Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Anal Chem. Author manuscript; available in PMC 2015 July 29.

Published in final edited form as:

Anal Chem. 2015 March 17; 87(6): 3177–3186. doi:10.1021/ac504012a.

Selective Paired Ion Contrast Analysis: A Novel Algorithm for Analyzing Postprocessed LC-MS Metabolomics Data Possessing High Experimental Noise

Tytus D. Mak[a], **Evagelia C. Laiakis**[b], **Maryam Goudarzi**[b], and **Albert J. Fornace Jr**[a,b]

[a]Lombardi Comprehensive Cancer Center, New Research Building E504/508, 3970 Reservoir Road, NW, Washington, DC 20057, United States

[b]Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, New Research Building E504/508, 3970 Reservoir Road, NW, Washington, DC 20057, United States

Abstract

One of the consequences in analyzing biological data from noisy sources, such as human subjects, is the sheer variability of experimentally irrelevant factors that cannot be controlled for. This holds true especially in metabolomics, the global study of small molecules in a particular system. While metabolomics can offer deep quantitative insight into the metabolome via easy-to-acquire biofluid samples such as urine and blood, the aforementioned confounding factors can easily overwhelm attempts to extract relevant information. This can mar potentially crucial applications such as biomarker discovery. As such, a new algorithm, called Selective Paired Ion Contrast (SPICA), has been developed with the intent of extracting potentially biologically relevant information from the noisiest of metabolomic datasets. The basic idea of SPICA is built upon redefining the fundamental unit of statistical analysis. Whereas the vast majority of algorithms analyze metabolomics data on a single-ion basis, SPICA relies on analyzing ion-pairs. A standard metabolomic data set is reinterpreted by exhaustively considering all possible ion-pair combinations. Statistical comparisons between sample groups are made only by analyzing the differences in these pairs, which may be crucial in situations where no single metabolite can be used for normalization. With SPICA, human urine data sets from patients undergoing total body irradiation (TBI), and from a colorectal cancer (CRC) relapse study were analyzed in a statistically rigorous manner not possible with conventional methods. In the TBI study, 3530 statistically significant ion-pairs were identified, from which numerous putative radiation specific metabolite-pair biomarkers that mapped to potentially perturbed metabolic pathways were elucidated. In the CRC study, SPICA identified 6461 statistically significant ion-pairs, several of which putatively mapped to folic acid biosynthesis, a key pathway in colorectal cancer. Utilizing support vector machines (SVMs), SPICA was also able to unequivocally outperform binary classifiers built from classical single-ion feature based SVMs.

Current address and contact information of corresponding author: Biomolecular Measurement Division, National Institute of Standards and Technology, 100 Bureau Drive, Stop 8362 Gaithersburg, MD 20899-8362.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Introduction

The rise of metabolomics as a primary "−omics" platform in high throughput quantitative biology has enabled the exploration of biological systems at an unprecedented level of insight. With the capability to quantify thousands of small molecule signatures in a particular system, liquid chromatography (LC) coupled with mass spectrometry (MS) based untargeted metabolomics is a powerful tool for exploring and characterizing metabolic processes, as well as biomarker discovery.[1] However, there are both positive and negative aspects to the platform that make data analysis unique challenge.

The sensitivity and flexibility of the metabolomics platform vastly increases the range of sample types and sources from which samples can be acquired for analysis. Sample types such as urine, blood, cell lysates, feces, and saliva can easily be fed into the metabolomics workflow. Furthermore, biofluids, such as urine, can be sampled from mice and other small animal models at multiple time points without compromising survivability, unlike multiple blood draws. However, this flexibility can also introduce a myriad of confounding factors that were never an issue for platforms with more restrictive sample requirements, such as microarray based transcriptomics. While ostensibly an ideal sample type for analysis via metabolomics, urine samples from experiments utilizing animal models in ideal environmental and dietary conditions will result in metabolomics data that, by the standards of other −omics platforms, exhibit an exceptionally high degree of variability and fluctuation.[2] This is in large part due to the high sensitivity of the urine metabolome to virtually any stimulus, especially when analyzed via metabolomics. This problem is exacerbated when the experiment involves human subjects, where factors such as diet, environment, genotype, age, and sex cannot always be controlled for, especially when sample sizes are low.

These problems are compounded by several confounding characteristics that are inherent idiosyncrasies of metabolomics data. Raw LC-MS metabolomics data, in the form of chromatograms, must first undergo a pre-processing stage in which the chromatographic peaks are identified and selected in order to produce the more familiar postprocessed high dimensional quantitative data resembling outputs from other −omics platforms. A large part of the pre-processing stage involves mitigating issues such as retention time drift, proper peak alignment across multiple samples, and correcting for external environmental variables that may affect the results, such as room temperature fluctuations.[3] These factors can certainly affect the final postprocessed output, and add to the overall difficulty of analyzing metabolomics data. The postprocessed data itself poses a serious challenge for bioinformaticians due to a number of peculiarities. Variables in the data often have very different variances when compared to one another, making many classical biostatistical methods invalid due to their inherent assumption of equivariance. Perhaps the defining attribute of metabolomics data is the "missing" data issue, which is typically defined as a zero value in the relative abundance for a given ion.[4] While missing data is not a new problem, it is the magnitude and inexplicable pattern of this "missingness" that introduces new problems during analysis. Many mathematical procedures and operations simply fail during these circumstances, and standard solutions, such as value imputation, become questionable when the numbers of values that need to be imputed comprise such a large

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

fraction of the total data set. Taken together, these factors pose as serious obstacles when attempting to normalize a data set.

Nonetheless, the considerable potential of the metabolomics platform as a tool for non-hypothesis driven research and biomarker discovery necessitates the development of specialized algorithms that are robust enough to extract potentially biologically relevant information from its noisy and idiosyncratic data sets. The current repertoire of techniques and algorithms for analyzing these data rely on classical univariate and multivariate procedures, such as standard statistical tests, principal component analysis (PCA), and orthogonal projections to least squares (OPLS). Workflows that have been developed for handling metabolomics data sets, including MetaboAnalyst,[5] as well as MetaboLyzer,[6] which our group recently developed, rely on these standard techniques, which are not explicitly designed for the data, and may not fully exploit its unique characteristics. As such, a new algorithm, called Selective Paired Ion Contrast Analysis (SPICA), has been developed for the express purpose of analyzing metabolomics data, taking into full account its advantages and shortcomings, so that potentially biologically relevant information can be extracted even from the most inscrutable of data sets.

SPICA is built upon utilizing pairs of ions as the fundamental unit of statistical analysis, rather than individual ions. All possible ion-pair combinations are first generated from the ion features in a dichotomous data set. Statistically significant differentiating ion-pair features are then identified when comparing data from the control versus treated sample groups. By conducting analysis in this pairwise fashion, numerous potential normalization issues are mitigated, as well as exposing possible latent structures in the data that would otherwise be missed when only analyzing the data in the traditional single-ion fashion. Biological interpretation of ion-pair features is conducted by augmenting traditional metabolic pathway analysis techniques for this new paradigm. Analysis was conducted on a radiation metabolomics study that analyzed urine samples collected from cancer patients before and after undergoing total body irradiation (TBI), as well as a separate study that compared differences between urine samples from colorectal cancer (CRC) patients who eventually relapsed versus those who did not. In both cases SPICA was able to identify numerous statistically significant ion-pair features that putatively mapped to metabolic pathways directly linked to the stressors. When adapted into a support vector machine (SVM) based binary classifier, SPICA was able to unequivocally outperform its single-ion based counterpart.

Methods and Tools

The prototype code for the SPICA algorithm and all supporting scripts were written in Python. The code implementation relies upon numerous open source libraries and tools, namely SciPy,[7] RPy2,[8] Matplotlib,[9] and PyOpenCL.[10] Numerous statistical tests implemented in SciPy were utilized in the SPICA code. Via RPy2, many functions and libraries in the R statistical computing environment[11] were also used, namely Kernlab[12] and ROCR.[13] Matplotlib was utilized for its graphing capabilities. OpenCL,[14] via PyOpenCL, was instrumental in making several components of SPICA feasible by reducing calculation times. SPICA is open source and is freely available at https://sites.google.com/a/

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

georgetown.edu/fornace-lab-informatics/home/spica along with detailed installation instructions (Supplement 2).

All urine samples in both the TBI and CRC data sets were stored at −80 °C and analyzed utilizing Ultra Performance Liquid Chromatography coupled to time-of-flight mass spectrometry, utilizing a Waters Corporation QTOF Premier. Samples were run in both positive and negative ionization modes. The TBI study chromatograms were preprocessed with MarkerLynx (Waters, Inc.), while the CRC study chromatograms were preprocessed with XCMS.[3]

Initial Data Reprocessing

SPICA's novel reinterpretation and reorganization of a standard dichotomous metabolomic data set (e.g. control vs. treated) begins with an initial data reprocessing stage (illustrated in Figure 1). This step precedes the ion-pair reinterpretation stage, and functions as an initial filtering and standardization step. These procedures are necessary for reducing the intrinsic variability and noise found in metabolomic data, and serve in laying a foundation for ion-pair formation. While most of these procedures are standard, one novel component involves merging the positive and negative mode data sets, which are generated from the positive and negative electrospray ionization modes used during LC-MS based metabolomics data acquisition, into a unified data frame. This maximizes the number of derived ion-pairs, and eliminates the unnecessary dichotomization of the overall data. The initial preprocessing stage is multistep, and involves data-wide transformation, normalization, and filtering procedures based on both statistical and putative biochemical properties of the ions.

The reprocessing workflow begins with filtering out all ions in both ionization modes that are missing in a user-defined percentage (a zeros threshold Z_{thr}) of the samples in at least one of the data subsets (e.g. control or treatment). This categorizes the ions as either partial-presence, wherein the ions appear above the Z_{thr} in only one set, or complete-presence, wherein the ions appear above the Z_{thr} in both sets. Data for both the complete- and partial-presence ions is then log transformed (base *e*) and Gaussian normalized, which involves performing a statistical Z standardization, wherein the estimated mean and standard deviation parameters used during standardization are estimated on a per-sample basis, and only utilizing the complete-presence data for estimation. Both the zeros threshold filtering and the Gaussian normalization procedures are discussed at length in our previous work.[6] These procedures are executed independently for the positive and negative mode data sets, meaning that the Z standardization parameters are derived on a per-sample as well as a per-mode basis. The data are then simply concatenated into a unified set comprising data from both positive and negative ions, which, to reiterate, have been transformed and normalized based on their originating ionization mode data set. As a final filtering step, ions may be excluded based on their assigned putative biochemical identities via integration of the Kyoto Encyclopedia of Genes and Genomes (KEGG),[15] the Human Metabolome Database (HMDB),[16] and the BioCyc small molecule databases.[17] Putative identity assignment is discussed in our previous work.[6] The prototype SPICA code implements both a porous and restrictive rule set for filtering, and is detailed in Supplement 1.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Ion-Pair Reinterpretation

The fundamental concept behind SPICA is the novel reinterpretation and translation of standard metabolomics data into a paired feature paradigm. These paired features are generated by exhaustively generating every possible ion-pair combination in a standard metabolomic data set. For instance, if there are 1,000 ions in a data set, then 499,500 ion-pair features will be generated. All subsequent analysis is conducted only through the meta-data generated from this process. The reinterpreted data for these ion-pairs are constructed as either continuous or discrete features, and undergo a rigorous statistical filtering procedure that evaluates for missingness, normality, and outliers. While continuous features intrinsically contain more information, discrete features are by nature more statistically robust, allowing for information to be extracted from otherwise untenable sources. It is this heterogeneous model (illustrated in Figure 2), utilizing both continuous and discrete features, which allows SPICA to fully exploit the advantages of parametric and non-parametric statistics.

Initially, the complete-presence ion data from the transnormalized data set produced in the Initial Data Reprocessing step is utilized to construct the continuous ion-pair features. The continuous ion-pair feature ($\delta_{ionA|ionB}$) for any two complete-presence ions (*ionA*, *ionB*) in a given sample (*X*) is calculated by the signed arithmetic difference between their associated transnormalized abundance values ($Ab_{ionA,X}$, $Ab_{ionB,X}$):

$$\delta_{ionA|ionB,X} = Ab_{ionA,X} - Ab_{ionB,X}$$

If either abundance value is missing in the sample, then $\delta_{ionA|ionB,X}$ cannot be calculated, and is considered missing. The delta values for all complete-presence ion-pair combinations are initially calculated for all samples in both data subsets (e.g. control and treated), and then filtered based on missingness, utilizing the same user-defined zeros threshold Z_{thr} in the previous reprocessing procedure. This is necessary because any two complete-presence ions (which by definition have non-missing abundance values in at least Z_{thr} percent of the samples in both data subsets) can easily produce an ion-pair feature that has a missingness that is below Z_{thr} due to the computational requirement that both ions must be non-missing within the same sample. Only ion-pair features that are above the Z_{thr} are kept, while the remaining features are separated for possible discretization in the next phase. The continuous ion-pair feature set is further reduced via outlier filtering. This process involves removing features whose constituent data loss after outlier removal (via non-parametric 1.5 interquartile range filtering) exceeds a user-defined percentage (L_{thr}) in either subset. Features that are removed at this stage are again kept for possible discretization. Finally, the remaining features may be subjected to an omnibus test for normality, depending on whether the Welch's t-test for statistical significance will be used during the Data Comparison step. If this option is selected, features will be further reduced (and kept for eventual discretization) based on the outcome of the omnibus test (and the associated p-value threshold). All remaining continuous ion-pair features are considered fit for analysis in the Data Comparison step.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Discrete ion-pair features are constructed from both the partial- and complete-presence ion subsets, as well as from the continuous ion-pair features that were filtered out in the previous phase. Discrete ion-pairs are constructed by pairing partial-presence ions with other partial-presence ions, but also by pairing with complete-presence ions as well. The discrete ion-pair feature ($\beta_{ionA|ionB}$) for any two ions (*ionA*, *ionB*) in a given sample (*X*) is a binary variable determined by which associated transnormalized abundance value ($Ab_{ionA,X}$, $Ab_{ionB,X}$) is greater:

$$\beta_{ionA|ionB,X} = \begin{cases} 1, & Ab_{ionA,X} > Ab_{ionB,X} \\ -1 & Ab_{ionA,X} < Ab_{ionB,X} \end{cases}$$

This equation also serves to discretize the continuous ion-pair features discarded from the previous phase. Missing abundance values are given a value of negative infinity ($-\infty$), thus $\beta_{ionA|ionB,X}$ can still be calculated if either abundances are missing, but cannot be calculated if both are. Thus, $\beta_{ionA|ionB,X}$ is only considered missing if both abundance values are also missing, or in the extremely rare case where the abundance values are identical. Discrete ion-pair feature missingness serves as the basis for filtering, wherein missing $\beta_{ionA|ionB,X}$ values are considered outliers, and features are removed based on whether the percentage of data loss exceeds the user-defined L_{thr} cutoff in either data subset. All remaining discrete ion-pair features are kept for analysis in the Data Comparison step. Overall, this discrete ion-pair feature model may serve as a tenable solution for extracting useful information from "missing" portions of metabolomic data that are otherwise unused.

Data Comparison

The statistical procedure for comparing SPICA generated ion-pair features in two sample subsets follows a workflow (outlined in Supplemental Figure 1) utilizing classical biostatistical techniques as well as newer computational approaches designed for high throughput −omics data. The primary objective of this procedure is the identification of statistically significant ion-pair features that differentiate between the two sample subsets in a typical dichotomous data set. The workflow follows a standard procedure, which encompasses initial statistical testing, followed by multiple testing correction (MTC), and finally data visualization. However, the heterogeneous nature of the ion-pair features necessitates using old tools in new ways via the development of a modified principal component analysis (PCA) based procedure known as heterogeneous PCA for visualizing statistically significant ion-pair data.

Identifying statistically significant ion-pairs involves initial statistical testing followed by multiple testing corrections (MTC) procedures to control for the false-positive rate (Type I error). Continuous ion-pair features are evaluated for statistical significance either by the parametric Welch's t-test, or the non-parametric Kolmogorov-Smirnov (K-S) test. Discrete ion-pair features are evaluated by the Fisher's exact test. MTC is then conducted separately for continuous and discrete features. Either a Monte Carlo implementation of Westfall and Young's maxT step-down permutation resampling procedure,[18] or a standard false discovery rate (FDR) procedure[19] is used for MTC in both feature types. The permutation

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

resampling procedure is widely used in both genome wide association studies[20] and gene expression data analysis,[21] which place high emphasis on strongly controlling for Type I errors. However, this procedure may prove too conservative in some cases, producing far fewer results than desired, thus necessitating the use of an FDR based procedure. The prototype code implements a simple and widely used FDR step-up p-value adjustment method, which results in less stringent Type I error control, but greater statistical power.[22]

Data visualization is a common means by which statistically significant features in high-throughput biological data are qualitatively evaluated, with PCA being the most prevalent procedure in metabolomics. Heterogeneous PCA, a new data visualization procedure based on PCA, was developed in order to accommodate both continuous and discrete ion-pair features during analysis. Standard PCA procedures are only able to analyze continuous features. An alternative procedure, known as polychoric PCA, is able to incorporate both continuous as well as discrete features by constructing a pseudo-correlation matrix wherein the polychoric correlation ($r_{polychoric}$) is used between two discrete features, the polyserial correlation ($r_{polychoric}$) is used between a discrete and a continuous feature, and finally the traditional Pearson's correlation ($r_{pearson}$) is used between two continuous features.[23] However, using a correlation matrix for PCA does not allow for the contribution of variables to be differentially emphasized during analysis (i.e. weighting), thus a new procedure, called heterogeneous PCA, was developed. Heterogeneous PCA builds on polychoric PCA, and constructs a weighted pseudo-correlation matrix (outlined in detail in Supplement 1), which maximizes the usage of information available in both the discrete and continuous ion-pair features. Heterogeneous PCA is utilized for data visualization, as well as for dimensionality reduction during SVM based classification (detailed in Supplement 1).

