UC San Diego

UC San Diego Previously Published Works

Title

BLANKA: an Algorithm for Blank Subtraction in Mass Spectrometry of Complex Biological Samples.

Permalink

https://escholarship.org/uc/item/15m2n7dd

Journal

Journal of the American Society for Mass Spectrometry, 30(8)

ISSN

1044-0305

Authors

Cleary, Jessica L
Luu, Gordon T
Pierce, Emily C
et al.

Publication Date

2019-08-01

DOI

10.1007/s13361-019-02185-8

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Am Soc Mass Spectrom. Author manuscript; available in PMC 2020 August 01.

Published in final edited form as:

J Am Soc Mass Spectrom. 2019 August ; 30(8): 1426–1434. doi:10.1007/s13361-019-02185-8.

BLANKA: an algorithm for blank subtraction in mass spectrometry of complex biological samples

Jessica L Cleary[1], **Gordon T Luu**[1], **Emily C Pierce**[2], **Rachel J Dutton**[2], **Laura M Sanchez**[1]

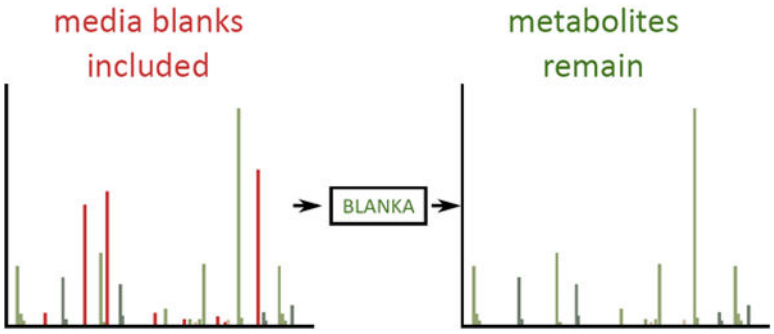
[1]Department of Medicinal Chemistry and Pharmacognosy, University of Illinois at Chicago, 833 S Wood St, Chicago, IL 60612, USA

[2]Division of Biological Sciences, University of California, San Diego, 9500 Gilman Dr, La Jolla, CA 92093, USA

Abstract

Multispecies microbiome systems are known to be closely linked to human, animal, and plant life processes. The growing field of metabolomics presents the opportunity to detect changes in overall metabolomic profiles of microbial species interactions. These metabolomic changes provide insight into function of metabolites as they correlate to different species presence and the observed phenotypic changes, but detection of subtle changes is often difficult in samples with complex backgrounds. Natural environments such as soil and food contain many molecules that convolute mass spectrometry-based analyses, and identification of microbial metabolites amongst environmental metabolites is an informatics problem we begin to address here. Our microbes are grown on solid or liquid cheese curd media. This medium, which is necessary for microbial growth contains high amounts of salts, lipids, and casein breakdown products which make statistical analyses using LC-MS/MS data difficult due to the high background from the media. We have developed a simple algorithm to carry out background subtraction from microbes grown on solid or liquid cheese curd media to aid in our ability to conduct statistical analyses so that we may prioritize metabolites for further structure elucidation.

Graphical Abstract



Address reprint requests to Laura M Sanchez, 833 S Wood St, MC 781, Room 539, University of Illinois at Chicago, Chicago, IL 60612, phone: (312) 996-0842, sanchelm@uic.edu.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Keywords

LC-MS/MS; Cheese curd media; bacteria; fungi; multispecies interactions

Introduction

Elucidation of chemical species directly involved in a given microbiome's formation and their exact role in subsequent microbial interactions is often difficult to assess because of the large number of abiotic and biotic variables in complex multi-domain microbial communities.[1–4] Despite these difficulties, chemical elucidation of specialized metabolites that govern these interactions has proven valuable,[5, 6] such as the recently described studies involving crop pathogens and the production and expression of the small molecules ralsolamycin and bikaverin.[7, 8] Ralsolamycin was found via imaging mass spectrometry to be important for how *Ralstonia solanacearum* exhibits an endofungal lifestyle potentially allowing it to persist in the environment in the absence of a plant host, whereas bikaverin protects specific *Fusarium* and *Botrytis* spp from invasion by this crop bacterial pathogen. Bikaverin is a weak antibiotic and ralsolamycin has antifungal properties. It is our expectation that identifying known and unknown secondary metabolites from microbial communities in a system with reduced complexity will similarly lead to further understanding of microbial chemical ecology and increase discovery of therapeutically or industrially relevant molecules.[2, 9]

Cheese rind-derived microbes allow for a simplified model system and can be used as a means to study the mechanisms behind microbial community formation.[2] Aged cheeses can be inoculated with desirable microbes yet many microbial species present at the end of the aging process are not those found in starter cultures and inoculations. The ability of similar genera to consistently colonize cheese rinds worldwide suggests that there are underlying mechanisms driving the formation of these microbiomes. Highly reproducible patterns of microbial community succession have been observed on cheese rinds with very little regional variation, indicating that the process of formation in this model system is not purely stochastic. Instead, community formation is heavily dependent upon observable factors such as environmental stressors and microbial interactions.[10] Elucidating these factors is feasible with cheese rind microbiomes mainly because of the limited number of variables present.[11] On average, a cheese rind contains 10 −12 different species of bacteria and fungi and the steps prior to aging are tightly controlled. Abiotic factors such as salt and pH content can easily be measured and manipulated while temperature and humidity are closely regulated throughout aging.[2] Previous work has demonstrated that biotic interactions are also crucial for proper species succession and there are likely metabolites that are unique to those biotic interactions.[10]

It is well established that production of metabolites is also dependent on microbial natural environments and growing partners.[2, 12, 13] Therefore it is important to mimic those natural environments in the laboratory as closely as possible. Metabolomics experiments are commonly performed on complex human and mammalian samples in a variety of applications and myriad tools exist for analysis of this data.[14–16] Often times, these

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

experiments are limited to known biomarkers or previous knowledge of the metabolites of interest.[17] Metabolomics is challenging for experiments that delve deeply into understudied systems which lack a wealth of standards and/or genomic information from the producing organisms, such as fermented cheese derived species. It is important in these cases to retain and primarily focus on *m/z* values that represent unknown metabolites associated with specific phenotypes.[18] At the same time, metabolomics performed on complex samples, such as extractions of cheese curd media with microbial growing partners, presents a challenge to sort unknown metabolites from noise and high background of proteins, peptides, and lipids.

Current metabolomics literature highlights the wide variety of online tools and the applications and ease with which users can access their potential.[19, 20] In order to properly use existing online platforms for metabolomic analysis of mass spectrometry data, it is often necessary to translate spectra that are collected into a list of *m/z* values found in each sample with intensity and, for liquid chromatography with tandem mass spectrometry (LC-MS/MS) data, retention times. Many tools exist to generate these lists, however most of these tools include all major peaks found in a spectrum and many of those peaks are not of any biological interest in our case given the high background from our media. Therefore, it is not always beneficial to take fold change over media controls to indicate thatsignals are uniquely produced metabolites as it is likely that microbes alter the concentration of media metabolites in the environment (ie the breakdown of casein to generate unique peptides over time). The online MetaboAnalyst platform has become a very useful tool for analysis of metabolomics data and is capable of a variety of statistical tests.[21] In our case, MetaboAnalyst tools such as principal component analysis (PCA, Figure 1a) are confounded by the presence of media metabolites as evidenced by the loadings (Figure 1b) that are strongly driven by media derived metabolites (*m/z* values 1461, 605, 414, 804).

To specifically identify *m/z* values that represent metabolites produced by microbes grown on/in complex media, it would be advantageous to completely eliminate *m/z* signals that are also found in media controls regardless of the abundance. Thousands of spectra are generated during one LC-MS/MS run and metabolomics experiments require many LC-MS/MS runs with biological and technical replicates for each sample. Manual curation of these large data sets is not possible or necessary when automation can be used to perform noise and blank or media control subtraction. There are existing online platforms to deal with different types of mass spectrometry data. For example, this can be accomplished with online tools such as the global natural product social molecular networking (GNPS) for LC-MS/MS data by inputting all data into molecular networks and manually subtracting all media and blank nodes post networking in Cytoscape.[22] However, GNPS is not capable of removing these media controls prior to molecule networking which lengthens analysis nor is it capable of processing matrix-assisted laser desorption/ionization coupled with time-of-flight mass spectrometry (MALDI-TOF MS) or LC-MS data and therefore is limited. Many online and offline tools are similarly capable of some level of blank and media subtraction but the process can be somewhat convoluted. SubtractMZ is a function found in the msPurity R package developed by Lawson *et al.* that performs blank subtraction.[23] Schiffman *et al.* have also incorporated blank subtraction into a metabolomics pipeline.[24] However, while both algorithms perform blank removal, the knowledge and ability to write/modify code is

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

required to implement blank removal. Moreover, the output of these tools are incompatible for the input in other online tools such as GNPS or MetaboAnalyst.[19, 22] Emerging technologies for utilizing MALDI-TOF MS data to establish metabolomic profiles[25] highlight the need to first remove media signals from data before undergoing extensive analysis. We have created an algorithm for subtracting noise, blanks, and media controls from mass spectra data files without reliance on expertise with online platforms or proprietary/commercial software, and have performed subsequent analysis on LC-MS/MS and MALDI data from cheese curd media microbial extracts.

Experimental

A. Microbial Culturing

All bacterial cultures were grown overnight in brain-heart infusion (Bacto® BHI) liquid media (BD) at room temperature. Liquid cultures were normalized to an optical density (OD_{600}) of 0.1 and bacterial cultures were diluted 10^{-1} for further experiments. Fungal cultures were grown on plate count agar milk salt (PCAMS; 1 g/L whole milk powder, 1 g/L dextrose, 2.5 g/L yeast extract, 5 g/L tryptone, 10 g/L sodium chloride, 15 g/L agar). Plates were kept at room temperature and spores were harvested at 7 days (or until sporulation was observed) of growth for subsequent experiments. Spores harvested from fungi were put into 1X PBS and normalized to an O.D. of 0.1 for further experiments.

