Journal of Visualized Experiments

Mass Spectrometry-Guided Genome Mining as a Tool to Uncover Novel Natural Products

--Manuscript Draft--

Article Type:	Methods Article - JoVE Produced Video
Manuscript Number:	JoVE60825R2
Full Title:	Mass Spectrometry-Guided Genome Mining as a Tool to Uncover Novel Natural Products
Section/Category:	JoVE Chemistry
Keywords:	Genome Mining; Molecular Networking; MS-Guided Genome Mining; natural products; Streptomyces; Target Molecular Network; whole genome sequencing
Corresponding Author:	Luciana de Oliveira Universidade Estadual de Campinas Campinas, São Paulo BRAZIL
Corresponding Author's Institution:	Universidade Estadual de Campinas
Corresponding Author E-Mail:	lucianag@unicamp.br
Order of Authors:	Renata Sigrist
	Bruno Sacchetto Paulo
	Celio Fernando Figueiredo Angolini
	Luciana de Oliveira
Additional Information:	
Question	**Response**
Please indicate whether this article will be Standard Access or Open Access.	Standard Access (US$2,400)
Please indicate the **city, state/province, and country** where this article will be **filmed**. Please do not use abbreviations.	Campinas and São Paulo Cities, São Paulo State, Brazil

Prof. Dr. Luciana G. de Oliveira,
Department of Organic Chemistry
Institute of Chemistry
University of Campinas - UNICAMP
Caixa Postal 6154 – CEP 13083-970
Fone: +55 19 35210193
FAX: +55 19 35213023

University of Campinas, 26th of September 2019

Dear Dr. Benjamin Werth,
Science Editor - Chemistry | Biochemistry JoVE

Please, find enclosed the manuscript entitled "**MS-Guided Genome Mining as Tool to Uncover New Natural Products**", authored by Renata Sigrist, Bruno S. Paulo, Célio F. F. Angolini and Luciana G. de Oliveira, and submitted to be considered for publication in *JoVE* as a Methods Article – JoVEProduced Video.

Useful and fast methodologies to perform a wide-ranging evaluation of natural products are often desired and, since the introduction of genome mining as a powerful tool to promote dereplication of secondary metabolites, research in Natural Product has reached an ultimate level. "MS-Guided Genome Mining" particularly is very advantageous to target new molecules backbones or even to harvest a metabolic profile in order to identify analogues from already known compounds. In this article, we describe a protocol that provides the correlation of Genome Mining and Molecular Network to direct the discovery of new natural entities from crude metabolic extracts in which the main components are previously elected, and structurally related new candidates can also be associated with the genome sequence annotation.

All authors have contributed equally to the elaboration of this manuscript and we don't have any competing and/or relevant financial interest to disclose. It is also important to mention that this document is not under consideration by any other journal.

We are looking forward to hearing from you soon.

Sincerely yours,

Prof. Luciana Gonzaga de Oliveira

TITLE:

Mass Spectrometry-Guided Genome Mining as a Tool to Uncover Novel Natural Products

AUTHORS AND AFFILIATIONS:

Renata Sigrist[1], Bruno S. Paulo[1], Célio F. F. Angolini[2], Luciana G. De Oliveira[1]

[1]Department of Organic Chemistry, University of Campinas (UNICAMP), Campinas, Sao Paulo, Brazil

[2]Center for Natural and Human Sciences, Federal University of ABC (UFABC), Santo André, Sao Paulo, Brazil

All authors contributed equally to this work.

Corresponding Author:

Luciana G. de Oliveira (lucianag@unicamp.br)

Email Addresses of Co-Authors:

Renata Sigrist (renata.sigrist@lnbio.cnpem.br)

Bruno S. Paulo (bruno.paulo@iqm.unicamp.br)

Célio F. F. Angolini (celio.fernando@ufabc.edu.br)

KEYWORDS:

genome mining, molecular networking, mass spectrometry-guided genome mining, natural products, *Streptomyces*, target molecular network, whole genome sequencing

SUMMARY:

A mass spectrometry-guided genome mining protocol is established and described here. It is based on genome sequence information and LC-MS/MS analysis and aims to facilitate identification of molecules from complex microbial and plant extracts.

ABSTRACT:

The chemical makeup of many natural products is immense and widely unrecognized. Therefore, convenient methodologies to perform wide-ranging evaluation of their functions in nature and potential human benefits (e.g., for drug discovery applications) are desired. This protocol describes the combination of genome mining (GM) and molecular networking (MN), two contemporary approaches that match gene cluster-encoded annotations in whole genome sequencing with chemical structure signatures from crude metabolic extracts. This is the first step towards the discovery of new natural entities. These concepts, when applied together, are defined here as MS-guided genome mining. In this method, the main components are previously designated (using MN), and structurally related new candidates are associated with genome sequence annotations (using GM). Combining GM and MN is a profitable strategy to target new molecule backbones or harvest metabolic profiles in order to identify analogues from already known compounds.

INTRODUCTION:

Investigations of secondary metabolism often consist of screening crude extracts for specific biological activities followed by purification, identification, and characterization of the constituents belonging to active fractions. This process has proved to be efficient, promoting the isolation of several chemical entities. However, nowadays this is seen as unfeasible, mainly due to the high rates of rediscovery. As the pharmaceutical industry revolutionized without knowledge of the roles and functions of specialized metabolites, their identification was carried out under laboratory conditions that did not accurately represent nature[1]. Today, there is a better understanding of natural signaling influences, secretion, and the presence of most targets at undetectably low concentrations. Additionally, regulation of the process will help the academic community and pharmaceutical industry to take advantage of this knowledge. It will also benefit research involving the direct isolation of metabolites related to silent biosynthetic gene clusters (BGCs)[2].

In this context, advances in genomic sequencing have renewed interest in screening microorganism metabolites. This is because analyzing the genomic information of uncovered biosynthetic clusters can reveal genes encoding novel compounds not observed or produced under laboratory conditions. Many microbial whole genome projects or drafts are available today, and the number is growing every year, providing massive prospects for uncovering novel bioactive molecules through genome mining[3,4].

The Atlas of Biosynthetic Gene Clusters is the current largest collection of automatically mined gene clusters as a component of the Integrated Microbial Genomes Platform of the Joint Genome Institute (JGI IMG-ABC)[2]. Most recently, the Minimum Information for Biosynthetic Gene Clusters (MIBiG) Standardization Initiative has promoted the manual reannotation of BGCs, providing a highly curated reference dataset[5]. Nowadays, plenty of tools are available to enable computational mining of genetic data and their connection to known secondary metabolites. Different strategies have also been developed to access new bioactive natural products (i.e., heterologous expression, target gene deletion, in vitro reconstitution, genomic sequence, isotope-guided screening [genomisotopic approach], manipulation of local and global regulators, resistance target-based mining, culture independent mining, and, more recently, MS-guided/code approaches[2,6-15]).

Genome mining as a singular strategy requires efforts to annotate a single or small group of molecules; thus, gaps in the process remain in which new compounds are prioritized for isolation and structure elucidation. In principle, these approaches target only one biosynthetic pathway per experiment, thereby resulting in a slow discovery rate. In this sense, using GM along with a molecular networking approach represents an important advance for natural product research[14,15].

The versatility, accuracy, and high sensitivity of liquid chromatography-mass spectrometry (LC-MS) make it a good method for compound identification. Currently, several platforms have invested algorithms and software suites for untargeted metabolomics[16–20]. The core of these programs includes feature detection (peak picking)[21] and peak alignment, which allows

matching of identical features across a batch of samples and searching for patterns. MS pattern-based algorithms[22,23] compare characteristic fragmentation patterns and matching MS^2 similarities generate molecular families sharing structural features. These features can then be highlighted and clustered, conferring the ability to rapidly discover known and unknown molecules from a complex biological extract by tandem MS[2,24,25]. Therefore, tandem MS is a versatile method to gain structural information of several chemotypes contained in a large amount of data simultaneously.

The Global Natural Products Social Molecular Networking (GNPS)[26] algorithm uses the normalized fragment ions intensity to construct multidimensional vectors, in which similarities are compared using a cosine function. The relationship between different parent ions are plotted in a diagram representation, in which each fragmentation is visualized as a node (circles), and the relatedness of each node is defined by an edge (lines). The global visualization of molecules from a single source is defined as a molecular network. Structurally divergent molecules that fragment uniquely will form their own specific cluster or constellation, whereas related molecules cluster together. Clustering chemotypes allows the hypothetical connection of similar structural features to their biosynthetic origins.

Combining both chemotype-to-genotype and genotype-to-chemotype approaches is powerful when creating bioinformatics links between BGCs and their small molecule products[27]. Therefore, MS-guided genome mining is a rapid method and low material-consuming strategy, and it helps bridge parent ions and biosynthetic pathways revealed by WGS of one or more strains under diverse metabolic and environmental conditions.

The workflow of this protocol (**Figure 1**) consists of feeding WGS data into a biosynthetic gene cluster annotation platform such as antiSMASH[28–30]. It helps estimate the variety of compounds and class of compounds encoded by the genome. A strategy to target a biosynthetic gene cluster encoding a chemical entity of interest must be adopted, and culture extracts from a wild type strain and/or heterologous strain containing the BGC can be analyzed to generate clustered ions based on similarities using GNPS[26,31]. Consequently, it is possible to identify new molecules that associate with the targeted BGC and are unavailable in the database (mainly unknown analogues, sometimes produced in low titers). It is relevant to consider that these platforms are contributed to by users and that the availability of bioinformatics and MS/MS data is increasing rapidly. This drives the constant development and upgrade of effective computational tools and algorithms to guide efficient connections of complex extracts with molecules.

