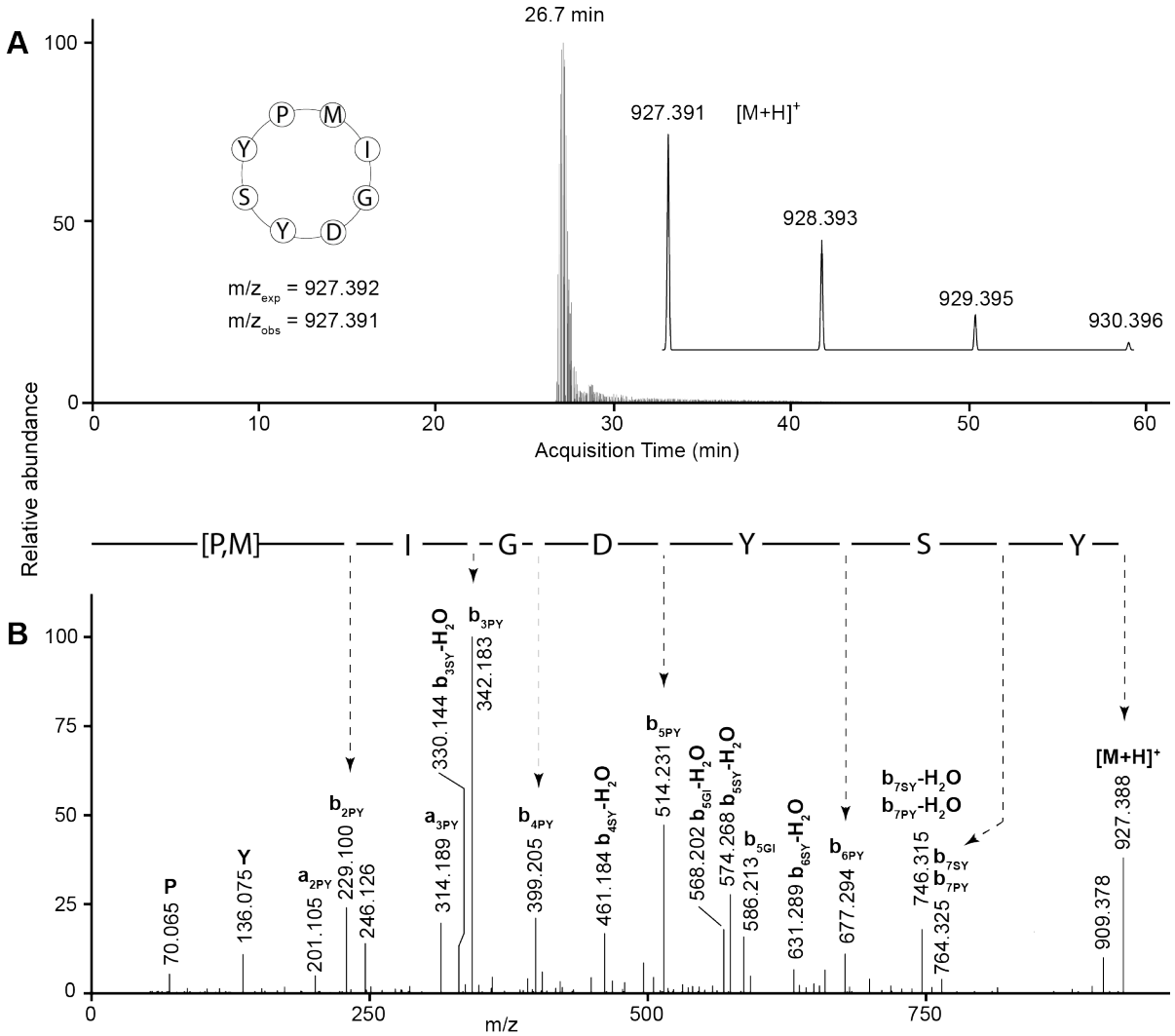
Supplementary Figures



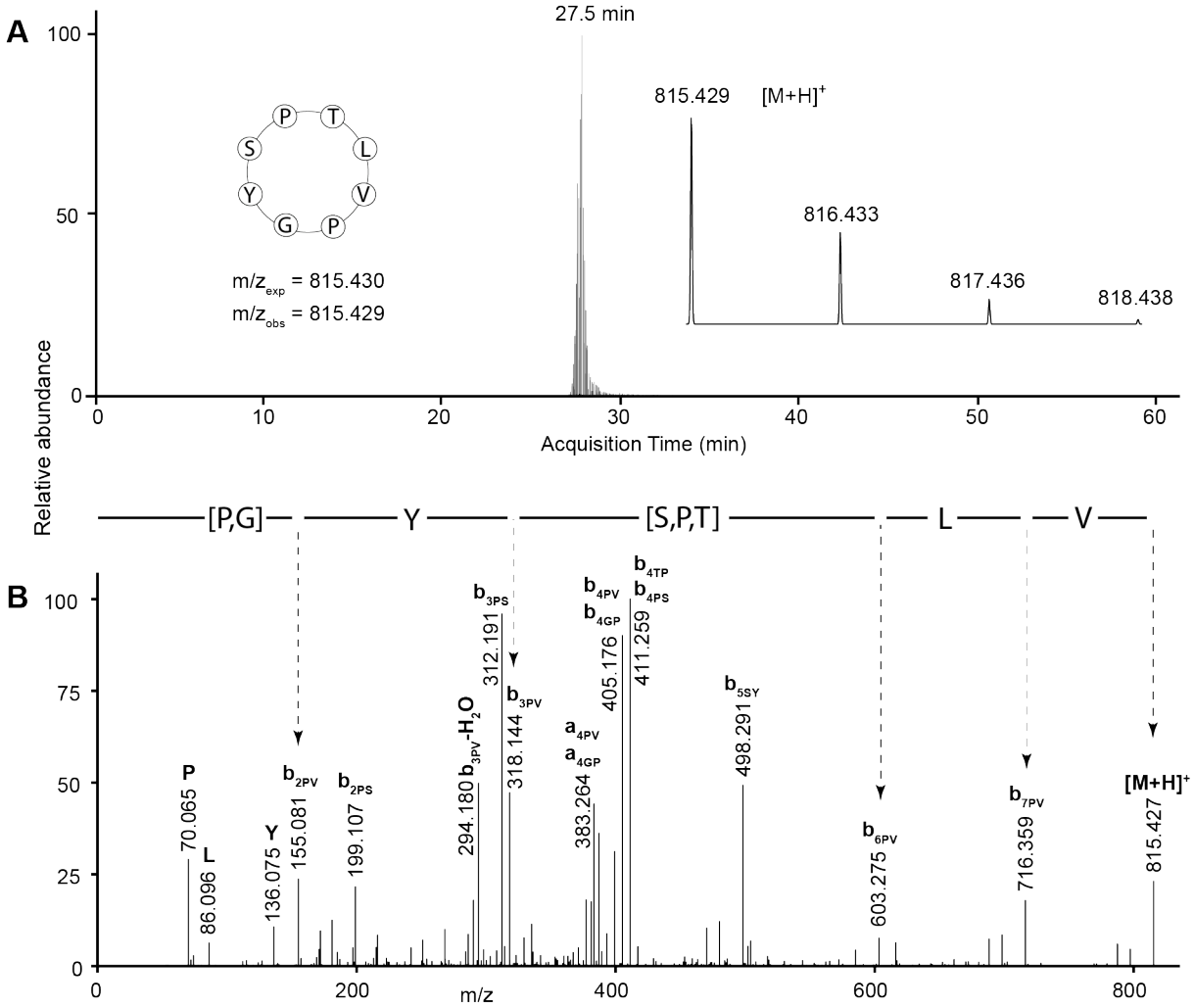
Supplementary Figure 1. LC-MS data for annomuricatin E with sequence cyclo-GLSAVTP. **A**: Extracted ion chromatogram showing acquisition time of the peptide, with (inset left) peptide sequence with expected and observed mass-to-charge ratios (m/z) and (inset right) peptide mass spectrum. **B**: Tandem 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.



