analytical chemistry

Article

Cite This: *Anal. Chem.* 2019, 91, 12108−12115

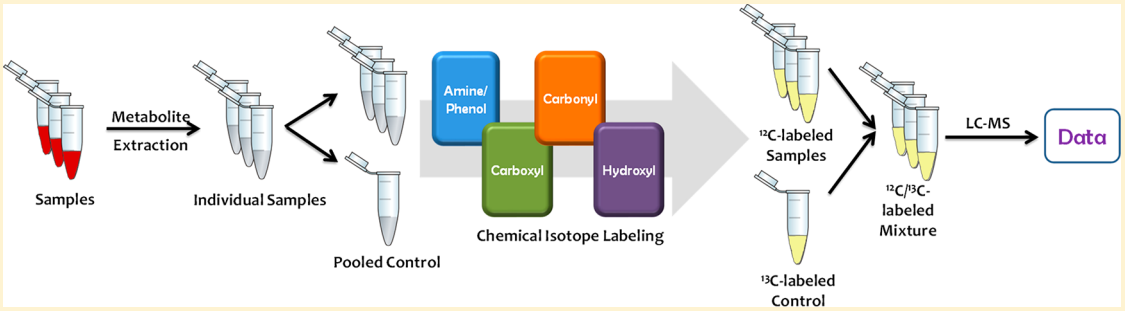
pubs.acs.org/ac

Metabolomic Coverage of Chemical-Group-Submetabolome Analysis: Group Classification and Four-Channel Chemical Isotope Labeling LC-MS

Shuang Zhao,[†] Hao Li,[†] Wei Han, Wan Chan, and Liang Li*

Department of Chemistry, University of Alberta, Edmonton, Alberta T6G 2G2, Canada

Supporting Information



ABSTRACT: Chemical isotope labeling (CIL) liquid chromatography mass spectrometry (LC-MS) is a powerful technique for in-depth metabolome analysis with high quantification accuracy. Unlike conventional LC-MS, it analyzes chemical-group-based submetabolomes and uses the combined results to represent the whole metabolome. Due to analysis time and cost constraint, not all submetabolomes can be profiled and thus knowledge of chemical group classification is important in guiding submetabolome selection. Herein we report a study of determining the distribution of functional groups of compounds in a database and then examine how well we can experimentally analyze the major chemical groups in two representative samples (i.e., human plasma and yeast). We developed a computer algorithm to classify chemical structures according to their functional groups. After removing lipids which are targeted molecules in lipidomic analysis, inorganic species and other molecules that are unique to drug, food, plant, and environmental origins, five groups (i.e., amine, phenol, hydroxyl, carboxyl, and carbonyl) are found to be the dominant classes. In the databases of MCID (2683 filtered metabolites), HMDB (5506), KEGG (11598), YMDB (1107), and ECMDB (1462), 94.7%, 85.7%, 86.4%, 85.7%, and 95.8% of the filtered metabolites belong to one or more of the five groups, respectively. These groups can be analyzed in four-channel CIL LC-MS where hydroxyls (H), amines and phenols (A), carboxyls (C), and carbonyls or ketones/aldehydes (K) are separately profiled as individual channels using dansyl and DmPA labeling reagents. A total of 7431 peak pairs were detected with 6109 unique-mass pairs from plasma, while 5629 pairs with 4955 unique-mass pairs were detected in yeast. Compared to group distributions of database compounds, hydroxyl-containing metabolites were severely underdetected, which might indicate that the current method is less than optimal for analyzing this group of metabolites. As a result, the overall experimental coverage is likely significantly lower than the database-derived coverage. In short, this study has shown that high metabolome coverage is theoretically attainable by analyzing only the H, A, C, and K submetabolomes and the group classification information should be helpful in guiding future analytical method development and choices of submetabolomes to be analyzed.

Because of great diversity of chemical and physical properties of metabolites present in a complex metabolome sample, the conventional liquid chromatography mass spectrometry (LC-MS) approach of metabolome analysis relies on the use of multiple LC and MS conditions to increase the number of metabolites detectable or the metabolome coverage. For example, the combination of a reversed-phase (RP) LC column for separation of relatively hydrophobic metabolites and a hydrophilic interaction liquid chromatography (HILIC) column for separation of relatively hydrophilic metabolites, along with positive and negative ion MS detection, allows the detection of different types of metabolites.[1] This approach has the advantage of using a simple workflow with readily available instrumentation and software for metabolite detection and data analysis and thus can be easily implemented. However, this approach has the shortcomings of limited metabolome coverage due to low detectability of many metabolites and limited quantification accuracy due to the lack of suitable internal standards for a vast majority of metabolites. Chemical isotope labeling (CIL) LC-MS offers a means of overcoming these limitations.[2]

Received: July 28, 2019
Accepted: August 23, 2019
Published: August 23, 2019

DOI: 10.1021/acs.analchem.9b03431
Anal. Chem. 2019, 91, 12108−12115

Derivatization reaction/class	Patterns of functional groups	SMARTS substructure patterns
Amine/phenol labeling	Primary or secondary amines, N-H bond in aromatic environment, phenols	[NX3;H2,H1;!$(NC=O)], [nX3;H2,H1;!$(nc=O)], [OX2H][cX3]:[n,c]
Carbonyl labeling	Aldehydes, ketones	[CX3H1](=O), [#6][CX3](=O)[#6], [#6][cX3](=O)[#6]
Hydroxyl labeling	Hydroxyls, thiols	[CX4][OX2H], [CX3,NX2]=[CX3][OX2H], [#16!H0]
Carboxyl labeling	Carboxyls and the conjugated bases	[CX3](=O)[OX1H0-,OX2H1]
Esters	Esters	[#6][CX3](=O)[OX2H0][#6], [#6][cX3](=O)[oX2H0][#6]
Amides	Amides	[nX3][cX3](=[OX1]), [NX3][CX3](=[OX1])

Figure 1. Targeted functional groups for each reaction or class and SMARTS substructure patterns for determining chemical groups.

