Brazilian Journal of Analytical Chemistry
2020, Volume 7, Issue 29, pp 18-44
doi: 10.30744/brjac.2179-3425.RV-03-2020

BrJAC

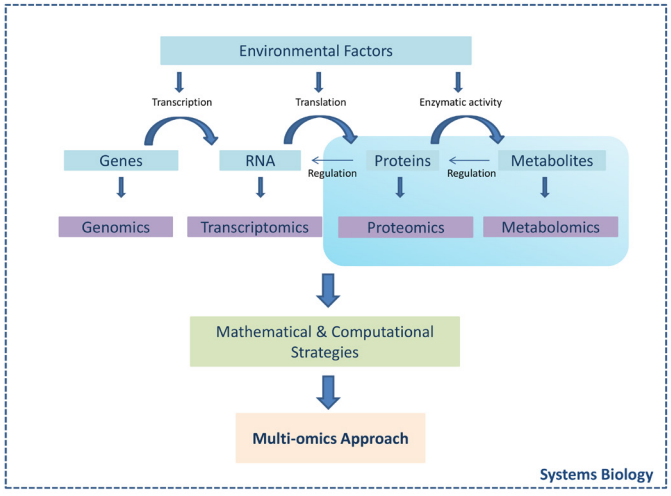
REVIEW

Multi-omics: An Opportunity to Dive into Systems Biology

Emerson Andrade Ferreira dos Santos[1] Elisa Castañeda Santa Cruz[1] Henrique Caracho Ribeiro[1] Luidy Darllan Barbosa[1] Flávia da Silva Zandonadi[1] Alessandra Sussulini[1,2,*]

[1]Laboratory of Bioanalytics and Integrated Omics (LaBIOmics), Department of Analytical Chemistry, Institute of Chemistry, University of Campinas (UNICAMP), P.O. Box 6154, 13083-970, Campinas, SP, Brazil

[2]National Institute of Science and Technology for Bioanalytics – INCTBio, Institute of Chemistry, University of Campinas (UNICAMP), P.O. Box 6154, 13083-970, Campinas, SP, Brazil



Omics data integration employing multi-omics approach is an outstanding opportunity to design a reliable picture of the biochemistry and dynamics of biological systems, as well as prioritize strategies for biomarker discovery. Biological functions are characterized by complex interaction networks, in which the dynamics of biomolecules manages physical and biochemical processes. However, since the emergence of omic sciences, researchers are still looking for the most accurate method to classify and determine the identity and function of biomarkers that describe a system and the ongoing biological processes. Thus, according to the strategies unveiled in the multi-omics literature, this is considered a challenging science field. Therefore, this review describes a workflow example regarding multi-omics data integration, indicating mathematical and computational tools in analysis pipelines that use various methods to perform a sequence of tasks, which would be able to describe biological processes within the systems biology context.

Keywords: multi-omics, data integration tools, omics

INTRODUCTION

The application of computational and mathematical modeling towards a deeper and broader understanding of biological systems is called systems biology. It proposes the study of biology through the use of holistic approaches, in opposition to reductionism, which focuses on the study of subsystems. Systems biology is highly dependent on the biological information acquired by molecular biology and/or omic strategies, which often provide a hypothesis that demand confirmation using a complementary reductionist approach. One important strategy to gain new insights and also assist in the experimental design is data integration [1]. Thus, computational biology can be used to accomplish two main tasks: (1)

Cite: dos Santos, E. A. F.; Santa Cruz, E. C.; Ribeiro, H. C.; Barbosa, L. D.; Zandonadi, F. S.; Sussulini, A. **Multi-omics: An Opportunity to Dive into Systems Biology**. *Braz. J. Anal. Chem.,* 2020, 7 (29), pp 18-44. doi: http://dx.doi.org/10.30744/brjac.2179-3425.RV-03-2020

Submitted 11 February 2020, Resubmitted 22 May 2020, 2nd time Resubmitted 26 June 2020, Accepted 29 June 2020, Available online 22 July 2020.

knowledge discovery, which is performed by the analysis of large amounts of experimental data in order to reveal unknown patterns that usually result in a hypothesis, and (2) simulation-based analysis, with the application of *in silico* experiments affording predictions to be further confirmed by experimental assays [2]. Within this context, the term multi-omics was proposed as a combination of methods to integrate data obtained from different omic approaches, aiming at gaining insight on how the different biomolecules (*e.g.*, proteins, RNAs, metabolites) are interconnected and how the flow of biological information occurs [3].

The term omics comes from the Latin suffix *ome*, which means mass or many [4]. The difference between omics and molecular biology approaches is, therefore, that the first one englobes a larger number of measurements per endpoint rather than one or a few. Despite the number of parameters measured per analysis is increased in omics, the number of replicates is decreased. In part, it happens because of the costs and time necessary for the experiments, and also due to the super estimation of methods, since there is a belief that more measurements would compensate a small number of samples [5]. Frequently, single-omic studies attempt to address specific biological issues without requiring a prior understanding of the biological bases involved [6]. However, experimental limitations, such as sample size (*e.g.*, rare samples), imperfect sequence identifications (*e.g.*, proteins) in databases, or representation of kinetic models from “static” data (*e.g.*, biochemical interactions) may generate gaps in the response of biological questions, which can be filled by multi-omics analysis.

After the advent of genomics, the scientific community has been trying to establish a correlation between the genotype and the phenotype in cells and living organisms [7]. Even with the development of strategies that provide information closer to the phenotype description, like transcriptomics [8], proteomics [9], and metabolomics [10], the individualized data provided by each one of these omics alone do not answer how the different biological processes are correlated [11] and how to explain this complexity. With the purpose of revealing these connections, an alternative is to integrate all (or most of) the available omics data. Through these connections, it is possible to provide complementary information from each omics strategy by the observation and understanding of how these relationships behave in a biological system, *i.e.* the study of genes and their products (RNAs, proteins, and metabolites) could provide a broader view of the modulations at genotype and phenotype levels of a system undergoing a specific biological process, such as a disease. Therefore, data integration from different omic sciences is a promising tool for the early detection of illnesses, as well as to study different treatments and their effects on patients, helping to choose the right medication, within the personalized medicine context [6].

Recently, the strategies in data integration were conducted from proteomic and metabolomic datasets, providing promising and significant results. In 2015, Del Boccio *et al.* [12] integrated these two omics to assess the differential pathways and networks between protein and metabolites in multiple sclerosis. In 2018, Cambiaghi *et al.* [13] employed a different approach integrating targeted metabolomics and proteomics data by using a correlation algorithm in order to observe the importance of circulating lipids and coagulation cascade in septic shock patients, evaluating the progression of the disease. Furthermore, in 2018, Gui *et al.* [14] described a disturbance in the phospholipid metabolism pathway in major depressive disorder and reported 74 differential proteins and 28 metabolites related to this specific metabolic pathway.

Although biology has always been a science of complex properties, in order to perform multi-omics data integration and ensure the data quality, some parameters (*e.g.*, list of genes, proteins, lipids, metabolites) need to be defined for subsequent data integration. For the study of a biological system, single-omic analyses are initially performed in order to identify sets of biomolecules, such as proteins and metabolites, which discriminate the evaluated conditions (*e.g.*, depression patients *vs.* healthy controls). Subsequently, these lists are submitted to multi-omics integration, which aims to reveal how the different types of biomolecules interact and are related to the phenotype. Multi-omics acquired the status of a new scientific area, partly because of the computational mathematics development for high-throughput data integration, while bioinformatics was previously used to treat data from genomics, transcriptomics, metabolomics, lipidomics, and other omic approaches separately. The complete workflow involved in a multi-omics study can be reached in two stages. The first is the acquisition of omics data and their subsequent treatment

by bioinformatics tools and the second is the integration of the parameters previously obtained from the isolated omic approaches by computational mathematics models.

It is clear how systems biology is strongly influenced by the data obtained from omic strategies and by how they are chemometric treated (bioinformatics), and mathematically modeled (multi-omics). If scientists want to improve their models and get more accurate results in systems biology, they first need to look at the difficulties found in the individual omic approaches, since they determine the parameters used in multi-omics analysis and carry possible errors in their results when integrating data. There is a myriad of different platforms for omics integration. However, some of them offer little support or do not possess a clear example dataset as a guide for new users. Moreover, review articles in this area focus on presenting the different applications for each platform without deepening in analyzing the interface and assessing their pros and cons. In this context, this review aims to describe different omics integration platforms for scientists who plan to start working on multi-omics with no experience in the area, focusing on the platform interfaces and their particularities, especially considering proteomics and metabolomics (other relevant platforms used in different multi-omic approaches were previously discussed by Pinu *et al*. [15], Misra *et al.* [16], and Fondi *et al.* [17]). Furthermore, the primordial steps of data preprocessing and pretreatment are addressed in this review, as well as the software used for this purpose and for chemometric analysis and current problems and pitfalls found in omic approaches, since good quality of omics data is indispensable for valuable results when integrated and correlated to the biology of the system.

SAMPLE COLLECTION AND PREPARATION

For omics clinical analysis, serum, saliva, urine and cerebrospinal fluid samples are generally used. Because omic approaches aim to have an insight in the entire ome, it is extremely important to warrant sample stability. There may be active enzymes in the biological specimens, which continue to present activity, thus degrading part of the biological content [18]. Hence, after collection, samples should be correctly stored until analysis. Blood needs to be left to coagulate, following centrifugation and separation of serum, and must be kept at -80 ºC [19]. In the case of urine, phosphate buffer should be added and then it can be stably stored at -25 ºC. For storage at 4 ºC, sodium azide should be added [20]. Saliva can be collected through expectoration in collecting tubes, with the help of a cotton swab or placing a cup over a particular salivary duct, followed by storage at -80 ºC [21,22]. Cerebrospinal fluid represents the most invasive sample type. It is obtained by lumbar puncture, and needs to be centrifuged for removing blood content before freezing at -80 ºC [23].

Samples should be analyzed as soon as possible. If it is not the case, they need to be properly stored and the cycles of freezing-thawing must be controlled, since biological samples are sensitive to environmental changes [24]. The samples may be used for multiple analyses. Biofluids such as blood serum and cerebrospinal fluid can be fractionated by liquid-liquid extraction in an organic and an aqueous layer and also in a protein pellet [25-27]. In saliva and urine (which lack an organic layer), the protein content may be separated from the aqueous layer by different techniques, such as protein precipitation with acetonitrile or acetone [28,29]. The extractions should be executed according to the protocol chosen amongst the multiple possibilities described in the literature, in which the amount of sample necessary varies from quantities such as 30 µL up to 20-50 mL, depending also on the sample type [20,22,25-27,29-31].

PREPROCESSING, PRETREATMENT AND STATISTICAL PLATFORMS

From proteomic and metabolomic approaches, a list of differentiating proteins and metabolites accompanied by their signal intensities are generated, respectively. These lists compose the information required to accomplish integration by characterizing the compounds that present different relative concentrations when different biological conditions are compared, indicating which metabolic pathways have divergent activity. Thereby, for relative quantification, the respective software calculates the fold changes for each biomolecule. For metabolomics, quantification is derived from signal intensities, and for proteomics, spectral counting. Consequently, protein and metabolite absolute quantification is

not a necessary step for multi-omic approaches. There is also neither a mandatory minimum number of biomolecules for the integration, nor the selection of specific metabolites and proteins (the important biomolecules are the ones that discriminate the compared biological conditions). However, the higher the number of identifications, the richer is the biological information obtained from multi-omics analysis.

In lists generated by proteomics, the influence of post-translational modifications (PTM) in proteins depends mainly on the steps involving specific sample preparation techniques and setting parameters for data processing and analysis, though it is not compulsory in a multi-omics approach. Obviously, for researchers interested in PTM analysis, it must be kept in mind that this type of information will be almost completely dissolved when observing the total proteome. In this sense, the necessity for PTM investigation will depend on the biological relevance and the techniques previously applied, as PTM enrichment and purification methods (for examples, see Supplemental Table I) added to the steps of data acquisition and processing downstream information.

Before starting data integration, the first step is to perform processing and treatment of the obtained information, and the second one is to determine the best approach for data integration, in order to guarantee a proper data quality analysis. There is a pipeline, in multiple stages, that is essential to follow when performing the data processing (Figure 1) [32]. In this review, some tools, approaches, and statistical methods from these two essential stages are discussed.

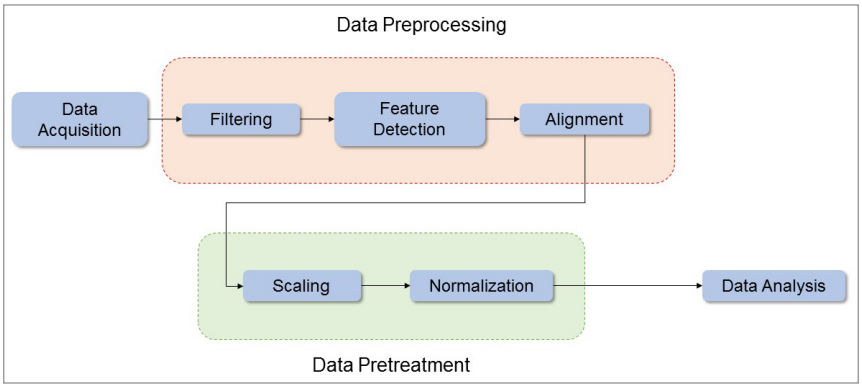


Figure 1. Workflow for the data analysis in multiple stages.

Preprocessing

In mass spectrometry analysis, usually, a specific software (*e.g.*, LabSolutions LCMS, Chromeleon, MassLynx, Analyst, etc.) is employed to extract and export the necessary information from the mass spectrometer to the computer. After exported, the information is mostly presented as files with an extension known as “raw file”. Table I presents the different mass spectrometry vendors’ file formats.

Table I. Raw file extensions from different mass spectrometry vendors

Company	Extension
Agilent/Bruker	.d, .YEP
Bruker	.BAF, .FID, .TDF
ABI/Sciex	.WIFF, .t2d
Thermo Scientific	.RAW
Waters	.PKL, .RAW
Shimadzu	.LCD

Depending on their formats, the files may be provided with different extensions. This is one of the bottlenecks on data integration: data are provided with different extensions or may appear as folders or even as files. In this way, some data types are unable to be opened by any other program than the one which exported them. In order to solve this issue, and redefine the extensions, free tools are available to convert them into the desired format, *e.g.,* MSConvert from ProteoWizard [33].

The MSConvert (http://proteowizard.sourceforge.net) is a simple tool to transform different data formats. Figure 2 presents one of the initial configurations to set up. When converting data, it is possible to apply different filters. Some of these filters and their descriptions are listed in Table II.

Table II. Description of some filters available on MSConvert. More information can be found on http://proteowizard.sourceforge.net/tools/filters.html

Filters	Description
Polarity	Keeps only the spectra with the selected polarity: positive or negative
msLevel	Keeps only spectra with the selected msLevel
scanNumber	Select spectra by scan number
Analyzer	Keeps only spectra with the selected analyzer
Analyzer Type	Filter by mass analyzer type

For Mascot users, the final data conversion format should be .cms1, .cms2, .ms1, .ms2, .text, .mz5, .mgf; while for XCMS (discussed later), they should be .mzML or .mzXML.

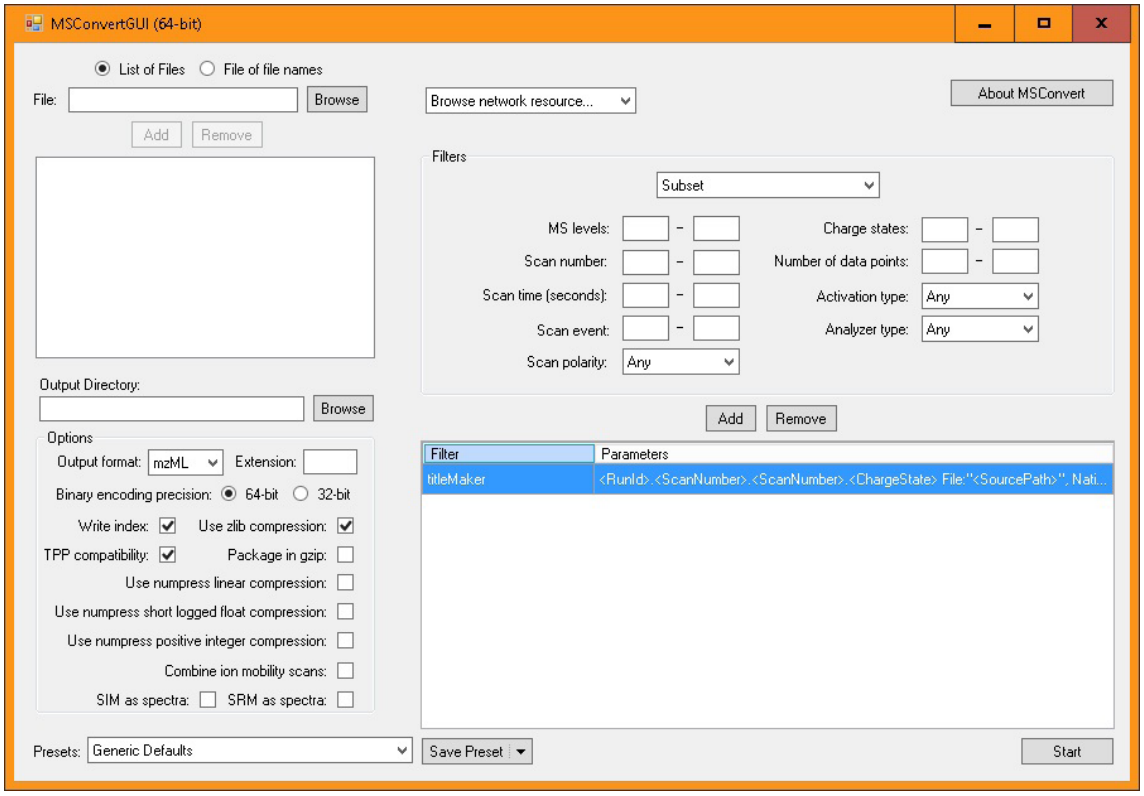


Figure 2. MSConvert interface for Windows.

Another necessary step is to remove the noise generated by uncontrollable variables (*e.g.*, instrumental fluctuations, analyst-related errors, or method inaccuracies), which can happen in any chemical analysis,

through application of filtering methods, which improve the signal-to-noise ratio (*i.e.*, the ratio between the intensity of the measured signal and that of interferences) [34]. Filtering is one of the most important stages of data preprocessing. However, inadequate use of filters can affect subsequent statistical analyses, generating false positives and results without biological value or end up discarding peaks lacking good quality, but which contains reliable biological information [35]. Therefore, it is common to use software to make it automatic and faster. One of them is called XCMS.

XCMS Online (https://xcmsonline.scripps.edu) is a software developed by the Siuzdak Lab at Scripps Research Institute (La Jolla, California, USA). It is used to perform statistical and identification analyses from liquid or gas chromatography coupled to mass spectrometry data for untargeted metabolomics - or targeted analyses by XCMS-MRM. This software performs, besides statistical analysis, chromatographic peak alignment, feature detection, and feature matching in the METLIN database. The software was initially developed for use in R software as a package called 'xcms' [36]. There are two additional help packages: IPO [37], which performs XCMS parameters optimization (*e.g.,* retention time - alignment of the deviation of the time that same compounds present in different samples when eluting from the chromatographic column -, and bandwidth - correction of the standard deviation of the Gaussian model for the shapes of peaks) [38] and CAMERA [39], available only for liquid chromatography coupled to mass spectrometry data, which performs extraction of compound spectra, annotate isotope and adduct peaks and propose the accurate compound mass in highly complex data. All packages are free to download at https://www.bioconductor.org [40].

For beginners, XCMS is used on the online version, providing some default filtering parameters based on the equipment used for analysis. On the other hand, in R software, an Integrated Development Environment (IDE) (discussed later), these parameters can be obtained and optimized by the IPO package [37]. This choice is beneficial for the ones who have programming language knowledge and join the preprocessing steps along the stages. As an additional step procedure, for the R software approach, there are filtering peaks and instrumental (*e.g.*, mass resolution - ability to distinguish two peaks of slightly different *m/z* ratios*,* which depends on the mass analyzer used -, and polarity) and statistical (*e.g.*, *p*-value) parameter settings. For data preprocessing, in this way, this software uses data from quality control (QC) samples. QC is a pool composed of the same amount from each sample used in the assay. Since the QC is analyzed together with the samples, the idea is that all the data from QC analyses do not present significant differences among the measurements, being reproducible [41]. Therefore, if that information does not vary, it means that regardless of the sequence and number of times the same sample was analyzed, the conclusion is "the method is robust and the dataset is ready for analysis".

After preprocessing, data is saved in tabular format files, which contain information on detected features (*e.g.*, retention time, *m/z*, number of peaks detected, etc.) and their intensity for each sample. Then, the data need to go through pretreatment steps so that multivariate chemometric methods can be applied to extract biological information.

Pretreatment

From the tabular format files, two types of information are obtained from previous treatment. The first is composed of desirable (*e.g.*, an important metabolite detected) and undesirable (*e.g.*, a feature with no biological information, such as solvent signal) information, and the second, comprising stochastic contributions. The latter is called noise, which are all uncontrollable variables. Therefore, the pretreatment step must be applied to samples or matrix variables to scaling, normalizing, and treating the missing values imputations.

Scaling

Scaling methods are methods applied to variables and are procedures that precede chemometric analyses. Scaling means that, depending on the method applied, some variables have higher or the same importance as others (*i.e.*, all variables are relevant for the study) [42]. Following are described some

available scaling methods for use according to the type of assay.

i. *Mean center scaling*. This method increases relevance for the most intense peaks rather than the smaller (closer to the noise) in order to diminish the error in multivariate analysis. When using this scale, there is a translation of the data towards the average value of each one.
ii. *Autoscaling*. This method is applied when the data matrix has different dimensions and make all variables contribute with the same importance. This approach makes the data dimensionless - independent of the unit.
iii. *Pareto scaling*. This method is very similar to the previous one. The difference is that instead of using the standard deviation, the square root of the standard deviation is used. This makes the scaling less impactful than the previous one.

The mean center scaling approach is commonly used in spectroscopy. Even though autoscaling is well used for metabolomics and proteomics, an alternative is Pareto scaling, which causes a decrease in the influence of noise and is sensitive to small concentrations of sample constituents [43].

Normalization

At this stage, the values of each variable in each sample are divided by a normalization factor, which can be the average or median of all samples for this variable, putting all samples on a predetermined scale and maintaining the qualitative information. The main objective of this step is to remove the systematic bias, which can be caused by the degradation of the sample components, variations in the amount of sample injected, measurement errors, among others, and this is verified by the QC sample [44,45].

Missing value input

When processing any data, the analyst must be prepared to face some missing data in his data matrix. These missing values can be the result of equipment malfunction, stochastic variations, or even a rigid preprocessing process that excludes the data. Another reason for this effect may be the concentration of compounds lower than the limit of quantification of the equipment. For dealing with this, some methods can be applied [46]:

i. *Feature exclusion*. The first and simplest approach is to remove the entire feature, which presents more than 20 % of its values as NaN, following the “80 % rule” [47].
ii. *LOQ Filling*. The second approach is to fill the missing values with the limit of quantification (LOQ) of the equipment [48]. Missing values, in this approach, can be replaced by 0, LOQ, LOQ/2, or LOQ/$\sqrt{2}$, the most commonly used [49].
iii. *kNN (k-Nearest Neighbors Imputation)*. This approach estimates the value of missing values when finding k samples that do not present missing values and are closest to that not determined. It is calculated based on a simple average or a weighted average [50].
iv. *SVD*. This makes use of principal components analysis or the decomposition by singular values (SVD). In general, this method assigns a value to the missing data and principal component analysis (PCA) is applied. With this, the main components that are most significant are selected and a new value is attributed to the missing data. After that, consecutive PCAs are performed until the value is small enough [51].
v. *Mean/Median*. Used to replace the missing values with the average of the non-missing values of other samples, or with the median of those values.

The choice of an appropriate pretreatment is essential for a successful chemometric analysis. For untargeted metabolomics, for example, it is recommended to use kNN for missing value imputation, according to Do *et al.*, as well described in a metanalysis study [52].

Statistical Platforms

After data preprocessing and pretreatment, statistical analyses are performed. Some of the most used platforms in metabolomics and proteomics will be further discussed.

MetaboAnalyst

One user-friendly free online platform for metabolomics, and nowadays for proteomics, is MetaboAnalyst (https://www.metaboanalyst.ca) [53], which contains different tools for analysis, interpretation, and integration of omics data. Figure 3 presents the start screen with all MetaboAnalyst tools.

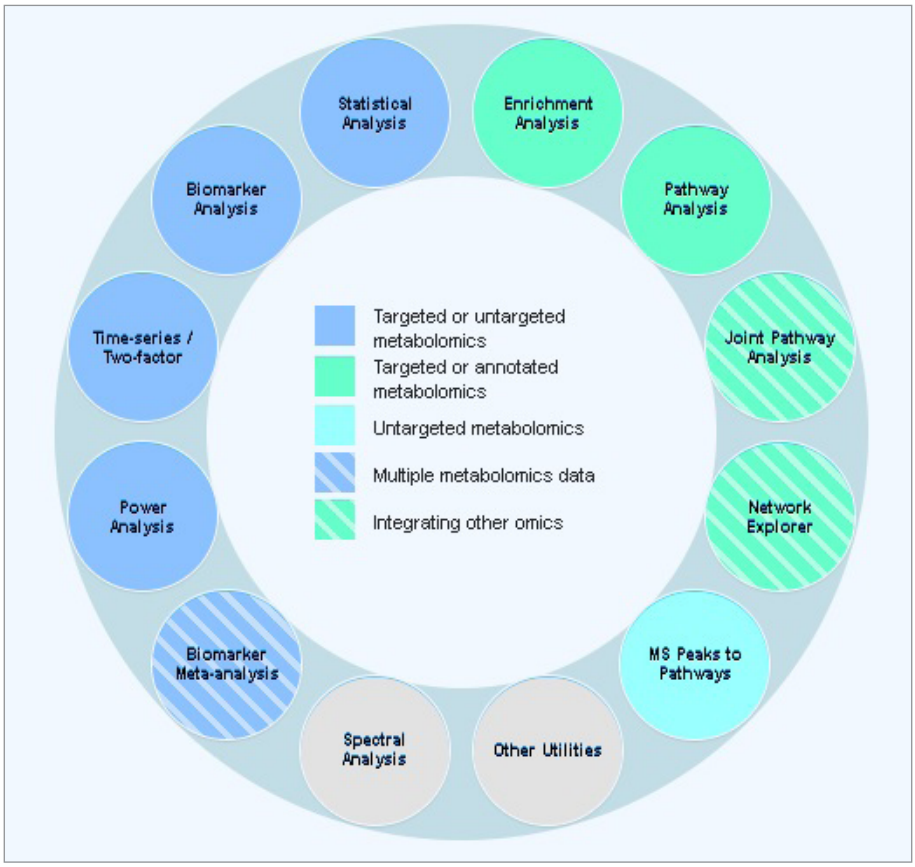


Figure 3. MetaboAnalyst modules screen showing the available modules and the main application of each one.

Following, some of these modules are detailed:

i. *Statistical Analysis*. This module has options for pretreatment such as normalization, scaling, missing value estimation, data filtering and data transformation. Besides, after the pretreatment step, this module can perform univariate (e.g., fold change, t-test, ANOVA), cluster (e.g., dendrogram, heatmaps) and chemometrics (e.g., principal component analysis - PCA -, and partial least squares discriminant analysis - PLS-DA) analysis, among others.

ii. *MS Peaks to Pathways*. This module performs a metabolic pathway enrichment analysis and visual exploration based on mummichog algorithm, with LC-MS spectral peak data. In the visual exploration, it is possible to know which metabolic pathways and which metabolites are significant. Human, mouse and zebrafish are some of the 21 organisms supported in this module.

iii. *Pathway Analysis*. This module performs a pathway analysis (enrichment and pathway topology analysis) for targeted metabolomics analysis. Here, 21 organism models are available (the same organisms from MS Peaks to Pathways module).

iv. *Network Explorer*. This module performs integration between metabolites (metabolomics) and genes (transcriptomics) or metagenomics data. Network Explorer allows the user to upload the data in a list format that is transformed in a visual network based on mummichog, showing their interconnections.

Figure 4. R command history screen of MetaboAnalyst showing some commands of the *Statistical Analysis* module.

For learning and exploring how these modules work, MetaboAnalyst offers sample datasets. Information on other modules can be found on the same page as the module selection panel. In addition to the online version, the R package (MetaboAnalystR) is also available [53]. For each task performed on the online platform, a box (on the right corner of the browser containing coding lines) is displayed, as demonstrated in Figure 4. These codes can be used on the MetaboAnalyst package in R software and describe what was done on the online platform.

Despite having similar functions to XCMS, the big difference between these two is that XCMS uses data in formats as .mzXML, .mzData, .mzData.XML, .netCDF, .cdf, .wiff, .wiff.scan, while MetaboAnalyst works with files already filtered and preprocessed in tabular formats (.txt or .csv). Although XCMS performs data preprocessing, it does not provide options for database queries and some pretreatment methods, while MetaboAnalyst offers options for pretreatment methods and database options. Hence, the idea is to complement the information using the best tools from each software. This can be done directly in RStudio, which has packages with the functions of XCMS and MetaboAnalyst in script format. For the use of all these methods, it is common to use statistical tools and software. The most common is the Integrated Development Environment (IDE). IDEs are software designed to gather in one place everything the developer needs, and often contain syntax checkers, task automakers, prompts, and more [54]. In this context, for omic sciences and statistics, two IDEs stand out: MATLAB [55] and RStudio [56].

MATLAB

MATLAB (MATrix LABoratory) is a powerful mathematical tool and IDE for developing mathematical methods. Created by Clever Moler, it is one of the most known worldwide and uses a C / C ++ derived language. MATLAB offers some expansions known as ToolBoxes that allow for higher expertise in some areas of method development. Although this software requires user knowledge of programming, it presents itself as a user-friendly platform, with clean layouts and functions named in an easy-to-remember way. More information can be found on the MathWorks website (https://www.mathworks.com).

RStudio

RStudio (https://rstudio.com) is a free IDE for R programming language and statistical graphics. Being an open-source IDE, it allows the community to participate in the development of new packages and methods. Although it has a straightforward layout, it does not present menus and buttons, requiring the development of lines of code to use functions. Thus, R users need a little more experience with coding. One of the significant advantages, besides its open-source, is that it is free. In general, this IDE has built-in console, syntax proofing and graph plotting tools, among other functions.

When starting RStudio for the first time, it presents itself with four main windows, in addition to the menus on the top bar. Figure 5 shows the RStudio initial screen. In the upper left window, it is possible to write the script of the functions to be used, depicting where the main work is carried out.

In the upper right window, one can see three tabs: environment, history and connections. The first

shows information about variables, functions, imported data frames and everything that need to be stored. The second shows the history of used functions, imported packages, etc. The third is where one can manage connections to data sources. In the lower left window, there are three tabs: the console, showing the results of processing and executing codes; the terminal, which is used to set up access to the system shell from within RStudio IDE; and Jobs, which manages the jobs RStudio is running. In the lower right window, one can see five tabs: Files, Plots, Packages, Help and Viewer. The 'Files' tab shows computer paths and can be used to import files and to set the Working Directory. The 'Plots' tab displays all plots made. 'Packages' shows installed packages and allows to import them and to install new ones. 'Help' shows RStudio documentation contents for packages and datasets when the help function (signalized as "?") is called followed by the package or the dataset's name (for example, ?ggplot2). 'Viewer' is used to view local web content.

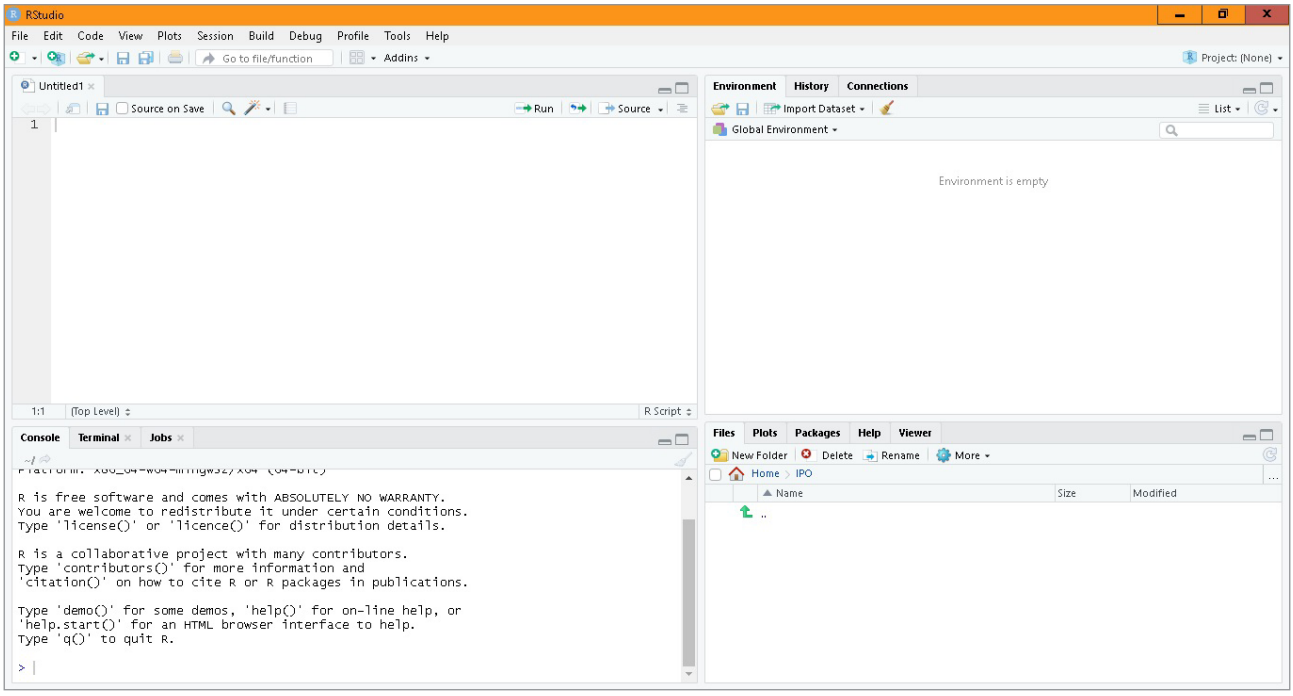


Figure 5. RStudio (version 1.2.5001) initial screen. RStudio layout can be customized according to the user's preferences. The 'View' menu assists in this customization.

The main differences between these two IDEs are the syntax of each program and that one is free (RStudio), while the other is paid (MATLAB).

Similar to MATLAB, which has toolboxes, RStudio has packages made by developers, which can be installed. These packages have plots, data handling, filters, escalation functions, among others. MATLAB also has packages for data preprocessing, accepting files of different formats such as .mzCDF, .mzXML, .JCAMP, among others. One of these packages can be accessed at the main address (https://www.mathworks.com/help/bioinfo/mass-spectrometry-and-bioanalytics.html) or at the most accessible data analysis host in the tool package: https://www.mathworks.com/help/bioinfo/ug/mass-spectrometry-data-analysis.html. Hence, the IDE chosen to perform the described procedures are the user's preference. The main points to be considered for this are the investment to acquire MATLAB and the flexibility of developing the methods in RStudio.

Unlike metabolomics, proteomics data preprocessing and pretreatment are performed using vendor software, which employ their algorithms during data acquisition. Since data processing algorithms are not fully documented and usually are restricted to one instrument platform, it limits the portability to other data

processing tools and comparison of results [57]. Furthermore, most software are available commercially. Following, open and free software for proteomics data analysis are described.

MaxQuant & Perseus

MaxQuant (https://www.maxquant.org) is a free software that analyzes shotgun proteomics data sets obtained by liquid chromatography coupled to mass spectrometry to identify and quantify peptides or proteins. It is well recognized due to its availability to improve the mass accuracy of peptide features through computational techniques [58]. Beyond analyzing data from labeling and label-free strategies, it has a default setup compatible with leading brands of equipment on the market such as Thermo Fisher Scientific, Bruker Daltonics, AB Sciex, and Agilent Technologies [59]. Perseus (https://www.maxquant.org/perseus/) is a platform that helps interpret the data obtained from MaxQuant. There are a variety of statistical tools, such as covering normalization, pattern recognition, multiple-hypothesis testing. Like MaxQuant, Perseus comes in a user-friendly format and is also free [60].

For beginners, it is recommended starting data analysis of large mass-spectrometric data sets following MaxQuant website recommendations (http://coxdocs.org/doku.php?id=maxquant:start). The software includes the Andromeda (peptide search engine based on probabilistic scoring), as well as the Viewer application for inspection of raw data, identification and quantification results. For statistical analysis of its output data, the Perseus platform can be used. Additionally to the installation support, there is important information on the webpage concerning to guiding users on how to process the data from input raw files (MaxQuant & Andromeda) to output processing data (Viewer & Perseus), as well as a link to a forum in Google Groups that discusses questions not found in the software documentation.

DATA INTEGRATION

The platforms discussed in this section apply different strategies to integrate proteomics and metabolomics datasets. Omics integration platforms present two different approaches: (1) chemometric analysis and (2) pathway integration plus visualization. One platform for each approach will be discussed: MixOmics [61] and OmicsNet [62], respectively.

OmicsNet

OmicsNet is a new omics integration platform built by the same group responsible for MetaboAnalyst [53] and has a well-explained and straightforward interface. The tutorials explain the functionalities step-by-step and are very useful for beginners (it can be accessed in the option of the tutorial at https://www.omicsnet.ca/). The input data for proteomics must be structured as a list of differential proteins accompanied by their log-transformed fold change (logFC), where the fold change is defined as how many times the expression has increased or decreased (expression level in condition 1 divided by expression level in condition 2, also known as a ratio). The logFC input is not mandatory; however, it can offer new insights about pathway regulation after the data is integrated. Data with zero in a specific condition (logFC = 0) mean that the fold change is equal to 1. When a specific protein/metabolite shows this value, we can affirm that this specific variable presented no difference (increase or decrease) between the two analyzed groups.

The platform accepts several data formats (Entrez, Ensembl, Uniprot, Official gene symbol ID), as well as ten different organisms. The same scheme applies to metabolomics, although using different databases (such as KEGG [63], PubChem, and HMDB [64]). Figure 6 presents the initial screen, with options to select a proteomics-metabolomics integration and the input options.

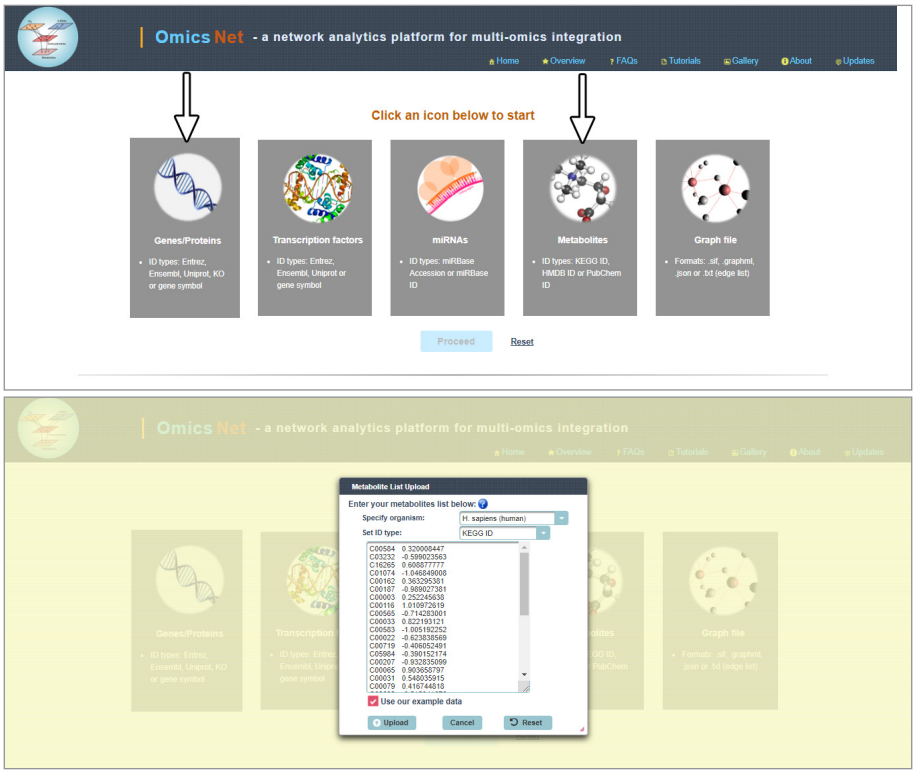


Figure 6. OmicsNet initial screen with the selections for a proteomics-metabolomics integration and input options [62].

After data inputting, the platform leads the user into a second screen, where the network details can be selected before its construction. Here, it is possible to select which interaction will be prioritized over the network construction. For proteomics and metabolomics data, there are two interaction options available: protein-protein interactions (PPI) and metabolite-protein interactions. OmicsNet tutorials suggest using metabolite-protein as primary interaction, as this procedure identifies which enzymes are interacting with the metabolites and the respective PPI [62], as demonstrated in Figure 7.

Depending on the input, the output network can be extensive, and sometimes confusing. The complexity can be reduced in the second screen by controlling the degree (number of connections among nodes) and the betweenness (the measure of the centrality in the graph based on the shortest paths going through the node). Besides, by reducing the network to a “minimum network”, the seeds and other essential non-seeds maintain the network working at minimum. Seeds are the inputted data on the first screen (list of differential proteins/metabolites), while non-seeds are all the other nodes that are used to establish connections between the seeds and, then, build a network with the results. This option is useful to analyze critical connections between both datasets but might reduce the identification of potential pathways in the “network viewer” (next step). After adjusting the settings and submitting the data, it proceeds forwards.

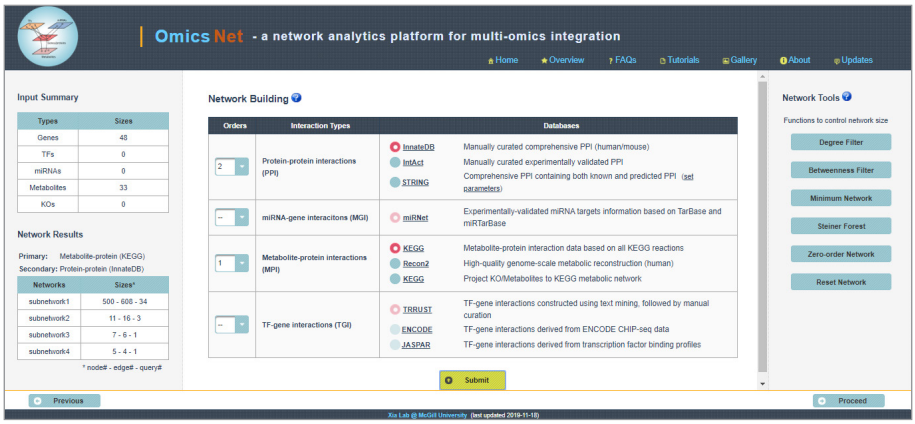


Figure 7. OmicsNet network building screen, with the building results on the left side, the building options in the center, and the network reducing tools on the right side [62].

Network viewer

In this stage, the visualization of the network previously built is possible and manageable. There are some aesthetic options, such as changing the color of the nodes, view options, layout, etc. Sometimes, after a network generation, it creates smaller subnetworks that are accessible in the upper left corner. In the left panel, the platform shows a table with all the current nodes, ranking them by their degree and expression levels using only the logFC input of the first step. Degree is the number of connections between nodes. Nodes with higher degree act as important hubs in the built network. Expression levels showed on the network viewer is a general term to address all the possible alterations of metabolites/proteins between groups (this value is only viewable if the user inputted the logFC in the first screen). The authors of the platform also address this optional second column on the input screen as abundance levels [65].

In the right panel, there are explorer modules and enrichment analysis, which shows the differential pathways in the data. Finally, the save option for the pathway enrichment analysis in a .csv format for further evaluations is possible, as well as downloading the built network for a presentation or an article. The network viewer screen is reproduced in Figure 8.

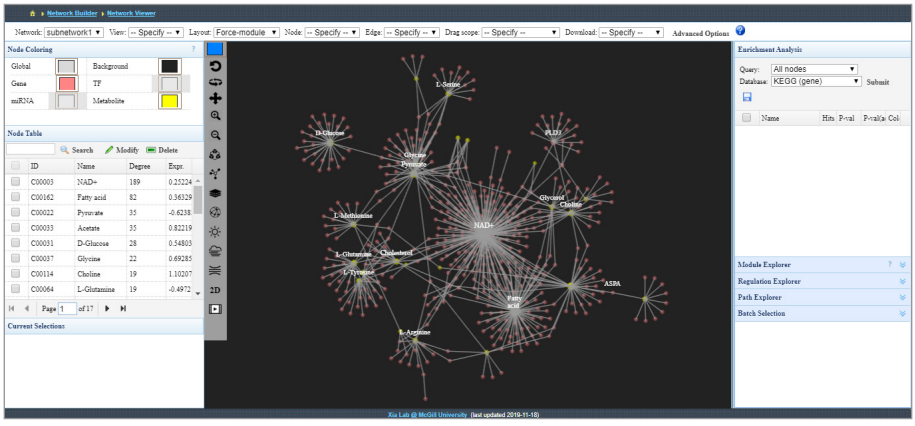


Figure 8. Network viewer screen showing the different modules, node table, and aesthetic options [62].

Regarding many journals that usually demand high-resolution images, one pitfall is the low-resolution images produced in comparison to MetaboAnalyst, which offers image exportation options (such as 300 dpi and 600 dpi in tiff format). Thus, researchers have to use other options to get an appropriate image. Providentially, the platform can export the network in .json format, which is accepted by other visual

systems biology platforms like Cytoscape [66]. Nevertheless, this format loses its expression information in the case of inputted logFC and has to be re-inputted manually.

MixOmics (R package)

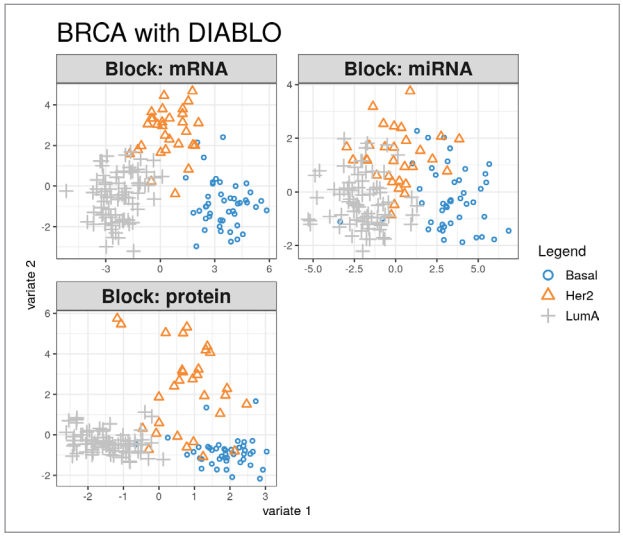
MixOmics [61] is an R package with various statistical and chemometric tools for omics data, which presents a focus on variable selection. The package is composed of nineteen different multivariate methodologies, such as PCA, PLS, PLS-DA, and sparse partial least squares discriminant analysis (sPLS-DA), among others. In this section, we will focus on a specific integration methodology from this package, named DIABLO [67].

DIABLO is an acronym for **D**ata **I**ntegration **A**nalysis for **B**iomarker discovery using **L**atent c**O**mponents. DIABLO method generalizes PLS for multiple matching datasets, and it is named as a N-Integration method by the authors who suggested the primordial analysis of the datasets to be integrated through sPLS-DA and PLS in order to evaluate the major sources of variation and guide the user through the integration processes. For the proper performance of DIABLO, the proteomics and metabolomics datasets must be in .csv or .tsv format to be uploaded into R. In this review, we used RStudio version 1.1.423 (R version 3.5.2), with MixOmics package version 6.10.6.

The datasets must be precisely composed of the same samples as the focus is analyzing the correlation between both omic approaches. After checking it, the next step is the proper analysis.

For the first step of this procedure, the package from the Bioconductor website needs to be downloaded (https://www.bioconductor.org/packages/release/bioc/html/mixOmics.html) and installed on RStudio. After installing and loading the package, the datasets must be also loaded as two different variables (samples on rows, variables on columns). When working with imported .csv files, the data, sometimes, might be in a different format (MixOmics accepts data frames with only numeric values) than the accepted by the package. HINT! If the data set is not being recognized, it is recommended to compare it with the provided data example by the toolkit using the command: `data <-breast.TCGA` and observe the differences between them.

After verifying if datasets are in the accepted format, it is necessary to load a vector Y within the R environment as a factor using “as.factor” before the vector name in order to set it in the correct format. Each of the omics used will be treated as a “block”. DIABLO also requires that the used datasets are structured as a list. Use the list command to build a new one with both datasets and then load this new list and the vector Y into the package using the function:




```
MyResult.diablo <- block.splsda(X,Y)
```

Then, with the `plotIndiv(MyResult.diablo)` function, it is possible to visualize the dispersion of each data, as shown in Figure 9.

With the `plotVar` function, it is possible to observe the correlation circle plots between the datasets, which shows the contribution of each variable to each correlation component [68], as observed in Figure 10.

Figure 9. Dispersion of each block using the `plotIndiv` function in DIABLO package [67].

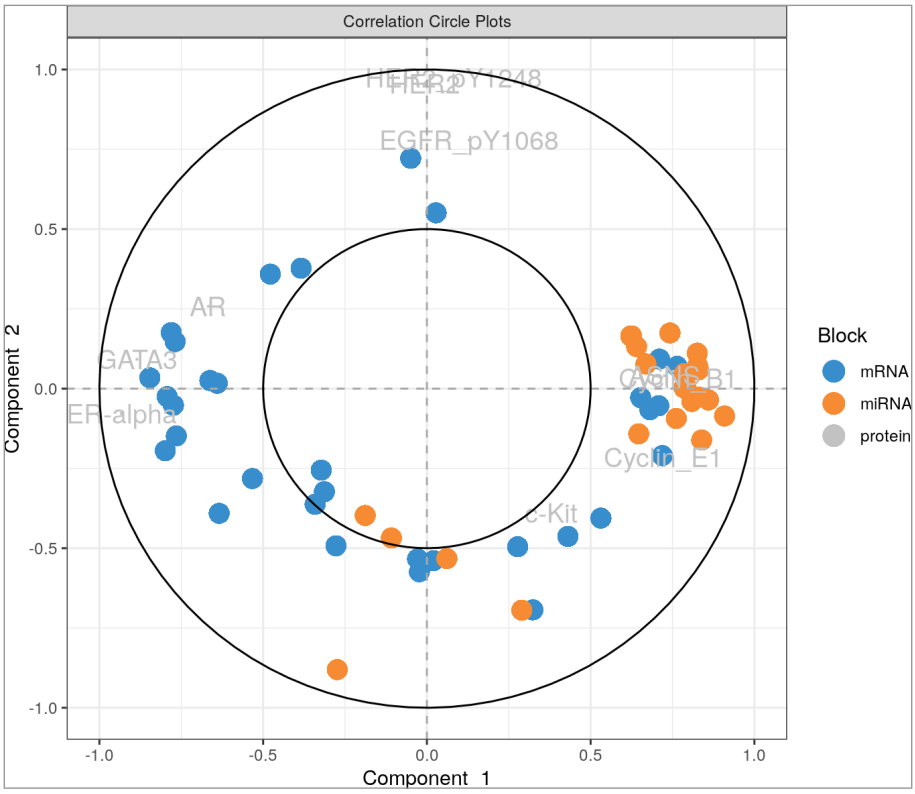


Figure 10. Correlation circle plots of inputted data plotted with `plotVar` function [68].

The `plotDiablo` function allows analyzing the correlation between inputted datasets, as the example in Figure 11.

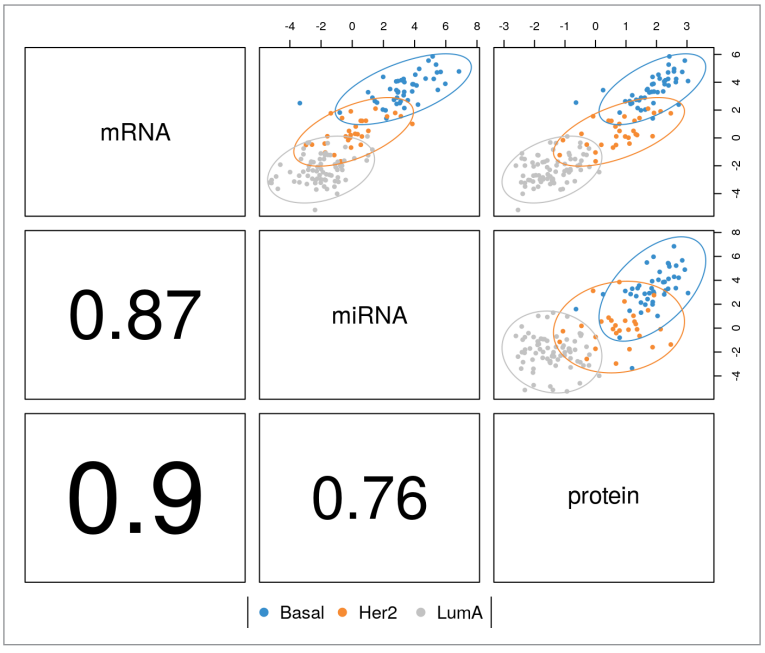


Figure 11. Correlation structure using the `plotDiablo` function [67].

The most useful function of this platform is undoubtedly the `circosPlot`. This plot represents the correlations between the variables represented on the side quadrants, and it is based on a similarity matrix extended to multiple datasets [62]. An example of the `circosPlot` function is presented in Figure 12.

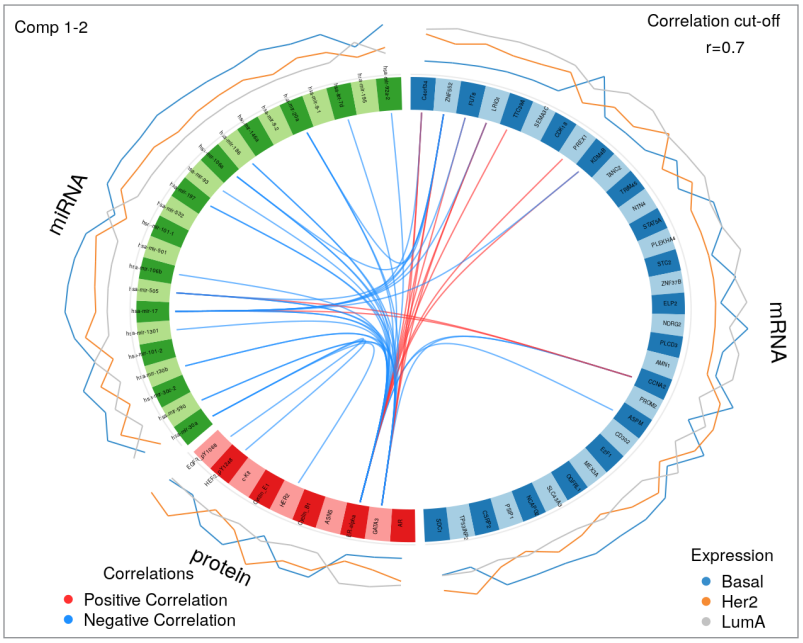


Figure 12. Correlation between variables using the `circosPlot` function [62].

The variables' names are small and can be challenging to observe. However, by using RStudio export function, a file in .pdf format can be exported with enough resolution to allow the observation of which variables correlate to each other, whether it is a positive or negative correlation and the expression of each dataset.

The main drawback of these platforms is the necessity of a previous identification and validation of the metabolites/proteins in order to visualize the possible interactions appropriately among the omic experiments. Another drawback of MixOmics is that all data need to be acquired from the same samples. OmicsNet is excellent at seeing a general linkage between the datasets and identifying the differential pathways, visually. MixOmics can present the correlations between variables through two datasets, indicating the most robust interaction, so the results obtained in OmicsNet can be filtered, showing the most relevant interactions among the metabolites and proteins. They can be used individually; however, together, they can provide a large amount of relevant biological interactions, which can unravel new biochemical mechanisms in several studies worldwide.

BIOLOGICAL INTERPRETATION: ID CROSS-CONNECTIONS

Multi-omics data provide a holistic view of the biological system. In this review, the description on how to create a multi-omics dataset and some platforms to perform multi-omics analysis were presented. Currently available methods approach the integration and interpretation at the downstream process, *i.e.*, using the identifiers (ID) from genomics, proteomics, and metabolomics, to understand the gaps left from single-omic studies [69]. Important biological information has been obtained, especially from integrative approaches using metabolomics and proteomics data. Previous studies have provided evidence concerning the organizational aspects of biomolecules [70], with great relevance to systems knowledge, such as the network topologies and spatial organization of enzyme interactions that correlate with metabolic efficiency analysis [70].

In this section, some publicly available-databases and the most common workflows used for biological interpretation are described. In this context, Yugi and collaborators [71] provided an important overview about the connection steps between all interaction procedures. They classified the available methods based on regulation levels into five categories: (1) metabolic, (2) transcriptional factors, (3) kinase-substrate relationship (KSR), (4) protein-protein interaction (PPI), and (5) enzyme allosteric regulation by small molecules (Figure 13A) [71]. For each molecule identified, the ID is used to integrate the single-omic approaches (Figure 13B).

Amongst metabolomics researchers, the Kyoto Encyclopedia of Genes and Genomes (KEGG) is the main option for ID conversion, regarding the cross-connection along with the other omic strategies (Figure 13B), *i.e.*, the IDs are used to perform multi-omics analysis and interpretation. KEGG database plays a pivotal role in connecting multiple omics data by ID manipulation, allowing conversion among molecular entities, giving access to different platforms to build pathway maps in order to understand the interaction levels. Another option is MetaboAnalyst [53], a web-based tool suite that host a platform to integrate the omics data based on changes in both gene/protein expression and metabolite concentrations concerning the biochemical pathway and phenotype.

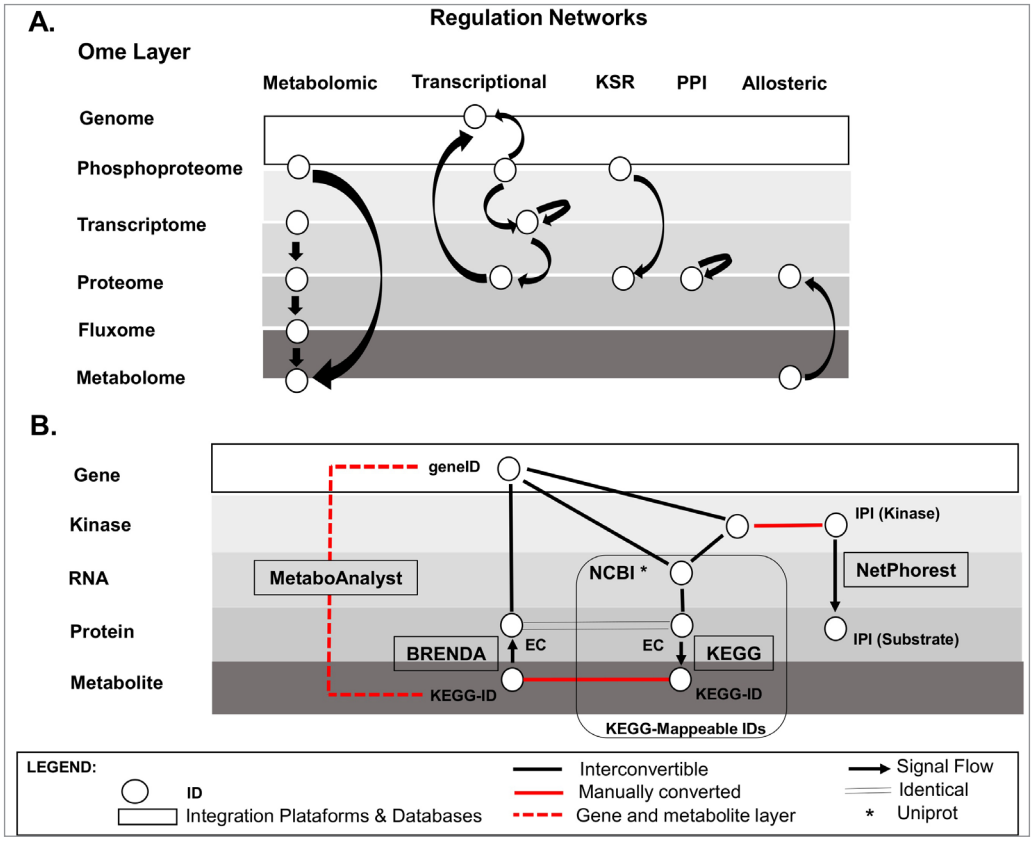


Figure 13. Cross-connection and multi-omics interpretation. **(A)** Single-omic classification and connected layers by regulation IDs. From metabolic to allosteric regulation and omics correlation are indicated in horizontal and vertical positions, respectively. The arrows indicate the directions of regulation processes across the layers. **(B)** Connecting IDs (circles) across multiple omic layers. Lines drawn between circles indicate conversion between IDs. Black lines indicate that an ID association or conversion can be performed by using cross-reference tables provided by, *e.g.*, KEGG. Red lines indicate manual conversions required for IDs. Abbreviations: IPI, International Protein Index; KSR, kinase–substrate relationship; PPI, protein–protein interaction; BRENDA: BRaunschweig ENzyme Database (modified from [71]).

CHALLENGES IN OMIC APPROACHES

The simultaneous integration of multi-omic approaches represents a powerful strategy to disclose the mechanisms connecting identified genetic variations to several conditions. Nevertheless, many sources of variability are combined into statistical models when identifying key drivers and pathways that aim to represent the most significant contributions to a biological process.

As mentioned before, the problems observed in multi-omics analysis begin with the individual omic approaches. In this way, bias is a relevant drawback to emphasize when developing an omics project. Ransohoff has defined bias as "the systematic erroneous association of some characters with a group in a way that distorts the comparison with another group" [72]. It is likely to be introduced because omic measurements are usually performed without a specific hypothesis in mind for lacking some biological reasoning. In reductionist approaches such as the measurement of glucocorticoid levels in response to acute stress [73] or the methylation of DNA as an indicator of repressed gene expression [74], for example, some of the biological relationships these markers present are established and underlie the experiment's initial hypothesis. In this way, the variables are mostly known and tightly controlled. In omic approaches, which refer to the analysis of global sets of biomolecules (*omes*), this is not the case. The complex interactions between genes, proteins, lipids, among other biomolecules, and the interplay with environmental, social, psychological and etiological factors make the associations of cause and effect difficult; thus, many omics-based experiments are poorly defined at the beginning [6,24]. Therefore, the findings from omic approaches need further confirmation by reductionist experiments.

The high potential of biomarkers in the clinical field creates an atmosphere in science urging to discover them as quickly as possible, since these biomarkers can detect diseases and potentially save lives. Although this rush, it is essential that measurements and experiments are designed as carefully as possible with their most addressed drawbacks and limitations. Their implications in the social and health care systems are significant and must be performed with attention.

Moreover, bias is complicated to solve; it cannot be intentionally introduced and is even harder to address its sources. It can be brought if a group is treated differently from other ones. That includes different collection protocols for test and control groups, different protocols for sample extraction and preparation, as well as the number of freezing-thaw cycles and different storage conditions, since the samples can change due to molecular responses to altering conditions [24]. Besides, during analysis, bias might be carried by no proper selection of subjects or participants, such as match of age range or gender among the conditions. For example, if a group of patients has a median age of 70 years and the control group a median of 25, then bias can be introduced due to age. Ethical, genre and socio-economic data should also be addressed and reported, since they are factors that can influence the experiment outcomes [75]. Another source is analytical bias when the measured signal shifts over time. Though, in this case, sample randomization is a good practice to avoid this effect [76].

Different analysis approaches also have different sources of bias (*e.g.*, specimen collection, sample preparation). For example, studies involving disease prognosis by RNA analysis might be affected by the specimen collection, while for the same experiment conducted with DNA, it would be less influenced, since DNA is more stable than RNA. There is no statistical way to solve bias, nor has it any relationship with reproducibility. The best way to minimize its impact is firmly controlling the experimental design and, if possible, addressing its probable sources. If researchers can detect and consider the most likely magnitude and direction of impact of bias, then it can be judged as being present, but not relevant. In any case, bias does not need to be entirely managed; nevertheless, it needs to be considered and the protocols reported in detail, so reviewers and other scientists can interpret the reliability of the study [72].

Another problem with the omic sciences in regard to statistics is called the "Anna Karenina effect". It occurs when the null hypothesis is rejected without any true observation and can also be called "chance". It means that the statistical significance does not equate to biological relevance. Chance is overfitting of the data in a way that discriminates with high precision between test and control groups, which is indicated by the *p*-value. This determines if an observation occurs by chance [77,78]. Many researchers believe that

a low *p*-value is a guarantee of the reproducibility and strength of the study; however, it is not always true. The reproducibility of a statistically significant result, when the probability is true, is substantially lower than one might expect. It does not relate directly to sample size effect, but varies along experimental replicates generated from the same population [79].

Lay Jr. *et al.* [6] suggest that most researchers in the omics field believe that a large number of measurements involved within the approaches could somehow compensate the small sample number. It does not occur, even assuming very low *p*-values such as 0.005. If there is no real association, even increasing the number of samples infinitely cannot render it valid. In order to overcome this issue, overfitting can be assessed through a random dataset for validation and further with external validation – application of the method within an external group to prove its validity [72]. As an alternative, cross-validation should be performed when the latter is not satisfied [80].

The difficulty in reproducing omic results obtained in different laboratories, in order to propose new biomarkers, brings barriers to the development of the field, since biomarkers are not thoroughly validated (Box 1). Therefore, it points out a need for better control and descriptions to follow the protocols applied in the research. Therefore, scientists can have a better understanding of the experimental designs and help to detect the sources of errors and develop alternatives to minimize them.

In order to overcome this issue, the proper design of the study and the control of the pre-analytical variables must be assessed and the influences of the analytical techniques should be reduced through the application of internal controls and standards for calibration within quality control [81]. A random sample selection from a studied population is ideal and should also be matched: the same median age, genres, routines, diet, types of medication, etc. Samples with uncertainty in response diagnoses or other clinical data should be excluded. It is highly indicated sample blocking since it can avoid the introduction of analytical and technical bias. All the steps should be done carefully and described accurately; moreover, standard operating procedures (SOPs) should be established in each stage. Figures of methods: variability, sensitivity and specificity should also be determined [82].

Once again, it is highly essential to demonstrate intra-laboratory and inter-laboratory reproducibility and standardization, reporting a detailed work in all stages. Some researchers claim the codes used in data treatment, which should be submitted together with the scientific articles containing omic studies [83]. Both experts and scientists of diverse fields, such as biostatisticians, program developers, and bioinformaticists, need to work together in order to overcome the difficulties found in omic sciences [6]. It does not seem to be an easy work; however, it brings reliable results. Moreover, it is necessary for the evolution of multi-omic approaches for further clinical application.

Box 1. Stages for biomarker validation [84-87]

Stage I - Pre-clinical exploratory studies
Biomolecule detection strategies (such as proteomics and metabolomics) enable the discovery of candidate biomarkers, achieved by screening through modern imaging techniques and other high-throughput techniques. Those identified markers are prioritized based on diagnostic, prognostic, or predictive characteristics, which may suggest their development to be of clinical use.

Stage II - Development of clinical trial for a disease (validation phase)
Clinical trial, based on the use of non-invasively obtained (non-surgical) samples, has two essential components. The first must record clinical utility, which assays need to be validated for reproducibility and demonstrated to be portable among different laboratories (full validation). The second consideration is that the assessment should analyze for clinical performance in terms of "sensitivity" and "specificity" within the determined preliminaries for the intended clinical use. Very often, biomarkers do not continue beyond this stage due to the lack of reliable and accurate assay tests or validation studies do not confirm whether the markers have a proper sensitivity or specificity to continue their development.

Stage III - Retrospective longitudinal repository studies
Researchers evaluate the sensitivity and specificity of the test for detecting diseases that have not been clinically detected yet. Diseased patients are compared with control patients before their clinical diagnosis, providing evidence of the biomarker ability to detect preclinical disease. If biomarker levels in the studied individuals only present a little divergence from those in control subjects near the time of clinical diagnosis, then the biomarker reveals little promise for screening. Otherwise, when it exhibits significative different levels between disease and control patients within a broader time before diagnosis, the biomarker validation is fulfilled, and it is ready for clinical use.

Stage IV - Screening studies perspective
A positive test triggers procedure for a definitive diagnostic, which is often invasive, and it could lead to an increased burden of economic health care. Thus, the study consists in determining the operational characteristics (disease state at the time of detection) of the biomarker-based screening test in a relevant population by determining the proportion of detection and the proportion of false references.

Stage V - Disease control studies
The final phase is to evaluate how the biomarker test performs in a population. In order to determine whether the screening test reduces the burden (on morbidity and mortality) of the disease in a predefined community, large-scale studies are necessary.

In this review, (1) a description on how multi-omics can enhance the understanding of the biological complexity under a systems perspective was provided; (2) the application of omics data integration to fill gaps generated from individual omic studies and (3) a workflow detailing how multi-omic approaches can be incorporated into filtering protocols that aim at identifying molecular candidates of specific biological processes were discussed; and (4) essential considerations and future directions that are relevant to the success of molecular targets selection supported by multi-omics data were pointed out, in view of the systems biology concept. It is important to emphasize that the omic strategies, as well as mathematical (*e.g.*, statistical) and computational tools, and the integrative methods mentioned in this review, are a subset

of the current methods available and additional ones can also be used to identify biomarker candidates and describe biological systems successfully.

FINAL CONSIDERATIONS

Since the beginning and development of omic sciences, investigations have been directed towards the discovery of biomarkers and their biochemical ability to explain biological conditions, *e.g.*, health-disease, host-parasite, even molecular factors that promote systems' dynamics. These search methods are perpetuated until the present day; however, for some researchers, with many caveats. The main one is the way that the experimental designs have been conducted in order to determine specific molecular patterns based on reductionist methods [88-91], even though these same targets belong to a physiological mechanism of high complexity, being part of a system.

Hence, methods for the integrative analysis of multi-omics data arise as an opportunity to thoroughly describe ideas regarding data mining using an integration algorithm to test the target molecules and data generated from omic approaches under a complex context. At the same time, biomarker discovery approaches require a completer and more accurate picture of the molecular systems' dynamics. The complexity of biological systems, the technological boundaries, the large number of biological variables and the relatively low number of biological samples analyzed are challenges in multi-omics, but it is still essential to recognize it as a potential tool to fill important gaps and respond questions not answered by omic strategies alone.

REFERENCES

1. Nielsen, J. *Annu. Rev. Biochem*., **2017**, *86*, pp 245–275 (http://doi.org/10.1146/annurev-biochem-061516-044757).
2. Kitano, H. *Nature*, **2002**, *420*, pp 206–210 (http://doi.org/10.1038/nature01254).
3. Hasin, Y.; Seldin, M.; Lusis, A. *Gen. Biol*., **2017**, *18*, pp 1–15 (http://doi.org/10.1186/s13059-017-1215-1).
4. Zhang, A.; Sun, H.; Wang, Z.; Sun, W.; Wang, P.; Wang, X. *Planta Med*., **2010**, *76*, pp 2026–2035 (https://doi.org/10.1055/s-0030-1250542).
5. Forshed, J. J. *Proteome Res*., **2017**, *16*, pp 3954-3960 (https://doi.org/10.1021/acs.jproteome.7b00418).
6. Lay Jr., J. O.; Borgmann, S.; Liyanage, R.; Wilkins, C. L. *Trends Anal. Chem*., **2006**, *25*, pp 1046-1056 (https:// 10.1016/j.trac.2006.10.007).
7. Villas-Bôas, S. G.; Mas, S.; Åkesson, M.; Smedsgaard, J.; Nielsen, J. *Mass Spec. Rev*., **2005**, *24*, pp 613-646 (https://doi.org/10.1002/mas.20032).
8. Garlow, S. J. *Neuron*, **2002**, *34*, pp 327-328 (https://doi.org/10.1016/S0896-6273(02)00680-3).
9. Hanash, S. *Nature*, **2003**, *422*, pp 226–232 (https://doi.org/10.1038/nature01514).
10. Fiehn, O.; Kopka, J.; Dörmann, P.; Altmann, T.; Trethewey, R. N.; Willmitzer, L. *Nat. Biotechnol*., **2000**, *18*, pp 1157-1161 (https://doi.org/10.1038/81137).
11. Kopczynski, D.; Coman, C.; Zahedi, R. P.; Lorenz, K.; Sickmann, A.; Ahrends, R. *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids*, **2017**, *1862*, pp 808-811 (https://doi.org/10.1016/j.bbalip.2017.02.003).
12. Boccio, P.; Del Rossi, C.; Cicalini, I.; Sacchetta, P.; Pieragostino, D. *Proteomics Clin. Appl*., **2016**, *10*, pp 470-484 (https://doi.org/10.1002/prca.201500083).
13. Cambiaghi, A.; Díaz, R.; Martinez, J. B.; Odena, A.; Brunelli, L.; Caironi, P.; Masson, S.; Baselli, G.; Ristagno, G.; Gattinoni, L.; De Oliveira, E.; Pastorelli, R.; Ferrario, M. *Sci. Rep*., **2018**, *8*, pp 1-12 (https://doi.org/10.1038/s41598-018-25035-1).
14. Gui, S.-W.; Liu, Y.-Y.; Zhong, X.-G.; Liu, X.; Zheng, P.; Pu, J.-C.; Zhou, J.; Chen, J.-J.; Zhao, L.-B.; Liu, L.-X.; Xu, G.; Xie, P. *Neuropsychiatr. Dis. Treat*., **2018**, *14*, pp 1451-1561 (https://doi.org/10.2147/NDT.S164134).
15. Pinu, F. R.; Beale, D. J.; Paten, A. M.; Kouremenos, K.; Swarup, S.; Schirra, H. J.; Wishart, D. *Metabolites*, **2019**, *9*, pp 1-31 (https://doi.org/10.3390/metabo9040076).

16. Misra, B. B.; Langefeld, C.; Olivier, M.; Cox, L. A. *J. Mol. Endocrinol.*, **2019**, *62*, pp R21-R25 (https://doi.org/10.1530/JME-18-0055).
17. Fondi, M.; Liò, P. *Microb. Res.*, **2015**, *171*, pp 52-64 (https://doi.org/10.1016/j.micres.2015.01.003).
18. Álvares-Sanchez, B.; Priego-Capote, F.; Luque de Castro, M. D. *Trends Anal. Chem.*, **2010**, *29*, pp 111-119 (https://doi.org/10.1016/j.trac.2009.12.003).
19. Tuck, M. K.; Chan, D. W.; Chia, D.; Godwin, A. K.; Grizzle, W. D.; Krueger, K. E.; Rom, W.; Sanda, M.; Sorbara, L.; Stass, S.; Wang, W.; Brenner, D. E. *J. Proteome Res.*, **2009**, *8* (1), pp 113-117 (https://doi.org/10.1021/pr800545q).
20. Lauridsen, M.; Hansen, S. H.; Jaroszewski, J. W.; Cornet, C. *Anal. Chem.*, **2007**, *79* (3), pp 1181-1186 (https://doi.org/10.1021/ac061354x).
21. Cuevas-Córdoba, B.; Santiago-García, *J. OMICS*, **2014**, *18* (2), pp 87-97 (https://doi.org/10.1089/omi.2013.0064).
22. Fan, X.; Peters, B. A; Min, D.; Ahn, J.; Hayes, R. B. *PLoS One*, **2018**, *13* (4), e0194729 (https://doi.org/10.1371/journal.pone.0194729).
23. Teunissen, C. D; Petzold, A.; Bennett, J. L; Berven, F. S.; Brundin, L.; Comabella, M; Franciotta, D.; Frederiksen, J. L.; Fleming, J. O.; Furlan, R.; et al. *Neurology*, **2009**, *73* (22), pp 1914-1922 (https://doi.org/10.1212/WNL.0b013e3181c47cc2).
24. Comes, A. L; Papiol, S.; Mueller, T.; Geyer, P. E.; Mann, M.; Schulze, T. G. *Transl. Psychiatry*, **2018**, *8*, p 160 (https://doi.org/10.1038/s41398-018-0219-2).
25. Salem, M.; Bernach, M.; Bajdzienko, K.; Giavalisco, P. *J. Vis. Exp.*, **2017**, *124*, e55802 (https://doi.org/10.3791/55802).
26. Coman, C.; Solaris, F. A.; Hentschel, A.; Sickmann, A.; Zahedi, R. P.; Ahrends, R. *Mol. Cell. Proteomics*, **2016**, *15* (4), pp 1453-1466 (https://doi.org/10.1074/mcp.M115.053702).
27. Breil, C.; Vian, M. A.; Zemb, T.; Kunz, W.; Chemat, F. *Int. J. Mol. Sci.*, **2017**, *18* (4), p 708 (https://doi.org/10.3390/ijms18040708).
28. Olszowy, P.; Buszweski, B. *J. Sep. Sci.*, **2014**, *37* (20), pp 2920-2928 (https://doi.org/10.1002/jssc.201400331).
29. Gutiérrez, A.; Cerón, J. J.; Razzazi-Fazeli, E.; Schlosser, S.; Tecles, F. *BMC Vet. Res.*, **2017**, *13* (1), p 375 (https://doi.org/10.1186/s12917-017-1296-9).
30. Yu, Y.; Suh, M.-J.; Sikorski, P.; Kwon, K.; Nelson, K. E.; Pieper, R. *Anal. Chem.*, **2014**, *86* (11), pp 5470-5477 (https://doi.org/10.1021/ac5008317).
31. Álvarez-Sánchez, B.; Priego-Capote, F.; Luque de Castro, M. D. *J. Chromatogr. A*, **2012**, *1248*, pp 178-181 (https://doi.org/10.1016/j.chroma.2012.05.029).
32. Katajamaa, M.; Orešič, M. *J. Chromatogr. A*, **2007**, *1158*, pp 318–328 (https://doi.org/10.1016/j.chroma.2007.04.021).
33. Holman, J. D.; Tabb, D. L.; Mallick, P. *Curr. Protoc. Bioinformatics*, **2014**, *46* (1) pp 1–9 (https://doi.org/10.1002/0471250953.bi1324s46).
34. Smit, H. C.; Walg, H. L. *Chromatographia*, **1975**, *8* (7), pp 311–323 (https://doi.org/10.1007/BF02350794).
35. Schiffman, C.; Petrick, L.; Perttula, K.; Yano, Y.; Carlsson, H.; Whitehead, T.; Metayer, C.; Hayes, J.; Rappaport, S.; Dudoit, S. *BMC Bioinformatics*, **2019**, *20* (1), pp 1–10 (https://doi.org/10.1186/s12859-019-2871-9).
36. Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G. *Anal. Chem.*, **2012**, *84* (11), pp 5035–5039 (https://doi.org/10.1021/ac300698c).
37. Libiseller, G.; Dvorzak, M.; Kleb, U.; Gander, E.; Eisenberg, T.; Madeo, F.; Neumann, S.; Trausinger, G.; Sinner, F.; Pieber, T.; Magnes, C. *BMC Bioinformatics*, **2015**, *16* (118), pp 1–10 (https://doi.org/10.1186/s12859-015-0562-8).
38. Petera, M.; Le Corguillé, G.; Martin, J.-F.; Guitton, Y. *LC-MS analysis* (Galaxy Training Materials). Available from: https://training.galaxyproject.org/training-material/topics/metabolomics/tutorials/lcms/tutorial.html [accessed May 15 2020].

39. Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. *Anal. Chem*., **2012**, *84*, pp 283–289 (https://doi.org/10.1021/ac202450g).
40. Gentleman, R. C.; Carey, V. J.; Bates, D. M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; et al. *Genome Biol*., **2004**, *5*, pp R80.1-R80.16 (https://doi.org/10.1186/gb-2004-5-10-r80).
41. Simonet, B. M. *Trends Anal. Chem*., **2005**, *24* (6), pp 525–531 (https://doi.org/10.1016/j.trac.2005.03.011).
42. Gromski, P. S.; Xu, Y.; Hollywood, K. A.; Turner, M. L.; Goodacre, R. *Metabolomics*, **2015**, *11* (3), pp 684–695 (https://doi.org/10.1007/s11306-014-0738-7).
43. Salek, R. M.; Maguire, M. L.; Bentley, E.; Rubstov, D. V.; Hough, T.; Cheeseman, M.; Nunez, D.; Sweatman, B. C.; Haselden, J. N.; Cox, R. D.; et al. *Physiol. Genomics*, **2007**, *29* (2), pp 99–108 (https://doi.org/10.1152/physiolgenomics.00194.2006).
44. Wang, P.; Tang, H.; Zhang, H.; Whitewalker, J.; Paulovich, A. G.; *McIntosh, M. Biocomputing*, **2006**, pp 315–326 (https://doi.org/10.1142/9789812701626_0029).
45. Sauve, A.; Speed, T. *Proc. Gensips*, **2004**, pp 1–4.
46. Wei, R.; Wang, J.; Su, M.; Jia, E.; Chen, S.; Chen, T.; Ni, Y. *Sci. Rep*., **2018**, *8* (1), pp 1–10 (https://doi.org/10.1038/s41598-017-19120-0).
47. Bijlsma, S.; Bobeldijk, I.; Verheij, E.; Ramaker, R.; Kochhar, S.; Macdonald, I. A.; Ommen, B. V.; Smilde, A. K. *Anal. Chem*., **2006**, *78* (2), pp 567–574 (https://doi.org/10.1021/ac051495j).
48. Pleil, J. D. J. *Breath Res*., **2016**, *10* (4) (https://doi.org/10.1088/1752-7155/10/4/045001).
49. Antweiler, R. C.; Taylor, H. E. *Environ. Sci. Technol*., **2008**, *42* (10), pp 3732–3738 (https://doi.org/10.1021/es071301c).
50. Troyanskaya, O.; Cantor, M.; Shelock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R. B. *Bioinformatics*, **2001**, *17* (6), pp 520–525 (https://doi.org/10.1093/bioinformatics/17.6.520).
51. Stacklies, W.; Redestig, H; Scholz, M.; Walther, D.; Selbig, J. *Bioinformatics*, **2007**, *23* (9), pp 1164–1167 (https://doi.org/10.1093/bioinformatics/btm069).
52. Do, K. T.; Wahl, S.; Raffler, J.; Molnos, S.; Laimighofer, M.; Adamski, J.; Suhre, K.; Strauch, K.; Peters, A.; Gieger, C; et al. *Metabolomics*, **2018**, *14* (10), pp 1–18 (https://doi.org/10.1007/s11306-018-1420-2).
53. Chong, J.; Wishart, D. S.; Xia, J. *Curr. Protoc. Bioinformatics*, **2019**, *68* (1), pp 1–128 (https://doi.org/10.1002/cpbi.86).
54. Bruch, M.; Bodden, E.; Monperrus, M.; Mezini, M. Proceedings of the FSE/SDP Work. Fut. Softw. Eng. Res. FoSER, **2010**, pp 53–57 (https://doi.org/10.1145/1882362.1882374).
55. Valdman, J. *Applications from Engineering with MATLAB Concepts*. InTech, Rijeka, Croatia, **2016**, Chapter 8, p. 171.
56. Pesaran, M. H. *J. Appl. Econom*., **2012**, *27*, pp 167–172 (https://doi.org/10.1002/jae).
57. Domon, B.; Aebersold, R. *Mol. Cell. Proteomics*, **2006**, *5*, pp 1921–1926 (https://doi.org/10.1021/jasms.8b04081).
58. Cox, J.; Michalski, A.; Mann, M. *J. Am. Soc. Mass Spectrom*., **2011**, *22*, pp 1373-1380 (https://doi.org/10.1021/jasms.8b04081).
59. Tyanova, S.; Temu, T.; Cox, J. *Nat. Protoc*., **2016**, *11* (12), pp 2301–2319 (https://doi.org/10.1038/nprot.2016.136).
60. Tyanova, S.; Temu, T.; Sinitcyn, P.; Carlson, A.; Hein, M. Y.; Geiger, T.; Mann, M.; Cox, J. *Nat. Methods*, **2016**, *13* (9), pp 731–740 (https://doi.org/10.1038/nmeth.3901).
61. Rohart, F.; Gautier, B.; Singh, A.; Lê Cao, K.-A. *PLoS Comput. Biol*., **2017**, *13*, pp 1-19 (https://doi.org/10.1371/journal.pcbi.1005752).
62. Zhou, G.; Xia, J. *Nuc. Acids Res*., **2018**, *46*, pp W514-W522 (https://doi.org/10.1093/nar/gky510).
63. Ogata, H.; Goto, S.; Sato, K.; Fujibuchi, W.; Bono, H.; Kanehisa, M. *Nuc. Acids Res*., **1999**, *27*, pp 29-34 (https://doi.org/10.1093/nar/27.1.29).
64. Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; et al. *Nuc. Acids Res*., **2013**, *41*, pp 801-D807 (https://doi.org/10.1093/nar/gks1065).

65. Zhou, G.; Xia, J. *Curr. Protoc. Bioinformatics*, **2019**, *65* (1), e69 (https://doi.org/10.1002/cpbi.69).
66. Shannon, P; Markiel, A; Ozier, O; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwilowski, B.; Ideler, T. *Genome Res*., **2003**, *13*, pp 2498-2504 (https://doi.org/ 10.1101/gr.1239303).
67. Singh, A.; Shannon, C. P.; Gautier, B.; Rohart, F.; Vacher, M.; Tebbutt, S. J.; Cao, Kim-Anh. L. *Bioinformatics*, **2019**, *35*, pp 3055–3062 (https://doi.org/10.1093/bioinformatics/bty1054).
68. González, I.; Cao, K. L.; Davis, M. J.; Déjean, S. *BioData Mining*, **2012**, *5*, pp 1-23 (https://doi.org/10.1186/1756-0381-5-19).
69. Conesa, A.; Beck, S. *Sci. Data*, **2019**, *6* (1), pp 1-4 (https://doi.org/10.1038/s41597-019-0258-4).
70. Durek, P.; Walther, D. *BMC Syst. Biol*., **2008**, *2* (1), p 100 (https://doi.org/10.1186/1752-0509-2-100).
71. Yugi, K.; Kubota, H.; Hatano, A.; Kuroda, S. *Trends Biotechnol*., **2016**, *34* (4), pp 276-290 (https://doi.org/10.1016/j.tibtech.2015.12.013).
72. Ransohoff, D. F. *Nat. Rev. Cancer*, **2005**, *5*, pp 142-149 (https://doi.org/10.1038/nrc1550).
73. Cattaneo, A.; Riva, M. A. *J. Steroid Biochem*., **2016**, *160*, pp 169-74 (https://doi.org/10.1016/j.jsbmb.2015.07.021).
74. Curradi, M.; Izzo, A.; Bandaracco, G.; Landsberger, N. *Mol. Cell Biol*., **2002**, *22* (9), pp 3157-3173 (https://doi.org/10.1128/MCB.22.9.3157-3173.2002).
75. Renzi, C.; Provencal, N.; Bassil, K. C.; Evers, K.; Kihlbom, U.; Radford, E. J.; Koupil. I.; Mueller-Myshok, B.; Hansson, M. G.; Rutten, B. P. F. *Progr. Mol. Biol. Transl*., **2018**, *158*, pp 299-323 (https://doi.org/10.1016/bs.pmbts.2018.04.011).
76. Suresh, K. *J. Hum. Reprod. Sci*., **2011**, *4*, pp 8-11 (https://doi.org/10.4103/0974-1208.82352).
77. Goodman, S. N. *Ann. Intern. Med*., **1999**, *130*, pp 995-1004 (https://doi.org/10.7326/0003-4819-130-12-199906150-00008).
78. Goodman, S. N. *Semin. Hematol*., **2008**, *45*, pp 135-140 (https://doi.org/10.1053/j.seminhematol.2008.04.003).
79. Goh, W. W. B.; Wong, L. *Trends Biotechnol*., **2018**, *36*, pp 488-498 (https://doi.org/10.1016/j.tibtech.2018.01.013).
80. Browne, M. W. *J. Math. Psy.*, **2000**, *44*, pp 108-132 (https://doi.org/10.1006/jmps.1999.1279).
81. Lumbreras, B.; Porta, M.; Marquez, S.; Pollán, M.; Parker, L. A.; Hernández-Aguado, I. *Proteomics Clin. Appl.*, **2009**, *3*, pp 173-184 (https://doi.org/10.1002/prca.200800092).
82. Parikh, R.; Mathai, A.; Parikh, S.; Sekhar, G. C.; Thomas. R. *Indian J. Ophtalmol*., **2008**, *56*, pp 45-50 (https://doi.org/ 10.4103/0301-4738.37595).
83. Baggerly, K.; Coombes, K. *Clin. Chem*., **2011**, *57*, pp 688-690 (https://doi.org/10.1373/clinchem.2010.158618).
84. Kumar, M.; Sarin, S. K. *Curr. Trends Sci*., **2010**, *15*, pp 403-417.
85. Pepe, M. S.; Etzioni, R.; Feng, Z.; Potter, J. D.; Thompson, M. L.; Thornquist, M. *J. Natl. Cancer Inst*., **2001**, *93*, pp 1054-1061 (https://doi.org/10.1093/jnci/93.14.1054).
86. Sataloff, R. T. *Ear Nose Throat J*., **2004**, *83*, pp 665-665 (https://doi.org/10.1177/014556130408301001).
87. EDRN. Early Detection Research Network Manual of Operations [Internet]. **2009**. Available from: https://edrn.nci.nih.gov/docs [accessed December 10, 2019].
88. Sarkar, S. *Genetics and Reductionism*. Cambridge University Press, Cambridge, **1998**.
89. Li, R.; Ma, T.; Gu, J.; Liang, X.; Li, S. *Sci. Rep*., **2013**, *3*, pp 1543-1549 (https://doi.org/10.1038/srep01543).
90. Chibbaro, S.; Rondoni, L.; Vulpiani, A. *Reductionism, Emergence and Levels of Reality*. Springer International Publishing, **2014**.
91. Calvani, R.; Picca, A.; Cesari, M.; Tosato, M.; Marini, F.; Manes-Gravina, E.; Berbanei, R.; Landi, F.; Marzetti, E. *Curr. Protein Pept. Sci*., **2018**, *19*, pp 639-642 (https://doi.org/10.2174/1389203718666170516115422).

Supplemental Table I. (Bio)chemical techniques to enrich and characterize the main PTM by mass spectrometry analyses

Techniques	Enrichment & Purification methods		PTMs		Large-scale analysis application	Published protocols (examples)*
Biochemical	Enzymatic Labeling		O-GLcNAc S-glutathionylation		Long preparation cycle & multiple steps	1-3
	Immunoaffinity		Acetylation - Lys		Yes	4,5
			Methylation	Lys	Yes, but poor specificity	6-8
				Arg	Yes, but poor specificity	
			Ubiquitination - Lys		Yes	9-11
			Phosphorylation	Tyr	Yes	12,13
				His	No, but poor specificity	14-16
Chemical	Chromatography	IMAC	Phosphorylation	Ser, Thr and Tyr	Yes	17 - 19
		HILIC	N-linked glycopeptides		Yes, relatively poor specificity	20 - 23
		Boric Acid	N-linked glycopeptides		No, poor specificity	23, 24
	Chemical Derivatization	Hydrazide chemistry	N-linked glycopeptides		Yes, high specificity	25,26
		Biotin switch technique	Redox Modification		Comprised by incomplete reaction and side reaction	27-29
		Direct reductive methylation	Cysteine Oxidation		Long preparation cycle & multiple steps	30,31

*In these articles, the authors provide other references concerning protocols and the main parameters for MS analyses.

Supplemental Table I – References

1. Pineda-Molina, E.; Klatt, P.; Vázquez, J.; Marina, A.; de Lacoba, M. G.; Pérez-Sala, D.; Lamas, S. *Biochemistry*, **2001**, *40*, pp 14134-14142 (https://doi.org/10.1021/bi011459o).
2. Cao, W.; Cao, J.; Huang, J.; Yao, J.; Yan, G.; Xu, H.; Yang, P. *PLoS ONE*, **2013**, *8*, e76399 (https://doi.org/10.1371/journal.pone.0076399).
3. Wu, H. Y.; Lu, C. T.; Kao, H. J.; Chen, Y. J.; Chen, Y. J.; Lee, T. Y. *BMC Bioinf*., **2014**, *15*, S1 (https://doi.org/10.1186/1471-2105-15-S16-S1).
4. Junot, C.; Pruvost, A.; Créminon, C.; Grognet, J. M.; Benech, H.; Ezan, E. *J. Chromatogr. B: Biomed. Sci. Appl.*, **2001**, *752*, pp 69-75 (https://doi.org/10.1016/S0378-4347(00)00520-X).
5. Thao, S.; Escalante-Semerena, J. C. *Curr. Opin. Microbiol*., **2011**, *14*, pp 200-204 (https://doi.org/10.1016/j.mib.2010.12.013).
6. Wang, Q.; Liu, Z.; Wang, K.; Wang, Y.; Ye, M. *Anal. Chim. Acta*, **2019**, *1068*, pp 111-119 (https://doi.org/10.1016/j.aca.2019.03.042).
7. Chen, Y. *Methods Mol. Biol*., **2016**, *1410*, pp 23-37 (https://doi.org/10.1007/978-1-4939-3524-6_2).
8. Wesche, J.; Kühn, S.; Kessler, B. M.; Salton, M.; Wolf, A. *Cell. Mol. Life Sci*., **2017**, *74*, pp 3305-3315 (https://doi.org/10.1007/s00018-017-2515-z).
9. Bustos, D.; Bakalarski, C. E.; Yang, Y.; Peng, J.; Kirkpatrick, D. S. *Mol. Cell. Proteomics*, **2012**, *11*, pp 1529-1540 (https://doi.org/10.1074/mcp.R112.019117).
10. Xu, G.; Paige, J. S.; Jaffrey, S. R. *Nat. Biotechnol*., **2010**, *28*, pp 868-873 (https://doi.org/10.1038/nbt.1654).
11. Valdés, A.; Bergström, S. L. *Proteomics*, **2020**, *20*, 1800425 (https://doi.org/10.1002/pmic.201800425).
12. Guo, A.; Gu, H.; Zhou, J.; Mulhern, D.; Wang, Y.; Lee, K. A.; Yang, V.; Aguiar, M.; Kornhauser, J.; Jia, X.; et al. *Mol. Cell. Proteomics*, **2014**, *13*, pp 372-387 (https://doi.org/10.1074/mcp.O113.027870).
13. Liang, Y.; Zhu, X.; Zhao, M.; Liu, H. *Methods*, **2012**, *56*, pp 174-179 (https://doi.org/10.1016/j.ymeth.2011.08.006).
14. Fuhs, S. R.; Meisenhelder, J.; Aslanian, A.; Ma, L.; Zagorska, A.; Stankova, M.; Binnie, A.; Al-Obeidi, F.; Mauger, J.; Lemke, G.; et al. *Cell*, **2015**, *162*, pp 198-210 (https://doi.org/10.1016/j.cell.2015.05.046).
15. Kee, J. M.; Oslund, R. C.; Perlman, D. H.; Muir, T. W. *Nat. Chem. Biol*., **2013**, *9*, pp 416-421 (https://doi.org/10.1038/nchembio.1259).
16. Adam, K.; Lesperance, J.; Hunter, T.; Zage, P. E. *Int J. Mol. Sci*., **2020**, *21*, p 3319 (https://doi.org/10.3390/ijms21093319).
17. Ruprecht, B.; Koch, H.; Medard, G.; Mundt, M.; Kuster, B.; Lemeer, S. *Mol. Cell. Proteomics*, **2015**, *14*, pp 205-215 (https://doi.org 10.1074/mcp.M114.043109).
18. Gan, C. S.; Guo, T.; Zhang, H.; Lim, S. K.; Sze, S. K. *J. Proteome Res*., **2008**, *7*, pp 4869-4877 (https://doi.org/10.1021/pr800473j).
19. Kielkopf, C. L.; Bauer, W.; Urbatsch, I. L. *Cold Spring Harbor Protocols*, **2020** (https://doi.org/10.1101/pdb.prot102194).
20. Shu, Q.; Li, M.; Shu, L.; An, Z.; Wang, J.; Lv, H.; Yang, M.; Cai, T.; Hu, T.; Fu, Y.; Yang, F. *Mol. Cell. Proteomics*, **2020**, *19*, pp 672-689 (https://doi.org/ 10.1074/mcp.RA119.001791).
21. Zhang, H.; Lv, Y.; Du, J.; Shao, W.; Jiao, F.; Xia, C.; Gao, F.; Yu, Q.; Liu, Y.; Zhang, Y.; et al. *Anal. Chim. Acta*, **2020**, *1098,* pp 181-189 (https://doi.org/10.1016/j.aca.2019.11.012).
22. Qing, G.; