Published in final edited form as:

Rapid Commun Mass Spectrom. 2013 September 30; 27(18): 2091–2098. doi:10.1002/rcm.6656.

Liquid Chromatography/Mass Spectrometry Methods for Measuring Dipeptide Abundance in Non-Small Cell Lung Cancer

Manhong Wu¹, Yue Xu², William L. Fitch¹, Ming Zheng¹, Robert E. Merritt², Joseph B. Shrager^{2,3}, Weiruo Zhang⁴, David L. Dill⁴, Gary Peltz¹, and Chuong D. Hoang^{2,3}

¹Department of Anesthesia, Stanford University School of Medicine

²Division of Thoracic Surgery, Department of Cardiothoracic Surgery, Stanford University School of Medicine

³Section of Thoracic Surgery, Veterans Affairs Palo Alto Health Care System

⁴Department of Computer Science, Stanford University School of Engineering

Abstract

RATIONALE—Metabolomic profiling is a promising methodology of identifying candidate biomarkers for disease detection and monitoring. Although lung cancer is among the leading causes of cancer-related mortality worldwide, the lung tumor metabolome has not been fully characterized.

METHODS—We utilized a targeted metabolomic approach to analyze discrete groups of related metabolites. We adopted a dansyl [5-(dimethylamino)-1-naphthalene sulfonamide] derivatization with liquid chromatography/ mass spectrometry (LC/MS) to analyze changes of metabolites from paired tumor and normal lung tissues. Identification of dansylated dipeptides was confirmed with synthetic standards. A systematic analysis of retention times (RT) was required to reliably identify isobaric dipeptides. We validated our findings in a separate sample cohort.

RESULTS—We produced a database of the LC retention time and MS/MS spectra of 361 dansyl dipeptides. Interpretation of the spectra is presented. Using this standard data, we identified a total of 279 dipeptides in lung tumor tissue. The abundance of 90 dipeptides was selectively increased in lung tumor tissue compared to normal tissue. In a second set of validation tissues, 12 dipeptides were selectively increased.

CONCLUSIONS—A systematic evaluation of certain metabolite classes in lung tumors may identify promising disease-specific metabolites. Our database of all possible dipeptides will facilitate ongoing translational applications of metabolomic profiling as it relates to lung cancer.

Keywords

metabolomics; lung cancer; dipeptides

INTRODUCTION

Lung cancer continues to be the number one cause of cancer death in the U.S. accounting for about 165,000 deaths in 2012 ^[1]. Worldwide, estimated lung cancer deaths affect over one million peoples ^[2]. Nearly 80% of all lung cancer is non-small cell lung carcinoma

Address correspondence to: gpeltz@stanford.edu or cdhoang@stanford.edu.

There are no conflicts of interest

(NSCLC). The ongoing and current low survival statistics of those afflicted with NSCLC are partially attributed to diagnosis at advanced stages. Despite extensive searches, there are no available biomarkers that lead to earlier detection, and thus improved clinical outcome, for NSCLC in routine use [3, 4]. While there are ongoing efforts to characterize the metabolomes unique in different types of human cancers, only a few metabolomic studies have examined lung cancer.

Early studies of lung cancer metabolomics indicated numerous metabolite species were differentially expressed in tumor specimens. Lung cancer cell lines (3 specimens) with differing chemotherapy resistance showed significant changes in cholines, lipids, lactate, and glycine among other metabolites [5]. Another study profiled ipsilateral bronchioloalveolar carcinoma lesions (2 specimens) after erlotinib treatment and resection [6]. A metabolomic profile based on characteristic differences in amino acids, phosphocholines, and Krebs cycle metabolites correlated with erlotinib sensitivity and positron emission tomography activity. More recently, a study used gas chromatography mass spectrometry to evaluate a heterogeneous mix of serum and tissue samples (7 lung cancer patients) in both non-small cell lung carcinoma and small cell lung carcinoma [7]. The authors identified 48 metabolites significantly changed in lung tumors and that a subset of metabolites had characteristic alterations in histologic subtype and disease stage of tumor. Another metabolomic study of both lung and prostate tumors confirmed hyperactive glycolysis in lung specimens [8]. They highlighted the potential of capillary electrophoresis time-of-flight mass spectrometry-based metabolomics combined with phosphorylated enzyme analysis for understanding tissue-specific tumor microenvironments. Clearly, there is a need for methods development addressing the technical constraints that limit rapid progress in this area of lung cancer care and diagnostics. We present here our exploratory work that could facilitate this process.

Because the extreme differences in physicochemical properties make it impossible to accurately measure changes in all cellular metabolites using a single analytic method, we chose to utilize methods that reliably analyze discrete groups of chemically related metabolites. We refined a recently developed dansyl derivatization method [9] with LC/MS analysis to analyze changes in a large number of amine or phenol containing metabolites in tumor and normal lung tissues. Dansylation increases metabolite detection sensitivity by 10-1000 fold, and improves polar metabolite retention and separation on reversed phase columns. Comparing tumor and normal tissues from the same individual eliminates confounding inter-individual differences, and increases the accuracy for identifying tumor-related changes.

