

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

Corresponding Author: Professor Roland Kersten

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The study submitted by Wang et al. aims to rationalize the use of an improved method for RiPP discovery based on the lack of incomplete (or low quality) assemblies of deposited sequence data. The discovery of novel peptides, is without doubt, a very timely and important field of research, and the experiments and data presented are extensive and of high quality. However, in its current form the manuscript lacks a clear answer to the following questions:

- why is there a need for this optimized mining approach?

depending on the size/length of these peptides it is not entirely clear why they cannot be found in non-assembled SRAs, since the length of the deposited sequence normally exceeds the length of these short peptides... surely this wont give you the full precursor, but at least you might find the peptides...

- what exactly are the advantages of this approach, and how does it compare to alternative approaches?

compare and comment on the use of alternative methods e.g. targeted peptidomics-based discoveries (<https://www.nature.com/articles/s41467-023-39905-4>) or the use of AI/ML ([https://www.cell.com/cell/fulltext/S0092-8674\(24\)00522-1](https://www.cell.com/cell/fulltext/S0092-8674(24)00522-1))

- has the approach been validated with other datasets (except SRAs)

how does the mining approach compare to when using datasets with already assembled contigs, i.e. can you accelerate or extend the number of IDs you obtain, or is the novelty in the way you assemble transcriptomes, irrespective of any previous (automated assemblies)?... in case this is a misunderstanding, it clearly exemplifies the need for a more thorough description of the rationale and the method in itself

- what exactly are the benefits of discovery of these diverse peptides with respect to activity, medicinal chemistry/structure-activity and potential translation?

surely this is a critical point, and in its current form is not addressed to satiety: firstly, what's the reference compounds for bioassays (known peptides and current drugs)? Not clear how an IC50 in the μM range can be called "improved bioactivity"; how do these natural variants compare, i.e. since you appear to use a high-content screening unit at 384 well format, it should easily be possible to obtain a larger picture, and how these naturally occurring variants be rationalized to guide SAR and MedChem efforts in the future; s.d. appear to be a more appropriate way of data presentation for these type of assays (doi: 10.4103/2229-3485.100662)

Along this line, the presented Figures 2-4 are somehow repetitive. For instance, these could be combined into two multipanel Figures highlighting one example, i.e. the important discovery, e.g. moroidins, for which diverse peptide have been found. This could be presented as a "discovery" (containing the assemblies, the cloning and gene structure and diversity of the discovered peptides, including the cyclases) and "activity" (containing the isolation/analytcs, production, structure and bioactivity of all these peptides) Figure, by moving parts from the SI in the main Figures, and extending the study (assays) where necessary. Further, the authors should focus and extend on the actual workflow (Fig. 1) and the comparative study (Table 1) to other assemblies and datasets, while with presenting more numbers... e.g. how many peptides and in how many families were discovered; how does this compare to other mining approaches?

Other comments:

- In the light of the above said, the title appears too imprecise and should be improved: what is “repository-scale”?... and did you really address the “improved bioactivity”?
- The abstract should be carefully revised to allow the reader to understand that the different classes of peptides are known, but in that the authors present a workflow to identify novel peptides (how many in total, distributed amongst how many phylogenetic clades?)

Reviewer #2

(Remarks to the Author)

This manuscript assessed three de novo short-read assemblers for generating precursor peptides from RNAseq datasets of RiPP-producing plants. The most effective assembler was applied to identify moroidin analogs with novel second-ring structures and uncover the cyclases responsible for their biosynthesis. Through repository-scale transcriptome mining, the authors discovered two stephanotic acid-type burpitides with unique second-ring crosslinks and isolated a new moroidin analog with potent in vitro cytotoxicity, showing selective toxicity against H1437 lung adenocarcinoma cells. This work significantly expands the medicinal chemistry toolbox for the chemical and biological exploration of eukaryotic RiPP lead structures, highlighting its scientific importance and application potential. However, the following points warrant further discussion with the authors:

1. The authors have optimized the RNA-seq workflow. Besides improvements in speed and memory usage, what other advantages do these optimizations offer compared to the original workflow?
2. Although the manuscript compares different assembly tools, how accurate are these tools in handling repetitive sequences and short-read sequences? Could assembly errors potentially affect the results?
3. The improved RNA-seq workflow has been successful in diversifying and enhancing stephanotic acid compounds through transcriptome mining. Is this improved method applicable to the discovery of other types of compounds?
4. The manuscript mentions the cytotoxicity of some compounds against non-small cell lung cancer cells. Have these compounds been broadly tested against other types of cancer cells or non-cancerous cells to verify their selectivity and safety?
5. The manuscript proposes new biosynthetic pathways. Have these pathways been experimentally validated, or are they solely based on bioinformatics predictions?
6. Had the authors built a flowchart or document of the de novo NGS assembler method to make it easier for readers to use?
7. In Figure 4F, the error bar of Moroidin-QLLVWRNH dose-response assay was a little high.

Reviewer #3

(Remarks to the Author)

The NCBI sequence read archive (SRA) has the RNA-seq datasets of nearly 10,000 plant species, which proved a valuable resource for identifying new functional genes. In this paper, the authors developed a workflow to mine the new moroidin analogs from RNA-seq data in SRA. The first step of their workflow is to de novo assemble short NGS RNA-seqs reads into longer transcripts as most of the plants do not have reference genomes yet. In order to identify a suitable de novo NGS assembler for large-scale plant transcriptome assembly, the authors evaluated the performance of three assemblers: Trinity, SPAdes, and MEGAHIT. And the results showed that SPAdes is the best assembler for the task. By leveraging SPAdes, the authors identified a diverse set of moroidin cyclases and potential new burpitide chemotypes. These findings open avenues for the bioengineering of bicyclic RiPPs with enhanced therapeutic potential.

Comments:

1. The authors used 138 plant RiPP precursor genes to evaluate the de novo assemblers. It is unclear how many of them come from the genome with reference sequences. Could authors evaluate the performance of de novo assembler using RNA-seq datasets with reference genome? and comparing the de novo assembler methods with the reference genome-based methods.
2. Did the authors evaluate the parameter of tblastn? How will they affect the identification of morindin analogs? Did the authors use other search tools? For example, PHI-blast, as the moroidin has motifs.
3. All numbers should be in scientific format, for example, 10000 to 10,000; 27000 to 27,000.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

All earlier comments have been addressed to full satisfaction. Great paper!

Reviewer #2

(Remarks to the Author)

This manuscript has been greatly improved and I think it has met the publication standards.

Reviewer #3

(Remarks to the Author)

The authors have addressed all my concerns. Thank you and congratulations!

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Response to reviewer comments

We thank all reviewers for their evaluation of our manuscript and answer their questions and suggestions in blue color in this response document and in green color for corresponding text inserted into the manuscript and Supporting Information.

Reviewer #1 (Remarks to the Author):

The study submitted by Wang et al. aims to rationalize the use of an improved method for RiPP discovery based on the lack of incomplete (or low quality) assemblies of deposited sequence data. The discovery of novel peptides, is without doubt, a very timely and important field of research, and the experiments and data presented are extensive and of high quality. However, in its current form the manuscript lacks a clear answer to the following questions:

We thank the reviewer for the evaluation of our manuscript and for raising important questions for its improvement.

- why is there a need for this optimized mining approach?
depending on the size/length of these peptides it is not entirely clear why they cannot be found in non-assembled SRAs, since the length of the deposited sequence normally exceeds the length of these short peptides... surely this wont give you the full precursor, but at least you might find the peptides...

We are grateful for asking if assembly is needed for our peptide discovery approach. We tested this important question based on our target compound class, i.e. stephanotic acid-type burpitides.

For this, we benchmarked searching a short peptide motif such as the stephanotic acid motif QLLVW in unassembled plant RNA-seq data against searching the QLLVW motif in assembled plant RNA-seq data. QLLVW was chosen as a motif because it resembles the first ring of moroidin and is the core peptide motif for large-scale transcriptome search of stephanotic acid-type peptides in our manuscript. For benchmarking, we used 1137 RNA-seq datasets from our large-scale assembly representing 890 plant species (~10% of all species in our study) which also have tandem-mass spectrometry-based metabolomic data deposited in a searchable plant metabolomic repository called plantMASST (<https://doi.org/10.1101/2024.05.13.593988>). Therefore, these species can be queried for the presence of predicted peptide analytes such as the target stephanotic acid-QLLVW in plantMASST.

First, we searched the unassembled data for the presence of short peptide motif QLLVW by (1) a blastn search with the nucleotide sequence of the QLLVW sequence from moroidin precursor gene KjaBURP as a query sequence and (2) a seqkit2-grep search of 6frame-translated protein sequences of the unassembled RNA-seq data with the protein sequence QLLVW as a query.