CIL LC-MS metabolome analysis is a divide-and-conquer approach where the metabolites are divided into different chemical groups (e.g., amines, acids, etc.), instead of dividing them according to physical properties such as hydrophobicity and ionic property.[3,4] Each group of metabolites are chemically labeled with a suitable reagent, followed by LC-MS analysis. Many reagents have been developed for both targeted and untargeted metabolome analysis.[5−10] With rational design of chemical structures of the labeling reagents, concomitant improvement in both metabolite separation and ionization can be achieved, resulting in significant enhancement in metabolite detectability and hence much higher metabolome coverage.[3] Using differential isotope labeling (e.g., ^{12}C-reagent labeled individual samples spiked with a ^{13}C-reagent labeled reference or pooled sample, followed by LC-MS analysis of the resultant mixtures), accurate relative quantification of all labeled metabolites in comparative samples can be performed.[11]

The presumed disadvantage of CIL LC-MS is the requirement of chemical derivatization that may add a complication in sample processing. However, sample processing for metabolome analysis often involves multiple steps[1,2] and thus a robust chemical reaction (e.g., by merely adding a reagent to a sample) may be seamlessly incorporated into the overall workflow, just as it is done for protein precipitation (e.g., by adding a solvent to precipitate proteins and then removing them), sample normalization (e.g., by creatinine measurement for urine samples), cell lysis (e.g., by adding a lysis reagent), metabolite extraction (e.g., by adding a solvent for liquid−liquid extraction), etc. Thus, performing chemical labeling, if properly done, should not inconvenience the sample handling process. The benefits of improving metabolite detectability and quantification accuracy significantly outweigh the addition of an extra labeling step. However, a more fundamental question is actually related to the number of labeling reactions we need to do for a given sample in order to cover the whole chemical space of the metabolome. Addressing this question will allow us to understand the chemical group diversity of a metabolome, prioritize the development efforts on labeling chemistries to target certain groups of metabolites, and examine the deficiency of current labeling methods to guide future method optimization or new labeling method development.

In this study, we report our investigation of chemical group diversity of compound entries in some commonly used metabolome databases. We developed and applied a high-performance four-channel chemical labeling approach, based on dansylation for analyzing amines and phenols,[3] base-activated dansylation for hydroxyls,[12] DmPA bromide labeling for carboxylic acids,[13] and dansylhydrazine (DnsHz) labeling for carbonyl metabolites,[4] for the analysis of human plasma as well as yeast cells in order to examine the current coverage of these different groups of metabolites in representative complex metabolome samples. By comparing the distribution of chemical groups of database compounds with those detected in four-channel CIL LC-MS, we discussed some limitations of the current methods that we hope will stimulate future research activities to meet the ultimate goal of using CIL LC-MS for whole metabolome profiling.

EXPERIMENTAL SECTION

Chemical Group Classification. A Java-based program was developed to classify the compounds in a database according to their chemical groups. The workflow of metabolite classification contains five steps. First, we downloaded the five selected metabolome databases: MyCompoundID (MCID),[14] HMDB,[15] KEGG,[16] YMDB,[17] and ECMDB.[18] MCID is an evidence-based database, including the metabolites detected from human sources and the predicted compounds generated from these human metabolites after subjecting them to one or two common metabolic reactions (MCID-1R as the one-reaction library and MCID-2R as the two-reaction library). In this work, we used MCID zero-reaction library for group classification, as the structures of all entries are known. HMDB collects detailed information about small molecules such as chemical property data, and clinical and biochemical data with the initial focus on human metabolome, but has expanded to include compounds that may be associated with human (e.g., food, drugs, and chemicals of environmental sources) as well as some predicted metabolites. KEGG database includes small molecules along with their biological processing information such as reactions, pathways, and related enzymes as well as biological connectivity among the compounds. YMDB and ECMDB focus on the metabolites and pathway information on two widely used model organisms, *Saccharomyces cerevisiae* (yeast) and *Escherichia coli* (*E. coli*), respectively.

The second step was to extract compound information, including compound names and chemical structures, database

DOI: 10.1021/acs.analchem.9b03431
Anal. Chem. 2019, 91, 12108−12115

ID and other information such as source of compound (e.g., drug and food) and actual or predicted compound. We then built the chemical substructure patterns of functional groups including different patterns for aliphatic and aromatic atoms in molecules (see Figure 1 for the list). Note that thiols are grouped into hydroxyls, as they can be labeled using similar chemical labeling reaction condition. To accurately determine the functional groups (e.g., amines vs amides), the SMARTS (SMiles ARbitrary Target Specification) program (www.daylight.com/dayhtml/doc/theory/theory.smarts) was used to construct the exact substructure patterns (see Figure 1).

