Metabolomics (2023) 19:64
https://doi.org/10.1007/s11306-023-02027-5

ORIGINAL ARTICLE

Statistical considerations and database limitations in NMR-based metabolic profiling studies

Imani L. Ross[1] · Julie A. Beardslee[2] · Maria M. Steil[3] · Tafadzwa Chihanga[4] · Michael A. Kennedy[2]

Received: 15 February 2023 / Accepted: 12 June 2023

Abstract

Introduction Interpretation and analysis of NMR-based metabolic profiling studies is limited by substantially incomplete commercial and academic databases. Statistical significance tests, including p-values, VIP scores, AUC values and FC values, can be largely inconsistent. Data normalization prior to statistical analysis can cause erroneous outcomes.

Objectives The objectives were (1) to quantitatively assess consistency among p-values, VIP scores, AUC values and FC values in representative NMR-based metabolic profiling datasets, (2) to assess how data normalization can impact statistical significance outcomes, (3) to determine resonance peak assignment completion potential using commonly used databases and (4) to analyze intersection and uniqueness of metabolite space in these databases.

Methods P-values, VIP scores, AUC values and FC values, and their dependence on data normalization, were determined in orthotopic mouse model of pancreatic cancer and two human pancreatic cancer cell lines. Completeness of resonance assignments were evaluated using Chenomx, the human metabolite database (HMDB) and the COLMAR database. The intersection and uniqueness of the databases was quantified.

Results P-values and AUC values were strongly correlated compared to VIP or FC values. Distributions of statistically significant bins depended strongly on whether or not datasets were normalized. 40–45% of peaks had either no or ambiguous database matches. 9–22% of metabolites were unique to each database.

Conclusions Lack of consistency in statistical analyses of metabolomics data can lead to misleading or inconsistent interpretation. Data normalization can have large effects on statistical analysis and should be justified. About 40% of peak assignments remain ambiguous or impossible with current databases. 1D and 2D databases should be made consistent to maximize metabolite assignment confidence and validation.

Keywords Metabolomics · Metabolic profiling · Metabolite databases · Statistical significance · NMR · Data normalization

1 Introduction

One goal of NMR-based metabolic profiling studies is to determine statistically significant differences in the intensities of NMR resonances in sets of NMR spectra belonging to distinct groups, e. g. between a healthy control group and a diseased study group (Beckonert et al., 2007; Emwas et al., 2019; Markley et al., 2017; Zhang et al., 2012). Several measures of statistical significance are commonly used in NMR-based metabolic profiling studies (Chihanga et al., 2018b, 2018c; Saccenti et al., 2014; Schmahl et al., 2018), including the Student's t-test p-value (Goodpaster et al., 2010; Worley & Powers, 2013), the variable importance in projection (VIP) score generated in partial least square-discriminant analysis (PLS-DA) (Chihanga et al., 2018b; Schmahl et al., 2018; Worley & Powers, 2013) and the accuracy of a given resonance for determining group belonging determined from the area under the curve (AUC) calculations from receiver operator characteristic (ROC) curves (Goodpaster et al., 2010; Worley & Powers, 2013). In many studies, only a single metric of statistical

✉ Michael A. Kennedy
kennedm4@miamioh.edu

1 Department of Chemistry and Biochemistry, University of California, San Diego, CA 92093, USA

2 Department of Chemistry and Biochemistry, Miami University, Oxford, OH 45056, USA

3 Division of Plastic Surgery, University of Texas Medical Branch, Galveston, TX 77555, USA

4 Division of Oncology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA

Published online: 28 June 2023

significance analysis is used, but in some studies, multiple metrics are used in the analysis of a single dataset, and in these cases, there can be a lack of consistency in the statistical significance of NMR resonance intensity differences determined by p-values, VIP scores and AUC values (Chihanga et al., 2018b, 2018c; Schmahl et al., 2018). One goal of this study is to establish a quantitative understanding of the degree of correlation and the corresponding lack of consistency in the statistical significance of NMR resonance intensity differences as determined by p-values, VIP scores and AUC values using an experimental dataset collected from an NMR-based metabolic profiling study, and to determine the impact that data normalization to total intensity had on the correlation of these metrics.

Another significant limitation of NMR-based metabolic profiling studies is the inability to make complete assignments of the NMR resonances using the most commonly used database tools, specifically the Chenomx software package (https://www.chenomx.com/), the human metabolite data base (HMDB) (Wishart et al., 2007, 2009, 2013, 2018), and the Complex Mixture Analysis by NMR (COLMAR) software package (Bingol et al., 2016). This problem is significant given that it is not uncommon for the unassigned fraction of NMR resonances in spectra of urine, human cell line extracts and other biological samples to reach or exceed 40% of detectable resonances. This aspect of data analysis, however, is rarely reported or addressed in published manuscripts, and the potential impact on the interpretation of the data and the limitations to a given study are generally not discussed. While the number of metabolites identified in NMR-based metabolic profiling studies typically falls in the range of 30–70 compounds (Chihanga et al., 2018b, 2018c; Schmahl et al., 2018; Yang et al., 2008), a "high bar" number, that has rarely been achieved, has been reported at 209 compounds by Wishart and co-workers (Bouatra et al., 2013). Another goal of this manuscript is to provide a quantitative understanding of the extent to which NMR spectra of mouse urine can be assigned using the combination the Chenomx, HMDB and COLMAR databases, and to determine the extent of overlap and uniqueness of these three databases.

A final issue addressed in this manuscript is a quantitative assessment of the confidence in the assignments that can be tentatively made using the Chenomx, HMDB and COLMAR databases. While it is common for investigators to report tables of metabolites assigned from NMR-based metabolic profiling data, the uncertainty or confidence in the reported assignments is rarely addressed. We have recently introduced the RANCM scheme (Joesten & Kennedy, 2019) that can be used to rank the confidence in metabolite assignments. This tool can then be used to assess and quantify the uncertainty of NMR-based metabolite assignments reported in NMR-based metabolic profiling studies, which was another goal of the study reported in this manuscript.

2 Materials and methods

2.1 Institutional approval of mouse studies

All procedures involving mice were approved by both the ethics committee and the Institutional Animal Care and Use Committee at Miami University (Animal Welfare Assurance Number: D16-00100). The approved protocols were assigned IACUC Project Numbers: 889 and 893. All procedures were conducted in such a way as to minimize any suffering or discomfort experienced by the mice. Daily health monitoring of mice was conducted by the Miami University's Animal Resources and Care Facility. Researchers were notified if any immediate action needed to be taken to care for the animals. For orthotopic surgeries, mice were anesthetized using isoflurane so that the mice were unconscious and unable to experience pain during the surgery. To initiate anesthesia, the mice were placed in a drop box containing 3–5% atmospheric isoflurane. Once anesthetized, the mice were placed on the surgical table and the surgery conducted while a nose cone administered 1–3% isoflurane to maintain anesthesia. After surgery, 0.05–0.10 mg/kg buprenorphine was injected subcutaneously between the mouse's shoulder blades following suturing of the incision as a pain reliever. The mice were then allowed to recover for 15 min after surgery in a cage atop a warming blanket. An additional dose of 0.05–0.10 mg/kg buprenorphine was injected subcutaneously 12 h post-operation to minimize pain and suffering of the mice during the study.

2.2 Orthotopic xenograft mouse model of pancreatic cancer

NOD.CB17-Prkdcscid/J SCID mice were purchased from Taconic (Hudson, NY) and maintained in a barrier facility Miami University according to institutional guidelines. MiaPaCa-2 were purchased from the American Type Culture Collection (Manassas, VA). MiaPaCa-2 cells were grown on the recommended medium, i. e. high glucose Dulbecco's Modified Eagle Medium, (DMEM) supplemented with 10% fetal bovine serum (FBS) and 1% penicillin–streptomycin (ThermoFischer, Pittsburgh PA). MiaPaCa-2 cells were scraped from the culture flasks and the suspended cells were drawn into insulin needles. Each mouse was administered 5% isofluorane in a sealed box to induce anesthesia. Once the mouse was unconscious, it was placed on the surgical table with a nose cone administering isofluorane at a concentration of 1–3% to maintain anesthesia during surgery. The mouse's left side from a dorsal view was shaved and

the spleen was located and used to find the pancreas. The site of the incision was scrubbed with butadiene and then rinsed with 70% alcohol three times. A 1 cm incision was made through both cutaneous layers and the pancreas was exposed. A 10 µl volume of MiaPaCa-2 cells suspended in DMEM high glucose media was injected into the pancreas. For the control mice, sham surgery was performed and a 10 µl volume of DMEM high glucose media injected into the pancreas. A cotton swab was used to apply pressure in order to minimize leakage of cells into abdominal cavity. The pancreas was repositioned in the abdominal cavity and the incision was sutured closed. A 0.10 mg/kg dose of buprenorphine was injected subcutaneously after the incision was closed to mitigate pain. Mice were fed ad libitum, and were monitored daily for signs of distress. After eight weeks of urine collection, as described below, mice were anesthetized and euthanized by a terminal blood draw collected by cardiac puncture according to approved procedures.

2.3 Sample collection and processing

Following sham surgery or MiaPaCa-2 injection into pancreata, mice were places in metabolism cages and urine samples collected at 1 week, 3 weeks, 5 weeks and 7 weeks after surgery. Urine samples were stored at −80 °C after collection. Samples were thawed on ice then prepared for NMR analysis as follows. A 1-ml aliquot of each sample was pH corrected to 7.4, centrifuged and buffered as previously published by Romick-Rosendale et al. (Romick-Rosendale et al., 2009, 2014).

2.4 NMR data collection

All NMR spectra were recorded on a Bruker Avance TM III spectrometer operating at 600 MHz. All experiments were conducted at 298 K using 5 mm NMR tubes (Norell, Morganton, NC). A standard ^{1}H 1D presaturation (zgpr) experiment was collected to assess the sample shimming, which was considered acceptable when the TSP peak linewidth was < 1 Hz. Once the sample was adequately shimmed, the 1D first increment of a NOESY (noesypr1D) was collected to determine if there were changes in lipid composition and the CPMG (cpmgpr1d) experiments were recorded were collected to allow quantitation of metabolite changes. All spectra were processed as previously reported by Romick-Rosendale et al. (). 2D ^{1}H-^{1}H TOCSY and ^{1}H-^{13}C HSQC spectra were collected as previously described by Chihanga et al. (). In brief, 2D ^{1}H-^{1}H TOCSY and ^{1}H-^{13}C HSQC spectra were recorded at a constant temperature of 298 K using the HSQC (hsqcetgpsi). The 2D ^{1}H-^{1}H TOCSY was collected using the TOCSY (mlevgpph19) pulse sequence at 298 K. The data was processed with Topspin 3.2 (Bruker BioSpin, Billerica MA).

2.5 Multivariate statistical analysis and statistical significance of individual buckets

The Bruker AMIX software v.3.9 (Analysis of MIXtures software, Bruker Biospin) manual bucketing tool was used to bucket spectral resonances. Manual bucketing was performed after stacking all of the spectra in both the control and study groups to ensure that any shifts of peaks of interest in individual spectra due to differences in pH, salt, etc. did not cause peaks from individual spectra to fall outside the bucketing range for individual buckets. In the normalized datasets, bucket areas in each spectrum were normalized to total intensity prior to statistical analysis. P-values were calculated using a modified Student's t-test that allows for different variances and numbers of samples in each group, known as a Welch's t test (Goodpaster et al., 2010). Per the standard application of a t test, the null hypothesis being tested was that there was no difference in the bucket spectral areas between the control and study groups. Return of a p-value from the t test less than 0.05 was used as a threshold to reject the null hypothesis, and therefore consider that the difference in the bucket areas were significant by the t test. For the purpose of this study, statistical significance of the other metrics was defined as follows based on applications used in past studies (Chihanga et al., 2018b, 2018c; Schmahl et al., 2018), namely a variable importance for predicting in partial least square discriminant analysis (PLS-DA) score of > 1.0, i.e. VIP > 1.0, an accuracy for predicting group belonging of greater of 70% (i.e. AUC > 0.7), and fold changes of greater than a factor of 2 (FC > 2.0). While the FC is not a statistical test, and no confidence levels can be assigned to a particular outcome, the FC is widely used to identify qualitative changes large datasets, such as microarray datasets, often in the form of volcano plots (Cui & Churchill, 2003). PLS-DA was performed using the SIMCA-P ver. 11.0 software package (Umetrics, Umea, Sweden).

2.6 Identification and quantification of metabolites

Metabolite identification and quantification were assessed as previously described (Chihanga et al., 2018a, 2018b, 2018c; Petrova et al., 2019; Romick-Rosendale et al., 2009, 2014; Schmahl et al., 2018; Standage et al., 2021; Watanabe et al., 2012) and the confidence in the metabolite assignments was evaluated using RANCM scores (Joesten & Kennedy, 2019). Briefly, initial metabolite assignments were made manually, and did not rely on a peak assignment algorithm, by comparison of the experimental spectra with the reference spectra databases included in Chenomx v.8.1 (Chenomx Inc., Alberta, Canada) augmented by the HMDB database (Wishart et al., 2007, 2009, 2013, 2018) and confirmed using the Complex Mixture Analysis by NMR (COLMAR) software (Bingol et al., 2016) that involved matching

experimental spectra to reference databases of metabolites. Two-dimensional NMR spectra (^{1}H-^{1}H TOCSY and ^{1}H-^{13}C HSQC,) were also used to confirm assignments and to define the RANCM scores (Joesten & Kennedy, 2019).

2.7 Database overlap analyses

For the database analyses, the 600 MHz Chenomx (reference library 10) and HMDB 4.0 databases and the 800 MHz Chenomx (reference library 10) and 850 MHz HMDB 4.0 databases were considered.

3 Results

3.1 Consistency of p-values, VIP scores, AUC values and fold changes across NMR spectral bins in a representative normalized and un-normalized dataset

As discussed above, several metrics are used to determine the statistical significance of differences in peak intensities in NMR-based metabolic profiling studies, including p-values, AUC values VIP scores and fold changes (FC). Here, we analyzed how consistent these measures were (i.e. the extent to which the various metrics indicated the same conclusions regarding whether or not the differences in bucket areas between groups were significant or not) in an example dataset obtained using an orthotopic mouse model of pancreatic cancer using both un-normalized and normalized data (Fig. 1). In the unnormalized dataset, a total of 184 bins were considered. Of these, 157 were significant by p-value, 14 by VIP 169 by AUC and 57 by FC (Fig. 1A). Four combinations of these measures were also considered, which yielded 13 bins significant by P + VIP, 157 by P + AUC, 57 by P + FC, 14 by VIP + AUC, 5 by VIP + FC and 56 by AUC + FC. Combinations of three metrics yielded 14 significant by P + VIP + AUC, 5 significant by P + VIP + FC and 5 significant by VIP + AUC + FC. Only five bins out of 157 deemed significant by both P and AUC were significant by all four metrics. This analysis indicated that the P-values (157) and accuracy/AUC values (169) were most consistent among the four metrics, in strong contrast to P + VIP, which shared only 13 common bins and VIP + AUC, which shared only 14 common bins. An intermediate number of bins were significant by P + FC. Reexamination of the same dataset after normalization to total intensity (Fig. 1B) yielded a similar distribution of consistency among the four metrics, however, the absolute number of bins significant by each metric changed significantly. For example, the number of bins significant by P-value dropped to 43 from 157 in the unnormalized dataset. Similarly, the number of bins significant by AUC dropped to 53 in the normalized dataset compared to 169 in the normalized dataset. The number of bins significant by P + AUC dropped to 37 in the normalized dataset from 157 in the unnormalized dataset. An overview of the impact on the absolute number of significant bins in each category is represented in Fig. 1C. Overall, this analysis highlights the strong influence that normalization to total intensity can have on the evaluation and assessment of NMR-based metabolic profiling data.