[Place Figure 1 here]

This protocol describes a rapid and efficient workflow to combine genome mining and molecular networking as starting point for the natural product discovery pipeline. Although many applications are able to visualize the composition and relatedness of MS-detectable molecules in one network, several are adopted here to visualize structurally similar clustered molecules. Using this strategy, novel cyclodepsipeptide products observed in metabolic extracts

of *Streptomyces* sp. CBMAI 2042 are successfully identified. Guided by genome mining, the whole biosynthetic gene cluster encoding for valinomycins is recognized and cloned into the producer strain *Streptomyces coelicolor* M1146. Finally, following a MS pattern-based molecular networking, the molecules detected by MS are correlated with BGCs responsible for their biogenesis[32].

PROTOCOL:

1. Genome mining for biosynthetic gene clusters

1.1. Perform whole genome sequencing (WGS) as the first step to electing a biosynthetic gene cluster (BCG) for MS-guided genome mining. The whole genome draft of the strain of interest (bacteria) can be obtained by Illumina MiSeq technology using the following with high quality genomic DNA: shotgun TruSeq PCR-Free library prep and Nextera Mate Pair Library Preparation Kit[33].

NOTE: After sequencing, the Illumina shotgun library and Illumina mate pair library can be assembled using the Newbler v3.0 (Roche, 454) assembler program (found at <https://ngs.csr.uky.edu/Newbler>) and annotated using a pipeline based on FgeneSB (found at http://www.softberry.com/berry.phtml?topic=fgenesb_annotator&group=help&subgroup=pipelines>), as described previously[33]. Microbiology Resource Announcements (MRA) is a fully open access journal with articles publishing the availability of any microbiological resource deposited in an available repository (found at <https://mra.asm.org>). The candidate protein-coding genes are identified using the RAST server annotation[34], and the Whole Genome Shotgun (WGS) project is deposited in the DDBJ/ENA/GenBank (found at <https://www.ncbi.nlm.nih.gov/genbank/>) and Gold (found at <https://gold.jgi.doe.gov>) sequence databases.

1.2. To obtain in silico information about secondary metabolism gene clusters annotations from a complete sequenced genome, submit the sequence file (GenBank/EMBL or FASTA format) to an antiSMASH platform (found at <https://antismash.secondarymetabolites.org/>).

1.3. Select the gene cluster of interest from output data (**Figure 2**) based on the most similar known cluster.

NOTE: First, it is routine to explore gene-by-gene and conduct individual searches (blastp) to evaluate which functions are associated with the desired biosynthetic gene groups. This procedure can also help to determine which BGC is likely associated with the production of a desired compound, even if it is a low percentage. An antiSMASH prediction considers all genes within a cluster to make percentage coverage, which can represent a global low percentage of similarity for the aimed BGC. However, when analyzing gene-by-gene, it is possible to obtain more accurate information using the most similar known cluster. Second, antiSMASH has two options to refine a search: 1) detection strictness: the degree of strictness to which the biosynthetic gene cluster must be to be considered a hit. For this option, the user should use

the following parameters: a) strict: detects exclusively well-defined clusters containing all required regions, insusceptible to errors about genetic information; b) relaxed: detects partial clusters missing one or more functional region, which also works for detecting the strict feature; or c) loose: detects poorly defined clusters and clusters that likely match primary metabolites, which can lead to appearance of false positives or poorly defined BGCs. The other option is 2) extra features: the type of information the platform must search for and show in the output. In general, these two options can save time after the prediction. However, the antiSMASH job requires a longer time period.

[Place Figure 2 here]

1.4. Based on DNA sequence information of the BGC, design primers (20–25 nt) flanking the gene cluster for ESAC (*E. coli*/*Streptomyces* Artificial Chromosome) library screening.

NOTE: Different methods[35,36] can be used to capture the whole biosynthetic gene cluster from DNA. Here, the method used is construction of a representative ESAC library[37,38] from *Streptomyces* sp. CBMAI 2042 containing clones with average size fragments of ~95 kb.

2. Heterologous expression of whole biosynthetic gene cluster from the ESAC library

2.1. Move the ESAC vector from *E. coli* DH10B to *E. coli* ET12567 by triparental conjugation[32].

2.1.1. Inoculate *E. coli* ET12567 (CamR), TOPO10/pR9604 (CarbR), DH10B/ESAC4H (AprR) in 5 mL of Luria-Bertani (LB) medium containing chloramphenicol (25 µg/mL), carbenicillin (100 µg/mL), and apramycin (50 µg/mL).

2.1.2. Incubate the culture overnight at 37 °C and 250 rpm.

2.1.3. Inoculate 500 µL of the overnight culture in 10 mL of LB medium containing a half-concentration of antibiotics.

2.1.4. Incubate the culture at 37 °C and 250 rpm until reaching an A_{600} of 0.4–0.6.

2.1.5. Harvest the cells by centrifugation at 2,200 x *g* for 5 min.

2.1.6. Wash the cells 2x with 20 mL of LB medium.

2.1.7. Resuspend the cells in 500 µL of LB medium.

2.1.8. Mix 20 µL of each strain in a microcentrifuge tube and drip into an agar plate with LB medium lacking antibiotics.

2.1.9. Incubate the plates at 37 °C overnight.

2.1.10. Streak the grown cells onto a fresh LB agar plate containing antibiotics and incubate at 37 °C overnight.

3. *Streptomyces*/*E. coli* conjugation

3.1. To obtain the recombinant heterologous organism, perform conjugation[32] between *E. coli* ET12567 containing the ESAC vector, helper plasmid pR9604, and *Streptomyces coelicolor* M1146 or another selected host strain[39].

3.2. Day 1: Inoculate isolated colonies of *S. coelicolor* M1146 in 25 mL of TSBY medium in a 250 mL Erlenmeyer flask fitted with an inox-spring at 30 °C and 200 rpm for 48 h.

3.3. Day 2/3: Inoculate ET12567/ESAC/pR9604 in 5 mL of LB medium containing chloramphenicol (25 µg/mL), carbenicillin (100 µg/mL), and apramycin (50 µg/mL) overnight at 37 °C and 250 rpm.

3.4. Day 3/4: Inoculate 500 µL of the overnight culture in 10 mL of 2TY (in a 50 mL conical tube) containing half-working concentrations of antibiotics. Incubate at 37 °C and 250 rpm until reaching an A_{600} of 0.4–0.6.

3.5. Centrifuge the cultures (ET12567/ESAC/pR9604 and M1146) at 2200 x *g* for 10 min.

3.6. Wash the pellets 2x in 20 mL of 2TY medium and resuspend in 500 µL of 2TY.

3.7. Aliquot 200 µL of the *S. coelicolor* M1146 suspension and dilute in 500 µL of 2TY (suspension A).

3.8. Aliquot 200 µL of suspension A and dilute in 500 µL of 2TY (suspension B).

3.9. Aliquot 200 µL of suspension B and dilute in 500 µL of 2TY (suspension C).

3.10. Aliquot 200 µL of the ET12567/ESAC/pR9604 suspension and mix with 200 µL of suspension C.

3.11. Plate 150 µL of the conjugation mixture on an SFM agar plate lacking antibiotics.

3.12. Incubate at 30 °C for 16 h.

3.13. Cover plates with 1 mL of antibiotic solution (according to plasmid resistance). After drying, incubate at 30 °C for 4–7 days.

NOTE: Here, a solution containing 1.0 mg/mL thiostrepton and 0.5 mg/mL nalidixic acid was prepared.

3.14. Streak putative exconjugants onto SFM agar plates containing thiostrepton (50 mg/mL) and nalidixic acid (25 mg/mL). Incubate at 30 °C.

3.15. Streak exconjugants onto an SFM agar containing only nalidixic acid.

3.16. Perform PCR analysis with isolated colonies to confirm that the entire gene cluster has been transferred to the *S. coelicolor* M1146 host.

4. Strain cultivation

4.1. To obtain the metabolic profile, inoculate 1/100 of the strain's pre-culture in appropriate fermentation media and under the appropriate culture conditions.

4.2. Centrifuge cultures at 2200 x *g* for 10 min.

4.3. Perform the extraction according to the class of the compound of interest[40].

5. Acquiring mass spectra and preparation for GNPS analysis

5.1. To acquire MS/MS data, program suitable HPLC and mass spectrometry methods using the control software. Both high and low resolution data-dependent mass spectrometry analysis (DDA) can be analyzed.

NOTE: Generally, a 1 mg/mL solution of complex crude extract samples is ideal. Dilutions are needed for less complex extracts. It should be noted that MS/MS networking is the detectable molecular network under the given mass spectrometric conditions.

5.2. Convert mass spectra to .mzXML format using MSConvert from Proteowizard (found at <http://proteowizard.sourceforge.net/>). The input parameters for the conversion are illustrated in **Figure 3**. Data from software of almost all companies are compatible.

[Place Figure 3 here]

5.3. Upload the converted LC-MS/MS files into the GNPS database. Two options are available: using a file transfer protocol (FTP) or directly in a browser through the online platform.

NOTE: Detailed information on how to install and transfer data to GNPS is available at <https://ccms-ucsd.github.io/GNPSDocumentation/fileupload/>.

6. GNPS analysis

6.1. After creating an account in GNPS (found at <https://gnps.ucsd.edu/>), log in to the created account select **Create Molecular Network**. Add a job title.

6.2. Basic options: select the mzXML files to perform the molecular network. They can be organized into up to six groups. Select the libraries for the dereplication routine (**Figure 4**).

NOTE: These groups do not interfere with molecular network construction. This information will be used only for the graphical representation.

[Place Figure 4 here]

6.3. Select the precursor ion mass tolerance and fragment ion mass tolerance of 0.02 Da and 0.05 Da, respectively.

NOTE: GNPS has different types of strictness available based on 1) how accurate the MS/MS data is and 2) how accurate the association must be. Basic options: in this folder, it is possible to set **Precursor Ion Mass Tolerance** and **Fragment Ion Mass Tolerance**. These parameters are used as a guide to determine how precise the precursor ion and fragment ion must be. The selected mass tolerances depend on the resolution and accuracy of the mass spectrometer that is used.