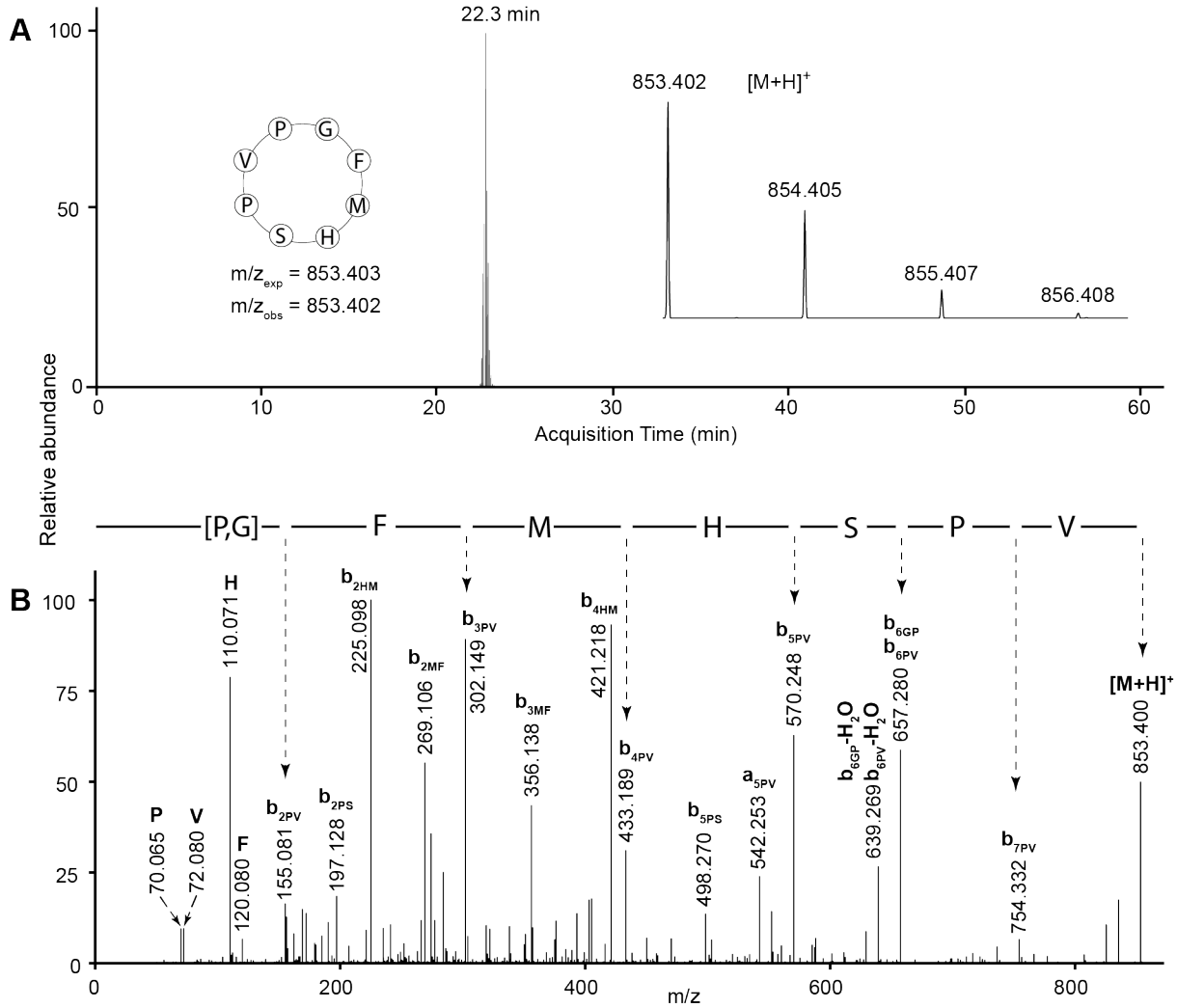
Supplementary Figure 2. LC-MS data for annomuricatin F with sequence cyclo-GLSAVTP. **A**: Extracted ion chromatogram showing acquisition time of the peptide, with (inset left) peptide sequence with expected and observed mass-to-charge ratios (m/z) and (inset right) peptide mass spectrum. **B**: Tandem 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.