Interestingly, when we compared our results of significantly altered metabolites between tumor and normal lung specimens, among global LC/MS and targeted, dansylated LC/MS techniques, we observed consistent, pronounced changes of dipeptides. The precise role of dipeptides in normal physiology is still being investigated. In a eukaryotic model organism, dipeptides are important for protein nutrition, growth, and development [10]. Dipeptide levels are regulated by membrane proteins of the peptide transporter superfamily [11]. The peptide transporters are widely distributed among organs such as brain, pancreas, intestines, kidney, lymphatics, lung, mammary gland and spleen [12]. The role and significance of dipeptide(s) alterations in malignancy are unknown. Since the human metabolome database contained very few validated dipeptide spectra, we recognized the utility of developing a more complete catalogue of dipeptides that may be identified in human specimens. Therefore, our initial effort has focused on characterizing the chromatographic properties of all possible dipeptides as a reference database to facilitate ongoing metabolomic investigations.

EXPERIMENTAL

Chemicals

Individual amino acids, an amino acid standard mixture LAA21, dansyl chloride and sodium decahydrate were purchased from Sigma-Aldrich (St. Louis, Missouri). Dipeptides were purchased from JPT Peptide Technologies GmbH (Berlin, Germany), and dipeptides containing a proline at the C-terminus were obtained from AnaSpec Inc. (Fremont, California). HPLC-grade acetonitrile and water were from Honeywell Burdick and Jackson (Morristown, New Jersey). LC/MS-grade formic acid was obtained from ProteoChem, Inc. (Denver, Colorado).

Human specimens

All tissues for this study were obtained from the Stanford University Hospital Tissue Bank. Informed patient consents were obtained as part of approved IRB protocols. Appropriate institutional review occurred before all IRB protocols were approved. Nine males with NSCLC were included in this study as the testing set. To minimize confounding factors, we limited our analysis to the most common histologic type of NSCLC, i.e. adenocarcinoma. Each specimen was analyzed by a board-certified pathologist and was diagnosed as an adenocarcinoma. Three spatially distinct samples were obtained from each individual: 1) a cross-sectional slice of tissue through a bisected plane that includes the tumor; 2) grossly normal appearing lung tissue immediately adjacent to the tumor; and 3) a portion of normal appearing lung located at least >3 cm away from the primary tumor (Figure S1 in Supplement section). For the testing set, we included the adjacent-to-tumor as an added control sample to maximize our ability to identify tumor-specific metabolites. The samples were snap frozen within 30 minutes to one hour after dissection, and were kept in -80°C before processing. For the validation set, a separate, distinct group of 10 males diagnosed with adenocarcinoma were chosen and the same procedures followed except that adjacent tissue was not analyzed. An adjacent tissue specimen was not needed in the validations because we were performing a targeted metabolomic analysis.

Similarly, paired liver cancer and normal liver tissues were obtained from six different individuals that were included in this study to demonstrate specificity of our metabolomic findings in NSCLC. A pathologist reviewed the tissues, and all were diagnosed as hepatocellular carcinoma. The tumor and normal liver tissues were snap frozen, and stored at -80°C before processing. Metabolomic analysis of all tissues was conducted under a separate approved IRB protocol. All patients enrolled into this study were selected based solely on having tissue amounts (lung and liver specimens) sufficient for metabolomic analysis.

Tissue extraction and metabolite dansylation

Frozen lung tissue samples were weighed and homogenized with stainless steel beads by Qiagen TissueLyser II (Valencia, California), at a frequency of 20/second for 30 seconds - the homogenization was repeated four times with a standard solvent mixture^[13]. The upper layer of the chloroform:methanol:water mixture, representing the polar metabolites extracted from the tissues, were removed and dried in a Speedvac. The dried extract was re-suspended in water. Fifty μL of polar metabolite extract, 25 μL volume of 0.1M sodium tetraborate buffer and 50 μL of 20 mM dansyl chloride (in acetonitrile) were combined and vortexed as described^[9]. The mixture was incubated at room temperature for 30 minutes before 50 μL of 0.5% formic acid were added to stop the reaction. The supernatant of the reaction mixture was removed into an autosampler vial.

