analytical chemistry

Article

pubs.acs.org/ac

Autonomous Metabolomics for Rapid Metabolite Identification in Global Profiling

H. Paul Benton,[†] Julijana Ivanisevic,[†] Nathaniel G. Mahieu,[‡] Michael E. Kurczy,[†] Caroline H. Johnson,[†] Lauren Franco,[§] Duane Rinehart,[†] Elizabeth Valentine,[#] Harsha Gowda,[†,¶] Baljit K. Ubhi,[∫] Ralf Tautenhahn,[†,∥] Andrew Gieschen,[⊥] Matthew W. Fields,[§] Gary J. Patti,*[,‡] and Gary Siuzdak*

[†]Scripps Center for Metabolomics and Mass Spectrometry, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, United States

[‡]Departments of Chemistry, Genetics and Medicine, Washington University, One Brookings Drive, St. Louis, Missouri 63130, United States

[§]Department of Microbiology and Immunology and Center for Biofilm Engineering, Montana State University, 109 Lewis Hall, Bozeman, Montana 59717, United States

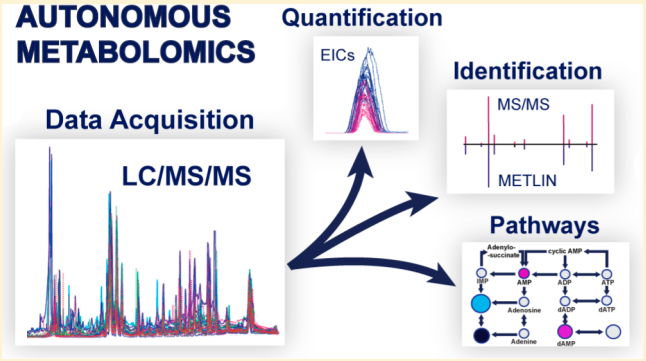
[∫]AB SCIEX, 1201 Radio Road, Redwood City, California 94065, United States

[⊥]Agilent Technologies, 11011 North Torrey Pines Road, La Jolla, California 92037, United States

[#]The Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, United States

Supporting Information

ABSTRACT: An autonomous metabolomic workflow combining mass spectrometry analysis with tandem mass spectrometry data acquisition was designed to allow for simultaneous data processing and metabolite characterization. Although previously tandem mass spectrometry data have been generated on the fly, the experiments described herein combine this technology with the bioinformatic resources of XCMS and METLIN. As a result of this unique integration, we can analyze large profiling datasets and simultaneously obtain structural identifications. Validation of the workflow on bacterial samples allowed the profiling on the order of a thousand metabolite features with simultaneous tandem mass spectra data acquisition. The tandem mass spectrometry data acquisition enabled automatic search and matching against the METLIN tandem mass spectrometry database, shortening the current workflow from days to hours. Overall, the autonomous approach to untargeted metabolomics provides an efficient means of metabolomic profiling, and will ultimately allow the more rapid integration of comparative analyses, metabolite identification, and data analysis at a systems biology level.



Untargeted metabolomic experiments performed on complex biological matrices with high-resolution, high-throughput, and sensitive mass spectrometry (MS) technology have enabled the detection of thousands of metabolite features from a single experiment.[1−4] However, the identification of these features presents a major bottleneck in the metabolomics workflow. It is not only a time-consuming process, taking weeks to carry out, but often results in a low yield of correctly identified metabolites. This is partly due to the manual interpretation required by the investigator and also the number of metabolites currently characterized in metabolite databases.[5] This process can be potentially shortened by integrating metabolite profiling and identification into a single autonomous workflow.

Current metabolomic studies typically adopt a multistep workflow (Figure 1). Comparative profiling is first carried out in MS mode; the data is then processed using bioinformatics software such as XCMS[6,7] to reveal features of interest that show statistically significant differences. The features are then subjected to tandem mass spectrometry (MS/MS) acquisition and identified through MS/MS matching to standards in metabolite databases such as METLIN.[8,9] This conventional approach typically involves the generation of a list of dysregulated metabolite features from an initial set of experiments followed by statistical analysis and manual selection of precursor ions, which are then fragmented to obtain mass spectra used for metabolite characterization.[10]

Received: July 11, 2014
Accepted: December 12, 2014

DOI: 10.1021/ac5025649
Anal. Chem. XXXX, XXX, XXX−XXX

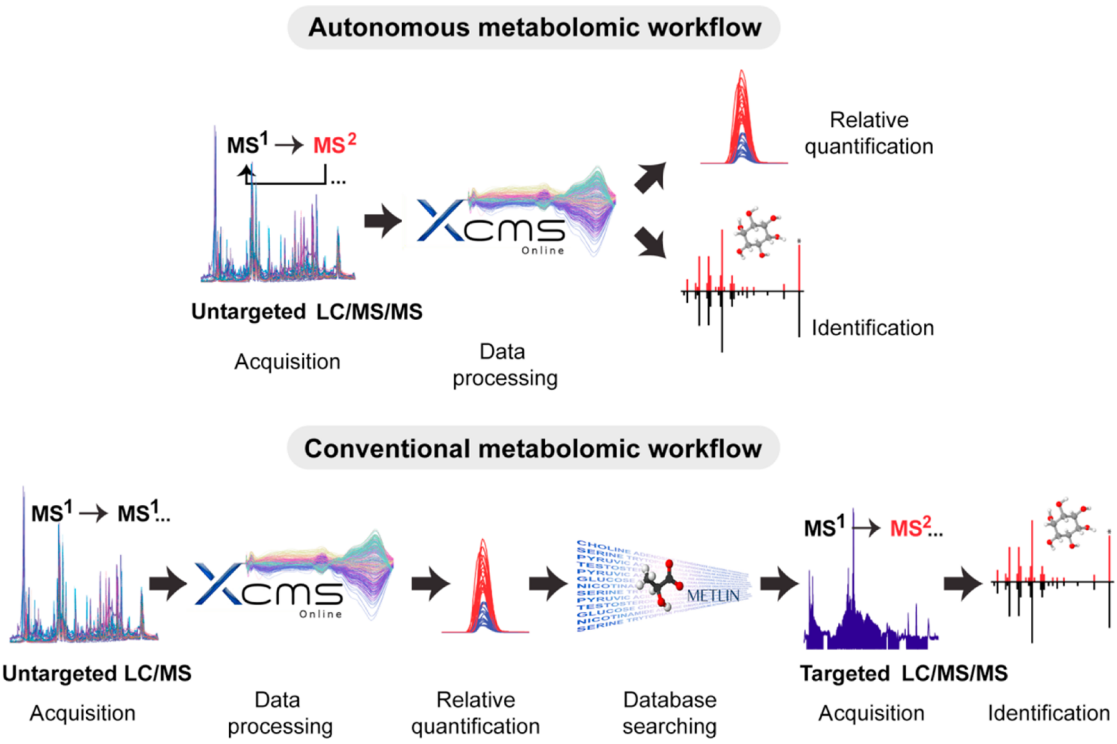


Figure 1. An autonomous vs conventional mass spectrometry-based metabolomic workflow. The autonomous workflow is based on parallel untargeted MS^1 and MS^2 data acquisition, where cycle time is optimized as a compromise for high peak definition (MS^1 for comparative quantitative analysis) and high quality MS^2 (for MS/MS matching to facilitate metabolite identification). The conventional workflow, the untargeted MS^1 data acquisition for comparative quantitative analysis and targeted MS^2 data acquisition of dysregulated features of interest (for metabolite identification) are performed in two subsequent steps. The autonomous workflow can be accomplished through the direct link between XCMS software and METLIN metabolite database.

Similar to the early days of proteomics, this time-consuming metabolomic workflow relies heavily on investigator input and manual data analysis.

To facilitate a more efficient and autonomous approach toward 'omic scale profiling, we have been developing XCMS Online, a software package for untargeted metabolomics to enable the integration of metabolite profiling and identification into one step (Figure 1). XCMS Online has integrated algorithms for peak detection, retention time correction, chromatogram alignment, and relative quantification based on peak area. In addition, a number of univariate and multivariate statistical tools have been incorporated into the workflow to aid in elucidating the features that have the highest statistical significance and drive sample clustering.[11] However, one of the most important aspects of XCMS Online is its integrated metabolite database, METLIN, a spectral library with MS/MS data available for more than ten thousand endogenous and exogenous metabolites, enabling automatic putative identification of metabolites.

The basis of the autonomous metabolomic workflow is high resolution sequential MS and MS/MS data acquisition (Figure 1). High-quality MS and MS/MS data are a prerequisite to extract accurate quantitative information and simultaneously characterize metabolite features by MS/MS data matching.[12] This one-step autonomous approach requires coordinated acquisition of MS and MS/MS data to ensure efficient and accurate comparative analysis and identification. Mass spectrometers that have high scan speeds and can alternate between MS and MS/MS modes in the same run are essential for this type of simultaneous data acquisition. Indeed, most modern quadrupole-time-of-flight (Q-TOF) instruments have high sensitivity and rapid scan speeds ensuring the acquisition of a sufficient amount of data points across a chromatographic peak while simultaneously acquiring fragmentation data which is necessary for metabolite characterization.

Key to the autonomous approach has been the development, over the last ten years, of databases to facilitate identification of metabolites in metabolomic experiments. The METLIN tandem mass spectrometry database,[13,14] Human Metabolome Database (HMDB),[15] MassBank,[16] and LipidMaps[17] databases have been generated largely from the analysis of pure standard compounds to provide for accurate identification of metabolites from biological samples. Most of these repositories allow researchers to compare MS/MS data from their research samples to MS/MS data recorded from standards one at a time for metabolite identification.

Here, we have designed an autonomous untargeted metabolomic workflow that acquires MS and MS/MS data sequentially. Quantitative information is extracted from MS data using XCMS Online and metabolite features are simultaneously characterized by matching MS/MS data to the METLIN database. The effectiveness of this workflow is demonstrated using standard mixture and a bacterial extract samples.

EXPERIMENTAL SECTION

Chemicals. Ammonium acetate (NH_4Ac) and ammonium hydroxide (NH_4OH) were purchased from Sigma-Aldrich (St. Louis, MO, USA). LC-MS grade methanol (MeOH) was purchased from Honeywell (Muskegon, MI, USA). LC-MS

DOI: 10.1021/ac5025649
Anal. Chem. XXXX, XXX, XXX−XXX

grade acetonitrile was purchased from Fisher Scientific (Morris Plains, NJ, USA) and water was purchased from J.T. Baker (Phillipsburg, NJ, USA).

Bacterial Strain and Biofilm Growth Conditions. *Desulfovibrio vulgaris* Hildenborough (DvH) was obtained from Dr. Romy Chakraborty (Lawrence Berkeley National Lab). DvH was grown in LS4D medium, which contains lactate as the carbon source and electron donor and sulfate as the electron acceptor[18] (modified to 2.5 μM resazurin and 130 μM riboflavin in modified Thauer's vitamins stock).[19] Lactate and sulfate concentrations were altered to create balanced and electron acceptor limited conditions. The balanced condition was defined as 60 mM sodium lactate and 30 mM sodium sulfate and the electron acceptor limited condition was defined as 50 mM sodium lactate and 10 mM sodium sulfate. DvH was grown as a biofilm under continuous flow conditions in a modified CDC reactor. Exponential phase cells were inoculated into a reactor containing balanced or electron acceptor limited LS4D medium and were grown in batch mode for 48 h. Reactors were grown at room temperature (20−23 °C), with a dilution rate of 0.04^{-hr}, stirred at 60 rpm, and the headspace was continually sparged with sterile N_2 gas to maintain anaerobic conditions. Coupons of aclar (7.8 mil thickness) (Electron Microscopy Sciences, Hatfield, PA) supported by glass slides were submerged in the reactor body as a surface for biofilm growth. Aclar coupons were removed from the reactor at 144 h, rinsed three times in ice-cold phosphate buffered saline, wicked dry, and flash frozen in liquid nitrogen. Samples were stored at −70 °C until further analysis.