The next step was to filter out the unconventional metabolites which we defined as lipids (particularly long-chain lipids), inorganic species, and other molecules that are unique to drug, food, plant, and environmental origins. Although lipids can also be considered as metabolites, they can be extracted and analyzed using methods that are different from CIL methods. Thus, in this study, we excluded the lipids. HMDB contains superclass information such as those denoted as lipids and lipid-like compounds. We used the superclass information to remove lipid and lipid-like compounds. Considering that some lipids were not removed using the superclass information, compounds containing equal to or more than eight-carbon chains with no class information were also filtered out as lipids. Then class information was used to remove flavonoid-, coumarin-, and lignan-related plant compounds. Compounds not containing any carbon were filtered out as inorganic compounds. In addition, we removed drugs and environmental compounds. For KEGG, any compounds without SMILES structure information were removed. Generic compounds representing homologous series were filtered out according to the "Comment" entry. Inorganic compounds were removed. Because KEGG also contains lipids, phytochemical compounds, and others, BRITE, manually generated functional hierarchies[16] (Supporting Information Table S1), was used to filter out the unconventional compounds. Compounds meeting the three following criteria were kept: (1) compounds containing any information on reaction or pathway or module, (2) compounds belonging to 08001, and (3) compounds not containing BRITE information.

The final step of the program was to determine the functional group(s) in a compound structure by matching it to different group substructure patterns. We wrote a Java-based program, SubstrcMatch, which uses chemical structure files (in SMILES format) and substructure patterns as input and then generates a .txt file to indicate whether a compound contains the targeted functional group. The group classification results from different databases were used for metabolomic coverage analyses.

Four-Channel Labeling. The general workflow for metabolome analysis using four-channel CIL LC-MS is shown in Supporting Information Figure S1. It includes the following steps: (1) sample pretreatment and metabolite extraction, (2) generation of a pooled sample by mixing aliquots of all individual samples, (3) dividing a sample into four aliquots, (4) applying four isotope labeling chemistries targeting different submetabolomes, (5) LC-UV quantification of dansyl-labeled metabolites for pre-data-acquisition normalization,[19] (6) mixing of equal moles of ^{12}C-labeled samples and ^{13}C-labeled pooled sample, (7) high-resolution RPLC-MS analysis of ^{12}C-/^{13}C-mixtures, (8) data processing including peak pair picking and peak ratio measurement,[11,20,21] and (9) metabolite identification based on the use of labeled standard library for positive identification[22] and the use of other compound libraries for putative identification.[12,23]

In this work, to demonstrate the performance of the combined analyses of the four submetabolomes, human plasma and yeast samples were labeled using the four chemistries in experimental triplicates. In this case, a sample was divided into two aliquots. One was labeled with ^{12}C-reagent and the other was labeled with ^{13}C-reagent, followed by mixing and LC-MS analysis. Supporting Information Note N1 provides detailed information on sample preparation and labeling.

LC-MS Analysis. The ^{12}C-/^{13}C-labeled mixtures from individual channels were analyzed using a Bruker Compact quadrupole time-of-flight (QTOF) mass spectrometer (Bruker, Billerica, MA) linked to an UltiMate 3000 UHPLC (Thermo Scientific, MA). Supporting Information Note N1 shows the LC-MS conditions used for the analysis. The injection volume for each channel was determined by injection amount optimization experiments (Supporting Information Figure S2)

Data Processing and Metabolite Identification. The resulting LC-MS data were processed using a set of in-house developed software (Supporting Information Note N1). Metabolite identification was carried out at three different levels of confidence, or three tiers, using IsoMS Pro software and database (Nova Medical Testing Inc., Edmonton, Canada). Positive identification as the first tier was based on accurate mass and retention time (RT) search against the labeled standard library currently composed of 1060 unique human endogenous metabolites, including 711 amines/phenols, 187 carboxyls, 85 hydroxyls, and 77 carbonyls. The second tier identification was based on searching against the Linked Identity (LI) Library containing metabolic-pathway-related metabolites (2500 entries extracted from the KEGG database) with accurate mass and predicted RT information. These second tier matches were considered to be high-confidence putative identification. For the third tier, accurate masses of peak pairs were searched against compound entries in metabolome databases, resulting in putative matches; 5506 entries in HMDB were searched for the plasma samples and 1123 entries in YMDB were searched for the yeast cell samples. The remaining unmatched peak pairs were mass-searched against the predicted metabolome libraries (i.e., MCID one- and two-reaction libraries).

RESULTS AND DISCUSSION

Group Classification of Database Entries. We selected some of the commonly used databases for chemical group classification, including MCID, HMDB, KEGG, YMDB, and ECMDB. We applied the group classification program to examine the structures of database compounds and then classify them into different chemical groups. We were particularly interested in the hydroxyl, amine, phenol, carboxyl, and carbonyl groups, as we have already developed the robust labeling methods for labeling these groups. These five groups are covered in four channels of submetabolome profiling: hydroxyl (H), amine/phenol (A), carboxylic acid (C), and ketone/aldehyde (K)-channel. Thus, using the combined results obtained from four-channel CIL LC-MS, we could compare the group coverage of the experimental data with those of the database data.