B. Extraction of cultures

For extraction of solid agar plates, 5uL of working cultures were spotted onto 10% cheese curd agar (CCA: 100 g/L lyophilized cheese curd, 5 g/L xanthan gum, 30 g/L NaCl, 17 g/L agar, pH adjusted to 7.0). After at least 7 days of growth, agar was removed from the petri plate and placed into 50 mL falcon tubes. Five mL of acetonitrile was added to each tube and all were sonicated for 30 minutes. All falcon tubes were centrifuged and liquid was removed from the solid agar pieces and put into 15 mL falcon tubes. The falcon tubes containing agar were then centrifuged and liquid was removed from any residual solid debris and put into scintillation vials. These liquid extractions were then dried using a steady stream of air. Dried extracts were then weighed and diluted with methanol to obtain 1 mg/mL solutions which were put into HPLC vials and analyzed on a Thermo LCQ advantage max ion trap and a Bruker Impact II qTOF.

C. Mass spectrometry data collection

Low-resolution LC-MS/MS analysis was done on a Thermo Finnigan LCQ Advantage Max mass spectrometer coupled to an HP1050 HPLC. A gradient of 10–100% methanol with 0.02% formic acid over 25 minutes was used for separation. The ESI conditions were set with the source voltage at 5 kV and capillary temperature at 250°C. The detection window was set from 200 to 2000 Da, collision energy was at 35%, isolation width was 3 *m/z*, with three data dependent MS^2 events per MS^1 and dynamic exclusion. High resolution LC-MS/MS data was collected on a Bruker impact II qTOF in positive mode with the detection window set from 50 to 1500 Da, on a UPLC gradient of 10–100% acetonitrile with 0.02% formic acid over 17 minutes. The ESI conditions were set with the capillary voltage at 4.5 kV. The detection window was set from 50 to 1500 Da and the top three precursor ions from

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

each MS^1 scan were subjected to collision energies of 12 eV, 48eV, and 60eV for a total of nine data dependent MS^2 events per MS^1 with dynamic exclusion.

D. BLANKA

BLANKA (https://github.com/gtluu/blanka) is a command line script written in Python that removes noise and background (control) media signals without the need for user written code. (documentation found in Online Resource 2). It currently supports LC-MS (LC-MS/MS) and MALDI-TOF MS spectra, and has been tested using data from a Thermo Finnigan LCQ Advantage Max, Bruker MaXis, and a Bruker AutoFlex Speed LRF. Raw data formats generated during data collection or .mzXML can be used as input for BLANKA. Users may specify a parent folder containing sample and control data, and all subfolders will be searched for data. Multiple sample datasets can be processed in one run as long as data is found under the parent folder, and multiple control datasets can be used to allow for technical/biological replicates. LC-MS data should consist of one .mzXML file per LC-MS run. MALDI data should consist of one .mzXML file per spot. In addition to the files, it is necessary to have the metadata for each file in an Excel template containing the coordinates and identities of each sample. If the user specifies raw data as the input, MSConvert[26] is used to convert the data into .mzXML format as the first step of the BLANKA algorithm. Once control and sample datasets have been loaded using the mzXML module in Pyteomics,[27] noise removal is first performed on each dataset based on a user defined threshold, we recommend at least a 4:1 signal to noise ratio as a starting cutoff, followed by removal of signal peaks from the experimental spectrum based on corresponding control spectra. By default, BLANKA removes both noise and background signals, but the user may choose to forego either step and perform only noise removal or only blank removal. Several files are generated when running BLANKA: 1) raw data in .mzXML format (if original input was not .mzXML format), 2) an .mgf file with the noise/blank removed MS^2 spectra, 3) an .mgf file with the noise/blank removed MS^1 and MS^2 spectra 4) an .mgf file with lists of only removed background peaks from each spectra, 5) an .mgf file with the noise removed MS^2 spectra, and 6) an .mgf file with the noise removed MS^1 and MS^2 spectra (Online Resource 2). If the user performs only blank removal, no noise removal file will be output and vice versa. All files are output to a user specified directory, and in the event that no directory is specified, files are output to the directory that the input data was found in. The amount of files that can be simultaneously processed by BLANKA and the amount of time required is dependent upon computer hardware and data file size and as we continue to test and develop BLANKA, general limitations will become clear. For the analysis performed here, three blank files were removed from six sample files in under two minutes for the low resolution files and two blank files were removed from two sample files in under five minutes for the high resolution data.

Results and Discussion

Noise Removal

To perform noise removal, BLANKA first calculates baseline noise by averaging the n least intense peaks in a given spectrum (n defined in Equation 1)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$n = 0.05 \times \text{number of peaks in spectrum} \quad (1)$$

Once the baseline noise has been calculated, peaks that are less than or equal to the signal to noise ratio (SNR) specified are removed from from the spectrum, as illustrated in Figure 2b.

LC-MS(/MS) Blank Removal

For each given spectrum in a sample dataset, a corresponding spectrum with a matching retention time (rt) within a rt tolerance window (MS^1) is identified from the control dataset, which can be comprised of one or more LC-MS runs. In the case of LC-MS/MS data, matching rt within a rt tolerance window as well as precursor ion mass within a precursor ion mass tolerance window are identified from the control dataset. Tolerance levels for both rt and precursor ion mass may be specified by the user to adjust the algorithm for various instrument specifications. In the event that multiple ion matches in a rt threshold are found, the spectrum with the closest matching criteria is selected as the control ion. If no ion matches in a designated rt are found, the spectrum remains unmodified. It is important to point out that BLANKA does not perform a peak picking and rt alignment step which is common to many metabolomics experiments. In BLANKA, peak picking is not usually necessary because the data is centroided either prior to input or during the conversion step if raw files are used as the input. While the inspiration for this algorithm was to aid with our metabolomics experiments, we intend BLANKA to be for general use with mass spectrometry data sets and users should be able to perform blank subtraction without needing technical or biological replicates. If rt drift outside the defined tolerances is expected and users have replicates, it would be beneficial to perform rt alignment for LC-MS files using existing metabolomics tools, such as XCMS. This is not as much of a concern for LC-MS/MS files considering that the fragmentation data associated with precursor ions is informative along with rt. Tandem spectra data would still be identified to form a consensus spectrum in tools such as GNPS.