3.2 Analysis of distributions of statistically significant p-values, VIP scores and AUC values in un-normalized versus normalized datasets

In order to analyze the potential magnitude of the impact of normalization to total intensity on the statistical significance analysis of NMR-based metabonomics datasets, we evaluated four different weeks of mouse urine NMR data collected from an orthotopic mouse model study of pancreatic cancer. Urine samples were collected after 1 week, 3 weeks, 5 weeks, and 7 weeks after implanting human MiaPaCa2 cells in the pancreas of immunocompromised mice. Three metrics were assessed to determine the statistical significance of differences in metabolite NMR peak intensities: 1) Welch's t-test p-values, 2) VIP scores obtained from Partial Least Squares-Discriminant Analysis (PLS-DA) and 3) group-belonging prediction accuracy obtained from receiver operator characteristic (ROC) area under the curve (AUC) values. Thresholds to define statistically significant differences in peak/feature intensities were p-values < 0.05, VIP scores > 1.0 and AUC values are > 0.70. The results are summarized in Fig. 2. The number of statistically significant differences in peak intensities in the NMR datasets was strongly affected by normalization. It should be noted that all NMR samples in this study were prepared from identical amounts of urine collected from mice using metabolism cages. In the week 1 dataset, only 15 out of 137 peaks/features (~ 11%), were significant by p-value in both normalized and un-normalized datasets, 45 peaks/features (~ 33%) were only significant in un-normalized data, six peaks/features (~ 4%) were only significant in normalized data and 71 peaks/features (~ 52%) were not significant in either normalized or un-normalized data (Fig. 2A). Similar trends were observed for the p-value distributions in weeks 3, 5 and 7 (Fig. 2A). The impact of normalization on VIP scores was similar to that observed on p-values (Fig. 2B). For example, in week 1, only seven peaks/features (~ 5%) were significant with or without normalization, 17 peaks/features (~ 12%) were significant only in un-normalized data, one peak/feature (< 1%) was only significant only in normalized data and 112 peaks/features (~ 82%) were not significant in either normalized or un-normalized data (Fig. 2B). Again, this trend was similar in weeks 3, 5, and 7 (Fig. 2B). Finally, the accuracy of the peaks/features for distinguishing between control and study

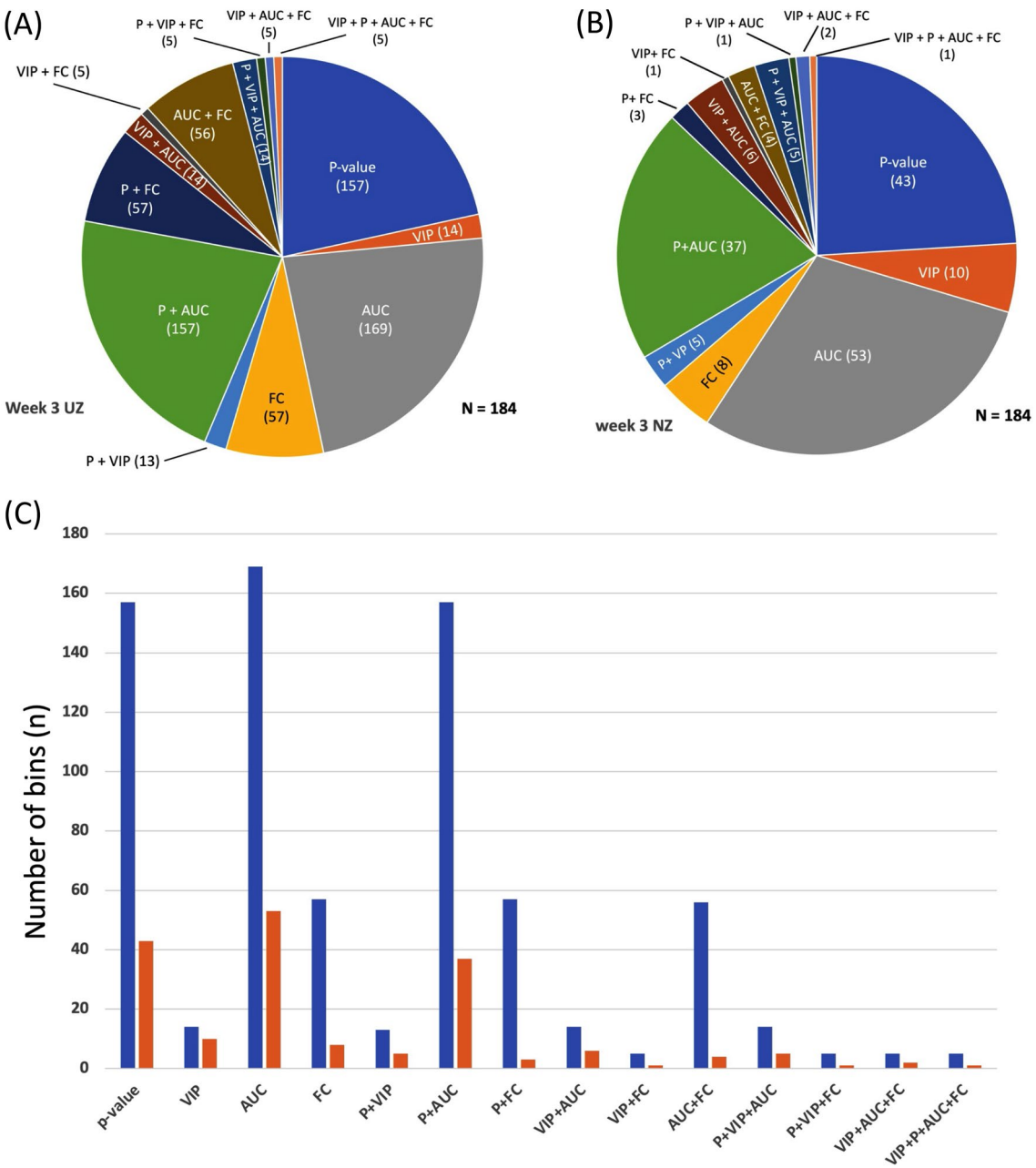

Fig. 1 Analysis of Consistency of Statistical Significance Metrics in Un-Normalized and Normalized NMR-Based Metabolic Profiling Data. Pie chart analysis of comparison of **A** un-normalized and **B** normalized week-three NMR spectra obtained from an orthotopic mouse model of pancreatic cancer. The pie graph shows the number of spectral bins that are significant by p-value, VIP, AUC or FC alone, and for all permutation of combinations. **C** Bar graph plot of the number of statistically significant bins in each category for the un-normalized dataset (blue) and the normalized dataset (orange)

groups was similarly impacted by normalization. In week 1, 20 peaks/features (~15%) were significant regardless of normalization, 19 peaks/features (~14%) were only significant in un-normalized data, 10 peaks/features (~7%) were significant only in normalized data and 88 peaks/features (~64%) were not significant in either normalized or un-normalized data (Fig. 2C).

Figure 3A shows how peak intensity distributions were affected by normalization for the three most significant peaks/features in the un-normalized dataset that changed to not significant after normalization. In the unnormalized data set, the peak at 8.12 ppm had very high accuracy for distinguishing between the control and cancer group based on an AUC = 0.893 supported by a p-value = 1.19×10^{-5}