Our analysis showed that:

- a. 11 of 890 tested plant species had a stephanotic acid-matching analyte in their metabolomic data from plantMASST search and LCMS-based chemotyping (**Figure S2**).
- b. Searching unassembled RNA-seq data via blastn for QLLVW identified that 688 of 890 tested plant species had a sequence match (False discovery rate: 98%)

- c. Searching unassembled RNA-seq data via seqkit2 for QLLVW identified that 800 of 890 tested plant species had a sequence match (False discovery rate: 99%)

As searches of unassembled data for our short peptide motif resulted in very high false discovery rates, we searched unassembled and assembled data of the corresponding plant species for homologs of moroidin cyclase KjaBURP (**Figure 1D**) and the presence of the motif QLLVW via PHI-BLAST, a protein BLAST with sequence motif search option (as suggested by reviewer 3).

Our benchmarking showed that:

- d. Searching unassembled RNA-seq data via PHI-BLAST for homologs of BURP-domain protein KjaBURP with search motif QLLVW identified 7 stephanotic acid-producing plants (False discovery rate: 46%) whereas searching assembled data identified 9 stephanotic acid-producing plants (False discovery rate: 36%)
- e. Assembly resulted in reduction of BLAST database size by 99.8%, BLAST time by 99.5% and BLAST memory by 67% (with a set maximum memory of 48 GB for both searches on the same computational resource). The final BLAST database for unassembled data search was 14 TB compared to 27 GB for assembled data search.

In conclusion, our benchmarking analysis showed that *de novo* RNA-seq assembly can reduce database size, search time, and search cost for peptide discovery and increase search hit numbers. Therefore, *de novo* RNA-seq data assembly is advantageous for mining RNA-seq data at scale for biosynthetic genes.

We summarized the benchmarking for assembled and unassembled RNA-seq data for short plant peptide motifs with the following additions to the manuscript:

Abstract:

p.1, I.34-37: “Nearly 10,000 plant species are represented by RNA-seq datasets in the NCBI sequence read archive (SRA), which are difficult to access in unassembled format for large-scale gene discovery due to database size. Here, we apply an optimized RNA-seq assembly workflow to transform most of this public RNA-seq data to a searchable database for biosynthetic gene discovery.”

Results:

p.3, I.119-136: “Towards large-scale transcriptome mining of plant RiPPs such as stephanotic acid-type burpitides, we wanted to first test if sequencing data assembly is advantageous for peptide discovery from RNA-seq data. We searched unassembled and SPAdes-*de novo* assembled transcriptomes of 890 plant species for the presence of the stephanotic acid core peptide QLLVW (**Data S1**). The target plant species had metabolomic data deposited in plantMASST⁴⁸, a repository of tandem mass spectrometry-based metabolomic data, and which can thus be tested for the presence of the stephanotic acid-QLLVW analyte. The unassembled RNA-seq data were first searched for the presence of the QLLVW motif as nucleotide or protein queries via blastn or seqkit2⁴⁹, respectively. These searches of the stephanotic acid motif yielded too many false positives (98-99%) for effective peptide discovery (**Figure 1E, Data S1**). Next, we searched homologs of moroidin cyclase KjaBURP with a QLLVW motif in both unassembled and assembled data via PHI-BLAST, a protein BLAST with target motif query⁵⁰. Searching assembled data yielded 9 of 11 stephanotic acid-containing plants based on metabolomic data compared to 7 hits in the unassembled data query (**Figure S2**). In addition, the assembled data search had a

lower false discovery rate (36% versus 46%) compared to the unassembled data search. The BLAST database size of assembled transcriptomes was 0.2% of unassembled RNA-seq data and search time and memory was 0.5% and 33%, respectively, with the same computational resource for assembled RNA-seq data (**Figure 1E, Figure S2, Data S1**). Given the smaller database size, lower search cost, and better search results, we focused on applying RNA-seq assembly for plant peptide discovery from short sequence motifs.”

We added the following **Figure 1E** regarding this analysis to the manuscript:

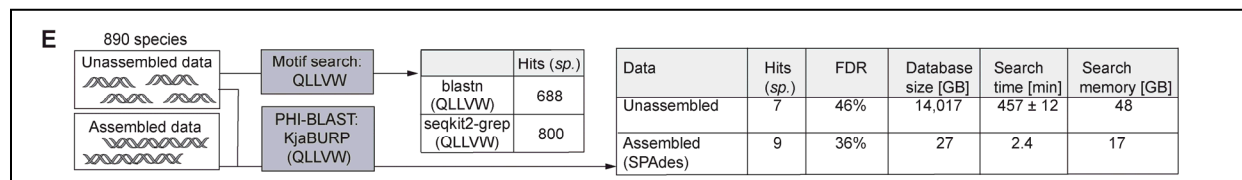


Figure 1 [...] (E) Comparison of unassembled and assembled RNA-seq data search for discovery of stephanotic acid-type burpitides.

To support this analysis, we also added:

- **Data S1 | Benchmarking assembled versus unassembled data search.**
- **Figure S2 | Detection of stephanotic acid (QLLVW) in plantMASST-LCMS2 data and manually analyzed LCMS1 data.**

- what exactly are the advantages of this approach, and how does it compare to alternative approaches?

compare and comment on the use of alternative methods e.g. targeted peptidomics-based discoveries (<https://www.nature.com/articles/s41467-023-39905-4>) or the use of AI/ML ([https://www.cell.com/cell/fulltext/S0092-8674\(24\)00522-1](https://www.cell.com/cell/fulltext/S0092-8674(24)00522-1))

As mentioned above, the main advantage of our approach compared to searching unassembled data is a smaller database size, lower search cost, and better search results and, thus, a higher peptide discovery rate.

The main advantage of our approach compared to the mentioned alternative peptidomic & AI/ML methods (HypoRiPPAtlas and AMPSphere, respectively) is that our approach can use unassembled sequencing data through *de novo* assembly for discovery, whereas HypoRiPPAtlas and AMPSphere require assembled transcriptome, genome or metagenome sequences for peptide discovery. HypoRiPPAtlas requires assembled genomes for microbial RiPP discovery and assembled transcriptomes for burpitide discovery, AMPSphere requires assembled genomes and metagenomes for peptide discovery.

Because our approach can use unassembled data for peptide discovery, it can search more plant species for peptides (7569 species) as it would be possible by searching assembled data such as the 1kp plant transcriptome database (966 species of higher plants, **Data S4**). The 1kp plant transcriptome database consists of plant transcriptomes assembled by a different *de novo* RNA-seq assembler called SOAPdenovo-trans, which enables less plant peptide discovery based on added comparison to our SPAdes-based RNA-seq assembly (**Table S2**).

We added the following sentences regarding this advantage to the manuscript:

Results:

p.4, l.166-173: “This assessment was further realized by a scaled *de novo* transcriptome assembly of the 1kp higher plant RNA-seq data with SPAdes and comparison of moroidin precursor search results via PHI-BLAST with the published datasets assembled by *de novo* transcriptome assembler SOAPdenovo-trans (**Data S4**)^{21,29}. SPAdes-assembled 1kp data enabled identification of six fused moroidin cyclases with twelve different moroidin core peptides in five species, whereas published 1kp data contained only three fused moroidin cyclases with five different core peptides in three species (**Table S3**), showing that *de novo* SPAdes assembly can yield more predicted peptides at scale than existing plant transcriptomes from other assemblers.”

and added **Data S4** and the following **Table S3**:

Table S3 | Comparison of scaled transcriptome mining for fused moroidin cyclases and moroidin core peptides in published 1kp SOAPdenovo-trans assemblies and SPAdes-assemblies.

Assembly	Species	Detected fused moroidin cyclases	Detected moroidin core peptides
SOAPdenovo-trans (1kp study ⁴)	3 (<i>Lomandra longifolia</i> , <i>Bauhinia tomentosa</i> , <i>Kerria japonica</i>)	3	5 (QLLVWREH, QLLVWRSH, QLLFWFGRH, QLLVWFARH, QLLVWRAH)
SPAdes (v3.15.5, this study)	5 (<i>Kerria japonica</i> , <i>Lomandra longifolia</i> , <i>Orchidantha maxillarioides</i> , <i>Bauhinia tomentosa</i> , <i>Cercis canadensis</i>)	6	14 (QLLFWFGRH, QLLFWTRH, QLLVWFGHR, QLLVWPKH, QLLVWQKH, QLLVWRAH, QLLVWREH, QLLVWRGH, QLLVWRH, QLLVWRNH, QLLVWRSH, QLLVWSRH, QLLVWTDH, QLLVWTRH)

- has the approach been validated with other datasets (except SRAs)

Before large-scale transcriptome assembly and mining, we further validated our SPAdes-based *de novo* transcriptome mining approach by using unassembled RNA-seq data from the 1kp database for moroidin discovery as described above. The comparison of search results for fused moroidin cyclases and their moroidin core peptides showed that the SPAdes assemblies resulted in detection of more moroidin cyclases and moroidin core peptides than from the published SOAPdenovo-trans-based assemblies (**Table S2**).

how does the mining approach compare to when using datasets with already assembled contigs, i.e. can you accelerate or extend the number of IDs you obtain, or is the novelty in the way you assemble transcriptomes, irrespective of any previous (automated assemblies)?... in case this is a misunderstanding, it clearly exemplifies the need for a more thorough description of the rationale and the method in itself.