Putative Metabolic Pathway Analysis

For the biologist, perhaps the most tangible way in which to examine the statistically significant ion-pair features identified in the Data Comparison step is by examining the metabolic pathways that are putatively associated with these features. As mentioned in the Initial Data Reprocessing step, the constituent ions of all ion-pair features are examined for any putative matches to biologically relevant small molecules and any associated metabolic pathways via KEGG and BioCyc (only KEGG results shown during analysis for brevity). With this information, a novel approach to conducting pathway analysis was developed, utilizing a conventional pathway analysis procedure[25] that has been augmented for ion-pair features. By necessity, only ion-pair features in which both constituent ions possess a putative identification and associated KEGG or BioCyc pathway are utilized. It is important to emphasize that all results are putative in nature, as they rely on putative ion identifications, and must eventually be verified via tandem mass spectrometry. Furthermore, more than one identity can potentially be assigned to an ion, and thus mapped to more than one pathway. In these cases, all identities (and the resulting combination of pairs) are separately considered during analysis. Analysis is separated into two stages (outlined in Figure 3), with the first stage focusing on ion-pair features in which both constituent ions putatively identify to the same pathway, and the second stage focusing on ion-pair features in which both ions identify to two different pathways.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The first stage of metabolic pathway analysis focuses on the subset of ion-pair features that map to only one pathway per feature (referred to as single-mapped features). Conceptually, these single-mapped features are the easiest to interpret from a biological standpoint, because a statistically significant ion-pair feature in which both constituent ions putatively identify to the same pathway implies that this pathway has been somehow perturbed. Quantifying the degree of the perturbation of a pathway relies on identifying all single mapped features, both statistically significant and non-significant, that map to it, and tabulating the fraction of this total mapped ion-pair set that is significant. This is essentially the classical approach to pathway analysis, as conducted in gene expression analysis. In following this methodology, SPICA utilizes a hypergeometric test for statistical significance in order to assign p-values to perturbed metabolic pathways, and conducts MTC via application of the false discovery rate (FDR) method. In doing so, perturbed pathways can then be ranked by p-value. Pathways possessing corrected p-values lower than a user-defined threshold ($\alpha_{pathway}$) can be interpreted as being perturbed beyond the statistical expectation, with respect to all other perturbed pathways analyzed. However, it is important to emphasize that a perturbed pathway with a p-value above the threshold does not imply that it is not potentially biologically relevant, only that it is not perturbed more than what is statistically expected. This procedure is conducted separately for both KEGG and BioCyc pathways, and the results are displayed in a simple bar graph.

The second stage of metabolic pathway analysis examines the ion-pair features that map to two different pathways, one for each constituent ion (referred to as dual-mapped features). The difficulty in biologically interpreting statistically significant dual-mapped features lies in the inability to identify which metabolic pathway has been perturbed, or if indeed both have been perturbed. The problem is the lack of a reference point, which is not an issue for single-mapped features. The approach to this problem involves grouping and analyzing dual-mapped features on a per-pathway basis. For a given metabolic pathway, all dual-mapped features in which either constituent ion maps to the pathway are aggregated. From this, the fraction of features that are statistically significant are then tabulated, from which a p-value quantifying the statistical significance of the perturbation of the pathway can be calculated using the hypergeometric test. Once all p-values have been calculated and aggregated for all pathways, MTC is conducted via FDR. These steps mirror the analysis of single-mapped features, and essentially provide the same results pertaining to pathway perturbation quantification. However, the nature of dual-mapped feature data allows for additional analysis to be conducted beyond basic pathway perturbation assessment. When dual-mapped features are instead grouped on a double pathway basis and then subsequently analyzed via FDR corrected hypergeometric testing, a distance matrix can be constructed based on this data. For two metabolic pathways (X, Y), elements ($d_{X,Y}$) in the distance matrix are assigned depending on the corrected hypergeometric p-value ($p_{X,Y}$) and the user defined p-value threshold ($\alpha_{pathway}$):

$$d_{X,Y} = \begin{cases} 1, & p_{X,Y} < \alpha_{pathway} \\ 0, & p_{X,Y} \geq \alpha_{pathway} \end{cases}$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

This distance matrix can then be analyzed with the classical multidimensional scaling (MDS) algorithm to create a 3-dimensional visualization, which maps each pathway as a point in Euclidean space. The arrangement of the pathways in the MDS plot is representative of their absolute perturbation as well as their perturbation in relation to one another. The farther a pathway is from the center of the MDS plot, the greater degree of its perturbation, but even more importantly, the spatial relationship between pathways is indicative of their differential perturbation with respect to one another. This gives the investigator a more complete depiction of the multifaceted interactions between pathways, instead of looking at pathway perturbation on a one-dimensional level.

Analysis of Total Body Irradiation Data

The challenges in studying the effects of radiation exposure in biological systems primarily stem from the intrinsically stochastic nature of the stressor.[25] These problems can be compounded in high-sensitivity platforms such as metabolomics. SPICA was used to analyze a data set from a previously reported radiobiology study,[26] which consisted of urine samples collected from 36 patients undergoing total body irradiation (TBI) collected before, and 6 hours post-exposure to a single dose of 125 cGy of radiation. A zeros threshold (Z_{thr}) of 0.50 was used, on the qualitative basis that in the worst-case scenario, a continuous ion-pair feature would be present, and therefore applicable for at least 50% of the samples in both groups. A practical outlier removal threshold (L_{thr}) of 0.9 was utilized, which excludes a feature if more than 90% of the data is found to be an outlier via 1.5 IQR filtering. A porous biological filtering rule set was also utilized for removing irrelevant ions. SPICA was able to identify 33 statistically significant ($P < 0.05$, step-down permutation corrected) continuous ion-pair features via the K-S test (example shown in Supplemental Figure 2A), and 3497 statistically significant ($P < 0.05$, step-down permutation corrected) discrete ion-pair features via the Fisher's exact test (example shown in Supplemental Figure 2B). These features were utilized to construct a heterogeneous PCA scores plot, shown in Figure 4A, which illustrates a very strong separation between the pre-exposure (red circles) versus the post-exposure (blue triangles) samples. When comparing these results with the PCA scores plot (Figure 4B) generated from a traditional single-ion approach (see Supplement 1 for more detail on parameters) for identifying statistically significant features via the K-S test ($P < 0.05$, uncorrected), the qualitative separation is markedly more difficult to discern. Notably, no multiple testing correction was conducted for identifying the 307 statistically significant single-ion features, as it would have reduced the significant feature count to impractically small numbers for any appreciable analysis. Thus, in comparison to SPICA, the single-ion results may exhibit a far higher false positive rate.

Putative metabolomic pathway analysis was also conducted on the 3530 statistically significant features, which yielded a number of significantly perturbed KEGG pathways via single-mapped (Figure 5A) and dual-mapped (Figure 5B) feature analysis. The graphs plot the - 1*log10 transformed p-values, both uncorrected (blue bars) and FDR corrected (red bars), via the hypergeometric test for significance. The yellow line is the threshold for a FDR of 0.1. Analysis of both the single-mapped features, i.e. ion-pairs in which both constituent ions map to the same metabolic pathway, as well as the dual-mapped features, i.e. ion-pairs in which the ions map to different pathways, yielded lysine biosynthesis

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(ko00300) and lysine degradation (ko00301) as statistically significant ($P < 0.10$, FDR corrected). Poly-L-lysine has been linked to hematopoietic stem cell differentiation, and amino acids in general play a role in forming free radicals as a result of ionizing radiation exposure.[27] Analysis of the dual-mapped features also yielded folate biosynthesis (ko00790) as being statistically significant, which has been directly linked to radiation exposure.[28] Furthermore, an MDS plot (Figure 5C) produced from analysis of dual-mapped features indicates that the folate biosynthesis pathway is differentially perturbed when compared to the perturbation of the two lysine pathways, which are nearly identical (most likely the result of many shared ion-pair features). This may suggest that the ion-pair features mapped to the folate pathway may reveal different underlying mechanisms from that of the lysine pathways.