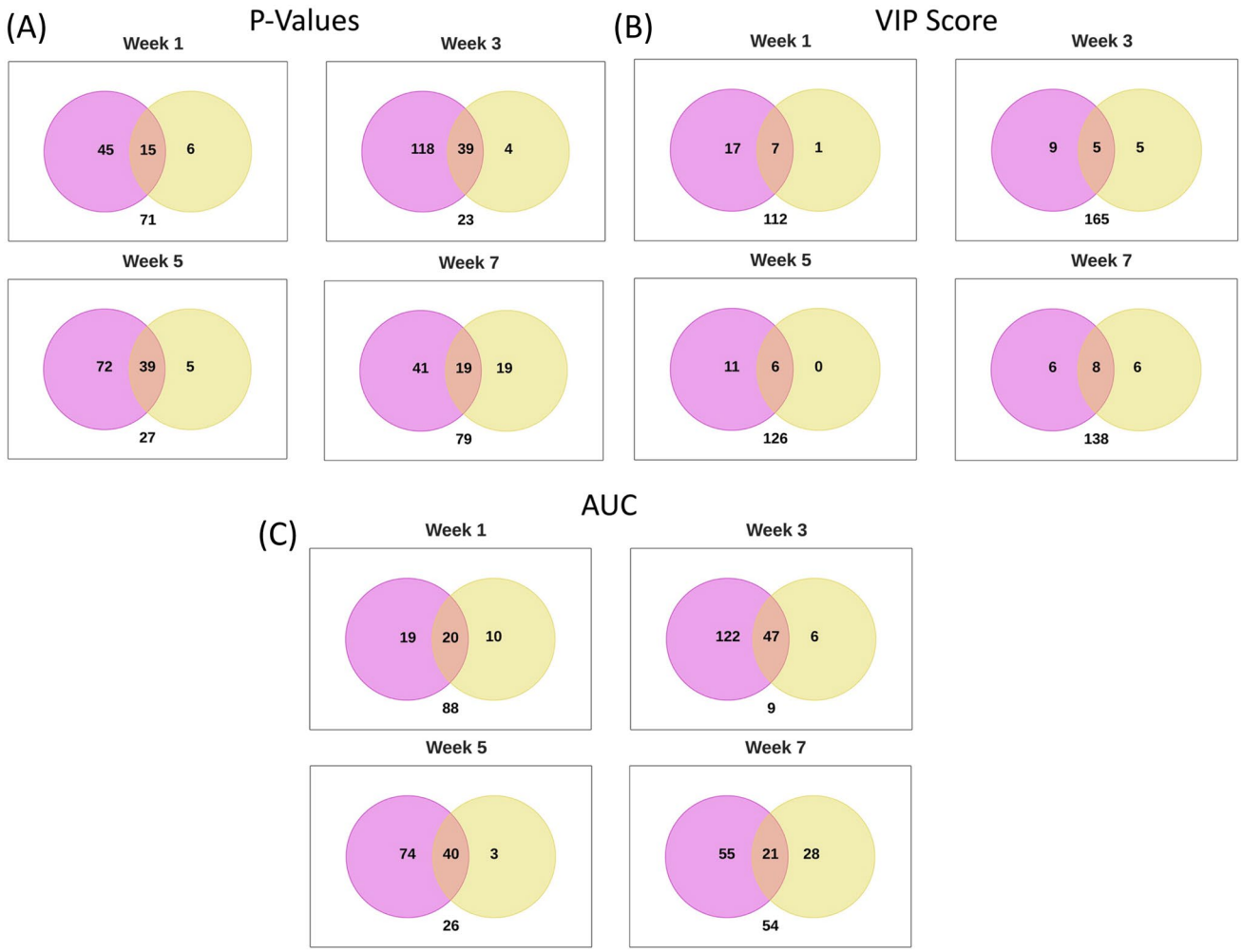

Fig. 2 Impact of normalization on significance analysis of p-values, VIP scores and AUC values. A Venn diagram analysis is shown for four weeks of data collection of mice urine in an orthotopic mouse model of pancreatic cancer. The impact of data normalized is shown for **A** p-values, **B** VIP scores and **C** AUC values. The number of peaks that are significant in both normalized and un-normalized data are shown in the intersection of each diagram (colored orange), those significant in un-normalized data shown in the left-hand circle (pink) and those significant in normalized data only shown in the right-hand circle (yellow). The number of peaks not significant in either normalized or un-normalized data are indicated by the number in the box outside the circles

(Fig. 3A, Top, Left) that dropped to low accuracy, near that of a random prediction, with an AUC = 0.546 with a p-value = 0.976 ppm after normalization (Fig. 3A, Bottom, Left). Similarly, the peak at 6.21 ppm had an AUC = 0.918 with a p-value = 3.61×10^{-5} prior to normalization (Fig. 3A, Top, Middle) that dropped to an AUC = 0.515 with a p-value = 0.787 after normalization (Fig. 3A, Bottom, Middle). Finally, the peak at 3.046 ppm had an AUC = 0.918 with a p-value = 3.30×10^{-5} in the un-normalized data (Fig. 3A, Top, Right) that dropped to an AUC = 0.531 with a p-value = 0.870 (Fig. 3A, Bottom, Right). The associated boxplots for each comparison showed that prior to normalization the peak intensity distributions were completely separated whereas after normalization the peak distributions were completely overlapped. These three examples are representative of the 118 peaks/features that were significant based on their p-values and the 122 peaks/features that were significant based on their AUC prior to normalization that were no longer significant by either measure after normalization to total intensity.

Peaks/features that were not significant in the un-normalized dataset that changed to significant following normalization are shown in Fig. 3B. The peak at 7.88 ppm had an AUC = 0.500 and a p-value = 0.530 prior to normalization (Fig. 3B, Top, Left) that changed to an AUC = 0.776 and a p-value = 0.019 after normalization (Fig. 3B, Bottom, Left). The peak at 2.88 ppm had an AUC = 0.679 and a p-value = 0.130 in the un-normalized data (Fig. 3B, Top, Middle) that changed to an AUC = 0.74 and a p-value = 0.032 after normalization (Fig. 3B,