6.4. Advanced network options: select the parameters according to **Figure 5**. These parameters directly influence the network cluster size and form. Another parameter in the remaining tabs section are for advanced users; thus, leave the default values.

NOTE: Advanced parameters can be read in GNPS documentation (found at <https://ccms-ucsd.github.io/GNPSDocumentation/>).

[Place Figure 5 here]

6.5. Choose an e-mail address to receive an alert when the work is done, and submit the job.

7. Analysis of GNPS results

7.1. Log in to GNPS. Select **Jobs** > **Published job** > **Done** to open the job. A webpage will open as illustrated in **Figure 6**. All results obtained from molecular networking will be displayed.

7.2. Select **View Spectral Families (In Browser Network Visualizer)** to see all network clusters (red box, **Figure 6**).

[Place Figure 6 here]

7.3. A list will be displayed with all generated molecular networking clusters. If a library search was selected to generate the findings, tentative molecules identification will be displayed in AllIDs. Select **Show** to visualize them.

NOTE: The data analyses can be driven for other results (i.e., genome mining, biological assays,

library dereplication molecules, etc.).

7.4. To analyze the molecular network cluster, select **Visualize Network**.

NOTE: Each cluster is composed of nodes (circles) and edges, which represents molecules and molecular similarity, respectively. Dereplicated molecules will be highlighted as a blue node in the online browser network visualizer.

7.5. In the node labels box, select **parent mass** (red box, **Figure 7**).

7.6. In the edge labels box, select **Cosine** or **DeltaMZ** to observe node similarity or mass difference between nodes, respectively (yellow box, **Figure 7**).

7.7. In the case of multigroup analyses, click **Draw pies** in the node coloring box to observe the frequency at which each node appears in each group (blue box, **Figure 7**).

NOTE: Other choices are possible, but those suggested above are optimal for annotating cluster nodes and unraveling their structures.

[Place Figure 7 here]

7.8. To see all library hits, select **View all library hits** (blue box, **Figure 7**).

NOTE: Also, the MNW can be downloaded in "Direct Cytoscape Preview/Download" (yellow box, **Figure 7**), and the file can be opened in the Cytoscape platform (found at <https://cytoscape.org/>) for more options in graphical structure.

7.9. Manual confirmation of dereplicated compounds and structure elucidation of related compounds are needed. Open the fragmentations spectra directly in the GNPS platform or in original raw files.

REPRESENTATIVE RESULTS:

The protocol was successfully exemplified using a combination of genome mining, heterologous expression, and MS-guided/code approaches to access new specialized valinomycin analogues molecules. The genome-to-molecule workflow for the target, valinomycin (VLM), is represented in **Figure 8**. *Streptomyces* sp. CBMAI 2042 draft genome was analyzed in silico, and the VLM gene cluster was then identified and transferred to a heterologous host. Heterologous and wild type strains were cultivated in proper fermentation conditions, partitioned with ethyl acetate, and concentrated to generate the crude extract. From the product, MS/MS data was acquired to generate a tandem MS metabolite profile for molecular networking. **Figure 9** represents the clustered ions obtained from MS/MS data from *Streptomyces* sp. CBMAI 2042 crude extract, in which characteristic fragmentation patterns and corresponding MS similarities suggest the occurrence of a molecular family sharing structural features[2]. Following known biosynthetic logic and bioinformatics insights, and supported by pattern-based MS/MS spectra, the structure

of four originally reported cyclodepsipeptides were elucidated, and their origins were correlated with the same biosynthetic gene clusters responsible for VLM assembly[32].

Molecular networking data (found at <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=6f97aa4addfa4d20b505fdb4328b088c>) was processed in a GNPS platform and deposited in a MASSIVE repository (MSV000083709). For dereplication, two strategies were selected to populate the network with previously described compounds: 1) Dereplicator (found at <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=1a55e768d02649aaa09d78d0d4778ef3>) and 2) a peptide natural product identification tool called VarQuest (found at Our previous publication provides further details[32].

[Place Figure 8 here]

[Place Figure 9 here]

FIGURE AND TABLE LEGENDS:

Figure 1: Overview of the entire workflow. Shown is an illustration of the bioinformatic, cloning, and molecular networking steps involved in the described MS-guided genome mining approach to identify new metabolites.

Figure 2: Output from antiSMASH platform. Secondary metabolism in silico analysis from whole genome sequence annotation.

Figure 3: **Using MsConvert to convert MS files to mzXML extension.** The correct parameter for GNPS analysis is displayed. The instructions are as follows: add all MS files in box 1 and add the filter **Peak Picking** in box 2; for this filter, use the algorithm vendor; press **start** and the processes of conversion will follow.

Figure 4: **Using online GNPS platform to perform molecular network analysis.** Selection of mzXML files is done by clicking in box 1. In the open dialog box, the files can be selected from personal folder (box 2) or be uploaded in the second tab using the drag-and-drop file uploader (less than 20 MB). The files can be grouped into up to six groups.

Figure 5: **Using GNPS to perform molecular network analysis (advanced options).** Min Pair Cos will directly influence the size of clusters, as high values will result in combining closely-related compounds and low values in combining distantly-related compounds. Using values that are too low should be avoided. Minimum matched fragment ions represent the number of shared fragments between two fragmentation spectra to be linked in the network. Together, both parameters guide the network format; lower values will cluster more distantly-related compounds and vice-versa. Using the proper values will greatly help the compound elucidation.

Figure 6: Using GNPS to visualize molecular network results. All related compound clusters can

be seen in view spectral families (red box). To visualize only library hits, “view all library hits” (blue box) should be selected. For better graphical representation of molecular network results, “Direct Cytoscape Preview” (yellow box) should be downloaded, and the latest version of Cytoscape should be used.

Figure 7: Using GNPS to visualize molecular cluster results. After opening the molecular clusters for better data visualization, the following should be chosen: “Parent mass” as node labels (red box); “DeltaMZ” as edge labels (yellow box); and “Draw pies” as node coloring (blue box). Navigate through the molecular cluster and try to annotate all nodes.

Figure 8: Workflow from in silico genome sequence analysis to MS data acquisition. (A) A draft from *Streptomyces* sp. CBMAI 2042 genome is obtained by Illumina MiSeq sequencing. **(B)** Valinomycin BGC identification and annotation. (**C)** After transferring the whole gene cluster to an appropriate host, the strain is cultivated. The ethyl acetate extract from culture is analyzed by LC to obtain a profile of produced secondary metabolites. The chromatogram shows that valinomycin, montanastatin, and five analogues are produced by VLM BGC expression in a heterologous host.

Figure 9: Molecular networking results. (**A**) Molecular networking from *Streptomyces* sp. CBMAI 2042 extract. Molecular networking ions corresponding to valinomycin, an already known compound with the corresponding BGC annotated in *Streptomyces* sp. CBMAI 2042 genome, are clustered with ions related to analogues firstly described for VLM BGC. (**B**) MS spectra and chemical structures for valinomycin and related analogues are shown.

DISCUSSION:

The strongest advantage of this protocol is its ability to rapidly dereplicate metabolic profiles and bridge genomic information with MS data in order to elucidate the structures of new molecules, especially structural analogues[2]. Based on genomic information, different natural products chemotypes can be investigated, such as polyketides (PK), nonribosomal peptides (NRP), and glycosylated natural products (GNP), as well as cryptic BGCs. Metabolomic screening yields evidence of activated BGC profiles and chemical diversity produced by a specific strain under laboratory conditions. Thus, a BGC can be cloned to direct production of a new compound or unknown analogues related to an already known BGC, facilitated by similarities discovered by molecular networking. Therefore, this procedure helps to distinguish valuable compounds produced by natural sources and can be used as a guide for future isolation steps, which are common in natural product pipelines.

MS-guided genome mining was firstly described in the fields of peptidogenomics[41] and glycogenomics[42]. To estimate the extent of peptide natural product chemical diversity, Dorrestein and colleagues developed an automated method using MS and genomics to visualize the connection between expressed natural products (chemotype) and their gene clusters (genotype). The concept of MS-guided genome mining was then described while using peptide specialized metabolites. Here, a method for the identification of microbial glycosylated natural products (GNP) using a GM approach and tandem MS was applied as tool to rapidly connect

GNP chemotypes (from microbial metabolomes) with their corresponding biosynthetic genotypes following sugar footprints.

The concept of peptidogenomics has been applied to reveal stenothricin gene clusters in *Streptomyces roseosporus,* providing the first insights into the broad utility of GNPS as a platform[43]. Pattern-based genome mining and molecular networking was finally combined with the GNPS platform[26] to facilitate the dereplication of new compounds, known compounds, detection of new analogs, and structure elucidation of 35 *Salinispora* strains. This led to the isolation and characterization of retimycin A, a quinomycin-type depsipeptide[44]. After the introduction of GNPS, integrated metabolomics and genome mining approaches have become the most versatile avenue to connect molecular networks with biosynthetic capabilities[45–50].

This protocol reinforces the feasibility of using genomic and metabolomic analyses to investigate the production of known and unknown chemically analogous compounds in a few steps while consuming low levels of materials. The model presented here is related to valinomycin analogue identification from crude extracts through molecular networking dereplication. The structure of analogues is deduced by MS/MS fragmentation and follows the biosynthetic logic of cloned VLM BGCs.

Different software is available for mining secondary metabolite biosynthetic gene clusters[51] and for metabolite elucidation, but open source options have the advantages because of constant updates, and they are open to the scientific community. In this sense, antiSMASH and the GNPS platform are the most popular choices.

This general procedure can be modified for other extraction methodologies based on the natural source explored. More than one method of extraction can also be combined according to metabolite properties (i.e., polarity, hydrophobicity, the capability to form micelles), and even similar properties, different solvents, or resin can achieve enhanced results. Usually, extracts are prepared from liquid medium cultivation, but there is a plethora of extraction methods available to isolate enriched extracts and screen any biological sample of interest.

