AMPOLINI, BRIGITTE G. M.S. Bioinformatic Discovery of New Ribosomally Synthesized and Post-translationally Modified Peptides in Plants and Fungi. (2023)
Directed by Dr. Jonathan Chekan. 36 pp.

Ribosomally synthesized and post-translationally modified peptides (RiPPs) are a chemically diverse class of natural products with exciting potential. The genomic basis of RiPPs creates a unique opportunity to discover new natural products bioinformatically using genome and transcriptome mining. Class specific features of RiPP gene clusters can be used to guide the bioinformatic analysis to discover new molecules and enzymes. Here, we searched for new fungal dikaritins by genome mining with a diagnostic biosynthetic enzyme, leading to the discovery of 77 putative new dikaritins. Applying a similar genome mining strategy to plants, we created a custom hidden Markov model to define a new class of plant cyclopeptides called burptides and to search through all Viridiplantae genomes for the potential to make new burptides. Ultimately, this bioinformatic analysis led to the discovery of a new molecule from *Coffea arabica*: arabipeptin A. To continue looking for new burptides, we used our custom hidden Markov model to search publicly available raw transcriptomic data. This process involved the creation of a pipeline that automates the process of downloading, assembling, and analyzing transcriptomic data with the custom burptide HMM. The results of the transcriptome search yielded 67 potential producers of novel burptides. RiPP cores bioinformatically seen in potential producer *Gardenia jasminoides*, were later verified by mass spectrometry, validating our transcriptome mining approach. This approach to bioinformatic mining has led to the identification of numerous potential molecules in both plants and fungi that will aid in the search for new RiPP natural products.

BIOINFORMATIC DISCOVERY OF NEW RIBOSOMALLY SYNTHESIZED AND POST-TRANSLATIONALLY MODIFIED PEPTIDES IN PLANTS AND FUNGI

by

Brigitte G. Ampolini

A Thesis
Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Greensboro

2023

Approved by

Dr. Jonathan Chekan
Committee Chair

DEDICATION

To my parents: Your unwavering love and support in the face of life-altering illness made this thesis, and the past two years, possible. Thank you.

"You never know what worse luck your bad luck has saved you from."

— Cormac McCarthy, No Country for Old Men

APPROVAL PAGE

This thesis written by Brigitte G. Ampolini has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair

Dr. Jonathan Chekan

Committee Members

Dr. Nicholas Oberlies

Dr. Ethan Taylor

June 6, 2023

Date of Acceptance by Committee

June 6, 2023

Date of Final Oral Examination

ACKNOWLEDGEMENTS

I am extremely grateful to my advisor, Dr. Jon Chekan for his flexibility, mentorship, and expertise. His kindness and friendship throughout this process was invaluable.

I would also like to thank Dr. Nicholas Oberlies for serving on my committee. I am thankful for his patience, sincere interest in my future and kind advice.

I am appreciative of Dr. Ethan Taylor for generously giving his time to serve on my committee.

Finally, I would like to thank the entire Chekan lab for creating a friendly, supportive environment and for their dedication to our research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

Natural products are molecules produced by an organism's secondary, nonessential metabolism.[1] Secondary metabolites are used to adapt to the environment the organism lives in and have functions ranging from defense against predators to signaling.[2] These molecules have evolved over millions of years to excel at their functions. Consequently, they possess unmatched chemical diversity, unique modifications, and a myriad of functions. Natural products have been used by different cultures in traditional medicine for thousands of years and have inspired over half of modern day drugs such as penicillin, codeine, Taxol, artemisinin and ivermectin.[2–4]

Ribosomally synthesized and post-translationally modified peptides (RiPPs) are a rapidly growing, diverse class of natural products. RiPPs offer extensive chemical diversity, bioactivity and potential for pharmaceutical development.[5] The genetic basis of precursor peptides and frequent proximity to tailoring enzymes that are also genetically encoded make RiPPs conducive to genome mining.[6] However, genome mining in eukaryotes poses unique challenges. Unlike prokaryotes, eukaryotic genes contain introns and biosynthetic genes are not always clustered together. As a result, far fewer RiPPs have been identified in eukaryotes and the space remains underexplored.

I.A. Background

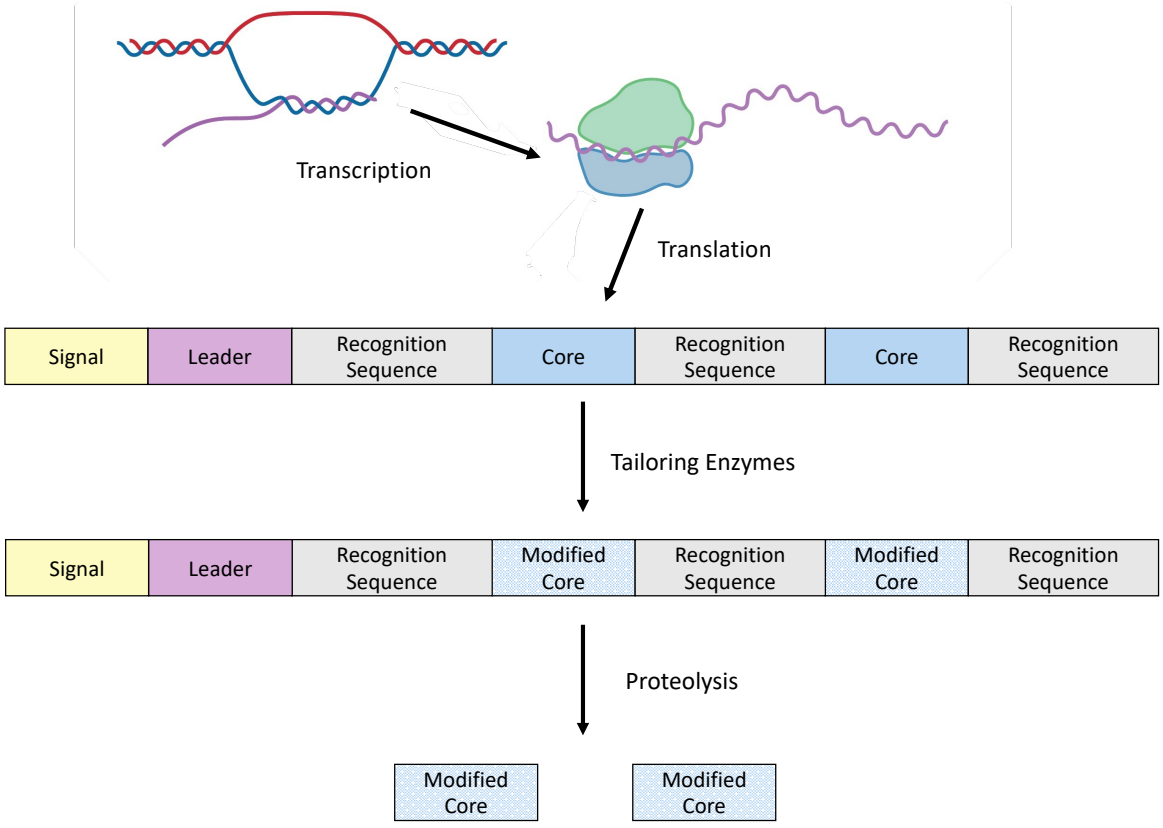
I.A.1 Ribosomally Synthesized and Post-translationally Modified Natural Products:

Ribosomally synthesized and post-translationally modified peptides (RiPPs) are a large, growing class of secondary metabolites. Their biosynthesis begins with transcription of the precursor peptide gene, followed by translation by the ribosome. The resulting linear precursor peptide undergoes post-translational modifications of the core peptide by tailoring enzymes and then proteolysis to ultimately yield the mature natural product (**Figure 1**).[6] Genes for these

tailoring enzymes are frequently, but not always, found nearby the gene for the precursor peptide.[5,6]

RiPP precursor peptides in eukaryotes are characterized by the presence of signal, leader and core sequences.[5] The signal sequence ensures the peptide is directed to the correct location in the cell. The leader is attached to the N-terminus of the core and is used for recognition by tailoring enzymes. Finally, the core sequence is the region that will undergo modification. Eukaryotes often utilize precursor peptides that have multiple core sequences (**Figure 1)**. In these cases, the cores are demarcated by recognition sequences that are used for proteolysis and recognition by the tailoring enzymes.[6]

Figure 1: RiPP Biosynthesis



After transcription and translation, the precursor peptide undergoes modification by tailoring enzymes before proteolysis excises the modified cores.

Eukaryotic RiPPs are understudied compared to their prokaryotic counterparts, but they have been identified in fungi and plants (**Figures 2** and **3**).

I.A.2 RiPP Classes in Fungi and Plants

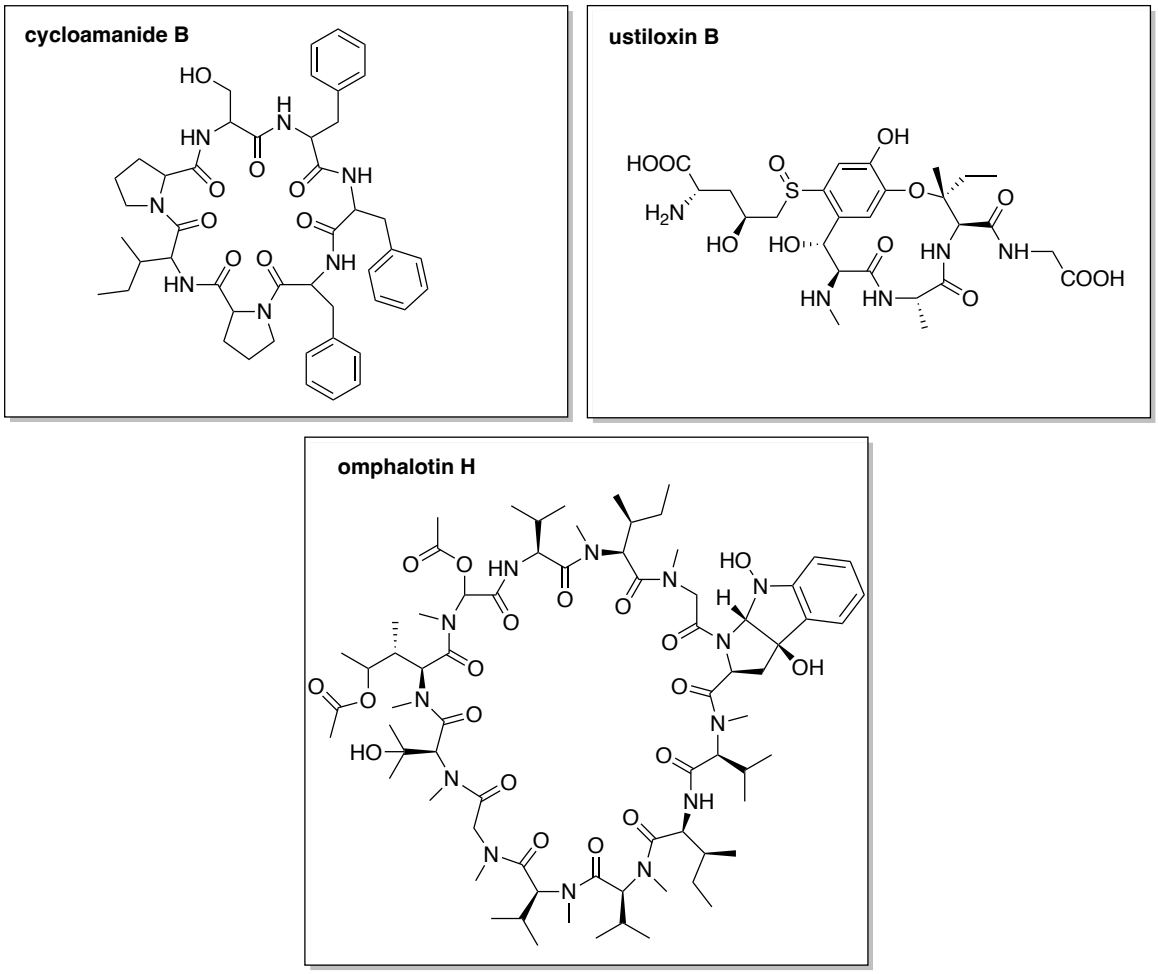
Fungi: Cycloamanides, borosins, & dikaritins

Cycloamanides: The founding family of fungal RiPPs, the cycloamanides (**Figure 2**), come from basidiomycetes. Formerly known as the MSDINs after the five N-terminal amino acids shared by its members, this family includes phallotoxins, amatoxins and virotoxins.[7] They are characterized by macrocyclization and proteolysis of the core by the enzyme prolyl oligopeptidase B (POPB).[7–9]

Borosins: The borosin family (**Figure 2**) also comes from basidiomycetes. These RiPPs are head-to-tail cyclized peptides with significant backbone methylation.[10,11] The methylation is carried out by a methyltransferase domain fused to the precursor peptide.[7,12,13]

Dikaritins: The dikaritin family (Figure 2) includes ustiloxins, phomopsins, asperipin-2a, victorins, and epichloëcyclins.[14–18] Dikaritins are the first example of RiPPs from ascomycetes. They are characterized by Kex-2 protease cleavage sites and DUF3328 catalyzed cyclization.[15]

Figure 2: Fungal RiPP Classes



Cycloamanide B is from the cycloamanide family, ustiloxin B from the dikaritins, and omphalotin H from the borosins.

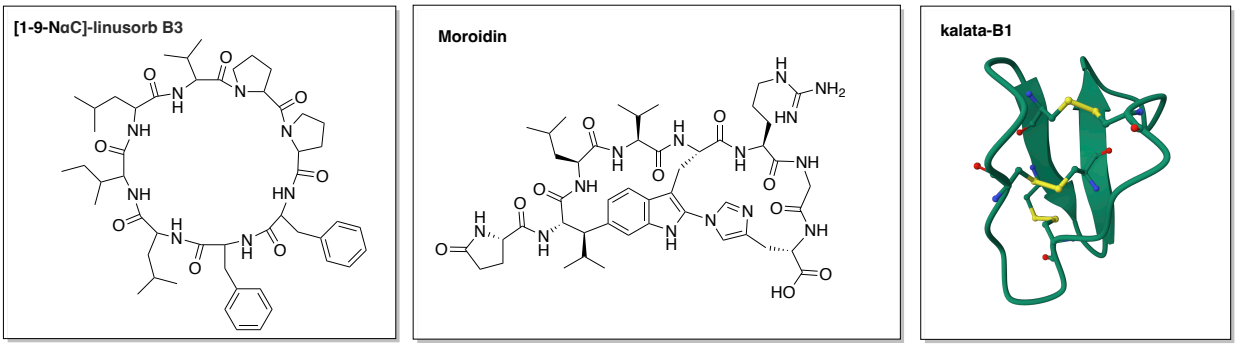
Plants: Cyclotides, orbitides & burptides

Cyclotides: Cyclotides are large, head-to-tail cyclized peptides with a distinctive cyclic-cysteine knot motif (**Figure 3**).[19,20] They are known for their insecticidal, antiviral, antimicrobial and cytotoxic bioactivities.[21] Cyclization takes place in the endoplasmic reticulum by asparaginyl endoprotease.[6,8]

Orbitides: Orbitides are small N-to-C cyclized peptides that lack disulfide bonds (**Figure 3**). These peptides range from 5 to 12 amino acids in length and are typically comprised of hydrophobic amino acids. They typically do not have any modifications other than cyclization.[22] Orbitides antimalarial, antiviral, antibacterial, and immunosuppressive bioactivities.[6,8,23]

Burptides: Burptides are macrocyclic rings with crosslinks between amino acid side chains (**Figure 3**). These peptides are post-translationally modified by BURP peptide cyclases.[24] Many have tyrosine or tryptophan as the C-terminal amino acid, some such as moroidins have a C-terminal histidine. The side chain cross links can be between aromatic side chains or aromatic side chains and the β-carbon of another amino acid in the chain.[25,26] Prior to this work, cyclopeptide alkaloids, hibispeptins and arabipeptins were not identified as burptides.