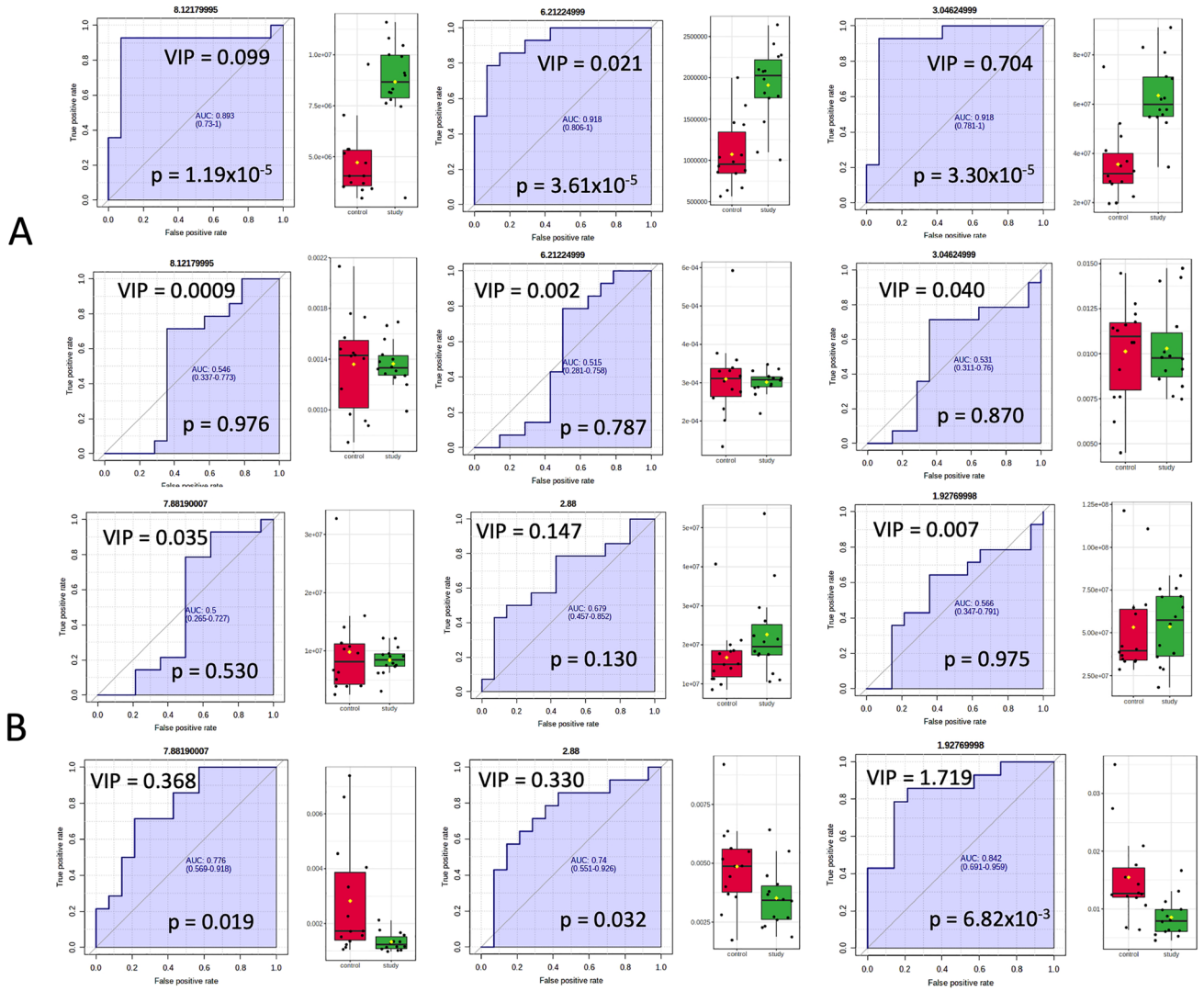

Fig. 3 Illustration of how normalization impacts peak intensity distributions and corresponding p-values, VIP scores and AUC values change for NMR resonances with the most statistically significant differences in normalized and unnormalized mouse urine datasets. **A** The top row shows ROC plots for the three NMR resonances with the most statistically significant differences in the unnormalized dataset (8.12 ppm (left), 6.21 ppm (middle) and 3.046 ppm (right)) shown at top with the ROC curves, AUC values, VIP scores and p-values indicated at left and the histogram of the peak intensities shown at right. The bottom row shows how the plots, values and peak intensity distributions change upon normalization. **B** The bottom row shows ROC plots for the three NMR resonances with the most statistically significant differences in the normalized dataset (7.88 ppm (left), 2.88 ppm (middle) and 1.92 ppm (right)) shown at top with the ROC curves, AUC values, VIP scores and p-values indicated at left and the histogram of the peak intensities shown at right. The top row shows how the plots, values and peak intensity distributions change without normalization

Bottom, Middle). Finally, a peak at 1.92 ppm that had an AUC = 0.566 and a p-value = 0.975 prior to normalization (Fig. 3B, Top, Right) that changed to an AUC = 0.842 and a p-value = 6.83×10^{-3} after normalization (Fig. 3B, Bottom, Right). In these cases, the associated boxplots for each comparison showed increased separation of peak intensity distributions after normalization. These three examples were representative of just four peaks that were not significant in un-normalized data that were significant following normalization based on p-values (Fig. 2A) and four peaks that switched to significant following normalization to total intensity based on AUC values (Fig. 2C).

3.3 Impact of normalization of statistical significance analysis in a simple test sample case

In order to illustrate the potential impact of *normalization to total intensity* in a simple test case, two sets of well-defined samples were analyzed. One set contained five compounds

(L-alanine, L-cysteine, L-leucine, taurine, and L-threonine) and a second set contained six compounds (The same five compounds just listed plus p-cresol). All compounds common to both sample sets were prepared at the same concentration in both sets of samples. Six experimental replicates were prepared for both the five-compound and the six-compound sets. Representative spectra from the five-compound and six-compound samples are shown in Fig. 4A and B respectively, illustrating the absence of resonance peaks for compound p-cresol. Naturally, the total intensity of the six-compound spectra, which is the sum of all peak areas, will be greater than that of the five-compound spectra. Consequently, matched peak intensities should be about the same in the five-compound and six-compound spectra when comparing un-normalized data sets, as is observed in Fig. 4C. In contrast, after normalization, matched peaks would no longer be expected to have the same intensities, since the total intensity of the six-compound spectra will be greater than that of the five-compound spectra, as discussed above, despite that fact that the compound concentrations were prepared to be the same in the two sets of samples. The effect that *normalization to total intensity* can have on individual peak intensities is illustrated in Fig. 4D, where the matched peak at 1.32 ppm in the normalized six-compound spectrum is noticeably smaller than that observed in the five-compound spectrum. As expected, there is no significant difference in the peak intensities of the five-compound spectra and six-compound spectra as indicated by their boxplot distributions and supported by an AUC value of 0.69 (Fig. 4E), in contrast to the clear difference in peak intensity distributions indicated by the boxplot analysis and associated AUC value of 1.0 (Fig. 4F). This simple example illustrates the risk of unjustified *normalization to total intensity* that is routinely and widely applied in NMR-based metabolic profiling studies. Clearly, *normalization to total intensity* has the potential of altering peak intensity distributions for peaks that are not actually different between comparison groups. For example, in the case where new compounds show up or disappear in a disease group in comparison to a healthy control group, thus altering the total intensity of the disease group spectra, *normalization to total intensity* has the possibility of generating statistically significant differences in the intensities of resonances metabolites that have no actual difference in the biological samples.

3.4 Quantification of resonance peaks with no matches in the Chenomx, HMDB and COLMAR databases and assessment of uncertainty in metabolite assignments using RANCM

An attempt was made to quantify the typical number of peaks that had relatively confident metabolite assignments (i.e. those with significant diagnostic matching features with metabolites in the existing databases), those that had ambiguous assignments (i. e. those peaks that could be explained by more than one metabolite in the databases, and those that had no possible matching peaks in the databases. The analysis was performed on urine from a mouse containing an orthotopic, xenograft tumor established with human Mia-PaCa-2 pancreatic cancer cells and the soluble cell extracts from two different human pancreatic cancer cell lines: Mia-PaCa-2 and Panc-1, and the results are shown in Fig. 5. It was observed that 52.8% of the peaks in the urine from the orthotopic, xenograft mouse model of pancreatic cancer had at least a putative assignment while 25% of the peaks had an ambiguous assignment and 22.2% of the peaks had no matching peaks in the database (Fig. 5A). Quite similar patterns were observed for the two human pancreatic cancer cell lines with 63.5% of peaks having a putative assignment, 25.2% having ambiguous peak assignments and 11.3% of peaks having no matching peak assignments in the Mia-PaCa-2 cell line extracts (Fig. 5B) and 53.3% of peaks having a putative assignment, 31.1% having ambiguous peak assignments and 15.6% of peaks having no matching peak assignments in the Panc-1 cell line extracts (Fig. 5C).

3.5 Evaluation of the overlap of metabolite space between the Chenomx, HMDB and COLMAR databases

Several large public databases are commonly used to aid in the assignment of metabolites in NMR-based metabolic profiling studies. One of the most widely used commercially available databases is available from the Chenomx company (https://www.chenomx.com/) which can be supplemented with the human metabolite database (HMDB) (https://hmdb.ca/). Another very useful resource for assigning two-dimensional homonuclear ^{1}H-^{1}H TOCSY spectra and ^{1}H-^{13}C heteronuclear HSQC spectra is the COLMAR software and database (http://spin.ccic.ohio-state.edu/index.php/colmar). We have developed the RANCM strategy to assign confidence levels to NMR-based metabolite assignments (Joesten & Kennedy, 2019) that uses all three software packages. A limitation of this approach is that all three databases are significantly incomplete with 40–50% of observed NMR peaks in human and mouse urine samples and human cancer cell extracts have no matching peaks in the ChenomX, HMDB or COLMAR databases. Here we provide a quantitative analysis of the overlap between the three databases. The ChenomX database included 339 compounds, 323 of which were unique (16 were found to be redundant, i. e. the same compounds listed with different synonyms). The HMDB database included 657 compounds, 614 of which were unique (43 were the same compounds listed with different synonyms). The COLMAR database included 701