When acquiring MS data, data dependent acquisition (DDA) analysis should be used. This issue is important when a larger number of compounds are being evaluated in a single injection. While performing DDA, the maximum number of MS/MS spectra of each precursor ion and maximum number of different precursor ion should be compensated. When using fast scan rate equipment, this can be achieved with higher scan rates (~6–10 MS/MS scans per cycle). However, in lower scan rate equipment, MN performance can be only increased with better chromatographic resolution. The most comprehensive data to populate the molecular networking should be obtained. For MS data acquisition, fixed collision energy is possible, but ramp energies are suitable to yield improved results. There are no optimal conditions that will perfectly work for all samples. Achieving sufficient MS analysis is crucial to the following steps. Henceforth, the molecular network clusters should be generated and dereplicated according to the procedure.

A frequent troubleshooting error is missing intensities for masses. Normally, this can be solved by introducing higher collision energy during analysis. Sometimes, no correlations are observed between the spectra and GNPS library, which is very uncommon. In this case, ensure that the folder opens properly in the previsualization MS software as errors can sometimes be created during the conversion step to .mzXML files.

Regarding genome mining, the most precise output from gene cluster annotation platforms will be provided for higher quality whole genome sequencing for both, single strain, or culture independent mining. High quality sequencing will generate high quality bioinformatic insights for dereplication of biosynthetic pathways. In contrast, although BGC prediction bioinformatics software has been rapidly developing, exact predictions of gene function and putative products is still difficult, especially when investigating novel biosynthetic pathways and features that cannot be predicted in silico. Also, some biosynthetic machinery is strikingly conserved, while enzymology that is involved in hybrid systems, *trans*-AT modular PKs, and NRPSs are recognized as exceptions of the colinearity rule. In this sense, heterologous expression and refinements in bioinformatic output software can help elucidate unpredictable enzyme functions and unusual biochemistry[52–54]. The enrichment of public databases will lead to more precise predictions and discovery of novel specialized metabolites, as the cost for WGS does not represent the handicap for genome mining.

Finally, the strongest advantages of integrated metabolomic and genome mining approaches are related to their feasibility to perform genotype and chemotype dereplication via automated and high throughput analysis connecting genomic, transcriptomic, and metabolomic data to efficiently connect genes with molecules.

ACKNOWLEDGMENTS:

The financial support for this study was provided by São Paulo Research Foundation - FAPESP (2019/10564-5, 2014/12727-5 and 2014/50249-8 to L.G.O; 2013/12598-8 and 2015/01013-4 to R.S.; and 2019/08853-9 to C.F.F.A). B.S.P, C.F.F.A., and L.G.O. received fellowships from the National Council for Scientific and Technological Development - CNPq (205729/2018-5, 162191/2015-4, and 313492/2017-4). L.G.O. is also grateful for the grant support provided by the program For Women in Science (2008, Brazilian Edition). All authors acknowledge CAPES (Coordination for the Improvement of Higher Education Personnel) for supporting the post-graduation programs in Brazil.

DISCLOSURES:

The authors have nothing to disclose.

REFERENCES:

1. Davies, J. Specialized microbial metabolites: functions and origins. *The Journal of Antibiotics (Tokyo).* **66** (7), 361–364 (2013).
2. Ziemert, N., Alanjary, M., Weber, T. The evolution of genome mining in microbes – a review. *Natural Product Reports.* **33** (8), 988–1005 (2016).

3. Zerikly, M., Challis, G. L. Strategies for the Discovery of New Natural Products by Genome Mining. *ChemBioChem* **10** (4), 625–633 (2009).
4. Gomez-Escribano, J. P., Bibb, M. J. Heterologous expression of natural product biosynthetic gene clusters in *Streptomyces coelicolor*: from genome mining to manipulation of biosynthetic pathways. *Journal of Industrial Microbiology & Biotechnology.* **41** (2), 425–431 (2014).
5. Medema, M. H. et al. Minimum Information about a Biosynthetic Gene cluster. *Nature Chemical Biology.* **11** (9), 625–631 (2015).
6. Lautru, S., Deeth, R. J., Bailey, L. M., Challis, G. L. Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. *Nature Chemical Biology.* **1** (5), 265–269 (2005).
7. Chiang, Y.-M. et al. Molecular Genetic Mining of the *Aspergillus* Secondary Metabolome: Discovery of the Emericellamide Biosynthetic Pathway. *Chemistry & Biology.* **15** (6), 527–532 (2008).
8. Huang, T. et al. Identification and Characterization of the Pyridomycin Biosynthetic Gene Cluster of *Streptomyces pyridomyceticus* NRRL B-2517. *Journal of Biological Chemistry.* **286** (23), 20648–20657 (2011).
9. Udwary, D. W. et al. Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. *Proceedings of the National Academy of Sciences.* **104** (25), 10376–10381 (2007).
10. Gross, H. et al. The Genomisotopic Approach: A Systematic Method to Isolate Products of Orphan Biosynthetic Gene Clusters. *Chemistry & Biology.* **14** (1), 53–63 (2007).
11. Spohn, M., Wohlleben, W., Stegmann, E. Elucidation of the zinc-dependent regulation in *Amycolatopsis japonicum* enabled the identification of the ethylenediamine-disuccinate ([S,S]-EDDS) genes. *Environmental Microbiology.* **18** (4), 1249–1263 (2016).
12. Thaker, M. N., Waglechner, N., Wright, G. D. Antibiotic resistance–mediated isolation of scaffold-specific natural product producers. *Nature Protocols.* **9** (6), 1469–1479 (2014).
13. Katz, M., Hover, B. M., Brady, S. F. Culture-independent discovery of natural products from soil metagenomes. *Journal of Industrial Microbiology & Biotechnology*. **43**, 129–141 (2016).
14. Quinn, R. A. et al. Molecular Networking as a Drug Discovery, Drug Metabolism, and Precision Medicine Strategy. *Trends in Pharmacological Sciences.* **38** (2), 143–154 (2017).
15. Yang, J. Y. et al. Molecular Networking as a Dereplication Strategy. *Journal of Natural Products.* **76** (9), 1686–1699 (2013).
16. Lommen, A. MetAlign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing. *Analytical Chemistry.* **81** (8), 3079–3086 (2009).
17. Katajamaa, M., Miettinen, J., Oresic, M. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* **22** (5), 634–636 (2006).
18. Pluskal, T., Castillo, S., Villar-Briones, A., Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11** (1), 395 (2010).
19. Tautenhahn, R., Patti, G. J., Rinehart, D., Siuzdak, G. XCMS Online: A Web-Based Platform

to Process Untargeted Metabolomic Data. *Analytical Chemistry.* **84** (11), 5035–5039 (2012).
20. Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R., Neumann, S. CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Analytical Chemistry.* **84** (1), 283–289 (2012).
21. Katajamaa, M., Orešič, M. Data processing for mass spectrometry-based metabolomics. *Journal of Chromatography A.* **1158**, 318–328 (2007).
22. Liu, W.-T. et al. Interpretation of Tandem Mass Spectra Obtained from Cyclic Nonribosomal Peptides. *Analytical Chemistry.* **81** (11), 4200–4209 (2009).
23. Ng, J. et al. Dereplication and de novo sequencing of nonribosomal peptides. *Nature Methods.* **6** (8), 596–599 (2009).
24. Liaw, C. et al. Vitroprocines, new antibiotics against *Acinetobacter baumannii*, discovered from marine *Vibrio* sp. QWI-06 using mass-spectrometry-based metabolomics approach. *Scientific Reports.* **5** (1), 1–11 (2015).
25. Kang, K. Bin et al. Targeted Isolation of Neuroprotective Dicoumaroyl Neolignans and Lignans from *Sageretia theezans* Using in Silico Molecular Network Annotation Propagation-Based Dereplication. *Journal of Natural Products.* **81** (8), 1819–1828 (2018).
26. Wang, M. et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology.* **34** (8), 828–837 (2016).
27. Doroghazi, J. R. et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nature Chemical Biology.* **10** (11), 963–968 (2014).
28. Medema, M. H. et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research.* **39**, W339–W346 (2011).
29. Weber, T. et al. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Research.* **43**, W237–W243 (2015).
30. Blin, K. et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research.***47**, W81–W87 (2019).
31. Watrous, J. et al. Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences.* **109** (26), E1743–E1752 (2012).
32. Paulo, B. S., Sigrist, R., Angolini, C. F. F., De Oliveira, L. G. New Cyclodepsipeptide Derivatives Revealed by Genome Mining and Molecular Networking. *ChemistrySelect* **4** (27), 7785–7790 (2019).
33. Gonzaga de Oliveira, L., Sigrist, R., Sachetto Paulo, B., Samborskyy, M. Whole-Genome Sequence of the Endophytic *Streptomyces* sp. Strain CBMAI 2042, Isolated from *Citrus sinensis*. *Microbiology Resource Announcements.* **8** (2), 1–2 (2019).
34. Aziz, R. K. et al. The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics.* **9** (1), 75 (2008).
35. Nah, H.-J., Pyeon, H.-R., Kang, S.-H., Choi, S.-S., Kim, E.-S. Cloning and Heterologous Expression of a Large-sized Natural Product Biosynthetic Gene Cluster in *Streptomyces* Species. *Frontiers in Microbiology.* **8**, 1–10 (2017).
36. Zhang, J. J., Tang, X., Moore, B. S. Genetic platforms for heterologous expression of