Figure 3: Plant RiPP classes



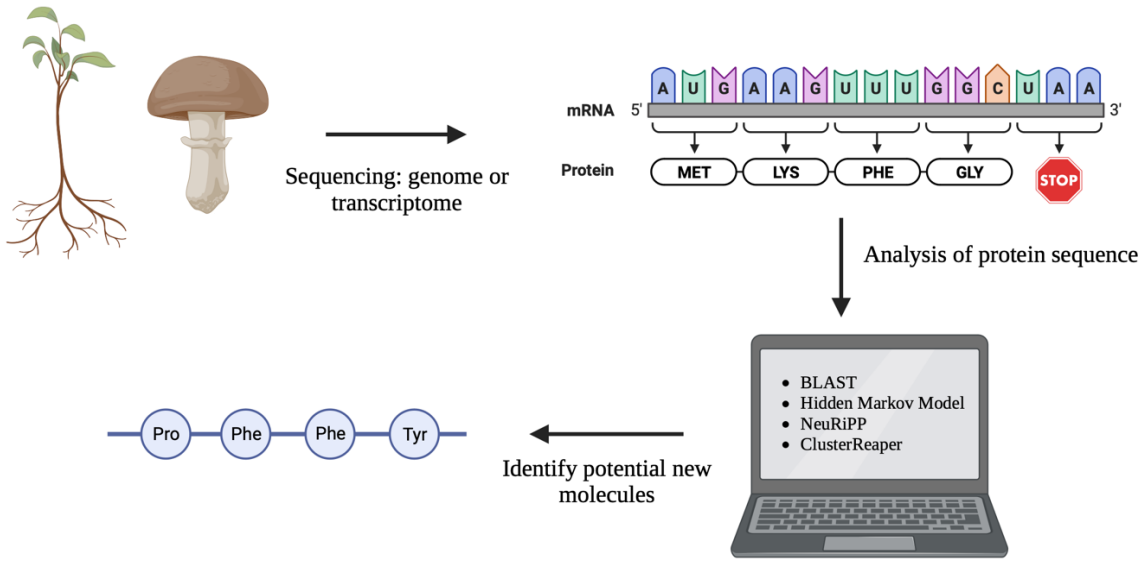
[1-9-NαC]-linusorb B3 is an example of an orbitide, consisting of hydrophobic amino acids and no modifications other than cyclization. Moroidin is a burptide. The crosslink occurs between tryptophan and the β- carbon of leucine. Kalata-B1 exhibits the cysteine knot motif seen in cyclotides. (Image taken from https://doi.org/10.2210/pdb1NB1/pdb)

I.B. Basis for Study

I.B.1 Genome Mining

Genome mining is the bioinformatic process of screening an organism's genome for new natural products or biosynthetic pathways (**Figure 4**). The genomic origin of RiPPs makes them well suited to this approach. RiPP biosynthetic logic relies on conserved leader, core and recognition sequences, as well as tailoring enzymes that are genetically encoded. The sequence of any one of these features can be used to search for homologous sequences in other genomes. Three tools were used for the bioinformatic studies: BLAST, hidden Markov models and sequence similarity networks.

Figure 4: Genome mining strategy



Sequencing data can be analyzed with a variety of software in order to identify potential new molecule candidates. (Created with BioRender.com)

BLAST: The Basic local alignment search tool (BLAST) is a web driven search platform that takes a protein or nucleotide query sequence and compares it to a database of subject sequences in order to find homologous sequences.[27] BLAST searches its database of subject

sequences using three amino acid long "words" from the query sequence. When a match is found, the alignment is extended forward and backward by two additional "words" as long as the alignment score increases or until a critical drop-off value is reached.[28] Alignment scores are assigned to each letter of the query sequence as it is aligned with a letter in the subject sequence. These scores are then summed over the length of the alignment.[29] To address amino acid substitutions and gaps, BLAST creates a matrix that contains scores for all possible amino acids at each position. If the substitution is likely, it receives a positive score. If it is unlikely, it receives a negative score.[27–29]

Hidden Markov model: Profile hidden Markov models (HMMs) take multiple sequence alignments and create a position specific scoring system that can be used to query other datasets for homologous sequences.[30] Match, insert, and delete states are used to account for sequence variability and deletions. Match and insert states have 20 emission probabilities, one for each amino acid. Delete states have no emission probabilities. This allows the finished profile to model both the frequency of an amino acid occurring at each position and the transition state between amino acids.[30,31] Once made, the model can be run against large numbers of sequences to rapidly identify homologs. Hidden Markov models can be made and used to search sequence databases with the software HMMER.[32]

Sequence Similarity Network: Sequence similarity networks (SSNs) are an efficient way of visually sorting a dataset of proteins into clusters of related sequences. Each protein sequence is represented as a node in the network. All-by-all BLAST is used to generate edge values, and an edge is drawn between two nodes if the BLAST pairwise similarity score is above the defined threshold.[33] The resulting network visually shows clusters of nodes that contain homologous protein sequences.

I.C. Significance

Many useful small molecules and drugs are derived from natural products. RiPPs represent a vast, underexplored wealth of potential new natural products with diverse cyclizations and modifications. Their genetic basis is advantageous in the search for new molecules. Instead of needing to blindly isolate a new natural product from its source, its presence can be predicted by looking at the organism's genome and can then be validated by isolation or heterologous production for structural elucidation. This genome mining project will help discover new RiPP natural product leads in eukaryotes, as well as provide insight into their biosynthesis.

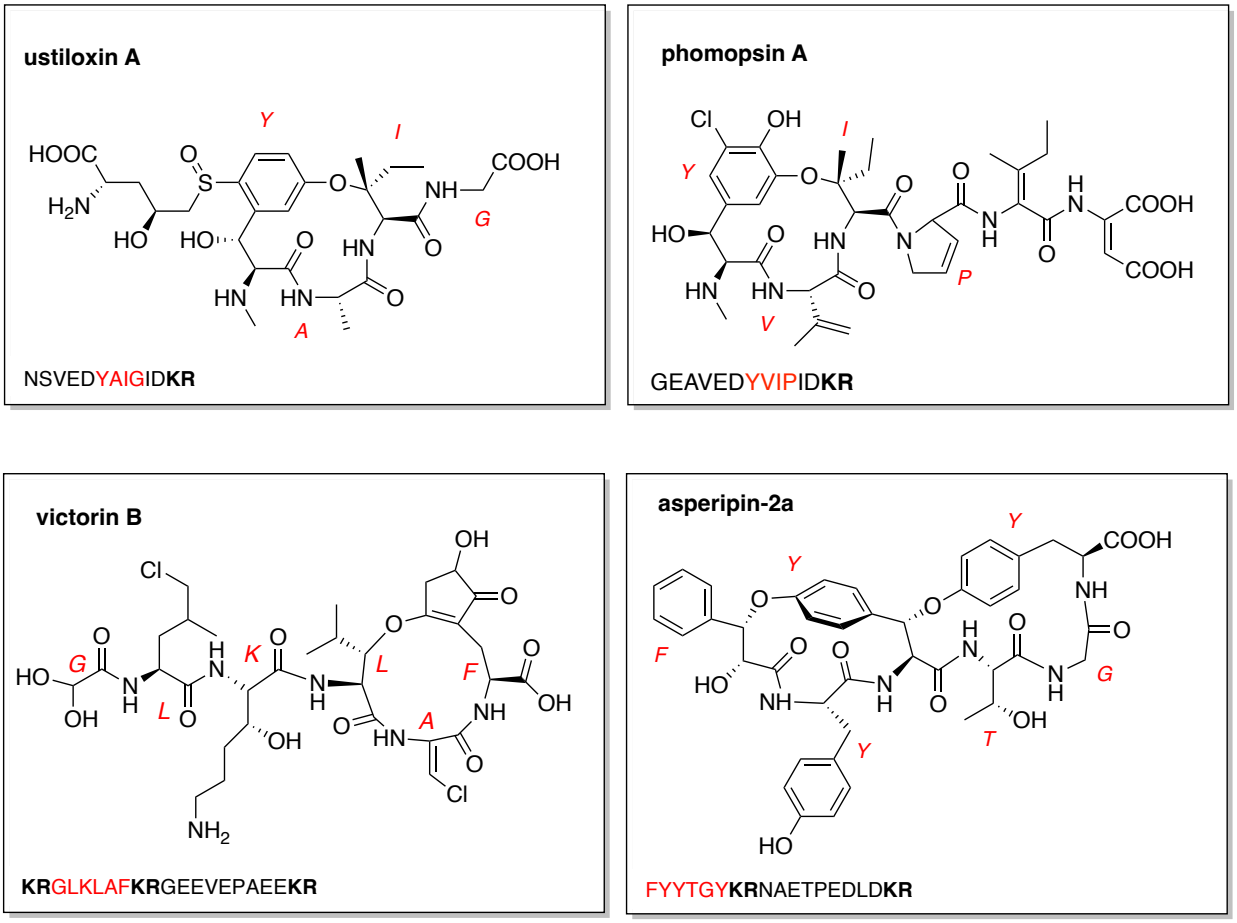
CHAPTER II: GENOME MINING IN FUNGI

II.A. Approach

There are currently three families of fungal RiPPs: dikaritins, cycloamanides, and borosins. Of these, dikaritins are particularly interesting due to their diverse modifications and proximity to the tailoring enzyme. We chose to target the dikaritins for genome mining. The family of dikaritins includes ustiloxins, phomopsins, asperipin-2a, victorins, and epichloëcyclins (**Figure 5**).[7,14–18] These natural products are derived from precursor peptides that have an N-terminal signal peptide followed by repeats containing the core sequences. These repeats are separated by recognition sequences and at least one Kex-2 protease recognition site consisting of the amino acid pairs KK, RR or KR.[7,34]

Biosynthetically, a hallmark of dikaritins is the presence of DUF3328 enzymes in close proximity to precursor peptides.[26] The function of DUF3328 has yet to be clearly elucidated, but it is hypothesized to play a role in a diverse range of modifications including ether bond cyclizations and chlorinations.[7,14,16,35] Because of its hypothesized role in the modifications seen in this RiPP class and its close genomic proximity to precursor peptides, DUF3328 guided genome mining is an appealing method for searching for new fungal RiPPs. Using a combination of hidden Markov models and sequence similarity network bioinformatic tools, all deposited fungal identical protein groups were searched for the presence of new dikaritins.

Figure 5: Examples of known dikaritins



Core sequences are shown in red. Kex-2 protease sites are bolded. Chlorination is seen in victorin B and phomopsin A. Ether linkages are seen in all.

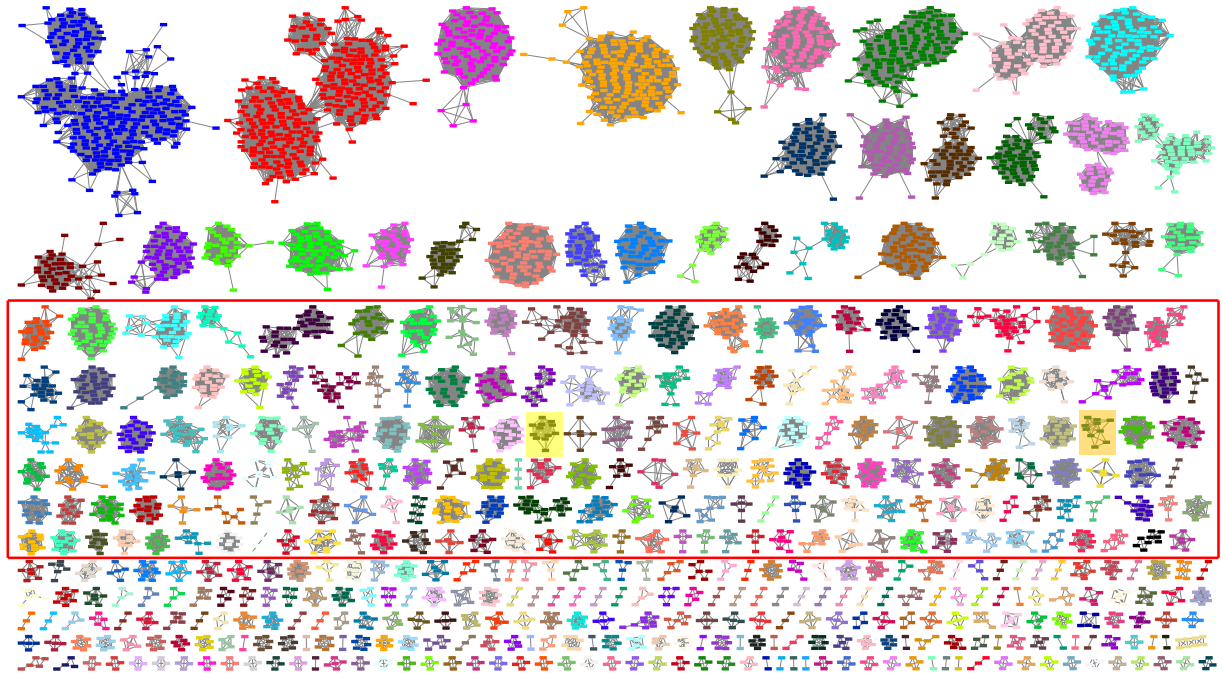
To identify these new dikaritins, all identical fungal protein groups were downloaded from the National Center for Biotechnology Information (NCBI) public database. The HMMER function *hmmsearch* was used to evaluate these protein groups with the existing hidden Markov model for the DUF3328 protein family.[36] The resulting hits were used to make a sequence similarity network with the Enzyme Function Initiative sequence similarity tool and visualized in Cystoscape.[33] Each of the large clusters was run through ClusterReaper, a tool developed by the Chekan lab for easy identification and exploration of biosynthetic gene clusters. Sequences from

precursor peptide core regions, if present, were extracted for each node in the cluster and compared to known fungal RiPP cores. Due to a scarcity of research, the exact core sequences of these putative dikaritins are unclear.