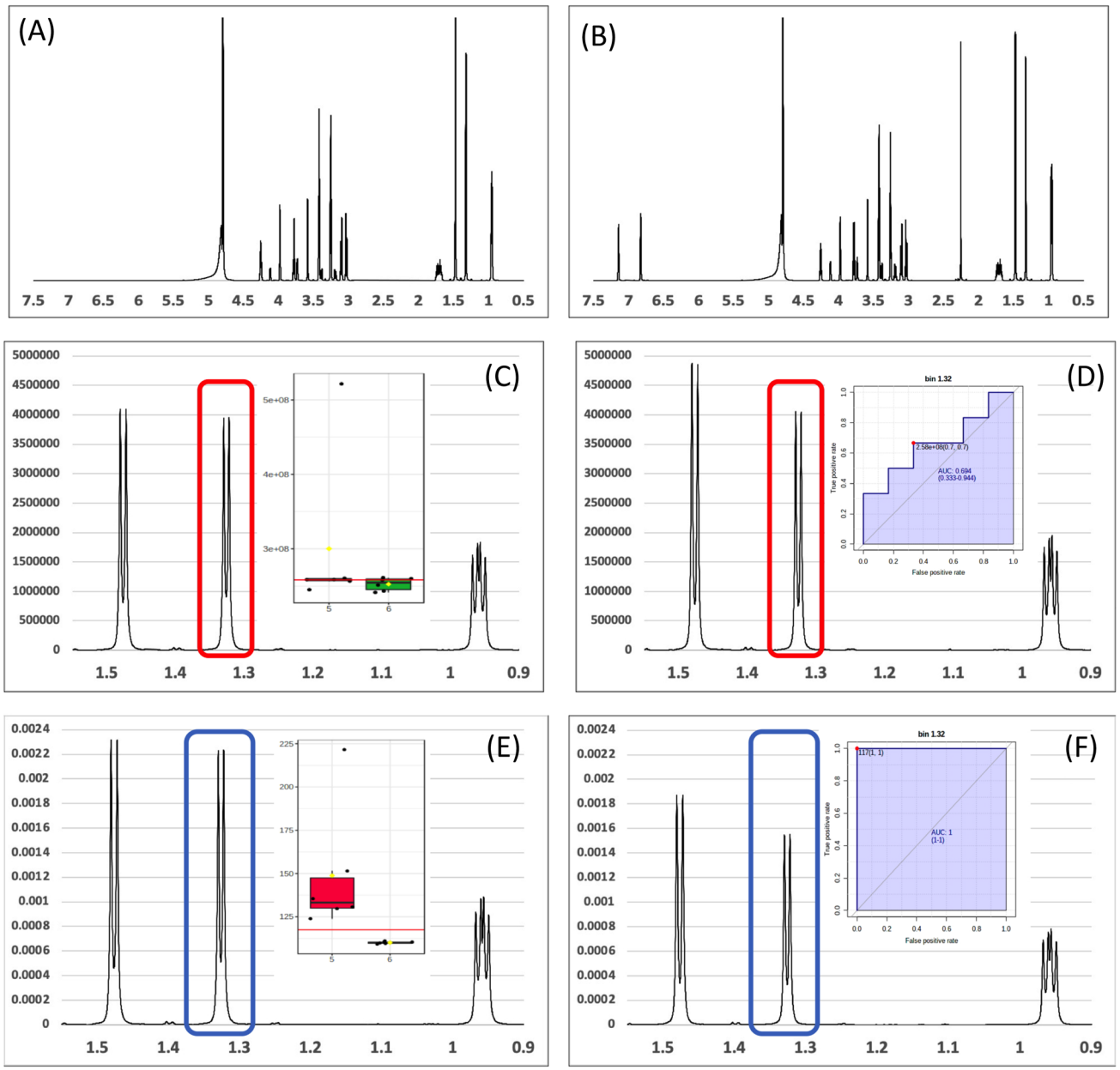

Fig. 4 Illustration of how normalization can impact peak intensities, p-values, VIP scores and AUC values in a simple well-defined set of NMR samples. **A** Representative spectrum of the 5-compound sample set. **B** Representative spectrum of the six-compound sample set. **C** L-threonine methyl peak intensity (1.32 ppm, boxed in red) in a representative spectrum of the five-compound set without normalization. **D** L-threonine methyl peak intensity (1.32 ppm, boxed in red) in a representative spectrum of the six-compound set without normalization. C-Inset) A boxplot showing the comparison of the peak intensity distributions between the 5-compound and 6-compound data sets without normalization. D-Inset) A ROC plot showing the accuracy of the L-threonine methyl peak intensity (1.32 ppm) for distinguishing between the 5-compound and 6-compound sets of spectra without normalization. **E** L-threonine methyl peak intensity (1.32 ppm, boxed in blue) in a representative spectrum of the five-compound set after normalization. **F** L-threonine methyl peak intensity (1.32 ppm, boxed in red) in a representative spectrum of the five-compound set after normalization. E-Inset) A boxplot showing the comparison of the peak intensity distributions between the 5-compound and 6-compound data sets after normalization. F-Inset) A ROC plot showing the accuracy of the L-threonine methyl peak intensity (1.32 ppm) for distinguishing between the 5-compound and 6-compound sets of spectra after normalization

compounds, 673 of which were unique (28 were the same compounds listed with different synonyms). There were 254 compounds common to all three databases, 16 compounds shared between the Chenomx and the HMDB databases but absent in COLMAR, 25 compounds common between the Chenomx and COLMAR databases that were absent in the

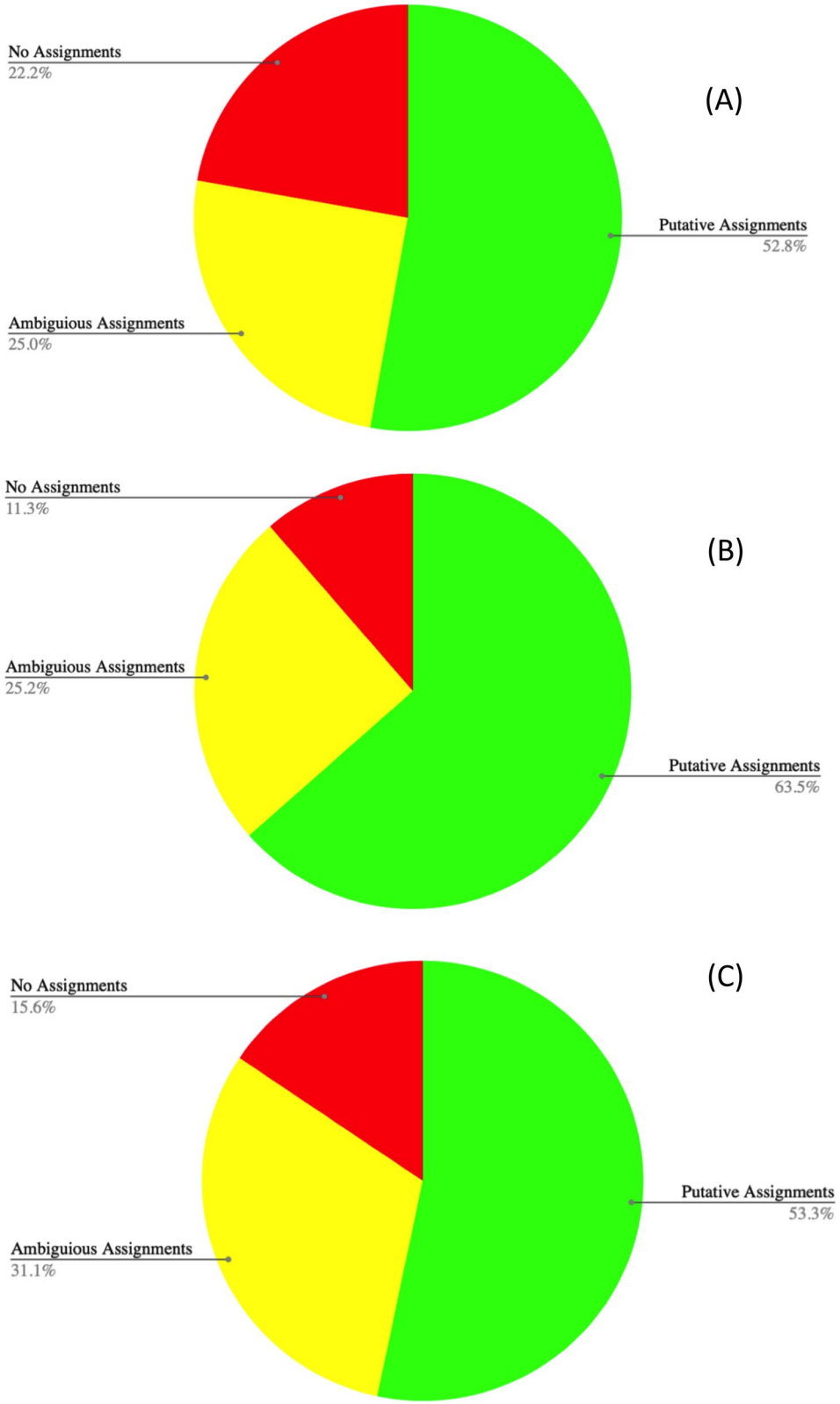

Fig. 5 Analysis of distribution of putative, ambiguous and no assignment of peaks in mouse urine and two different pancreatic cancer cell lines. Putative (green), ambiguous (yellow) and no assignments (red) in **A** mouse urine from a mouse with a pancreatic tumor, **B** the soluble fraction from the human MiaPaCa-2 pancreatic cancer cell line, and **C** the soluble fraction from the human Panc-1 pancreatic cancer cell line