Growth of Credentialed *Escherichia coli* Standards. *Escherichia coli* cultures were grown in a rotary shaker at 37 °C. A preculture of K12 MG1655 strain was grown in LB broth for 16 h. M9 minimal salts were prepared and 100 mL aliquoted into 2 sterile 1 L Erlenmeyer flasks. To each aliquot 2 mL of 200 mg/mL glucose was added via fresh filtered syringe. One aliquot received $U^{13}C$-labeled glucose and the second received natural abundance glucose. Each aliquot was then inoculated with the equivalent of 1 mL of OD600 = 0.6 preculture as determined by dilution. Cultures were grown to OD600 = 1.0 at which point they were harvested. Cell quench solution of 10 mM ammonium acetate in 3:2 methanol/water was aliquoted into 50 conical tubes, 30 mL per tube. Tubes were chilled to −60 °C using a dry ice and 60% ethylene glycol in ethanol bath. Upon reaching OD600 = 1.0 flasks were removed from the shaker and rapidly aliquoted into the prechilled tubes. Pairs of tubes were prepared, with one being mixed at 8:11 and the other 11:8 ^{12}C culture to ^{13}C culture, respectively. Cells were pelleted by centrifugation at 0 °C and 3200 RCF for 15 min.

Metabolite Extraction. Bacteria biofilm and cell pellets (DVH and *E. coli*) were extracted using a MeOH/ACN/H_2O (2:2:1, v/v) solvent mixture. A volume of 1 mL of cold solvent was added to each pellet, vortexed for 30 s and incubated in liquid nitrogen for 1 min. Samples were allowed to thaw and sonicated for 10 min. This cycle of cell lysis in liquid nitrogen combined with sonication was repeated three times. To precipitate proteins, the samples were incubated for 1 h at −20 °C, followed by 15 min centrifugation at 13,000 rpm and 4 °C. The resulting supernatant was removed and evaporated to dryness in a vacuum concentrator. The dry extracts were then reconstituted in 100 μL of ACN:H_2O (1:1, v/v), sonicated for 10 min and centrifuged for 15 min at 13000 rpm and 4 °C to remove insoluble debris. The supernatants were transferred to HPLC vials and stored at −80 °C prior to LC/MS analysis.

LC/MS/MS Analysis. Analyses were performed using an HPLC system (1200 series, Agilent Technologies) coupled to a quadrupole time-of-flight (AB Sciex TripleTOF 5600, and Agilent Q-TOF 6530). Samples were analyzed using a Phenomenex (Torrance, CA, USA) Luna Aminopropyl column for HILIC-MS analysis.[20] The mobile phase A = 20 mM NH_4Ac and 20 mM NH_4OH in 95% water and B = 95% acetonitrile was used. A linear gradient elution from 100% B (0−5 min) to 100% A (45−50 min) was used. The 10 min postrun was applied to ensure column re-equilibration and maintain the reproducibility. The total gradient time was 60 min. The flow rate was 50 μL/min and the sample injection volume was 5 μL. DVH biofilm samples and standard mixture were analyzed using AB Sciex TripleTOF 5600 instrument whereas credentialed *E. coli* standards were analyzed using Agilent Q-TOF 6530.

ESI source conditions on TripleTOF were set as following: Ion Source Gas 1 (GS1) as 35, Ion Source Gas 2 (GS2) as 35, Curtain gas (CUR) as 30, source temperature 550 °C, IonSpray Voltage Floating (ISVF) −4500 V in negative mode. In MS only acquisition, the instrument was set to acquire over the m/z range 50−1000 Da with TOF MS scan, and the accumulation time for TOF MS scan was set at 0.25 s/spectra. In auto MS/MS acquisition, the instrument was set to acquire over the m/z range 50−1000 Da for TOF MS scan and the m/z range 25−1000 for product ion scan. The accumulation time for TOF MS scan was set at 0.25 s/spectra and product ion scan at 0.05 s/spectra (cycle time = 1 sec). The product ion scan is acquired using information dependent acquisition (IDA) with high sensitivity mode selected. IDA triggers MS/MS during the full scan experiment based on a set of criteria that the user inputs. The unit resolution is selected for precursor ion selection, and the collision energy (CE) was fixed at −30 V with ±15 spread. Declustering potential (DP) was set as −100 V. IDA settings were set as following: charge state 1 to 1, intensity 500 cps, exclude isotopes within 4 Da, mass tolerance 10 ppm and maximum number of candidate ions 15. The "exclude former target ions" was set as 9 s after 2 occurrences. In IDA Advanced tab, "dynamic background subtract" was also chosen.

ESI source settings on Agilent 6530 instrument were as follows: gas temp, 250 °C; gas flow, 6 L/min; nebulizer pressure, 25 psi; sheath flow, 9 L/min; sheath temperature, 350 °C; fragmentor, 120 V; capillary, 2000 V; nozzle, 1500 V. All MS and MS/MS acquisition settings were as listed above for TripleTOF.