Finally, a Monte Carlo cross validation (MCCV) procedure was conducted on the TBI data utilizing a novel support vector machine (SVM) based SPICA derived prediction algorithm called Adaptive Paired Ion Contrast Classification Analysis, or APICCA (detailed in Supplement 1). Approximately 10% of the samples were randomly removed from the original data set, while the remaining 90% were utilized to train APICCA. The 10% would then be classified to evaluate the accuracy of the prediction model. To reduce computation time, the top 1000 statistically significant (uncorrected p-value < 0.01) continuous ion-pair features ranked by K-S derived p-values, and the top 1000 statistically significant (uncorrected p-value < 0.01) discrete features ranked by Fisher's exact test were utilized for classification. To evaluate the efficacy of APICCA, a traditional single-ion based SVM classifier utilizing filtering and normalization procedures that mirror SPICA's workflow was utilized (detailed in Supplement 1). A K-S test was also used for selecting statistically significant ($P < 0.01$, uncorrected) single-ion features. The data was resampled 100 times, from which a total of 800 predictions from each classifier were used to construct Receiver Operating Characteristic (ROC) curves, which plot the sensitivity (true positive rate) versus the specificity (false positive rate) of a classifier (Figure 5D). From the ROC curves, it is apparent that APICCA (AUC 0.924) outperforms the baseline model (AUC 0.817) by a wide margin, and indicates that the predictive capability may potentially be leveraged in practical scenarios such as identifying exposed individuals in radiological emergencies for triage and treatment.

Analysis of Colorectal Cancer Relapse Data

Identifying the mechanisms of cancer recurrence is especially difficult, given that the samples must originate from human cohort studies with numerous potential confounding factors. SPICA was used to analyze a colorectal cancer (CRC) data set consisting of 20 non-relapse urine samples and 20 relapsed urine samples collected from patients at the time of surgery, prior to any treatment, and with a minimum follow up time of five years.[29] Many of the same parameters used during analysis of the TBI data were used, including a Z_{thr} of 0.50, an L_{thr} of 0.9, and a porous biological rule set, however the FDR procedure was used for MTC instead of permutation, as the latter was too conservative, and did not produce adequate results. SPICA was able to identify 590 statistically significant ($P < 0.05$, FDR corrected) continuous ion-pair features via the K-S test and 5871 statistically significant ($P < 0.05$, FDR corrected) discrete ion-pair features via the Fisher's exact test. These features

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

were utilized to construct a heterogeneous PCA scores plot, shown in Figure 6A, which indicates a strong clustering of non-relapse samples (red circles) versus the comparatively spread out distribution of the relapse (blue triangles) samples. This suggests a lower variance in the statistical distributions of the significant features for non-relapse cases, in contrast to a higher variance in the distributions of these same features for relapsed patients. As with the TBI data, the PCA scores plot (Figure 6B) generated from a traditional single-ion feature approach exhibits a much lower qualitative degree of separation between the two sample groups. Again, no multiple testing correction was utilized in identifying the 236 ($P < 0.05$, uncorrected) statistically significant single-ions via the K-S test, as it would have reduced the number of results to impractical levels.

Putative metabolic pathway analysis was conducted only on the dual-mapped ion-pair features, as there were insufficient single-mapped features for analysis. The results (Figure 7A) indicate that by far the most significantly perturbed KEGG pathway was folate biosynthesis (ko00790). Increased dietary folate intake has been strongly linked with a decrease in CRC risk.[30] Other pathways found to be statistically significant ($P < 0.10$, FDR corrected) include propanoate metabolism (ko00640),[31] as well as GABAergic synapse (ko04727),[32] sphingolipid metabolism (ko00600),[33] and several other pathways, many of which have been identified in the literature as being associated with colorectal cancer. The MDS plot (figure 7B) suggests that the perturbation of the folate biosynthesis pathway is distinctly different from the perturbation of other pathways, and also suggests that tryptophan metabolism (ko00380), not seen in the more traditional pathway analysis procedure, may be significant, and indeed the literature suggests that tryptophan levels are associated with quality of life in CRC patients.[34]

As with the TBI data set, MCCV was also conducted with the CRC data via APICCA. The same parameters used during validation with the TBI data set were used for the CRC data as well, for both APICCA and the baseline model. Due to the lower sample count, cross validation was repeated 200 times instead, for a total of 800 predictions, which were utilized in constructing an ROC curves (Figure 7C). These curves again show APICCA (AUC 0.891) outperforming the baseline model (AUC 0.833) by a considerable, if less impressive margin than in the TBI data.

Discussion

There is a categorical need for algorithms specifically designed for the burgeoning field of metabolomics, and SPICA is one of the first attempts at developing highly specialized tools for postprocessed LC-MS metabolomic data sets. Its ability to incorporate the idiosyncrasies of metabolomic data sets, rather than simply working around it, sets it apart from the techniques that have been utilized thus far. Furthermore, SPICA represents a comprehensive workflow that provides both exploratory capabilities, via its heterogeneous PCA implementation and its putative pathway analysis, as well as predictive capacity, via the SVM based APICCA. At its core, SPICA's ion-pair paradigm is not limited to the analytical procedures used in the workflow described herein, and can be used in tandem with any statistical and computational technique for conducting data analysis, potentially with

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

application to other "−omics" data as well.. These aspects make SPICA a powerful tool that serves to enrich the metabolomics platform as a whole.

SPICA's defining characteristic of using paired features in lieu of the customary single feature based analysis allows it to more fully utilize a metabolomic data set during analysis. Whereas traditional single feature methods will stumble when there is too much missing data, SPICA's ability to interpret its paired features as discrete variables allows information to be extracted even when the available data is sparse. For instance, in a dichotomous data set with 10 samples each in the control and treated groups, if an ion's abundance value is only non-missing in 2 of the control samples and in 6 of the treated samples, traditional statistical hypothesis testing would fail due to insufficient data. SPICA, however, would pair this ion with another ion feature whose data is not as sparse, converting it into a discrete ion-pair feature that may potentially yield relevant information. We speculate that it is primarily this efficacy in data utilization that allows APICCA to outperform the baseline single-ion procedure, rather than the characteristics afforded by utilizing ion-pair features. Nonetheless, it is this pairwise approach that facilitates maximizing the extraction of relevant information from a data set. This underscores the need to develop specialized algorithms for metabolomics data, as many valuable insights can be overlooked from using all-purpose tools for analysis.

Perhaps a less obvious benefit of SPICA's paired feature approach is its ability to circumvent normalization issues that often plague metabolomics studies. For instance, the sample-to-sample variation in urine concentration levels may be a major confounding factor in a study. Even minor pipetting errors during sample preparation for less variable sample types such as cell lysate extracts can cause unforeseen issues during data analysis. While tried and true procedures such as TIC normalization may alleviate the issue, SPICA's use of ion-pair features bypasses the problem due to the fact that only the difference between abundance values for any two given ions is analyzed, rather than the absolute abundance values themselves. Thus, inter-sample concentration variations are effectively removed from consideration, as absolute abundance values are never directly compared to one another during SPICA's analysis. A common normalization technique for urine metabolomics involves normalizing all abundance values by creatinine, which is one of the most abundant metabolites in urine, and a good indicator of kidney function. In a sense, it is a form of ion-pair analysis, as every ion is paired to creatinine. SPICA essentially generalizes this concept by "normalizing" every ion by every other ion in the sample via exhaustively generating all possible ion-pair features. SPICA's normalization property may be crucial in some sample types, such as serum, saliva, and many other biofluids where there is no obvious or biologically relevant metabolite that can be used for normalization.