We are grateful for this idea and added a section about comparison of *de novo* assembly to genome-guided assembly for 16 of our benchmarking RNA-seq datasets for which a reference genome is available. We compared the genome-guided assemblers StringTie (v2.2.1) and Trinity (v2.15.1) with our previously tested *de novo* assemblers (SPAdes, MEGAHIT, Trinity). In short, genome-guided assembly via StringTie is faster, assembles more repeats with core peptides and leads to identification of more total core peptides than SPAdes *de novo* assembly. However, SPAdes assemblies include more unique cores, i.e. different peptides, and higher similarity and

identity values for assembled precursor peptides than genome-guided assemblies. Overall, our *de novo* assembly approach has the advantage that it does not require a high-quality annotated genome sequence for assembly and, therefore, it can explore more plant species for peptides given the larger number of RNA-seq datasets (>9000) compared to annotated genome sequence assemblies (414) in NCBI GenBank (**Table S1**, May 2024).

We added the following sections describing this comparison of our *de novo* approach to the genome-guided approach:

Results:

p.4, I.174-183: “An alternative to *de novo* RNA-seq assembly is genome-guided transcriptome assembly which requires an annotated reference genome ⁶⁸. With increasing numbers of high-quality plant genomes complementing available RNA-seq data, we also compared genome-guided transcriptome assembly with StringTie ⁶⁹ or Trinity to *de novo* assemblers regarding precursor peptide assembly (**Table 2, Data S5**). Overall, genome-guided assembly via StringTie outperformed *de novo* SPAdes assembly in time, total detected cores, and assembly of cores in sequence repeats but SPAdes assemblies of precursor peptides were superior to genome-guided assembly by StringTie or Trinity in the number of detected unique cores and similarity and identity scores (**Table 2**). Given the lower number of high-quality plant genomes compared to plant RNA-seq datasets (**Table S1**), we proceeded with a *de novo* assembly strategy with SPAdes.”

And we added the following **Table 2**:

p.24, I.949: Table 2 | Benchmarking results of *de novo* and genome-guided transcriptome assembly for plant RiPP precursor discovery. 16 RNA-seq datasets with corresponding annotated genomes were trimmed and assembled which include 20 precursor genes with 98 core peptides and 38 unique core peptides. Each RNA-seq dataset was assembled in triplicate (n=3). Displayed values are means with one standard deviation shown as an error.

Assembler	Approach	Total time [min]	Total memory usage [GB]	Total detected cores (98 total cores)	Detected unique correct cores (38 total unique cores)	Detected repetitive cores (87 total repetitive cores)	Similarity [sum %] (total similarity score: 3600)	Identity [sum %] (total identity score: 3600)
MEGAHIT (v1.2.9)	<i>De novo</i>	940 ± 44	123 ± 1	42 ± 0	24 ± 0	32 ± 0	1582 ± 0	1587 ± 0
SPAdes (v3.15.5)	<i>De novo</i>	1185 ± 39	180 ± 1	66 ± 0	32 ± 0	56 ± 0	1728 ± 0	1713 ± 0
Trinity (v2.15.1)	<i>De novo</i>	13683 ± 621	485 ± 6	41 ± 0	25 ± 0	30 ± 1	2618 ± 10	2597 ± 7
StringTie (v2.2.1)	Genome-guided	922 ± 621	1635 ± 2	77 ± 0	27 ± 0	65 ± 0	1565 ± 0	1545 ± 0
Trinity (v2.15.1)	Genome-guided	10545 ± 417	1787 ± 4	62 ± 0	28 ± 0	51 ± 0	1795 ± 1	1691 ± 1

and Supplementary **Data S5 - Benchmarking assembly – *de novo* vs genome-guided**.

- what exactly are the benefits of discovery of these diverse peptides with respect to activity, medicinal chemistry/structure-activity and potential translation?
surely this is a critical point, and in its current form is not addressed to satiety: firstly, what's the reference compounds for bioassays (known peptides and current drugs)? Not clear how an IC₅₀ in the μM range can be called "improved bioactivity"; how do these natural variants compare, i.e. since you appear to use a high-content screening unit at 384 well format, it should easily be possible to obtain a larger picture, and how these naturally occurring variants be rationalized to guide SAR and MedChem efforts in the future; s.d. appear to be a more appropriate way of data presentation for these type of assays (doi: 10.4103/2229-3485.100662)

Regarding bioactivity and SAR of the peptides: We introduce the bioactivity and SAR of stephanotic acid-type peptides in the **Introduction** as follows:

p.2/3, I.98-100: "Moroidin and its analog celogentin C exhibit moderate *in vitro* cytotoxicity against non-small cell lung cancer whereas stephanotic acid-[LV] has no such detectable cytotoxicity, indicating that the second ring of moroidin is required for its bioactivity ^{2,18}."

Regarding the improvement of bioactivity: As stated above, the second ring of moroidin appears to determine its bioactivity in H1437 non-small cell lung cancer cell line, which is why we focused on discovery of stephanotic acid-type bicyclic peptides with different second ring structures. Moroidin-QLLVWRNH is 9-fold more cytotoxic in H1437 non-small cell lung cancer compared to moroidin (core: QLLVWRGH, IC₅₀ = 12 μM , Reference ¹⁸) and therefore constitutes an improved bioactivity of the targeted peptide scaffold *in vitro*. To further highlight the bioactivity of moroidin-QLLVWRNH in H1437 cells, we screened the peptide in several additional cancer cell lines (A549 lung adenocarcinoma, HUH7 hepatocellular carcinoma, and LNCaP prostate cancer), with a counterscreen in multiple non-cancerous fibroblast cell lines. Overall, moroidin-QLLVWRNH was most potent in H1437 cells (IC₅₀ = 1.4 μM), and demonstrated no significant dose-responsive cytotoxicity in non-cancerous cell lines. We summarized these additional assays as follows in the manuscript:

Abstract:

p.1/2, I.49-52: "Furthermore, transcriptome mining enabled isolation of a moroidin analog from water chickweed (*Stellaria aquatica*), which exhibits higher cytotoxic effects than moroidin against several cancer cell lines, particularly against H1437 lung adenocarcinoma cells."

Results:

p.7/8, I.342-351: "High-content screening of moroidin-QLLVWRNH against H1437, H1299, and U2OS cancer cell lines revealed significant cytotoxicity in H1437 non-small cell lung adenocarcinoma cells (IC₅₀ = 1.4 μM , **Figure 5A**), exhibiting a five-fold lower IC₅₀ than celogentin C, the most potent previously reported moroidin analog ¹⁸. To evaluate selectivity and bioactivity, the peptide was orthogonally screened in three additional cancer cell lines (A549, LNCaP, and HUH7) and three non-cancerous fibroblast lines (IMR-90, HFF, and MEF, **Figure 5A & 5B**). Moroidin-QLLVWRNH was found inactive in H1299, U2OS, and LNCaP cells, but exhibited moderate dose-responsive cytotoxicity in A549 (IC₅₀ = 12.7 μM) and HUH7 (IC₅₀ = 26.3 μM) cells. Notably, no significant dose-responsive cytotoxicity below IC₅₀ > 40 μM was observed in any non-cancerous fibroblast lines, underscoring its promising cancer cell specificity - an essential factor for clinical translation."

and in bioactivity **Figure 5**:

