Published in final edited form as:

Anal Chem. 2019 February 05; 91(3): 1838–1846. doi:10.1021/acs.analchem.8b03132.

Peak annotation and verification engine (PAVE) for untargeted LC-MS metabolomics

Lin Wang^{†,‡}, Xi Xing^{†,‡}, Li Chen^{†,‡}, Lifeng Yang^{†,‡}, Xiaoyang Su^{†,§}, Herschel Rabitz[‡],
Wenyun Lu^{†,‡}, and Joshua D. Rabinowitz^{†,‡}

[†]Lewis Sigler Institute for Integrative Genomics, Princeton University, New Jersey 08544, USA

[‡]Department of Chemistry, Princeton University, New Jersey 08544, USA

[§]Department of Medicine, Robert Wood Johnson Medical School, Rutgers University, New Brunswick, NJ 08904, USA

Abstract

Untargeted metabolomics can detect more than 10,000 peaks in a single LC-MS run. The correspondence between these peaks and metabolites, however, remains unclear. Here we introduce a Peak Annotation and Verification Engine (PAVE) for annotating untargeted microbial metabolomics data. The workflow involves growing cells in ¹³C and ¹⁵N isotope-labeled media to identify peaks from biological compounds and their carbon and nitrogen atom counts. Improved de-isotoping and de-adducting is enabled by algorithms that integrate positive mode, negative mode and labeling data. To distinguish metabolites and their fragments, PAVE experimentally measures the response of each peak to weak in-source collision induced dissociation, which increases the peak intensity for fragments while decreasing it for their parent ions. The molecular formulae of the putative metabolites are then assigned based on database searching using both *m/z* and C/N atom counts. Application of this procedure to *S. cerevisiae* and *E. coli* revealed that more than 80% peaks do not label, i.e. are environmental contaminants. More than 70% of the biological peaks are isotopic variants, adducts, fragments, or mass spectrometry artifacts yielding ~2,000 apparent metabolites across the two organisms. About 650 match to a known metabolite formula based on *m/z* and C/N atom counts, with 220 assigned structures based on MS/MS and/or retention time to match to authenticated standards. Thus, PAVE enables systematic annotation of LC-MS metabolomics data, with only ~4% of peaks annotated as apparent metabolites.

Graphical Abstract

Correspondence: Wenyun Lu: wlu@princeton.edu; Joshua D. Rabinowitz: josh@princeton.edu.

Supporting Information

Supporting Information Available:

Supporting experimental methods for cell culture and metabolite extraction, LC-MS, and data analysis. Supporting tables of liquid media composition, PAVE peak annotation as a function of peak height, annotation of peaks with too large mass relative to C-count, isotope and adduct masses used in JUNKREMOVER, comparison of PAVE with other approaches for annotating full scan LC-MS metabolomics data, 432 metabolite standards and their retention times, full data for evaluation of the in-source CID method for distinguishing metabolites and fragments, and annotation of all peaks. Supporting figures of total ion chromatograms, metabolite molecular weight as a function of C-count, ringing artifact, formula assignment workflow, implementation of the in-source CID method on orbitrap and TOF instruments, and MS/MS spectrum of the unknown glutathione condensation product.

Metabolomics, the large-scale analysis of small molecule metabolites, is playing an increasing role in biological research, as well as more applied areas such as medicine and bioenergy¹. Multiple instrument platforms are available for metabolomics. Among these, liquid chromatography-electrospray ionization-high resolution mass spectrometry (LC-MS) has emerged as a popular choice, with time-of-flight or orbitrap instruments commonly used for the mass spectrometry step. LC-MS metabolomics data contains thousands of peaks, defined by their mass-to-charge ratio (m/z) and chromatographic retention time (RT). Open source software, such as XCMS²⁻⁴, MZmine^{5,6}, MetAlign⁷ and MS-DIAL⁸, are available for peak alignment and subsequently discovering peaks that change statistically significant across samples.

While useful for focusing on biologically interesting peaks, such an approach does not address the full scope of metabolites present. One major challenge is that many peaks can arise from the same metabolite, due to isotopic variants, adducts, in-source fragmentation, etc. A key need is to properly annotate these peaks, which can then either be used collectively for metabolite quantitation or structure elucidation or discarded to focus on only the metabolite $[M+H]^+$ or $[M-H]^-$ ions. Several bioinformatics tools have been designed to group together such peaks, based on shared chromatographic retention time (RT), peak shape, and intensity correlation across samples⁹⁻¹¹. In one example, an algorithm known as CAMERA reduced the number features by approximately 60%¹². In another case, MS-FLO removed or flagged roughly 15% of features that passed through the filters of MS-DIAL¹³. Thus, these software achieve substantial data simplification, although the completeness of their annotation remains unclear.

Another approach that can help with peak annotation is stable isotope labeling. Labeling distinguishes peaks of biological origin (biological peaks) from environmental contaminants. Moreover, it can be used to determine atom counts based on the shift of mass between different labeled samples^{14,15}. NTFD is a widely used software for isotopically labeled GC-MS data¹⁶. Similarly, software including MAVEN^{17,18}, mzMatch-ISO^{19,20}, X¹³CMS²¹, and ALLocator²² were developed to facilitate analysis of isotope-labeled LC-MS data. To circumvent possible errors in matching unlabeled and ¹³C-labeled peaks, the IROA approach was developed²³⁻²⁵. Unlabeled and ¹³C-labeled samples are mixed together for uniform extraction, sample preparation and LC-MS analysis. Peaks showing mirrored isotopes are identified, along with their carbon atom count, which facilitates formula identification. MetExtractII^{26,27} builds on these concepts with additional tools for tracking the fate of specific ¹³C-labeled inputs and for identifying labeled product ions in MS/MS data.

Recently, Mahieu and Patti²⁸ combined ¹³C-labeling with computational de-isotoping, de-adducting, and removal of common neutral loss fragmentation peaks. Using data from *E. coli*, they provided evidence that less than 10% of LC-MS peaks are metabolite ions,

suggesting that typical metabolomics experiments measure less than 1000 metabolites, rather than the > 10,000 suggested by the raw number of peaks. Here we build on this progress by developing a Peak Annotation and Verification Engine (PAVE) that (i) combines ^{15}N and ^{13}C labeling, (ii) integrates both labeling patterns and positive and negative mode data for more effective de-adducting, (iii) annotates certain MS artifacts, such as “ringing peaks” in orbitrap data,²⁹ which had been overlooked in previous software, and (iv) introduces a new experimental approach for fragment identification, which was previously a particularly difficult aspect of untargeted metabolomics data annotation (Figure 1). Application of PAVE to *E. coli* and *S. cerevisiae* supports the general claim of Mahieu and Patti that less than 10% of peaks in LC-MS metabolomics data correspond to metabolite ions. This conclusion is strengthened by PAVE, unlike the approach of Mahieu and Patti, effectively correctly identifying most well-known core metabolites.