The advancement of the field of metabolomics through the development of specialized informatics techniques is just as crucial as its other aspects. It is through such tools that metabolomics can mature into a more standardized platform that can gain a wider user base. The development of SPICA, a novel methodology expressly for postprocessed LC-MS metabolomics data, has been predicated on these goals.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Conclusion

The development of SPICA represents a concerted effort to create novel algorithms expressly for postprocessed metabolomics data, which are designed to address many of the obstacles that make analysis difficult. SPICA specifically addresses the issue of “noisy” data that is especially exacerbated for metabolomics, which is due in to the platform’s high sensitivity and capability to process a far wider gamut of biological sample types. Taken together, these techniques we have proposed represent a comprehensive workflow that provides both exploratory and classification capabilities which encompasses many of the goals for which metabolomics was initially envisioned to fulfill. While the concept of an ion-pair may be more difficult to understand than a single ion, we have effectively demonstrated that augmenting traditional methods of pathway analysis can be used to produce potentially biologically relevant results that are palatable to investigators without a strong informatics background. With metabolomics, and “−omics” platforms in general being touted as paving the way for non-hypothesis driven biomedical research, SPICA’s ability to translate often inscrutable high-throughput data into comprehensible results with biological meaning in a statistically rigorous manner is crucial for furthering these goals.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the Grants U19AI067773 and R01AI101798; the NIAID Grant U19AI067773 was crucial in supporting this effort. George Luta, John Moult, Subha Madhavan, and Anton Wellstein provided vital guidance and consultation in the development of the approaches. Subha Madhavan also provided support for the analysis of the CRC data set. The project was also supported by Award P30CA051008 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

References

1. Daviss B. The Scientist. 2005; 19:25–28.
2. Scalbert A, Brennan L, Fiehn O, Hankemeier T, Kristal BS, van Ommen B, Pujos-Guillot E, Verheij E, Wishart D, Wopereis S. Metabolomics. 2009; 5:435–458. [PubMed: 20046865]
3. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. Analytical chemistry. 2006; 78:779–787. [PubMed: 16448051]
4. Hrydziuszko O, Viant MR. Metabolomics. 2012; 8:161–174.
5. Xia J, Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS. Nucleic acids research. 2012; 40:W127–W133. [PubMed: 22553367]
6. Mak TD, Laiakis EC, Goudarzi M, Fornace AJ Jr. Analytical chemistry. 2013; 86:506–513. [PubMed: 24266674]
7. Jones E, Oliphant T, Peterson P. J. 2001 http://www.scipy.org/.
8. Gautier L. 2008
9. Hunter JD. Computing in Science & Engineering. 2007:90–95.
10. Klöckner A, Pinto N, Lee Y, Catanzaro B, Ivanov P, Fasih A. Parallel Computing. 2012; 38:157–174.
11. Team RDC. R Foundation Statistical Computing. 2008
12. Karatzoglou A, Smola A, Hornik K, Zeileis A. 2004

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

13. Sing T, Sander O, Beerenwinkel N, Lengauer T. Bioinformatics. 2005; 21:3940–3941. [PubMed: 16096348]
14. Group KOW. A. Munshi, Ed. 2008
15. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. Nucleic acids research. 2012; 40:D109–D114. [PubMed: 22080510]
16. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S. Nucleic acids research. 2007; 35:D521–D526. [PubMed: 17202168]
17. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M. Nucleic acids research. 2010; 38:D473–D479. [PubMed: 19850718]
18. Westfall, PH.; Young, SS. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley-Interscience; 1993. p. 360
19. Benjamini Y, Hochberg Y. Journal of the Royal Statistical Society. Series B (Methodological). 1995:289–300.
20. Johnson RC, Nelson GW, Troyer JL, Lautenberger JA, Kessing BD, Winkler CA, O'Brien SJ. BMC genomics. 2010; 11:724. [PubMed: 21176216]
21. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistica sinica. 2002; 12:111–140.
22. Shaffer JP. Annual review of psychology. 1995; 46:561–584.
23. Kolenikov S, Angeles G. Chapel Hill: Carolina Population Center, University of North Carolina. 2004
24. Evangelou M, Rendon A, Ouwehand WH, Wernisch L, Dudbridge F. PloS one. 2012; 7:e41018. [PubMed: 22859961]
25. Brenner DJ, Doll R, Goodhead DT, Hall EJ, Land CE, Little JB, Lubin JH, Preston DL, Preston RJ, Puskin JS. Proceedings of the National Academy of Sciences. 2003; 100:13761–13766.
26. Laiakis EC, Mak TD, Anizan S, Amundson SA. Radiation. 2014
27. Park K-S, Ahn J, Kim JY, Park H, Kim HO, Lee S-H. Tissue Engineering Part A. 2014Schaich KM, Pryor WA. Critical Reviews in Food Science & Nutrition. 1980; 13:189–244. [PubMed: 6254726]
28. Batra V, Devasagayam TPA. Food and Chemical Toxicology. 2012; 50:464–472. [PubMed: 22154853] Leopardi P, Marcon F, Caiola S, Cafolla A, Siniscalchi E, Zijno A, Crebelli R. Mutagenesis. 2006; 21:327–333. [PubMed: 16950805]
29. Madhavan S, Gusev Y, Natarajan TG, Song L, Bhuvaneshwar K, Gauba R, Pandey A, Haddad BR, Goerlitz D, Cheema AK. Frontiers in genetics. 2013; 4
30. Sanjoaquin MA, Allen N, Couto E, Roddam AW, Key TJ. International Journal of Cancer. 2005; 113:825–828.
31. Hague A, Elder DJE, Hicks DJ, Paraskeva C. International Journal of Cancer. 1995; 60:400–406.
32. Kleinrok Z, Matuszek M, Jesipowicz J, Matuszek B, Opolski A, Radzikowski C. Journal of physiology and pharmacology: an official journal of the Polish Physiological Society. 1998; 49:303–310. [PubMed: 9670113]
33. Berra B, Colombo I, Sottocornola E, Giacosa A. European journal of cancer prevention. 2002; 11:193–197. [PubMed: 11984139]
34. Huang A, Fuchs D, Widner B, Glover C, Henderson DC, Allen-Mersh TG. British journal of cancer. 2002; 86:1691–1696. [PubMed: 12087451]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

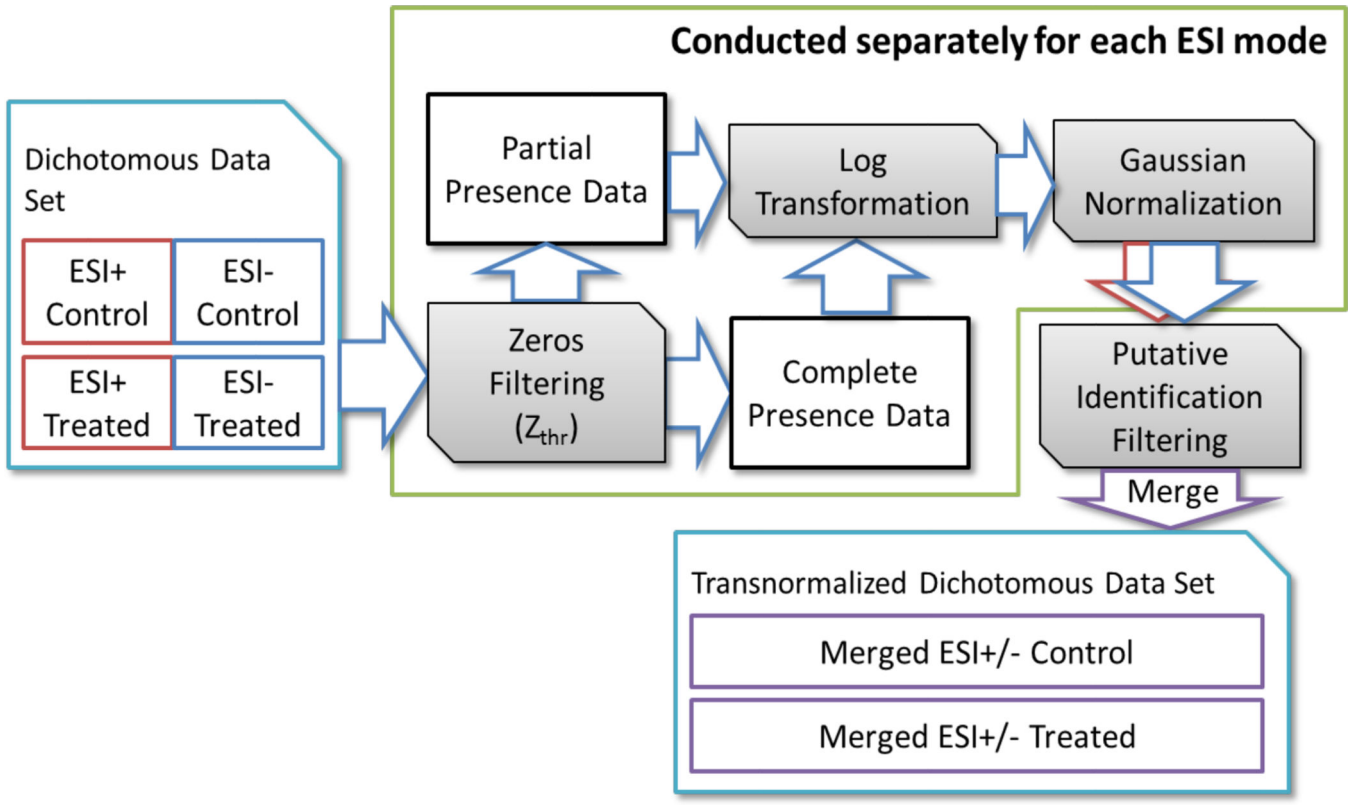

Figure 1.