Data Processing of DVH Biofilm Samples. The raw MS data (wiff.scan files) were converted to mzXML files using ProteoWizard MSConvert and processed using XCMS for feature detection, retention time correction and alignment. The parameters in XCMS were set as follows: centWave settings for feature detection (Δ m/z = 15 ppm, minimum peak width = 10 s and maximum peak width = 120 s) and mzwid = 0.015, minfrac = 0.5, and bw = 5 for chromatogram alignment. After careful evaluation retention time alignment was shown not to be required. Isotopic peaks and adducts were detected using CAMERA.[21] Paramater ID was 12480 and the data sets are shared as public data sets ID: 1039270 (Auto MS/MS job) and 1039276 (MS Only job). Each mzXML file was processed separately to ensure features were not missed. For each file a global noise intensity level was calculated using the XCMS noise function on each scan. The median from the distribution was used as a filter to remove peak below this value in each spectrum. The precursor was then matched to METLIN at 20

DOI: 10.1021/ac5025649
Anal. Chem. XXXX, XXX, XXX−XXX

ppm accuracy. The database currently contains MS/MS data for more than 12 000 metabolites and 61 000 MS/MS spectra. Matches with MS/MS data were scored using a cosine similarity metric (0−1) to the closest collision energy in the database. Spectral mirror plots were created for conformation of positive matches.

Data Processing of Credentialed *E. coli* Standards. A pair of *E. coli* samples (8:11 and 11:8) were analyzed as above in both MS1 only mode and auto MS/MS mode. These samples were credentialed using the github pattilab/credential R package (http://pattilab.wustl.edu/software/credential/). Parameters used were mixed_ratio_factor = 4, mixed_ratio_ratio_factor = 1.8, mpc_factor = 1.1. The list of credentialed features was then formatted as a preferred MS/MS list and the samples were analyzed again in Preferred Auto MS/MS mode.

RESULTS AND DISCUSSION

Autonomous Workflow Optimization: Auto MS/MS method. The overall design of the autonomous workflow approach is to simultaneously acquire MS and MS/MS data during each LC/MS run (Figure 1). The quality of both MS and MS/MS data were of equal importance and a balance between the amounts of time spent on the acquisition of each needed to be made. While MS/MS data were necessary to generate fragmentation patterns for metabolite identification, too long accumulation times to maximize signal-to-noise and collect high quality MS/MS data can compromise MS data quality. MS data acquisition is crucial for metabolite feature detection, profile alignment, comparative and statistical analyses; however excessive time acquiring MS data could limit the number of MS/MS spectra and weaken MS/MS spectral quality. Fortunately, modern Q-TOF mass spectrometers have fast scan rates and high sensitivities making it possible to collect both MS and MS/MS data simultaneously, throughout the same run, without compromising overall data quality.

In this study, biological samples were extracted and analyzed on a Q-TOF mass spectrometer coupled to a capillary LC. The MS/MS spectra were generated using automated Information Dependent Acquisition (IDA), whereby the instrument software continuously evaluates the profile of the precursor ions (MS data) in the full survey scan and triggers MS/MS acquisition based on a set of criteria that the user inputs into the method. Instrument settings including acquisition rate, number of precursor candidate ions, collision energy, and dynamic exclusion time were optimized for IDA. In the optimized Auto MS/MS method, MS/MS acquisition was triggered on selected precursor ions that met the defined criteria of intensity and charge state. A cycle time of 1 s was used for data acquisition where each MS scan was followed by 15 MS/MS events. MS data were acquired at 4 Hz (250 ms accumulation time for each spectrum) while MS/MS data were acquired at 20 Hz (50 ms accumulation time for each spectrum, and 750 ms in total for 15 MS/MS events). This ensured an approximate duty cycle of 25% on MS acquisition while simultaneously acquiring up to 15 MS/MS spectra in the remaining time period. Chromatographic base peak widths for ions in HILIC HPLC mode were 30 s on average. A cycle time of 1 s ensured at least 15−30 MS data points acquired for each extracted ion chromatogram (EIC) peak. Multiple MS data points along the chromatographic peak are crucial for reliable peak detection, peak area integration, comparative and statistical analyses. The accumulation time for both MS and MS/MS acquisitions by UHPLC separation can be reduced 3 to 5 times as features tend to be 3 to 5 times narrower when compared to capillary LC separation. However, shortening the accumulation time compromises the sensitivity and data quality.

Some of the crucial parameters for precursor ion selection in IDA mode included a precursor ion intensity threshold of 500 counts, monoisotopic precursor selection and charge state screening to exclude any feature that was not singly charged (see Methods). Dynamic exclusion in IDA mode has been enabled to further reduce the redundancy of precursor ion selection and expand the MS/MS acquisition coverage. Every precursor ion that qualified using the preset criteria, was selected for fragmentation two consecutive times, and was dynamically excluded for the next nine seconds. Repeated acquisition on the same precursor ion and short exclusion times were chosen to maintain sufficient redundancy for MS/MS data acquisition providing higher quality data.

Validation of Autonomous Workflow: Auto MS/MS versus MS Only. To test the efficacy of our autonomous workflow a complex biological matrix from bacterial biofilm culture was investigated. *Desulfovibrio vulgaris* Hildenborough (DvH) biofilm samples were grown in balanced (BAL, $n = 4$) and in electron acceptor limited conditions (EAL, $n = 4$) to obtain insight into the altered metabolic activity and understand how it could be associated with a particular phenotype. The metabolic response of DVH biofilms grown in EAL conditions versus balanced conditions were characterized and compared using the optimized MS sampling rate and duty cycle as described above. Analysis was carried out in both MS Only mode (i.e., only MS^1 acquisition) and Auto MS/MS mode (i.e., MS^1 and MS^2 acquisition). Out of total aligned features, thirty six percent of the features that were observed using MS only method were also observed using the auto MS/MS method (Figure 2A). Forty nine percent of additional unique features detected by MS only method were likely due to frequent MS^1 data acquisition, and therefore higher MS^1 signal intensity and peak definition. Furthermore, 29% of the dysregulated features (p-value ≤ 0.05, intensity ≥ 1000) detected with the MS only method were also observed using the Auto MS/MS acquisition (Figure 2A). The other 28% were related to chemical noise and false positives or negatives, likely because of lower peak definition quality. Fold-change is an important parameter to characterize the degree of metabolite change in perturbed biological system. An example of one down-regulated metabolite feature characterized by m/z 132.030 and retention time of 24.6 min acquired with both methods was used to demonstrate the consistency of the Auto MS/MS method compared to the MS only method (Figure 2B). Both acquisition methods showed similar fold change and statistical significance (Welch t test). MS/MS data acquired using the Auto MS/MS method allowed for identification of the selected metabolite feature as aspartic acid. Together, these data demonstrated that Auto MS/MS acquisition with a 25% duty cycle for MS data collection can reduce data quality for metabolite feature detection, although the majority of dysregulated features from comparative analysis were correctly assigned and identified. In general, the results from the autonomous or conventional untargeted workflow should be validated using targeted quantification of identified dysregulated metabolites of interest.[22]