Before we applied our group classification program to the compounds in all databases, we examined the classification accuracy using the relatively small database, YMDB, where

DOI: 10.1021/acs.analchem.9b03431
Anal. Chem. 2019, 91, 12108−12115

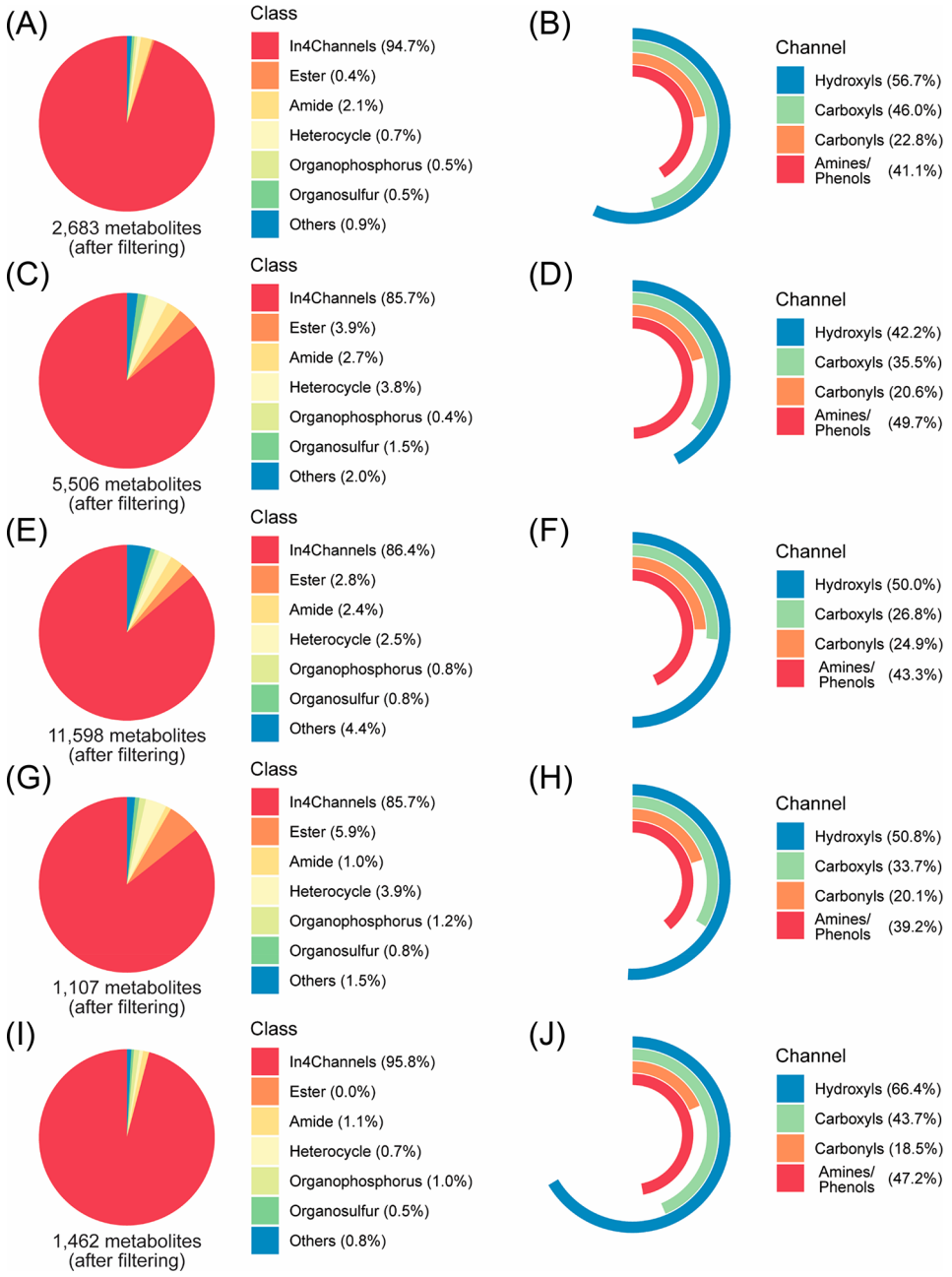


Figure 2. Classification of chemical groups of (A) MCID zero-reaction library, (C) HMDB, (E) KEGG, (G) YMDB, and (I) ECMDB. Sequential class-elimination approach was used to determine the remaining groups (i.e., after removing all the four-channel metabolites, a small number of the remaining metabolites contain the ester group. After removing four-channel metabolites and ester-containing metabolites, a few remaining metabolites contain the amide group). Percent distributions of metabolites belonging to the four channels including overlapped metabolites with two or more functional groups in (B) MCID, (D) HMDB, (F) KEGG, (H) YMDB, and (J) ECMDB.

manual checking of the program-generated classification results was manageable. Out of the 1107 filtered metabolites (i.e., yeast database entries minus the lipids, inorganic species, and hydrocarbons), only four metabolites were misclassified, indicating an error of 0.4%. These four misclassifications were caused by wrong SMILES or resonance structure, as shown in Supporting Information Table S2. Thus, the program was deemed to be very accurate in classifying chemical structures into different chemical groups.

Figure 2 shows the classification results. In all databases except HMDB, the hydroxyl or H-channel covers the highest percentage, i.e., 56.7%, 42.2%, 50.0%, 50.8%, and 66.4% for MCID, HMDB, KEGG, YMDB, and ECMDB, respectively. In contrast, the carbonyl (ketone/aldehyde) or K-channel covers the least, i.e., 22.8%, 20.6%, 24.9%, 20.1%, and 18.5% for MCID, HMDB, KEGG, YMDB, and ECMDB, respectively. In total, the four channels can cover 94.7% of the metabolites in the MCID database containing 2683 filtered metabolites. Similarly, for HMDB (5506 filtered metabolites), KEGG (11598), YMDB (1107), and ECMDB (1462), the four-channel coverage is 85.7%, 86.4%, 85.7%, and 95.8%, respectively. Lower percentages found in HMDB, KEGG, and ECMDB correlate with increased percentages of the ester, amide, and heterocycle groups. Note that, for the groups of ester, amide, heterocycle, organophosphorus, organosulfur, and others shown in Figure 2, we used a sequential group-