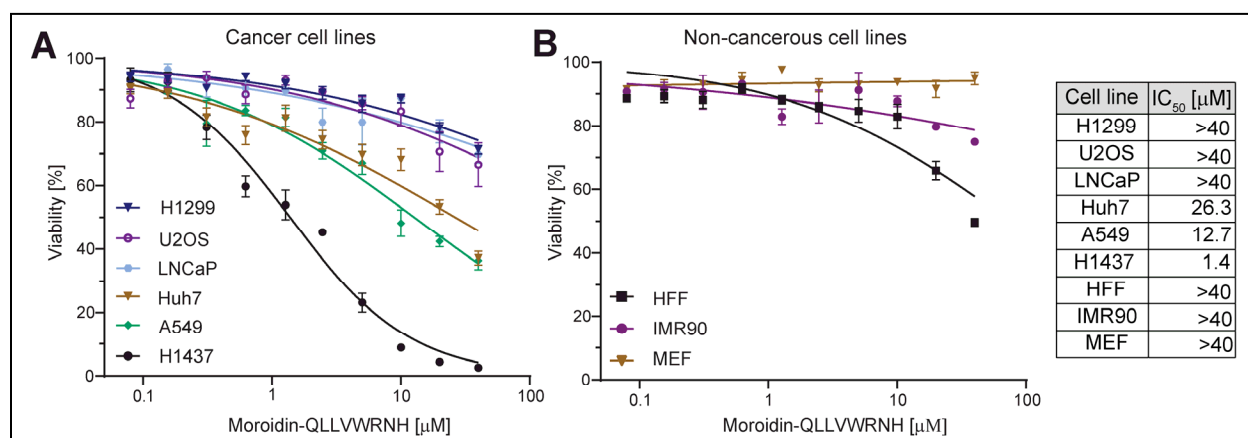


Figure 5 | Bioactivity of moroidin-QLLVWRNH. (A) *In vitro* anticancer assays of moroidin-QLLVWRNH. Cell viability of six cancer cell lines is plotted for the peptide across ten concentrations. The data points represent the mean of triplicate samples and error bars represent one standard deviation of the mean. (B) *In vitro* counterscreen of moroidin-QLLVWRNH in three non-cancerous cell lines is plotted for the peptide across ten concentrations. The data points represent the mean of triplicate samples (n=3) and error bars represent one standard deviation of the mean. Corresponding IC₅₀ values of A and B are listed.

Regarding potential translation and future SAR efforts: While the discovered stephanotic acid-type peptides glechomanin and mercurialin with new second ring crosslinks show no improvement in cancer cell cytotoxicity (**Figure S41**), their discovered cyclases could enable more diversification of the second ring of bioactive moroidin bicycles by changing the Trp-His-crosslink to other crosslinks. This is described in the **Discussion**:

p.8, I.383-386: “Although glechomanin and mercurialin did not exhibit *in vitro* dose-responsive cytotoxicity (**Figure S41**), their characterized cyclases could increase the number of biosynthetically feasible second-ring analogs of moroidin in the future.”

Regarding assay references: As specified in the methods section, DMSO in the negative control matched the highest DMSO volume used in treated cells (40 μM peptide) and taxol (10 μM) served as the positive control:

p.17, I.823-826: “Compounds were normalized to the highest class DMSO volume of 1%, then incubated at 37 °C and 5% CO₂ for 48 h. Each cell line had n=16 positive controls (10 μM taxol) and vehicle (DMSO).”

Regarding data visualization: Data points in the dose-response plots represent the mean of three replicates and error bars now represent one standard deviation of the mean as suggested. We added this information to the corresponding figure legends, for example:

p.23, I.921-926: “**Figure 5 | Bioactivity of moroidin-QLLVWRNH.** (A) *In vitro* anticancer assays of Moroidin-RNH. Cell viability of six cancer cell lines is plotted for the peptide across ten concentrations. The data points represent the mean of triplicate samples and error bars represent one standard deviation of the mean. (B) *In vitro* counterscreen of moroidin-QLLVWRNH in three non-cancerous cell lines is plotted for the peptide across ten concentrations. The data points

represent the mean of triplicate samples (n=3) and error bars represent one standard deviation of the mean. Corresponding IC₅₀ values of **A** and **B** are listed.”

Regarding SAR of natural analogs: We added the following sentence for comparison to the previously highest cytotoxicity by a natural moroidin analog (celogentin C), which was determined by the same high-content screening assay in reference 18:

p.7/8, I.342-345: “High-content screening of moroidin-QLLVWRNH against H1437, H1299, and U2OS cancer cell lines revealed significant cytotoxicity in H1437 non-small cell lung adenocarcinoma cells (IC₅₀ = 1.4 μM, **Figure 5A**), exhibiting a five-fold lower IC₅₀ than celogentin C, the most potent previously reported moroidin analog ¹⁸.

SAR of more natural and unnatural moroidins with diversification in the second ring will be the focus of future studies.

Along this line, the presented Figures 2-4 are somehow repetitive. For instance, these could be combined into two multipanel Figures highlighting one example, i.e. the important discovery, e.g. moroidins, for which diverse peptide have been found. This could be presented as a “discovery” (containing the assemblies, the cloning and gene structure and diversity of the discovered peptides, including the cyclases) and “activity” (containing the isolation/analytcs, production, structure and bioactivity of all these peptides) Figure, by moving parts from the SI in the main Figures, and extending the study (assays) where necessary.

We agree that **Figures 2-4** have a similar structure but we think that combining them in steps of gene discovery and peptide characterization would break with the manuscript structure of highlighting each peptide discovery and its second ring diversification system (glechomanin, mercurialin and moroidin) in a specific result section. We agree that the bioactivity should be highlighted more and generated an additional **Figure 5** with added bioassay data about the *in vitro* cytotoxicity of moroidin-QLLVWRNH towards cancer cells and non-cancerous cells and removed bioactivity data from **Figure 4**.

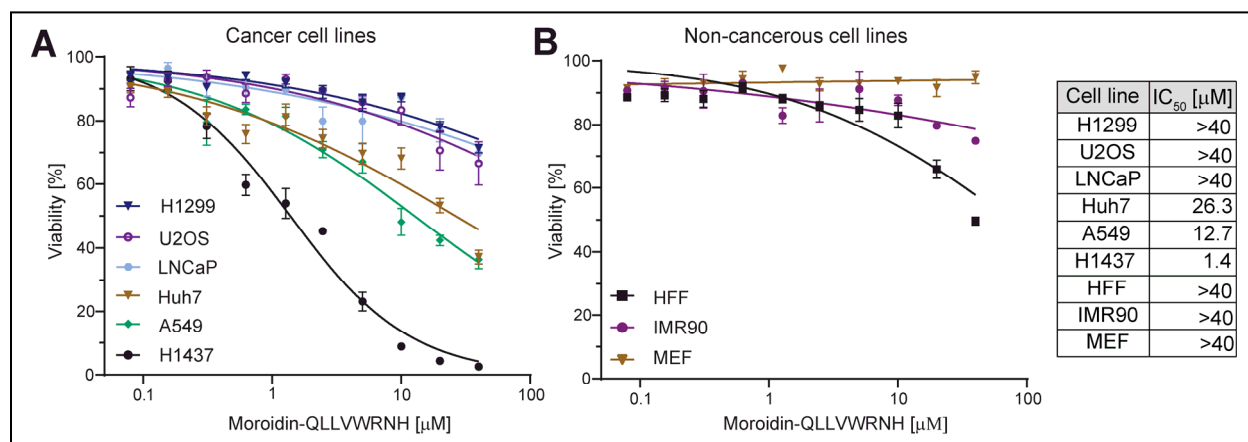


Figure 5 | Bioactivity of moroidin-QLLVWRNH. (A) *In vitro* anticancer assays of moroidin-QLLVWRNH. Cell viability of six cancer cell lines is plotted for the peptide across ten concentrations. The data points represent the mean of triplicate samples and error bars represent one standard deviation of the mean. (B) *In vitro* counterscreen of moroidin-QLLVWRNH in three non-cancerous cell lines is plotted for the peptide across ten concentrations. The data points

represent the mean of triplicate samples and error bars represent one standard deviation of the mean. Corresponding IC₅₀ values of **A** and **B** are listed.

Further, the authors should focus and extend on the actual workflow (Fig. 1) and the comparative study (Table 1) to other assemblies and datasets, while with presenting more numbers... e.g. how many peptides and in how many families were discovered; how does this compare to other mining approaches?

We agree that more comparison to other datasets and mining approaches can improve our manuscript. As stated above, we added:

- (a) a comparison of mining *de novo* assembled data from the 1kp database (SOAPdenovo-trans assembly) compared to our SPAdes-assemblies for moroidin-type core peptides.
 - o The 1kp databases includes 966 higher plant transcriptomes assembled by SOAPdenovo-trans and is included in our SPAdes-assembled transcriptomes. Searching the 1kp data assembled either by SOAPdenovo-trans (1kp reference: ²¹) or SPAdes (this study) for moroidin core peptides showed that more moroidin core peptides are identified in SPAdes assemblies (14) of the same datasets than in the original SOAPdenovo-trans assemblies (5). As stated before, we added the following sentence to the manuscript:

p.4, l.164-173: “Ultimately, our benchmarking identified SPAdes as the *de novo* assembler, which generated precursor peptide transcripts with most unique core peptides at low computational cost and high speed. This assessment was further realized by a scaled *de novo* transcriptome assembly of the 1kp higher plant RNA-seq data with SPAdes and comparison of moroidin precursor search results via PHI-BLAST with the published datasets assembled by *de novo* transcriptome assembler SOAPdenovo-trans (**Data S4**) ^{21,29}. SPAdes-assembled 1kp data enabled identification of six fused moroidin cyclases with twelve different moroidin core peptides in five species, whereas published 1kp data contained only three fused moroidin cyclases with five different core peptides in three species (**Table S3**), showing that *de novo* SPAdes assembly can yield more predicted peptides at scale than existing plant transcriptomes from other assemblers.”

and added **Data S4** and the following **Table S3**:

Table S3 | Comparison of scaled transcriptome mining for fused moroidin cyclases and moroidin core peptides in published 1kp SOAPdenovo-trans assemblies and SPAdes-assemblies.