LC/MS analysis

The samples were randomized into a single batch for LC/MS analysis. All samples were analyzed on an Agilent (Santa Clara, California) accurate mass quadrupole time-of-flight (QTOF) 6520 coupled with an Agilent UHPLC Infinity 1290 system. Chromatography method A was done on a Zorbax C18 column (0.5 × 150 mm). The flow rate was 5 μ L/min with a gradient from 5% solvent B to 95% B in 30 minutes and held at 95% B for 5 minutes. Solvent A was HPLC water with 0.1% formic acid and Solvent B was LC/MS-grade acetonitrile with 0.1% formic acid. All data were acquired by positive electrospray ionization with Agilent MassHunter acquisition software. The heated capillary temperature in the source was held at 325°C. Full scan (mass-to-charge (m/z) 110–1000) spectra were collected. Chromatography method B was run on a Phenomenex Kinetex (Torrance, California) C18 column (2.1 × 100 mm, 2.6 μ m particles, 100Å pore size). The 30 minutes gradient at 0.5 mL/min was as follows: t = 0.5 minute, 5% B; t = 20.5 minutes, 60% B; t = 25 minutes, 95% B; t = 30 minutes, 95% B (using the same mobile phases as Method A). The column was re-equilibrated at 5% B for 5 minutes.. A chromatographic Method C was briefly evaluated using the Kinetex column using a gradient at 0.5 mL/min as follows: t = 0.5 minute, 10% B; t = 2.5 minutes, 20% B; t = 17.5 minutes, 60% B; t = 18 minutes, 95% B (using the same mobile phases as Method A). The column was re-equilibrated at 5% B for 5 minutes.

MS/MS analysis

Subsequent runs were performed to collect MS/MS spectra for each compound. Data-dependent and targeted MS/MS analyses were also acquired on QTOF 6520 mass spectrometer with the same HPLC gradient as above. The injection volume was 2–4 μ L. The acetonitrile in the injection solution was required to maintain the hydrophobic components in solution and its inclusion did not affect the early eluting peaks. For targeted MS/MS, a retention time (RT) window was set at 0.6 minutes. The collision energy was set at 26V, isolation width 4 m/z, MS acquisition rate at 5 spectra/second and MS/MS acquisition rate at 3 spectra/second. The collision energies for the highest molecular weight derivatives like HH and YY were set at 32V.

MS Data analysis

Agilent MassHunter Qualitative software was utilized. For untargeted data analysis, the Molecular Feature Extractor was utilized to find features in the 27 raw data files. Extracted peaks were retention time aligned using Mass Profiler Professional and unique features detected by least squares analysis. The Agilent version of the Metlin database [14] was utilized to identify results from untargeted metabolomic assessment. For targeted dipeptide analyses, the dipeptide database and the Agilent MassHunter “Find By Molecular Formula” feature were utilized. Parameters were set to use the MS/MS integrator and a mass error of 10 ppm.

Quality control

For analysis of each set of samples, a standard mixture of 26 amino acids was derivatized and analyzed at the beginning and again at the end of the run. This quality control step was performed to ensure that the reaction worked, that the instrument performed well, and that consistent RTs were obtained. We found that the standards had a reproducible RT (within 6 seconds) between runs.

Statistical analysis

Since three different tissues were obtained from the same individual in the testing set, a paired, two-sided Wilcoxon's rank sum test [15] was applied to assess the significance of the

abundances in tumor tissues compared to those in normal tissues (adjacent-to-tumor lung and normal lung). Since only nine individuals were included in this initial portion of our study and the non-parametric Wilcoxon's test usually has low power, none of the dipeptides would have significant p value after multiple-testing adjustment. Nevertheless, the significance of the differential abundance for all dipeptides pooled together can still be assessed using a binomial test. Under the null hypothesis that none of the dipeptides is differentially abundant, each of them has a 0.05 probability to have a raw p value that is < 0.05 . Therefore, the p value for the situation where we observed 105 dipeptides having a raw p value that is < 0.05 out of 334 dipeptides investigated is calculated from a binomial distribution. Since the lung tumor and normal tissues were from the same patient, an analysis of the necessary sample size to detect differential metabolites with significance level set at 5% and power at 80% yields an estimate of nine to 15 sets of matched human specimens (refer to Supplement). Sample size calculation was performed using SAS v9.3 (Cary, NC).

RESULTS/ DISCUSSION

We profiled metabolomic changes in tumor, adjacent-to-tumor, and normal tissues obtained from nine individuals with NSCLC. Tissue was extracted and dansylated to yield samples amenable to LC/MS analysis of amine and phenolic metabolites. Initial metabolomic comparison revealed three features with significant differences between normal and tumor. The m/z values corresponded to neutrals of monoisotopic molecular weight 822.1212, 452.1865 and 422.1752. MS/MS analysis on a pooled tumor sample showed the m/z 823 ion to be an in-source adduct associated with dansyl-glycerophosphoethanolamine. After removal of the dansyl mass (233.0510), a Metlin database search indicated the others might be dansyl-dipeptides. MS/MS supported that the structures were dansyl-VT and dansyl-GL. Analysis of the full scan MS and MS/MS data indicated that more than one dipeptide isomer was present for each mass. Based on this observation, it appeared that many dipeptides might be tumor related. Therefore, we characterized the chromatographic behavior of 360 dipeptide standards and performed a targeted re-analysis of this metabolomic dataset. Where necessary, other results including figures and tables are located in the Supplement section.