DOI: 10.1021/acs.analchem.9b03431
Anal. Chem. 2019, 91, 12108−12115

elimination approach to determine each percentage for clarify (i.e., 100% in total). For example, in Figure 2A, after eliminating 94.7% of the filtered metabolites (2683) which can be analyzed by the four-channel LC-MS method, a small number of the remaining metabolites (0.4% of the total) contain the ester group. Thus, if an ester submetabolome profiling channel is developed in the future, it can only increase the overall metabolome coverage by 0.4%, assuming all hydroxyls, carboxyls, carbonyls, amines, and phenols, including those also containing ester group, have already been covered by the four-channel method. After eliminating the four-channel metabolites and the ester group, a small number of the remaining metabolites (2.1%) contain the amide group. This elimination process applies to the remaining groups sequentially. These analyses indicate that, based on the current entries of the studied databases, very high coverage of the chemical space, ranging from ~86% to 96%, can be achieved using the four-channel profiling approach.

Four-Channel Labeling LC-MS Results. Supporting Information Figure S1 shows a schematic of the four-channel LC-MS approach. As an example, we present the metabolome analysis results obtained from human plasma. In the ion chromatograms of the four-channel submetabolome analyses, many peaks across the entire separation window in RPLC were detected, showing both the chemical diversity and enhanced detectability (Supporting Information Figure S3). The labeling methods allow the conversion of metabolites not retainable in RPLC into relatively hydrophobic derivatives that can be efficiently separated using RPLC. In addition, chemical labeling allows the enhancement of ionization efficiency by ~10 to ~1000 fold.[3] There are smaller enhancements for readily ionizable metabolites and larger enhancements for not-well-ionizable metabolites, resulting in all labeled metabolites having similar detectability (see Supporting Information Figure S4 from the analysis of an equal-mole mixture of 22 labeled standards). Equalizing metabolite ionization efficiency means that the MS peak intensity of a labeled metabolite is more reflective of the metabolite concentration. As a consequence, peak intensity differences for different labeled metabolites are mainly caused by concentration differences, rather than ionization efficiency differences.

Figure 3A shows the distribution of the absolute intensities of metabolite peak pairs detected from the labeled plasma samples. Within the detection dynamic range of the instrument, there is a clear trend of an increase in the number of peak pairs detectable as the peak intensity decreases. Thus, we can increase the metabolome coverage significantly by using a highly sensitive CIL LC-MS method to detect an increasing number of lower concentration metabolites. Interestingly, the four submetabolomes have similar distributions, indicating similar concentration distributions of these different groups of metabolites in plasma.

To further compare the number of metabolites detected in the four channels, Figure 3B shows the Venn diagram of the number of peak pairs detected in each channel. Because of lack of structure identities for many of the detected metabolites, in this comparison, we assumed that the same metabolite was detected in two channels if the same accurate mass of the intact metabolite [i.e., mass of a labeled metabolite minus the mass of labeling tag(s)] was found in the two channels. For example, there are 29 peak pairs detected in all four channels, as these peak-pair masses minus the mass of labeling tag(s) give the same mass; each one of them is deemed to be from the

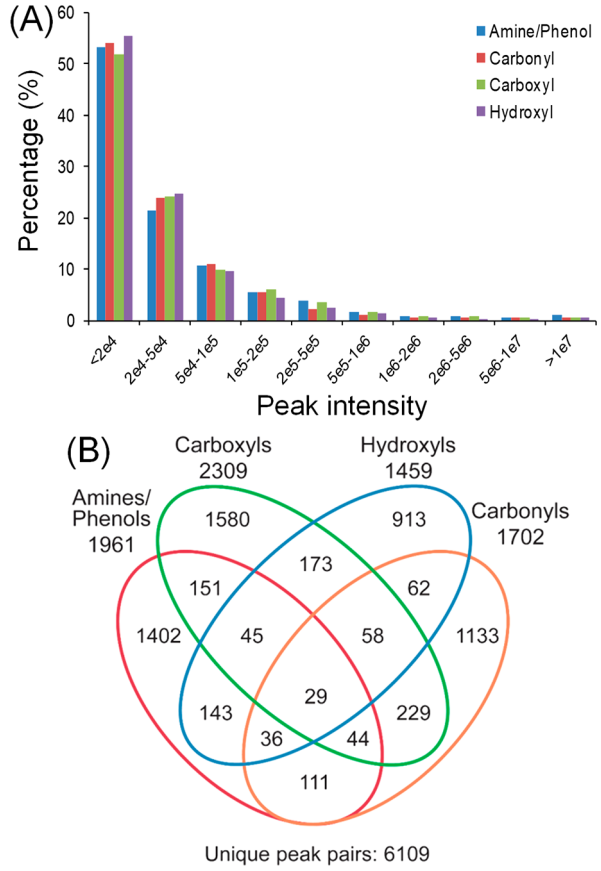


Figure 3. (A) Percentage of peak pair detected in four-channel LC-MS analysis of plasma as a function of peak intensity. (B) Venn diagram of the numbers of peak pairs detected in four channels.

same metabolite that were detected four times. This is a conservative approach of determining the unique metabolites detected, as one would expect that metabolites with the same mass may have different structures (e.g., isomers; see below) and thus belong to different molecules. There are 1961, 2309, 1702, and 1459 peak pairs detected in the amine/phenol, carboxyl, carbonyl, and hydroxyl submetabolome, respectively. If we only count the overlapped metabolites as one unique-mass metabolite, out of a combined total of 7431 peak pairs detected from the four channels, there are 6109 unique-mass peak pairs.