Experimental Section

Materials.

HPLC-grade water, acetonitrile, and methanol (Optima) were from Thermo Fisher. Metabolite standards and all media components were from Sigma-Aldrich. U- ^{13}C -Glucose (99%) and ^{15}N -(NH_4) $_2\text{SO}_4$ and ^{15}N - NH_4Cl were from Cambridge Isotope Laboratories.

Cell Culture and Metabolite Extraction.

S. cerevisiae strain FY4 was grown at 30 °C in YNB without amino acids (Difco 291940) and 2% glucose. Labeling was carried out for > 10 generations. Cells were harvested in exponential phase (OD_{600} of 0.45) by filtering onto a 50 mm nylon membrane filter (Millipore), which was immediately transferred into -20°C extraction solvent (1.3 mL 40:40:20, acetonitrile:methanol:water with 0.5% formic acid) in a Petri dish. The dish was kept at -20°C for 3 min. Then 108 μL 15% NH_4HCO_3 was added to neutralize the samples³⁰. The final solution was kept at -20°C for 15 min and the resulting mixture was transferred into an Eppendorf tube and spun down at 16,000 g for 15 min at 4°C . The supernatant was taken for LC-MS analysis. *E. coli* strain NCM3722 was handled as *S. cerevisiae*, except growth temperature was 37°C , media was Gutnick minimal media containing 0.4% glucose and 10 mM NH_4Cl ³¹, and OD was measured at 650 nm. A procedure blank was generated identically with water replacing the cell culture. Liver extracts were generated from 20 mg flash frozen mouse liver, which was weighed and ground with a cryomill (Retsch) at 25 Hz for 30 s and extracted as above.

LC-MS Analysis.

LC used a Vanquish UHPLC system (Thermo Fisher) and Xbridge BEH Amide HILIC column (Waters) with 25 min gradient from acetonitrile to pH 9.5 aqueous buffer. Injected sample volume was 5 μL . LC was coupled by electrospray ionization (± 3.3 kV) to a Q-Exactive Plus mass spectrometer (Thermo Fisher). Raw LC/MS data were converted to mzXML format using the command line “msconvert” utility.³² Peak extraction of the raw data for cells growing in unlabeled media was performed using the EIMaven software package (<https://elucidatainc.github.io/EIMaven>). For details, see Supplementary Methods.

Fragment Identification.

The significance of peak intensity changes upon application of in-source CID was assessed by t-test. Peaks were annotated as fragments if they showed a significant signal increase ($p < 0.05$) at either 2 eV or 4 eV.

Computer Code.

The code for PAVE is written with Matlab R2017b package. It is available at <http://github.com/xxing9703/PAVE>.

Results and Discussion

PAVE strategy.

Our approach to annotate untargeted LC-MS data begins with growing cells in four different isotopic conditions: unlabeled, ^{13}C , ^{15}N , and $^{15}\text{C}+^{13}\text{N}$ (Figure 1). Mass shifts across these four conditions are used by an algorithm called ATOMCOUNT to identify biological peaks and their carbon and nitrogen atom counts. The biological peaks are then subjected to thorough de-adducting and de-isotoping by an algorithm called JUNKREMOVER. This algorithm identifies adducts and isotopic variants by relating them to their co-eluting metabolite ions ($[\text{M}-\text{H}]^-$ or $[\text{M}+\text{H}]^+$), which may appear in the same ionization mode or (less commonly) only in the opposite ionization mode. Adducts have the same labeling patterns as their metabolite peak, facilitating accurate annotation. JUNKREMOVER also filters out peaks whose C counts are too low for the observed mass and ringing peaks which occur on both sides of intense ion peaks in orbitrap data. The remaining peaks are divided into fragment versus metabolite ions, based on the intensity of fragments increasing under a mild in-source CID voltage, whereas the intensity of the corresponding metabolite peaks decline. The resulting substantially verified metabolite peaks are assigned formulae based on their m/z and C/N atom counts, and associated with known metabolites based on RT and/or MS/MS match.

Data generation.

To implement this approach, *S. cerevisiae* strain FY4 and *E. coli* strain NCM3722 were each grown in minimal essential glucose-ammonia media in four different isotopic conditions: unlabeled, ^{15}N , ^{13}C , and $^{15}\text{C}+^{13}\text{N}$ (Table 1) and harvested in exponential phase. Triplicate biological samples for each condition were analyzed by untargeted HILIC LC-MS using an amide column and high-resolution full-scan MS1 analysis on a Q-Exactive instrument, with ~15,000 peaks detected in negative mode and ~30,000 peaks in positive mode.

Identifying biological peaks and their C/N counts with ATOMCOUNT.

The core concept of ATOMCOUNT is that, if peak i of m/z M_i in the unlabeled sample is of biological origin, it should show a characteristic mass shift across labeled conditions. Moreover, the corresponding peaks ($M_i + N_i$, $M_i + C_i$, $M_i + N_i + C_i$) in the labeled samples should match the intensity of peak i in the unlabeled sample. For each peak i in the unlabeled sample, ATOMCOUNT records, from the raw data of the labeled samples at the unlabeled peak's RT (± 20 s; chromatographic shifts of < 20 s accordingly do not impact the

procedure), the highest signal intensity for masses (± 10 ppm) corresponding to every integer combination of N_i and C_i with $N_i = \{0, 1, 2, \dots, N_{max}\}$ and $C_i = \{1, 2, \dots, C_{max}\}$ with N_{max} arbitrarily chosen to be 10 and C_{max} the highest possible number of carbon atoms given the mass. For each combination of C_i and N_i , it computes the Pearson correlation (ρ) between the observed and expected intensity pattern (Figure 2). In cases where no values of N_i and C_i yield $\rho > 0.75$, the peak is discarded. In cases where multiple values of N_i and C_i yield $\rho > 0.75$, those yielding the highest value for ρ are accepted. An algorithmic advantage of ATOMCOUNT is retrieving the peaks in labeled samples, at the relevant masses and RT, directly from the raw data. This contrasts with prior approaches that compare list of picked peaks across the unlabeled and labeled samples. Directly querying the raw data avoids false negatives due to imperfect peak picking³³.