MALDI Dried Droplet Blank Removal

In the case of MALDI-TOF MS data, the user will define the control spectrum spot in the algorithm and the corresponding spectrum is used for the subtraction and noise removal. In the case where multiple technical and/or biological replicates are present in the dataset, signal averaging is used to create a single consensus control spectrum for subtraction. Each experimental spectrum is then compared to the consensus control spectrum, and matching peaks found in the control spectrum that correspond to a signal within the signal ion mass tolerance in the experimental spectrum are then removed from the experimental spectrum.

BLANKA Performance

To begin to assess BLANKA’s performance, the GNPS molecular networking workflow was employed using six files which correspond to three biological replicates each of *Penicillium* sp. #12 (fungus) with either *E. coli* K12 (bacteria #1) or *Pseudomonas psychrophila* sp. JB418 (bacteria #2) as growing partners.[2] Original data sets including controls which were comprised of extracted cheese curd agar and those with the controls removed using

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

BLANKA were run through identical workflow parameters and compared in Figure 3 to confirm that BLANKA was capable of removing nodes resulting from blank and media controls. Networking parameters can be viewed in Online Resource 2.

In the resulting molecular network without BLANKA subtraction (Figure 3a), 24 out of all 66 nodes were found in controls leaving 42 nodes that were only found in fungal cultures. Fifteen of the control nodes were present also in samples and should be removed by BLANKA in our processed data sets. Inspection of the networks shows that all but two of those control nodes (*m/z* 1461 and 378) were removed by BLANKA. Manual inspection of the raw data demonstrated that *m/z* 1461 and 378 were indeed found in control files and not removed with BLANKA due to differences in retention times of the precursor ions. Theoretically, the BLANKA processed data set would all contain 42 nodes from the original data set that were considered fungal metabolites and the loss of nodes is likely due to the wide tolerance settings used in BLANKA for a low resolution instrument. Data reduction is to be expected and the user can set tolerances according to the tradeoffs between adequate control removal and loss of real data.

The discrepancy here highlights a limitation to this algorithm that can occur due to experimental error. In the case observed here, the retention time matching was just out of the user defined tolerance window which can occur in our system due to the lack of inline degasser on the LC. For this issue, the inclusion of technical replicates with rt drift alignment performed using tools such as MZmine or XCMS would not only circumvent this problem but would also aid in identification of contaminants inherent to PEEK tubing and individual instruments.[28] Taking this into consideration, the user should be aware that BLANKA removes noise and media controls based on matching retention times and should be used only when retention times are comparable and take care to set appropriate retention time thresholds.

Statistical analysis of the resulting BLANKA processed molecular networks showed that removal of media blanks and noise resulted in a different set of *m/z* values that are considered significant (Figure 4) compared to the unprocessed data. Three out of eleven identified *m/z* values in the unprocessed data were due to media components (Figure 4a) as opposed to one out of nine *m/z* values identified by the processed data (Figure 4b) that was not removed by BLANKA, as described above. This serves to highlight that as with all data visualization tools, manual inspection of the raw data is necessary before further validation of the ions or metabolites is carried out. We would also point out that removal of control peaks will result in different absolute values for statistical analysis when a normalization step is included. For example, the fold change of *m/z* values in these plots differ because BLANKA treatment removes what are likely to be intense control spectral peaks. The normalization applied in MetaboAnalyst thus considers a different set of peaks and subsequent calculations reflect that. The data displayed in Figures 3 and 4 was processed using BLANKA settings adjusted for a low-resolution instrument as the default settings are appropriate for high-resolution instruments only. Tailoring these settings for different datasets acquired on the users instruments is highly advised in order to enhance the output.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The LC-MS/MS data acquired on a 3D ion trap unprocessed data consists of six sample files (two different conditions with three biological replicates of each) and five control files (three media technical replicates and two solvent technical replicates). LC-MS/MS data acquired on a qTOF unprocessed data consists of two sample files and two control files (one media and one solvent). Volcano plots are not displayed for qTOF data as only one replicate of each sample was obtained and statistical analyses would therefore not be appropriate. A comparison of the number of clusters and library hits found by GNPS listed in Table 1 highlights the differences in processing data with BLANKA versus including controls in data sets.