The workflow of SPICA's Initial Data Reprocessing stage, which conducts a preliminary attenuation of potentially confounding factors in the data. All Ion features of a dichotomous data set are initially categorized as either partial- or complete-presence features, and subsequently undergoes transformation and Gaussian normalization. Biological filtering based on the putative identification of the ion features is then conducted, and the data from both ESI modes is then merged into a unified transnormalized dichotomous data set

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

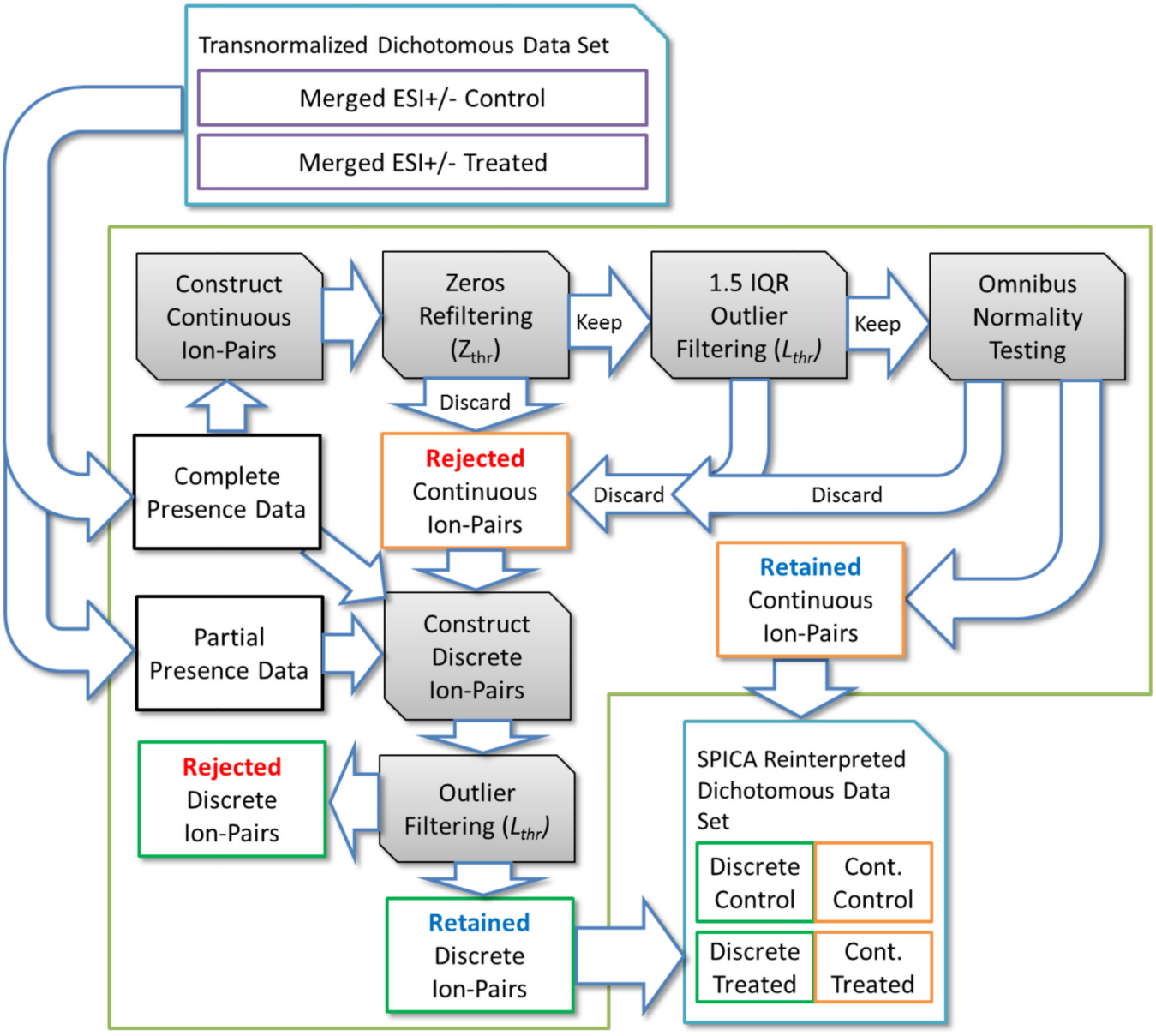

Figure 2.

The workflow for the Ion-Pair Reinterpretation stage of SPICA. Ion features from transnormalized dichotomous data, which has been categorized as either complete-presence or partial-presence, are exhaustively paired to form all possible ion-pair features. Pairs formed from two complete-presence ions are constructed as continuous ion-pair features, while all other pairing combinations are used to construct discrete ion-pair features. Continuous ion-pair features undergo a filtering process that examines the degree of missingness (zeros refiltering Z_{thr}), and outlier percentage (L_{thr}). Continuous features that pass are retained as continuous variables, but those discarded during this process are reinterpreted as discrete ion-pairs. All discrete ion-pair features also go through an outlier filtering process. The continuous and discrete ion-pair features that are ultimately retained by the end of this process are considered suitable for further analysis via SPICA.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

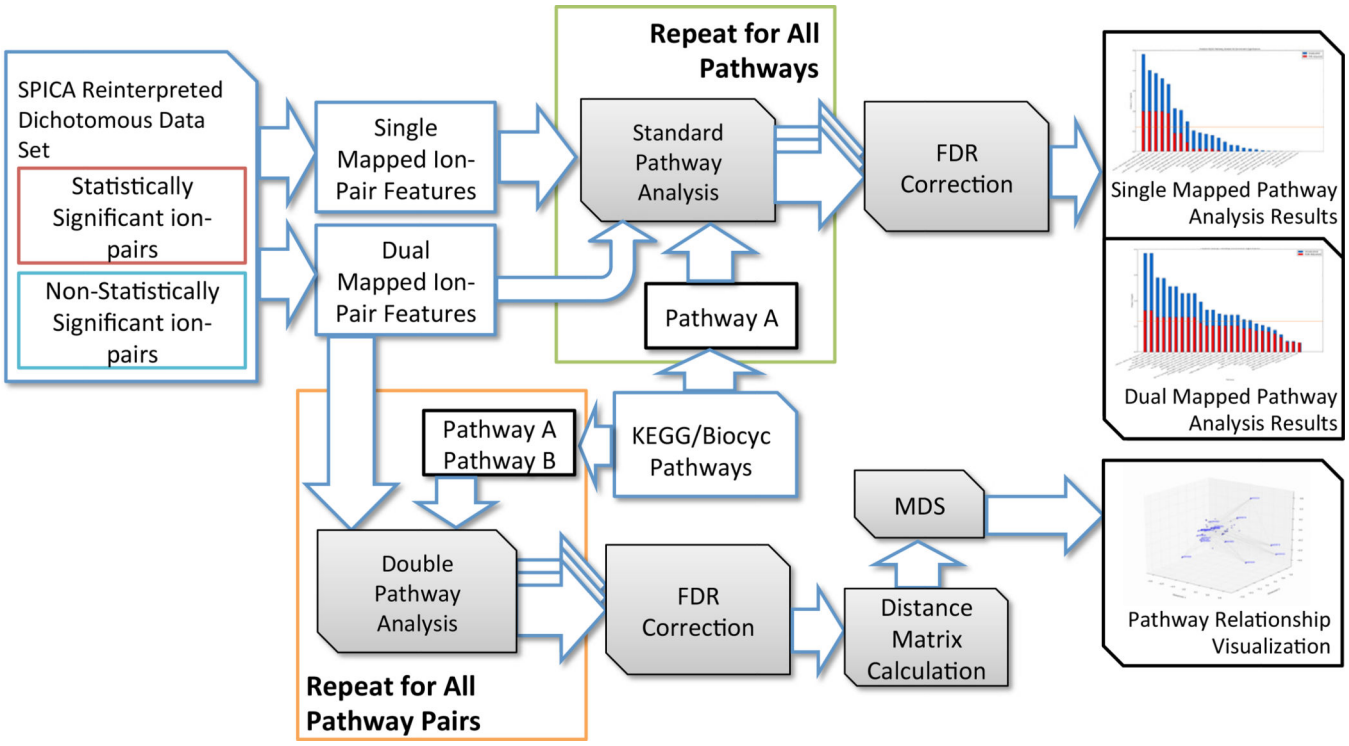

Figure 3.