ATOMCOUNT identified less than 20% of peaks as having a logical labeling pattern (i.e. being biological peaks) (Table 2). Thus, a primary driver of the very large number “unknown” metabolites in many prior estimates is miscounting of environmental contaminants (i.e. chemicals not made by the microorganism, but instead found in water, tubes, etc.). Consistent with most of the discarded peaks being environmental contaminants, we observed $\sim 80\%$ of the discarded peaks in procedure blanks at intensities $> 50\%$ of that observed in the cellular samples. In contrast, only 10% of peaks retained by ATOMCOUNT were found at this level in the blank. Simple blank substrate, however, cannot substitute for the ATOMCOUNT procedure: even excluding peaks found in the blank, there are a roughly equal number of peaks that labeled and do not label. Moreover, ATOMCOUNT effectively annotates biological peaks with a C and N atom count that facilitates subsequent data analysis.

Removal of adducts, natural isotopic variants and ringing peaks with JUNKREMOVER.

In addition to the metabolite peaks ($[M-H]^-$ or $[M+H]^+$), the list of biological peaks generated by ATOMCOUNT includes adducts, natural isotopic variants, and fragments. JUNKREMOVER systematically evaluates every peak returned by ATOMCOUNT and discards adducts and isotopic variants. As is typical¹³, this process begins with a pre-determined list of relevant adducts and isotopic variants and their associated mass deviations (D) from the metabolite ion. The list of adducts was initially obtained by literature search³⁴ and was augmented by examination of our data, looking for common mass shifts among co-eluting peaks that could reasonably be explained by some form of adduct formation. With these lists in hand (Table S5 and S6), for each biological peak i in the unlabeled sample, JUNKREMOVER records from the raw data of the unlabeled samples (in both ionization modes), at the relevant RT, the peak intensity at $M_i - D$ (one-by-one, for every candidate adduct and isotopic variant in this list, correcting by the mass of 2H when checking in the opposite mode).

For de-adducting, the peak i is discarded if (i) the peak intensity at $M_i - D$ is at least 20% of that at M_i , (ii) the peak at $M_i - D$ passes the ATOMCOUNT criteria for a biological metabolite, and (iii) the C/N counts of the two peaks are equivalent. The value of 20% was based on our finding that, for known metabolites, either the $[M+H]^+$ or $[M+H]^-$ peak is reliably at least 20% of the size of the largest adduct peak. Note that C and N atoms

introduced by adduct formation, e.g. from formate or ammonia in the running buffer, are not of biological origin and therefore do not affect the C/N counts returned by ATOMCOUNT. As an example of the utility of labeling data for de-adducting, glucose gives rises to peaks at m/z 119.0349 $[M-H-C_2H_4O_2]^-$, 179.0560 $[M-H]^-$, and m/z 239.0772 $[M-H+C_2H_4O_2]^-$ in negative mode (Figure 3 A). The labeling data make clear that m/z 239.0772 is an adduct (as it matches the carbon count of m/z 179.0560) while m/z 119.0349 is a fragment (as it has two fewer carbons). As an example of the value of combined positive and negative mode analysis, the $[M+Na]^+$ peak of glucose (Figure 3 B) was detected in positive mode and correctly recognized as an adduct by comparing to the $[M-H]^-$ peak in negative mode, despite the glucose $[M+H]^+$ peak not being detectable. Overall, we observed more than 150 adducts whose proper assignment required comparison to metabolite peaks in the opposite polarity mode (Table 2).

For candidate isotopic variants, the approach is identical to that for adducts, except that there is no requirement for ^{13}C and ^{15}N natural isotopic variants to match C/N counts, and for ^{18}O and ^{34}S , the peak intensity at mass $i-D$ is required to be substantially greater than at mass i (20 times larger for ^{18}O and 5 for ^{34}S).

To focus only on $[M+H]^+$ and $[M-H]^-$ peaks, we also annotated doubly charged ions, which are relatively common in positive mode (Table 2). Similarly, we annotated homodimers (i) by searching for a peak at mass $(i-1)/2$ (in negative mode) or $(i+1)/2$ (in positive mode), where the C/N counts differ by exactly a factor of two. Collectively, these above steps in JUNKREMOVER annotate roughly 60% of the peaks found by ATOMCOUNT.

Manual examination revealed that many of the remaining peaks had a surprisingly low C/N atom counts relative to the mass. We believe that these peaks with inappropriately low C/N count arise, at least in part, due to ions that are some combination of contaminants (such as solvent or plastic impurities) with an endogenous metabolite. This blending may occur during sample preparation (i.e. due to formation of covalent condensation products) or in the ESI source (i.e. due to adduct formation). To filter out such peaks, we eliminated those whose mass exceeded (i) the 99th percentile mass of unique endogenous metabolite molecular formulae in the HMDB database³⁵ for that C count ± 2 , and (ii) the cut-off for any lesser value of C. Especially in positive ion mode in *S. cerevisiae*, we observed many peaks exceeding this threshold (Figure 3 C, D). The number of peaks removed in this analysis depends on the percentile cutoff, which is a user parameter in JUNKREMOVER. The 99th percentile cutoff likely retains some peaks that are artifacts, while losing ~ 1% of real metabolite peaks. Using a 95th percentile cutoff would remove roughly an additional 20% of peaks (Table S4).

In addition to the low C/N atom count peaks, we also observed ringing peaks which are a known artifact of Fourier transform-MS instruments including orbitraps. The ringing peaks occur in mass spectra symmetrically around high intensity peaks, shifted in mass by roughly ± 500 ppm (Figure S6A). The mass shift and relative intensity of ringing peaks is independent of MS resolution and ionization polarity (Figure S6B and C). To annotate ringing peaks, for each peak, we check if there is a larger peak at the same RT whose intensity is at least 100x higher and occurs within ± 500 ppm; if so, the smaller peak is

annotated as a ringing peak. Based on this analysis, about 2,000 of the initial peaks are ringing artifacts. While most of these were too noisy to be selected by ATOMCOUNT, 93 advanced into and were removed by JUNKREMOVER.

Annotation of in-source fragments.