It is worth noting that the significant data reduction achieved through processing with BLANKA on a small subset of samples will likely be enhanced for experiments with many sample files. To test this proposal we ran BLANKA on an online public data set containing 26 files from a high resolution instrument and compared the processed and unprocessed data using GNPS (online resource 4). The data from this MassIVE dataset (MSV000080540) explores how *Fusarium fujikuroi* metabolically responds to wild type and a mutant strain of *Ralstonia solanacearum* as well as high and low nitrogen conditions.[7] This data set also contains media and solvent controls and biological replicates making it an ideal test for high resolution LC-MS/MS data. We found on average a 73% reduction in the number of nodes found in both sample and media and a 21% reduction in the number of nodes found in samples only (Online Resource 2**, Items 5 and 6**). The reasons for the discrepancy between 100% and 73% reduction of undesirable nodes and 0% and 21% reduction of desirable nodes are likely multifaceted, but ultimately due to the fact that GNPS considers fragmentation and not retention time while clustering spectra together while BLANKA consider retention time but not fragmentation while removing spectra. This highlights the value in orthogonal use of GNPS to screen BLANKA processed files for media components that may not match retention times (perhaps due to pH or choice of stationary phase) but have similar or identical fragmentation. In general, using BLANKA to reduce the amount of data that is input into GNPS results in networks that are smaller and thus easier to navigate, but it should not be considered a stand-alone blank removal step.

We continue to identify and correct BLANKA performance as our own sample size increases. Future iterations of BLANKA will include the capacity to export files in .mzXML format which will allow direct import into MetaboAnalyst and circumvent the need to export clusters from GNPS. This capacity will also allow users to input files into XCMS which performs peak picking and retention time alignment for direct comparison of two different conditions, such as with cloud plots,[29] without the complication of media components.

Conclusions

Blank and noise subtraction is necessary for the analysis of data from nutrient-complex samples in order to quickly prioritize signals for further validation based on statistical analyses from the metabolomic information. Statistical tools for the analysis of metabolomics datasets are useful for extracting valuable information from large amounts of data. Removing data points that represent media artifacts completely allows us to run statistical analysis with more confidence in results, and we will continue to develop

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

BLANKA as we expand to larger datasets. This simple algorithm prepares data for further analysis using existing online platforms such as GNPS and MetaboAnalyst with minimal effort by the user and is ideal for users that have complex nutrient or media requirements to culture their cells or microbial samples

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This publication was funded in part by UIC Startup funds (L.M.S.). Research reported in this publication was supported in part by the National Center for Complementary and Integrative Health of the National Institutes of Health under Award Number T32AT007533 (J.L.C) and the National Institute of General Medicine of the National Institutes of Health under Award Number T32GM7240–40 (E.C.P). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was supported by NSF grant 1817955 and 1817887 (to L.M.S. and R.J.D.).

References

1. Nemergut DR, Schmidt SK, Fukami T, O'Neill SP, Bilinski TM, Stanish LF, Knelman JE, Darcy JL, Lynch RC, Wickey P, Ferrenberg S: Patterns and processes of microbial community assembly. Microbiol. Mol. Biol. Rev 77, 342–356 (2013) [PubMed: 24006468]
2. Wolfe BE, Button JE, Santarelli M, Dutton RJ: Cheese rind communities provide tractable systems for in situ and in vitro studies of microbial diversity. Cell. 158, 422–433 (2014) [PubMed: 25036636]
3. Gibbons SM, Gilbert JA: Microbial diversity--exploration of natural ecosystems and microbiomes. Curr. Opin. Genet. Dev 35, 66–72 (2015) [PubMed: 26598941]
4. Zuñiga C, Zaramela L, Zengler K: Elucidation of complexity and prediction of interactions in microbial communities. Microb. Biotechnol 10, 1500–1522 (2017) [PubMed: 28925555]
5. Liu J, Fu K, Wang Y, Wu C, Li F, Shi L, Ge Y, Zhou L: Detection of Diverse N-Acyl-Homoserine Lactones in Vibrio alginolyticus and Regulation of Biofilm Formation by N-(3-Oxodecanoyl) Homoserine Lactone In vitro. Front. Microbiol 8, 1097 (2017) [PubMed: 28670299]
6. Schoenian I, Spiteller M, Ghaste M, Wirth R, Herz H, Spiteller D: Chemical basis of the synergism and antagonism in microbial communities in the nests of leaf-cutting ants. Proc. Natl. Acad. Sci. U. S. A 108, 1955–1960 (2011) [PubMed: 21245311]
7. Spraker JE, Wiemann P, Baccile JA, Venkatesh N, Schumacher J, Schroeder FC, Sanchez LM, Keller NP: Conserved Responses in a War of Small Molecules between a Plant-Pathogenic Bacterium and Fungi. MBio. 9, (2018). doi:10.1128/mBio.00820-18
8. Spraker JE, Sanchez LM, Lowe TM, Dorrestein PC, Keller NP: Ralstonia solanacearum lipopeptide induces chlamydospore development in fungi and facilitates bacterial entry into fungal tissues. ISME J. 10, 2317–2330 (2016) [PubMed: 26943626]
9. Cho I, Blaser MJ: The human microbiome: at the interface of health and disease. Nat. Rev. Genet 13, 260–270 (2012) [PubMed: 22411464]
10. Kastman EK, Kamelamela N, Norville JW, Cosetta CM, Dutton RJ, Wolfe BE: Biotic Interactions Shape the Ecological Distributions of Staphylococcus Species. MBio. 7, (2016). doi:10.1128/mBio.01157-16
11. Cleary JL, Kolachina S, Wolfe BE, Sanchez LM: Coproporphyrin III Produced by the Bacterium Glutamicibacter arilaitensis Binds Zinc and Is Upregulated by Fungi in Cheese Rinds. mSystems. 3, (2018). doi:10.1128/mSystems.00036-18
12. Adnani N, Chevrette MG, Adibhatla SN, Zhang F, Yu Q, Braun DR, Nelson J, Simpkins SW, McDonald BR, Myers CL, Piotrowski JS, Thompson CJ, Currie CR, Li L, Rajski SR, Bugni TS: Coculture of Marine Invertebrate-Associated Bacteria and Interdisciplinary Technologies Enable