The putative metabolic pathway analysis workflow. After statistical analysis has determined which ion-pair features are significant, all features are analyzed via putative ion identification as either single-mapped or dual-mapped. Both feature types can be utilized in a standard pathway analysis procedure (green box) that determines whether a metabolic pathway has a statistically greater than expected fraction of statistically significant ion-pair features mapped to it. Statistical significance for a pathway is determined via the hypergeometric test, and once all pathways are analyzed in this fashion, p-values are corrected via FDR. Results for the single- and dual-mapped features are displayed in separate bar graphs that plot the - 1*log10 of corrected and uncorrected p-values for each pathway. Dual-mapped features are further examined in a procedure (orange box) that is similar to the aforementioned pathway analysis, but instead determines whether there is a greater than expected fraction of significant ion-pair features for any two given pathways, instead of just one. This allows for a distance matrix to be constructed, and a MDS plot to be created from it, which allows for pathway perturbation to be visualized in relation to other pathways. This gives magnitude, as well as direction to a pathway's perturbation, where the previous analysis only supplied magnitude (via p-values).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

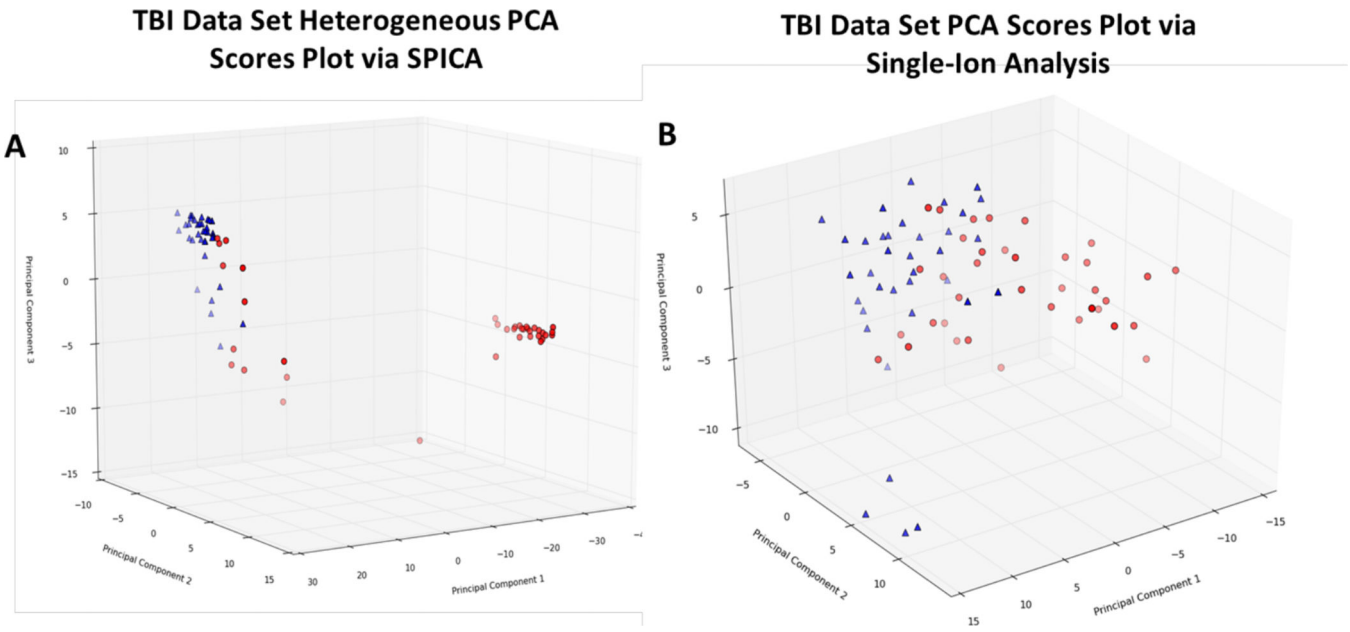

Figure 4.

PCA scores plots generated from the statistically significant features identified from two analyses of urine samples collected from 36 patients before and 6 hours after exposure to 125 cGy of total body gamma radiation (TBI). (A) The plot generated from SPICA's heterogeneous PCA indicates a very strong separation between the pre-exposure (red circles) and post-exposure (blue triangles) samples as defined by the 3530 statistically significant total ion-pair features. Both sample groups also exhibit a high degree of clustering as well. (B) Separation between the two groups and clustering within each group in the standard PCA plot is markedly more difficult to discern when conducting a traditional single-ion analysis, which identified 307 statistically significant single-ion features.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

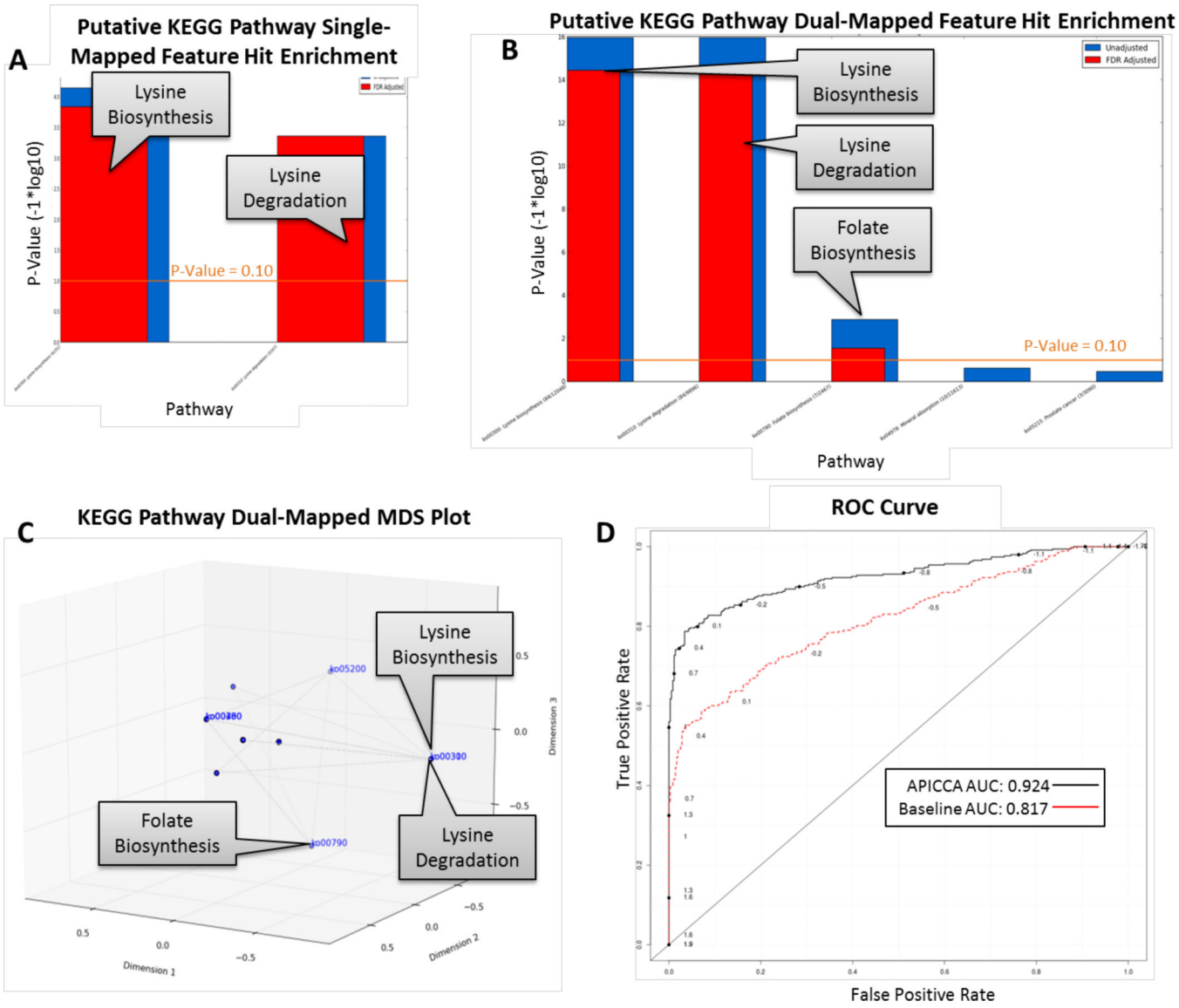

Figure 5.