Generation of the dipeptide database

The dipeptides were produced by array based combinatorial synthesis at JPT Peptide Technologies GmbH, and the proline-containing (XP) dipeptides, which could not be synthesized by this method, were obtained from AnaSpec Inc. Cysteine dipeptides were not included in this study so 361 individual dipeptides are reported in this paper. Aliquots containing 7 to 30 nanomoles of each of the 18 dipeptides with a common C-terminus were combined and diluted to 10 μ M in 50% acetonitrile and water. Each standard mixture was derivatized by dansylation and analyzed by LC method B with full scan MS. Our mass spectra of dansyl dipeptides were typically dominated by singly protonated ions. Minimal sodium adducts or dimeric species were observed under these conditions. Basic side chains led to increased abundance of doubly or triply charged molecular species. Subsequent runs were performed to collect MS/MS spectra for each dipeptide. For spectral alignment we developed a method to adjust for the small but significant variability in RT between different LC/MS runs. Since amino acids are universal analytes in biological samples, we used amino acids as standards for inter-run comparisons. The RT of the amino acids and dipeptide standards were highly correlated; polynomial alignment of the amino acid retention times was the best method for aligning the dipeptide RT. The full data set of aligned RT, formulae, number of dansyl groups and the tissue abundances of these dipeptides in the initial 27 lung tumor, adjacent and normal tissues are in Table S1 (located in the Supplement section due to its size).

Identification of dipeptides in lung tissue

Of the 361 dipeptides, only 161 have unique chemical formulas. Only the 17 homo-dipeptides are single isomers (e.g. AA); while 118 formulas describe two different dipeptides (e.g. YA vs AY), 21 formulas describe four different dipeptides (e.g. AD, DA, GE, EG) and four formulas describe six dipeptides (e.g. LG, IG, VA, AV, GL, GI). With the full set of RT available to us, it became apparent that several would not be identifiable by full scan MS alone, so that MS/MS analysis would be required. As detailed below, MS/MS is not always able to distinguish these isomeric dansylated dipeptides. And neither data dependent or mass-targeted MS/MS experiments could collect product ion spectra for multiple low level analytes. Using a cutoff of 0.1 min, there are 22 dipeptide masses that are associated with more than one dipeptide. Figure 1 illustrates the difficulties in evaluating these isomers. The six isomeric dipeptides are resolvable, but the identification of several pairs of these dipeptides (IG vs AV; GI vs GL) will require careful alignment of their retention characteristics. Manual reanalysis of a pooled tumor sample using Method B indicated 279 dipeptides were present. Of these, 234 had mass accuracies within 5 ppm of predicted, and 45 others had mass accuracy error of < 10 ppm. Figure S2 shows the RT differences for these 279 identifications versus their predicted RT from the dipeptide database. Most are within 0.1 min but arginine-containing dipeptides behave differently in standards versus the tumor sample, possibly due to ion pairing in a complex biological extract on C18 columns [16]. A special case is diproline. A standard for RT comparison was not available but the predicted mass for dansyl-PP was detected at the appropriate RT in the pooled tumor sample and was designated as real and included in the RT database.

MS/MS dipeptide spectra

The collision induced dissociation MS/MS spectra of dansylated dipeptides are often uninformative, since they are dominated by dansyl ions (Figure 2). The $C_{12}H_{12}N$ cation formed from dansyl cleavage produces the base peak in most dansyl-dipeptide spectra with an average m/z value of 170.0964 ($\sigma = 0.0011$, $n = 57$ spectra, accuracy of 3.3 ppm), and 95% of measured values would fall within 0.022 mDa or 13 ppm. Using this number to identify the formulae of fragment ions, all reported structures were within 13 ppm of the correct value. The related radical species $C_{12}H_{13}N$ (m/z 171.105) is typically present at 30-100% of the m/z 170 ion. Other dansyl specific ions observed include: m/z 234.059 ($C_{12}H_{12}NSO_2$), which is universally present in product ion spectra of singly charged ions but at highly variable relative intensity; m/z 252.069 ($C_{12}H_{14}NSO_3$) is commonly present in oxygen containing dipeptides; and a minor m/z 186.092 ($C_{12}H_{12}NO$) ion is also common. Figure 3 shows the cleavage sites for dansylated peptides, and the diagnostic fragments are summarized in Table 2. Other than these species, the informative ions are the a ions and derived fragments from residue specific neutral losses, and the I ions. Only rarely are the classic b, c, x, y or z ions observed. The d cleavage (loss of H_2O+CO) is specific to W. All 360 MS/MS spectra along with raw uncorrected RT for the dipeptides and the amino acids in each standard run is available as online supplement.