II.B. Results

Out of the 15,692,220 total proteins downloaded from the fungal identical protein groups NCBI database, 11,152 scored above the default HMMER inclusion threshold for being a member of the DUF3328 protein family. Further analysis of these putative DUF3328 family members using an SSN clustered them into 1,424 isofunctional groups, of which 136 were evaluated (**Figure 6)**. Possible precursor peptides were identified in 77 of these clusters.

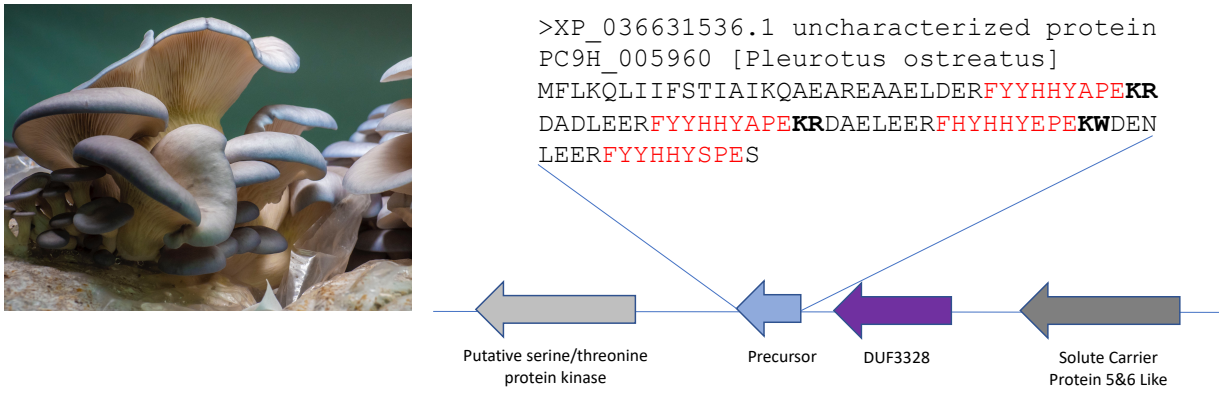
Figure 6: Sequence similarity network generated by DUF3328 guided genome mining



Clusters investigated are boxed in red. Cluster containing a putative precursor peptide from *Pleurotus ostreatus* is highlighted in yellow. Cluster containing a putative precursor peptide from *Aspergillus nomiae* is highlighted in orange.

Of interest, a potential precursor peptide was noted in *Pleurotus ostreatus*, commonly known as oyster mushrooms (**Figure 7**). Oyster mushrooms are basidiomycetes and there are currently no examples of dikaritins in this fungal subkingdom. These commonly eaten mushrooms appear to have a precursor peptide containing Kex-2 cleave sites (KR/KW) and conserved recognition sequences between potential core regions. This putative precursor peptide is in close proximity to a DUF3328, suggesting it may be modified into a dikaritin product.

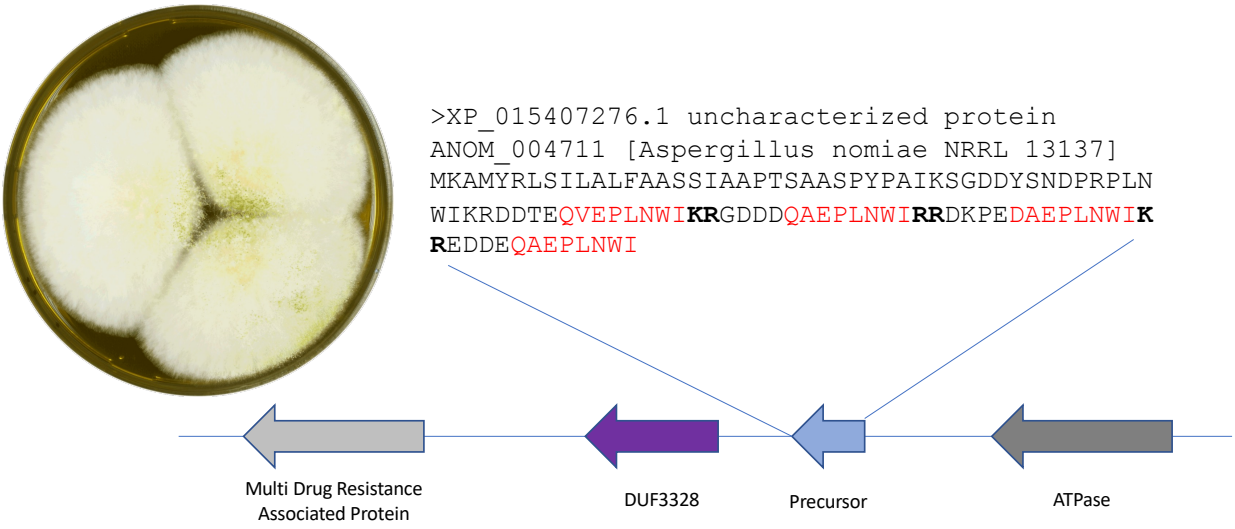
Figure 7: Putative gene cluster for an uncharacterized RiPP from the widely eaten mushroom *Pleurotus ostreatus*



Potential core regions are red, Kex-2 protease sites are bold. (Image by Preston Keres. https://tinyurl.com/yzh7vvcf)

Another example of a potentially interesting precursor peptide is found in *Aspergillus nomiae* NRRL 13137. The precursor has KR and RR cleavage sites, a relatively short recognition sequence and a DUF3328 is nearby (**Figure 8**). While the exact core sequence is unknown, of the sequence of the putative core region does not match any of the known dikaritins. This hit is advantageous because it comes from an NRRL strain. This simplifies future analysis as it can be ordered directly from the ARS Culture Collection Database.

Figure 8: Putative gene cluster for an uncharacterized RiPP from *Aspergillus nomiae* NRRL 13137



Potential core regions are red, Kex-2 protease sites are bold. (Image taken from https://commons.wikimedia.org/w/index.php?title=File:Aspergillus_nomius_meaox.png&oldid=637708472)

II.B. Conclusion

Numerous putative dikaritin precursor peptides were identified using a DUF3328 guided genome mining strategy. These dikaritin precursors were seen in both Dikarya subkingdoms, Ascomycota and Basidiomycota. Basidiomycetes were not previously known to produce dikaritins. Of interest, novel precursor peptides were bioinformatically seen in *Pleurotus ostreatus* as well as *Aspergillus nomiae.*

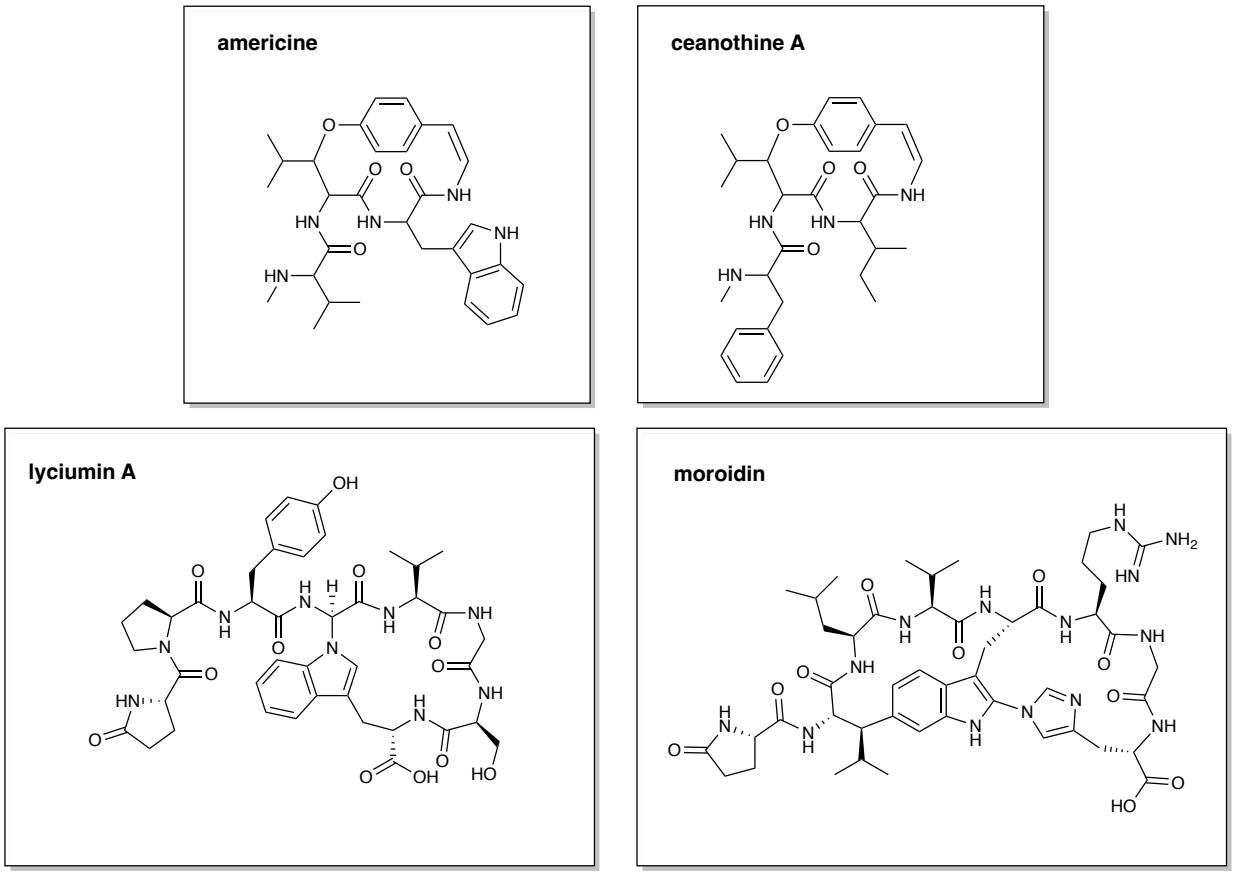
CHAPTER III: GENOME MINING IN PLANTS

III.A. Approach

Cyclopeptide alkaloids are cyclic peptides produced by many plant families, particularly the Rhamnaceae family.[37] They are macrocyclic rings composed of 4 or 5 amino acids, characterized by an ether linkage between tyrosine and the β-carbon the neighboring amino acid (**Figure 9**). This tyrosine undergoes decarboxylation and desaturation to form a styrylamine moiety.[37,38] There are currently over 200 known cyclopeptide alkaloids, with over 100 coming from the *Ziziphus* genus.[37] Plants that produce these peptides have been used in traditional medicine for centuries. A variety of bioactivities have been shown including sedative, analgesic, antibacterial, antifungal and antidiabetic properties.[37,38]

The Chekan lab was investigating the biosynthetic basis of cyclopeptide alkaloids using *Ceanothus americanus,* a well-known producer of cyclopeptide alkaloids (**Figure 9**).[22] Transcriptomic data was used to link precursor peptides to their cyclopeptide alkaloid products. The ether linkage in cyclopeptide alkaloids shares some structural similarities with burptides, suggesting they may share biosynthetic commonalities. Previous research had shown BURP peptide cyclases are responsible for the cyclization seen in some burptides such as moroidins and lyciumins **(Figure 9)**.[25,39] For this reason, it was hypothesized that a BURP peptide cyclase may be responsible for installing the side-chain crosslink in cyclopeptide alkaloids

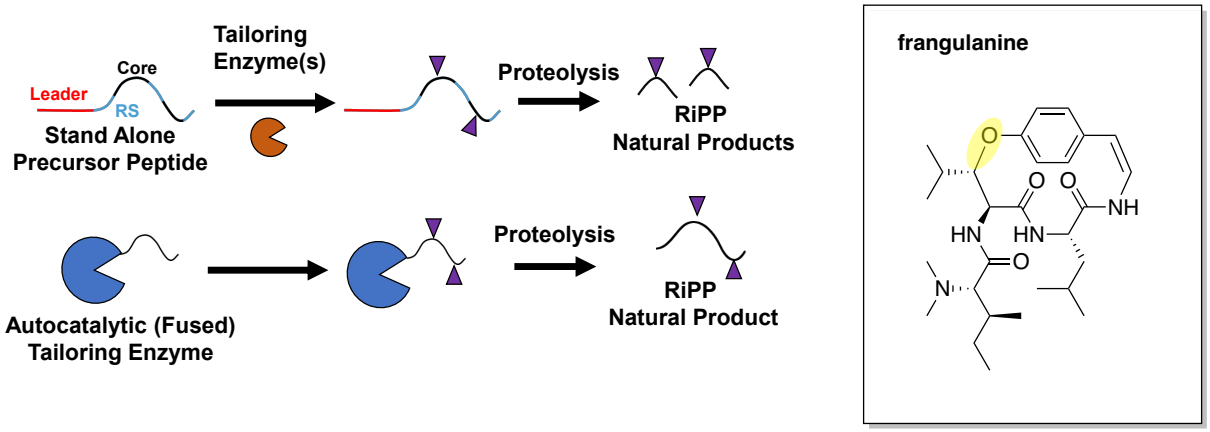
Figure 9: Cyclopeptide alkaloids



Cyclopeptide alkaloids from C. americanus and the burptides moroidin and lyciumin A. Characteristic side chain cross links are seen in all structures.

BURP peptide cyclases are copper dependent enzymes that carry out the macrocyclization of burptides through a crosslink between amino acid side chains.[37,39] In place of an autocatalytic "fused" system, a standalone precursor peptide was identified. Separately, BURP peptide cyclases were found nearby in a "split" biosynthetic system (**Figure 10**).

Figure 10: Split (top) and fused (bottom) burptide biosynthetic routes



In a split system, the precursor peptide and the tailoring enzyme are transcribed and translated as independent peptides. In a fused system, the precursor peptide is transcribed and translated with the tailoring enzyme as one peptide. The tailoring enzyme is ultimately cleaved from the modified peptide during proteolysis.[24] Right: Frangulanine, a cyclopeptide alkaloid from *C. americanus* derived from a split BURP system, with the ether linkage highlighted.

The predicted precursor peptide from *C. americanus* was used to search the Viridiplantae NCBI identical protein groups for homologs. This was accomplished using BLAST to search for additional precursor peptides similar in sequence to those from *C. americanus*. Using these, a custom hidden Markov model was built that searches for burptide precursor peptides across all plants. A sequence similarity network using these precursor peptides are generated and manually annotated for the presence of core sequences and known molecules. Finally, the species represented in the HMM results were used to determine their distribution in the Viridiplantae by building a cladogram. Ultimately, this bioinformatic approach enabled the identification of new molecules, the biosynthetic system used to make them, and its phylogenetic distribution.