Autonomous Metabolomic Analysis of Bacterial Biofilm. Further data analysis of DVH biofilm grown in EAL conditions and using Auto MS/MS acquisition in negative

DOI: 10.1021/ac5025649
Anal. Chem. XXXX, XXX, XXX−XXX

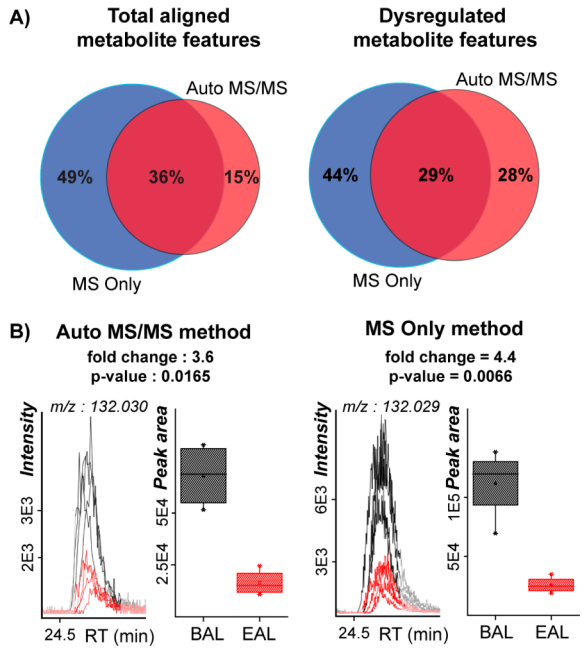


Figure 2. Characterization of the metabolic response of *Desulfovibrio vulgaris* (DVH) biofilm grown in the electron acceptor limited (EAL) conditions, using both auto MS/MS and MS-only acquisition method. (A) The overlap between metabolite features acquired with auto MS/MS method vs MS only method. Dysregulated metabolite features were defined using following parameters: p-value ≤ 0.05, intensity ≥ 1000. (B) Metabolite feature characterized by m/z 132.030 and retention time 24.6 min (MS/MS match, aspartic acid), its extracted ion chromatograms (EICs), p-values and fold-changes when acquired with auto MS/MS and MS only method. Red lines, biofilm grown in EAL conditions ($n = 4$); black lines, biofilm grown in balanced conditions ($n = 4$).

ionization mode revealed 34% of significantly dysregulated metabolite features (339 features with p-value ≤ 0.05, and intensity ≥ 1000) out of total aligned metabolite features (Figure 3A). Sixty seven of these had MS/MS matches (20%) to the METLIN metabolite database and 46 were identified using matching score ≥ 0.5 (Figure 3B). Of these, 36 had KEGG IDs which were used for biochemical pathway mapping using KEGG small molecules database. The results of pathway mapping suggested several perturbed pathways at the cell level, relevant to DVH as a response to biofilm growth in EAL conditions. The affected purine metabolism was chosen as an example to demonstrate the effectiveness of the autonomous workflow for identifying biologically relevant metabolites (Figure 3C). Specifically, 6 metabolites were mapped onto the purine metabolism pathway (Figure 3C) and all of which were down-regulated as much as 6.5 fold (e.g., hypoxanthine). Statistical information (e.g., fold-change and p-value) and identification (MS/MS spectral match, METLIN, and KEGG ID) can be extracted from data set (acquired in Auto MS/MS mode) for each metabolite in the pathway to facilitate interpretation (Figure 3B). As shown here, the development of the autonomous metabolomic workflow offered simultaneous metabolite profiling and metabolite identification, further supporting the use of metabolomics as a systems biology tool.

A false discovery rate (FDR) used in proteomics is not feasible here for the MS/MS spectral matching as no appropriate decoy metabolite databases can be constructed,[23] therefore the results need to be further manually assessed by the user. The relatively low match percentage (up to 20%) is likely due to incompleteness of the METLIN database and metabolite databases in general. However, METLIN is rapidly expanding and the number of successful matches is likely to improve over time. In the past 3 years METLIN has added more than 6,000 metabolites with MS/MS data. As of September 2014 METLIN housed more than 12 000 compounds with MS/MS data. This demonstrates the importance of having a comprehensive metabolite database and its impact on the metabolite identification process. As more and more fragmentation data from metabolites and their derivatives are deposited into METLIN, we expect that our ability to identify metabolites from biological samples will greatly improve. The scoring algorithm can also affect the results. Here, we have used the cosine similarity score for its ease of implementation and interpretation. However, there are many other scoring algorithms in use today, such as X-rank[24] weighted cosine,[16] and others.

