Published in final edited form as:
Anal Chem. 2014 January 7; 86(1): 506–513. doi:10.1021/ac402477z.

MetaboLyzer: A Novel Statistical Workflow for Analyzing Post-Processed LC/MS Metabolomics Data

Tytus D. Mak[1], **Evagelia C. Laiakis**[2], **Maryam Goudarzi**[2], and **Albert J. Fornace Jr.**[1,2]

[1]Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, New Research Building E504/508, 3970 Reservoir Rd, NW, Washington, DC 20057

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

Abstract

Metabolomics, the global study of small molecules in a particular system, has in the last few years risen to become a primary –omics platform for the study of metabolic processes. With the ever-increasing pool of quantitative data yielded from metabolomic research, specialized methods and tools with which to analyze and extract meaningful conclusions from these data are becoming more and more crucial. Furthermore, the depth of knowledge and expertise required to undertake a metabolomics oriented study is a daunting obstacle to investigators new to the field. As such, we have created a new statistical analysis workflow, MetaboLyzer, which aims to both simplify analysis for investigators new to metabolomics, as well as provide experienced investigators the flexibility to conduct sophisticated analysis. MetaboLyzer's workflow is specifically tailored to the unique characteristics and idiosyncrasies of postprocessed liquid chromatography/mass spectrometry (LC/MS) based metabolomic datasets. It utilizes a wide gamut of statistical tests, procedures, and methodologies that belong to classical biostatistics, as well as several novel statistical techniques that we have developed specifically for metabolomics data. Furthermore, MetaboLyzer conducts rapid putative ion identification and putative biologically relevant analysis via incorporation of four major small molecule databases: KEGG, HMDB, Lipid Maps, and BioCyc. MetaboLyzer incorporates these aspects into a comprehensive workflow that outputs easy to understand statistically significant and potentially biologically relevant information in the form of heatmaps, volcano plots, 3D visualization plots, correlation maps, and metabolic pathway hit histograms. For demonstration purposes, a urine metabolomics data set from a previously reported radiobiology study in which samples were collected from mice exposed to gamma radiation was analyzed. MetaboLyzer was able to identify 243 statistically significant ions out of a total of 1942. Numerous putative metabolites and pathways were found to be biologically significant from the putative ion identification workflow.

Introduction

Metabolomics has in recent years risen to become a popular platform for biological research. Utilizing various techniques of liquid chromatography (LC) in tandem with mass spectrometry (MS) technologies, metabolomics offers an unprecedented level of quantitative characterization of the metabolome via biofluid and tissue analysis[1]. A wide variety of biological samples can be utilized in metabolomics research, ranging from obvious sources such as urine and serum to less common origins such as feces. This flexibility in analysis, coupled with its abilities as an untargeted platform for non-hypothesis driven research make metabolomics particularly appealing. However, there are many obstacles that make entering the field very difficult for investigators new to the field.

Regardless of the source of the sample, data from metabolomics studies will almost always exhibit an exceptionally high degree of variability and fluctuation that make quantitative analysis a major challenge for bioinformaticians. Analysis begins with the preprocessing stage, in which the raw chromatograms produced in the metabolomics LC/MS workflow is deconvoluted into post-processed high dimensional quantitative data. Numerous software packages, both commercial and open source, such as Waters' MarkerLynx, XCMS[2], and MZmine[3], specialize in pre-processing chromatograms, which typically involves peak picking, alignment, integration, and normalization. The resulting post-processed data, which consists of high dimensional matrices resembling those from other high-throughput –omics fields, can then be statistically analyzed. Analyzing post-processed metabolomics datasets is difficult due to its many statistically confounding characteristics. For instance, variables in datasets often have highly dissimilar variances from one another, which contravenes a major assumption of many techniques used in biostatistics, that of equal variance amongst variables. Furthermore, there is the inherent issue of frequent and often inexplicable "missing" data points, which is defined as a zero value in the relative abundance for a particular ion, and is endemic to most metabolomics-derived datasets[4]. This "missingness" is not purely random, however, and depending on the dataset, can be potentially correlated with numerous observable and latent factors. The issue is compounded by the fact that many fundamental mathematical operations are simply not meant for handling this kind of data (such as logarithmic transformations). This renders many established statistical analysis techniques infeasible, or at the very least negatively impacts their efficacy. The primary methods of dealing with "missingness" thus far have been methods that attempt to impute the missing values, however one can argue that this only sidesteps the problem instead of incorporating this "missingness" as a fundamental feature of metabolomics data. Investigators new to the field of metabolomics who are unaware of these issues may attempt to blindly apply many standard statistical procedures without realizing that it may produce erroneous results.

Furthermore, even in its post-processed state, metabolomics data is difficult to translate into biologically relevant results. This is due to the instrumentation utilized in metabolomics, which conducts analysis by ionizing compounds, and identifying ions only by their mass over charge (*m/z*) and retention time. While this information offers valuable clues as to the true chemical identity of the ion via small molecule database searching, it is by no means a definitive process. Multiple putative matches can be elucidated from a single *m/z* value, and due to current limitations in LC/MS spectra databases, retention times for the most part cannot be leveraged to aid in identification. Definitive identification can only be conducted through a tedious validation procedure in which an ion in a sample is isolated and fragmented via LC-MS/MS and compared to the fragmentation spectra of the pure chemical compound that the ion is putatively identified as. The issue is compounded by the fact that a significant portion of the metabolome is simply unknown, meaning that many ions cannot be identified due to an incomplete knowledge of all metabolic pathways and processes[5]. These factors alone may deter investigators from attempting to enter the field of metabolomics entirely.