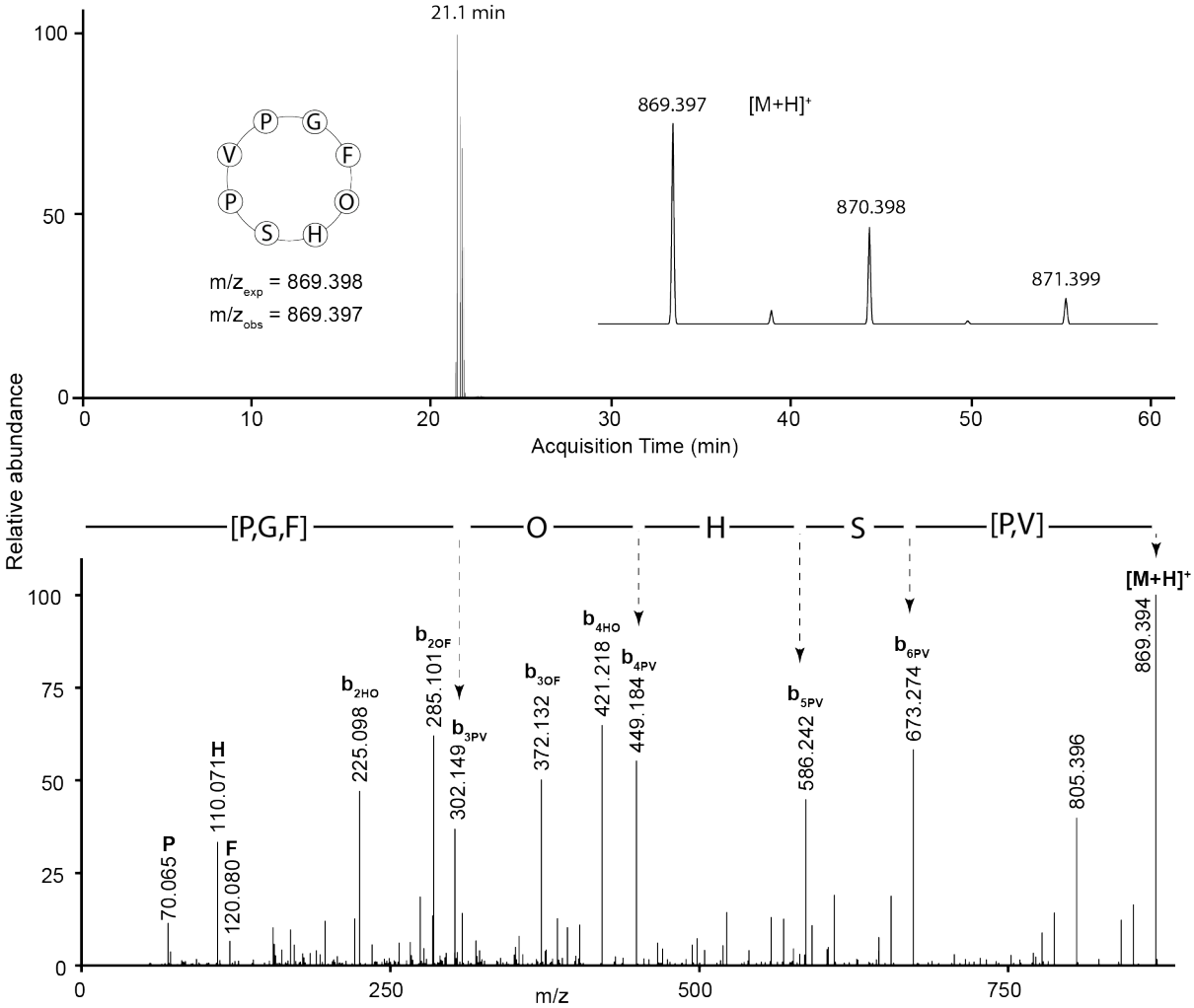
Supplementary Figure 3. LC-MS data for annomuricatin G with sequence cyclo-MIGDYSYP. **A**: Extracted ion chromatogram showing acquisition time of the peptide, with (inset left) peptide sequence with expected and observed mass-to-charge ratios (m/z) and (inset right) peptide mass spectrum. **B**: Tandem 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.