Addressing the Coverage of Autonomous MS/MS Spectra Acquisition. Additional experiments were performed to demonstrate the effectiveness of the autonomous workflow for metabolite identification through METLIN tandem mass spectral matches. In addition to the limitation presented by the incomplete database two major factors limit the number of matched features; MS/MS spectral coverage and the match percentage. To determine the coverage of the autonomous metabolomic approach, we measured a standard metabolite mixture (40 metabolites, Supporting Information Table 1) by optimized Auto MS/MS method, and the results revealed that MS/MS spectra were acquired for 85% of the metabolites present in the standard mixture. All of these metabolites could be identified through MS/MS matching against METLIN metabolite database. The future improvements in instrument scan speed as new instruments are developed will likely increase this already high MS/MS spectral coverage. However, increasing the scan speed comes at the detriment of acquiring good quality MS/MS spectra, therefore, in our case; a 20 Hz acquisition rate was employed as a compromise to achieve higher numbers of metabolite identifications (Figure 4).

To maximize scan time, and therefore maximize MS/MS spectral quality, we opted to narrow down the number of candidate features via "credentialing" or defining the real, biologically relevant nonartifactual features. In an autonomous metabolomic experiment, the objective is to acquire MS/MS data for all monoisotopic ions of biological origin. Out of the thousands of ions that are typically detected in an untargeted metabolomic experiment only a small percentage of these will be relevant, biological ions. Selecting the appropriate ions for MS/MS analysis, thus, is a challenge. As demonstrated herein, modern instruments are capable of intelligently excluding isotopes, charge states, and low abundance ions from MS/MS analysis. Still, using auto MS/MS, many ions will be analyzed which are irrelevant to the experiment. This is due to many monoisotopic, high abundance ions which are the result of contamination and chemical noise rather than the biological sample. These peaks waste scans, taking time away from true, relevant metabolites and effectively decreasing MS/MS coverage. To investigate and address the challenge of coverage in auto MS/MS experiments we utilized a credentialing approach. In this approach we performed auto MS/MS on an extract of *E. coli* specifically prepared for the credentialing informatic workflow. This workflow allows the distinction of biological

DOI: 10.1021/ac5025649
Anal. Chem. XXXX, XXX, XXX−XXX

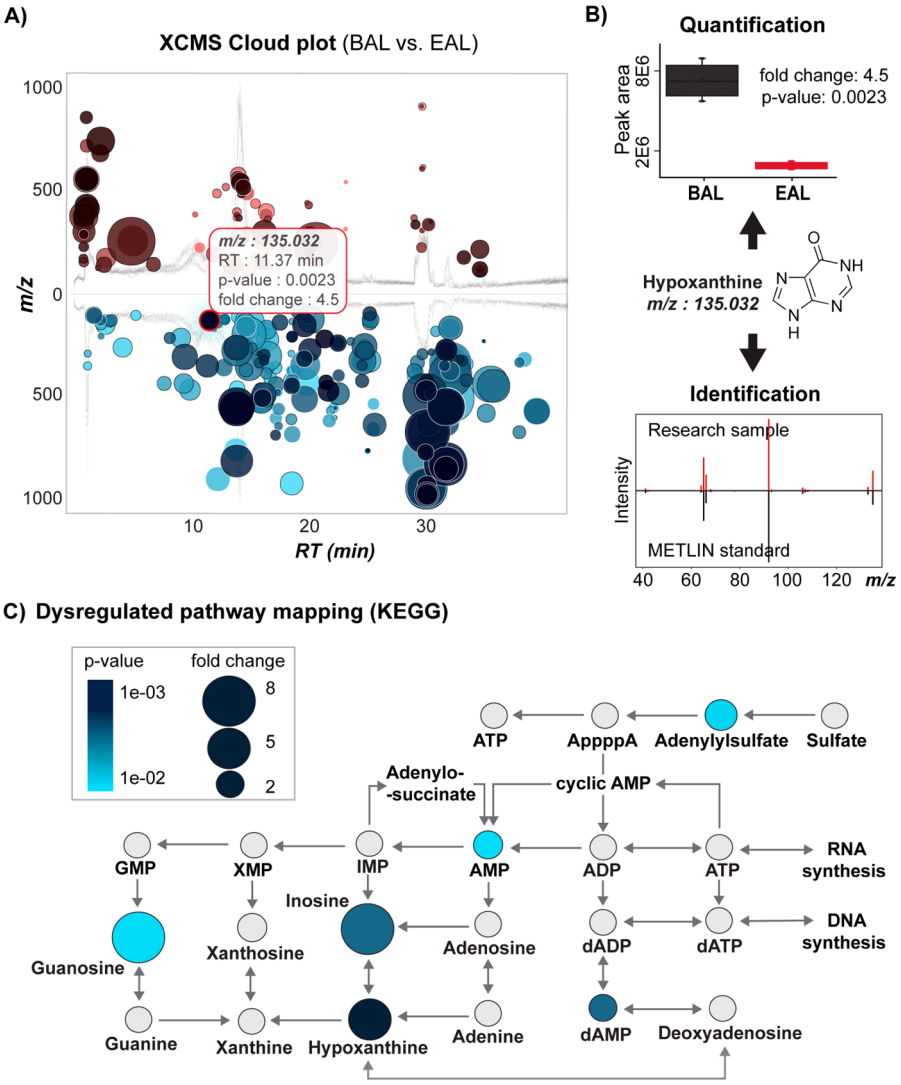


Figure 3. Autonomous metabolomic approach for simultaneous comparative analysis and identification of metabolites. (A) XCMS Cloud plot representation of the dysregulated metabolite features from *Desulfovibrio vulgaris* biofilm grown in electron acceptor limited conditions: red bubbles represent up-regulated features and blue bubbles represent down-regulated features. (B) Relative quantification and MS/MS matching identification of hypoxanthine as intermediary metabolite in purine metabolism. (C) Pathway mapping of dysregulated metabolite features using KEGG. Identified metabolites and their role in the purine metabolism pathway: blue circles represent down-regulated metabolites and the size of the circle represents the fold change.

features and artifactual features (such as contamination and chemical noise).[25]

Using this information, we assess the coverage of Auto MS/MS and whether this coverage can be improved with guided precursor ion selection (preferred auto MS/MS). As shown in Figure 5, after deisotoping and credentialing, the 12 420 total original features decreased to 836 credentialed features. The removed artifactual features correspond to various noise sources, contamination, and isotopes. The remaining 836 credentialed features represent the biologically relevant signals of which MS/MS data is desired.