Assembly	Species	Detected fused moroidin cyclases	Detected moroidin core peptides
SOAPdenovo-trans (1kp study ⁴)	3 (<i>Lomandra longifolia</i> , <i>Bauhinia tomentosa</i> , <i>Kerria japonica</i>)	3	5 (QLLVWREH, QLLVWRSH, QLLFWFGRH, QLLVWFARH, QLLVWRAH)
SPAdes (v3.1.5.5, this study)	5 (<i>Kerria japonica</i> , <i>Lomandra longifolia</i> , <i>Orchidantha maxillarioides</i> , <i>Bauhinia tomentosa</i> , <i>Cercis canadensis</i>)	6	14 (QLLFWFGRH, QLLFWTRH, QLLVWFGRH, QLLVWPKH, QLLVWQKH, QLLVWRAH, QLLVWREH, QLLVWRGH, QLLVWRH, QLLVWRNH, QLLVWRSH, QLLVWSRH, QLLVWTDH, QLLVWTRH)

(b) a comparison of mining RNA-seq data for moroidin core peptides assembled by SPAdes for split versus fused burpitide precursor peptide genes.

- A recent study (reference ¹⁹) mined SPAdes-assembled RNA-seq data of 700 plant species for the discovery of burpitides derived from split precursor peptides, i.e. precursor peptides without a BURP domain. We added the following **Table S8** about the outcomes of this and our study in terms of detected moroidin core peptides and characterized burpitides (biosynthesis and chemistry).
- A recent study (reference ¹⁸) mined SPAdes-assembled RNA-seq data of 793 plant species for the discovery of moroidins derived from fused precursor peptides.

Our study mined about one order of magnitude more plant species for burpitides and could characterize more burpitides with new posttranslational modifications and their underlying burpitide cyclases than the previous studies. We added the following Table and text to the manuscript:

Table S10 | Comparison of burpitide transcriptome mining.

Comparison of RNA-seq studies	Datasets	Detected stephanotic acid core peptides	Characterized new burpitides (NMR, biosynthesis)	Characterized new PTMs (NMR, biosynthesis)
Kersten <i>et al.</i> ³	700	3	1	1
Kruger <i>et al.</i> ⁵	793	5	0	0
This study	8893	55	3	2

Discussion:

p.8, l.372-374: “SPAdes has been applied in transcriptome mining of burpitides from fused and split precursor peptides but at one order of magnitude smaller data scale and with smaller outcome in terms of characterized burpitide chemistry (**Table S10**).”

Other comments:

- In the light of the above said, the title appears too imprecise and should be improved: what is “repository-scale”?... and did you really address the “improved bioactivity”?

We changed the title and result sub-title to “large-scale” instead of “repository-scale”. We keep the term “improved bioactivity” given that we improved the IC₅₀ value of moroidin (QLLVWRGH, IC₅₀ (H1437) = 12 µM, reference ¹⁸) by 9-fold through the discovery of moroidin-QLLVWRNH (IC₅₀ (1437) = 1.4 µM).

- The abstract should be carefully revised to allow the reader to understand that the different classes of peptides are known, but in that the authors present a workflow to identify novel peptides (how many in total, distributed amongst how many phylogenetic clades?)

We revised the abstract to inform about the known chemistry of moroidins and burpitides:

p.1, l.37-41: “We test our transcriptome mining pipeline towards the diversification of moroidins, which are plant ribosomally-synthesized and posttranslationally-modified peptides (RiPPs) called burpitides ¹ biosynthesized from copper-dependent peptide cyclases ^{2,3}. Moroidins are bicyclic structures containing a stephanotic acid scaffold with a Leu-C β -Trp-indole-C6-crosslink and a second, C-terminal ring formed by a Trp-indole-C2-His-imidazole-N1-crosslink.”

The abstract further informs about application of our workflow towards discovery of moroidins and the phylogenetic distribution of peptides discovered by our approach:

p.1/2, l.45-54: “We applied our RNA-seq discovery workflow to identify moroidin analogs with novel second ring structures and the cyclases responsible for their biosynthesis. A search of >27000 candidate burpitide cyclases revealed two stephanotic acid-type burpitides with new second ring crosslinks derived from posttranslational modification: Glechomanin from ground ivy (*Glechoma hederacea*) with a C-terminal tryptophan-crosslink, and mercurialin from annual mercury (*Mercurialis annua*) featuring a C-terminal tyrosine-crosslink. Furthermore, transcriptome mining enabled isolation of a moroidin analog from water chickweed (*Stellaria aquatica*), which exhibits higher cytotoxic effects than moroidin against several cancer cell lines, particularly against H1437 lung adenocarcinoma cells. Our study expands burpitides to grasses, Lowiaceae, mints, pinks, and spurge while demonstrating that large-scale transcriptome mining can broaden the medicinal chemistry toolbox for chemical and biological exploration of eukaryotic RiPP lead structures such as moroidins.”

Reviewer	#2	(Remarks	to	the	Author):
----------	----	----------	----	-----	----------

This manuscript assessed three de novo short-read assemblers for generating precursor peptides from RNAseq datasets of RiPP-producing plants. The most effective assembler was applied to identify moroidin analogs with novel second-ring structures and uncover the cyclases responsible for their biosynthesis. Through repository-scale transcriptome mining, the authors discovered two stephanotic acid-type burpitides with unique second-ring crosslinks and isolated a new moroidin analog with potent in vitro cytotoxicity, showing selective toxicity against H1437 lung adenocarcinoma cells. This work significantly expands the medicinal chemistry toolbox for the chemical and biological exploration of eukaryotic RiPP lead structures, highlighting its scientific importance and application potential.

We thank the reviewer for their evaluation of our manuscript and their suggestions for improvement.

However, the following points warrant further discussion with the authors:

1. The authors have optimized the RNA-seq workflow. Besides improvements in speed and memory usage, what other advantages do these optimizations offer compared to the original workflow?

Regarding memory: The requested memory (48 GB) was optimized to enabled (1) an efficient job distribution on a computational cluster, (2) lowest computational costs at given rates of the cluster per requested memory unit and (3) high success rates of job completion. For example, only 30 of 8893 datasets required higher memory for completion of SPAdes assembly.

Regarding scale: Parallel submission of assembly jobs via SLURM job arrays on a computational cluster has improved our scale to assemble thousands of RNA-seq datasets in several weeks.

[Read trimming](#) of RNA-seq data before assembly has improved core peptide detection and precursor peptide assembly based on SPAdes assemblies. We added the following **Table S2** and the following sentence to highlight this improvement in our workflow.

Table S2 | Benchmarking results of *de novo* transcriptome assembly for plant RiPP precursor discovery. 27 RNA-seq datasets were assembled which include 36 precursor genes with 138 core peptides and 65 unique core peptides. Each RNA-seq dataset was assembled in triplicate (n=3) with one standard deviation shown as an error.

<i>De novo</i> assembler	Total time [min]	Total memory usage [GB]	Total detected cores (138 total cores)	Detected unique correct cores (65 total unique cores)	Detected repetitive cores (120 total repetitive cores)	Similarity [sum %] (total similarity: 3600%)	Identity [sum %] (total identity: 3600%)	Gaps [sum %] (total gaps: 3600%)
SPAdes (v3.15.5)	2041 ± 273	269.9 ± 6.6	88	44	69	2948	2915	609
Trimalore (v0.6.7) + SPAdes (v3.15.5)	(439 + 1628 = 2067) ± 241	(63 + 241 = 304) ± 3	93	48	74	3012	2981	545

Results:

p.4, l.163-164: “Read trimming before assembly resulted in more detected core peptide and more accurate precursor peptide assembly (**Table S2**).”

[Search](#) has improved in our workflow by:

- Sequenceserver updates which enable fast BLAST database generation, tblastn search and hit fasta file download for further core peptide prediction via RepeatFinder.
- Commandline PHI-BLAST search on a computational cluster (as suggested by reviewer 3) which enables scalable search of fused burpitide precursors with a target motif such as the stephanotic acid core (QLxxW, x = any proteinogenic amino acid).