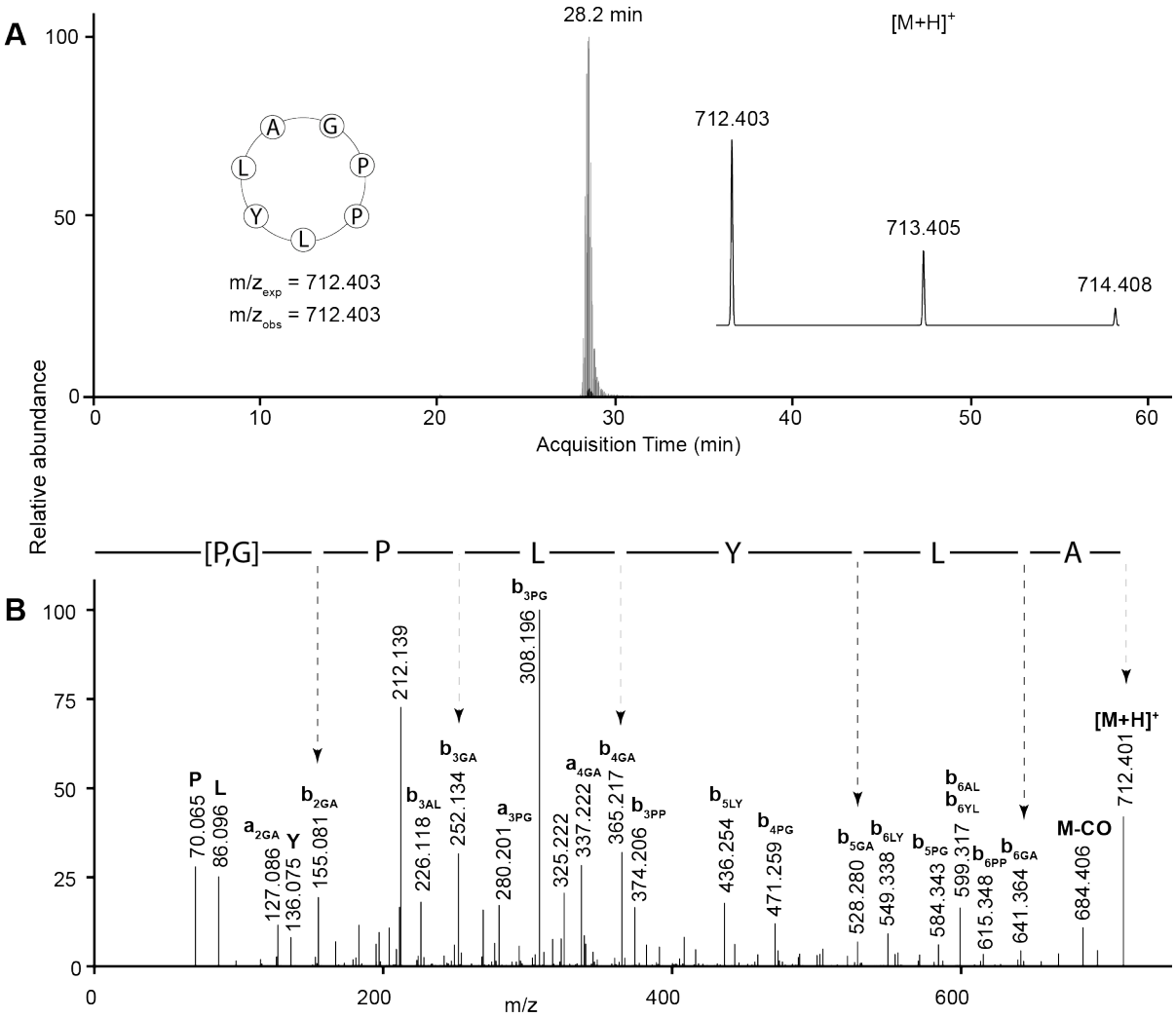
Supplementary Figure 4. LC-MS data for annomuricatin H with sequence cyclo-TLVPGYSP. **A**: Extracted ion chromatogram showing acquisition time of the peptide, with (inset left) peptide sequence with expected and observed mass-to-charge ratios (m/z) and (inset right) peptide mass spectrum. **B**: Tandem 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.