The staggering volume of quantitative data that metabolomics-based studies yield, coupled with issues that are unique to the field, emphasizes the need for specialized workflows that make analysis easier for new metabolomics investigators, and also expedites tedious steps for experts. The most common techniques for analyzing post-processed metabolomics data are multivariate analysis procedures such as principle component analysis (PCA) and orthogonal projections to latent structures (OPLS). These procedures are general statistical techniques that treat the data like any other high dimensional dataset, and do not specifically tailor to the unique characteristics and idiosyncrasies of metabolomics data. As such, a new statistical workflow software package, MetaboLyzer, has been developed with the express

purpose of addressing many of the aforementioned issues. MetaboLyzer incorporates and adapts some of the most common approaches and techniques in biostatistics, as well as several novel statistical procedures, into a comprehensive analytic workflow that also integrates rapid putative identification and associated metabolic pathway information aggregation. MetaboLyzer takes raw post-processed LC/MS metabolomic data as an input and outputs statistically significant and potentially biologically relevant results that are easy to interpret. The workflow is designed with consideration for individuals who are new to the field of metabolomics, but possess a basic understanding of biostatistics and data analysis techniques. Thus, in order to expedite analysis, we have incorporated into the workflow many default preset parameters and hints that we have found to produce reasonable results in most scenarios. At the same time, the workflow gives many options to deviate from these defaults, allowing more experienced investigators to tailor a more sophisticated and customized analysis. These aspects make MetaboLyzer a powerful statistical toolbox and a "one-stop-shop" for the majority of metabolomics data analysis.

Methods and Tools

MetaboLyzer was written in the Python programming language utilizing a large array of open source libraries and tools. MetaboLyzer heavily relies on the SciPy[6], RPy2[7], and Matplotlib[8] libraries for many of its statistical calculations and visualization procedures. Via RPy2, MetaboLyzer also utilizes many functions and libraries available in the R statistical computing environment[9] including *Kernlab*[10], *gplots*[11], *fastICA*[12], and *MASS*[13]. The *Kernlab* package was utilized for kernel PCA procedures, and the *gplots* library was utilized for all heatmap construction procedures.

MetaboLyzer was developed under Ubuntu 12.04.2 LTS and requires a computer with a minimum of 8 GB of RAM and a 2.00 GHz quad core x86 microprocessor. MetaboLyzer has not been tested on any other platform except for Debian and Red Hat based Linux distributions. MetaboLyzer is open source and is freely available at https://sites.google.com/a/georgetown.edu/fornace-lab-informatics/home/metabolyzer along with detailed installation instructions, and a comprehensive tutorial.

General Workflow Overview

The backbone of MetaboLyzer is its statistical workflow that is specifically tailored for analysis of metabolomics datasets. A diagram of the workflow is illustrated in Figure 1. The basic goal of this workflow is to identify ions that are shown to exhibit statistically significant differences when comparing two input post-processed LC/MS datasets, which are typically control and treatment sets. Analysis is restricted to comparisons between only two groups. The "missing data" issue endemic to metabolomic datasets is dealt with in the first step by filtering out all ions that do not show up in a user-defined percentage (a zeros threshold Z_{thr}) of at least one of the datasets. Data is segmented into two distinct groups: partial-presence (i.e. data for ions that only appear above the Z_{thr} for a single dataset) and complete-presence (i.e. data for ions that appear above the Z_{thr} for both datasets). The higher the Z_{th}, the more conservative the ion selection will be. The user is then given the option to transform and normalize the data for the ions that have not been filtered out through a set of user selectable methods.

From this point, data is interpreted via two separate pipelines, depending on the aforementioned segmentation. Partial-presence data is treated as categorical in nature and analyzed via discrete statistics. Ion abundance values are reinterpreted as binary discrete variables, wherein a value is either zero or non-zero, and analyzed via Fisher's exact test for statistically significant differences. For complete-presence data, a far more thorough

analysis is conducted which encompasses classical parametric statistics as well as multivariate techniques. Initially, the data undergoes transformation, normalization, and outlier filtering via user selected procedures. Then, a battery of common statistical tests and calculations are performed, which includes the Student's t-test, Welch's t-test, Wilcoxon rank-sum (Mann-Whitney U) test, Kolmogorov-Smirnov test, Anderson-Darling test for normality, confidence intervals, as well as basic fold change, mean, and standard deviation calculations. While the results of all tests are available to the user for perusal, only the results from one (user defined) "primary" test are taken into consideration, from which the final decision to filter out ions that are not statistically significant (based on p-value) is made.

Upon completion of statistical testing, a volcano plot is first produced comparing the fold-change versus the p-value (for the primary test) for all complete-presence ions. The list of statistically significant complete-presence ions and their (transformed) data is then recompiled for high dimensional data visualization. All non-significant and partial presence data are excluded from this stage of the analysis. Various multivariate techniques, including principal component analysis (PCA), multidimensional scaling (MDS), kernel PCA, and independent component analysis (ICA), can be selected, from which a three dimensional visualization of the data is constructed, allowing the user to discern differences and similarities between the two input datasets in Euclidean space. Finally, a heatmap is constructed from the statistically significant ions, which sheds further light on the degree and magnitude of the aforementioned differences and similarities.

Transformation and Normalization Procedures

The foundation of MetaboLyzer's functionality lays in its data transformation and normalization procedures. Such procedures are always necessary in high throughput quantitative biological data, especially metabolomics, due to its many factors that make for intrinsically noisy data. The transformation procedures utilized in the software are well known, however the normalization procedure that is proposed is novel to the field of metabolomics, and was developed from primarily a statistical, rather than from a chemical or biological approach (such as sample spiking). Figure 2 provides a diagram of the workflow for data transformation and normalization.

Log transformation and inverse hyperbolic sine (IHS) transformation are two powerful techniques that greatly aid in proper statistical analysis. Log transformation is extremely common in bioinformatics, as well as most other fields involving quantitative analysis such as finance and econometrics. It is the classic solution to normalizing skewed data. Metabolomics data from any given sample will always possess far more metabolites in the low abundance range than the higher range, thus necessitating a log transform to eliminate the impact of the minority high abundance metabolites. MetaboLyzer's log transform procedure simply applies a natural logarithm to all non-zero abundance values, which by mathematical necessity removes all zeros from the data. IHS transformation, however is less common, but generally possesses the same characteristics as a classical log transform, such as large value suppression. IHS transformation is very common in econometrics[14], but is far rarer in the biological sciences. The primary difference between the IHS and log transformations is that while a standard logarithm is undefined at zero, the IHS function is fully defined, and thus will not remove zeros from the data during transformation. MetaboLyzer's IHS transform procedure simply applies the standard inverse hyperbolic sine function to the data.