Many of the detected peak pairs can be identified or matched to metabolome databases. Table 1 shows the number of identified peak pairs from each channel in three tiers with the lists shown in Supporting Information Tables S3–S6. For the plasma sample, out of the 7431 pairs detected, we positively identified 326 peak pairs based on accurate mass and retention time matches in tier 1. A few of the peak pairs could be matched to the same metabolite (e.g., carbonyl_594, carbonyl_635, and carbonyl_657 matched to butanal). These matches were manually checked from the LC-MS data and are likely structural isomers of one chemical formula. This example suggests that using mass-match to filter out overlap peaks from two or more channels, as discussed above, might remove some same-mass metabolites with different structures. In tier 2 where authentic standards are not available, but accurate mass and predicted RT data are available in the Linked Identity (LI) library, we identified 344 peak pairs by mass and RT matches. Thus, a total of 670 peak pairs (9.0%) can be identified as

DOI: 10.1021/acs.analchem.9b03431
Anal. Chem. 2019, 91, 12108–12115

Table 1. Summary of the Number of Peak Pairs Identified or Matched against Different Compound Libraries from the Human Plasma Samples Analyzed Using Four-Channel LC-MS

	A-channel	C-channel	H-channel	K-channel	total per tier	Supporting Information table
tier 1	208	54	23	41	326	Table S3
tier 2	71	157	68	48	344	Table S4
tier 3-HMDB	774	760	449	645	2628	Table S5
tier 3-MCID1R	570	1014	609	658	2851	Table S6
tier 3-MCID2R	168	242	165	202	777	Table S6
total per channel	1791	2227	1314	1594	6926	

Table 2. Summary of the Number of Peak Pairs Identified or Matched against Different Compound Libraries from the Yeast Cell Samples Analyzed Using Four-Channel LC-MS

	A-channel	C-channel	H-channel	K-channel	total per tier	Supporting Information table
tier 1	123	68	20	32	243	Table S7
tier 2	69	80	18	21	188	Table S8
tier 3-YMDB	297	261	163	159	880	Table S9
tier 3-MCID1R	944	1255	610	633	3442	Table S10
tier 3-MCID2R	99	155	116	144	514	Table S10
total per channel	1532	1819	927	989	5267	

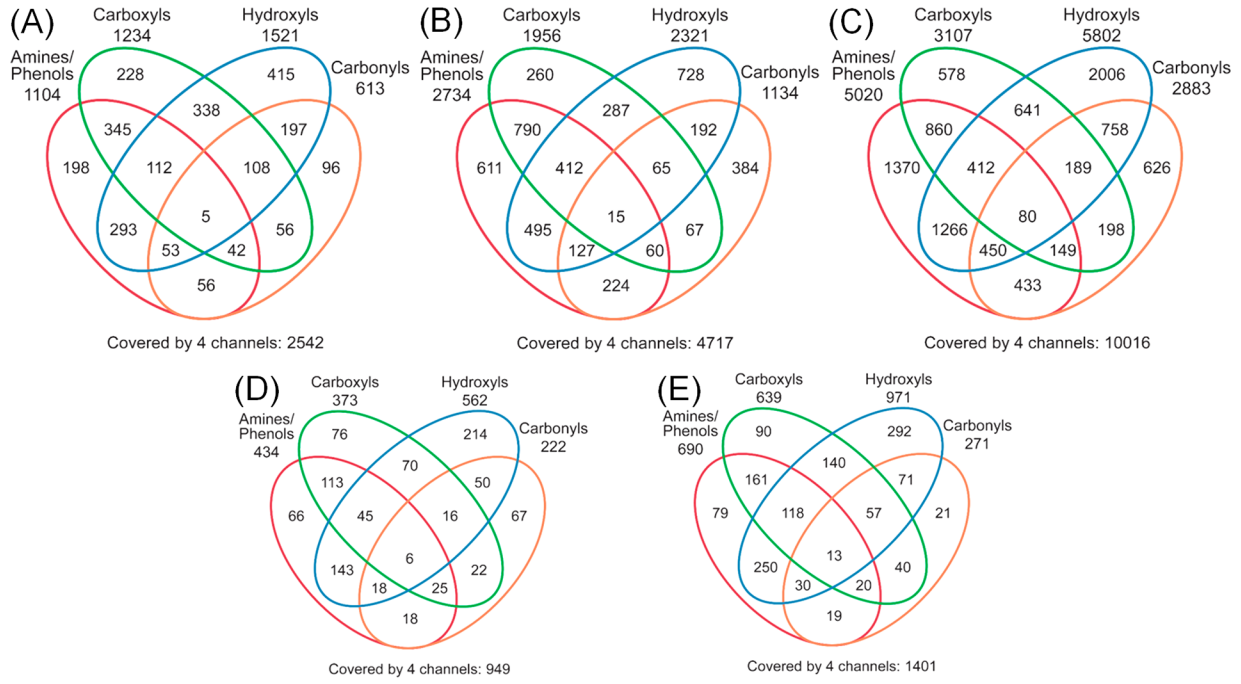


Figure 4. Venn diagram of the numbers of metabolites in four channels from the compound entries in (A) MCID, (B) HMDB, (C) KEGG, (D) YMDB, and (E) ECMDB.

high-confidence results (tier 1 and tier 2). In tier 3, the remaining peak pairs not identified in tiers 1 and 2 were mass-searched against the HMDB, MCID-1R, and MCID-2R libraries in sequence. There were 2628, 2851, and 777 peak pairs (35.4%, 38.4%, and 10.5%) matched to the three libraries, respectively. In total, 6926 peak pairs (93.2%) were either identified or matched to databases. The remaining 6.8% of detected pairs may belong to metabolites that are not included in any of the searched databases.