microbial natural products. *Natural Product Reports.* **36** (9), 1313–1332 (2019).
37. Alduina, R. et al. Artificial chromosome libraries of *Streptomyces coelicolor* A3(2) and *Planobispora rosea*. *FEMS Microbiology Letters.* **218** (1), 181–186 (2003).
38. Jones, A. C. et al. Phage P1-Derived Artificial Chromosomes Facilitate Heterologous Expression of the FK506 Gene Cluster. *PLoS One.* **8** (7), e69319 (2013).
39. Gomez-Escribano, J. P., Bibb, M. J. Engineering *Streptomyces coelicolor* for heterologous expression of secondary metabolite gene clusters. *Microbial Biotechnology* **4** (2), 207–215 (2011).
40. Cannell, R. J. P. *Natural Products Isolation*. Humana Press. Totowa, NJ (1998).
41. Kersten, R. D. et al. A mass spectrometry–guided genome mining approach for natural product peptidogenomics. *Nature Chemical Biology.* **7** (11), 794–802 (2011).
42. Kersten, R. D. et al. Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. *Proceedings of the National Academy of Sciences.* **110** (47), E4407–E4416 (2013).
43. Liu, W. et al. MS/MS-based networking and peptidogenomics guided genome mining revealed the stenothricin gene cluster in Streptomyces roseosporus *The Journal of Antibiotics (Tokyo).* **67** (1), 99–104 (2014).
44. Duncan, K. R. et al. Molecular Networking and Pattern-Based Genome Mining Improves Discovery of Biosynthetic Gene Clusters and their Products from *Salinispora* Species. *Chemistry & Biology.* **22** (4), 460–471 (2015).
45. Cao, L. et al. MetaMiner: A Scalable Peptidogenomics Approach for Discovery of Ribosomal Peptide Natural Products with Blind Modifications from Microbial Communities. *Cell Systems.* (2019).
46. Chen, L.-Y., Cui, H.-T., Su, C., Bai, F.-W., Zhao, X.-Q. Analysis of the complete genome sequence of a marine-derived strain *Streptomyces* sp. S063 CGMCC 14582 reveals its biosynthetic potential to produce novel anti-complement agents and peptides. *PeerJ.* **7** (1), e6122 (2019).
47. Kim Tiam, S. et al. Insights into the Diversity of Secondary Metabolites of *Planktothrix* Using a Biphasic Approach Combining Global Genomics and Metabolomics. *Toxins.* 11 (9), 498 (2019).
48. Özakin, S., Ince, E. Genome and metabolome mining of marine obligate *Salinispora* strains to discover new natural products. *Turkish Journal of Biology.* **43** (1), 28–36 (2019).
49. Trivella, D. B. B., de Felicio, R. The Tripod for Bacterial Natural Product Discovery: Genome Mining, Silent Pathway Induction, and Mass Spectrometry-Based Molecular Networking. *mSystems.* **3** (2), e00160-17 (2018).
50. Maansson, M. et al. An Integrated Metabolomic and Genomic Mining Workflow To Uncover the Biosynthetic Potential of Bacteria. *mSystems.* **1** (3), 1–14 (2016).
51. Blin, K., Kim, H. U., Medema, M. H., Weber, T. Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Briefings in Bioinformatics.* **20** (4), 1103–1113 (2019).
52. Fisch, K. M. Biosynthesis of natural products by microbial iterative hybrid PKS–NRPS. *RSC Advances.* **3** (40), 18228–18247 (2013).
53. Tatsuno, S., Arakawa, K., Kinashi, H. Analysis of Modular-iterative Mixed Biosynthesis of Lankacidin by Heterologous Expression and Gene Fusion. *The Journal of Antibiotics*

(Tokyo). **60** (11), 700–708 (2007).
54. Helfrich, E. J. N., Piel, J. Biosynthesis of polyketides by *trans*-AT polyketide synthases. *Natural Product Reports.* **33** (2), 231–316 (2016).

Figure 1

Click here to access/download;Figure;Figure1.psd

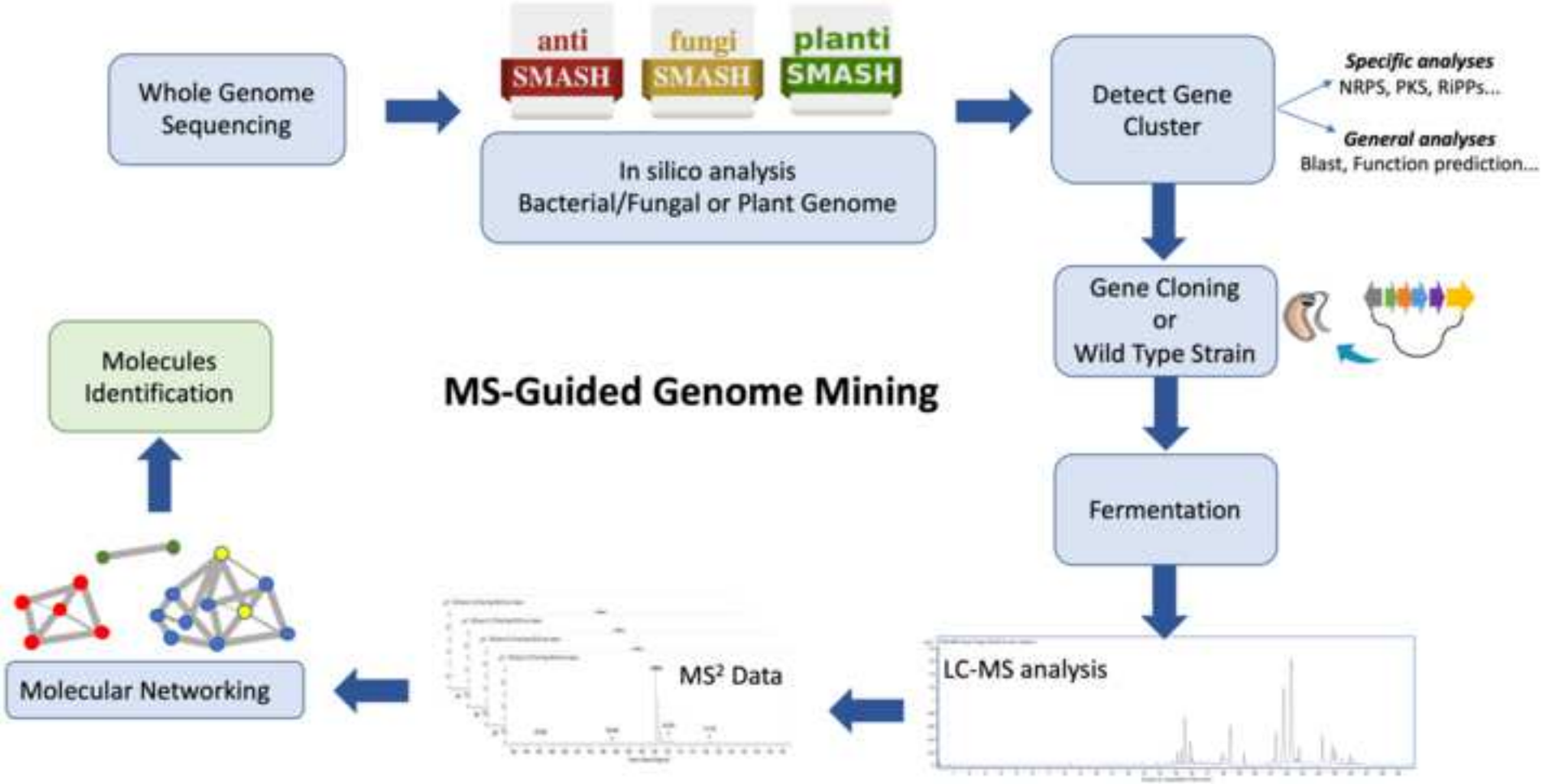
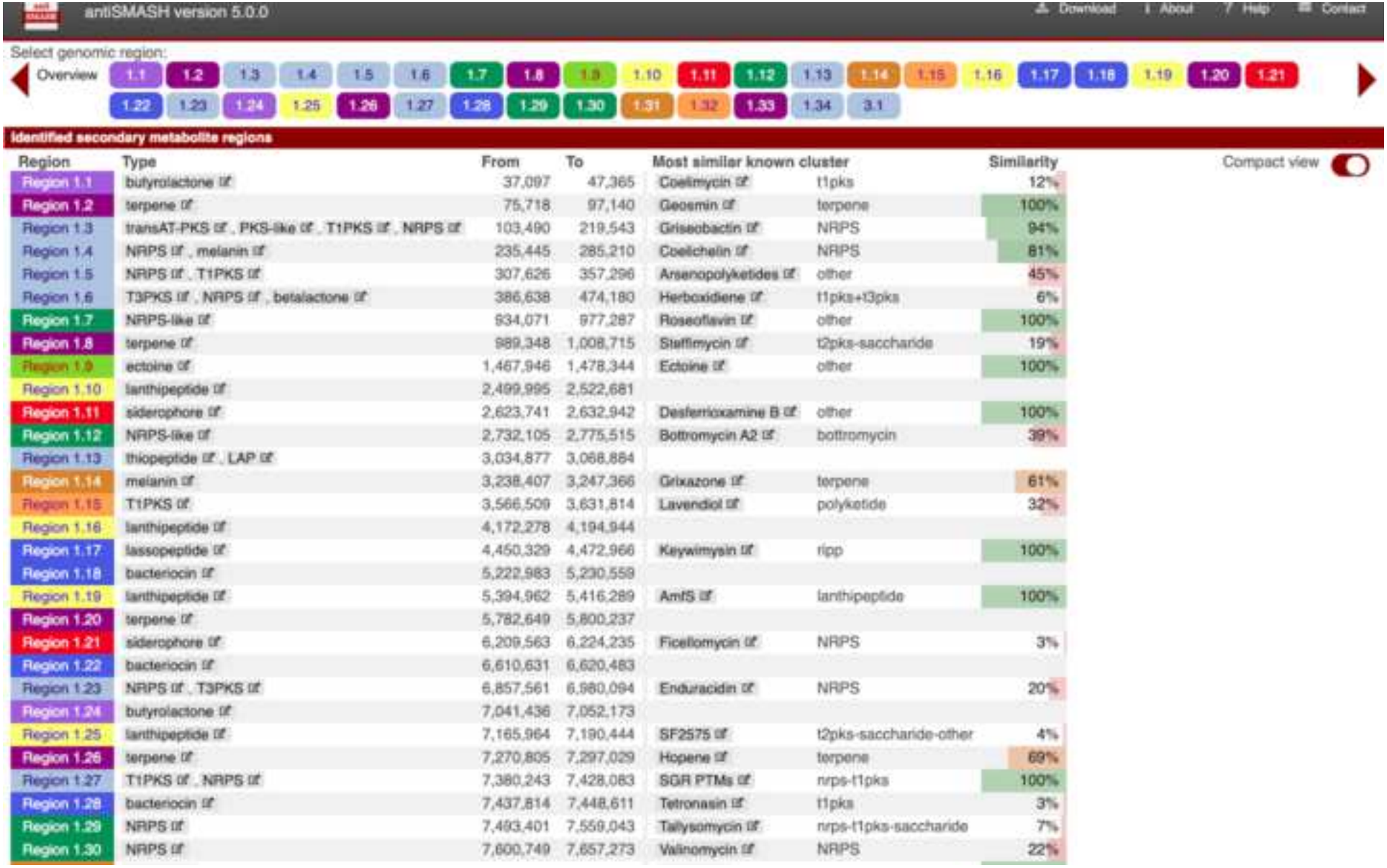