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Biosynthesis and Discovery of a New Antibiotic, Keyicin. ACS Chem. Biol 12, 3093–3102 (2017) [PubMed: 29121465]

13. Nai C, Meyer V: From Axenic to Mixed Cultures: Technological Advances Accelerating a Paradigm Shift in Microbiology. Trends Microbiol. 26, 538–554 (2018) [PubMed: 29191399]
14. Sheikh KD, Khanna S, Byers SW, Fornace A Jr, Cheema AK: Small molecule metabolite extraction strategy for improving LC/MS detection of cancer cell metabolome. J. Biomol. Tech 22, 1–4 (2011) [PubMed: 21455475]
15. Chen Y, Ma Z, Li A, Li H, Wang B, Zhong J, Min L, Dai L: Metabolomic profiling of human serum in lung cancer patients using liquid chromatography/hybrid quadrupole time-of-flight mass spectrometry and gas chromatography/mass spectrometry. J. Cancer Res. Clin. Oncol 141, 705–718 (2015) [PubMed: 25293627]
16. Gosetti F, Mazzucco E, Gennaro MC, Marengo E: Ultra high performance liquid chromatography tandem mass spectrometry determination and profiling of prohibited steroids in human biological matrices. A review. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci 927, 22–36 (2013)
17. Ulmer CZ, Koelmel JP, Ragland JM, Garrett TJ, Bowden JA: LipidPioneer : A Comprehensive User-Generated Exact Mass Template for Lipidomics. J. Am. Soc. Mass Spectrom. 28, 562–565 (2017) [PubMed: 28074328]
18. Jeffryes JG, Colastani RL, Elbadawi-Sidhu M, Kind T, Niehaus TD, Broadbelt LJ, Hanson AD, Fiehn O, Tyo KEJ, Henry CS: MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. J. Cheminform 7, 44 (2015) [PubMed: 26322134]
19. Spicer R, Salek RM, Moreno P, Cañueto D, Steinbeck C: Navigating freely-available software tools for metabolomics analysis. Metabolomics. 13, 106 (2017) [PubMed: 28890673]
20. Misra BB, Fahrmann JF, Grapov D: Review of emerging metabolomic tools and resources: 2015–2016. Electrophoresis. 38, 2257–2274 (2017) [PubMed: 28621886]
21. Chong J, Xia J: MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data. Bioinformatics. 34, 4313–4314 (2018) [PubMed: 29955821]
22. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu W-T, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu C-C, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw C-C, Yang Y-L, Humpf H-U, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, P CAB, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodríguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard P-M, Phapale P, Nothias L-F, Alexandrov T, Litaudon M, Wolfender J-L, Kyle JE, Metz TO, Peryea T, Nguyen D-T, VanLeer D, Shinn P, Jadhav A, Müller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BO, Pogliano K, Linington RG, Gutiérrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N: Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat. Biotechnol 34, 828–837 (2016) [PubMed: 27504778]
23. Lawson TN, Weber RJM, Jones MR, Chetwynd AJ, Rodrı Guez-Blanco G, Di Guida R, Viant MR, Dunn WB: msPurity: Automated Evaluation of Precursor Ion Purity for Mass Spectrometry-Based Fragmentation in Metabolomics. Anal. Chem 89, 2432–2439 (2017) [PubMed: 28194963]
24. Schiffman C, Petrick L, Perttula K, Yano Y, Carlsson H, Whitehead T, Metayer C, Hayes J, Edmands WMB, Rappaport S, Dudoit S: Data-adaptive pipeline for filtering and normalizing metabolomics data, https://www.biorxiv.org/content/10.1101/387365v1, (2018)
25. Clark CM, Costa MS, Sanchez LM, Murphy BT: Coupling MALDI-TOF mass spectrometry protein and specialized metabolite analyses to rapidly discriminate bacterial function. Proc. Natl. Acad. Sci. U. S. A 115, 4981–4986 (2018) [PubMed: 29686101]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

26. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak M-Y, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P: A cross-platform toolkit for mass spectrometry and proteomics. Nat. Biotechnol 30, 918–920 (2012) [PubMed: 23051804]

27. Levitsky LI, Klein J, Ivanov MV, Gorshkov MV: Pyteomics 4.0: five years of development of a Python proteomics framework. J. Proteome Res. (2018). doi:10.1021/acs.jproteome.8b00717

28. Caesar LK, Kvalheim OM, Cech NB: Hierarchical cluster analysis of technical replicates to identify interferents in untargeted mass spectrometry metabolomics. Anal. Chim. Acta 1021, 69–77 (2018) [PubMed: 29681286]