For yeast samples, a similar approach was applied for peak pair detection and metabolite identification using four-channel LC-MS. In total, we detected 5641 peak pairs, including 1747 from A-channel, 1867 from C-channel, 1006 from H-channel, and 1021 from K-channel (Supporting Information Figure S5). After filtering the same-mass metabolites detected in two or more channels, we have 4955 unique-mass peak pairs. Table 2 summarizes the identification results. From the 5641 peak pairs detected, 243 and 188 peak pairs were identified in tier 1 (Supporting Information Table S7) and tier 2 (Supporting Information Table S8), respectively. Thus, a total of 431 peak pairs (7.6%) can be identified as high-confidence results (tier 1 and tier 2). In tier 3, we found 880, 3442, and 514 peak pairs (15.6%, 61.0%, 9.1%) matched to YMDB, MCID-1R, and MCID-2R libraries, respectively (Supporting Information Tables S8 and S9). In total, 5267 peak pairs (93.3%) were either identified or matched to databases.

It should be noted that the peak intensity ratios measured in the triplicate analysis of 1:1 ^{12}C-/^{13}C-labeled plasma can be used to gauge the accuracy and precision for relative quantification of this particular mixture. When plotting the

DOI: 10.1021/acs.analchem.9b03431
Anal. Chem. 2019, 91, 12108−12115

distribution of peak pairs detected as a function of the average peak ratio and their relative standard deviation (RSD) (Supporting Information Figure S6), most of the peak pairs in four submetabolome profiling gave the ratio value close to the expected ratio of 1.0, demonstrating high accuracy. The RSD values are less than 20% for more than 95% of the pairs with an average RSD of 5.1% and thus the analytical precision was also very high. We note that we did not study the interday and intraday repeatability in this work. However, all the individual labeling methods have been used in a number of published metabolomics studies where quality control (QC) samples were used to gauge interday and intraday repeatability over a number of days. QC samples were clustered tightly, indicating excellent repeatability.[24] The linearity of peak ratio measurement has been addressed in previously reports such as the original paper published on dansylation labeling for amine/phenol submetabolome profiling.[3] Over 100-fold relative changes could be measured.[3]

Overlaps of Multifunctional Metabolites. Figure 4 shows the Venn diagrams of the numbers of database metabolites belonging to individual channels and overlaps among different channels. There are clearly many metabolites belonging to two or more channels. Taking the MCID database as an example (Figure 4A), there are five metabolites in all channels. Most of the overlaps occur for metabolites containing two functional groups. However, in our four-channel LC-MS results, most of the metabolites were uniquely detected in only one channel.

Much smaller overlaps found can be attributed to the fact that, by design, we developed the four-channel methods with due consideration of minimizing redundant analyses of the same metabolite in different channels. For example, in the analysis of the carboxyl acid submetabolome, we used 6 M HCl to acidify the sample, followed by organic solvent extraction of the acids. The amines and some phenols would be positively charged at this low pH and thus not be extracted by the organic solvent. This can be inferred from the analysis of amino acids; all 20 amino acids can be readily detected in the amine/phenol channel, but only 2−3 can be detected in the carboxylic acid channel. Similarly, for the analysis of the hydroxyl submetabolome, we used an organic solvent to extract the neutral metabolites containing hydroxyl groups from the highly acidified sample. In the analysis of the carbonyl submetabolome, the labeling solution is acidic under which dansylhydrazine preferentially reacts with neutral metabolites containing carbonyl groups. The charged species such as metabolites containing both amine and carbonyl groups may not react with dansylhydrazine. Another contributing factor might be related to the changed reactivity of a functional group in metabolites with two or more groups. For example, we found that several keto-acids with carbonyl and carboxyl groups conjugated together (e.g., oxaloacetic acid and acetoacetic acid) are difficult to be labeled in the carbonyl channel.

Group Under-representation. While the compound entries in a database are by no means perfect in terms of coverage (i.e., not including all metabolome compounds of an organism) and trueness (i.e., the presence of false entries), the group classification shown in Figure 2 for several databases gives consistent group distributions. For example, the hydroxyl group is by far the largest group, except in HMDB where it is the second largest. However, in our experimental data set, for both plasma and yeast samples, the number of hydroxyl-containing metabolites detected is smaller than the other groups. This suggests that the overall coverage achieved by the current four-channel experiments is lower than the database-derived theoretical coverage; the exact percentage of reduction is unknown. It appears that the sample preparation workflow or labeling reaction of the hydroxyl channel is not fully optimized. More development work should be devoted to optimize the hydroxyl submetabolome profiling.

CONCLUSIONS

After filtering out the lipids, inorganic species, and hydrocarbons that are not targeted for analysis by CIL LC-MS, we found that 86% to 96% of the metabolites in the studied databases contain one or more of the five functional groups: amine, phenol, hydroxyl, carbonyl, and carboxyl. Thus, in-depth profiling of these chemical groups can generate a very high coverage of the metabolome. We described a four-channel CIL LC-MS approach to analyze the hydroxyl (H), amine/phenol (A), carboxyl (C), and carbonyl (K) submetabolomes, separately.