A novel procedure is implemented in the software, named Gaussian normalization. The procedure is dependent on the characteristics of log transformed data, and thus is only

available if it is applied. Gaussian normalization is essentially a statistical Z standardization of all log transformed abundance values in a given sample (*Z*), which involves estimating the mean ($x_{\bar{Z}}$) and standard deviation (s_z) of all transformed ion abundance values, and then normalizing each transformed abundance value (A_{log}) in the sample as follows:

$$A_{normed} = (A_{log} - \overline{x}_z)/s_z$$

It is important to emphasize that this procedure is applied on a sample-by-sample basis, meaning that the $x_{\bar{Z}}$ and s_z parameters are unique to each biological sample. The impetus behind Gaussian normalization stems from the resemblance of the distribution of any given sample's log transformed data to a standard Gaussian distribution (hence the name of this normalization procedure). Gaussian normalization is simply an attempt to convert the distribution of a sample's data into a standard Gaussian curve with a mean of 0 and a standard deviation of 1. When this procedure is applied to all samples in a set, it can be thought of as "aligning" them, much like peak alignment procedures for chromatogram deconvolution.

Small Molecule Database Integration and Putative Identification

For the biologist, perhaps the most compelling aspect of MetaboLyzer is its tight integration of the Kyoto Encyclopedia of Genes and Genomes[15] (KEGG), Human Metabolome Database[16] (HMDB), Lipid Maps[17], and BioCyc[18] small molecule databases, which allows for direct translation of often cryptic quantitative data into putatively biologically relevant information. During the statistical testing phase of the MetaboLyzer workflow, the program will output the putative ion identification and its associated metabolic pathways (via KEGG and BioCyc) for every statistically significant ion found (in both complete- and partial-presence sets). This is accomplished by elucidating an ion's original neutral mass by considering all possible positive or negative mode adducts for a given *m/z* value. The KEGG, HMDB, Lipid Maps, and BioCyc databases are then utilized to find a putative match for this elucidated neutral mass, within user defined weight tolerances (via user defined parts per million threshold). Furthermore, MetaboLyzer attempts to identify only relevant putative matches by identifying a metabolite match's relevance to a user specified organism (rat, mouse, or human). KEGG and BioCyc offer even greater insight through their meticulously curated metabolic pathway databases, which are fully exploited by MetaboLyzer. The integration of MetaCyc, a vast BioCyc sub-database which curates experimentally verified data on pathways that are microbial in origin, is particularly useful in the field of microbiomics, the study of metabolism associated with gut flora. Once all putative identification is complete, MetaboLyzer compiles the most frequent KEGG and BioCyc pathway "hits" (i.e. pathways that have the most putatively identified ions belonging to it) and outputs this information in a histogram for each database. These histograms offer valuable initial insight into the possible biological mechanisms behind the differences observed between the two datasets.

Correlation Analysis Workflow

Ancillary to the primary statistical workflow is a separate procedure that looks for statistically significant differences via analysis of correlation. Whereas the primary workflow looks for statistically significant changes directly in individual ion abundance levels themselves, correlation analysis instead looks for statistically significant changes in the correlation between any two given ions, which can be biologically interpreted as increased or decreased co-regulation. Due to statistical requirements for calculating a Pearson product-moment correlation coefficient (*r*), this analysis is restricted to the

complete-presence ion dataset, with certain features only available if the sample size is a minimum of 10. Figure 3 diagrams correlation analysis, which is separated into two stages. All results produced in the correlation analysis workflow are separate from the results produced in the primary workflow.

The first stage of the correlation analysis workflow entails the construction of two correlation-based heatmaps. These two heatmaps are visual representations of the global metabolite co-regulation structure of the control and the treatment datasets. By visually inspecting both heatmaps in tandem, a great deal of information can be discerned regarding the organization and regulation of the metabolites, and how the treatment impacts these patterns. The first step involves calculating the *r* values between all complete-presence ions for both the control and treatment datasets. This results in two correlation matrices (control and treatment), which can be directly used to create two correlation heatmaps, or modified into two dissimilarity matrices to create two dissimilarity heatmaps. The dissimilarity matrices are calculated directly from the correlation matrices via element-by-element application of the following equation:

$$D_{X,Y}=(1-r_{X,Y})/2$$

Both types of maps allow for direct visual comparisons of global and/or localized correlation pattern changes. Construction of any heatmap requires feature reorganization based on dendrogram clustering. Hierarchical clustering analysis of the ions is conducted on either the control or treatment dataset, which produces a single dendrogram to be used in both heatmaps. For instance, a dendrogram can be created from the control set, which is then used when creating the heatmaps for both the control and treated datasets. The decision as to which dataset to create the dendrogram from depends on which dataset the user wishes to consider as the “canonical” organization, which is typically the control set.

The second stage of the correlation analysis workflow attempts to identify the statistically significant correlation differences when comparing the control and treatment correlation matrices. This stage is only possible when the minimum sample count of 10 is met, and is otherwise bypassed. Whereas the heatmaps created in the first stage are generally qualitative in nature, the second stage attempts to bring quantitative rigor to its results. A correlation difference matrix, which quantifies the differences in *r* values in the control and treatment matrices, is initially constructed. The difference matrix is calculated via element-by-element application of the following equation (with element being defined as a value in a matrix):

$$\delta_{X,Y}=|r_{X,Y,treatment}|-|r_{X,Y,control}$$

A differential correlation heatmap from this correlation difference matrix is constructed in a similar fashion as the previous heatmaps, using the same dendrogram for organization calculated in the first stage of the workflow. A p-value is also calculated for each element in the correlation difference matrix so that the most statistically significant correlation shifts can be extracted for further investigation. This p-value is calculated by using a simple Z-test on the original *r* values after a Fisher transformation has been applied. In doing so, the statistically significant differential correlation values can be identified, and utilized to create a new differential correlation heatmap that only highlights these statistically significant results.

Analysis of Experimental Data

For demonstration purposes, MetaboLyzer was used to analyze data from a previous radiobiology study consisting of urine samples collected from 11 C57BL/6 8–10 week old male mice 24 hours after whole body exposure to a single 8 Gy dose of gamma radiation [19]. For a baseline comparison, urine samples were also collected from 10 unexposed male mice of the same strain and age. The individual urine samples were then 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 chromatograms were then pre-processed with MarkerLynx (Waters, Inc.) to extract the final metabolomic dataset matrix and each sample was normalized to its respective creatinine levels.