Figure 2

Click here to access/download;Figure;Figure2.psd



Region	Type	From	To	Most similar known cluster		Similarity
Region 1.1	butyrolactone	37,097	47,365	Coelimycin	t1pks	12%
Region 1.2	terpene	75,718	97,140	Geosmin	terpene	100%
Region 1.3	transAT-PKS , PKS-like , T1PKS , NRPS	103,490	219,543	Griseobactin	NRPS	94%
Region 1.4	NRPS , melanin	235,445	285,210	Coelichelin	NRPS	81%
Region 1.5	NRPS , T1PKS	307,626	357,296	Arsenopolyketides	other	45%
Region 1.6	T3PKS , NRPS , betalactone	386,638	474,180	Herboxidiene	t1pks+t3pks	6%
Region 1.7	NRPS-like	934,071	977,287	Roseoflavin	other	100%
Region 1.8	terpene	989,348	1,008,715	Steffimycin	t2pks-saccharide	19%
Region 1.9	ectoine	1,467,946	1,478,344	Ectoine	other	100%
Region 1.10	lanthipeptide	2,499,995	2,522,681			
Region 1.11	siderophore	2,623,741	2,632,942	Desferrioxamine B	other	100%
Region 1.12	NRPS-like	2,732,105	2,775,515	Bottromycin A2	bottromycin	39%
Region 1.13	thiopeptide , LAP	3,034,877	3,066,864			
Region 1.14	melanin	3,238,407	3,247,366	Grixazone	terpene	61%
Region 1.15	T1PKS	3,566,509	3,631,814	Lavendiol	polyketide	32%
Region 1.16	lanthipeptide	4,172,278	4,194,944			
Region 1.17	lassopeptide	4,450,329	4,472,966	Keywimysin	ripp	100%
Region 1.18	bacteriocin	5,222,983	5,230,559			
Region 1.19	lanthipeptide	5,394,962	5,416,289	AmfS	lanthipeptide	100%
Region 1.20	terpene	5,782,649	5,800,237			
Region 1.21	siderophore	6,209,563	6,224,235	Ficellomycin	NRPS	3%
Region 1.22	bacteriocin	6,610,631	6,620,483			
Region 1.23	NRPS , T3PKS	6,857,561	6,980,094	Enduracidin	NRPS	20%
Region 1.24	butyrolactone	7,041,436	7,052,173			
Region 1.25	lanthipeptide	7,165,964	7,190,444	SF2575	t2pks-saccharide-other	4%
Region 1.26	terpene	7,270,805	7,297,029	Hopene	terpene	69%
Region 1.27	T1PKS , NRPS	7,380,243	7,428,083	SGR PTMs	nrps-t1pks	100%
Region 1.28	bacteriocin	7,437,814	7,448,611	Tetronasin	t1pks	3%
Region 1.29	NRPS	7,493,401	7,559,043	Tallysomycin	nrps-t1pks-saccharide	7%
Region 1.30	NRPS	7,600,749	7,657,273	Valinomycin	NRPS	22%