III.B. Results

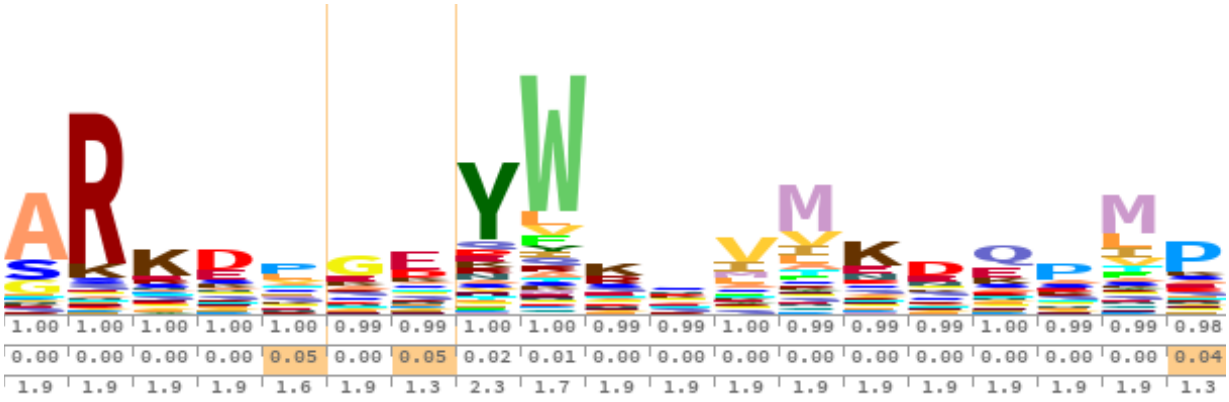
To evaluate if similar standalone precursor peptides were found in other plants, BLAST was used to search the available *Ziziphus jujuba* genomes for homologs of the *C. americanus* precursor. *Ziziphus jujuba* was chosen because it is a known producer of numerous cyclopeptide alkaloids.[40]

Core sequences of 15 known molecules, including jubanines F, G, H, I, J, and nummularine B could be mapped back to genes in the initial BLAST results from *Z. jujuba* (**Appendix A**).[40] Additionally, BURP peptide cyclases were found in close genomic proximity to 30 of 35 *Ziziphus* precursor peptides. These results suggested that a split biosynthetic route may be involved in cyclopeptide alkaloid biosynthesis, consistent with the observation of the standalone *C. americanus* precursor.

Next, a PSI-BLAST search with the *C. americanus* precursor peptide was carried out to create a set of similar sequences that could be used to search larger datasets for the presence of related precursor peptides with the machine learning algorithm NeuRiPP. NeuRiPP was chosen because it can be trained to identify RiPPs of any class in large datasets.[41] It requires a positive dataset of sequences with manually confirmed precursor peptides, and a negative "decoy" dataset of sequences that do not contain precursor peptides. The program builds a model and runs iterative training cycles until model accuracy is no longer improving. BLAST results were used for the positive dataset, and a list of monocots was used as the decoy set. Unfortunately, NeuRiPP was unable to consistently identify precursor peptides. It gave numerous false positives and false negatives. Adjusting the training weights, increasing the number of positive precursor sequences, and decreasing the number of decoys did not improve the quality of the results.

The list of positive precursor peptides was then used to construct a custom hidden Markov model that would look for these plant side-chain crosslinked cyclopeptides. The HMMER *hmmbuild* utility was employed for its creation.[36] The *hmmsearch* function was used to examine the 17,867,506 identical protein groups of the Viridiplantae clade deposited on NCBI with the new model.

Figure 11: Hidden Markov model visualization using Skylign



This conserved Ax_6YWx_7PMP motif is found in both autocatalytic BURP peptide cyclases and split BURP systems.

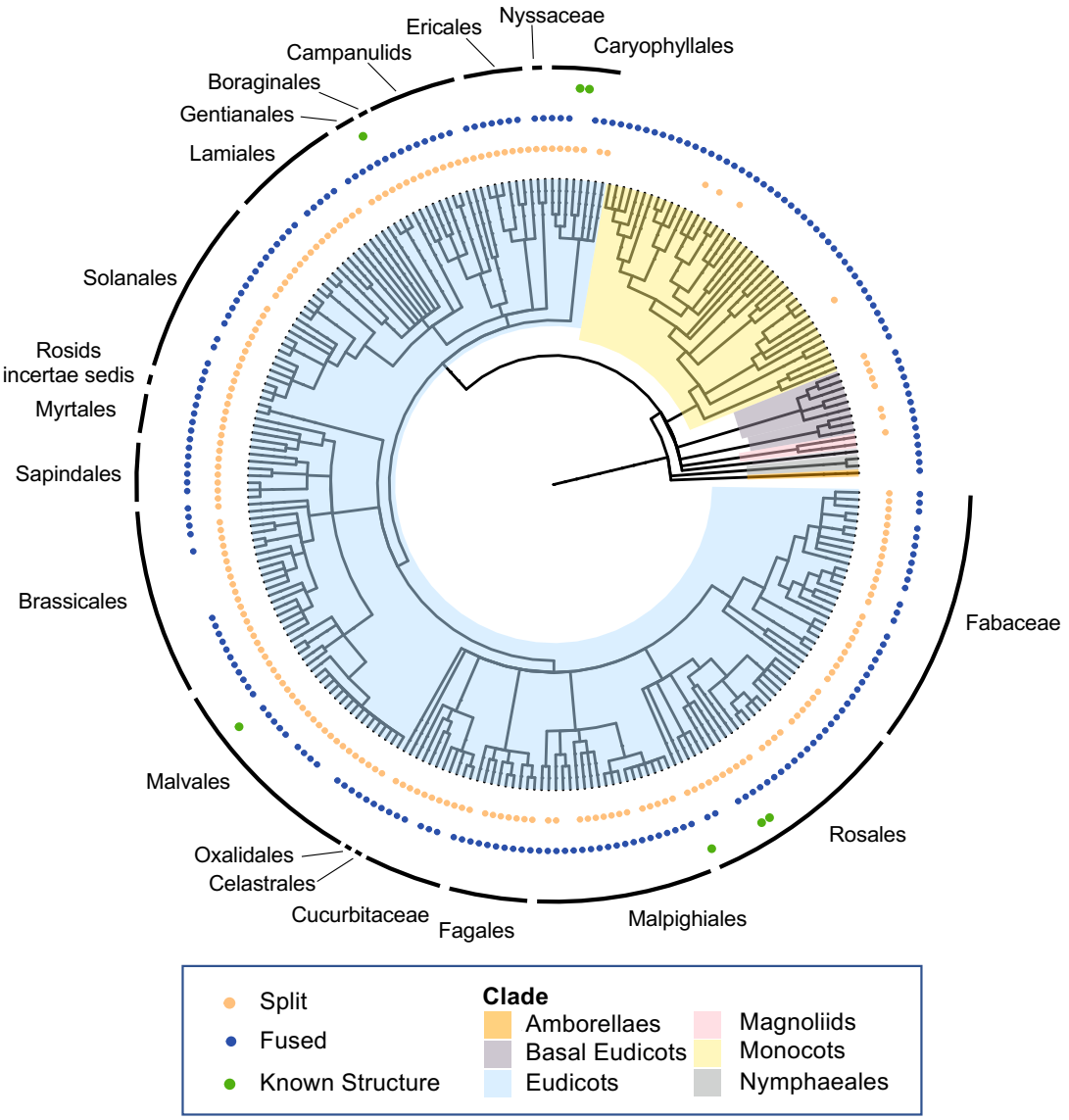
The results unexpectedly contained a significant number of hits annotated as BURP-domain containing proteins. This is likely due to the model detecting a conserved Ax_6YWx_7PMP motif in both the autocatalytic BURP peptide cyclases and split precursor peptides (**Figure 11**). To address this, the existing HMM for the BURP protein family was downloaded and run with the Viridiplantae identical protein groups.[42] These results were then compared to the results from the custom HMM and used to create lists of "fused" and "split" BURP peptide cyclases. The Hidden Markov models identified 1,423 split and 1,099 fused BURP cyclase systems.

Circular and linear cladograms were constructed from all species represented in the split and fused HMM results using the website phyloT.[43] The trees were annotated with clade,

presence of a known molecule, presence of a fused BURP peptide cyclase and presence of a split system in R with the ggtree and ggtreeExtra packages (**Figure 12**).[44,45]

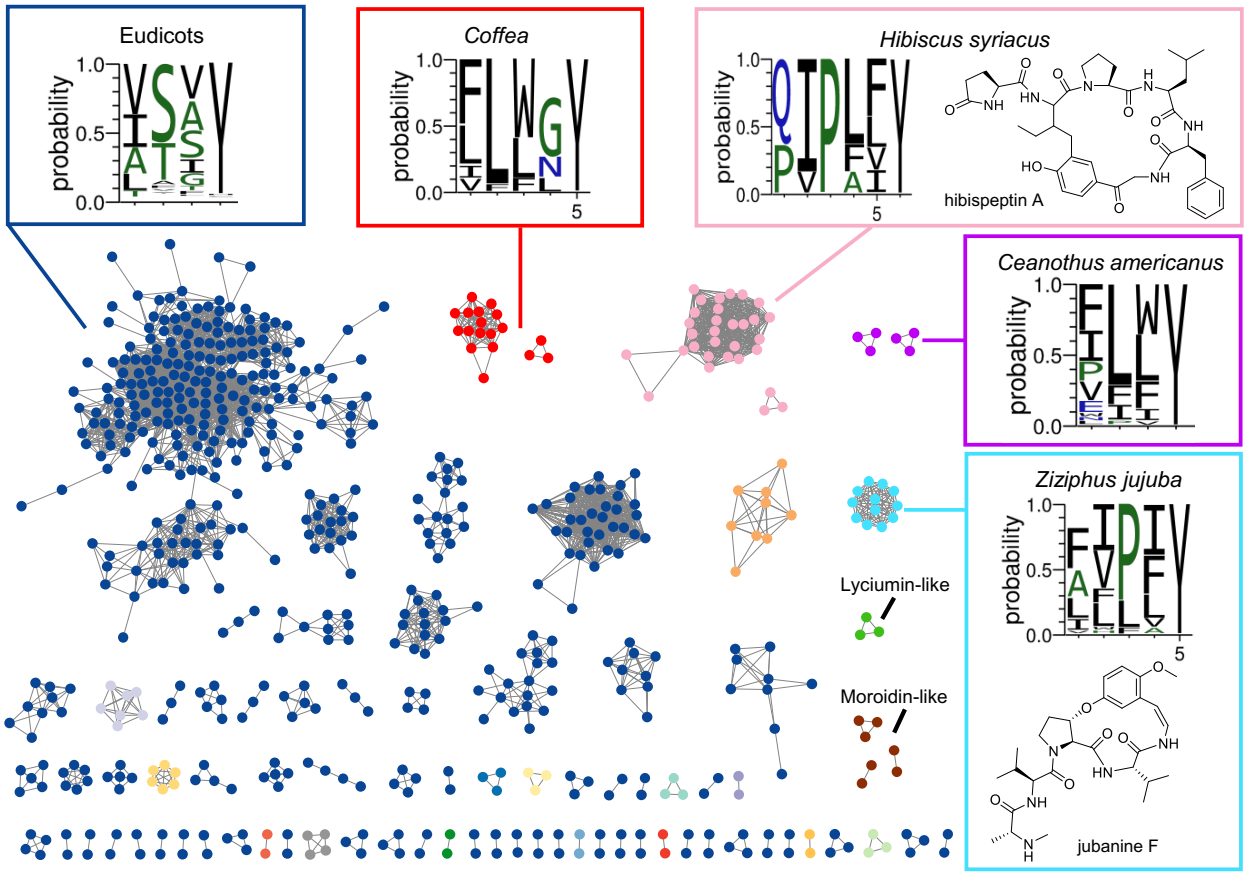
Percent abundance of the split and fused biosynthetic routes was calculated using annotated genomes available for eudicots and monocots. 89.1% of the 202 annotated eudicot genomes contain precursor peptides from split BURP systems. 79.2% contained fused BURP peptide cyclases as identified by our HMM. Fused BURP peptide cyclases were more prevalent in monocots with 83.7% of the 49 unique monocot species containing at least one. Split BURP cyclase systems were present in 32.7% of annotated monocot genomes (**Figure 12**).

Figure 12: Cladogram of fused and split burptides identified by the custom HMM



A sequence similarity network of the split precursor peptides above the burptide HMM inclusion threshold was made using the Enzyme Function Initiative Enzyme similarity tool (**Figure 13**).[33] The clusters were then manually annotated for core sequence and presence of known molecules.

Figure 13: Sequence similarity network of standalone precursor peptides using an alignment score of 70



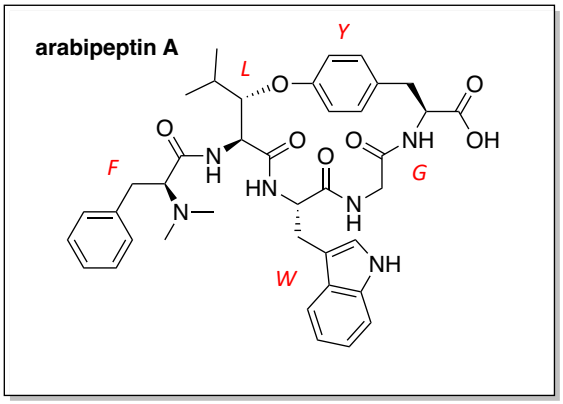
Precursor peptides with cores containing serine or threonine in the second position are represented in navy blue. Clusters from *H. syriacus* (pink), *Z. jujuba* (light blue), *Coffea* (red), and *C. americanus* (purple) are indicated along with those with lyciumin-like (green) and moroidin-like (brown) core sequences. Amino acid probabilities of the core sequences for these clusters are depicted.

The sequence similarity network of the split precursor peptides helped establish the presence of a handful of known molecules including hibispeptins A&B, jubanines F-G, moroidins and lyciumins (**Figure 13**). Numerous precursor peptides with cores not matching known molecules were noted in *Coffea arabica*. Ultimately one of these went on to be isolated,

purified and named arabipeptin A by Chekan lab member Dr. Stella de Lima Camargo (**Figure 14**). This demonstrated that the bioinformatically identified precursors not only explained known molecules but were predictive of new natural products.