HMDB and 246 compounds common to the HMDB and COLMAR that were absent from the Chenomx database. Finally, there were 28 compounds unique to the Chenomx database, 98 compounds unique to the HMDB database and 148 compounds unique to the COLMAR database. The overlap among these databases is summarized in a Venn diagram

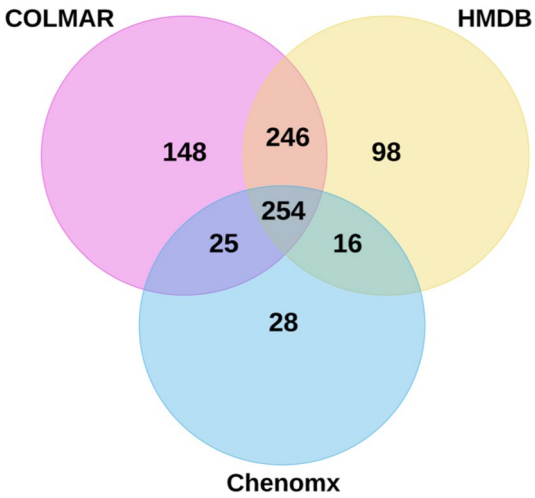

Fig. 6 Analysis of the overlap in the composition of commonly used metabolite databases. The Venn diagram shows the number of metabolites that are shared between all three databases (254), those shared between any two of the databases, and the number of compounds unique to each database

shown in Fig. 6. The conversion of all metabolites listed in all three databases to a common naming convention and the detailed analysis of the specific compounds that belong to each overlap category are included in Excel sheets available in the supplementary material and stored in a public database repository.

4 Conclusions

Here, we have addressed three important issues that continue to impact and limit NMR-based metabolic profiling studies. First, we addressed the inconsistency in various measures commonly used to assess statistically significant differences in NMR peak intensities when comparing sets of NMR spectra belonging to distinct groups, e. g. a healthy control group and a disease group. Our analysis indicated that p-values and AUC values appear strongly correlated whereas fold changes and VIP scores appear weakly or non-correlated with the p-values and AUC values, and appear to be less useful as a consequence.

Next, we quantitatively addressed the impact of data normalization to total intensity had prior to statistical analysis using urine samples obtained from an orthotopic, xenograft mouse model study of pancreatic cancer. The analysis indicated that the normalization had a large impact on the number and distribution of statistically significant resonances, and consequently, differences in metabolite concentrations. This practice has widespread potential significance in the field since investigators often routinely normalize NMR data prior to statistical analysis and reviewers often require some form of data normalization prior to statistical analysis when reviewing manuscripts. While experimental factors sometimes justify data normalization, automatic normalization of samples containing identical amounts of sample is not necessarily justified and may introduce increased variation into the dataset. Unjustified data normalization can also lead to false conclusions regarding the statistical significance of apparent differences in metabolite concentrations. An example where data normalization would be justified would be in a case where nine urine samples from a study group were used to prepare NMR samples using 540 µL of urine plus 60 µL of D_2O, however, only 100 µL of urine was available to make up one sample, which could be prepared using 440 µL of buffer and 60 µL of D_2O. In order to include the dilute sample prepared from 100 µL of urine in the analysis along with the other nine samples prepared from 540 µl of urine, data normalization would be justified and required. Another example where data normalization would be justified and required might occur when performing NMR analysis of human cell culture extracts. For example, if five control NMR samples were prepared from a single Petri dish and five additional NMR control samples were prepared from two Petri dishes. Then, in order to combine the ten samples into a single control group, data normalization would be required and justified. Given that the decision of whether or not to normalize spectra collected from samples prepared from identical quantities can introduce a confounding variable, it should be justified with a solid rationale. For example, if the metabolite concentrations vary naturally in some range, as would be expected, normalization could potentially introduce variance into the data that did not exist prior to data normalization. When comparing control and study groups, there is a potential risk that normalization of the data, in the case where normalization cannot be justified, could cause differences in apparent concentrations of some metabolites that appear to be statistically different when they actually are not, as was observed in our synthetic sample test case. It is also possible that normalization of the data may mask real statistical differences in peak intensities that become statistically insignificant after normalization.

Finally, we have drawn attention to a serious limitation of most NMR-based metabolic profiling studies, i.e. the observation that most NMR resonances detected in NMR-based metabolic profiling studies cannot be assigned from the most widely used metabolite databases. In fact, it is not uncommon for > 40% of the resonances to go unassigned. This problem not only occurs when evaluating human urine samples, but is significantly exacerbated when biofluids from other organisms are examined, for example mouse and rat urine, or when other biological fluids are examined, e. g. human cancer cell line extracts. A major cause of this problem is that many of the NMR peaks observed in these samples are simply not represented in

the ChenomX and HMDB databases. A further problem is that often when a metabolite is putatively assigned in the one-dimensional NMR spectrum, the peaks cannot be confirmed in the multidimensional NMR COLMAR database, which is used to increase the confidence in the metabolite assignments in the one-dimensional NMR spectra from which the statistical significance test conclusions are drawn, because the compounds in the ChenomX and HMDB databases are not always present in the COLMAR database. To gain a better understanding of the overlap and intersection of the metabolite spaces of each database, we quantified the overlap and intersections of the ChenomX, HMDB and COLMAR databases, the results of which revealed substantial overlap among the three databases, but also significant unique metabolite spaces in each database as well. This analysis indicates that an effort should be not only to significantly expand the metabolite coverage of the ChenomX and HMDB databases, so that the completion of putative assignments of NMR resonances from the one-dimensional NMR spectra can be significantly increased, made to ensure that the COLMAR metabolite list covers the entire metabolite space present in the ChenomX and HMDB databases used to assign one-dimensional NMR spectra, so that assignment confidence can be routinely maximized in NMR-based metabolic profiling studies.

Our study does include a number of caveats. One caveat, for example, is that the choice of cutoffs for the metrics considered can also impact the analysis presented here, but the cutoffs used in this study were chosen based on values commonly found to be used in the literature. Consequently, while the analyses may change depending on the choice of cutoffs used for the various metrics, the analysis presented here should be representative and informative for many investigators who choose the same or similar cutoffs in their own studies. Another caveat is that we have only considered a single type of normalization scheme, namely *normalization to total intensity*. This raises the interesting question as to how other normalization schemes may or may not impact the resulting distributions of statistically significant changes in NMR-based metabolic profiling datasets, however, normalization to total intensity is a widely used normalization technique and comparison with other normalization schemes was beyond the scope of the current study. And yet another caveat of this analysis is the fact that the t tests and other tests have been applied to the "processed" data and not directly to the observed raw data, which is the free induction decays themselves. A consequence of this fact is that the processing step have the potential to impact the statistical analyses, including the analyses performed and described in this manuscript. The possibility exists that the conclusions drawn from this analysis could change if the raw data were processed in a different manner, and the reader should be aware of this possibility.