There are several other noteworthy features in the MS/MS spectra. Glycine dipeptides are undetectable in MS/MS, no G specific ions are readily observed. AX dipeptides all show the dansyl-imine species a at m/z 277.101. XA dipeptides do not reveal any A related fragment ions. Hydrophobic amino acid containing dipeptides also have a few special features: I, L, V, and F containing dipeptides very commonly display the classic I ions (when in either position 1 or 2) and the dansyl ions a when in position 1, but are suppressed in some dipeptides where ($aa_1=H,K,P$ or R^{+1}). P shows especially strong I and a ions and also shows a strong y ion at m/z 116 when in position 2. Tryptophan often shows the I ion sometimes accompanied by neutral losses to m/z 144 and 130. The presence of multiple dansyl groups in tyrosine make these compounds quite hydrophobic, which often present as doubly

charged ions and share many of the properties of the basic amino acids with strong a, I1 and I2 signals. The polar, non-basic residues are difficult to detect in MS/MS; especially if a preferred ionization partner (K,R,H,Y,P,F,I,L) is present. E,D,S and T are often observed with an I1 or I2, rarely with an a ion (often these are present as the loss of water ion from I or a). The N₁ residue often gives rise to an m/z 261 ion (C₁₃H₁₃N₂SO₂) via neutral loss of ammonia and ketene from the a ion. XN gives a strong ammonia loss from the molecular species when paired with K,H R or Y. XQ uniquely gives rise to a c ion through neutral loss of pyroglutamate. None of the polar aa are detectable in the spectra of doubly charged dansyl-dipeptides. Dipeptides with basic amino acids (R,K and H) have complex behavior. The R dipeptides are highly doubly charged and dominate the MS/MS spectral features. RX is indicated by a strong m/z 408 ion equivalent to dansylated-R, especially in singly charged spectra. Both RX and XR give rise to the m/z 156/157 cluster (C₁₁H₉N,C₁₁H₁₀N) in doubly charged spectra. The lysine dipeptides are also highly doubly charged and dominate with m/z 317 (a-C₁₂H₁₄N₂SO₂) and classic m/z 84 fragments. Similarly, the fully dansyl-protected histidine dipeptides dominate with doubly charged spectra and strong I1 and I2 species. HX doubly charged dipeptides often shown the same m/z 261 ion that is present in NX and MX compounds. A few of the HX and KX dipeptides showed a loss of dansylation from the molecular species. The KX, HX and RX dipeptides show stable molecular species that show strong molecular species and strong neutral losses from the X side chain. Methionine dipeptides have very complex spectra with a ions at m/z 337 further fragmenting to lose CH₃SH (289), CH₃SH+NH₃+SO₂ (208) and CH₃SH+C₂H₄ (261) along with the classic I ion at m/z 102 and I-CH₃SH ion at m/z 56.

Chemistry of the dansylation

Variable levels of derivatization or unstable derivatives can alter the results produced by any derivatization method. While the dansylation method is robust, we identified a few cases for special consideration. For example, histidine dansylation produces two separate LC peaks. The later eluting peak [ref. ^[9]] is due to the bis-dansyl derivative, which has singly charged (m/z 622.1789) and doubly charged (m/z 311.5931) ions, and it generates a strong in-source fragment at m/z 389.1278 due to loss of one dansyl group. We saw no evidence for loss of this dansyl group during chromatography. The earlier eluting monodansyl-histidine has strong singly charged (m/z 389.1278) and doubly charged (m/z 195.0675) ions; which is not observed in the full scan mass spectrum of bis-dansyl HIS. The bis-dansyl derivative slowly decomposes in aged samples, but we have not determined how the conversion takes place or whether it can be eliminated. Both histidine derivatives were used for RT normalization. For standard HX and XH dipeptides, the fully dansylated derivatives were dominant. For HH, the ratio of dansyl₃/dansyl₂/dansyl₁ was 9/4.5/0.2. After comparing our results with the variation in reported RT and masses ^[17], it is evident that histamine, carnosine and other imidazoles have the same issues. Purines also form multiple monodansylated derivatives, which we suspect is due to unique derivatizations of the tautomeric forms. Similarly, the Alberta database ^[9, 17] also indicates that there are variable levels of dansylation of dihydroxybenzenes, and dopamine is reported as a monodansyl derivative.

Dipeptide identification using different analytical conditions

To identify dipeptides in the 27 lung samples that were analyzed with LC method A, an algorithm was developed that could analyze metabolomic data obtained under different chromatographic conditions. Since amino acids are universal analytes in all biological samples, we investigated whether amino acids could be used as internal standards to enable comparison of data generated in different laboratories. RT of the amino acids and dipeptide standards were highly correlated. After testing multiple methods, we found that locally estimated scatter plot smoothing (LOESS) method ^[18] was the best for aligning dipeptide RT. Using this information, we developed a method to analyze metabolomic data produced

in different laboratories using different LC systems. The relative retention order for amino acids remained constant across three different LC columns and gradient systems (Figure S3). The LOESS method was used to predict the RT for dipeptides obtained under one condition using the amino acid data and the dipeptide RT measured with another method. The dipeptide RT predicted by this method agree well with manually identified dipeptide RT, with the exception being the dipeptides that had at least one basic arginine residue (Figure S4). Since the LOESS method and amino acid functionality cannot completely correct for the chromatographic differences between these two datasets, an adjustment function was derived using the data in Figure S4, which produced correction factors that accurately predicted the dipeptide RT in the other 26 tumor samples. Figure S5 shows the residuals for another tumor sample determined using these adjustment factors. Thus, when comparing data across LC platforms it is necessary to adjust both for minor changes in RT using a set of common standards, and to identify as many analytes as possible to correct for complex chemical interactions (i.e. arginine dipeptides).