Figure 14: arabipeptin A

>XP_027065604.1 uncharacterized protein LOC113691593 isoform X3 [Coffea arabica]
MASSITLIAVFSIALFACITEARKNPTDFLQSAVINEHTEDNH
HAESSLSNQKKTSNGNTLKDFESKPGS**FLWGY**QGNDAESKSKE
EKPLMKGFESKPGS**FLWGY**QGNDAESKSKEEKPLMKGFESKPG
S**FLWGY**QGNDVESKSKEEKPLTKDFESKPGS**FLWGY**QGNHAES
KSKKEKPLMKDFESKPGS**FLWGY**QGNHAEYKEKKPLVKDN



Bioinformatically discovered gene for arabipeptin A and its structure determined after isolation and structure elucidation.[24]

III.C. Conclusion

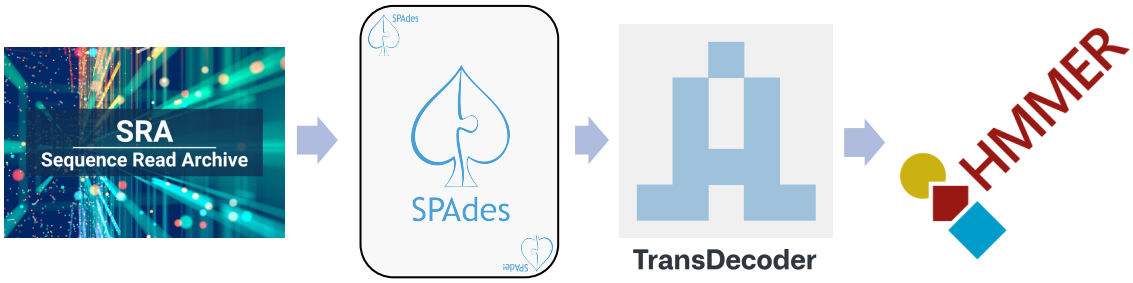
The Chekan lab identified a standalone cyclopeptide alkaloid precursor peptide from *Ceanothus americanus.* This precursor peptide was used to bioinformatically search for related standalone precursor peptides in the Viridiplantae family. BURP peptide cyclases were noted nearby many of these precursors, suggesting a split biosynthetic system was responsible for installing modifications seen in the mature natural product. Previously, BURP peptide cyclases were known as autocatalytic enzymes.[24,26,39] This split biosynthetic system is widely distributed among angiosperms. One of the bioinformatically identified precursor peptides from *Coffea arabica* was isolated and named arabipeptin A, validating our genome mining approach.

CHAPTER IV: TRANSCRIPTOME MINING PIPELINE

IV.A. Approach

Previously, only species with fully annotated genomes deposited to the NCBI were explored. However, raw sequencing data is frequently deposited into the publicly available Sequence Read Archive (SRA) and never submitted in an assembled and annotated form. This is required by most journals prior to publishing. Consequently, there is an abundance of unassembled RNA data for plants that can be searched with a custom HMM after assembly. This is accomplished by creating a pipeline that automatically downloads raw sequencing data and identifies searchable coding regions within the transcriptomes. This opens the search for new burptides to a large, unexplored set of data.

Figure 15: Transcriptome assembly pipeline



To generate an accurate, and fast pipeline for the de novo assembly of transcriptomes, it is important to select software that is both efficient and robust (**Figure 15**). The assembly software SPAdes was chosen for these reasons. SPAdes, short for St. Petersburg Genome Assembler was developed for de novo assembly of single cell sequencing data from small genomes. It contains pipelines that can assemble metagenomes, plasmids, RNA-Seq data, biosynthetic gene clusters, and transcriptomes.[46] SPAdes is appropriate due to its ability to assemble transcriptomes from organisms without well annotated reference genomes. SRA files sourced from RNA, with paired

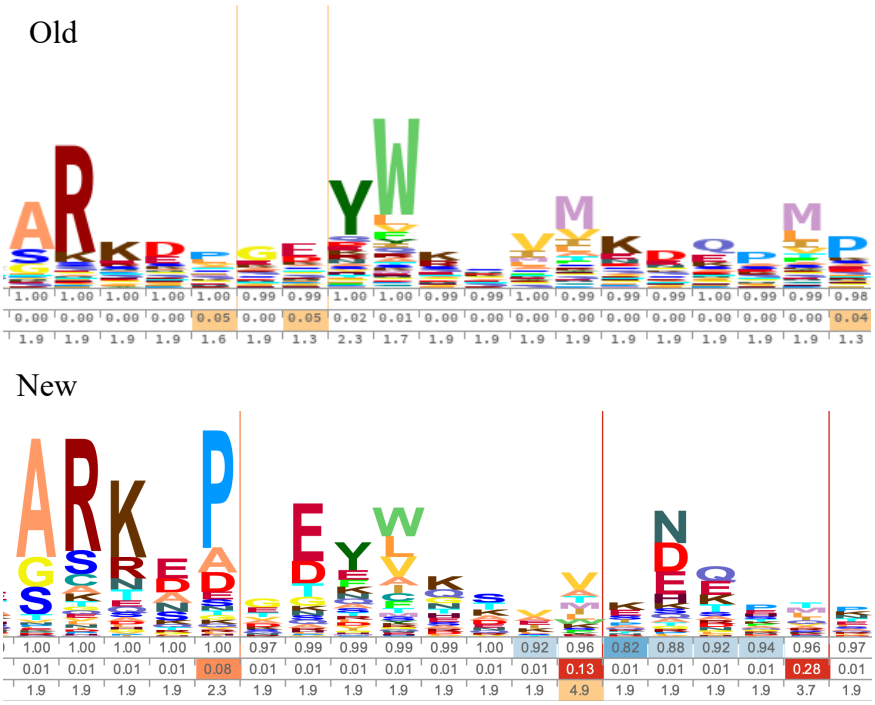
end library layouts and from Illumina sequencing were used for this project and subsequently assembled with SPAdes.

Coding regions of the assembled transcriptomes were identified using TransDecoder. TransDecoder predicts open reading frames (ORFs) within the transcript and scores them based on nucleotide composition, length and Pfam domain content.[47] The longest ORFs are reported back in FASTA format. In this transcriptome assembly pipeline, the script (**Appendix B**) developed automatically downloads user selected SRA files, assembles the transcriptomes with SPAdes and uses TransDecoder to identify coding regions (**Figure 15**).

Species with a high likelihood of producing burptides based on existing data of cyclopeptide alkaloid producers were prioritized for analysis. Furthermore, every order of the current angiosperm phylogeny with transcriptomes deposited in the SRA was represented in the analysis, with the addition of the gymnosperm order cycads.[48]

This list of SRAs was input into the script (**Appendix B**) and the results were analyzed with an updated custom hidden Markov model. The updated version of the HMM was biased towards small molecule discovery instead of searching for fused burptides and cores with no known corresponding subfamily (**Figure 16)**. This was accomplished by removing sequences such as xSxY and xAxY, represented in the "eudicot" Weblogo in **Figure 13** while adding sequences coding for known molecules such as hibispeptins, moroidins, jubanines, and lyciumins.

Figure 16: Old and new hidden Markov model visualization



The updated HMM (bottom) showed a highly conserved proline not seen in the prior model (top). Additionally, the Ax_6YWx_7PMP motif was less conserved.

IV.B. Results

442 accessions meeting the criteria for assembly by the SPAdes were selected for analysis. Of those, transcriptomes for 383 unique species were successfully assembled. These species represented 62 angiosperm orders and one order of gymnosperms. Three angiosperm orders, Dilleniales, Alismatales, and Picramniales were not represented in this analysis owing to them having no transcriptomic data deposited to the Sequence Read Archive. Analysis of the hidden Markov model results for each of these transcriptomes showed 67 species were potential producers. Species were labeled as potential producers if multiple repeats of a core sequence and a recognition sequence were seen in the HMM results. Cores with xSxY and xAxY sequences were not considered.

Potential producers were widely distributed throughout the angiosperm phylogenetic orders (**Figure 17**). The order Gentianales is home to the Rubiaceae and Rhamnaceae families and known burptide producers *Ceanothus americanus* and *Coffea arabica*.[24] Transcriptomic analysis revealed twelve new potential producers in this order. In the order Malvales where *Hibiscus syriacus* is found, six potential producers were identified.

Celosia argenta and *Amaranthus cruentus* are known moroidin producers.[24] Both of these species are from the order Caryophyllales. Moroidin and moroidin-like cores were seen in two new species of plants in the order Caryophyllales. Moreover, they were found in three species from the Icacinales, and two from the Lamiales. There was significant diversity in these core sequences, with most not matching any known moroidins (**Table 1**).[26]

Table 1: Moroidin-like producers

Accession	Species	Order	Potential Cores
SRR12006303	*Merrilliodendron megacarpum*	Icacinales	QLLVWKTH, QLLLWREH, KLLLWREH, QLQLWREH, QLLLCREH, QLLLGREH, HLLLWREH, QLLVWREH, QLKLLREH, QLLLWRQH, QLLLWHEH, QLLLLREH, QLLLWTDH, QLLLWREQ, QLLLWREL, QLLIWLIH, QLLVWKTH, QLLLWLEH
SRR11994211	*Pyrenacantha malvifolia*	Icacinales	QLLLWREH, QLLVWREH
SRR11994210	*Mappia racemosa*	Icacinales	PSYNY, QLLLWREH
SRR21095983	*Krascheninnikovia arborescens*	Caryophyllales	QLLVWRGH, QLFVWRNH, QLRVWLEH, QLLVSDAL
SRR13316945	*Atriplex imbricata*	Caryophyllales	PVLFWWQ, PNQVLYW, QLLVWRQG
SRR7806556	*Justicia pacifica*	Lamiales	QNRLAYH
SRR7848077	*Justicia gendarussa*	Lamiales	QLLVWRRH

Seven species produced moroidin-like compounds. Phylogenetic order and potential cores are shown.

This transcriptome mining approach was validated by Michael Pasquale in the Chekan lab when molecules matching core sequences FFFY and ILLY seen in *Gardenia jasminodies* (**Table 2**) were verified by mass spectrometry.

Table 2: Core sequences from *Gardenia jasminoides*

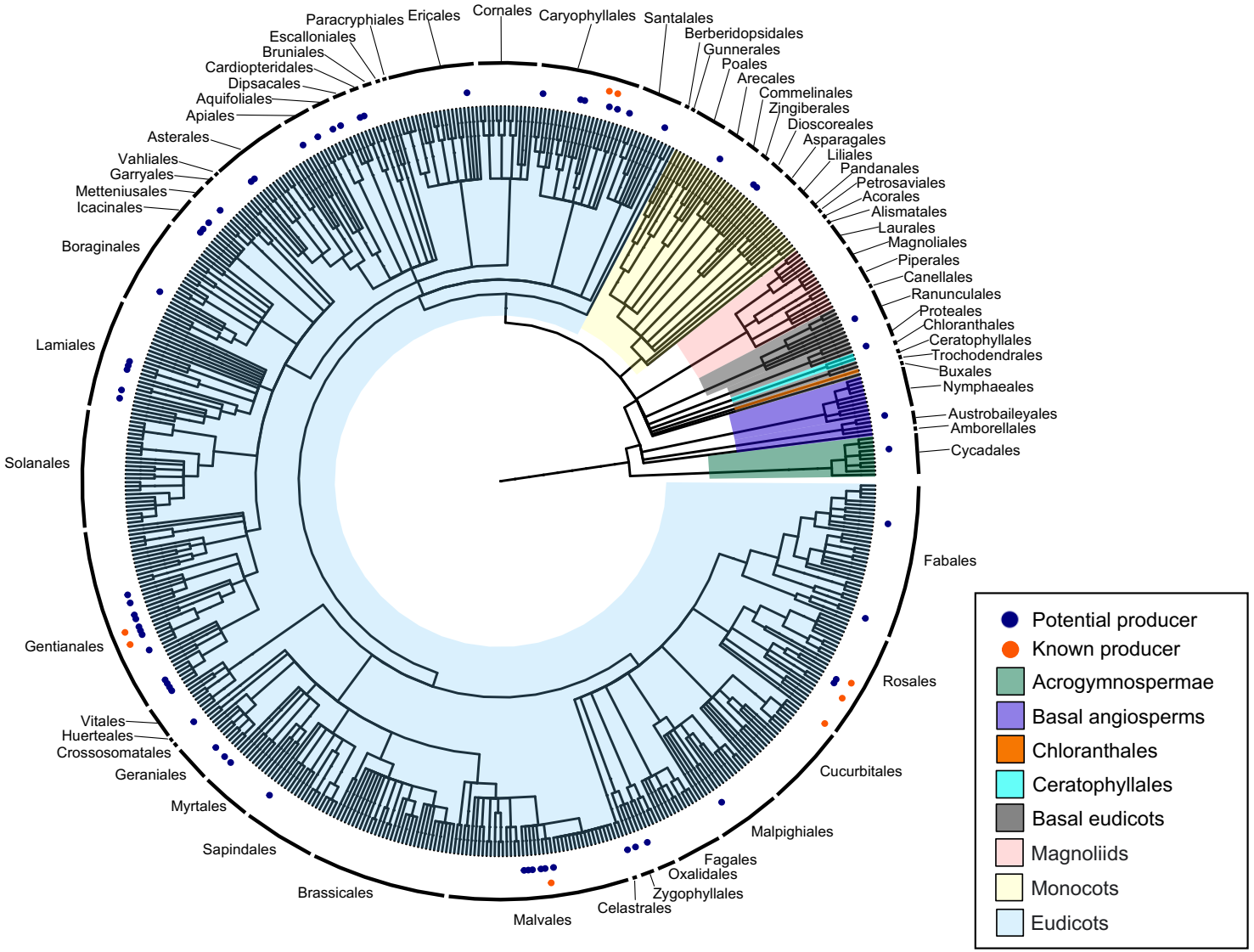
Species	Order (Family)	Potential Cores
***Gardenia jasminoides* (SRR19137756)**	Gentianales (Rubiaceae)	ILLY, FFFY, VLLY, PLFY, SLFY, LDIY, IFPY, HRRY, FFFF, FLIY, FLFY, FFIY, LYIY, IFLY, SVFY, FQPY, LFPY, LQKY, AVRY, AVRD

Twenty potential core sequences were seen bioinformatically. To date, ILLY and FFFY core sequences have been verified.

IV.C. Conclusion

Our search for new burptides had previously been confined to organisms with annotated genomes deposited to NCBI. In order to explore the abundance of raw transcriptomic data available in the NCBI Sequence Read Archive, a script was written that downloads, assembles and analyzes raw transcriptomic data. This was accomplished using the software SPAdes, TransDecoder and an updated version of our custom burptide HMM. This transcriptome mining approach identified 67 species that are potential burptide producers. These producers were widely distributed among the angiosperm phylogenetic orders. Finally, our transcriptome mining approached was validated when bioinformatically identified core sequences from *Gardenia jasminoides* were confirmed by mass spectrometry.