As a final note, it should be pointed out that the observations and conclusions drawn from this study reflected a limited space of the potential types of biological matrix samples that can be studied in NMR-based metabolic profiling studies. While the observations reported here were drawn from urine samples collected from an orthotopic, xenograft mouse model study of pancreatic cancer, and soluble cell extracts collected from two human pancreatic cancer cell lines, it is possible that the conclusions drawn from these studies may not be universally true when applied to other samples sets, for example human urine samples, tissue extracts, or other biological matrices. In order to answer these questions, further studies would need to be conducted to address these possibilities. However, the results presented here highlight some important limitations and caveats that persist in the analyses of NMR-based metabolic profiling data that should be considered when designing NMR-based metabolic profiling studies and when analyzing and interpreting NMR-based metabolic profiling datasets.

Acknowledgements This research was supported and funded by the National Science Foundation under Grant No. CHE-1851795. MAK acknowledges support of Miami University and the Ohio Eminent Scholar Program. The authors acknowledge the several undergraduate students who participated in conducting the orthotopic, xenograft mouse model of pancreatic cancer, including Hannah Ruby, Jenna Nicholson, Mark Bombacher, Kayla Hawkins, and Dan Fay.

Authors contribution MAK conceived and designed the study. ILR, JAB, MS and TC performed the research. ILR, JAB and MAK analyzed the data. MAK wrote the paper. All authors reviewed the manuscript.

Data availability All of the raw data is available at figshare: https://figshare.com/account/home#/projects/159752 or by request.

Declarations

Conflict of interest The authors declare no conflicts of interest.

Ethical statement All applicable international, national and institutional guidelines for the care and use of animals was followed.

References

Beckonert, O., Keun, H. C., Ebbels, T. M., Bundy, J., Holmes, E., Lindon, J. C., et al. (2007). Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols, 2*, 2692–2703. https://doi.org/10.1038/nprot.2007.376

Bingol, K., Li, D. W., Zhang, B., & Bruschweiler, R. (2016). Comprehensive metabolite identification strategy using multiple two-dimensional NMR spectra of a complex mixture implemented in the COLMARm web server. *Analytical Chemistry, 88*, 12411–12418. https://doi.org/10.1021/acs.analchem.6b03724

Bouatra, S., Aziat, F., Mandal, R., Guo, A. C., Wilson, M. R., Knox, C., et al. (2013). The human urine metabolome. *PLoS ONE, 8*, e73076. https://doi.org/10.1371/journal.pone.0073076

Chihanga, T., Hausmann, S. M., Ni, S., & Kennedy, M. A. (2018a). Influence of media selection on NMR based metabolic profiling of human cell lines. *Metabolomics, 14*, 28. https://doi.org/10.1007/s11306-018-1323-2

Chihanga, T., Ma, Q., Nicholson, J. D., Ruby, H. N., Edelmann, R. E., Devarajan, P., et al. (2018b). NMR spectroscopy and electron microscopy identification of metabolic and ultrastructural changes to the kidney following ischemia-reperfusion injury. *American Journal of Physiology-Renal Physiology, 314*, F154–F166. https://doi.org/10.1152/ajprenal.00363.2017

Chihanga, T., Ruby, H. N., Ma, Q., Bashir, S., Devarajan, P., & Kennedy, M. A. (2018c). NMR-based urine metabolic profiling and immunohistochemistry analysis of nephron changes in a mouse model of hypoxia-induced acute kidney injury. *American Journal of Physiology-Renal Physiology, 315*, F1159–F1173. https://doi.org/10.1152/ajprenal.00500.2017

Cui, X., & Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology, 4*, 210. https://doi.org/10.1186/gb-2003-4-4-210

Emwas, A. H., Roy, R., McKay, R. T., Tenori, L., Saccenti, E., Gowda, G. A. N., et al. (2019). NMR spectroscopy for metabolomics research. *Metabolites*. https://doi.org/10.3390/metabo9070123

Goodpaster, A. M., Romick-Rosendale, L. E., & Kennedy, M. A. (2010). Statistical significance analysis of nuclear magnetic resonance-based metabonomics data. *Analytical Biochemistry, 401*, 134–143. https://doi.org/10.1016/j.ab.2010.02.005

Joesten, W. C., & Kennedy, M. A. (2019). RANCM: A new ranking scheme for assigning confidence levels to metabolite assignments in NMR-based metabolomics studies. *Metabolomics, 15*, 5. https://doi.org/10.1007/s11306-018-1465-2

Markley, J. L., Bruschweiler, R., Edison, A. S., Eghbalnia, H. R., Powers, R., Raftery, D., et al. (2017). The future of NMR-based metabolomics. *Current Opinion in Biotechnology, 43*, 34–40. https://doi.org/10.1016/j.copbio.2016.08.001

Petrova, I., Xu, S., Joesten, W. C., Ni, S., & Kennedy, M. A. (2019). Influence of drying method on NMR-based metabolic profiling of human cell lines. *Metabolites*. https://doi.org/10.3390/metabo9110256

Romick-Rosendale, L. E., Goodpaster, A. M., Hanwright, P. J., Patel, N. B., Wheeler, E. T., Chona, D. L., et al. (2009). NMR-based metabonomics analysis of mouse urine and fecal extracts following oral treatment with the broad-spectrum antibiotic enrofloxacin (Baytril). *Magnetic Resonance in Chemistry, 47*(Suppl 1), S36-46. https://doi.org/10.1002/mrc.2511

Romick-Rosendale, L. E., Legomarcino, A., Patel, N. B., Morrow, A. L., & Kennedy, M. A. (2014). Prolonged antibiotic use induces intestinal injury in mice that is repaired after removing antibiotic pressure: Implications for empiric antibiotic therapy. *Metabolomics, 10*, 8–20. https://doi.org/10.1007/s11306-013-0546-5

Saccenti, E., Hoefsloot, H. C. J., Smilde, A. K., Westerhuis, J. A., & Hendriks, M. M. W. B. (2014). Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics, 10*, 361–374. https://doi.org/10.1007/s11306-013-0598-6

Schmahl, M. J., Regan, D. P., Rivers, A. C., Joesten, W. C., & Kennedy, M. A. (2018). NMR-based metabolic profiling of urine, serum, fecal, and pancreatic tissue samples from the Ptf1a-Cre; LSL-KrasG12D transgenic mouse model of pancreatic cancer. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0200658

Standage, S. W., Xu, S., Brown, L., Ma, Q., Koterba, A., Lahni, P., et al. (2021). NMR-based serum and urine metabolomic profile reveals suppression of mitochondrial pathways in experimental sepsis-associated acute kidney injury. *American Journal of Physiology. Renal Physiology, 320*, F984–F1000. https://doi.org/10.1152/ajprenal.00582.2020

Watanabe, M., Sheriff, S., Lewis, K. B., Cho, J., Tinch, S. L., Balasubramaniam, A., et al. (2012). Metabolic profiling comparison of human pancreatic ductal epithelial cells and three pancreatic cancer cell lines using nmr based metabonomics. *Journal of Molecular Biomarkers and Diagnosis*. https://doi.org/10.4172/2155-9929.S3-002

Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vazquez-Fresno, R., et al. (2018). HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Research, 46*, D608–D617. https://doi.org/10.1093/nar/gkx1089

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., et al. (2013). HMDB 3.0–the human metabolome database in 2013. *Nucleic Acids Research, 41*, D801–D807. https://doi.org/10.1093/nar/gks1065

Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., et al. (2009). HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Research, 37*, D603–D610. https://doi.org/10.1093/nar/gkn810

Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., et al. (2007). HMDB: the human metabolome database. *Nucleic Acids Research, 35*, D521–D526. https://doi.org/10.1093/nar/gkl923

Worley, B., & Powers, R. (2013). Multivariate analysis in metabolomics. *Curr Metabolomics, 1*, 92–107. https://doi.org/10.2174/2213235X11301010092

Yang, W., Wang, Y., Zhou, Q., & Tang, H. (2008). Analysis of human urine metabolites using SPE and NMR spectroscopy. *Science in China Series b: Chemistry, 51*, 218–225. https://doi.org/10.1007/s11426-008-0031-6

Zhang, A., Sun, H., Wang, P., Han, Y., & Wang, X. (2012). Modern analytical techniques in metabolomics analysis. *The Analyst, 137*, 293–300. https://doi.org/10.1039/c1an15605e