Only positive mode metabolomic data was used for the purposes of this demonstration. The fairly conservative zeros threshold (Z_{thr}) of 0.90 was used, which restricts analysis only to ions with non-zero abundance values in 90% or more of at least one group. Two analyses were conducted, one with log transformation coupled with Gaussian normalization, and another without any transformation or normalization whatsoever. 1.5 IQR based outlier filtering was used for both analyses. In the first analysis, MetaboLyzer identified 121 out of 1177 partial-presence ions to be statistically significant ($p < 0.05$) via the Fisher's exact test, and 122 out of 765 complete-presence ions to be statistically significant ($p < 0.05$) via the Mann-Whitney U test. In the second analysis, MetaboLyzer identified 121 out of 1177 partial-presence ions to be statistically significant ($p < 0.05$) via the Fisher's exact test, and 160 out of 765 complete-presence ions to be statistically significant ($p < 0.05$) via the Mann-Whitney U test. In this dataset, the transformation and normalization did not result in major differences; however they could play a major factor in other datasets, and these options are made available in MetaboLyzer for the user to conduct multiple analyses with different parameters. For the purposes of this demonstration, the results from the first analysis (with log transformation and Gaussian normalization) are further discussed.

A heatmap and volcano plot of the significant complete-presence ions, which both indicate significant increased as well as decreased excretion, are shown in Figure 4. The heatmap in the figure utilizes color assignments coupled with Euclidean based hierarchical clustering to visually represent the statistically significant changes found in the data. Samples under the red bar are non-irradiated, while samples under the blue bar are irradiated. In addition, a density plot of the distribution of the normalized abundances is shown in the top left hand corner of the heatmap. The volcano plot shows much of the same information as the heatmap, but in scatter plot form, plotting significance (p-value) versus fold change. Grey dots indicate complete-presence ions that are not statistically significant, while red dots are significant by the user defined statistical test and p-value cutoff.

Based on aggregate putative identification of the 243 (both partial- and complete-presence) statistically significant ions, many KEGG and BioCyc pathways were identified to potentially play a role in radiation response. Some significant BioCyc pathways include proline biosynthesis from arginine (PWY-4981) as well as tryptophan degradation to 2-amino-3-carboxymuconate semialdehyde (PWY-5651). KEGG pathways found to be significant include steroid hormone biosynthesis (ko00140) and 2-oxocarboxylic acid metabolism (ko01210). The putative pathway hit histogram in Figure 5 shows the complete list of KEGG pathway hits for metabolic pathways with 4 or more metabolites putatively identified as belonging to it.

Multivariate analysis was also conducted on the 122 statistically significant complete-presence ions. Singular value decomposition based linear principal component analysis

(PCA) as well as independent component analysis (ICA) procedures were conducted on the data, and a 3 dimensional visualization (first 3 principal components and first 3 dimensions, respectively) of the sample separation for both analyses are shown in Figure 6. In both plots, the non-irradiated (red dots) and irradiated (blue triangles) samples show visually distinguishable separation from one another, and give a general indication as to the discriminating qualities of the statistically significant complete-presence ions.

Finally, correlation analysis was conducted on all 765 complete-presence ions. Whereas the previously described analyses focused on the raw difference in abundance levels of individual ions though standard statistical testing, correlation analysis instead focuses on differences that can only be observed when considering the statistical relationship between two ions, known as correlation. When comparing the dissimilarity heatmaps for the non-irradiated (Figure 7A) versus the irradiated (Figure 7B) samples, it is clear that there are very stark changes to the correlation patterns that are the result of the radiation. Whereas the non-irradiated heatmap exhibits very distinct patterns and clusters of red, which are indicative of high correlation and suggests a high co-regulation of metabolic pathways, the irradiated heatmap appears almost homogenous, with no clear demarcations and unique features that the other map exhibited. This may be indicative of widespread disruption of the canonical metabolic pathway network that the non-irradiated heatmap suggests. The differential correlation heatmap (Figure 7C) combines both the non-irradiated and irradiated maps into a single map so that differences are more apparent, with the orange spectrum indicating a gain in correlation, while the blue spectrum indicates a loss. The prevalence of blue in the map essentially tells the same story that was discerned from comparing the two dissimilarity heatmaps, that of widespread loss of correlation structure and therefore disruption of metabolic regulation. Though there are indeed areas of orange on the map, which indicate a gain of correlation, statistical analysis indicates over three times the number of significant correlation losses at 17985 when compared to the 5651 correlation gains (p-value < 0.05). The final heatmap in Figure 7D, which only plots statistically significant correlation differentials, almost exclusively shows blue, loss of correlation.

Discussion

The comprehensive nature of MetaboLyzer is its core strength as a statistical metabolomics analysis platform. Though the majority of MetaboLyzer's methods are not novel, the integration of these classical biostatistics techniques into a focused and streamlined workflow, coupled with the incorporation of four major small molecule databases directly into an otherwise (for the biologist at least) dry statistical analysis makes MetaboLyzer a compelling platform. In a sense, MetaboLyzer can directly translate otherwise incomprehensible and cryptic numerical data directly into potentially biologically meaningful results. MetaboLyzer also provides novel approaches such as correlation analysis and zeros threshold based filtering.