Figure 17: Cladogram of species mined from both genomes and transcriptomes



Potential producers are represented by navy blue dots, known producers are marked with orange dots.

CHAPTER V: CONCLUSIONS

V.A. Genome Mining in Fungi

Using a DUF3328 guided genome mining approach, numerous new putative dikaritin precursor peptides were identified in fungi from both Ascomycota and Basidiomycota. Dikaritins were not previously known to be produced by basidiomycetes. One example of this is *Pleurotus ostreatus.* Better known as oyster mushrooms, these fungi are commonly eaten and bioinformatically appear to produce a new dikaritin.

V.B. Genome Mining in Plants

Using a standalone precursor peptide found in *Ceanothus americanus*, a custom hidden Markov model was developed to search for other related precursor peptides. The results of searching all available Viridiplantae genomes showed a widely distributed split biosynthetic system was responsible for installing post-translational modifications in a newly named class of peptides known as burptides that includes hibispeptins, cyclopeptide alkaloids, moroidins and lyciumins.[24,26,39] The enzymes involved, BURP peptide cyclases, were previously only known as autocatalytic cyclases.[39] This genome mining approach was validated when one of the bioinformatically discovered precursor peptides from *Coffea arabica* went on to be isolated and named arabipeptin A.

V.C. Transcriptome Mining Pipeline

A script was developed that downloads, assembles and analyzes transcriptomic data deposited to the Sequence Read Archive for novel burptides. Using this transcriptome mining approach, numerous prospective burptides were identified from plants across all orders of angiosperms and cycads. Bioinformatically identified core sequences from *Gardenia jasminoides*

transcriptomic data were later confirmed by mass spectrometry, validating our transcriptome mining approach.

V.D. Future Work

Isolation and structural elucidation of the bioinformatically derived cores from both plants and fungi will continue to yield new molecules and insight into the biosynthetic pathways that produce them.

REFERENCES

(1) Sorokina, M.; Steinbeck, C. Review on Natural Products Databases: Where to Find Data in 2020. *J. Cheminform.* **2020**, *12* (1), 20.

(2) Dias, D. A.; Urban, S.; Roessner, U. A Historical Overview of Natural Products in Drug Discovery. *Metabolites* **2012**, *2* (2), 303–336.

(3) Shen, B. A New Golden Age of Natural Products Drug Discovery. *Cell* **2015**, *163* (6), 1297–1300.

(4) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83* (3), 770–803.

(5) Ortega, M. A.; van der Donk, W. A. New Insights into the Biosynthetic Logic of Ribosomally Synthesized and Post-Translationally Modified Peptide Natural Products. *Cell Chem. Biol.* **2016**, *23* (1), 31–44.

(6) 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.; Göransson, 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.; Müller, 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.; Süssmuth, 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. Ribosomally Synthesized and Post-Translationally Modified Peptide Natural Products: Overview and Recommendations for a Universal Nomenclature. *Nat. Prod. Rep.* **2013**, *30* (1), 108–160.

(7) Kessler, S. C.; Chooi, Y.-H. Out for a RiPP: Challenges and Advances in Genome Mining of Ribosomal Peptides from Fungi. *Nat. Prod. Rep.* **2022**, *39* (2), 222–230.

(8) Luo, S.; Dong, S.-H. Recent Advances in the Discovery and Biosynthetic Study of Eukaryotic RiPP Natural Products. *Molecules* **2019**, *24* (8), 1541.

(9) Sgambelluri, R. M.; Smith, M. O.; Walton, J. D. Versatility of Prolyl Oligopeptidase B in Peptide Macrocyclization. *ACS Synth. Biol.* **2018**, *7* (1), 145–152.

(10) Mayer, A.; Anke, H.; Sterner, O. Omphalotin, A New Cyclic Peptide with Potent Nematicidal Activity FromOmphalotus OleariusI. Fermentation and Biological Activity. *Nat. Prod. Lett.* **1997**, *10* (1), 25–32.

(11) van der Velden, N. S.; Kälin, N.; Helf, M. J.; Piel, J.; Freeman, M. F.; Künzler, M. Autocatalytic Backbone N-Methylation in a Family of Ribosomal Peptide Natural Products. *Nat. Chem. Biol.* **2017**, *13* (8), 833–835.

(12) Quijano, M. R.; Zach, C.; Miller, F. S.; Lee, A. R.; Imani, A. S.; Künzler, M.; Freeman, M. F. Distinct Autocatalytic α- N-Methylating Precursors Expand the Borosin RiPP Family of Peptide Natural Products. *J. Am. Chem. Soc.* **2019**, *141* (24), 9637–9644.

(13) Ramm, S.; Krawczyk, B.; Mühlenweg, A.; Poch, A.; Mösker, E.; Süssmuth, R. D. A Self-Sacrificing N-Methyltransferase Is the Precursor of the Fungal Natural Product Omphalotin. *Angew. Chem. Int. Ed Engl.* **2017**, *56* (33), 9994–9997.

(14) Ye, Y.; Minami, A.; Igarashi, Y.; Izumikawa, M.; Umemura, M.; Nagano, N.; Machida, M.; Kawahara, T.; Shin-ya, K.; Gomi, K.; Oikawa, H. Unveiling the Biosynthetic Pathway of the Ribosomally Synthesized and Post-Translationally Modified Peptide Ustiloxin B in Filamentous Fungi. *Angew. Chem. Int. Ed Engl.* **2016**, *55* (28), 8072–8075.
(15) Ding, W.; Liu, W.-Q.; Jia, Y.; Li, Y.; van der Donk, W. A.; Zhang, Q. Biosynthetic Investigation of Phomopsins Reveals a Widespread Pathway for Ribosomal Natural Products in Ascomycetes. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (13), 3521–3526.
(16) Ye, Y.; Ozaki, T.; Umemura, M.; Liu, C.; Minami, A.; Oikawa, H. Heterologous Production of Asperipin-2a: Proposal for Sequential Oxidative Macrocyclization by a Fungi-Specific DUF3328 Oxidase. *Org. Biomol. Chem.* **2018**, *17* (1), 39–43.
(17) Kessler, S. C.; Zhang, X.; McDonald, M. C.; Gilchrist, C. L. M.; Lin, Z.; Rightmyer, A.; Solomon, P. S.; Turgeon, B. G.; Chooi, Y.-H. Victorin, the Host-Selective Cyclic Peptide Toxin from the Oat Pathogen Cochliobolus Victoriae, Is Ribosomally Encoded. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (39), 24243–24250.
(18) Johnson, R. D.; Lane, G. A.; Koulman, A.; Cao, M.; Fraser, K.; Fleetwood, D. J.; Voisey, C. R.; Dyer, J. M.; Pratt, J.; Christensen, M.; Simpson, W. R.; Bryan, G. T.; Johnson, L. J. A Novel Family of Cyclic Oligopeptides Derived from Ribosomal Peptide Synthesis of an in Planta-Induced Gene, GigA, in Epichloë Endophytes of Grasses. *Fungal Genet. Biol.* **2015**, *85*, 14–24.
(19) Craik, D. J.; Daly, N. L.; Bond, T.; Waine, C. Plant Cyclotides: A Unique Family of Cyclic and Knotted Proteins That Defines the Cyclic Cystine Knot Structural Motif. *J. Mol. Biol.* **1999**, *294* (5), 1327–1336.
(20) Colgrave, M. L.; Craik, D. J. Thermal, Chemical, and Enzymatic Stability of the Cyclotide Kalata B1: The Importance of the Cyclic Cystine Knot. *Biochemistry* **2004**, *43* (20), 5965–5975.
(21) Gerlach, S.; Mondal, D. The Bountiful Biological Activities of Cyclotides. *Chron. Young Sci.* **2012**, *3* (3), 169.
(22) Tan, N.-H.; Zhou, J. Plant Cyclopeptides. *Chem. Rev.* **2006**, *106* (3), 840–895.
(23) Ramalho, S. D.; Pinto, M. E. F.; Ferreira, D.; Bolzani, V. S. Biologically Active Orbitides from the Euphorbiaceae Family. *Planta Med.* **2017**, *84* (9–10), 558–567.
(24) Lima, S. T.; Ampolini, B. G.; Underwood, E. B.; Graf, T. N.; Earp, C. E.; Khedi, I. C.; Pasquale, M. A.; Chekan, J. R. A Widely Distributed Biosynthetic Cassette Is Responsible for Diverse Plant Side Chain Cross-Linked Cyclopeptides. *Angew. Chem. Int. Ed Engl.* **2023**, *62* (7), e202218082.
(25) Kersten, R. D.; Weng, J.-K. Gene-Guided Discovery and Engineering of Branched Cyclic Peptides in Plants. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (46), E10961–E10969.
(26) Kersten, R. D.; Mydy, L. S.; Fallon, T. R.; de Waal, F.; Shafiq, K.; Wotring, J. W.; Sexton, J. Z.; Weng, J.-K. Gene-Guided Discovery and Ribosomal Biosynthesis of Moroidin Peptides. *J. Am. Chem. Soc.* **2022**, *144* (17), 7686–7692.
(27) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215* (3), 403–410.
(28) Wheeler, D.; Bhagwat, M. BLAST QuickStart: Example-Driven Web-Based BLAST Tutorial. In *Comparative Genomics*; Humana Press: New Jersey, 2007; pp 149–176.
(29) Eric, S. D.; Nicholas, T. K. D. D.; Theophilus, K. A. Bioinformatics with Basic Local Alignment Search Tool (BLAST) and Fast Alignment (FASTA). *J. Bioinform. Seq. Anal.* **2014**, *6* (1), 1–6.

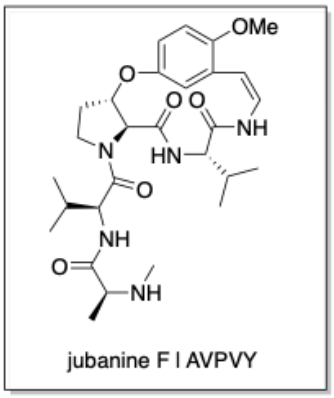
(30) Eddy, S. R. Profile Hidden Markov Models. *Bioinformatics* **1998**, *14* (9), 755–763.
(31) Eddy, S. R. What Is a Hidden Markov Model? *Nat. Biotechnol.* **2004**, *22* (10), 1315–1316.
(32) Finn, R. D.; Clements, J.; Eddy, S. R. HMMER Web Server: Interactive Sequence Similarity Searching. *Nucleic Acids Res.* **2011**, *39* (Web Server issue), W29-37.
(33) Zallot, R.; Oberg, N.; Gerlt, J. A. The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry* **2019**, *58* (41), 4169–4182.
(34) Umemura, M. Peptides Derived from Kex2-Processed Repeat Proteins Are Widely Distributed and Highly Diverse in the Fungi Kingdom. *Fungal Biol. Biotechnol.* **2020**, *7* (1), 11.
(35) Jiang, Y.; Ozaki, T.; Liu, C.; Igarashi, Y.; Ye, Y.; Tang, S.; Ye, T.; Maruyama, J.-I.; Minami, A.; Oikawa, H. Biosynthesis of Cyclochlorotine: Identification of the Genes Involved in Oxidative Transformations and Intramolecular O,N-Transacylation. *Org. Lett.* **2021**, *23* (7), 2616–2620.
(36) Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **2011**, *7* (10), e1002195.
(37) Morel, A. F.; Maldaner, G.; Ilha, V. Cyclopeptide Alkaloids from Higher Plants. *Alkaloids Chem. Biol.* **2009**, *67*, 79–141.
(38) Tuenter, E.; Exarchou, V.; Apers, S.; Pieters, L. Cyclopeptide Alkaloids. *Phytochem. Rev.* **2017**, *16* (4), 623–637.
(39) Chigumba, D. N.; Mydy, L. S.; de Waal, F.; Li, W.; Shafiq, K.; Wotring, J. W.; Mohamed, O. G.; Mladenovic, T.; Tripathi, A.; Sexton, J. Z.; Kautsar, S.; Medema, M. H.; Kersten, R. D. Discovery and Biosynthesis of Cyclic Plant Peptides via Autocatalytic Cyclases. *Nat. Chem. Biol.* **2022**, *18* (1), 18–28.
(40) Kang, K. B.; Ming, G.; Kim, G. J.; Ha, T.-K.-Q.; Choi, H.; Oh, W. K.; Sung, S. H. Jubanines F-J, Cyclopeptide Alkaloids from the Roots of Ziziphus Jujuba. *Phytochemistry* **2015**, *119*, 90–95.
(41) de Los Santos, E. L. C. NeuRiPP: Neural Network Identification of RiPP Precursor Peptides. *Sci. Rep.* **2019**, *9* (1), 13406.
(42) Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G. A.; Sonnhammer, E. L. L.; Tosatto, S. C. E.; Paladin, L.; Raj, S.; Richardson, L. J.; Finn, R. D.; Bateman, A. Pfam: The Protein Families Database in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D412–D419.
(43) Letunic, I. *phyloT : a phylogenetic tree generator*. https://phylot.biobyte.de (accessed 2023-05-26).
(44) Yu, G. Using Ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinformatics* **2020**, *69* (1), e96.
(45) Xu, S.; Dai, Z.; Guo, P.; Fu, X.; Liu, S.; Zhou, L.; Tang, W.; Feng, T.; Chen, M.; Zhan, L.; Wu, T.; Hu, E.; Jiang, Y.; Bo, X.; Yu, G. GgtreeExtra: Compact Visualization of Richly Annotated Phylogenetic Data. *Mol. Biol. Evol.* **2021**, *38* (9), 4039–4042.
(46) Prjibelski, A.; Antipov, D.; Meleshko, D.; Lapidus, A.; Korobeynikov, A. Using SPAdes DE Novo Assembler. *Curr. Protoc. Bioinformatics* **2020**, *70* (1), e102.
(47) Haas, B. J.; Papanicolaou, A.; Yassour, M.; Grabherr, M.; Blood, P. D.; Bowden, J.; Couger, M. B.; Eccles, D.; Li, B.; Lieber, M.; MacManes, M. D.; Ott, M.; Orvis, J.; Pochet, N.; Strozzi, F.; Weeks, N.; Westerman, R.; William, T.; Dewey, C. N.; Henschel, R.; LeDuc, R. D.; Friedman, N.; Regev, A. De Novo Transcript Sequence Reconstruction from

RNA-Seq Using the Trinity Platform for Reference Generation and Analysis. *Nat. Protoc.* **2013**, *8* (8), 1494–1512.

(48) The Angiosperm Phylogeny Group. An Update of the Angiosperm Phylogeny Group Classification for the Orders and Families of Flowering Plants: APG IV. *Bot. J. Linn. Soc.* **2016**, *181* (1), 1–20.