The initial experiment resulted in 2606 features being chosen during auto MS/MS mode, from the 12 420 total features, of these, 2288 (88%) were artifactual. After credentialing, only 836 features were detected compared to 12 420, and out of these 836, only 38% (318 features) were picked up by the original auto MS/MS experiment without credentialing. This low coverage of credentialed features, and the high number of noncredentialed features assayed, illustrates the challenge of prioritizing relevant ions for analysis.

To improve the coverage of relevant ions we applied a preferred ion list to guide MS/MS precursor selection. The 836 credentialed features were imported as a preferred ion list and the auto MS/MS experiment repeated. After addition of the preferred ion list 660 credentialed features were analyzed by MS/MS corresponding to a 207% increase. This represents 342 new credentialed features which were analyzed by MS/MS and a large increase in coverage of our method. The use of a credentialed, preferred auto MS/MS list is a valuable addition to the autonomous metabolomic workflow for any sample type. Biological ions from a standard *E. coli* extract are easily annotated using the feature credentialing workflow.[25] These credentialed features are then used to generate a preferred ion list and included in the autonomous method. While this preferred list was generated using *E. coli*, it is applicable to any biological sample. Preferred ions are only analyzed if they are present at the proper retention time, so extraneous ions in the

DOI: 10.1021/ac5025649
Anal. Chem. XXXX, XXX, XXX−XXX

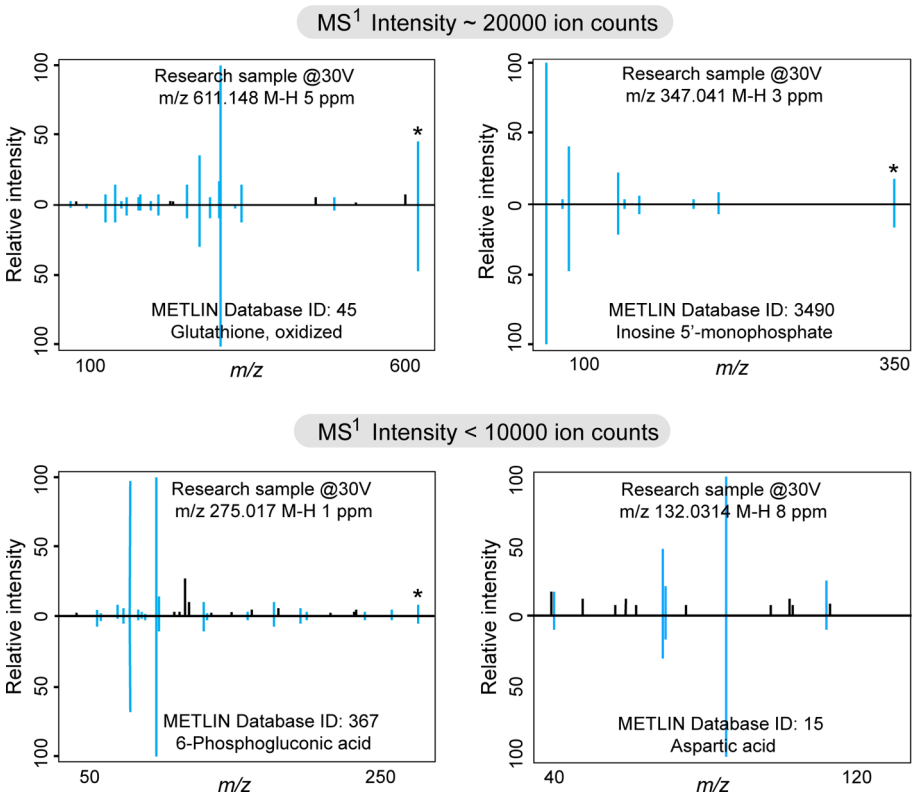


Figure 4. MS/MS spectral matching of the MS/MS data, acquired "on-the-fly", against METLIN metabolite database. MS/MS data were acquired at 50 ms scanning rate per spectra.

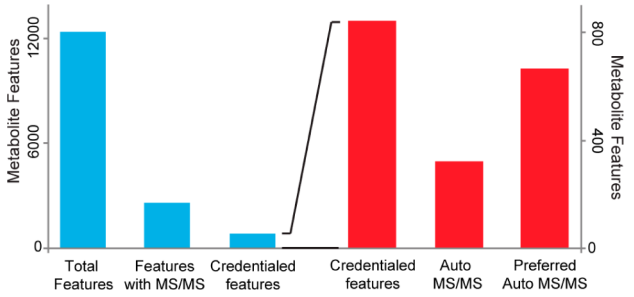


Figure 5. Preferred auto MS/MS increased MS/MS coverage of credentialed features relative to a naive experiment. Credentialed or biologically relevant features comprise only a portion of a metabolomic data set. The utilization of credentialed features to generate a preferred ion list increased auto MS/MS coverage to 78% of credentialed features.

list will not detract from the analysis. Further, the shared metabolite content of various biological samples ensures that many metabolites (central carbon metabolism, amino acids, etc.) are preferentially analyzed over artifactual peaks.

CONCLUSIONS

An autonomous workflow that allows for simultaneous MS and MS/MS data acquisition will enable relative quantification and structural characterization for the thousands of features observed using untargeted metabolomics. High quality data acquisition can be accomplished by taking advantage of the high sensitivity, high mass accuracy and resolution, and fast scanning speeds of modern mass spectrometers. Key to this workflow is utilizing the well-established XCMS bioinformatic platform and the largest metabolite database METLIN that facilitates feature assignment and identification. Our results revealed that Integrated MS and MS/MS data acquisition in the autonomous approach significantly improved the throughput performance, reducing the amount of sample and time required for the metabolomic experiments. An autonomous metabolomic approach will thus ultimately allow the more rapid integration of comparative analyses, metabolite identification, and data analysis at a systems biology level.