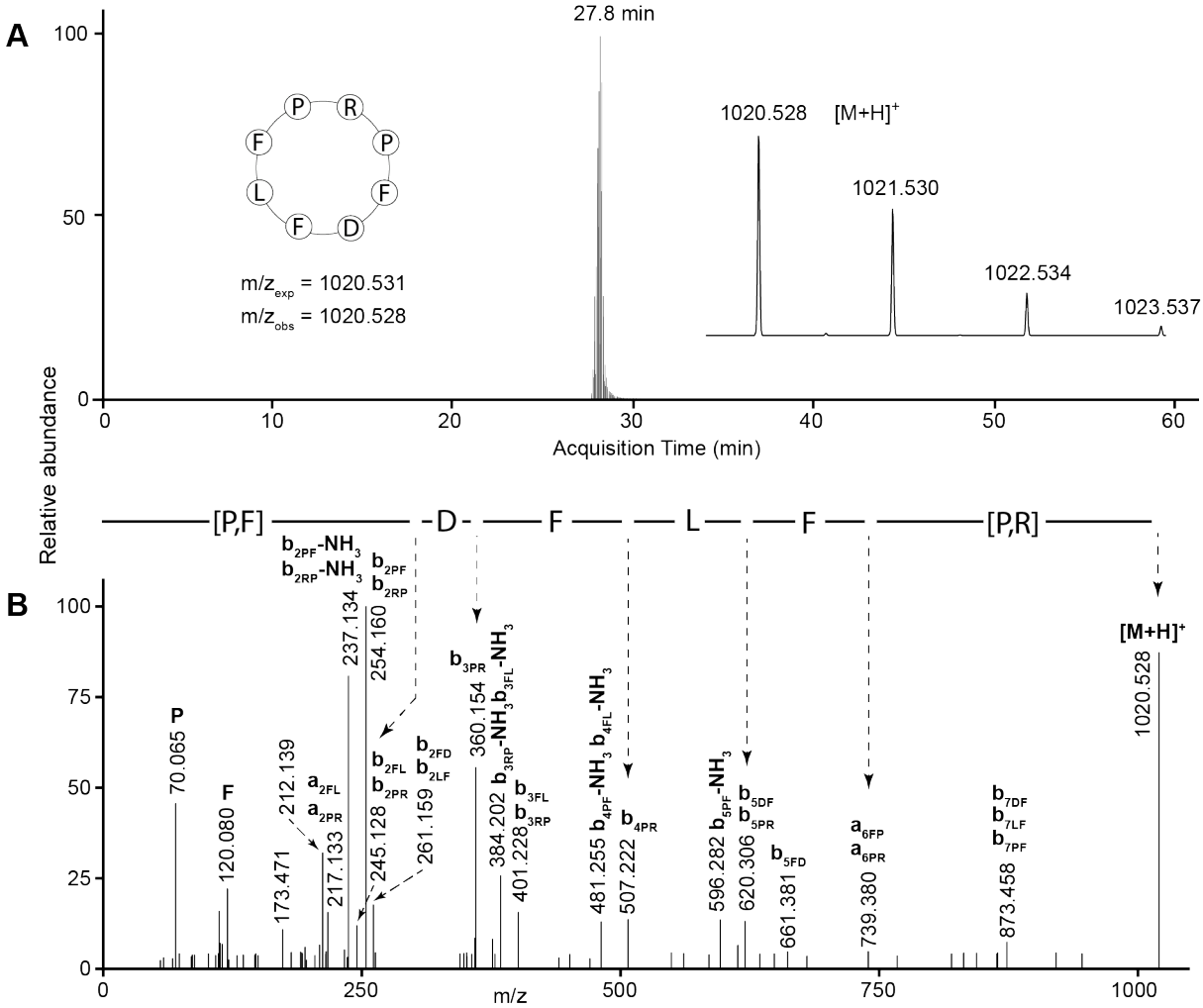
Supplementary Figure 5. LC-MS data for annomuricatin I with sequence cyclo-GFMHSPVP. **A**: Extracted ion chromatogram showing acquisition time of the peptide, with (inset left) peptide sequence with expected and observed mass-to-charge ratios (m/z) and (inset right) peptide mass spectrum. **B**: Tandem 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.