29. Gowda H, Ivanisevic J, Johnson CH, Kurczy ME, Benton HP, Rinehart D, Nguyen T, Ray J, Kuehl J, Arevalo B, Westenskow PD, Wang J, Arkin AP, Deutschbauer AM, Patti GJ, Siuzdak G: Interactive XCMS Online: simplifying advanced metabolomic data processing and subsequent statistical analyses. Anal. Chem 86, 6931–6939 (2014) [PubMed: 24934772]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

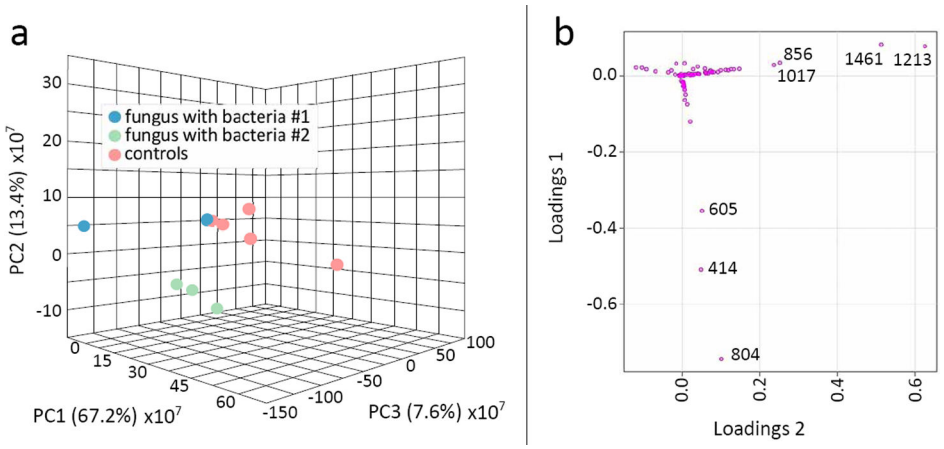


Figure 1.

a) Principal Component Analysis (PCA) was performed using the MetaboAnalyst platform with list outputs of clustered data from GNPS (available in Online Resource 1). b) Loading for the PCA plots point out *m/z* values that contribute to variability in samples. These *m/z* values are represented in their original networks in Figure 2b.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

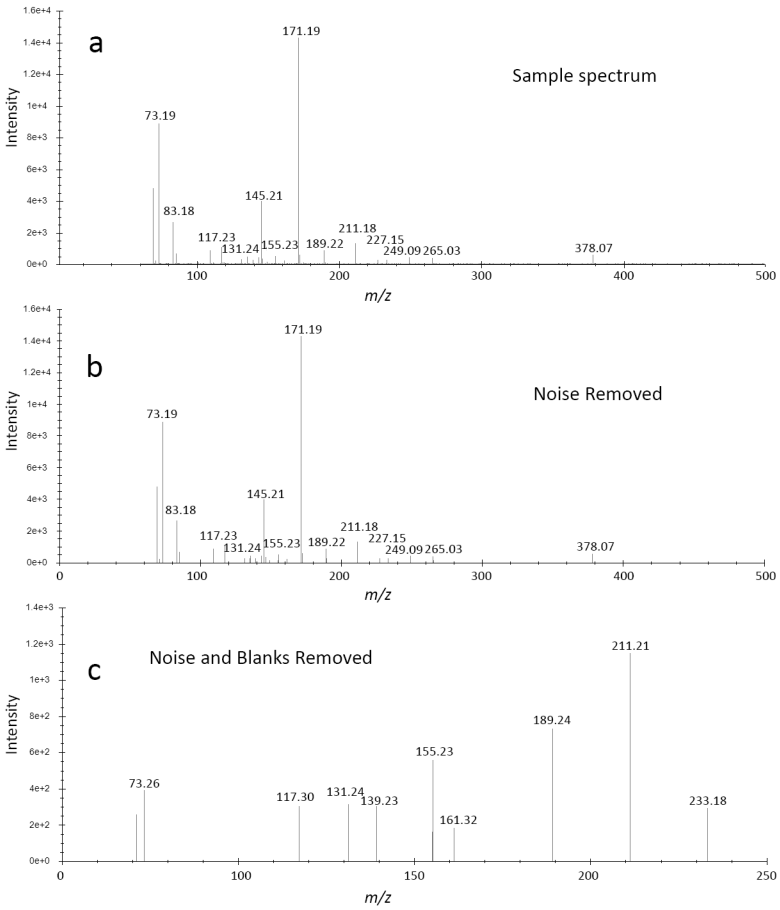


Figure 2.

MALDI mass spectra throughout the process of noise and blank removal. a) Baseline noise is visible in the original spectra obtained and b) removed by BLANKA. Media controls are considered blanks and after blank spectra are removed, the resulting spectra c) displays m/z values that are uniquely found in samples. The x-axis *m/z* scale in the processed spectra is smaller than previous scales because *m/z* values above 250 Da were also present in media controls and successfully removed by BLANKA.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

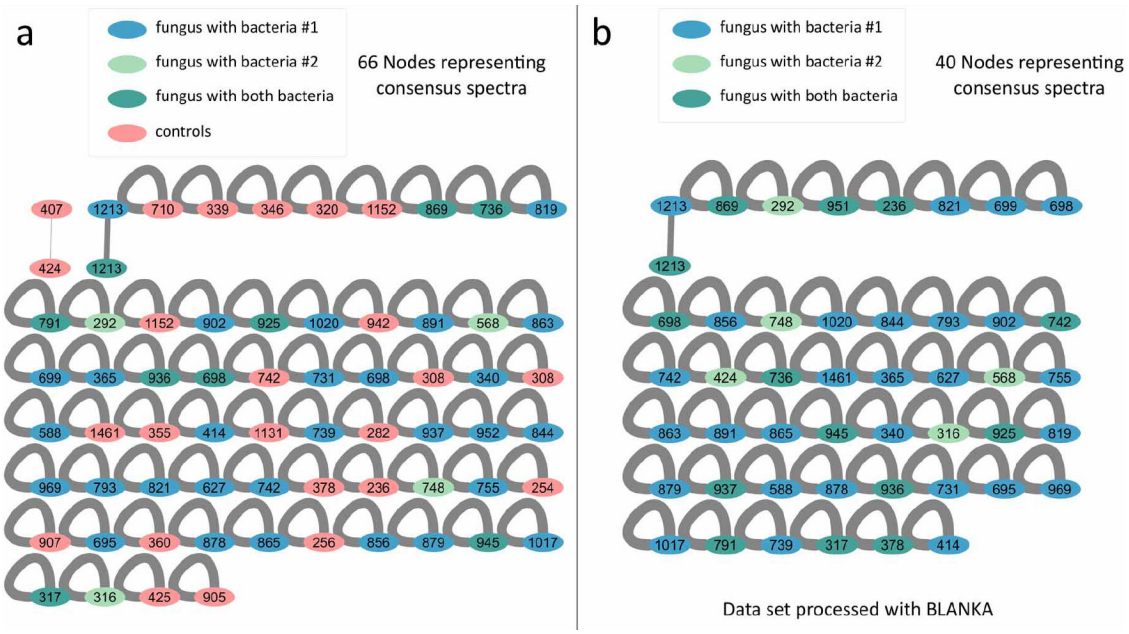


Figure 3.

Molecular networks using GNPS platform. Nodes represent consensus fragmentation spectra and are labeled with nominal precursor ion masses and color coded according to data sets. a) Data files with controls were input into GNPS and b) data files that were pre-processed with BLANKA were input and control files were left out. BLANKA setting for this instrument were defined as a rt tolerance of 10 seconds and precursor ion mass tolerance of 1.0 Da.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

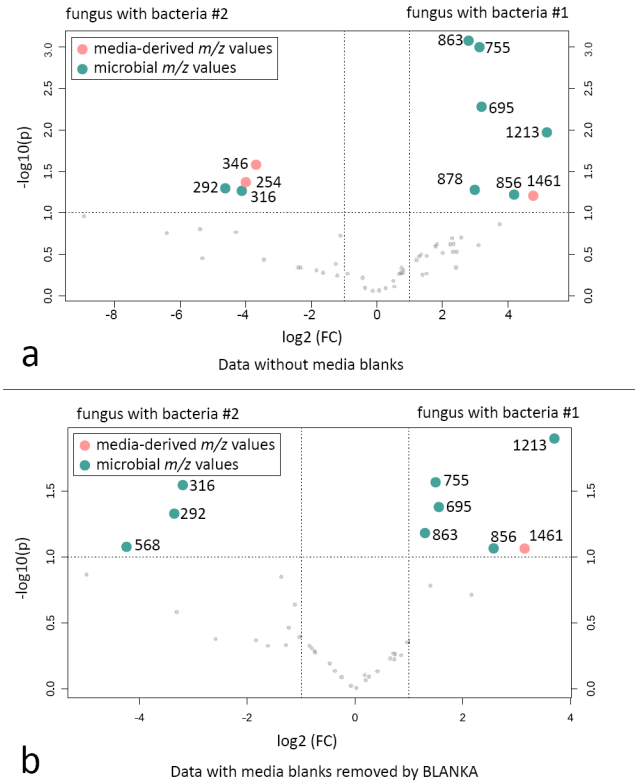


Figure 4.

Volcano plots of LC-MS/MS data show fold change between two sample sets on the x axis and p-values on the y axis. Data was exported from GNPS networks and reformatted for input into MetaboAnalyst (available in Online Resource 1). a) Fold changes between two different conditions highlights *m/z* values that are most significantly different between the two data sets. As only two conditions can be considered, samples are directly compared without media blank data. This plot represents the data from the molecular network in Figure 3a while b) represents the data from the molecular network in Figure 3b. Removal of media blank *m/z* values from all samples results in a similar but different set of *m/z* values from the original volcano plot and eliminated two *m/z* values from the original plot that were media signals.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1.

Comparison of processed and unprocessed data displays a reduction in the amount of metabolites that are considered in analyses. LC-MS/MS qTOF data was filtered to consider only *m/z* values from 200–2000 Da.

Data input	**# of Library hits in GNPS**	**# of MS/MS clusters**	***m/z* values in volcano plots identified as significant**
3D ion trap Unprocessed data	0	66	1461, 1213, 878, 863, 856, 755, 695, 346,316,292,254
3D ion trap Processed with BLANKA	0	40	1461, 1213, 863, 856, 755, 695, 568, 316,292
qTOF Unprocessed data	50	1675	-
qTOF Processed with BLANKA	40	1029	-