Targeted analysis of dipeptide abundance in lung tissue

Each of the original 27 lung tissue data files was queried for the presence of 21 dansylated amino acid standards, and this information was used to enable alignment of the RT for each of the 361 dipeptides using the Agilent Find by Formula targeted workflow. A total of 257 dipeptides were detected in at least one tissue extract, and each of the dipeptide identifications was based on a mass accuracy (< 5 ppm mass error in calculation of neutral species) and RT with a minimal variance from that predicted by the LOESS method (< 0.2 min). Then, least squares analysis was performed to identify dipeptides whose abundance was significantly different in the tumor samples. The dipeptides in the validation set of 20 samples were determined in the same way after LC/MS using method B. The results are summarized in Table 1, while all the results for the initial 27 samples are shown in Table S1 (dipeptide database) and the validation results are in Table S2 (both of these Tables are in Supplement section).

While 71 dipeptides were not detected in any of the 27 tissues (tumor, adjacent-to-tumor, and normal tissues from nine individuals) examined, 257 dipeptides were detected in at least one sample (Table S1). Notably, of the 253 dipeptides detected in tumor tissues, the abundance of 90 dipeptides was significantly (fold change > 2 and $p < 0.05$) increased in the tumor tissues relative to normal tissues. Again, the results are summarized in Table 1. The fold increases in dipeptide abundance in tumor tissues ranged from 2.1 to 23.1, with 58 and 11 dipeptides having greater than 5- or 10-fold increased abundances, respectively, in tumor tissues. Given the small sample size and the lower power of the non-parametric test for comparing samples obtained from the same individual, the abundance increase in any individual dipeptide could not reach metabolome-wide significance after a statistical adjustment for multiple testing. However, when the results from all of the dipeptides are examined, the overall measured increase in dipeptide abundance in tumor tissues has high statistical significance. Given the null hypothesis that none of the dipeptides are differentially abundant, only 16.7 (334×0.05) dipeptides would be expected to have a raw p value < 0.05 by chance. However, we observed that the abundance of 90 dipeptides was increased in tumor tissues, and each had a raw p value of < 0.05 . The combined p value for observing this number of dipeptides with a statistically significant increased abundance in tumor tissues is 2×10^{-53} .

Confirmation of dipeptide data

In a validation cohort with 10 paired tumor-normal samples, we compared the measured abundance of dipeptides in tumor relative to normal tissue using a one-sided Wilcoxon's test (Table S2). Adjacent-to-tumor tissue was not collected nor analyzed in this cohort as we

recognized that it offered little distinction from normal lung in the first cohort; as we stated earlier, this specimen for the testing set was included initially as a second negative control. Overall, the testing and validation tumors were similar from a clinical perspective (Table S3). The abundance of 145 dipeptides could be quantified in at least one sample, and the measured abundances of 22 dipeptides were significantly higher in tumor tissues (raw p value < 0.05). If none of the dipeptides actually had an increased abundance in tumor tissues, we would expect only seven dipeptides to have an increased abundance by chance (p value < 0.05); and the p value for observing an increase in 22 dipeptides (with a raw p value < 0.05) out of the 145 measured dipeptides is only 8×10^{-10} . Even when we assume that the abundance of the remaining 192 dipeptides that were not detected in any sample was not increased, the corresponding p value for observing an increase in 22 dipeptides is still 0.0058. By either method, these results indicate that the overall abundance of dipeptides was also significantly increased in tumor tissues in the validation cohort.

In addition to the statistical assessment, several other analyses indicated that the increased dipeptide abundance in NSCLC was selective, and was not due to differences in cell density, nor did a technical artifact cause it. First, of the 179 dipeptides detected in the adjacent-to-tumor tissues, the abundance of only three was significantly (fold-change >2 and p<0.05) increased relative to the corresponding normal tissues. We applied these significance criteria in selecting metabolites to optimize the balance in obtaining a stringent p value while identifying the highest number of biologically relevant metabolites. Second, when the levels of 19 amino acids in tumor and corresponding normal tissues were compared, there were only minimal increases in the abundance of any amino acid in tumor tissue (range 1.1 to 2.5), and none were significantly increased (fold change > 2, and p value <0.05) in tumor tissues (Table S1). Third, we did not observe a significant change in dipeptide abundance when hepatocellular carcinoma and normal liver tissue samples obtained from six other individuals were similarly analyzed. Only four dipeptides were differentially abundant in liver carcinomas relative to the corresponding normal tissues with a raw p value < 0.05. Moreover, the abundance of one of these four dipeptides was increased (1.6-fold) in tumor tissues (Table S4). Under the null hypothesis that none of the dipeptides are differentially abundant, the p value for obtaining this result is 0.9999, which indicates that these differences are clearly due to chance.