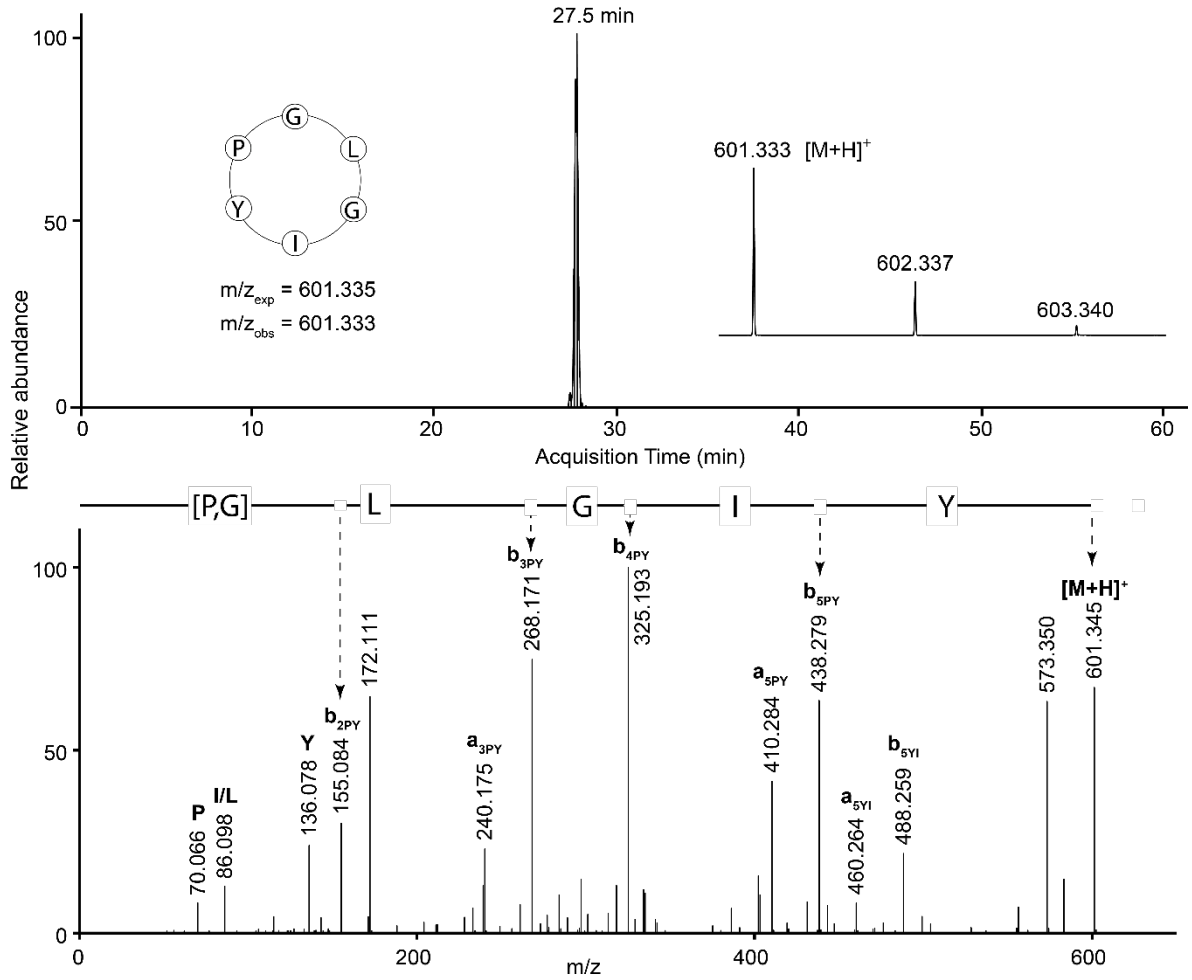
Supplementary Figure 6. LC-MS data for annomuricatin I with sequence cyclo-GFMHSPVP, although the Met was oxidized (hence O in figure). **A**: Extracted ion chromatogram showing acquisition time of the peptide, with (inset left) peptide sequence with expected and observed mass-to-charge ratios (m/z) 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.