Putative KEGG metabolic pathway analysis of both the single-mapped (A) and dual-mapped (B) ion-pair features of the TBI data yields lysine biosynthesis (ko00300) and degradation (ko00301) as being statistically significantly perturbed ($P < 0.1$, FDR corrected), though dual-mapped features reveal folate biosynthesis (ko00790) as being affected as well. Additional analysis of dual-mapped features produced an MDS plot (C) visualizing both magnitude and direction of perturbation, and reveals folate biosynthesis perturbation as being distinct from the perturbation direction of the lysine pathways. Finally, ROC curves (D) of the SPICA based SVM classifier (black line, AUC: 0.924) and a baseline single-ion based SVM classifier (red line, AUC: 0.817) demonstrate the powerful predictive potential of SPICA's ion-pair features, which unequivocally outperforms the traditional single-ion based classifier, and may even be strong enough for practical applications such as identifying exposed individuals in radiological emergencies.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

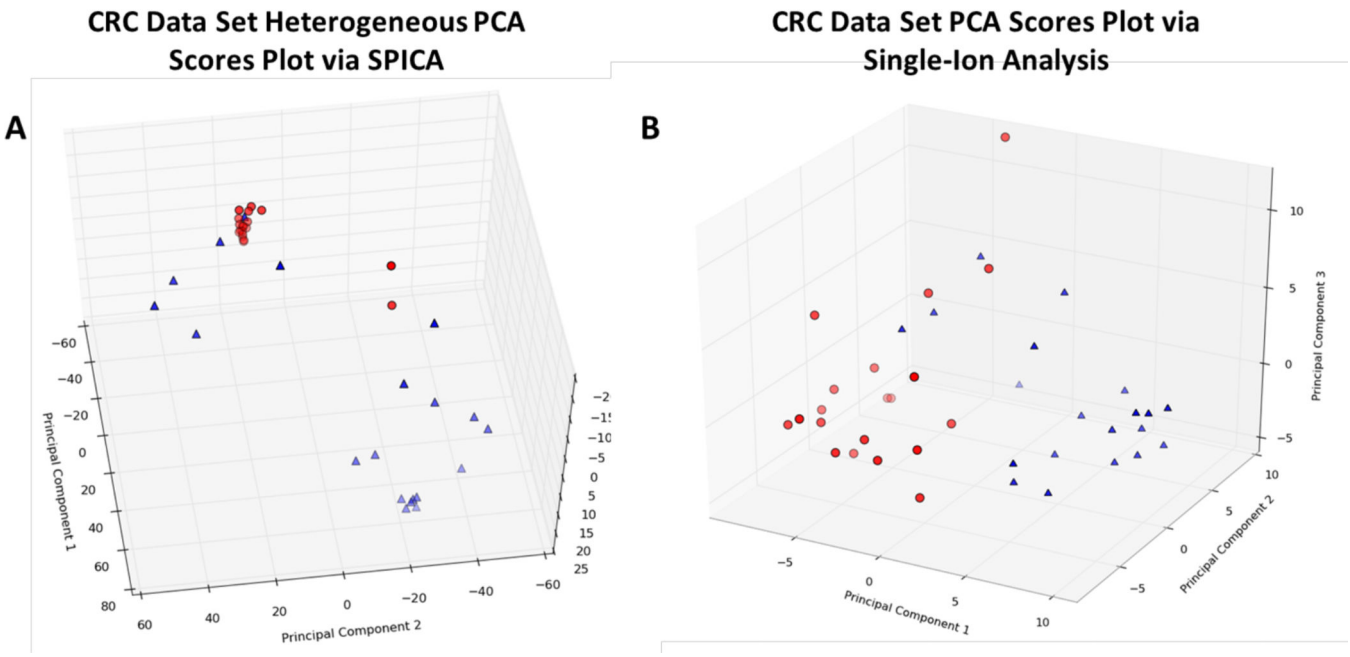

Figure 6.

PCA scores plots generated from the statistically significant features identified from two analyses of urine samples collected at the time of surgery from 20 colorectal cancer (CRC) patients with no recurrence, and 20 that eventually relapsed. (A) The plot generated from SPICA's heterogeneous PCA contrasts the strong clustering of the non-relapse samples (red circles) when compared to the relatively spread out relapse samples (blue triangles), which may be a valuable clue in uncovering the mechanisms behind CRC recurrence. Nonetheless, separation between the two groups is well defined. A total of 6461 statistically significant ion-pair features were identified during this analysis. (B) As with the TBI data, traditional analysis, which identified 236 statistically significant single-ion features, yielded unimpressive, but discernable separation between the two groups.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

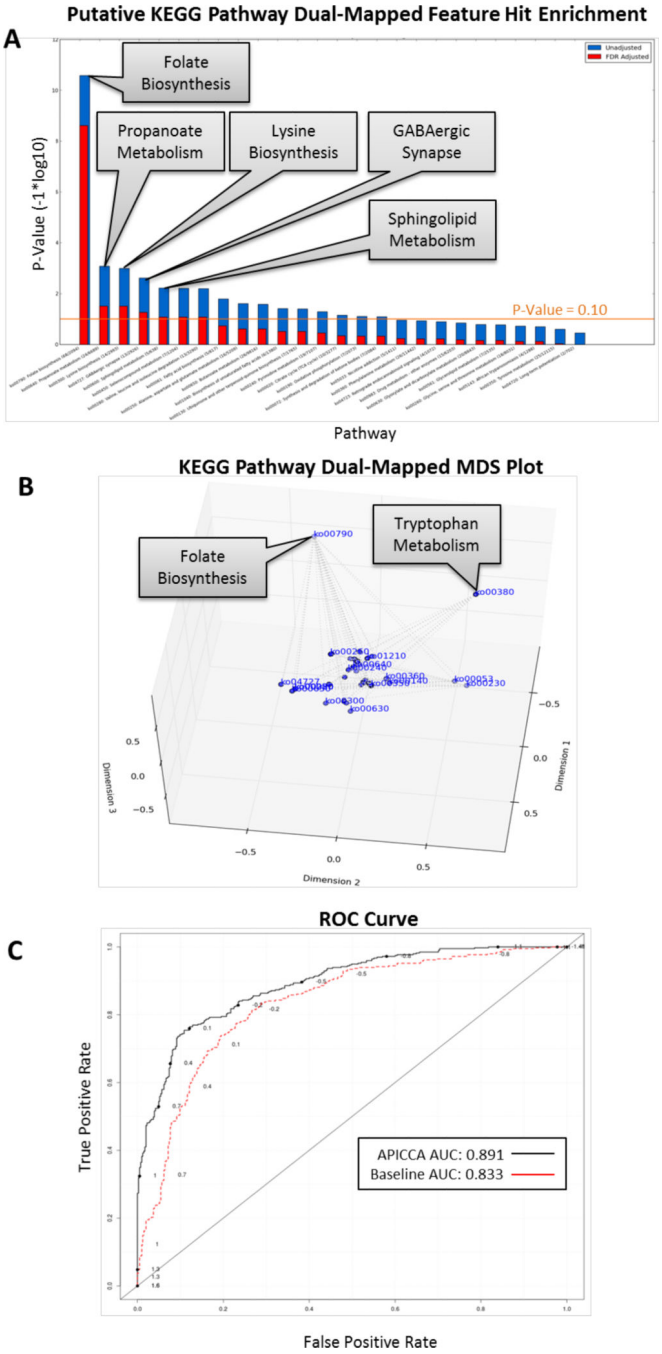

Figure 7.

Though there was an inadequate number statistically significant single-mapped ion-pair features for putative metabolic pathway analysis in the CRC data, dual-mapped features revealed several significantly perturbed ($P < 0.1$, FDR corrected) metabolic pathways (A) linked to colorectal cancer. The most significant pathway identified, folate biosynthesis (ko00790), is of particular interest, as a high folate diet has been strongly linked with a decrease in CRC risk. Additional analysis of dual-mapped features produced an MDS plot (B) visualizing both magnitude and direction of perturbation, and reveals an additional

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

pathway, tryptophan metabolism (ko00380), as potentially playing a role as well. ROC curves (C) of the SPICA based SVM classifier (black line, AUC: 0.833) and a baseline single-ion based SVM classifier (red line, AUC: 0.891) again demonstrates that the predictive potential of SPICA exceeds that of single-ions, though to a lesser degree than in the TBI data.