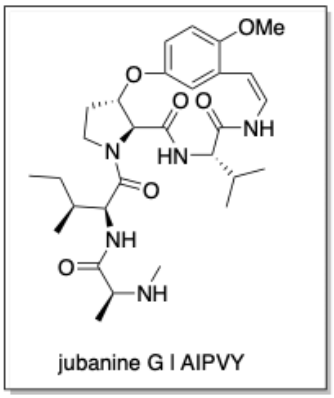
APPENDIX A: STRUCTURES FROM ZIZIPHUS JUJUBA

>KAH7517850.1 [Ziziphus jujuba var. spinosa]
MNLLDIILEHKGVGEKLARSCSGVVLVPPFALSSTIIARKEPVEY
VKTVVEQVAKDLSVDPSFAVPFYRSNKNGQSVDPSSAVPFHHGNK
NGLSIDPSSAIPVYRGNQNGQSIDPSAAIPIYHGNQNGQSIDPSS
ALLIYRINQNGLSIDPSSAVPFYHGNKNGFSIDPSSAVPFYHGNK
NGLSVDPSSAIPVYQGNQNGQSIDPSS**AVPVY**RSNQNGQSIDPSS
AVPIYHGNKNGLSVDPSSAIFVYRGNQNGQSIDPSSAIPVYRGNQ
NGQSIDPSSAIPVYRGNQNGQSIDPSSALLIYRINQNGLSIDPSS
AVPFYHGKKNGFFVDPSSAVPFYHGNKNGLSVDPSSAIPVYQGNQ
NGQSIDPSS**AVPVY**RSNQNGQSIDPSSAVPIYHGNKNGLSVDPSS
AIFVYRGNQNGQSIDPSSAIPIYHGNQNGQSIDPSS**AVPVY**RGNQ
NGQSIDPSSALLIYRINQNGLSIDPSSAVPFYHGNQNGLSMDLSS
AVPFYRGIKNGLSVDPSSAVPFYYENKNGLSVDPSSAVPFYRGNQ
NGEQNSKANEKGATKNLSTDQ



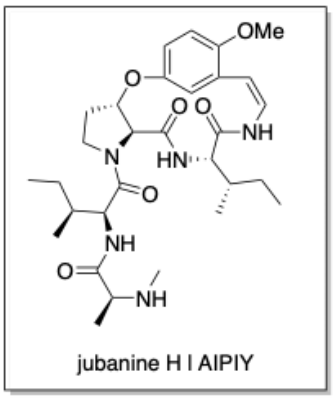
jubanine F I AVPVY

>KAH7517850.1 [Ziziphus jujuba var. spinosa]
MNLLDIILEHKGVGEKLARSCSGVVLVPPFALSSTIIARKEPVEY
VKTVVEQVAKDLSVDPSFAVPFYRSNKNGQSVDPSSAVPFHHGNK
NGLSIDPSS**AIPVY**RGNQNGQSIDPSAAIPIYHGNQNGQSIDPSS
ALLIYRINQNGLSIDPSSAVPFYHGNKNGFSIDPSSAVPFYHGNK
NGLSVDPSS**AIPVY**QGNQNGQSIDPSSAVPVYRSNQNGQSIDPSS
AVPIYHGNKNGLSVDPSSAIFVYRGNQNGQSIDPSS**AIPVY**RGNQ
NGQSIDPSS**AIPVY**RGNQNGQSIDPSSALLIYRINQNGLSIDPSS
AVPFYHGKKNGFFVDPSSAVPFYHGNKNGLSVDPSS**AIPVY**QGNQ
NGQSIDPSSAVPVYRSNQNGQSIDPSSAVPIYHGNKNGLSVDPSS
AIFVYRGNQNGQSIDPSSAIPIYHGNQNGQSIDPSSAVPVYRGNQ
NGQSIDPSSALLIYRINQNGLSIDPSSAVPFYHGNQNGLSMDLSS
AVPFYRGIKNGLSVDPSSAVPFYYENKNGLSVDPSSAVPFYRGNQ
NGEQNSKANEKGATKNLSTDQ



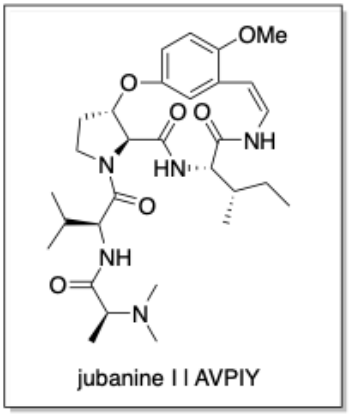
jubanine G I AIPVY

>KAH7517850.1 [Ziziphus jujuba var. spinosa]
MNLLDIILEHKGVGEKLARSCSGVVLVPPFALSSTIIARKEPVEYV
KTVVEQVAKDLSVDPSFAVPFYRSNKNGQSVDPSSAVPFHHGNKNG
LSIDPSSAIPVYRGNQNGQSIDPSA**AIPIY**HGNQNGQSIDPSSALL
IYRINQNGLSIDPSSAVPFYHGNKNGFSIDPSSAVPFYHGNKNGLS
VDPSSAIPVYQGNQNGQSIDPSSAVPVYRSNQNGQSIDPSSAVPIY
HGNKNGLSVDPSSAIFVYRGNQNGQSIDPSSAIPVYRGNQNGQSID
PSSAIPVYRGNQNGQSIDPSSALLIYRINQNGLSIDPSSAVPFYHG
KKNGFFVDPSSAVPFYHGNKNGLSVDPSSAIPVYQGNQNGQSIDPS
SAVPVYRSNQNGQSIDPSSAVPIYHGNKNGLSVDPSSAIFVYRGNQ
NGQSIDPSS**AIPIY**HGNQNGQSIDPSSAVPVYRGNQNGQSIDPSSA
LLIYRINQNGLSIDPSSAVPFYHGNQNGLSMDLSSAVPFYRGIKNG
LSVDPSSAVPFYYENKNGLSVDPSSAVPFYRGNQNGEQNSKANEKG
ATKNLSTDQ



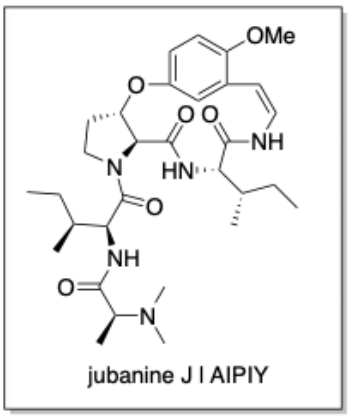
jubanine H I AIPIY

>KAH7517850.1 [Ziziphus jujuba var. spinosa]
MNLLDIILEHKGVGEKLARSCSGVVLVPPFALSSTIIARKEPVEY
VKTVVEQVAKDLSVDPSFAVPFYRSNKNGQSVDPSSAVPFHHGNK
NGLSIDPSSAIPVYRGNQNGQSIDPSAAIPIYHGNQNGQSIDPSS
ALLIYRINQNGLSIDPSSAVPFYHGNKNGFSIDPSSAVPFYHGNK
NGLSVDPSSAIPVYQGNQNGQSIDPSSAVPVYRSNQNGQSIDPSS
AVPIYHGNKNGLSVDPSSAIFVYRGNQNGQSIDPSSAIPVYRGNQ
NGQSIDPSSAIPVYRGNQNGQSIDPSSALLIYRINQNGLSIDPSS
AVPFYHGKKNGFFVDPSSAVPFYHGNKNGLSVDPSSAIPVYQGNQ
NGQSIDPSSAVPVYRSNQNGQSIDPSS**AVPIY**HGNKNGLSVDPSS
AIFVYRGNQNGQSIDPSSAIPIYHGNQNGQSIDPSSAVPVYRGNQ
NGQSIDPSSALLIYRINQNGLSIDPSSAVPFYHGNQNGLSMDLSS
AVPFYRGIKNGLSVDPSSAVPFYYENKNGLSVDPSSAVPFYRGNQ
NGEQNSKANEKGATKNLSTDQ



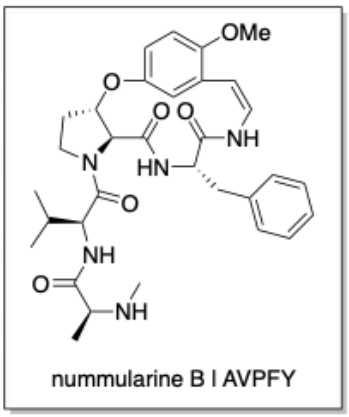
jubanine I | AVPIY

>KAH7517850.1 [Ziziphus jujuba var. spinosa]
MNLLDIILEHKGVGEKLARSCSGVVLVPPFALSSTIIARKEPVEY
VKTVVEQVAKDLSVDPSFAVPFYRSNKNGQSVDPSSAVPFHHGNK
NGLSIDPSSAIPVYRGNQNGQSIDPSA**AIPIY**HGNQNGQSIDPSS
ALLIYRINQNGLSIDPSSAVPFYHGNKNGFSIDPSSAVPFYHGNK
NGLSVDPSSAIPVYQGNQNGQSIDPSSAVPVYRSNQNGQSIDPSS
AVPIYHGNKNGLSVDPSSAIFVYRGNQNGQSIDPSSAIPVYRGNQ
NGQSIDPSSAIPVYRGNQNGQSIDPSSALLIYRINQNGLSIDPSS
AVPFYHGKKNGFFVDPSSAVPFYHGNKNGLSVDPSSAIPVYQGNQ
NGQSIDPSSAVPVYRSNQNGQSIDPSSAVPIYHGNKNGLSVDPSS
AIFVYRGNQNGQSIDPSS**AIPIY**HGNQNGQSIDPSSAVPVYRGNQ
NGQSIDPSSALLIYRINQNGLSIDPSSAVPFYHGNQNGLSMDLSS
AVPFYRGIKNGLSVDPSSAVPFYYENKNGLSVDPSSAVPFYRGNQ
NGEQNSKANEKGATKNLSTDQ



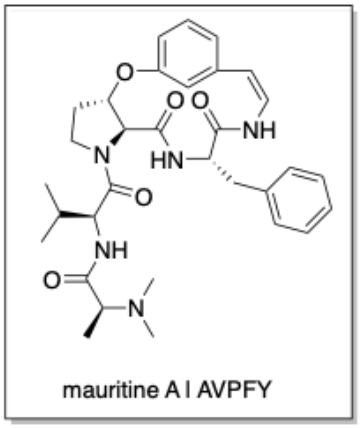
jubanine J | AIPIY

>KAH7517853.1 [Ziziphus jujuba var. spinosa]
MSINTESYTQPLHHSATTHLRTLPKKERIIASHFELFQIQLVIKMK
SFFALLAFSSLLLLSSTITARKEPAEYVKTVVEQVVKDLSVHPSS**A**
VPFYRSNKNGQSVDPSS**AVPFY**HENKNDLSVDPSS**AVPFY**RGNQND
QSVDPSS**AVPFY**HGNKNGLSVDPSSNKNGLSVDPSSALPFYRGNQN
GQSVDPSS**AVPFY**HGNKNGLSVDPSSVLPFYRGNQNGQSVDPSS**AV**
PFYHGNKNGLSVDPSSVLPFYRGNQNGQSVDPSS**AVPFY**HGNKNGL
SVDPSSVLPFYRGNQNGQSVDPSS**AVPFY**HGNKNGLSVDPSS**AVPF**
YRGNQNGQSVDPSS**AVPFY**HGNKNGLSVDPSS**AVPFY**RGNQNGQSV
DPSS**AVPFY**HGNKNGFFVDPSS**AVPFY**HGNQNGQSVDPSS**AVPFY**H
GNKNGFFVDPSS**AVPFY**HGNQNGQSVDPSS**AVPFY**HGNKNGFFVDP
SSAIPFYYGNKNGLYIDPSS**AVPFY**HSNQNNQSVDPSS**AVPFY**RGN
QNGDQNSKANEEGATKNLPLTNEM**AVPFY**RGNQNGQFVDPSS**AVPF**
YHGNKNSLSVDPSS**AVPFY**RSNQNGQSIDPSS**AVPFY**HR



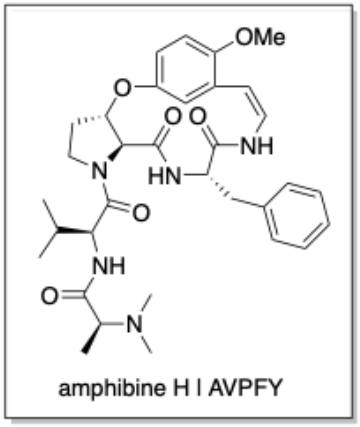
nummularine B | AVPFY

>KAH7517853.1 [Ziziphus jujuba var. spinosa]
MSINTESYTQPLHHSATTHLRTLPKKERIIASHFELFQIQLVIKM
KSFFALLAFSSLLLLSSTITARKEPAEYVKTVVEQVVKDLSVHPS
S**AVPFY**RSNKNGQSVDPSS**AVPFY**HENKNDLSVDPSS**AVPFY**RGN
QNDQSVDPSS**AVPFY**HGNKNGLSVDPSSNKNGLSVDPSSALPFYR
GNQNGQSVDPSS**AVPFY**HGNKNGLSVDPSSVLPFYRGNQNGQSVD
PSS**AVPFY**HGNKNGLSVDPSSVLPFYRGNQNGQSVDPSS**AVPFY**H
GNKNGLSVDPSSVLPFYRGNQNGQSVDPSS**AVPFY**HGNKNGLSVD
PSS**AVPFY**RGNQNGQSVDPSS**AVPFY**HGNKNGLSVDPSS**AVPFY**R
GNQNGQSVDPSS**AVPFY**HGNKNGFFVDPSS**AVPFY**HGNQNGQSVD
PSS**AVPFY**HGNKNGFFVDPSS**AVPFY**HGNQNGQSVDPSS**AVPFY**H
GNKNGFFVDPSSAIPFYYGNKNGLYIDPSS**AVPFY**HSNQNNQSVD
PSS**AVPFY**RGNQNGDQNSKANEEGATKNLPLTNEM**AVPFY**RGNQN
GQFVDPSS**AVPFY**HGNKNSLSVDPSS**AVPFY**RSNQNGQSIDPSS**A**
VPFYHR