Literature comparison

It is interesting to note that the many significant dipeptides were not detected in our initial untargeted metabolomic approach. We have found that, in general, untargeted data processing methods miss many metabolites because of overlapping peaks and the complexity of electrospray spectra (insource fragment ions, adducts, multiple charge states, etc.). But these methods often yield lead structures which are then better pursued with targeted analytical methods as described herein. More broadly, dipeptides in tissues have very rarely been examined [19-21]; and few previous studies have compared dipeptide abundance in normal and diseased human states. In fact, there are only 15 (protein-derived) dipeptides in the human metabolome database [22]. However, our study indicates that at least 260 more dipeptides are present in human lung tissue alone. Dansylation improves the sensitivity and resolving power for analysis of this important class of metabolites. This method, along with the database of dipeptide RT and other properties detailed here, should facilitate ongoing metabolomic studies of this class of biomolecules. For now, the biologic significance of increased dipeptide levels in human lung adenocarcinomas remains unknown and could not be addressed by the design of our study.

Because the dansylated derivatives do not have fully informative MS/MS behavior, a detailed analysis of cross platform retention time was undertaken. Since amino acids are universal analytes in all biological samples, amino acids could be used as internal standards

to enable comparison of data generated in different laboratories. We found that the RT of the amino acids and dipeptide standards were highly correlated. The RT of dansylated metabolites measured under multiple chromatographic conditions, and in two different laboratories, could be correlated when the amino acid RT were used for normalization. The 361-dansylated dipeptides characterized here, combined with the 220 dansylated metabolites characterized by Guo and Li ^[9], produce a database of 581 dansylated metabolites. This approach could enable subsequent metabolomic analyses to be performed without the costly and time-consuming requirement of analyzing reference standards for each experiment, and also enable comparison of metabolomic data produced in different laboratories.

Other results

Further description of experimental methods, results and discussions not detailed here are in the Supplement section.

CONCLUSIONS

We show the first systematic evaluation of dipeptides in lung tumors that may lead to identification of promising disease-specific metabolites. Our novel database of all possible dipeptides will facilitate ongoing translational applications of metabolomic profiling as it relates to lung cancer.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Bob Lewis for helpful discussions. G.P. was partially supported by funding from a transformative RO1 award (1R01DK090992) from the NIDDK. C.H. was supported by funding from the Bonnie J. Addario Lung Cancer Foundation.

Abbreviations

dansyl	5-(dimethylamino)-1-naphthalene sulfonamide
LC/MS	liquid chromatography coupled with mass spectroscopy
IRB	institutional review board
QTOF	quadrupole time-of-flight
HPLC	high performance liquid chromatography
m/z	mass-to-charge ratio
NSCLC	non-small cell lung cancer
RT	retention time

REFERENCES

1. Siegel R, Naishadham D, Jemal A. CA Cancer J Clin. 2012; 62:10. [PubMed: 22237781]
2. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. CA Cancer J Clin. 2011; 61:69. [PubMed: 21296855]
3. Sung HJ, Cho JY. BMB Rep. 2008; 41:615. [PubMed: 18823584]
4. Hassanein M, Callison JC, Callaway-Lane C, Aldrich MC, Grogan EL, Massion PP. Cancer Prev Res (Phila). 2012; 5:992. [PubMed: 22689914]