After peak scrubbing by JUNKREMOVER, we obtained a peak list comprised substantially of metabolites and in-source fragments. The chemistry of fragmentation is more complicated than adduct formation. While certain fragmentation events are common (e.g. loss of water, ammonia), others are not easily predicted. Moreover, whereas adduct formation retains the C and N atom counts from labeling data, fragmentation does not except in certain situations such as loss of water. This makes comprehensive assignment of fragmentation events based solely on computational analysis of untargeted LC-MS data difficult.^{36,37} Accordingly, we sought a straightforward experimental approach to differentiate fragments from their parent ions. We reasoned that the application of a weak in-source CID voltage should increase signal intensity for fragments while decreasing it for of parent ions. Taking glucose as an example, two fragment ions (m/z 119.0349, m/z 101.0245) were observed at the retention time of glucose, corresponding to loss of $C_2H_4O_2$ and $C_2H_6O_3$. The intensity of both fragments increased with 2 eV and 4 eV in-source CID energy, while the intensity of the glucose parent ion at m/z 179.0560 and its adducts decreased. Higher energies resulted in further fragmentation from m/z 119.0349 to m/z 101.0245, which decreased signal intensity at m/z 119.0349, highlighting the importance of using low CID voltages.

To test the general effectiveness of this approach, we examined how the intensities of metabolite ions and their fragments change in response to increasing in-source CID energies using 80 metabolite standards dissolved in extraction buffer (for individual compound data, see Table S9). Strikingly, each fragment ion increased at least somewhat at both 2 and 4 eV in-source CID energy, while every parent ion decreased (Figure 4). For 75 of the 80 known fragments, the increase was statistically significant at 4 eV ($p < 0.05$, t-test). This pattern held both for small robust compounds like malate and relatively labile-energy compounds like ATP.

In-source CID is a built-in capacity in Thermo's Q-Exactive Plus Orbitrap mass spectrometer. For other mass spectrometers, approaches such as all ion fragmentation can be used, and we verified that this approach works on Agilent's 6550 Q-TOF, with the basic trend of metabolite signals decreasing and fragment signals increasing robust across a range of collision energies (Figure S8).

The success of the in-source fragmentation approach depends on chromatographic resolution. For example, malate fragments to make an ion that mimics fumarate. Because malate and fumarate were chromatographically resolved³⁸, we correctly annotated both the fragmentation event (at the retention time of malate, 12.9 min) and the genuine fumarate metabolite peak (at 11.6 min). At the same time, because we did not effectively separate linoleic and arachidonic acid from larger lipids, we discarded their peaks, because their abundances rose upon application of the CID voltage due to fragmentation of the co-eluting lipids. In ongoing work, we have also noted cases where the peak at $[M+H]^+$ or $[M-H]^-$ rises in response to weak CID voltage due to an adduct peak breaking down to yield the

(de)protonation peak. This seems only to be a problem, however, when the adduct peak is much more abundant than the (de)protonation peak, in which case the metabolite is better measured in the opposite polarity mode. Thus, despite some complexities, the weak in-source CID procedure is effective for distinguishing metabolite from fragment peaks. We accordingly applied this procedure to all peaks that were not rejected by the JUNKREMOVER algorithm (Table 2). This procedure annotated ~200 fragments in *S. cerevisiae* and ~150 in *E. coli*, leaving a set of total ~1500 peaks in *S. cerevisiae* and ~600 peaks in *E. coli*.

Formula assignment for parent ions.

The first step in metabolite identification from high resolution mass spectrometry data is chemical formula assignment. Owing to the rapid expansion of the number of possible atomic compositions with increasing masses, even with ± 2 ppm mass accuracy, m/z alone is typically insufficient for chemical formula assignment.^{15,39} For the list of metabolite peaks, we searched for matches to their m/z and C/N atom counts in a diversity of metabolite databases (Metlin⁴⁰, Lipidmaps⁴¹, HMDB³⁵, YMDB⁴² and ECMDB⁴³).

Take m/z 171.006 as an example: at 5 ppm accuracy, there are three candidate formulae: $C_3H_9O_6P$, $C_6H_8N_2S_2$, $C_5H_4N_2O_5$, but only one of these matches the observed C count of 3 and N count of 0 (Figure S7). Another example is m/z 310.1634. Two formulae, $C_{12}H_{21}N_7O_3$ ($m/z=310.1633$) and $C_{20}H_{25}NS$ ($m/z=310.1635$), give the same mass error (< 0.5 ppm), but are distinguished by C/N count. Overall, we obtained a database formula match based on m/z and C/N count for 663 peaks, about 30% of metabolite ions. There was a strong enrichment for peaks with formula match among the more intense peaks, whereas the smaller peaks were enriched for unknowns (Figure 5 A, B). Some unknowns appear to be products of reactions between metabolites and environmental compounds. For example, the peak at m/z 512.2181 appears to be a condensation product of glutathione with an environmental compound, based on C/N count and MS/MS (Figure S9). Efforts to more fully annotate the unknown peaks are ongoing.

Metabolite identification.

For the 663 metabolite ions with a formula match, we attempted to confirm their structures based on RT match to authenticated standards run in our laboratory or MS/MS spectral match to reference libraries. For example, the metabolite of m/z 171.006 assigned above to formula $C_3H_9O_6P$ was annotated as glycerol-3-phosphate based on RT match to the purified standard. Finally, 221 peak identities were confirmed, 148 by RT match and 73 by database MS/MS spectral match (Table 2, Supporting Information Table S10). About 40% of these metabolites were detected in both organisms.

Evaluation of PAVE.

To evaluate PAVE, we used a list of 104 metabolites that were previously quantitated in *E. coli* using reversed-phase ion-pairing LC-MS with standard spiking ($N = 104$)⁴⁴ From this list, PAVE correctly detected and assigned C/N counts for 82. For the missing compounds, reasons included intensity below the selected PAVE threshold of 10^3 (e.g., guanine), high background (e.g., citrate), quantitative error leading to low ρ in ATOMCOUNT (e.g., IMP),

instability (e.g. acetoacetyl-CoA), and annotation as fragments (gluconate, dTDP, and dTTP, all of which are low abundance). In addition, adenosine and deoxyguanosine failed to chromatographically resolve and were recognized as one metabolite. As expected given our 99 percentile cut-off for carbon number versus mass, one compound (GTP) fell below this cut-off and was erroneously removed. For the 82 metabolites correctly assigned C/N counts, database search yielded the correct formula for 77.