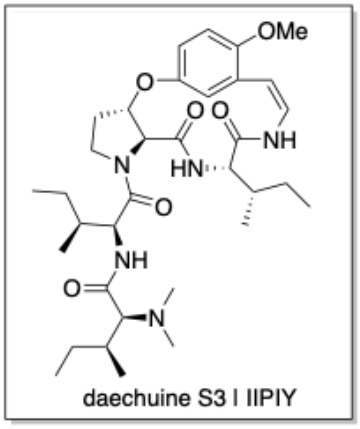
mauritine A I AVPFY

>KAH7517853.1 [Ziziphus jujuba var. spinosa]
MSINTESYTQPLHHSATTHLRTLPKKERIIASHFELFQIQLVIKM
KSFFALLAFSSLLLLSSTITARKEPAEYVKTVVEQVVKDLSVHPS
S**AVPFY**RSNKNGQSVDPSS**AVPFY**HENKNDLSVDPSS**AVPFY**RGN
QNDQSVDPSS**AVPFY**HGNKNGLSVDPSSNKNGLSVDPSSALPFYR
GNQNGQSVDPSS**AVPFY**HGNKNGLSVDPSSVLPFYRGNQNGQSVD
PSS**AVPFY**HGNKNGLSVDPSSVLPFYRGNQNGQSVDPSS**AVPFY**H
GNKNGLSVDPSSVLPFYRGNQNGQSVDPSS**AVPFY**HGNKNGLSVD
PSS**AVPFY**RGNQNGQSVDPSS**AVPFY**HGNKNGLSVDPSS**AVPFY**R
GNQNGQSVDPSS**AVPFY**HGNKNGFFVDPSS**AVPFY**HGNQNGQSVD
PSS**AVPFY**HGNKNGFFVDPSS**AVPFY**HGNQNGQSVDPSS**AVPFY**H
GNKNGFFVDPSSAIPFYYGNKNGLYIDPSS**AVPFY**HSNQNNQSVD
PSS**AVPFY**RGNQNGDQNSKANEEGATKNLPLTNEM**AVPFY**RGNQN
GQFVDPSS**AVPFY**HGNKNSLSVDPSS**AVPFY**RSNQNGQSIDPSS**A**
VPFYHR



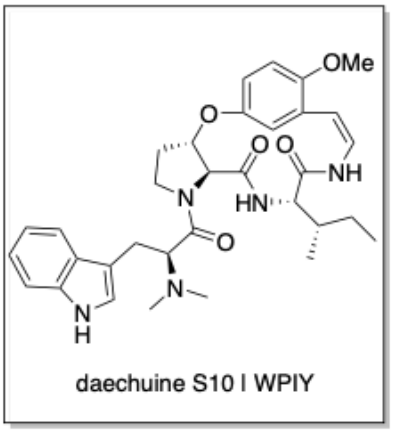
amphibine H I AVPFY

>KAH7517848.1 [Ziziphus jujuba var. spinosa]
MKSFFALLAFSSLFLFSSTITARKEPVEYVKTVVDQVAKDLYVDPS
SHILLYHGKQNGKDAAKDQSVDPSS**IIPIY**HGNKNGLSIDPSS**IIP**
IYHENRNGQSVDPLSFIPIYLGKKNGLSIDASS**IIPIY**HGQKNGFS
VDPSS**IIPIY**GNHNAKDHSVDPSSLIPIYHGNQNGQSVDPSSLIPI
YRDNQNGQSVDPLSLVLLYRGNQQNGQSIDPSSLIPFYRGNQNGQS
VDPLSLIPFYRGNQNGQSVDPSSLIPFYRGNQNGQSVDPSSLIPFY
RGKQQNGQFVDPSSLIPFYRGNQNGQSVDPSSLIPIYRGNQNGDQN
FKANEEGVAKSVSTDQ



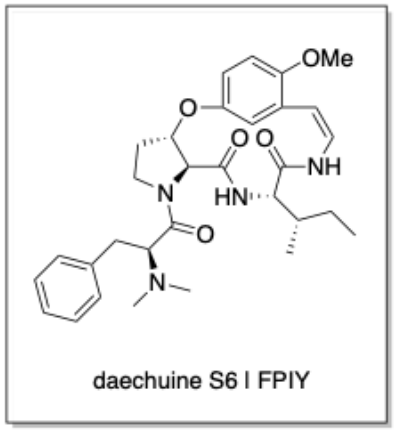
daechuine S3 I IIPIY

>XP_024932849.1 [Ziziphus jujuba]
MKSFFALLAFSSLLLLSSIVIARKEPVEYVKTVVEQVAKDLSIDP
SSR**WPIY**HGNQNGQSVDPLSR**WPIY**HGNQDQQFVDPLSR**WPIY**NG
NQNGQSVDPLSR**WPIY**HGNQNGQSVDPSSR**WPIY**HGNQNGQSVDP
SSR**WPIY**HGNQNGQSVDPSSR**WPIY**HGNRNGKSVDPSSW**WPIY**QN
GQSVDPSSW**WPIY**QNGQFVDPSSR**WPIY**HGNQNRQSVDPSSR**WPI**
DHGNQNGQSVDPSSR**WPIY**HGNQNGQSVDPSSK**WPIY**HGNQNGDQ
NSKANEEGAAKSASTDQ



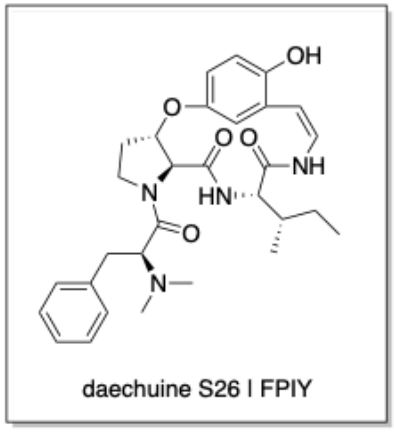
daechuine S10 | WPIY

>XP_024933341.1 [Ziziphus jujuba]
MKSFFALLAFSSLLLLSSIVIARKEPVEYVKTVVEQVAKDLSIDP
SSR**FPIY**HDNQNGQSVDTSSR**FPIY**HDNQNGQSIDPSSRFLIYHG
IDPFSR**FPIY**HGNQNGQSIDPSSR**FPIY**HGNQNGQSIDPSSQFLI
YRNGKSIDPSSR**FPIY**HGNQNGQSVDPSSR**FPIY**HDNQNGQSIDP
SSR**FPIY**HDNQNRQSIDPSSR**FPIY**HGNQNGQSIDPSSRFLIYHG
IDPSSR**FPIY**HGNQNGQSIDPSSR**FPIY**HGNQNGDQNSKANEEGA
AKSASTDQ



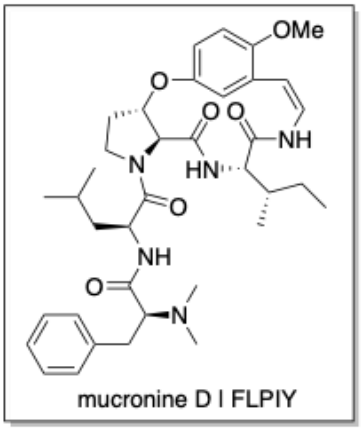
daechuine S6 | FPIY

>XP_024933341.1 [Ziziphus jujuba]
MKSFFALLAFSSLLLLSSIVIARKEPVEYVKTVVEQVAKDLSIDPS
SR**FPIY**HDNQNGQSVDTSSR**FPIY**HDNQNGQSIDPSSRFLIYHGID
PFSR**FPIY**HGNQNGQSIDPSSR**FPIY**HGNQNGQSIDPSSQFLIYRN
GKSIDPSSR**FPIY**HGNQNGQSVDPSSR**FPIY**HDNQNGQSIDPSSR**F**
PIYHDNQNRQSIDPSSR**FPIY**HGNQNGQSIDPSSRFLIYHGIDPSS
R**FPIY**HGNQNGQSIDPSSR**FPIY**HGNQNGDQNSKANEEGAAKSAST
DQ



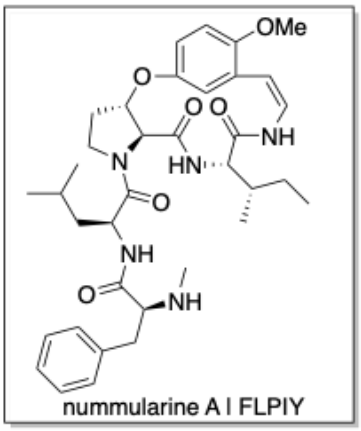
daechuine S26 | FPIY

>KAH7517857.1 [Ziziphus jujuba var. spinosa]
MKSFFALLAFSSLLLLSSTITARKEPGEYVKTVVEQVAEDLFVDP
SSIIPFYHKNKNGQSVDPLS**FLPIY**HGNQNGQSIDPSSLLLIYHG
NQIGQSVDPSS**FLPIY**HSNQNGQSVDPSS**FLPIY**HGNRNGQSVDP
SS**FLPIY**HGNRNGQSVDPSS**FLPIY**HGNRNGQSVDPSS**FLPIY**HG
NRNGQSVDPSS**FLPIY**HGNQNGQSVDPSS**FLPIY**QGNRNGQSVDP
SS**FLPIY**QGNRNGQSVDPSS**FLPIY**HGNLNRQSVDPSS**FLPIY**HG
NLNGQSVDPSS**FLPIY**HGNQNGQSVDPSSFHLLYHGNRNGQSVDP
SS**FLPIY**HGNLNGQSIDPSSRNGQSVDPSSLLLIYRDNQNGEQNP
KANEEGVAKSISTDQ



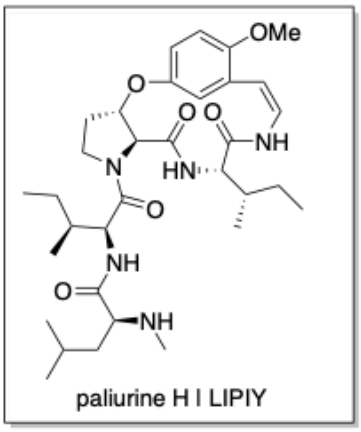
mucronine D I FLPIY

>KAH7517857.1 [Ziziphus jujuba var. spinosa]
MKSFFALLAFSSLLLLSSTITARKEPGEYVKTVVEQVAEDLFVDP
SSIIPFYHKNKNGQSVDPLS**FLPIY**HGNQNGQSIDPSSLLLIYHG
NQIGQSVDPSS**FLPIY**HSNQNGQSVDPSS**FLPIY**HGNRNGQSVDP
SS**FLPIY**HGNRNGQSVDPSS**FLPIY**HGNRNGQSVDPSS**FLPIY**HG
NRNGQSVDPSS**FLPIY**HGNQNGQSVDPSS**FLPIY**QGNRNGQSVDP
SS**FLPIY**QGNRNGQSVDPSS**FLPIY**HGNLNRQSVDPSS**FLPIY**HG
NLNGQSVDPSS**FLPIY**HGNQNGQSVDPSSFHLLYHGNRNGQSVDP
SS**FLPIY**HGNLNGQSIDPSSRNGQSVDPSSLLLIYRDNQNGEQNP
KANEEGVAKSISTDQ



nummularine A I FLPIY

>KAH7517848.1 [Ziziphus jujuba var. spinosa]
MKSFFALLAFSSLFLFSSTITARKEPVEYVKTVVDQVAKDLYVDPS
SHILLYHGKQNGKDAAKDQSVDPSSIIPIYHGNKNGLSIDPSSIIP
IYHENRNGQSVDPLSFIPIYLGKKNGLSIDASSIIPIYHGQKNGFS
VDPSSIIPIYGNHNAKDHSVDPSS**LIPIY**HGNQNGQSVDPSS**LIPI**
YRDNQNGQSVDPLSLVLLYRGNQQNGQSIDPSSLIPFYRGNQNGQS
VDPLSLIPFYRGNQNGQSVDPSSLIPFYRGNQNGQSVDPSSLIPFY
RGKQQNGQFVDPSSLIPFYRGNQNGQSVDPSS**LIPIY**RGNQNGDQN
FKANEEGVAKSVSTDQ



paliurine H I LIPIY

APPENDIX B: TRANSCRIPTOME ASSEMBLY PIPELINE & ANALYSIS SCRIPT

```
#Prompts for SRA accessions
echo Accession numbers\?

#Reads an input of accessions separated by a space into an array
read -a accession

for i in ${accession[@]} do cd /mnt/c/software/sratoolkit.3.0.0-ubuntu64/
#sratoolkit.3.0.0 is the directory where each accession folder will end up

#Makes a folder for each accession
mkdir $i

#Moves into the new folder
cd $i

pwd

#Downloading the SRA fastq file
../bin/fastq-dump-orig.3.0.0 --defline-seq '@$sn[_$rn]/$ri' --split-files
$i

#Assembling with SPAdes
/home/chekanlab/software/miniconda3/envs/spades/bin/rnaspades.py -o
//mnt/c/software/sratoolkit.3.0.0-ubuntu64/$i/spades -1
//mnt/c/software/sratoolkit.3.0.0-ubuntu64/$i/${i}_1.fastq -2
//mnt/c/software/sratoolkit.3.0.0-ubuntu64/$i/${i}_2.fastq -t 12 -m 50

#Running TransDecoder
/mnt/c/software/TransDecoder-TransDecoder-v5.5.0/TransDecoder.LongOrfs -m
75 -t
//mnt/c/software/sratoolkit.3.0.0-ubuntu64/$i/spades/transcripts.fasta -O
//mnt/c/software/sratoolkit.3.0.0-ubuntu64/$i/transdecoder

#Moves the .pep file into the corresponding accession folder
mv
/mnt/c/software/sratoolkit.3.0.0-ubuntu64/$i/transdecoder/longest_orfs.pep
/mnt/c/software/sratoolkit.3.0.0-ubuntu64/$i/${i}.pep

#Deleting files that are not needed
rm /mnt/c/software/sratoolkit.3.0.0-ubuntu64/$i/${i}_1.fastq rm
/mnt/c/software/sratoolkit.3.0.0-ubuntu64/$i/${i}_2.fastq rm -r
/mnt/c/software/sratoolkit.3.0.0-ubuntu64/$i/spades
#rm -r /mnt/c/software/sratoolkit.3.0.0-ubuntu64/$i/transdecoder
rm -r
/mnt/c/software/sratoolkit.3.0.0-ubuntu64/$i/transdecoder.
__checkpoints_longorfs

#Runs the burptide HMM
/home/chekanlab/software/miniconda3/envs/hmmer/bin/hmmsearch
```

```
//mnt/c/software/scripts/pscc_update.hmm
//mnt/c/software/sratoolkit.3.0.0-ubuntu64/$i/${i}.pep > ${i}_hmm.txt

#Runs the BURP HMM
/home/chekanlab/software/miniconda3/envs/hmmer/bin/hmmsearch
//mnt/c/software/scripts/burp.hmm
//mnt/c/software/sratoolkit.3.0.0-ubuntu64/$i/${i}.pep > ${i}_burp_hmm.txt

#Files will be found in the corresponding accession folder
done
```

Appendix B: Transcriptome assembly and analysis pipeline code utilizing SPAdes, TransDecoder and HMMER.