5. Gottschalk M, Ivanova G, Collins DM, Eustace A, O'Connor R, Brougham DF. NMR Biomed. 2008; 21:809. [PubMed: 18470962]
6. Fan TW, Lane AN, Higashi RM, Bousamra M 2nd, Kloecker G, Miller DM. Exp Mol Pathol. 2009
7. Hori S, Nishiumi S, Kobayashi K, Shinohara M, Hatakeyama Y, Kotani Y, Hatano N, Maniwa Y, Nishio W, Bamba T, Fukusaki E, Azuma T, Takenawa T, Nishimura Y, Yoshida M. Lung Cancer. 2011; 74:284. [PubMed: 21411176]
8. Kami K, Fujimori T, Sato H, Sato M, Yamamoto H, Ohashi Y, Sugiyama N, Ishihama Y, Onozuka H, Ochiai A, Esumi H, Soga T, Tomita M. Metabolomics. 2013; 9:444. [PubMed: 23543897]
9. Guo K, Li L. Anal Chem. 2009; 81:3919. [PubMed: 19309105]
10. Meissner B, Boll M, Daniel H, Baumeister R. J Biol Chem. 2004; 279:36739. [PubMed: 15155758]
11. Steiner HY, Naider F, Becker JM. Mol Microbiol. 1995; 16:825. [PubMed: 7476181]
12. Rubio-Aliaga I, Daniel H. Trends Pharmacol Sci. 2002; 23:434. [PubMed: 12237156]
13. Bligh EG, Dyer WJ. Can J Biochem Physiol. 1959; 37:911. [PubMed: 13671378]
14. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G. Ther Drug Monit. 2005; 27:747. [PubMed: 16404815]
15. Wilcoxon F. Biometrics Bulletin. 1945; 1:80.
16. Mant CT, Cepeniene D, Hodges RS. J Sep Sci. 2010; 33:3005. [PubMed: 21038458]
17. Guo K, Bamforth F, Li L. J Am Soc Mass Spectrom. 2011; 22:339. [PubMed: 21472593]
18. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, Brown M, Knowles JD, Halsall A, Haselden JN, Nicholls AW, Wilson ID, Kell DB, Goodacre R. Nat Protoc. 2011; 6:1060. [PubMed: 21720319]
19. Frey IM, Rubio-Aliaga I, Siewert A, Sailer D, Drobyshev A, Beckers J, de Angelis MH, Aubert J, Bar Hen A, Fiehn O, Eichinger HM, Daniel H. Physiol Genomics. 2007; 28:301. [PubMed: 17077276]
20. Jandke J, Spiteller G. J Chromatogr. 1986; 382:39. [PubMed: 3782411]
21. Tinoco AD, Saghatelian A. Biochemistry. 2011; 50:7447. [PubMed: 21786763]
22. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, Fung C, Nikolai L, Lewis M, Coutouly MA, Forsythe I, Tang P, Shrivastava S, Jeroncic K, Stothard P, Amegbey G, Block D, Hau DD, Wagner J, Miniaci J, Clements M, Gebremedhin M, Guo N, Zhang Y, Duggan GE, Macinnis GD, Weljie AM, Dowlatabadi R, Bamforth F, Clive D, Greiner R, Li L, Marrie T, Sykes BD, Vogel HJ, Querengesser L. Nucleic Acids Res. 2007; 35:D521. [PubMed: 17202168]

Figure 1.

The extracted ion chromatogram for a metabolite with m/z 422.1744 in a pooled tumor sample. The red bars indicate the predicted value for each indicated dipeptide based on the database (Table S1) and polynomial alignment using amino acids.

Figure 2.

MS/MS spectra of the dansylated AT dipeptide in a synthesized dipeptide standard (left) and in a pooled tumor sample (right). Comparison of these spectra reveals the high level of concordance for the major product ions due to the dansyl group (m/z 170,186,234,252) and the relatively small differences due to the residue specific I₂ and a₁ product ions.

Figure 3.

A diagram of the collision-induced dissociation of a dansyl dipeptide. The numbers and letters indicate the fragment nomenclature that is discussed in the supplemental text.

Table 1

Dipeptide levels in different lung regions.

Original cohort					Validation cohort	
T vs N	T vs A	A vs N	p value	BDL	T vs N	p value
6	0	0	<0.004	0	2	<0.004
29	8	0	<0.006	1	1	<0.006
19	2	0	<0.012	2	0	<0.012
22	8	0	<0.02	3	6	<0.02
14	8	1	<0.05	4	13	<0.05

The number of differentially abundant dipeptides present when tumor (T), adjacent-to-tumor (A) and normal (N) lung tissues obtained from individuals with NSCLC are compared. The number of dipeptides with a significant increase in the first tissue in each column relative to the other tissue is shown for the p-value in each row. Some of the dipeptides were below the detectable limit (BDL) in both of the tissues being compared, which reduced the total number of comparisons and produced a slightly lower p-value. The BDL column indicates the reduction in the number of comparisons for that row. The original cohort consisted of nine patients. The right-most columns show the data for the validation cohort (n = 10 patients).

Table 2

Diagnostic product ions in dansyl-dipeptide spectra.

Amino Acid	Position 1	Position 2
A	277 ¹	
D		88 ¹
E	317 ¹	102 ¹ ,84 ¹
F	353 ¹ ,120	120,d
G		
H	343,261 ² ,110	343 ¹ ,110
I	319 ¹ ,86	86,d
K	317,84	317,84
L	319 ¹ ,86	86,d
M	337 ¹ ,289 ¹ ,261 ¹ ,208,104,56	104 ¹ ,56 ¹
N	261 ¹	
P	303 ¹ ,70	70,116
Q	317 ¹	c ¹
R	408,156 ² ,70	156 ² ,70
S	293 ¹ ,60 ¹	60 ¹
T	307 ¹ ,289 ¹ ,74 ¹	74 ¹
V	305 ¹ ,72	72,d
W	392 ¹ ,261 ¹ ,159 ¹ ,144 ¹ ,130	159,144,130,d,d-C9H9N
Y	602 ¹ ,369	369

The superscripted number indicates that this fragment was only observed in singly or doubly charged spectra.