Figure 3

Click here to access/download;Figure;Figure3.psd

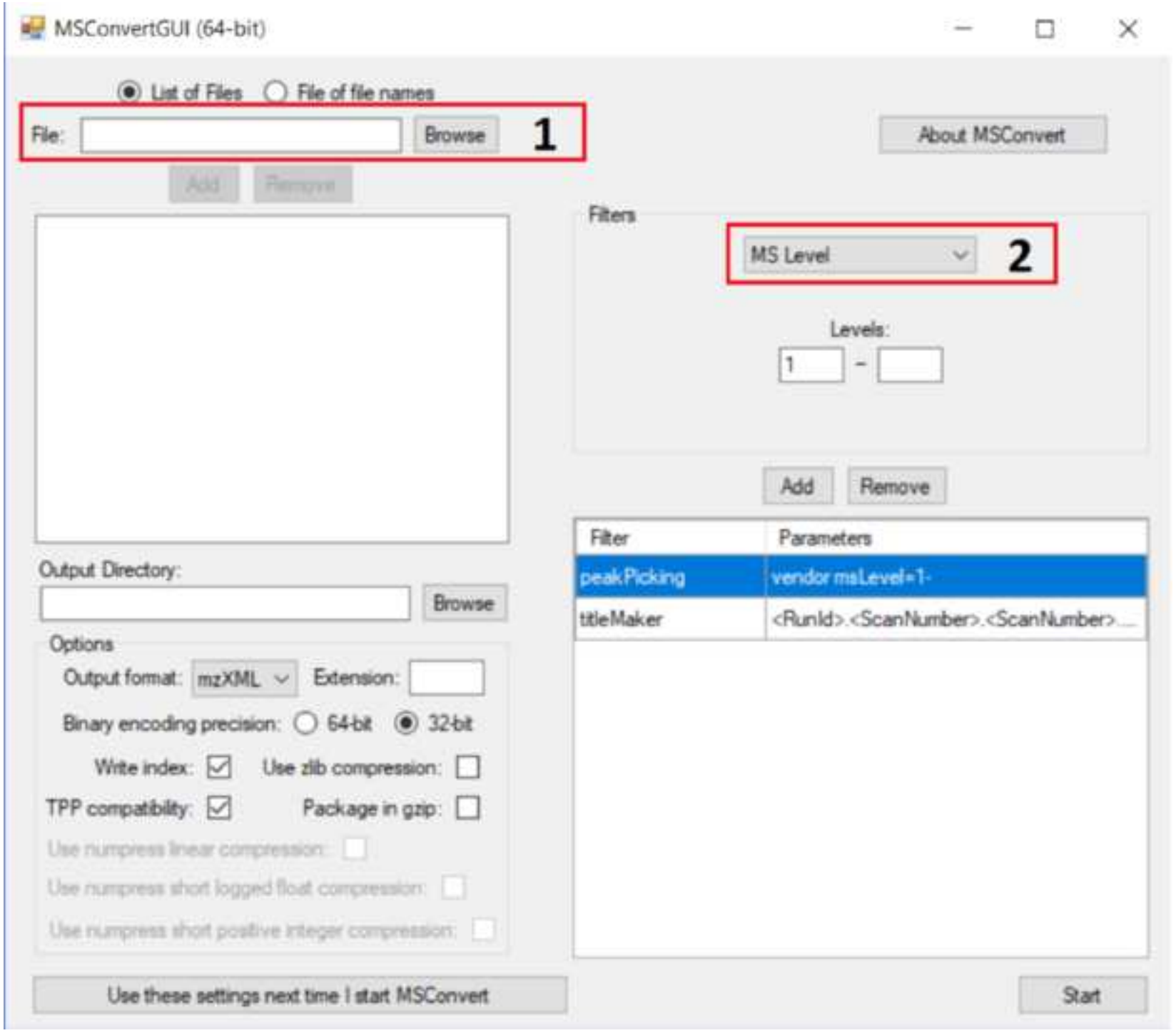


Figure 4

Click here to access/download;Figure;Figure4.psd

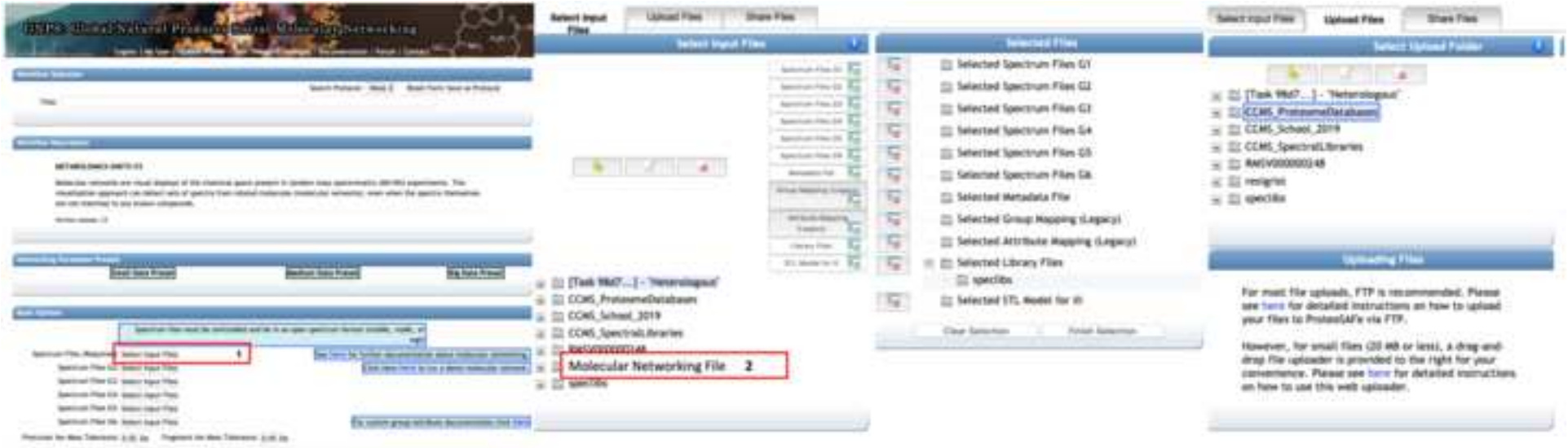


Figure 5

Click here to access/download;Figure;Figure5.psd

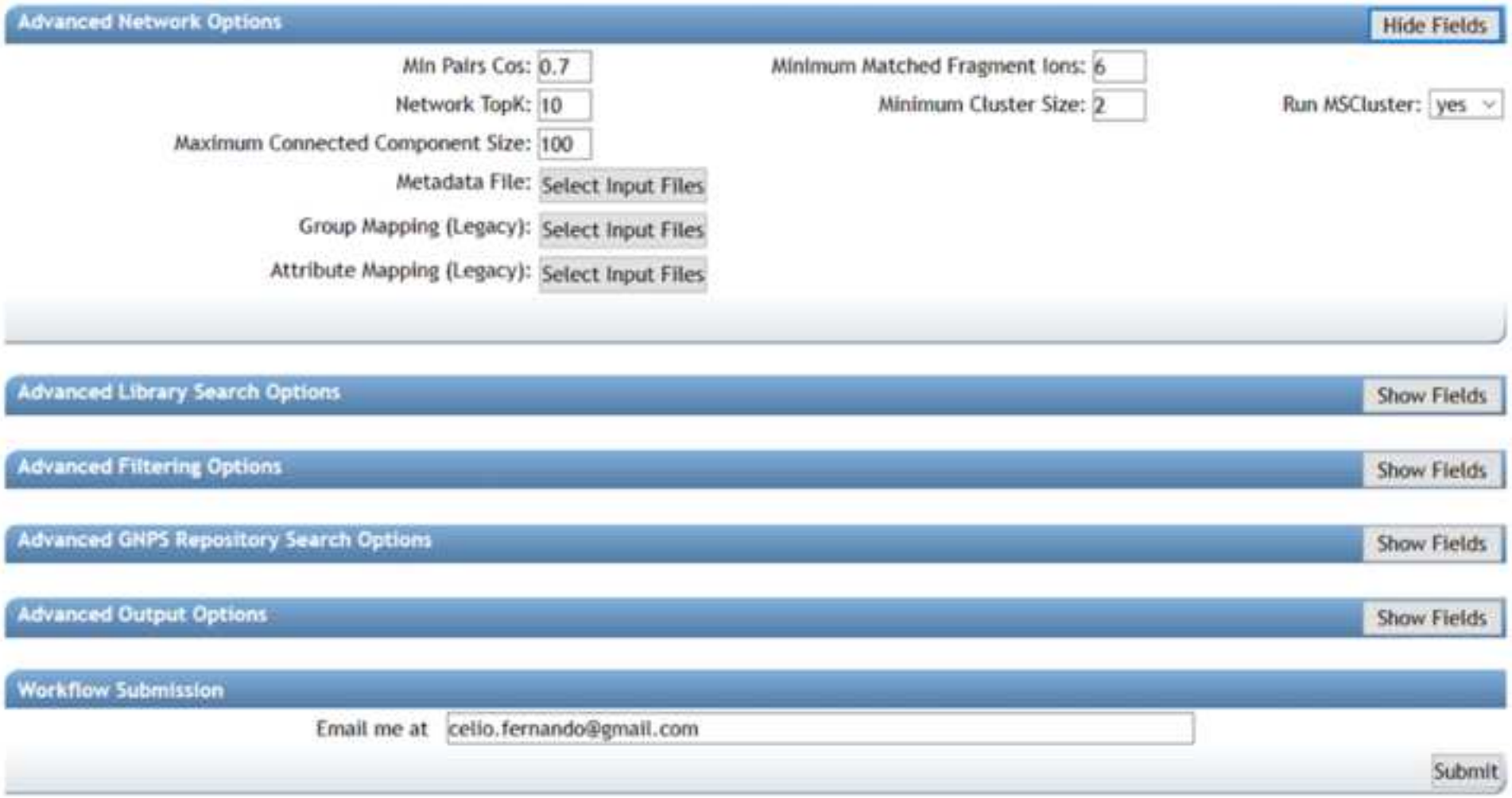


Figure 6

Click here to access/download;Figure;FIgure6.psd

Figure 7

Click here to access/download;Figure;Figure7.psd

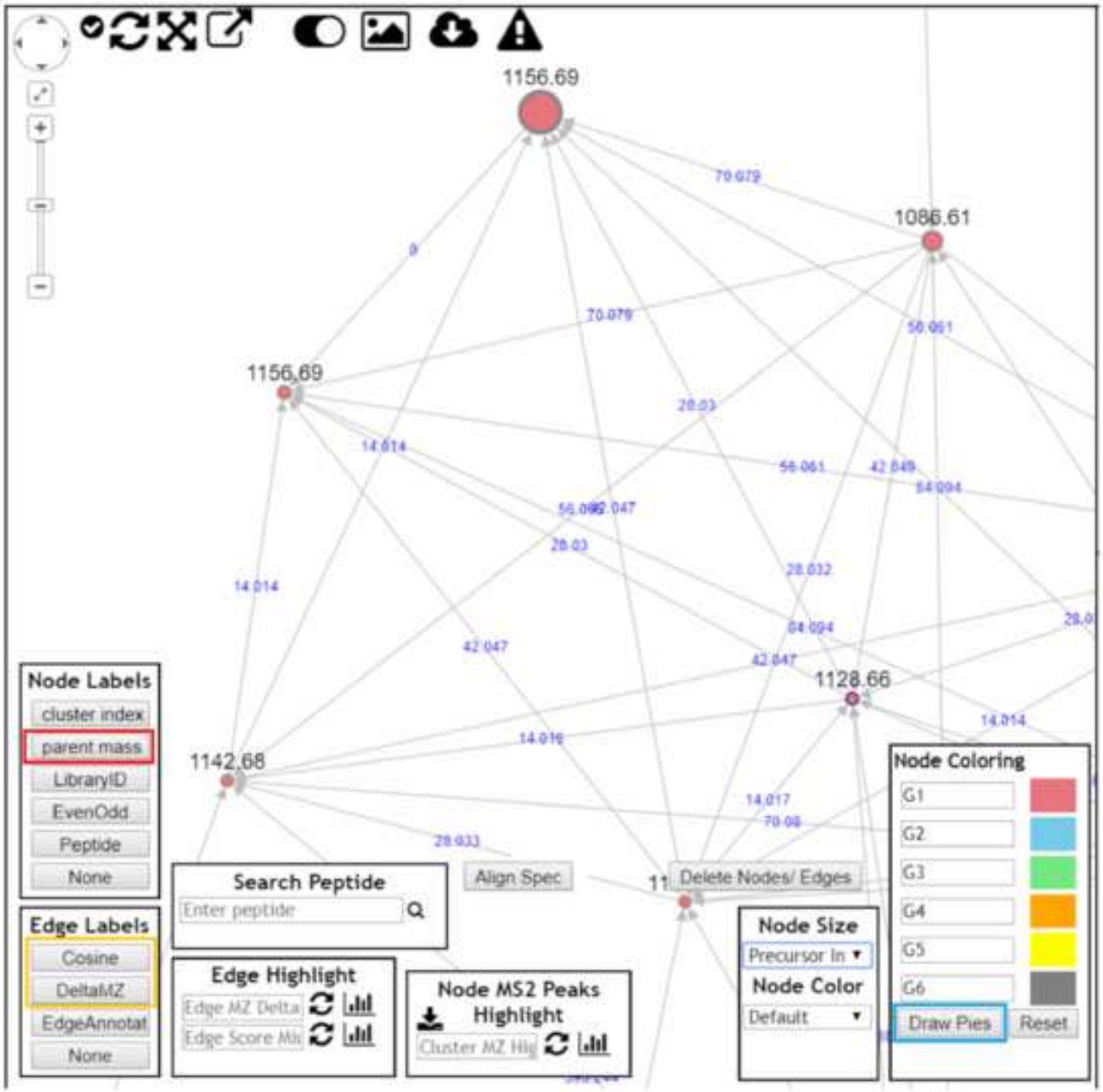


Figure 8

Click here to access/download;Figure;Figure8.psd

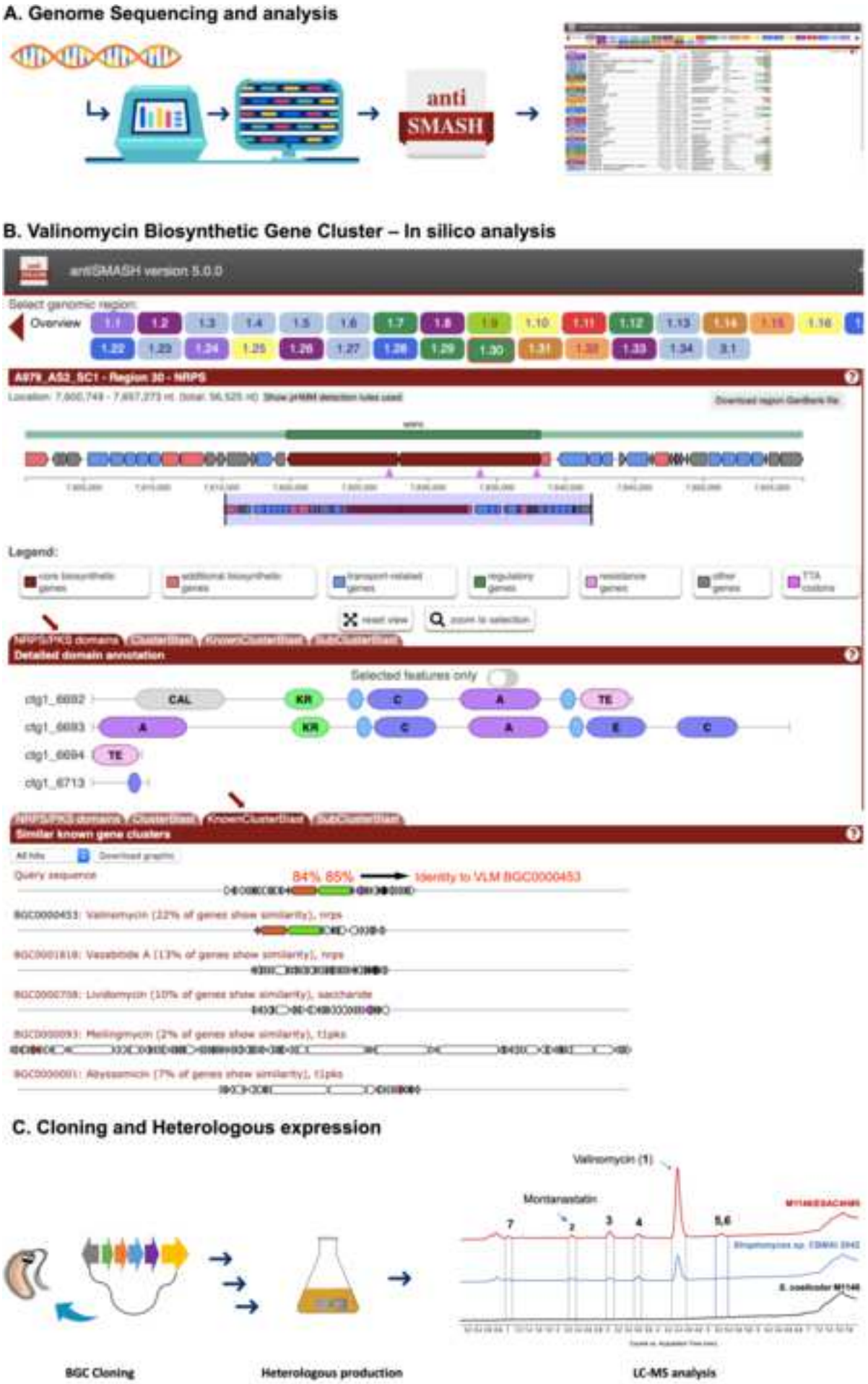
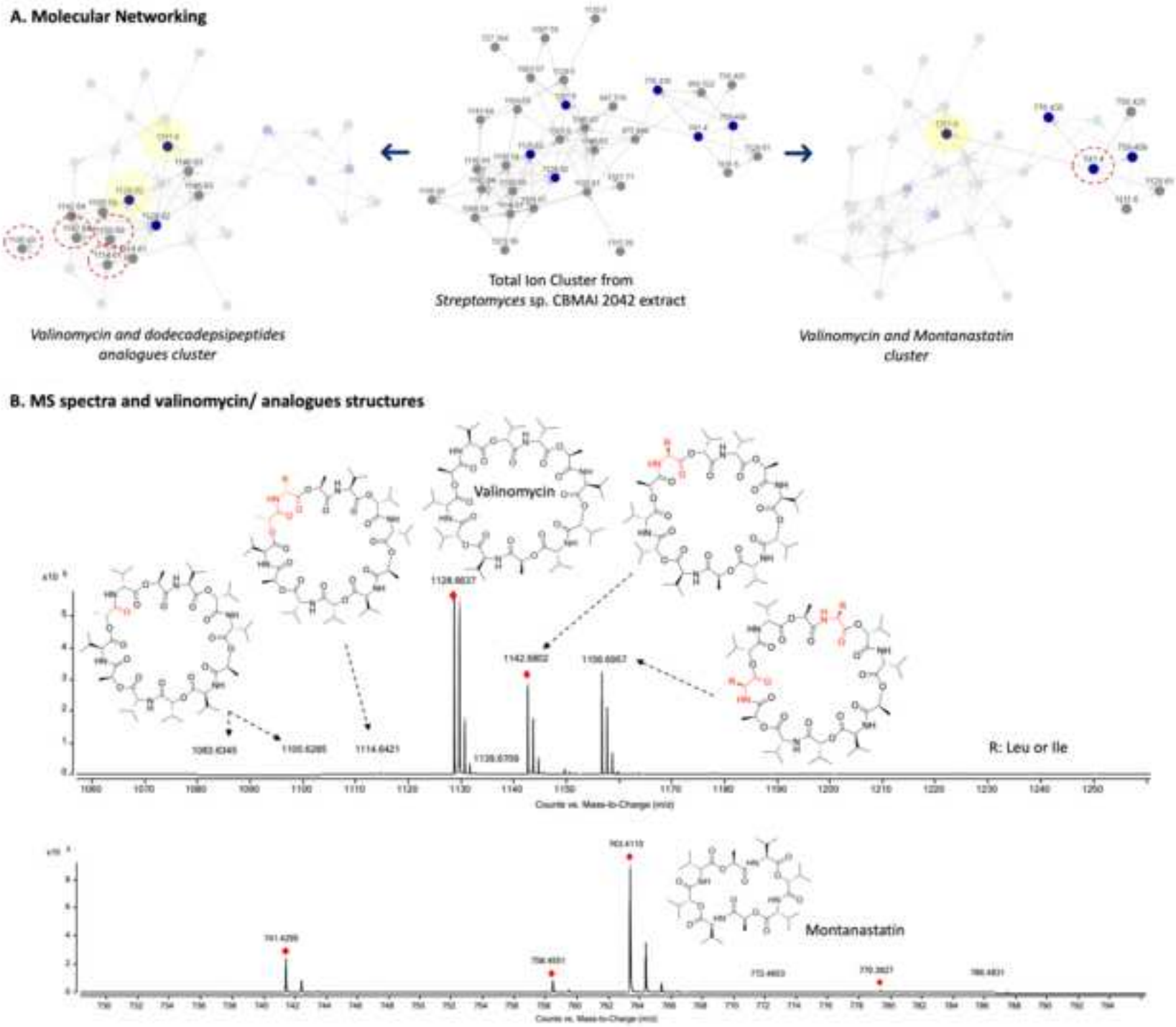


Figure 9

Click here to access/download;Figure;Figure9.psd



Name of Material/ Equipment	Company	Catalog Number
Acetonitrile	Tedia	AA1120-048
Agar	Oxoid	LP0011
Apramycin	Sigma Aldrich	A2024
Carbenicillin	Sigma Aldrich	C9231
Centrifuge	Eppendorf	NA
Chloramphenicol	Sigma Aldrich	C3175
Column C18	Agilent Technologies	NA
Kanamycin	Sigma Aldrich	K1377
Manitol P.A.- A.C.S.	Synth	NA
Microcentrifuge	Eppendorf	NA
Nalidixic acid	Sigma Aldrich	N4382
Phusion Flash High-Fidelity PCR Master Mix	ThermoFisher Scientific	F548S
Q-TOF mass spectrometer	Agilent technologies	NA
Sacarose P.A.- A.C.S.	Synth	NA
Shaker/Incubator	Marconi	MA420
Sodium Chloride	Synth	NA
Soy extract	NA	NA
Sucrose	Synth	NA
Thermal Cycles	Eppendorf	NA
Thiostrepton	Sigma Aldrich	T8902
Tryptone	Oxoid	LP0042
Tryptone Soy Broth	Oxoid	CM0129
UPLC	Agilent Technologies	NA
Yeast extract	Oxoid	LP0021

Comments/Description
HPLC grade
NA
NA
NA
5804
NA
ZORBAX RRHD Extend-C18, 80Å, 2.1 x 50 mm, 1.8 µm, 1200 bar pressure limit P/N 757700-902
NA
NA
5418
NA
NA
6550 iFunnel Q-TOF LC/MS
NA
NA