We added all scripts from our workflow of this study to a GitHub repository (<https://github.com/UM-KerstenLab4009/Moroidin-Transcriptomics>).

2. Although the manuscript compares different assembly tools, how accurate are these tools in handling repetitive sequences and short-read sequences? Could assembly errors potentially affect the results?

We tested the accuracy of repetitive sequence assembly in 18 known multi-core precursor peptides with a total of 120 sequence repeats in our benchmarking study and reported the results in **Table 1** in the column:

“Detected repetitive cores (120 total repetitive cores)” (see below)

p.24, I.943: Table 1 | Benchmarking results of *de novo* transcriptome assembly for plant RiPP precursor discovery. 27 RNA-seq datasets were trimmed and assembled which include 36 precursor genes with 138 core peptides and 65 unique core peptides. Each RNA-seq dataset was assembled in triplicate (n=3). Displayed values are means with one standard deviation shown as an error.

<i>De novo</i> assembler	Detected repetitive cores (120 total repetitive cores)
MEGAHIT (v1.2.9)	51 ± 0
SPAdes (v3.15.5)	74 ± 0
Trinity (v2.15.1)	44 ± 1

and in the manuscript as follows:

Results:

p.4, I.160/161: “[...] SPAdes assembled the most repetitive cores (62%), followed by MEGAHIT (43%) and Trinity (37%).”

We also added benchmarking data for genome-guided assembly of repetitive sequences in 10 known multi-core precursor peptides from **Table 1**, which have an annotated reference genome sequenced, and compared the assembly accuracy (**Table 2**) to *de novo* assembly (**Table 1**, **Data S5**) as follows:

p.4, I.178-181: “Overall, genome-guided assembly via StringTie outperformed *de novo* SPAdes assembly in time, total detected cores, and assembly of cores in sequence repeats but SPAdes assemblies of precursor peptides were superior to genome-guided assembly by StringTie or Trinity in the number of detected unique cores and similarity and identity scores (**Table 2**).”

Assembly errors likely affect the *de novo* assembly of repetitive sequences such as in multi-core precursor peptides from NGS data given that the read length can often be similar to the sequence repeat length in a repetitive gene sequence. Genome-guided assembly appears to reduce these errors because Trinity-based assembly of sequence repeats improved by 70% from *de novo* to genome-guided approach with a high-quality genome (30 versus 51 assembled repeats, respectively, **Table 2**). SPAdes (v3.15.5) *de novo* assembly performed best among *de novo* assemblers and better than genome-guided Trinity assembly (**Table 2**, ~10% more assembled repeats) and only worse than genome-guided StringTie assembly (**Table 2**, ~14% less assembled repeats).

3. The improved RNA-seq workflow has been successful in diversifying and enhancing stephanotic acid compounds through transcriptome mining. Is this improved method applicable to the discovery of other types of compounds?

Yes, the workflow can be applied to search for other burpitides and other plant RiPPs.

For mining other burpitides, there are two options in the revised manuscript:

- a. RepeatFinder-based prediction with new motifs: For RepeatFinder search, known motifs can be added for burpitide prediction with new core peptide motifs. We added a description of this step to our GitHub repository of this study under section **“Sequenceserver-based BLAST search and burpitide prediction”** and **4. RepeatFinder**.
- b. PHI-BLAST search with new motifs: Commandline PHI-BLAST search of transcriptome assemblies can be tailored to include new burpitide motifs in the phi_pattern.txt file and a query BURP domain protein sequence including the target motif. We added the description of this step to the GitHub repository of this study under section **“Commandline-PHI-BLAST search and burpitide prediction”** and **“d. PHI-BLAST”**.

<https://github.com/UM-KerstenLab4009/Moroidin-Transcriptomics>

For mining other plant RiPPs, the workflow can be adjusted by using protein queries from **Data S2** (Benchmarking sequences of assembled precursor peptides) corresponding to the target plant RiPP class. These queries can be used either in Sequenceserver or in PHI-BLAST. In PHI-BLAST, a RiPP-class specific motif would need to be defined in the phi_pattern.txt file as described in the GitHub repository above.

We also added a sentence to the **Materials and Methods** section ‘Transcriptome analysis’ regarding extended burpitide discovery in PHI-BLAST:

p.12, I.543-547: “In order to search for other core peptide motifs within BURP domain proteins, the phi_pattern should be changed accordingly and a query BURP-domain protein sequence should be selected which includes the target motif, e.g. for a moroidin (core: QLLVWRGH) PHI-BLAST search, use the phi_pattern QQLLVWRGH and KjaBURP (GenBank ID QIG55799.1) which includes a QLLVWRGH motif in its sequence.”

Addition of an additional Q in front of the stephanotic acid motif (QQLxxW) or moroidin motif (QQLLVWRGH) appeared to be needed for successful commandline PHI-BLAST runs. We also specified this in the GitHub repository of this manuscript under section **“Commandline-PHI-BLAST search and burpitide prediction”** and **“d. PHI-BLAST”**.

4. The manuscript mentions the cytotoxicity of some compounds against non-small cell lung cancer cells. Have these compounds been broadly tested against other types of cancer cells or non-cancerous cells to verify their selectivity and safety?

To provide a more comprehensive assessment of selectivity and bioactivity, we orthogonally screened three additional cancer cell lines (A549, LNCaP, HUH7) and three non-cancerous fibroblast lines (IMR-90, HFF, MEF). The results confirmed that moroidin-QLLVWRNH is most potent in H1437 cell culture, exhibits moderate cytotoxicity in A549 and HUH7 cells, and is inactive in all remaining cancerous and non-cancerous cell lines, supporting our previous statement that the peptide is likely selectively toxic towards cancer cells and in particular non-small cell lung adenocarcinoma cells. The updated cytotoxicity assay data is added as the following text and **Figure 5** to the manuscript:

p.7/8, I.342-351: “High-content screening of moroidin-QLLVWRNH against H1437, H1299, and U2OS cancer cell lines revealed significant cytotoxicity in H1437 non-small cell lung adenocarcinoma cells ($IC_{50} = 1.4 \mu M$, **Figure 5A**), exhibiting a five-fold lower IC_{50} than celogentin C, the most potent previously reported moroidin analog¹⁸. To evaluate selectivity and bioactivity, the peptide was orthogonally screened in three additional cancer cell lines (A549, LNCaP, and HUH7) and three non-cancerous fibroblast lines (IMR-90, HFF, and MEF, **Figure 5A & 5B**). Moroidin-QLLVWRNH was found inactive in H1299, U2OS, and LNCaP cells, but exhibited moderate dose-responsive cytotoxicity in A549 ($IC_{50} = 12.7 \mu M$) and HUH7 ($IC_{50} = 26.3 \mu M$) cells. Notably, no significant cytotoxicity was observed in any non-cancerous fibroblast lines, underscoring its promising cancer cell specificity - an essential factor for clinical translation.”

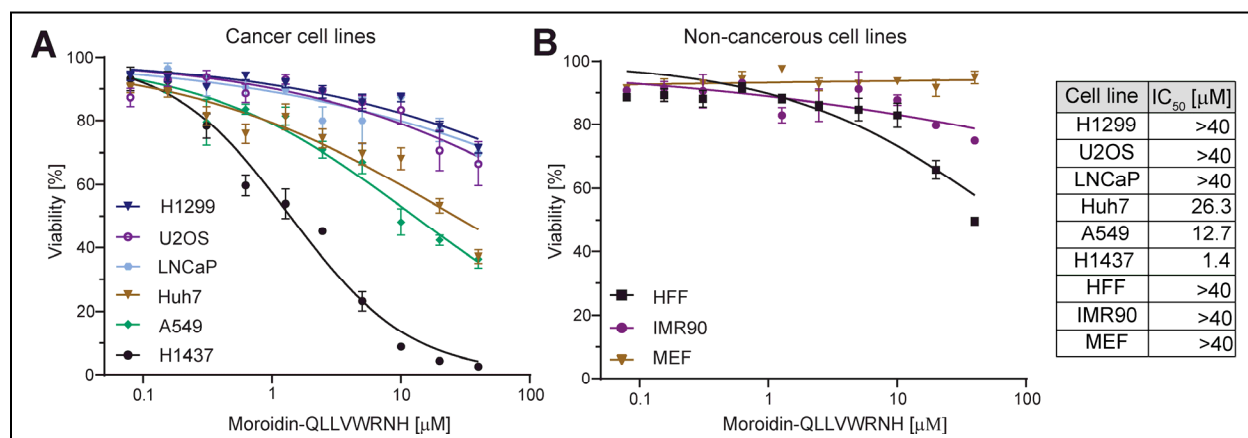


Figure 5 | Bioactivity of moroidin-QLLVWRNH. (A) *In vitro* anticancer assays of moroidin-QLLVWRNH. Cell viability of six cancer cell lines is plotted for the peptide across ten concentrations. The data points represent the mean of triplicate samples and error bars represent one standard deviation of the mean. (B) *In vitro* counterscreen of moroidin-QLLVWRNH in three non-cancerous cell lines is plotted for the peptide across ten concentrations. The data points represent the mean of triplicate samples (n=3) and error bars represent one standard deviation of the mean. Corresponding IC_{50} values of A and B are listed.

