analytical chemistry

Article

pubs.acs.org/ac

Credentialing Features: A Platform to Benchmark and Optimize Untargeted Metabolomic Methods

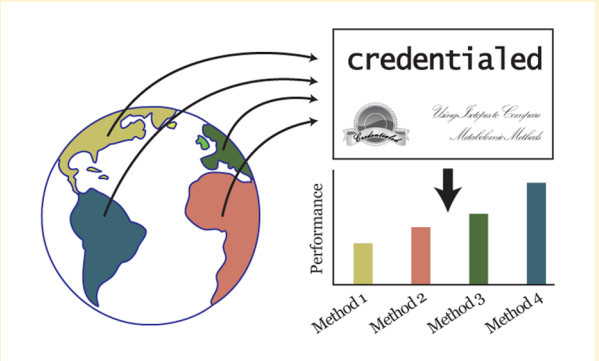
Nathaniel Guy Mahieu,[†,‡] Xiaojing Huang,[‡] Ying-Jr Chen,[†,‡] and Gary J. Patti*[,†,‡]

[†]Department of Chemistry, Washington University in St. Louis, St. Louis, Missouri 63130, United States

[‡]Departments of Genetics and Medicine, Washington University School of Medicine, St. Louis, Missouri 63110, United States

Supporting Information

ABSTRACT: The aim of untargeted metabolomics is to profile as many metabolites as possible, yet a major challenge is comparing experimental method performance on the basis of metabolome coverage. To date, most published approaches have compared experimental methods by counting the total number of features detected. Due to artifactual interference, however, this number is highly variable and therefore is a poor metric for comparing metabolomic methods. Here we introduce an alternative approach to benchmarking metabolome coverage which relies on mixed *Escherichia coli* extracts from cells cultured in regular and ^{13}C-enriched media. After mass spectrometry-based metabolomic analysis of these extracts, we "credential" features arising from *E. coli* metabolites on the basis of isotope spacing and intensity. This credentialing platform enables us to accurately compare the number of nonartifactual features yielded by different experimental approaches. We highlight the value of our platform by reoptimizing a published untargeted metabolomic method for XCMS data processing. Compared to the published parameters, the new XCMS parameters decrease the total number of features by 15% (a reduction in noise features) while increasing the number of true metabolites detected and grouped by 20%. Our credentialing platform relies on easily generated *E. coli* samples and a simple software algorithm that is freely available on our laboratory Web site (http://pattilab.wustl.edu/software/credential/). We have validated the credentialing platform with reversed-phase and hydrophilic interaction liquid chromatography as well as Agilent, Thermo Scientific, AB SCIEX, and LECO mass spectrometers. Thus, the credentialing platform can readily be applied by any laboratory to optimize their untargeted metabolomic pipeline for metabolite extraction, chromatographic separation, mass spectrometric detection, and bioinformatic processing.



The objective of untargeted metabolite profiling is to assay as many endogenous small molecules in a biological sample as possible.[1] Mass spectrometry-based metabolomics represents an established analytical platform that has been widely applied toward this goal and has already yielded many fundamental biological insights.[2−5] Nevertheless, experimental strategies to maximize the number of metabolites profiled are still being developed.[6−8] A major challenge in optimizing metabolomic methodologies has been the difficulty in comparing the number of metabolites profiled in each. Given that the size and identity of the complete metabolome is unknown, it is currently not possible to assess metabolome coverage directly. Consequently, the most common metric used to compare different experimental approaches has been the number of features detected in a sample.[6,9−12]

We show here that a method detecting a maximal number of features does not necessarily provide the greatest metabolome coverage. We present a solution for the evaluation of untargeted metabolomic method performance that enables us to distinguish between two types of features: artifactual features and biologically derived features. Artifactual features are peaks in metabolomic data that arise from contaminants, chemical noise, and bioinformatic noise. In contrast, biologically derived features are peaks that arise from metabolites in the biological sample being analyzed. We refer to the process of distinguishing artifactual features from features of biological origin as "credentialing". In the credentialing workflow (Figure 1), standard samples are prepared from *Escherichia coli* grown in either natural-abundance media or uniformly ^{13}C (U-^{13}C) enriched media. After performing metabolomic experiments utilizing the methods to be compared, our algorithm finds and credentials features based on expected isotope-intensity ratios. This number of credentialed features represents a more reliable metric of metabolome coverage than total feature count because credentialed features are known to be of biological origin and hence are representative of true metabolites. Upon optimizing our bioinformatic workflow by counting credentialed features, we reduce noise features by 15% and increase properly detected and grouped features by 20%. Further, we select several biological features for tandem mass spectrometry

Received: June 3, 2014
Accepted: August 26, 2014
Published: August 26, 2014

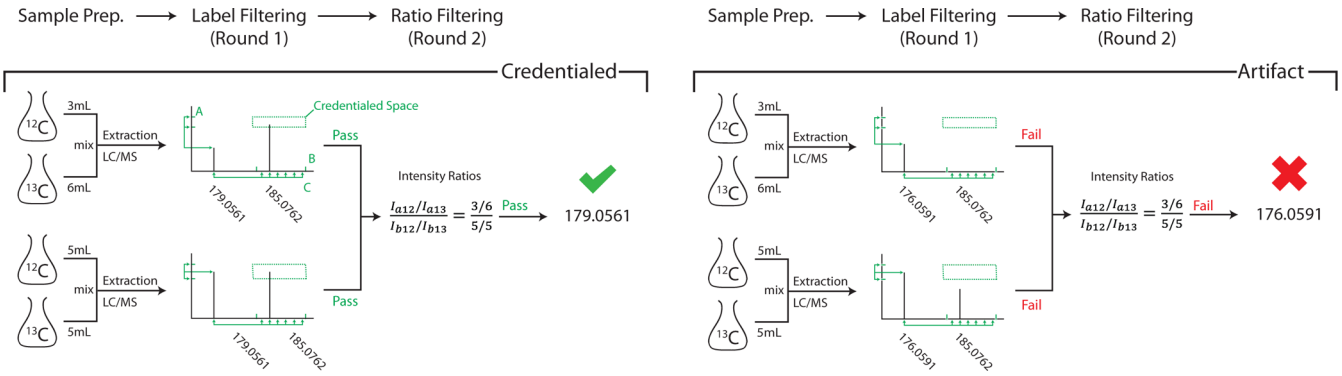


Figure 1. Overview of the feature credentialing process. A sample is generated from two cultures of *E. coli* grown in parallel, one grown on natural-abundance glucose and a second grown on ^{13}C-glucose as the sole carbon source. These two cultures are mixed in distinct ratios prior to harvesting, here 1:1 and 1:2. Extraction and LC/MS analysis is then performed on the standard samples. The resulting data are searched for pairs of coeluting peaks which satisfy the following requirements: (i) the intensities of the peaks must reflect the mixing ratio, (ii) the U-^{13}C peak must predict a feasible number of carbons for the mass in question, and (iii) the exact masses of the peaks must predict an integer number of carbons. These requirements define a "credentialed space" in which the apex of a second peak must be found to qualify as an acceptable isotope. These candidate peaks are then aligned and grouped between the two samples. Each peak pair is compared across samples and a second, stricter intensity check is performed. This requires that the ratios of each sample (I_{a12}/I_{a13} and I_{b12}/I_{b13}) are proportional to the mixed ratios of each sample. Peaks that pass these filters are considered credentialed.

(MS/MS) analysis without any prior knowledge of their identity or physiological significance. It is important to emphasize that the credentialing platform described herein is not intended to identify differences between various biological phenotypes (discovery profiling). Rather, the credentialing platform is designed only to compare the performance of different untargeted metabolomic methods. We provide a step-by-step protocol for performing credentialing with *E. coli*. While other cell types could potentially be used, *E. coli* is a simple model system whose optimized results will be applicable to the vast majority of metabolomic optimizations.

BACKGROUND

Metabolomic studies are complex, multistep experiments with a large number of parameters to optimize. The choice of sample extraction, chromatography, and ionization method strongly influences which metabolites are detected. Establishing protocols which survey the broadest number of metabolites during untargeted profiling has received detailed attention in recent years.[6,12−15] Previous studies have explored a multitude of experimental variations to improve global metabolome coverage that include the addition of ammonium fluoride and ion-pairing reagents to chromatographic mobile phases, separation strategies ranging from reversed-phase to hydrophilic interaction liquid chromatography (HILIC), different mass analyzers such as time-of-flight and the Orbitrap, and various informatic software solutions for subsequent data processing.[12,15−18] The extensive list of mutually exclusive experimental possibilities is confounding, particularly to scientists just entering the field of untargeted metabolomics. Yet, to date, comparisons of different methods have been impractical because there is no robust metric for performance evaluation.

Most published comparisons of mass spectrometry-based, untargeted metabolomic methods are evaluated by counting the total number of features detected. A feature is defined as a peak in the metabolomic data set with a unique retention time and mass-to-charge ratio. The number of features detected depends on numerous factors including sample type, metabolite extraction protocols, analyte separation, mass analyzer, and bioinformatic processing. For liquid chromatography/mass spectrometry (LC/MS)-based metabolomics, it is common to detect thousands of features from a biological sample. Importantly, a single metabolite often leads to many features[19] due to: (i) isotopic peaks from naturally occurring ^{13}C, (ii) adduct formation such as hydrogen, ammonium, and sodium adducts, (iii) neutral-loss fragments (loss of a hydroxyl group as water or a carboxylate as carbon dioxide), (iv) other fragmentation (breakage at labile bonds such as esters), (v) multiple-charge states, and (vi) chromatographic effects which result in a single metabolite eluting at more than one retention time.

Informatic solutions have been established to annotate isotopes, adducts, and neutral losses in untargeted metabolomic data sets.[17,20,21] Although these approaches are effective, they cannot distinguish signals as endogenous or artifactual. Thus, even after data reduction, a subset of the remaining features are likely the result of contaminants introduced during sample preparation, carryover from previous experiments, chemical noise, or bioinformatic error. These highly variable artifactual signals found in untargeted metabolomic data sets make it challenging to estimate the number of true biologically derived metabolites that are assayed by a particular untargeted LC/MS-based metabolomic experiment. There is therefore a great need to develop a robust metric to the evaluate performance of untargeted metabolomic methods.

EXPERIMENTAL SECTION

Our filtering process relies on the generation of standard samples derived from a mixture of *E. coli* grown on 100% natural-abundance glucose and *E. coli* grown on 100% U-^{13}C-glucose as the sole carbon source. Two standard samples are required for the filtering process; these are generated by mixing natural-abundance *E. coli* cultures and U-^{13}C-glucose *E. coli* cultures at either 5 mL/5 mL or 3 mL/6 mL ratios, respectively. The mixed *E. coli* samples are then extracted, yielding a standard sample for analysis and optimization.

Materials. U-^{13}C-D-Glucose was purchased from Cambridge Isotope Laboratories Inc. (Andover, MA). *E. coli* strain K12, MG1655 was purchased from ATCC (Manassas, VA). Lennox

dx.doi.org/10.1021/ac503092d | *Anal. Chem.* 2014, 86, 9583−9589

LB broth powder, 5× M9 salts, and all LC/MS-grade solvents were purchased from Sigma-Aldrich (St. Louis, MO). Cell culture was performed with ultrapure water provided by a Milli-Q system (Millipore).

Growth of *E. coli* Standards. Cultures were grown in a rotary shaker at 37 °C and 250 rpm. A preculture of *E. coli* was grown in LB broth for 16 h. Prior to inoculation, 3 mL of preculture was pelleted and resuspended to OD_{600} = 0.6 in M9 salts. M9 salts were prepared with the following concentrations in sterile Erlenmeyer flasks: 6.8 g/L $Na_2HPO_4{\cdot}7H_2O$; 3 g/L KH_2PO_4; 1 g/L NH_4Cl; 0.5 g/L NaCl; 240 mg/L $MgSO_4$; 11 mg/L $CaCl_2$. Salts were divided into two 100 mL aliquots, and to each aliquot, 2 mL of 20% glucose was added with a fresh-filtered syringe. The filter was rinsed with 2 mL of ultrapure water to ensure complete transfer of glucose. One aliquot received U-^{13}C-glucose and the second received natural-abundance glucose. The M9 media was then inoculated with 1 mL of the resuspended preculture per 100 mL of media. Cultures were grown to OD_{600} = 0.6, at which point they were harvested as described below.

Harvesting of *E. coli* Standards. Upon reaching OD_{600} = 0.6, flasks were removed from the shaker and placed on ice. Appropriate volumes of the ^{12}C and ^{13}C cultures were pipetted together into 15 mL centrifuge tubes, also on ice, generating samples with ratios of 1/1 of 1/2 ^{12}C/^{13}C culture. These mixtures established two distinct ratios of ^{12}C to ^{13}C feature intensities that could then be used in our credentialing algorithm, described below. Cells were pelleted by centrifugation at 2000*g* for 10 min at 4 °C. The supernatant was removed via pipet, and the cell pellets were snap-frozen in liquid nitrogen. In addition to the mixed ^{12}C and ^{13}C cultures, natural-abundance (^{12}C) cultures were used as controls. We refer to the mixed samples as "labeled" and the natural-abundance extracts alone as "unlabeled."

Metabolite Extraction. The mixed *E. coli* pellets were extracted as previously described.[6] Briefly, cells were lysed by three freeze−thaw cycles in 2/2/1 methanol/acetonitrile/water along with sonication and vortexing. The soluble portion was then vacuum concentrated and reconstituted in 100 μL of 1/1 acetonitrile/water for LC/MS analysis.

LC/MS Analysis. The data shown herein were obtained from an Agilent 6540 UHD QTOF interfaced with an Agilent 1260 Capillary LC. The column used for separation was a Phenomenex Luna NH_2 (150 mm × 1 mm, 3 μm). HILIC solvents were A, 95% water in acetonitrile with 10 mM ammonium acetate/10 mM ammonium hydroxide (pH 9.8), and B, 95% acetonitrile in water. HILIC was performed at 45 μL/min with the following linear gradient (minutes, %B): 0, 100%; 5, 100%; 45, 0%; 50, 0%; 51, 100%; 60, 100%. For all experiments, 5 μL of extract was injected. MS parameters were as follows: gas, 300 °C 9 L/min; nebulizer, 35 psi 1000 V; sheath gas, 350 °C 11 L/min; capillary, 3500 V; fragmentor, 175 V; scan rate, 1 scan/s.

To demonstrate the wide applicability of our credentialing approach to other metabolomic platforms, we also analyzed our samples and subsequently validated correct credentialing with multiple chromatographic and mass spectrometrometric technologies. In addition to the Agilent QTOF, we credentialed data from the Thermo QE, the AB SCIEX TripleTOF, and the LECO Pegasus GC-HRT. Chromatographic methods we credentialed include reversed-phase LC and HILIC. Effective parameters for credentialing each of these experimental platforms are listed in the Supporting Information Table S-3.

Data Analysis. Analysis was performed with a custom filtering script that utilizes the XCMS[16] and CAMERA[20] R[22] packages as well as the METLIN[23] database. The script is available on our laboratory Web site at http://pattilab.wustl.edu/software/credential/. The algorithm identifies features of biological origin through two rounds of data filtering, as depicted in Figure 1. Prior to filtering, features are detected from the MS raw data with the XCMS findPeaks.centWave algorithm. In the first round of filtering, coeluting peaks within a single sample are assessed for potential isotopologue pairs differing by $[(n)1.003355/z]$ Da in mass, where n is a whole number, z is the ion's charge, and the constant is the mass difference between ^{12}C and ^{13}C. Upper and lower bounds of n for each m/z in question were calculated from the distribution of mass per carbon number from the compounds in ECMDB[24] (*E. coli* Metabolome Database, Supporting Information Figure S-2). The ratios of the putative ^{12}C and ^{13}C peak intensities are then evaluated. Each measured ratio that is not within a set percentage of the mixture ratio of the ^{12}C and ^{13}C culture is disqualified. For credentialing, the default value of 400% is effective.

The two filtered samples with distinct mixture ratios of ^{12}C and ^{13}C are then taken together for a final round of filtering. Peaks from each sample are aligned and grouped. Surviving features found in both samples are evaluated such that

$$\frac{1}{e}\frac{x_1}{x_2} \le \frac{r_1}{r_2} \le e\frac{x_1}{x_2}$$

where x_i is the ^{12}C/^{13}C mixing ratio of the ith sample, r_i is intensity ratio (I_{12C}/I_{13C}) of the ith sample, and e (*ratio_tol*) sets the acceptable tolerance for the intensity ratio relative to the mixing ratio. This two-round intensity filter allows for features with varying ^{12}C and ^{13}C intensity ratios (due to the kinetic isotope effect or carbon fixation of atmospheric CO_2) to pass the relaxed first round and stricter second round as long as their intensities vary systematically between samples. All passing features are termed credentialed. Credentialed features are output as a summary table that includes all U-^{12}C peaks determined to be of biological origin.

RESULTS AND DISCUSSION

Each step of the untargeted metabolomic workflow can introduce artifactual signals that are not endogenous to the biological sample being analyzed. It is generally not possible to discriminate features of biological origin from artifactual features a priori, and thus, artifactual signals significantly complicate interpretation of untargeted metabolomic results. These artifactual signals can arise from sample contamination during metabolite extraction, carryover from previous experiments, background noise detected by the MS, or misannotation of data during bioinformatic processing. While efforts are made to minimize artifactual signals, it is not possible to completely eliminate them from the features list. We therefore attempted to filter out artifactual signals by using isotopic signatures of cellular metabolism that are easily identified by informatic analysis. We utilized the widely available and extensively characterized *E. coli* strain K12 to generate isotopically enriched biological extracts. Two cultures were prepared in parallel, one containing ^{12}C (natural-abundance) glucose and the other containing ^{13}C glucose as the sole carbon source in M9 minimal media. The cultures were mixed in defined ratios and processed through the metabolomic workflow together. By searching the

dx.doi.org/10.1021/ac503092d | *Anal. Chem.* 2014, 86, 9583−9589

resulting features list for pairs of unlabeled and fully labeled isotopologues and comparing their intensities to the values expected from the culture volume ratios, signals of biological origin can be distinguished from artifactual ones. The output of the approach is a list of credentialed features arising from the biological sample of interest. These features reflect the extent to which the methodology employed was able to capture the metabolome.

The power of stable isotope labeling in conjunction with mass spectrometry has long been leveraged to improve quantitative measurements. Mixing labeled and unlabeled samples has proven to be an effective approach to perform quantitation in proteomics,[25−27] and similar approaches have recently been extended to metabolomics.[28] Mashego et al. developed "mass isotopomer ratio analysis of U-^{13}C labeled extracts" (MIRACLE) in which U-^{13}C labeled metabolites obtained from yeast grown in defined culture medium are mixed with unlabeled sample extracts to improve quantitation.[29] More recently, an innovative variation of ^{12}C−^{13}C metabolite mixing was developed in which cells are grown in either 5% or 95% randomly enriched ^{13}C glucose. This experimental strategy, termed isotopic ratio outlier analysis or IROA, leads to a diagnostic isotopic pattern for naturally occurring compounds that can be used for quantitation and metabolite identification during untargeted profiling.[30,31] Here, we introduce another experimental approach which involves mixing ^{12}C and ^{13}C metabolic extracts. We then use the unique isotopic signals that result from the metabolic transformation of the label as a mechanism to identify features of biological origin.

Contrasting the Credentialing and IROA Platforms. It is worth distinguishing IROA from our credentialing approach. Fundamental to the distinction is that mixing a natural-abundance sample with a U-^{13}C labeled sample in a single ratio does not provide a specific enough signature to effectively discriminate features of biological origin from artifactual features. IROA introduces additional specificity to the isotopic pattern by enriching one sample with 5% ^{13}C and a second sample with 95% ^{13}C, instead of using natural-abundance and U-^{13}C samples. In contrast, credentialing introduces additional specificity to the isotopic pattern by mixing different ratios of natural-abundance and U-^{13}C samples. In credentialing, one sample is made by mixing natural-abundance and U-^{13}C cells at a ratio of 1/1 and a second sample is made by mixing natural-abundance and U-^{13}C cells at a ratio of 1/2. There are experimental benefits of each approach that make the platforms better suited for each of their unique experimental applications. IROA has been used to identify and quantitate differences between biological phenotypes during untargeted profiling. Given that the relative ratio of any given peak between biological phenotypes is unknown during untargeted profiling, the credentialing strategy based on defined ratios is incompatible with this type of discovery analysis. The objective of credentialing, on the other hand, is to identify features of biological origin exclusively from standard *E. coli* samples. While IROA could be used for this purpose in principle, the credentialing platform is not constrained by the aim of discovery analysis and therefore offers several advantages. First, media needed to produce labeled *E. coli* samples for credentialing is easily synthesized in any laboratory, whereas IROA media can only be obtained commercially. Second, the credentialing platform is better suited to identify low-intensity features of biological origin. In IROA, the signal intensity of any given metabolite is shifted away from the U-^{12}C peak and the U-^{13}C peak as a function of carbon number. For a metabolite with 10 carbons, as an example, 50% of the signal intensity is lost from the U-^{12}C peak or the U-^{13}C peak. This decrease in signal intensity prevents low-abundance *E. coli* derived metabolites that are detected in unlabeled samples from being detected with IROA. Because the credentialing platform only uses natural-abundance and U-^{13}C samples, it is not subject to this limitation. Indeed, detection of low-abundance metabolites is of particular importance when optimizing metabolomic methods as these compounds are the most challenging to measure, but can be of great biological importance. Finally, because credentialing only uses *E. coli* samples, the analysis of the resulting isotopic data can exploit the known relationship between mass and carbon number derived from ECMDB (Supporting Information Figure S-2).

Parameters for Credentialing. To accomplish the filtering of artifactual signals, we created a simple R package. The core function, credential(), has several adjustable parameters allowing various chromatographic and instrumental platforms to be credentialed. These parameters include (i) iso_ppm, the ppm tolerance when searching for ^{13}C isotopes, (ii) iso_rt, the retention-time window in which a peak and its isotope must elute, (iii) mix_tol, the tolerance for the intensity ratio of the ^{12}C and ^{13}C peak, (iv) ratio_tol, the tolerance for the ratio of the intensity ratios between two samples, and (v) mpc_tol, the tolerance for compounds with unusually high or low mass compared to the number of carbons they contain. (Details concerning the calculation of mass per carbon based on the ECMDB can be found in Supporting Information Figure S-2.)

We have determined effective parameters for reversed-phase and hydrophilic interaction liquid chromatography as well as for the Agilent QTOF, Thermo QE, AB SCIEX TripleTOF 5600^{+}, and the LECO Pegasus GC-HRT. These parameters have been experimentally validated and are listed in the Supporting Information Table S-3.

Evaluation of the filtering effectiveness was accomplished by comparing the number of credentialed features found in unlabeled and labeled extracts. In addition to the labeled extracts, natural-abundance (unlabeled) extracts were generated as controls. An unlabeled extract should have no credentialed features if it is not mixed with a labeled extract. Therefore, the number of passing features in an unlabeled extract represents the false positive rate. Initial experiments indicated that filtering based on a single mixed-extract sample was not sufficiently selective to remove the majority of artifactual peaks. We found that a two-sample, relative-intensity filter was most effective. As shown in Table 1, this filtering process is selective. The process credentialed only 0.6% of the negative-control features, whereas 9% of the ^{12}C/^{13}C mixture features were credentialed.

To further validate the filtering process, we examined the natural isotopic peaks that were credentialed in our ^{12}C/^{13}C sample. Consider that in a ^{12}C sample many peaks will contain a natural-abundance M + 1 peak which by definition satisfies the mass requirement to be an isotope. The filtering process credentials some of these natural isotopes along with the monoisotopic peak. These are easily detected and removed by established deisotoping methods, but these peaks allowed us to assess how often an M + 1 is credentialed when the M + 0 is not. If this occurs often, it would indicate that the algorithm is inappropriately disqualifying features. We detected 385 credentialed natural isotopes in our mixture sample. Out of

dx.doi.org/10.1021/ac503092d | Anal. Chem. 2014, 86, 9583−9589

Table 1. Performance of Feature Credentialing[a]

sample type	total features	credentialed features	percentage credentialed (%)
no injection	1564	13	0.8
extraction blank	2736	18	0.7
natural-abundance *E. coli*	18643	120	0.6
$^{12}C/^{13}C$ standard sample	23567	2192	9.3

[a]A summary of the results of the credentialing process after being applied to several different data sets. The rows labeled "no injection" and "extraction blanks" represent credentialed peaks due to carryover from previous credentialing runs. Natural-abundance *E. coli* is a negative control that estimates the false positive rate of the credentialing process.

the 385 credentialed, natural isotopes only one did not have a corresponding U-^{12}C in the final credentialed features list. This indicates the filtering approach is performing reliably.

Application: Reoptimization of a Previously Published XCMS Method. With an established method to credential features as biological in origin and exclude various noise sources, we set out to optimize our XCMS-based informatic workflow. XCMS is a widely used informatic package suited for the analysis of untargeted LC/MS data sets. The general XCMS workflow involves peak finding, peak grouping across samples, and retention-time alignment. Settings for each step in this process affect the quality of features returned and therefore the overall performance of the untargeted metabolomic workflow. For example, we found that settings for peak picking that cause the annotation of spurious noise peaks as features lower the quality of peak grouping and retention-time alignment (data not shown). Further, using poor grouping parameters can lead to XCMS splitting a single peak into multiple groups, thereby resulting in erroneous statistics.

To generate data for XCMS optimization, a previously published method was replicated.[6] The same LC/MS system, extraction method, and chromatography protocols were utilized as published and described in the Experimental Section. When processing the data, however, we varied several parameters of the XCMS functions findPeaks.centWave(), group(), and retcor(). As the filtering depends on each of these functions, the final number of credentialed features is representative of the quality of XCMS data processing. Previous approaches to optimizing untargeted metabolomic parameters such as these have relied on counting the total number of features detected. Here, we applied our filtering approach to instead count the number of credentialed features and use this as a benchmark for parameter optimization. Our results show that the published method parameters based on total number of features are suboptimal (Table 2). The published parameters do return a greater number of total features, but the number of features of biological origin accurately detected and grouped is substantially lower with these settings. These data highlight that a larger feature number does not necessarily indicate better metabolome coverage and therefore an improved untargeted metabolomic method.

Reoptimization of XCMS parameters resulted in a substantial improvement. Our XCMS parameters led to an increase of 20% in credentialed features (an increase of 342 features), while reducing the total number of features by 15% (a decrease of 4750 features). Parameters for findPeaks.centWave() were determined to be the most critical to the analysis, while further optimization of group and retcor qualified only an additional 41

Table 2. Reoptimization of Published XCMS Parameters[a]

XCMS parameter	published parameters	with optimized peak finding	with optimized retcor and group
ppm	15	12	12
peak width	10, 120	15, 140	15, 140
mzwid	0.015	0.015	0.015
bw	5	5	10
gapInit			0.6
total features	32010	27260	27260
credentialed features	1475	1776	1817

[a]Parameters used and the results of each step in the optimization process are shown. Published parameters were taken from a previously published method (ref 6). The column labeled "with optimized peak finding" shows results for the optimization of findPeaks.centWave().

peaks. It is notable that, prior to optimizing findPeaks.centWave(), optimization of group() parameters increased the number of credentialed features, partially overcoming the negative impact of artifactual signals.

Characterizing Features in Untargeted Metabolomic Data Sets. To translate metabolomic data into biochemical insight, the features generated in a typical untargeted experiment must first be structurally characterized. The standard workflow for structurally characterizing features requires matching MS/MS data of the features of interest to the MS/MS data of authentic standards. Identifying features is the most time-demanding step of the untargeted metabolomic workflow and is generally performed in a targeted manner. That is, MS/MS data are only acquired and interpreted for a handful of features determined to be interesting, usually on the basis of statistical thresholds. While this worfklow is often applied to identify tens of metabolites in a metabolomic study, attempting to identify each of the thousands of features detected in a typical sample with this approach is impractical. New technologies to reduce the time required to establish metabolite identifications are an active area of research, but high-throughput methods to structurally characterize metabolites are not widely available. Moreover, many of the MS/MS data are challenging to interpret. When the MS/MS pattern of a feature does not match any of the MS/MS patterns in metabolite databases, it is difficult to determine if the MS/MS data correspond to an unknown metabolite or merely MS/MS data from an artifactual feature.

The feature credentialing approach offers a mechanism to rapidly filter features that should not be pursued for identification, namely, those features that do not correspond to signals of biological origin. When we applied credentialing to *E. coli* extract, we reduced the number of features that represent candidates for MS/MS from 23 567 to 2192. The resulting subset of credentialed features can be targeted for MS/MS analysis with standard workflows. As an example, we performed targeted MS/MS on 250 compounds in a single experimental run. These data illustrate that MS/MS experiments could be performed on every feature of biological origin over a minimal and feasible number of analytical runs. Select data are presented in Figure 2A–C. The MS/MS data collected on these features were matched to the METLIN metabolite database and resulted in the identification of three metabolites: uracil, ADP (adenosine diphosphate), and UDP-GlcA (uridine diphosphate glucuronic acid). MS^1 spectra and chromatograms for these compounds can be found in Supporting Information Figure S-4.

dx.doi.org/10.1021/ac503092d | *Anal. Chem.* 2014, 86, 9583–9589

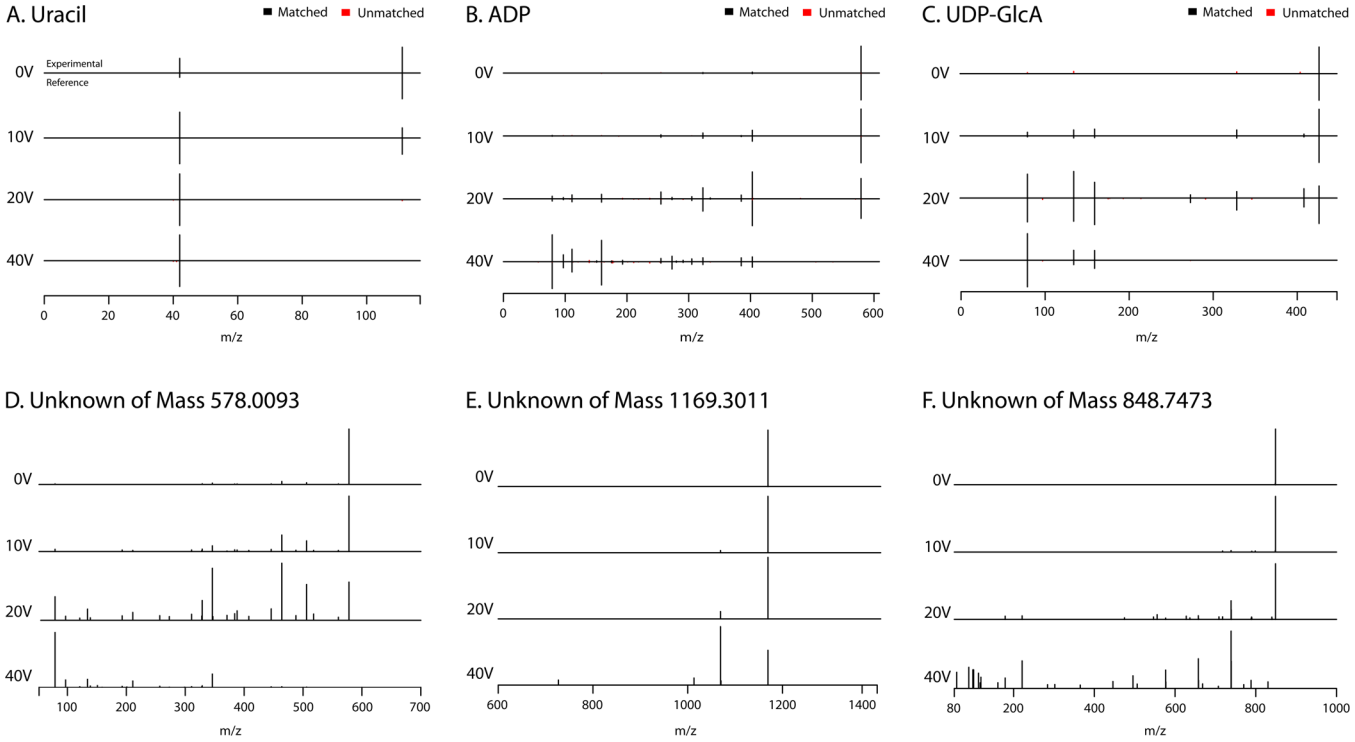


Figure 2. MS/MS spectra from six representative credentialed features. MS/MS spectra were collected at four collision energies (0, 10, 20, and 40 V) on six credentialed ions. Three of these ions (A) uracil, (B) ADP, and (C) UDP-GlcA were identified based on accurate mass, carbon number, and METLIN database hits. These identifications were confirmed by comparing the experimental MS/MS spectra to the METLIN MS/MS reference spectra as shown. The upper spectrum of each plot is the experimental data, and the lower spectrum is the METLIN reference data. Unmatched peaks are depicted in red. The second three ions (D) 578.0093, (E) 1169.3011, and (F) 848.7473 were classified as unknowns as they did not match any METLIN database entries as either a fragment or parent mass. The MS/MS spectrum of each ion is displayed as normalized intensity at the same four collision energies.

In addition to generating MS/MS data for metabolites included in databases, it is possible to reliably generate MS/MS data on biological peaks which currently cannot be annotated by metabolomic databases. Because credentialed features have passed our filtering rounds, we know that they are true metabolites of biological origin even if they do not return any database hits. Of the 1827 credentialed features, 392 were not found in METLIN or the METLIN fragment databases. Three such example features are seen in Figure 2D–F. Previously these features may have been discarded as artifacts, but the credentialing platform provides confidence in their authenticity such that they can be reported and referenced in future experiments.

CONCLUSION

The feature credentialing strategy presented here is a powerful platform to discriminate biological features from the various noise sources prevalent in untargeted metabolomic data. The process is experimentally straightforward and can be easily implemented in any metabolomic laboratory. Feature credentialing reliably removes artifactual features such as those arising from chemical and informatic noise, thereby resulting in a valuable list of features of biological origin. These credentialed features address many of the drawbacks associated with feature counting in comparing method performance on the basis of metabolome coverage. As such, counting credentialed features can be used in the development and optimization of untargeted metabolomic approaches as demonstrated by the reoptimization of XCMS parameters. Credentialing features is also an effective data reduction strategy for untargeted metabolomic results such that a smaller number of peaks can be targeted for MS/MS analysis. In summary, the feature credentialing platform introduced here represents a step toward defining optimal untargeted metabolomic platforms and provides a standard metric to facilitate collaboration between different metabolomic laboratories.

ASSOCIATED CONTENT

Supporting Information

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

AUTHOR INFORMATION

Corresponding Author

*E-mail: gjpattij@wustl.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health Grants R01 ES022181 (G.J.P.), L30 AG0 038036 (G.J.P.), and the Alfred P. Sloan Foundation.

REFERENCES

(1) Patti, G. J.; Yanes, O.; Siuzdak, G. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 263–269.

(2) Patti, G. J.; Yanes, O.; Shriver, L. P.; Courade, J.-P.; Tautenhahn, R.; Manchester, M.; Siuzdak, G. *Nat. Chem. Biol.* **2012**, *8*, 232–234.

(3) Khan, A. P.; Rajendiran, T. M.; Ateeq, B.; Asangani, I. A.; Athanikar, J. N.; Yocum, A. K.; Mehra, R.; Siddiqui, J.; Palapattu, G.; Wei, J. T.; Michailidis, G.; Sreekumar, A.; Chinnaiyan, A. M. *Neoplasia (N. Y., NY, U. S.)* **2013**, *15*, 491−501.
(4) Tang, W. H. W.; Wang, Z.; Levison, B. S.; Koeth, R. A.; Britt, E. B.; Fu, X.; Wu, Y.; Hazen, S. L. *N. Engl. J. Med.* **2013**, *368*, 1575−1584.
(5) Jain, M.; Nilsson, R.; Sharma, S.; Madhusudhan, N.; Kitami, T.; Souza, A. L.; Kafri, R.; Kirschner, M. W.; Clish, C. B.; Mootha, V. K. *Science* **2012**, *336*, 1040−1044.
(6) 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.
(7) Bajad, S. U.; Lu, W.; Kimball, E. H.; Yuan, J.; Peterson, C.; Rabinowitz, J. D. *J. Chromatogr., A* **2006**, *1125*, 76−88.
(8) Lu, W.; Bennett, B. D.; Rabinowitz, J. D. *J. Chromatogr., B* **2008**, *871*, 236−242.
(9) Masson, P.; Alves, A. C.; Ebbels, T. M. D.; Nicholson, J. K.; Want, E. J. *Anal. Chem.* **2010**, *82*, 7779−7786.
(10) Zelena, E.; Dunn, W. B.; Broadhurst, D.; Francis-McIntyre, S.; Carroll, K. M.; Begley, P.; O'Hagan, S.; Knowles, J. D.; Halsall, A.; HUSERMET Consortium; Wilson, I. D.; Kell, D. B. *Anal. Chem.* **2009**, *81*, 1357−1364.
(11) Geier, F. M.; Want, E. J.; Leroi, A. M.; Bundy, J. G. *Anal. Chem.* **2011**, *83*, 3730−3736.
(12) Yanes, O.; Tautenhahn, R.; Patti, G. J.; Siuzdak, G. *Anal. Chem.* **2011**, *83*, 2152−2161.
(13) Nordström, A.; Want, E.; Northen, T.; Lehtiö, J.; Siuzdak, G. *Anal. Chem.* **2008**, *80*, 421−429.
(14) Buescher, J. M.; Moco, S.; Sauer, U.; Zamboni, N. *Anal. Chem.* **2010**, *82*, 4403−4412.
(15) Lu, W.; Clasquin, M. F.; Melamud, E.; Amador-Noguez, D.; Caudy, A. A.; Rabinowitz, J. D. *Anal. Chem.* **2010**, *82*, 3212−3221.
(16) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−787.
(17) Chokkathukalam, A.; Jankevics, A.; Creek, D. J.; Achcar, F.; Barrett, M. P.; Breitling, R. *Bioinformatics* **2013**, *29*, 281−283.
(18) Mishur, R. J.; Rea, S. L. *Mass Spectrom. Rev.* **2012**, *31*, 70−95.
(19) Brown, M.; Dunn, W. B.; Dobson, P.; Patel, Y.; Winder, C. L.; Francis-McIntyre, S.; Begley, P.; Carroll, K.; Broadhurst, D.; Tseng, A.; Swainston, N.; Spasic, I.; Goodacre, R.; Kell, D. B. *Analyst* **2009**, *134*, 1322−1332.
(20) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. *Anal. Chem.* **2012**, *84*, 283−289.
(21) Alonso, A.; Julià, A.; Beltran, A.; Vinaixa, M.; Díaz, M.; Ibañez, L.; Correig, X.; Marsal, S. *Bioinformatics* **2011**, *27*, 1339−1340.
(22) R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2014; http://www.R-project.org.
(23) Smith, C. A.; 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.
(24) Guo, A. C.; Jewison, T.; Wilson, M.; Liu, Y.; Knox, C.; Djoumbou, Y.; Lo, P.; Mandal, R.; Krishnamurthy, R.; Wishart, D. S. *Nucleic Acids Res.* **2013**, *41*, D625−D630.
(25) Ong, S.-E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. *Mol. Cell. Proteomics* **2002**, *1*, 376−386.
(26) Wiese, S.; Reidegeld, K. A.; Meyer, H. E.; Warscheid, B. *Proteomics* **2007**, *7*, 340−350.
(27) Hebert, A. S.; Merrill, A. E.; Bailey, D. J.; Still, A. J.; Westphall, M. S.; Strieter, E. R.; Pagliarini, D. J.; Coon, J. J. *Nat. Methods* **2013**, *10*, 332−334.
(28) Birkemeyer, C.; Luedemann, A.; Wagner, C.; Erban, A.; Kopka, J. *Trends Biotechnol.* **2005**, *23*, 28−33.
(29) Mashego, M. R.; Wu, L.; Van Dam, J. C.; Ras, C.; Vinke, J. L.; Van Winden, W. A.; Van Gulik, W. M.; Heijnen, J. J. *Biotechnol. Bioeng.* **2004**, *85*, 620−628.
(30) De Jong, F. A.; Beecher, C. *Bioanalysis* **2012**, *4*, 2303−2314.
(31) Stupp, G. S.; Clendinen, C. S.; Ajredini, R.; Szewc, M. A.; Garrett, T.; Menger, R. F.; Yost, R. A.; Beecher, C.; Edison, A. S. *Anal. Chem.* **2013**, *85*, 11858−11865.