P. A. - ACS
NA
P. A. - ACS
Mastercycler Nexus Gradient
NA
NA
NA
1290 Infinity LC System
NA

Prof. Dr. Luciana G. de Oliveira,
Department of Organic Chemistry
Institute of Chemistry
University of Campinas - UNICAMP
Caixa Postal 6154 – CEP 13083-970
Fone: +55 19 35210193
FAX: +55 19 35213023

University of Campinas, 27th November 2019

Dear Dr. Phillip Steindel,
Review Editor of JoVE

Please, find enclosed the revised manuscript JoVE60825R1 "MS-Guided Genome Mining as Tool to Uncover New Natural Products", authored by Renata Sigrist, Bruno S. Paulo, Célio F. F. Angolini and Luciana G. de Oliveira, submitted for publication and video production in *JoVE*.

You will find below the answers to the Editorial comments and in the manuscript the protocol to be filmed highlighted in yellow.

We are looking forward to hearing from you soon.

Sincerely yours,

Prof. Luciana Gonzaga de Oliveira

Editorial comments:

1. Currently the protocol is too long to be filmed-we have strict limits on length, due to filming and video length concerns. Please highlight 2.75 pages or less of the protocol (complete sentences only; including headers and spacing) to be filmed.

Answer: the protocol was properly highlighted in yellow to fit in 2.75 pages for filming.

2. Step 4.3: This is vague; please include more detail and/or references.

Answer: As this a variable topic, concerning to sample used, it was added a reference.

3. Reviewer 2's comment about antiSmash and GNPS settings does not appear to be addressed in the manuscript itsef; it will probably be useful to include that information.

Answer: the notes were properly added in protocol steps and highlighted in green.