5. The manuscript proposes new biosynthetic pathways. Have these pathways been experimentally validated, or are they solely based on bioinformatics predictions?

The biosynthesis for new stephanotic acid-type burpitides glechomanin, mercurialin, and moroidin-QLLVWRNH were experimentally validated *in planta* by transient expression of their respective burpitide cyclase precursors GheBURP (**Figure 2D**), ManBURP (**Figure 3D**) and SaqBURP (**Figure 4D**) in *Nicotiana benthamiana*. In addition, fused glechomanin cyclase GheBURP (**Figure S17**) and fused mercurialin ManBURP (**Figure S29**) were each reconstituted *in vitro* and proteolytically converted to their corresponding burpitides. The proteases of the biosynthetic pathways of glechomanin, mercurialin, and moroidin-QLLVWRNH were not identified in this study. Biosynthetic experiments were described as follows:

Results:

p. 5, I. 223-226: “Transient expression of *GheBURP* in *Nicotiana benthamiana* via the pEAQ-HT expression vector ⁷² through *Agrobacterium tumefaciens* LBA4404 infiltration yielded the same QLFVWGW analyte in methanolic leaf extract of transgenic tobacco six days after infiltration (**Figure 2D**).”

p.6, I.270-279: “Incubation of GheBURP-1xQLFVWGW with 1 mM cupric chloride at pH 7 for 24 h and subsequent tryptic digest resulted in the detection of a tryptic core peptide species with a loss of four hydrogens in the peptide motif QLFVWGWK (**Figures S18**). The characterized MS/MS fragmentation of these tryptic core peptides indicated the formation of two macrocyclization bonds in the QLFVWGW core peptide (**Figure S18**). To confirm the mass loss of four hydrogens in the core peptide as the crosslinks of the glechomanin structure, the tryptic core peptide species was purified by HPLC from a scaled GheBURP-1xQLFVWGW enzyme assay with cupric chloride followed by trypsin proteolysis. The purified modified GheBURP peptide was then incubated with an aminopeptidase and a carboxypeptidase for 24 h. LC-MS analysis of the exopeptidase assay detected an analyte matching glechomanin (**Figure S18**) which shows that the mass loss observed after GheBURP enzyme assay represents the bicyclic bonds of glechomanin.”

p.6/7, I.289-294: “Transient expression of a full-length ManBURP in *N. benthamiana* yielded an analyte with the mass 1038.483 Da, which matched the core peptide sequence QLFFWRY after loss of four hydrogens corresponding to two putative macrocyclizations and loss of ammonia due to N-terminal pyroglutamate formation (**Figure S19**). The same analyte was detectable in an herbarium sample of *M. annua* (**Figures 3B & 3C**) but we could not obtain this source plant in sufficient quantities for peptide isolation.”

p.7, I.315-325: “A truncated construct of mercurialin cyclase ManBURP with one core of QLFFWRY was expressed and purified like glechomanin cyclase construct GheBURP-1xQLFVWGW (**Figure S29**). Bottom-up proteomic analysis of the ManBURP-1xQLFFWRY with LysC digestion identified an analyte with a mass of 4123.068 Da, corresponding to the sequence IPQNYNFQPEPNQLFFWRYQNRGNDIIRLPSMK, which includes the mercurialin core peptide (**Figure S30**). Incubation of ManBURP-1xQLFFWRY with 1 mM cupric chloride at pH 7 for 24 h, followed by LysC digestion, resulted in a loss of four hydrogens in the detected LysC-proteolytic core peptide (**Figure S30**). MS/MS fragmentation confirmed two macrocyclic bonds in the QLFFWRY core peptide (**Figure S30**). To verify the 4H loss as crosslinks in the mercurialin structure, the LysC-proteolytic core peptide was purified by HPLC and incubated with aminopeptidase and carboxypeptidase for 24 h. LC-MS detected a product consistent with mercurialin (**Figure S30**), confirming the 4H mass loss reflects the two macrocyclic crosslinks of mercurialin.”

p.8, I.339/340: “Moroidin-QLLVWRNH was purified (32 mg) and characterized by LC-MS/MS, NMR and transient tobacco expression of *Stellaria aquatica* burpitide cyclase SaqBURP[...].”

6. Had the authors built a flowchart or document of the de novo NGS assembler method to make it easier for readers to use?

Yes, we added a flowchart of the applied RNA-seq workflow for readability to **Figure 1F**.

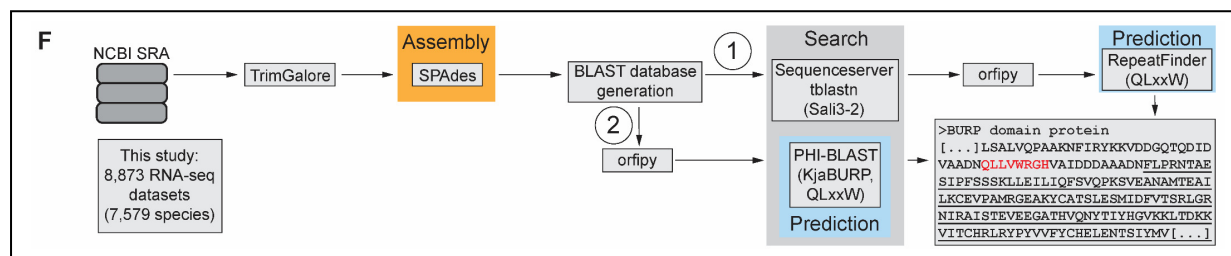


Figure 1 | Transcriptome mining of plant RiPPs. [...] (F) Transcriptome mining workflow of this study with either (1) Sequenceserver-based blast search or (2) commandline PHI-BLAST search.

7. In Figure 4F, the error bar of Moroidin-QLLVWRNH dose-response assay was a little high.

The H1437 screen was repeated with improved technical precision and computational analysis, utilizing CellPose instead of CellProfiler for enhanced cell segmentation and more accurate nucleus count. The revised **Figure 5** and **Materials and Methods** section “**Bioassay**” reflect these modifications.

Reviewer #3 (Remarks to the Author):

The NCBI sequence read archive (SRA) has the RNA-seq datasets of nearly 10,000 plant species, which proved a valuable resource for identifying new functional genes. In this paper, the authors developed a workflow to mine the new moroidin analogs from RNA-seq data in SRA. The first step of their workflow is to de novo assemble short NGS RNA-seqs reads into longer transcripts as most of the plants do not have reference genomes yet. In order to identify a suitable de novo NGS assembler for large-scale plant transcriptome assembly, the authors evaluated the performance of three assemblers: Trinity, SPAdes, and MEGAHIT. And the results showed that SPAdes is the best assembler for the task. By leveraging SPAdes, the authors identified a diverse set of moroidin cyclases and potential new burpitide chemotypes. These findings open avenues for the bioengineering of bicyclic RiPPs with enhanced therapeutic potential.

We thank the reviewer for their evaluation of our manuscript and their suggestions to improve it.

Comments:

1. The authors used 138 plant RiPP precursor genes to evaluate the de novo assemblers. It is unclear how many of them come from the genome with reference sequences. Could authors evaluate the performance of de novo assembler using RNA-seq datasets with reference genome? and comparing the de novo assembler methods with the reference genome-based methods.

Yes, we added a section about comparison of *de novo* assembly to genome-guided assembly for 16 of our benchmarking datasets for which an annotated reference genome is available. We compare the genome-guided assemblers StringTie and Trinity with our previously tested *de novo* assemblers (SPAdes, MEGAHIT, and Trinity). We added the following sections describing the results of this comparison:

Results:

p.4, I.174-183: “An alternative to *de novo* RNA-seq assembly is genome-guided transcriptome assembly which requires an annotated reference genome⁶⁸. With increasing numbers of high-quality plant genomes complementing available RNA-seq data, we also compared genome-guided transcriptome assembly with StringTie⁶⁹ or Trinity to *de novo* assemblers regarding precursor peptide assembly (**Table 2, Data S5**). Overall, genome-guided assembly via StringTie outperformed *de novo* SPAdes assembly in time, total detected cores, and assembly of cores in sequence repeats but SPAdes assemblies of precursor peptides were superior to genome-guided assembly by StringTie or Trinity in the number of detected unique cores and similarity and identity scores (**Table 2**). Given the lower number of high-quality plant genomes compared to plant RNA-seq datasets (**Table S1**), we proceeded with a *de novo* assembly strategy with SPAdes.”

and added the following **Table 2**:

p.24, I.949: Table 2 | Benchmarking results of *de novo* and genome-guided transcriptome assembly for plant RiPP precursor discovery. 16 RNA-seq datasets with corresponding annotated genomes were trimmed and assembled which include 20 precursor genes with 98 core peptides and 38 unique core peptides. Each RNA-seq dataset was assembled in triplicate (n=3). Displayed values are means with one standard deviation shown as an error.

Assembler	Approach	Total time [min]	Total memory usage [GB]	Total detected cores (98 total cores)	Detected unique correct cores (38 total unique cores)	Detected repetitive cores (87 total repetitive cores)	Similarity [sum %] (total similarity score: 3600)	Identity [sum %] (total identity score: 3600)
MEGAHIT (v1.2.9)	<i>De novo</i>	940 ± 44	123 ± 1	42 ± 0	24 ± 0	32 ± 0	1582 ± 0	1587 ± 0
SPAdes (v3.15.5)	<i>De novo</i>	1185 ± 39	180 ± 1	66 ± 0	32 ± 0	56 ± 0	1728 ± 0	1713 ± 0
Trinity (v2.15.1)	<i>De novo</i>	13683 ± 621	485 ± 6	41 ± 0	25 ± 0	30 ± 1	2618 ± 10	2597 ± 7
StringTie (v2.2.1)	Genome-guided	922 ± 621	1635 ± 2	77 ± 0	27 ± 0	65 ± 0	1565 ± 0	1545 ± 0
Trinity (v2.15.1)	Genome-guided	10545 ± 417	1787 ± 4	62 ± 0	28 ± 0	51 ± 0	1795 ± 1	1691 ± 1

and Supplementary **Data S5 - Benchmarking assembly - de novo vs genome-guided.**

Our *de novo* assembly approach has the advantage that it does not require a genome sequence for assembly and, therefore, it can explore more plant species for peptides given the larger number of RNA-seq datasets (>9000) in NCBI SRA compared to annotated genome sequence assemblies (414) in NCBI GenBank (**Table S1**).

2. Did the authors evaluate the parameter of tblastn? How will they affect the identification of morindin analogs?

For tblastn search, we kept parameters as default with an E-value of 1e-5 to ensure hits with a BURP domain. We found the suggestion of PHI-BLAST search very helpful and tested different E-values for PHI-BLAST search as a commandline approach with moroidin cyclase KjaBURP as a query and QLxxW as a search pattern, where x is any proteinogenic amino acid.

For default PHI-BLAST search settings (E-value 10), the stephanotic acid core peptide hits were the same as for the Sequenceserver-tblastn search (E-value 1e-5) with Sali3-2 as a BURP-domain protein query (**Table S4 & S5**). Reduction of the PHI-BLAST E-value to 1e-5 resulted in the same identified stephanotic acid core peptides (**Data S7**). PHI-BLAST enabled the detection of more core peptides of stephanotic acid-type burpitides (moroidin, glechomanin, mercurialin, stephanotic acid) from transcripts without or with incomplete BURP-domain sequence (i.e. none of the conserved CH of the 4xCH-motif are present as defined in our previous RepeatFinder analysis). The inclusion of transcripts without a defined BURP-domain sequences resulted in highest additional core peptide hits with PHI-BLAST E-values 0.01-10: 20 moroidin cores, 5 glechomanin cores, 1 mercurialin core, 2 stephanotic acid cores.

Did the authors use other search tools?

We tested seqkit2-grep for the identification of a stephanotic acid core peptide (QLLVW) from unassembled RNA-seq data, which resulted in too many false positives for effective RiPP discovery (**Figure 1E, Data S1**). This result is described in the revised manuscript:

p.3, l.124-127: “The unassembled RNA-seq data were first searched for the presence of the QLLVW motif as nucleotide or protein queries via blastn or seqkit2 ⁴⁹, respectively. These searches of the stephanotic acid motif yielded too many false positives (98-99%) for effective peptide discovery (**Figure 1E, Data S1**).”

For example, PHI-blast, as the moroidin has motifs.

We are grateful for this suggestion. We implemented PHI-BLAST into our RNA-seq workflow as an alternative to Sequenceserver-tblastn-search. PHI-BLAST enabled burpitide discovery without generating databases on Sequenceserver but on the same computational cluster used for RNA-seq data assembly. PHI-BLAST search with moroidin cyclase KjaBURP and target predefined motif QLxxW (x = any proteinogenic amino acid) with default PHI-BLAST search settings yielded

the same core peptide hits of stephanotic acid-type burpitides from fused burpitide cyclases as Sequenceserver-tblastn (Table S4 & S5). The addition of PHI-BLAST is shown in Figure 1F:

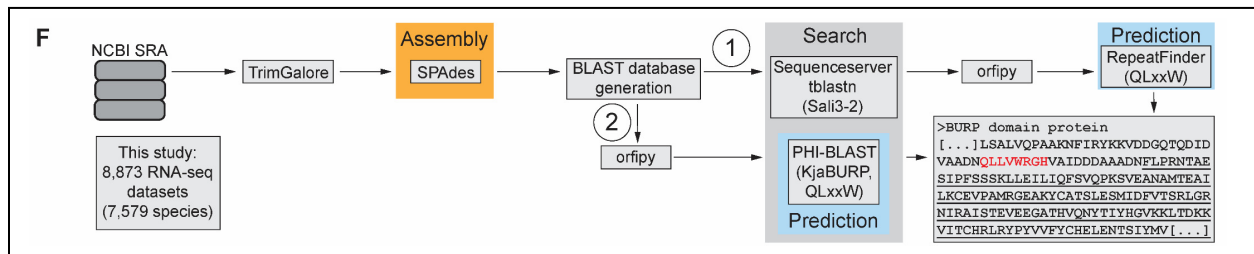


Figure 1 | Transcriptome mining of plant RiPPs. [...] (F) Transcriptome mining workflow of this study with either (1) Sequenceserver-based tblastn search or (2) commandline PHI-BLAST search.

We added the following text in the manuscript about PHI-BLAST search:

Results:

p.5, l.199-205: “In addition to Sequenceserver-tblastn search, we also applied PHI-BLAST search with moroidin cyclase KjaBURP and the stephanotic acid core QLxxW as a search motif. PHI-BLAST search yielded the same core peptide hits as Sequenceserver-tblastn search but PHI-BLAST resulted in 28 additional predicted stephanotic acid-type core peptides from transcripts without or with incomplete BURP-domain sequence (Table S4 & S5). Given the focus on fused burpitide cyclases in this study, we excluded these additional core peptides in subsequent RiPP discovery experiments.”

Methods:

p.11/12, l.533-543: “For commandline BLAST search, all SPAdes *de novo* assemblies (Data S3) were combined into a fasta file and then translated into protein sequences via orfipy (v0.0.4)⁹² with the following parameters: --pep --min 450 --between-stops. The resulting pep file was searched via PSI-BLAST (BLAST+ v2.16.0+) with phi_pattern QQL-(x)2-W and KjaBURP (GenBank ID QIG55799) on default settings (E-value 10). The resulting hits were analyzed by RepeatFinder as before. In our searches, a BURP-domain motif was defined by the presence of at least half the BURP-domain-conserved residues (i.e. FF/Fx-C-C-CH of FF/Fx-C-C-CH-CH-CH)³⁷. Hits with a single QLxxW core peptide at the first N-terminal α -helix motif of BURP domain proteins⁷³ were removed. PHI-BLAST results (Table S4) also included predicted stephanotic acid-type core peptides from transcripts without or incomplete BURP domain sequence. Reduction of the PHI-BLAST E-value 10 to 1e-5 had no effect on search results from fused burpitide cyclases but it reduced hits from incomplete or non-BURP domain transcripts (Data S7).”

and added all scripts on a [GitHub repository](#):

<https://github.com/UM-KerstenLab4009/Moroidin-Transcriptomics>

3. All numbers should be in scientific format, for example, 10000 to 10,000; 27000 to 27,000.

We updated all numbers to scientific format.

RESPONSE TO REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

All earlier comments have been addressed to full satisfaction. Great paper!

Reviewer #2 (Remarks to the Author):

This manuscript has been greatly improved and I think it has met the publication standards.

Reviewer #3 (Remarks to the Author):

The authors have addressed all my concerns. Thank you and congratulations!

[We thank the reviewers for their positive evaluation of our revised manuscript.](#)