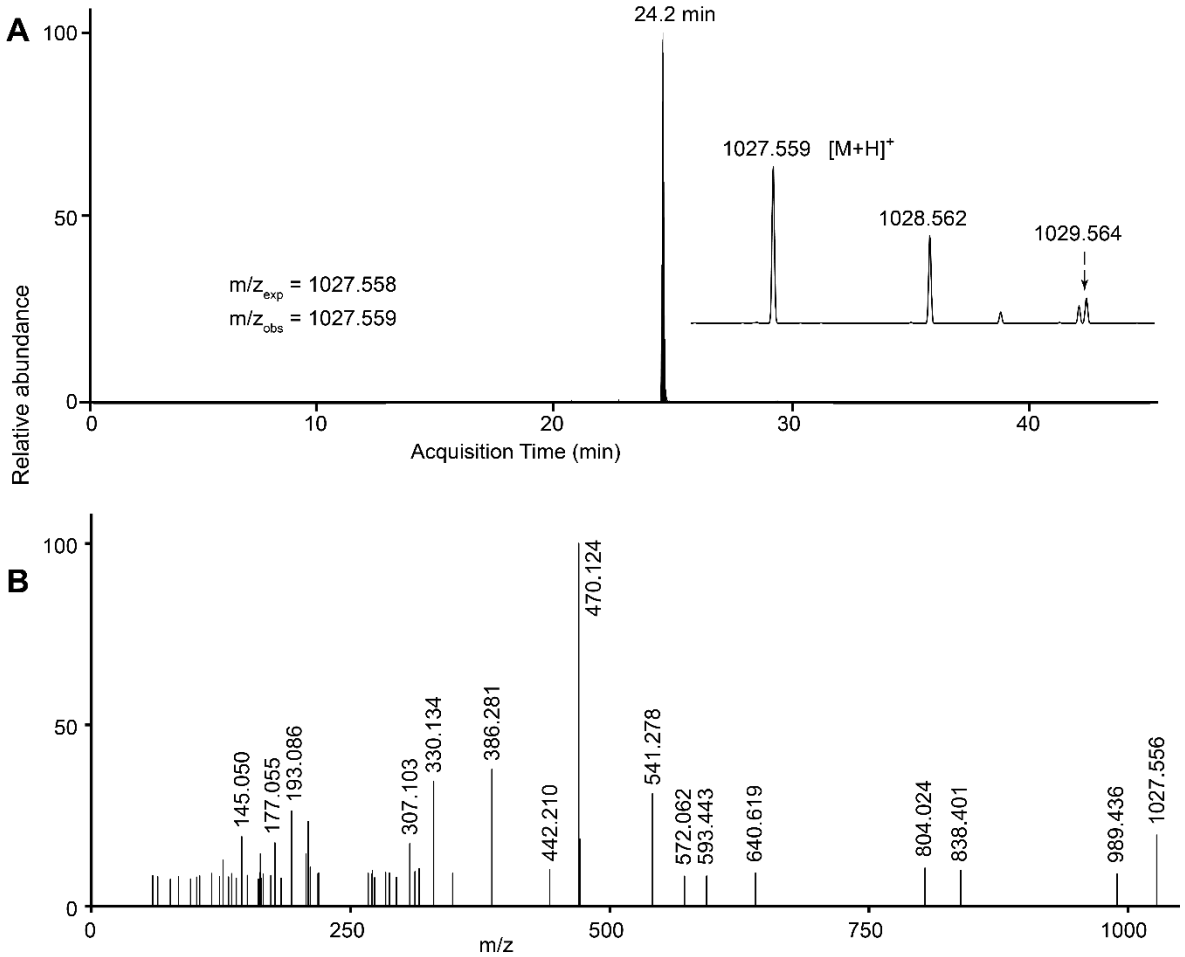
Supplementary Figure 7. LC-MS data for annomuricatin J with sequence cyclo-GPPLYLA. **A**: Extracted ion chromatogram showing acquisition time of the peptide, with (inset left) peptide sequence with expected and observed mass-to-charge ratios (m/z) and (inset right) peptide mass spectrum. **B**: Tandem 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.



Supplementary Figure 8. 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.



Supplementary Figure 9. 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.



Supplementary Figure 10. LC-MS data for putative peptide at m/z 1017.558. **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.