There are numerous software suites geared towards analyzing metabolomics data, however only a handful are tailored specifically for comprehensive post-processed data analysis. Some of the most well known software in the field includes XCMS, MetaboAnalyst[20], FIEmspro[21], Thermo Scientific's Sieve, Agilent Technologies' Mass Profiler Professional, Waters' MarkerLynx, and MZmine. Among these, only MetaboAnalyst and FIEmspro are specifically geared towards the type of post-processed analysis that MetaboLyzer focuses on. The other suites primarily focus on the various aspects of pre-processing (i.e. chromatogram deconvolution), and only have very basic post-processing analysis tools, such as PCA. MetaboAnalyst, an online resource that requires users to upload their data to their website in order to conduct analysis, offers a very impressive array of analysis tools, some of which are not available in MetaboLyzer. However, the sheer volume and breadth of

statistical information that is produced at the final stage can easily be overwhelming for users who do not have a strong background in statistics and bioinformatics, whereas the output of MetaboLyzer is more geared towards a more streamlined approach that focuses on the information that has the most potential for being biologically relevant. MetaboAnalyst relies on databases that are curated by the University of Alberta, including HMDB, DrugBank, and SMPDB for its putative ion identification and metabolic pathway analysis. The databases integrated into MetaboLyzer arguably have a wider breadth of coverage, especially for lipids (via Lipid Maps) as well as metabolites originating from microbes and plants via BioCyc/MetaCyc database integration, which gives MetaboLyzer the edge in microbiomics, an emerging sub-field of metabolomics that focuses on the microbiome. FIEmspro's analysis tools are extensive as well, and include several techniques, such as Random Forests, that MetaboLyzer lacks, but there is no putative ion identification workflow integration whatsoever. Table S-1 contains more information comparing the aforementioned software suites. Overall, MetaboLyzer's focus on a specific subset of statistical tools coupled with broad integration of major databases for elucidating potentially biologically relevant results makes it a very practical toolset.

A primary aspect of MetaboLyzer that differentiates it from virtually all other postprocessing analysis software is the method by which missing data is handled. Missing values in metabolomics datasets (i.e. zero abundance values) are typically "filled in" via imputation algorithms. However, the prevalence and sheer pervasiveness of missing values in metabolomics data can make these imputation methods a potentially hazardous solution. In essence, imputation methods are modifying the data to accommodate the tools available, and it is difficult to say at what point these modifications have gone too far, i.e. too much of the data is "made up". In contrast, MetaboLyzer's approach to this issue does not attempt to fill in missing data at all, but instead utilizes a user defined zeros threshold (Z_{thr}) to identify ions that have sufficient non-zero data for statistical analysis, eschewing the need to impute missing values. Zeros filtering is a simple and intuitive approach to the problem, and picking a proper Z_{thr} value is more of a qualitative endeavor that depends primarily on the sample size, as well as the degree of conservativeness desired by the user, akin to picking a lower or higher ppm error window for putative metabolilte identification. The tutorial, which is available online, provides more detailed guidelines for selecting a proper Z_{thr} value.

Two aspects that MetaboLyzer ignores are the incorporation of paired statistical tests and multiple testing correction. Paired statistical tests would be most applicable to experiments with pre-post designs, wherein samples are collected before and after treatment. This is due to the fact that MetaboLyzer has been designed primarily with biomarker identification in mind. When it is applicable to the data, the primary advantage with utilizing paired statistical tests is to increase statistical power by reducing confounding factors from inter-sample variation. Though unpaired statistical tests may have less statistical power when used to analyze paired data, the statistically significant results that are produced may be more robust and will not necessitate a baseline pre-treatment sample, as may be the case for results produced from paired tests. Furthermore, unpaired tests are applicable to all datasets, whereas paired tests are only applicable to a certain subset. Multiple testing correction is an issue that arises when conducting multiple statistical tests for a single experiment, which may result in the increase in the false positive rate. Basic solutions to these problems, such as the Bonferroni correction, are usually far too conservative for high throughput biological data such as those originating from most –omics fields, and indeed there has been a great deal of research into the most suitable methods of correction for these types of data. Multiple testing correction is a contentious issue, as there are doubts as to whether there is even a need for any correction to be made at all from a biological standpoint, and in some fields is not considered an absolute necessity[22]. In considering that the metabolomics field is

still relatively new, with many fundamental aspects of the data still unknown, we have chosen to not incorporate any multiple testing correction methods in the package.

Much of the innovation with regard to the quantitative aspects of the field of metabolomics has thus far focused on the pre-processing stage, i.e. chromatogram interpretation, peak identification, and alignment. Works such as XCMS and MZmine are major accomplishments that have greatly advanced the field and have gained wide acceptance. While there is still a great deal of work to be accomplished in these areas, progress in other aspects of metabolomics should not be neglected. We try to address this with the development of MetaboLyzer, an analytical toolset specialized for post-processed data.

Conclusion

The comprehensive integration of metabolomics-centric workflows, classical univariate and multivariate statistics, small molecule databases, and novel techniques specific to metabolomics into a single package are what make MetaboLyzer a useful and attractive platform. The sometimes overwhelming volume and often cryptic nature of metabolomics data can easily intimidate or confuse investigators new to the field. MetaboLyzer's ability to streamline analysis to rapidly translate quantitative data directly into potentially biologically relevant results is its core strength. At the same time, the standard workflow provides numerous options by which more advanced users can deviate from the default settings to conduct more sophisticated analyses. It is a powerful tool in the hands of the biologist and the biostatistician, as well as the metabolomics novice and the expert.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by U19AI067773 and R01AI101798; the NIAID grant U19AI067773 was crucial in supporting this effort. George Luta provided vital suggestions and consultation in regards to the statistical approaches utilized. The project was also supported by Award Number 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. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. Analytical chemistry. 2006; 78:779–787. [PubMed: 16448051]
3. Katajamaa M, Miettinen J, Orešič M. Bioinformatics. 2006; 22:634–636. [PubMed: 16403790]
4. Hrydziuszko O, Viant MR. Metabolomics. 2012; 8:161–174.
5. Baker M. Nature Methods. 2011; 8:117.
6. Jones, E.; Oliphant, T.; Peterson, P. J. 2001. http://www.scipy.org/
7. Gautier, L. 2008. http://rpy.sourceforge.net/rpy2.html/
8. Hunter JD. Computing in Science & Engineering. 2007:90–95.
9. Team, R. D. C. R Foundation Statistical Computing. 2008. Ripley, BD. Modern applied statistics with S. Springer; 2002.
10. Karatzoglou A, Smola A, Hornik K, Zeileis A. 2004
11. Warnes GR, Bolker B, Lumley T. R package version. 2009; 2
12. Koldovsky Z, Tichavsky P, Oja E. Neural Networks, IEEE Transactions on. 2006; 17:1265–1277.
13. Ripley B, Hornik K, Gebhardt A, Firth D. R package version. 2011; 7:3–716.