Using a ^{13}C -labeling strategy similar in some respects to PAVE, Mahieu and Patti reduced 25,000 positive mode peaks from reversed-phase LC-MS to a set of 776 likely metabolite peaks with carbon number using de-adducting, de-isotoping, and ^{13}C -based credentialing.²⁸ This extent of reduction is similar to that achieved with PAVE. There are, however, important methodological differences between the two studies, which resulted in striking differences in the verified peaks. Methodologically, the present work is distinguished by HILIC chromatography, positive and negative ion mode analysis, and algorithmic improvements. Presumably due to these differences, of the 1500 biological peaks (assigned C and N count) that we observe in *E. coli*, only 122 overlap with peaks (assigned C count) found by Mathieu and Patti. Manual inspection identified that a number of putative metabolite ions reported by Mathieu and Patti are obvious adducts (e.g. sodium, potassium). Moreover, of the list of 104 quantified water soluble metabolites⁴⁴, based on m/z match ± 10 ppm and C count, Mahieu and Patti found only ~ 7%, while PAVE detected ~ 80%.

Relevance to mammalian metabolism.

The ATOMCOUNT strategy is most easily applied to microorganisms that can be readily cultured in uniformly ^{15}N and/or ^{13}C media, and it should be immediately useful for examining different microbes. While complete labeling is more difficult to achieve in mammalian cells and intact multicellular organisms, upon feeding suitably labeled substrates^{45–48}, the strategy may also be applicable to them, although this depends on either achievement of sufficiently complete labeling or algorithmic advances to deal with incomplete labeling.

Importantly, even in the absence of such experiments, peak annotation from *S. cerevisiae* and *E. coli* immediately informs annotation of the same peaks from animal specimens. To this end, we looked in mouse liver data for peaks corresponding to 432 metabolite standards (that we routinely track in our lab, Table S8) augmented by an additional 663 metabolite ions with database formula match from the present analysis (Figure 5C). In liver, we found 158 peaks matching to the metabolite standards based on m/z and RT, and 324 peaks matching the additional metabolite ions from the present study. Thus, these efforts in microbes are already enhancing our ability to annotate untargeted LC-MS data from mammalian specimens, tripling the number of verified, readily measurable peaks that we can routinely track in liver.

Conclusion

Here we “PAVE” the way to comprehensive annotation of the metabolites detected in untargeted LC/MS metabolomics data. The PAVE strategy combines isotopic labeling and computational data analysis to pick out biological peaks, assign C and N atom counts, and

annotate adducts and natural isotopic variants. Remaining peaks are then discriminated from fragments based on a simple experimental strategy involving selective enhancement of fragment intensity by applying a weak in-source CID energy. PAVE annotates < 20% of all LC-MS peaks as biological, with about 25% of biological peaks recognized as putative metabolites and 30% of putative metabolites matching known formula. The PAVE workflow is immediately applicable to any organism that can be grown on uniformly ^{13}C and ^{15}N -labeled nutrients, opening up the possibility for metabolome annotation for less studied microbes that may have advantageous properties for industrial metabolic engineering or biofuel production.

Overall, PAVE discarded 96% of peaks and credentialed roughly 2000 peaks with C and N counts, a greater number of credentialed peaks with atom counts than prior efforts^{22,25,27,28}. The outcome of our peak annotation by PAVE supports the emerging consensus that common microbes contain at most a few thousand readily detectable metabolites. A critical question is whether most of the peaks annotated as unknown metabolites here are truly biological metabolites, or reflect currently unappreciated types of artifacts. We hope that the credentialed peak list reported here will spur further work to elucidate the molecular formulae and structures giving rise to these peaks. As annotation is carried out in different contexts, it seems likely that studies will converge on a relatively small set of true water-soluble metabolites, perhaps in the range of 1000.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank members of Rabinowitz lab for input. This research is supported by NIH grants R01CA163591, P30CA072720, DP1DK113643, P30DK019525 and DOE grants DE-SC0018420, DE-SC0018260 to JDR; NIH grant R50CA211437 to WL; and DOE grant DE-FG02-02ER15344 and ARO grant W911NF-16-1-0014 to HR.

References

- (1). Wishart DS Emerging Applications of Metabolomics in Drug Discovery and Precision Medicine. *Nature Reviews Drug Discovery* 2016, 15 (7), 473–484. 10.1038/nrd.2016.32. [PubMed: 26965202]
- (2). Gowda H; Ivanisevic J; Johnson CH; Kurczy ME; Benton HP; Rinehart D; Nguyen T; Ray J; Kuehl J; Arevalo B; et al. Interactive XCMS Online: Simplifying Advanced Metabolomic Data Processing and Subsequent Statistical Analyses. *Analytical Chemistry* 2014, 86 (14), 6931–6939. 10.1021/ac500734c. [PubMed: 24934772]
- (3). Huan T; Forsberg EM; Rinehart D; Johnson CH; Ivanisevic J; Benton HP; Fang M; Aisporna A; Hilmers B; Poole FL; et al. Systems Biology Guided by XCMS Online Metabolomics. *Nature Methods* 2017, 14 (5), 461–462. 10.1038/nmeth.4260. [PubMed: 28448069]
- (4). Smith CA; Want EJ; O'Maille G; Abagyan R; Siuzdak G XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry* 2006, 78 (3), 779–787. 10.1021/ac051437y. [PubMed: 16448051]
- (5). Katajamaa M; Miettinen J; Oresic M MZmine: Toolbox for Processing and Visualization of Mass Spectrometry Based Molecular Profile Data. *Bioinformatics* 2006, 22 (5), 634–636. 10.1093/bioinformatics/btk039. [PubMed: 16403790]

- (6). Pluskal T; Castillo S; Villar-Briones A; Orešić M MZmine 2: Modular Framework for Processing, Visualizing, and Analyzing Mass Spectrometry-Based Molecular Profile Data. *BMC Bioinformatics* 2010, 11 (1), 395. 10.1186/1471-2105-11-395. [PubMed: 20650010]
- (7). Lommen A MetAlign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing. *Analytical Chemistry* 2009, 81 (8), 3079–3086. 10.1021/ac900036d. [PubMed: 19301908]
- (8). Tsugawa H; Cajka T; Kind T; Ma Y; Higgins B; Ikeda K; Kanazawa M; VanderGheynst J; Fiehn O; Arita M MS-DIAL: Data-Independent MS/MS Deconvolution for Comprehensive Metabolome Analysis. *Nature Methods* 2015, 12 (6), 523–526. 10.1038/nmeth.3393. [PubMed: 25938372]
- (9). Domingo-Almenara X; Montenegro-Burke JR; Benton HP; Siuzdak G Annotation: A Computational Solution for Streamlining Metabolomics Analysis. *Analytical Chemistry* 2018, 90 (1), 480–489. 10.1021/acs.analchem.7b03929. [PubMed: 29039932]
- (10). Johnson CH; Ivanisevic J; Benton HP; Siuzdak G *Bioinformatics: The Next Frontier of Metabolomics*. *Analytical Chemistry* 2015, 87 (1), 147–156. 10.1021/ac5040693. [PubMed: 25389922]
- (11). Alonso A; Marsal S; Juliá A Analytical Methods in Untargeted Metabolomics: State of the Art in 2015. *Frontiers in Bioengineering and Biotechnology* 2015, 3 10.3389/fbioe.2015.00023.
- (12). Kuhl C; Tautenhahn R; Böttcher C; Larson TR; Neumann S CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Analytical Chemistry* 2012, 84 (1), 283–289. 10.1021/ac202450g. [PubMed: 22111785]
- (13). DeFelice BC; Mehta SS; Samra S; Cajka T; Wanciewicz B; Fahrman JF; Fiehn O Mass Spectral Feature List Optimizer (MS-FLO): A Tool To Minimize False Positive Peak Reports in Untargeted Liquid Chromatography–Mass Spectroscopy (LC-MS) Data Processing. *Analytical Chemistry* 2017, 89 (6), 3250–3255. 10.1021/acs.analchem.6b04372. [PubMed: 28225594]
- (14). Giavalisco P; Li Y; Matthes A; Eckhardt A; Hubberten H-M; Hesse H; Segu S; Hummel J; Köhl K; Willmitzer L Elemental Formula Annotation of Polar and Lipophilic Metabolites Using ¹³C, ¹⁵N and ³⁴S Isotope Labelling, in Combination with High-Resolution Mass Spectrometry: Isotope Labelling for Unbiased Plant Metabolomics. *The Plant Journal* 2011, 68 (2), 364–376. 10.1111/j.1365-3113.2011.04682.x. [PubMed: 21699588]
- (15). Hegeman AD; Schulte CF; Cui Q; Lewis IA; Huttlin EL; Eghbali H; Harms AC; Ulrich EL; Markley JL; Sussman MR Stable Isotope Assisted Assignment of Elemental Compositions for Metabolomics. *Analytical Chemistry* 2007, 79 (18), 6912–6921. 10.1021/ac070346t. [PubMed: 17708672]
- (16). Hiller K; Wegner A; Weindl D; Cordes T; Metallo CM; Kelleher JK; Stephanopoulos G NTFD--a Stand-Alone Application for the Non-Targeted Detection of Stable Isotope-Labeled Compounds in GC/MS Data. *Bioinformatics* 2013, 29 (9), 1226–1228. 10.1093/bioinformatics/btt119. [PubMed: 23479350]
- (17). Clasquin MF; Melamud E; Rabinowitz JD LC-MS Data Processing with MAVEN: A Metabolomic Analysis and Visualization Engine In *Current Protocols in Bioinformatics*; Baxevanis AD, Petsko GA, Stein LD, Stormo GD, Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2012 10.1002/0471250953.bi1411s37.
- (18). Melamud E; Vastag L; Rabinowitz JD Metabolomic Analysis and Visualization Engine for LC–MS Data. *Analytical Chemistry* 2010, 82 (23), 9818–9826. 10.1021/ac1021166. [PubMed: 21049934]
- (19). Scheltema RA; Jankevics A; Jansen RC; Swertz MA; Breitling R PeakML/MzMatch: A File Format, Java Library, R Library, and Tool-Chain for Mass Spectrometry Data Analysis. *Analytical Chemistry* 2011, 83 (7), 2786–2793. 10.1021/ac2000994. [PubMed: 21401061]
- (20). Chokkathukalam A; Jankevics A; Creek DJ; Achcar F; Barrett MP; Breitling R MzMatch–ISO: An R Tool for the Annotation and Relative Quantification of Isotope-Labelled Mass Spectrometry Data. *Bioinformatics* 2013, 29 (2), 281–283. 10.1093/bioinformatics/bts674. [PubMed: 23162054]

- (21). Huang X; Chen Y-J; Cho K; Nikolskiy I; Crawford PA; Patti GJX ¹³ CMS: Global Tracking of Isotopic Labels in Untargeted Metabolomics. *Analytical Chemistry* 2014, 86 (3), 1632–1639. 10.1021/ac403384n. [PubMed: 24397582]
- (22). Kessler N; Walter F; Persicke M; Albaum SP; Kalinowski J; Goesmann A; Niehaus K; Nattkemper TW ALLocator: An Interactive Web Platform for the Analysis of Metabolomic LC-ESI- MS Datasets, Enabling Semi-Automated, User-Revised Compound Annotation and Mass Isotopomer Ratio Analysis. *PLoS ONE* 2014, 9 (11), e113909. 10.1371/journal.pone.0113909.
- (23). de Jong FA; Beecher C Addressing the Current Bottlenecks of Metabolomics: Isotopic Ratio Outlier Analysis™, an Isotopic-Labeling Technique for Accurate Biochemical Profiling. *Bioanalysis* 2012, 4 (18), 2303–2314. 10.4155/bio.12.202. [PubMed: 23046270]
- (24). Qiu Y; Moir R; Willis I; Beecher C; Tsai Y-H; Garrett TJ; Yost RA; Kurland IJ Isotopic Ratio Outlier Analysis of the *S. Cerevisiae* Metabolome Using Accurate Mass Gas Chromatography/Time-of-Flight Mass Spectrometry: A New Method for Discovery. *Analytical Chemistry* 2016, 88 (5), 2747–2754. 10.1021/acs.analchem.5b04263. [PubMed: 26820234]
- (25). Qiu Y; Moir R; Willis I; Seethapathy S; Biniakewitz R; Kurland I Enhanced Isotopic Ratio Outlier Analysis (IROA) Peak Detection and Identification with Ultra-High Resolution GC-Orbitrap/MS: Potential Application for Investigation of Model Organism Metabolomes. *Metabolites* 2018, 8 (1), 9 10.3390/metabo8010009.
- (26). Bueschl C; Kluger B; Berthiller F; Lirk G; Winkler S; Krska R; Schuhmacher R MetExtract: A New Software Tool for the Automated Comprehensive Extraction of Metabolite-Derived LC/MS Signals in Metabolomics Research. *Bioinformatics* 2012, 28 (5), 736–738. 10.1093/bioinformatics/bts012. [PubMed: 22238263]
- (27). Bueschl C; Kluger B; Neumann NKN; Doppler M; Maschietto V; Thallinger GG; Meng-Reiterer J; Krska R; Schuhmacher R MetExtract II: A Software Suite for Stable Isotope-Assisted Untargeted Metabolomics. *Analytical Chemistry* 2017, 89 (17), 9518–9526. 10.1021/acs.analchem.7b02518. [PubMed: 28787149]
- (28). Mahieu NG; Patti GJ Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites. *Analytical Chemistry* 2017, 89 (19), 10397–10406. 10.1021/acs.analchem.7b02380. [PubMed: 28914531]
- (29). Mitchell JM; Flight RM; Wang QJ; Higashi RM; Fan TW-M; Lane AN; Moseley HNB New Methods to Identify High Peak Density Artifacts in Fourier Transform Mass Spectra and to Mitigate Their Effects on High-Throughput Metabolomic Data Analysis. *Metabolomics* 2018, 14 (10). 10.1007/s11306-018-1426-9.
- (30). Lu W; Wang L; Chen L; Hui S; Rabinowitz JD Extraction and Quantitation of Nicotinamide Adenine Dinucleotide Redox Cofactors. *Antioxidants & Redox Signaling* 2018, 28 (3), 167–179. 10.1089/ars.2017.7014. [PubMed: 28497978]
- (31). GUTNICK D; CALVO JM; KLOPOTOWSKI T; AMES BN Compounds Which Serve as the Sole Source of Carbon or Nitrogen for *Salmonella Typhimurium* LT-2. 1969, 100, 5.
- (32). Holman JD; Tabb DL; Mallick P Employing ProteoWizard to Convert Raw Mass Spectrometry Data: Employing ProteoWizard to Convert Raw Mass Spectrometry Data In Current Protocols in Bioinformatics; Bateman A, Pearson WR, Stein LD, Stormo GD, Yates JR, Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2014; pp 13.24.1–13.24.9. 10.1002/0471250953.bi1324s46.
- (33). Myers OD; Sumner SJ; Li S; Barnes S; Du X Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data. *Analytical Chemistry* 2017, 89 (17), 8689–8695. 10.1021/acs.analchem.7b01069. [PubMed: 28752757]
- (34). Keller BO; Sui J; Young AB; Whittall RM Interferences and Contaminants Encountered in Modern Mass Spectrometry. *Analytica Chimica Acta* 2008, 627 (1), 71–81. 10.1016/j.aca.2008.04.043. [PubMed: 18790129]
- (35). Wishart DS; Tzur D; Knox C; Eisner R; Guo AC; Young N; Cheng D; Jewell K; Arndt D; Sawhney S; et al. HMDB: The Human Metabolome Database. *Nucleic Acids Research* 2007, 35 (Database), D521–D526. 10.1093/nar/gkl923. [PubMed: 17202168]
- (36). Lu W; Su X; Klein MS; Lewis IA; Fiehn O; Rabinowitz JD Metabolite Measurement: Pitfalls to Avoid and Practices to Follow. *Annual Review of Biochemistry* 2017, 86 (1), 277–304. 10.1146/annurev-biochem-061516-044952.

- (37). Gabelica V; Pauw ED Internal Energy and Fragmentation of Ions Produced in Electrospray Sources. *Mass Spectrometry Reviews* 2005, 24 (4), 566–587. 10.1002/mas.20027. [PubMed: 15317019]
- (38). Xu Y-F; Lu W; Rabinowitz JD Avoiding Misannotation of In-Source Fragmentation Products as Cellular Metabolites in Liquid Chromatography–Mass Spectrometry-Based Metabolomics. *Analytical Chemistry* 2015, 87 (4), 2273–2281. 10.1021/ac504118y. [PubMed: 25591916]
- (39). Kind T; Fiehn O Metabolomic Database Annotations via Query of Elemental Compositions: Mass Accuracy Is Insufficient Even at Less than 1 Ppm. *BMC Bioinformatics* 2006, 10.
- (40). Tautenhahn R; Cho K; Uritboonthai W; Zhu Z; Patti GJ; Siuzdak G An Accelerated Workflow for Untargeted Metabolomics Using the METLIN Database. *Nature Biotechnology* 2012, 30 (9), 826–828. 10.1038/nbt.2348.
- (41). Fahy E; Sud M; Cotter D; Subramaniam S LIPID MAPS Online Tools for Lipid Research. *Nucleic Acids Research* 2007, 35 (Web Server), W606–W612. 10.1093/nar/gkm324. [PubMed: 17584797]
- (42). Jewison T; Knox C; Neveu V; Djoumbou Y; Guo AC; Lee J; Liu P; Mandal R; Krishnamurthy R; Sinelnikov I; et al. YMDB: The Yeast Metabolome Database. *Nucleic Acids Research* 2012, 40 (D1), D815–D820. 10.1093/nar/gkr916. [PubMed: 22064855]
- (43). Guo AC; Jewison T; Wilson M; Liu Y; Knox C; Djoumbou Y; Lo P; Mandal R; Krishnamurthy R; Wishart DS ECMDB: The E. Coli Metabolome Database. *Nucleic Acids Research* 2012, 41 (D1), D625–D630. 10.1093/nar/gks992. [PubMed: 23109553]
- (44). Bennett BD; Kimball EH; Gao M; Osterhout R; Van Dien SJ; Rabinowitz JD Absolute Metabolite Concentrations and Implied Enzyme Active Site Occupancy in *Escherichia Coli*. *Nature Chemical Biology* 2009, 5 (8), 593–599. 10.1038/nchembio.186. [PubMed: 19561621]
- (45). Hui S; Ghergurovich JM; Morscher RJ; Jang C; Teng X; Lu W; Esparza LA; Reya T; Le Zhan; Yanxiang Guo J.; et al. Glucose Feeds the TCA Cycle via Circulating Lactate. *Nature* 2017, 551 (7678), 115–118. 10.1038/nature24057. [PubMed: 29045397]
- (46). Sun RC; Fan TW-M; Deng P; Higashi RM; Lane AN; Le A-T; Scott TL; Sun Q; Warmoes MO; Yang Y Noninvasive Liquid Diet Delivery of Stable Isotopes into Mouse Models for Deep Metabolic Network Tracing. *Nature Communications* 2017, 8 (1). 10.1038/s41467-017-01518-z.
- (47). Faubert B; Li KY; Cai L; Hensley CT; Kim J; Zacharias LG; Yang C; Do QN; Doucette S; Burguete D; et al. Lactate Metabolism in Human Lung Tumors. *Cell* 2017, 171 (2), 358–371.e9. 10.1016/j.cell.2017.09.019. [PubMed: 28985563]
- (48). Jang C; Chen L; Rabinowitz JD Metabolomics and Isotope Tracing. *Cell* 2018, 173 (4), 822–837. 10.1016/j.cell.2018.03.055. [PubMed: 29727671]

Figure 1.

Flow chart for untargeted peak annotation by PAVE. *S. cerevisiae* and *E. coli* were grown in four conditions (unlabeled, ^{15}N , ^{13}C , and $^{15}\text{N}+^{13}\text{C}$). Metabolites were analyzed by HILIC LC- high resolution MS. For each peak in the unlabeled sample, ATOMCOUNT computationally assesses whether the labeled samples show a logical pattern of signals, generating a list of candidate biological metabolites with known nitrogen and carbon atom counts (N and C , respectively). JUNKREMOVER then computationally differentiates metabolite ions from natural isotopic variants, adducts, and other artifacts. To discriminate

between metabolite ions and fragments, all ions were subjected to a weak in-source CID energy, which increases the signal intensity for fragments while decreasing it for parent ions. The metabolite ions were then subjected to MS/MS in an effort to assign molecular structures based on database matching.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 2.

ATOMCOUNT. (A) Workflow. Every peak is assessed, one at a time, to determine whether it shows a logical labeling pattern for any choice of carbon atom count ($C \geq 1$) and nitrogen atom count ($N \geq 0$). (B) Idealized case (theoretical intensity pattern), where the unlabeled, carbon-labeled, nitrogen-labeled and dual labeled peaks all have the identical intensity. The entries in the matrix are peak intensities that are normalized to the peak intensity of the unlabeled sample. Zeroes in the matrix indicate the absence of unlabeled peaks in the labeled samples, and vice versa. (C) Example of peaks from environmental contaminants,

the most common type of peaks found in the data. Note that the signal does not shift with labeling. (D) Example of biological peak at m/z 505.9882 for the candidate atom counts of $N=4$ and $C=11$; the correlation is weak because these are not the correct atom counts. (E) Example of the same biological peak at m/z 505.9882, for $N=5$ and $C=10$, matching the molecular formula of $C_{10}H_{16}N_5O_{13}P_3$ (corresponding to ATP).

Figure 3.

JUNKREMOVER for de-adducting, de-isotoping and filtering out peaks with too low C count for the mass. For each peak found by ATOMCOUNT, JUNKREMOVER assesses whether there is a lower molecular weight peak at the same RT where the mass difference, C and N atom counts, and relative peak intensities suggest that the peak at the lower molecular weight is the metabolite ion. (A and B) Glucose adducts annotation benefits from C/N count matching and searching for the (de)protonated metabolite ion also in the opposite ionization mode data (for glucose, the $[M-H]^-$ but not $[M+H]^+$ ion is readily observed). In addition to the abundant adducts visible here, there are another ~45 less abundant glucose adducts (Table S10). (C and D) Molecular weight as a function of C count, for known metabolites from HMDB (gray dots) versus biological-derived peaks observed in *S. cerevisiae* (after de-adducting and de-isotoping) (blue and red dots). The black line defines the 99th percentile cutoff of known metabolites. Blue dots are retained peaks and red dots are discarded peaks.

Figure 4.

Impact of in-source CID voltage on the intensity of parent ions and fragments. (A) Mass spectrum for glucose and its adducts and fragments at in-source CID energies of 0, 2, 4, 6, 10, 14, 20 eV. (B) Peak intensity at 2, 4, 6, 15 eV (compared to 0 eV) for 80 metabolite standards and their related fragments. The standards were dissolved in extraction buffer. Note that application of an in-source CID energy universally decreases the parent ion intensity, whereas a small (2 or 4 eV) in-source CID energy consistently increases fragment intensity.

Figure 5.

Outcome of biological peak annotation by PAVE. (A, B) Annotation as a function of peak intensity (log10 intensity of 3 refers to all peaks with height between 10^3 to 10^4 , etc.). (C) Venn diagram showing number of metabolites with assigned formulae (found in PAVE + 432 standards commonly tracked in our lab) across *S. cerevisiae*, *E. coli* and mouse liver. (D) Venn diagram showing number of metabolite ions found in PAVE but without assigned formulae across *S. cerevisiae*, *E. coli* and mouse liver.

Table 1.

Labeling conditions and expected masses for a metabolite of mass M with C carbon atoms and N nitrogen atoms.

Media	Mass
Unlabeled	M
^{15}N -ammonia	$M+0.9970\text{ N}$
^{13}C -glucose	$M+1.0034\text{ C}$
^{15}N -ammonia, ^{13}C -glucose	$M+0.9970\text{ N}+1.0034\text{ C}$

Table 2.

Overview of peak annotation in PAVE.

Organism		<i>S. cerevisiae</i>		<i>E. coli</i>	
Ionization mode		Positive	Negative	Positive	Negative
Total peak number		30,212	14,927	28,761	14,921
ATOMCOUNT	Peaks in procedure blank	21,186	10,015	20,600	10,147
	Other peaks without labeling	3,591	2,393	6,312	3,422
	Labeling but $\rho < 0.75$	820	380	339	201
	Logical labeling (i.e. biological)	4,615	2,139	1,510	1,151
JUNKREMOVER	Isotopes	1,598	632	467	404
	Dimer or double charge	313	21	25	45
	Adducts (assigned using same polarity mode)	1,153	539	431	278
	Adducts (assigned only using opposite polarity mode)	138	76	55	11
	Too low C/N count for mass	363	70	32	12
	Ring peaks	28	47	7	11
In-source CID	Fragments	96	64	90	71
	Parent ions	926	690	403	319
Formula assignment	Formula match to metabolite	278	359	132	164
	No formula match in database	648	331	271	155
Metabolite structure identification (With formula)	RT match	144		92	
	MS/MS match	35		30	
	Unknown structure	359		133	