For future work, we will need to optimize the current method for hydroxyl submetabolome profiling and develop labeling methods to analyze other groups of metabolites currently not covered by the four-channel approach (e.g., esters and amides). We note that the compound entries in a current database may under- or over-represent certain groups of metabolites. For example, there may be the intermediate compounds of known metabolites that have not been documented, as evident from the mass-matches of many predicted metabolites from one or two metabolic reactions of known metabolites (i.e., MCID-1R and MCID-2R). As our knowledge of metabolites expands with the detection and identification of known unknowns and unknown unknowns, we will surely increase the coverage and reduce the false entries in a metabolome database of an organism. We envisage that the four-channel LC-MS approach, with perhaps additional channels, will play an important role in expanding our knowledge of chemical composition of a metabolome.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.9b03431.

Supplemental note N1 for experimental methods, Figures S1−S6 for workflow, and plots of experimental results (PDF)
Table S1 (XLSX)
Table S2 (XLSX)
Table S3 (XLSX)
Table S4 (XLSX)
Table S5 (XLSX)
Table S6 (XLSX)
Table S7 (XLSX)
Table S8 (XLSX)
Table S9 (XLSX)
Table S10 (XLSX)

AUTHOR INFORMATION

Corresponding Author

*E-mail: Liang.Li@ualberta.ca.

DOI: 10.1021/acs.analchem.9b03431
Anal. Chem. 2019, 91, 12108−12115

ORCID

Liang Li: 0000-0002-9347-2108

Author Contributions

†Equal contribution.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada, Canada Research Chairs, Canada Foundation for Innovation, Genome Canada, and Alberta Innovates.

■ REFERENCES

(1) Khamis, M. M.; Adamko, D. J.; El-Aneed, A. *Mass Spectrom. Rev.* **2017**, *36*, 115−134.
(2) Vuckovic, D. *Chem. Commun.* **2018**, *54*, 6728−6749.
(3) Guo, K.; Li, L. *Anal. Chem.* **2009**, *81* (10), 3919−3932.
(4) Zhao, S.; Li, L. *Anal. Chem.* **2018**, *90* (22), 13514−13522.
(5) Hao, L.; Johnson, J.; Lietz, C. B.; Buchberger, A.; Frost, D.; Kao, W. J.; Li, L. J. *Anal. Chem.* **2017**, *89* (2), 1138−1146.
(6) Leng, J. P.; Wang, H. Y.; Zhang, L.; Zhang, J.; Wang, H.; Guo, Y. L. *Anal. Chim. Acta* **2013**, *758*, 114−121.
(7) Tayyari, F.; Gowda, G. A. N.; Gu, H. W.; Raftery, D. *Anal. Chem.* **2013**, *85* (18), 8715−8721.
(8) Wong, J. M. T.; Malec, P. A.; Mabrouk, O. S.; Ro, J.; Dus, M.; Kennedy, R. T. *J. Chromatogr. A* **2016**, *1446*, 78−90.
(9) Yuan, W.; Edwards, J. L.; Li, S. W. *Chem. Commun.* **2013**, *49* (94), 11080−11082.
(10) Chu, J. M.; Qi, C. B.; Huang, Y. Q.; Jiang, H. P.; Hao, Y. H.; Yuan, B. F.; Feng, Y. Q. *Anal. Chem.* **2015**, *87* (14), 7364−7372.
(11) Huan, T.; Li, L. *Anal. Chem.* **2015**, *87* (14), 7011−7016.
(12) Zhao, S.; Luo, X.; Li, L. *Anal. Chem.* **2016**, *88* (21), 10617−10623.
(13) Guo, K.; Li, L. *Anal. Chem.* **2010**, *82* (21), 8789−8793.
(14) Li, L.; Li, R.; Zhou, J.; Zuniga, A.; Stanislaus, A. E.; Wu, Y.; Huan, T.; Zheng, J.; Shi, Y.; Wishart, D. S.; et al. *Anal. Chem.* **2013**, *85* (6), 3401−3408.
(15) Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Rosa, V.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; et al. *Nucleic Acids Res.* **2018**, *46*, 608−617.
(16) Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-kinoshita, K. F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M. *Nucleic Acids Res.* **2006**, *34*, 354−357.
(17) Jewison, T.; Knox, C.; Neveu, V.; Djoumbou, Y.; Guo, A. C.; Lee, J.; Liu, P.; Mandal, R.; Krishnamurthy, R.; Sinelnikov, I.; et al. *Nucleic Acids Res.* **2012**, *40*, 815−820.
(18) 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.* **2012**, *41*, 625−630.
(19) Wu, Y.; Li, L. *Anal. Chem.* **2012**, *84* (24), 10723−10731.
(20) Zhou, R.; Tseng, C.-L.; Huan, T.; Li, L. *Anal. Chem.* **2014**, *86* (10), 4675−4679.
(21) Huan, T.; Li, L. *Anal. Chem.* **2015**, *87* (2), 1306−1313.
(22) Huan, T.; Wu, Y.; Tang, C.; Lin, G.; Li, L. *Anal. Chem.* **2015**, *87* (19), 9838−9845.
(23) Huan, T.; Tang, C.; Li, R.; Shi, Y.; Lin, G.; Li, L. *Anal. Chem.* **2015**, *87* (20), 10619−10626.
(24) Han, W.; Sapkota, S.; Camicioli, R.; Dixon, R. A.; Li, L. *Mov. Disord.* **2017**, *32* (12), 1720−1728.

DOI: 10.1021/acs.analchem.9b03431
Anal. Chem. 2019, 91, 12108−12115