14. Ramirez OA, Moss CB, Boggess WG. Journal of Applied Statistics. 1994; 21:289–304.
15. Kanehisa M, Goto S. Nucleic acids research. 2000; 28:27–30. [PubMed: 10592173] 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. Sud M, Fahy E, Cotter D, Brown A, Dennis EA, Glass CK, Merrill AH, Murphy RC, Raetz CRH, Russell DW. Nucleic acids research. 2007; 35:D527–D532. [PubMed: 17098933]
18. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C. Nucleic acids research. 2008; 36:D623–D631. [PubMed: 17965431]
19. Laiakis EC, Hyduke DR, Fornace AJ Jr. Radiation Research. 2012; 177:187–199. [PubMed: 22128784]
20. Xia J, Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS. Nucleic acids research. 2012; 40:W127–W133. [PubMed: 22553367]
21. Beckmann M, Parker D, Enot DP, Duval E, Draper J. Nature protocols. 2008; 3:486–504.
22. Feise RJ. BMC Medical Research Methodology. 2002; 2:8. [PubMed: 12069695] Rothman KJ. Epidemiology. 1990; 1:43. [PubMed: 2081237]

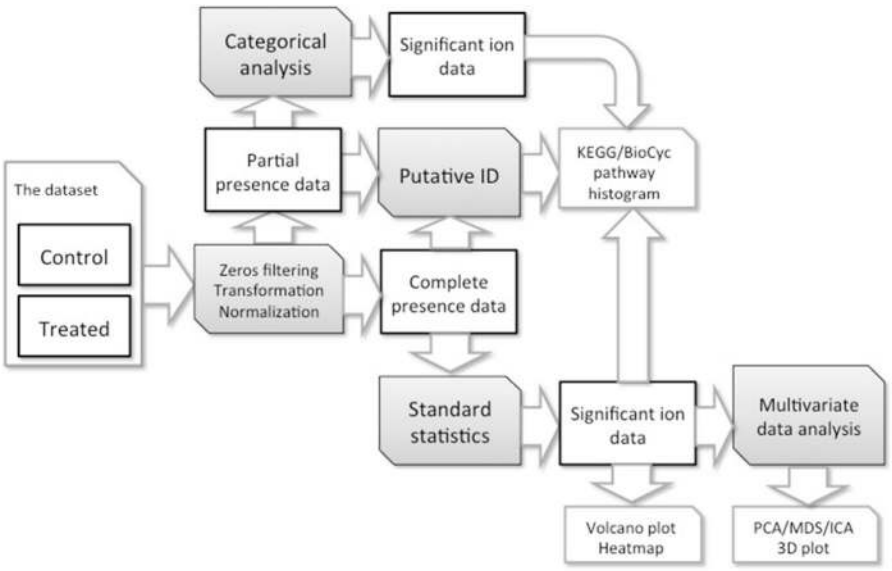

Figure 1.

A general overview of the MetaboLyzer statistical workflow. Two input metabolomic datasets are comprehensively analyzed utilizing a series of statistical techniques that involve filtering, transformation, normalization, categorization, and putative small molecule identification. The ultimate goal (as emphasized in Figure 1) is the output of easily interpretable, statistically, and biologically relevant information.

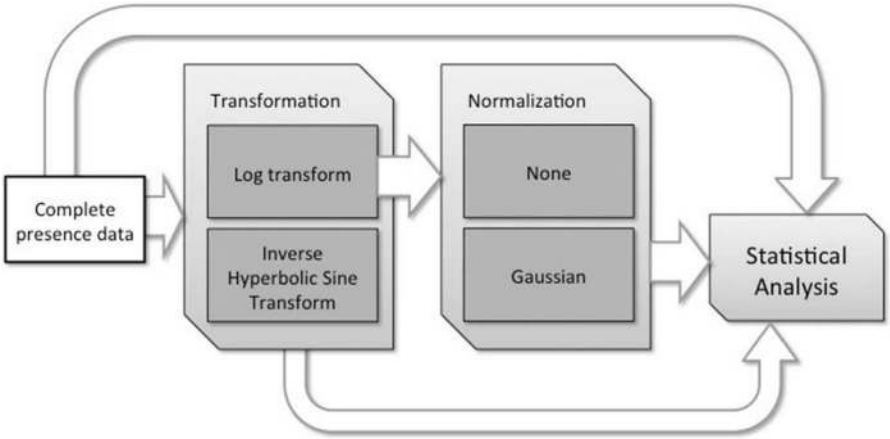

Figure 2.

MetaboLyzer's transformation and normalization workflow. Complete-presence data (ions that possess a non-zero value in a user defined percentage of both control and treatment datasets) are first transformed by either the standard log transformation or the inverse hyperbolic sine transformation. Normalization is an option only if the standard log transformation is chosen. After transformation and/or normalization, standard statistical testing will be conducted. The transformation and normalization stages can also be completely bypassed so that the original data can be analyzed.

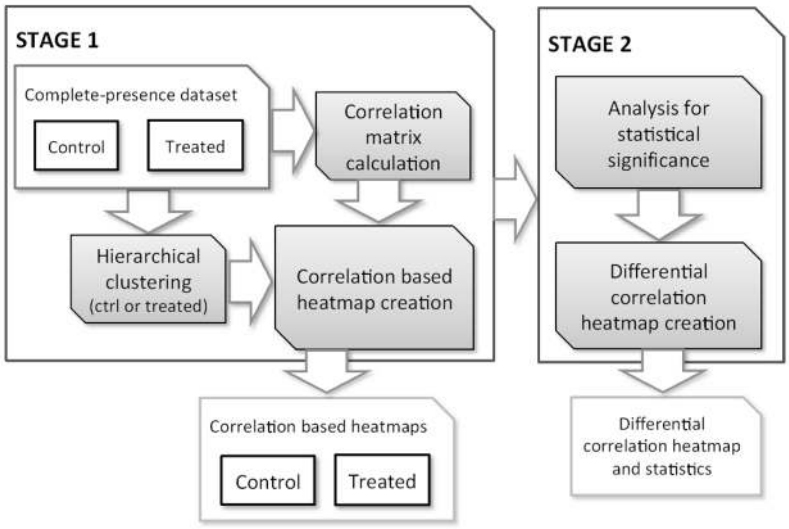

Figure 3.