ASSOCIATED CONTENT

Supporting Information

Additional material as described in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: gjpattij@wustl.edu.
*E-mail: siuzdak@scripps.edu.

Present Addresses

[¶]H.G.: Institute of Bioinformatics, International Tech Park, Bangalore 560066, India and
[‖]R.T.: Thermo Fisher Scientific, 355 River Oaks Parkway, San Jose, CA 95134.

Author Contributions

H.P.B. and J.I. contributed equally.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by US National Institutes of Health grants R01 CA170737 (G.S.), P30 MH062261 (G.S.), RC1

DOI: 10.1021/ac5025649
Anal. Chem. XXXX, XXX, XXX−XXX

HL101034 (G.S.), P01 DA026146 (G.S.) and R01 ES022181 (G.J.P.), L30 AG0 038036 (G.J.P.), and Department of Defense grant W81XWH-13-1-0402. This material by ENIGMA-Ecosystems and Networks Integrated with Genes and Molecular Assemblies (http://enigma.lbl.gov), a Scientific Focus Area Program at Lawrence Berkeley National Laboratory is also based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research under contract number DE-AC02-05CH11231 (G.S.).

REFERENCES

(1) Patti, G. J.; Yanes, O.; Siuzdak, G. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 263−269.
(2) Fiehn, O.; Kopka, J.; Dormann, P.; Altmann, T.; Trethewey, R. N.; Willmitzer, L. *Nat. Biotechnol.* **2000**, *18*, 1157−1161.
(3) Dettmer, K.; Aronov, P. A.; Hammock, B. D. *Mass Spectrom. Rev.* **2007**, *26*, 51−78.
(4) van der Greef, J.; van Wietmarschen, H.; van Ommen, B.; Verheij, E. *Mass Spectrom. Rev.* **2013**, *32*, 399−415.
(5) Johnson, C. H.; Gonzalez, F. J. *J. Cell. Physiol.* **2012**, *227*, 2975−2981.
(6) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−787.
(7) Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G. *Anal. Chem.* **2012**, *84*, 5035−5039.
(8) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. *Ther. Drug Monit.* **2005**, *27*, 747−751.
(9) Tautenhahn, R.; Cho, K.; Uritboonthai, W.; Zhu, Z.; Patti, G. J.; Siuzdak, G. *Nat. Biotechnol.* **2012**, *30*, 826−828.
(10) Zhu, Z.-J.; Schultz, A. W.; Wang, J. H.; Johnson, C. H.; Yannone, S. M.; Patti, G. J.; Siuzdak, G. *Nat. Protoc.* **2013**, *8*, 451−460.
(11) Gowda, H.; Ivanisevic, J.; Johnson, C. H.; Kurczy, M. E.; Benton, H. P.; Rinehart, D.; Nguyen, T.; Ray, J.; Kuehl, J.; Arevalo, B.; Westenskow, P. D.; Wang, J.; Arkin, A. P.; Deutschbauer, A. M.; Patti, G. J.; Siuzdak, G. *Anal. Chem.* **2014**, *86*, 6931−6939.
(12) Benton, H. P.; Wong, D. M.; Trauger, S. A.; Siuzdak, G. *Anal. Chem.* **2008**, *80*, 6382−6389.
(13) Smith, C. A.; Maille, G. O.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. *Ther. Drug Monit.* **2005**, *27*, 747−751.
(14) Tautenhahn, R.; Cho, K.; Uritboonthai, W.; Zhu, Z. J.; Patti, G. J.; Siuzdak, G. *Nat. Biotechnol.* **2012**, *30*, 826−828.
(15) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A. *Nucleic Acids Res.* **2013**, *41*, D801−D807.
(16) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom.* **2010**, *45*, 703−714.
(17) Fahy, E.; Subramaniam, S.; Murphy, R. C.; Nishijima, M.; Raetz, C. R. H.; Shimizu, T.; Spener, F.; van Meer, G.; Wakelam, M. J. O.; Dennis, E. A. *J. Lipid Res.* **2009**, *50*, S9−S14.
(18) Borglin, S.; Joyner, D.; Jacobsen, J.; Mukhopadhyay, A.; Hazen, T. C. *J. Microbiol. Meth.* **2009**, *76*, 159−168.
(19) Brandis, A.; Thauer, R. K. *J. Gen Microbiol* **1981**, *126*, 249−252.
(20) Ivanisevic, J.; Zhu, Z. J.; Plate, L.; Tautenhahn, R.; Chen, S.; O'Brien, P. J.; Johnson, C. H.; Marletta, M. A.; Patti, G. J.; Siuzdak, G. *Anal. Chem.* **2013**, *85*, 6876−6884.
(21) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. *Anal. Chem.* **2011**, *84*, 283−289.
(22) Johnson, C. H.; Ivanisevic, J.; Benton, H. P.; Siuzdak, G. *Anal. Chem.* **2014**, DOI: 10.1021/ac5040693.
(23) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. *J. Proteome Res.* **2007**, *7*, 29−34.
(24) Mylonas, R.; Mauron, Y.; Masselot, A.; Binz, P. A.; Budin, N.; Fathi, M.; Viette, V.; Hochstrasser, D. F.; Lisacek, F. *Anal. Chem.* **2009**, *81*, 7604−7610.
(25) Mahieu, N. G.; Huang, X.; Chen, Y. J.; Patti, G. J. *Anal. Chem.* **2014**, *86*, 9583−9589.

DOI: 10.1021/ac5025649
Anal. Chem. XXXX, XXX, XXX−XXX