A diagram of MetaboLyzer's correlation analysis workflow. The workflow is broken up into two stages. Whereas the first stage can be completed for sample sizes as low as 4, the second stage requires a sample size of at least 10. In the first stage, correlation matrices are calculated for both the control and treatment complete-presence subsets. Hierarchical clustering is conducted on only one of the sets (chosen by the user), which is subsequently utilized for correlation based heatmap creation. Heatmaps are created for both datasets, which can be qualitatively compared for changes and shifts in the global correlation structure. Stage two provides quantitative rigor to the analysis by utilizing Fisher transformations coupled with Z-tests to evaluate statistical significance in the correlation shifts when comparing the control versus treatment matrices. From this analysis, a differential correlation heatmap and statistics are produced.

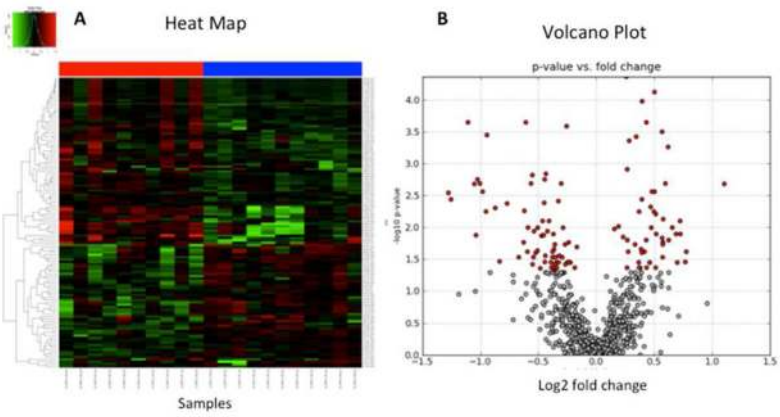

Figure 4.

A heatmap (A) and volcano plot (B) produced by MetaboLyzer during a demonstrative analysis of a previously published urine metabolomics dataset comparing samples collected from 10 non-irradiated versus 11 irradiated C57BL/6 male mice. Irradiated mice were exposed to 8 Gy of gamma radiation, and urine samples were collected 24 hours after exposure. Both figures indicate a very strong response, with a significant number of ions showing increased as well as decreased excretion. 122 out of 765 complete-presence ions were found to be statistically significant by the Kolmogorov-Smirnov test ($p < 0.05$). Red dots in the volcano plot (A) indicate statistically significant ions, while grey dots are insignificant. Using the data from these 147 ions, a heatmap (B) was constructed. Red squares in the heatmap indicate increased excretion, while green squares indicate decreased excretion. Samples under the red bar are non-irradiated, while samples under the blue bar are irradiated.

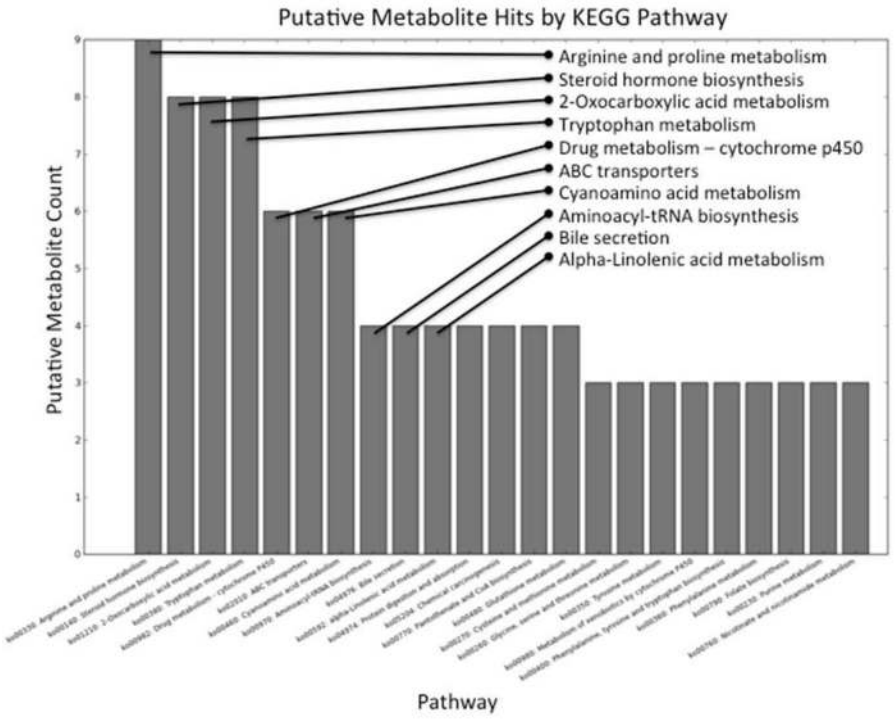

Figure 5.

A KEGG pathway hit histogram and significance plot produced by MetaboLyzer after putative identification of all 243 statistically significant ions. KEGG associates each small molecule with a metabolic pathway. Utilizing this information, significant pathways can be identified by aggregating the number of statistically significant ions that are identified with each pathway. Numerous relevant KEGG metabolic pathways have been identified in this experiment. The top 10 pathways have been specifically labeled for easier readability.

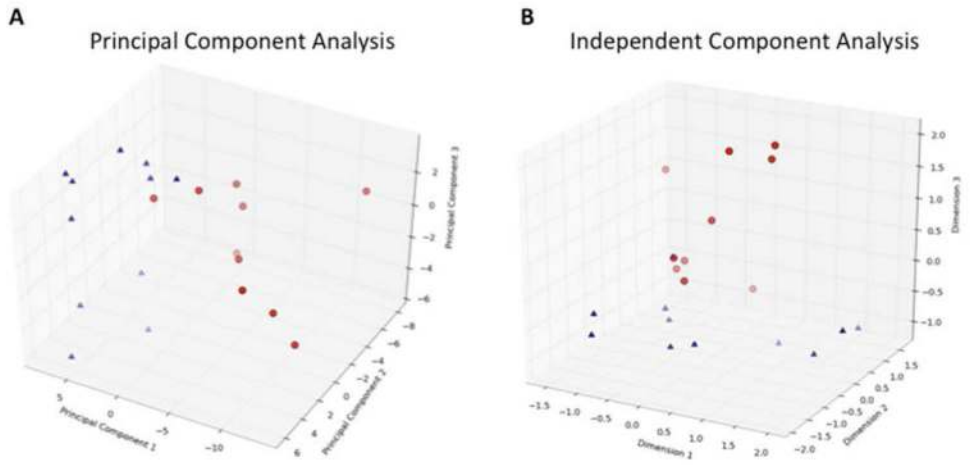

Figure 6.

PCA (A) and ICA (B) have been applied to the 122 complete-presence ions that were found to be statistically significant, from which visualizations of the data projected onto 3-dimensional Euclidean space are produced. The red dots are non-irradiated samples, while the blue triangles are the irradiated samples. In both visualizations, separation between the non-irradiated and irradiated samples is evident, and indicative of the potential discriminating power of the statistically significant ions identified.

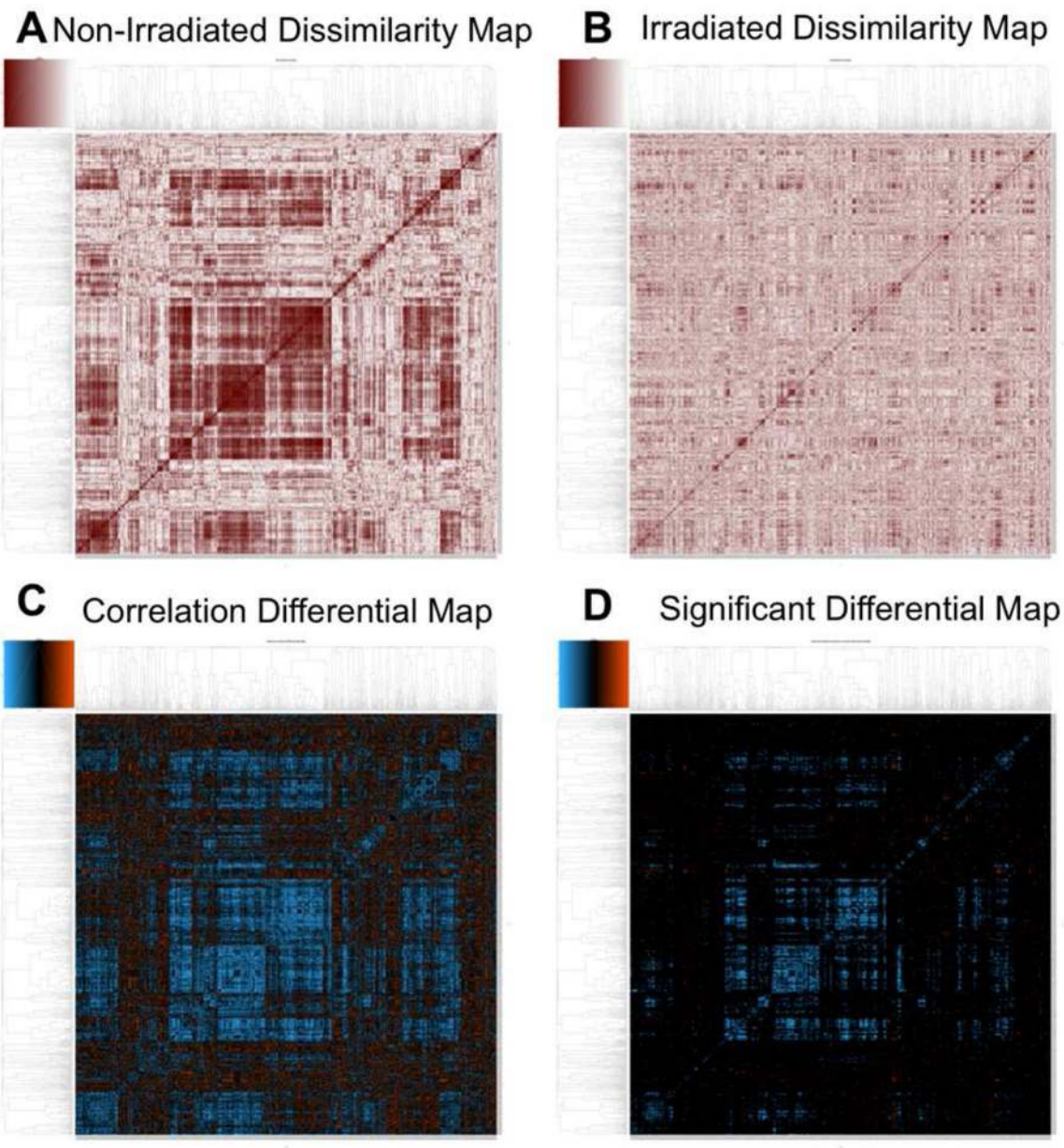

Figure 7.

Correlation based heatmaps created by MetaboLyzer's correlation analysis workflow. Two dissimilarity heatmaps were created, for both the non-irradiated (A) and irradiated (B) sets. Darker colors in these heatmaps indicate higher absolute Pearson's correlation values. The heatmap for the non-irradiated set (A) exhibits very distinct patterns and structures, which suggests a highly structured co-regulation of metabolic pathways. However, the heatmap for the irradiated set (B) is far less organized, with many of the patterns seen in the previous map either distorted or missing entirely, suggestive of a widespread disruption of the metabolic pathways as a result of exposure. The differential correlation heatmap (C) essentially combines the two dissimilarity heatmaps in order to reveal differences between the two, with the orange spectrum indicating gain of correlation (with respect to the non-irradiated), and the blue spectrum indicating loss. This heatmap reaffirms observations made with the dissimilarity heatmaps, with the prevalence of blue indicative of a widespread loss

of correlation. While indeed there is some gain of correlation observed in the differential correlation heatmap, as indicated by the areas in orange, when only the statistically significant correlation differentials are plotted (D), orange is almost entirely absent. This adds statistical rigor to the notion of widespread loss of correlation structure and therefore disruption of metabolic regulation caused by radiation exposure.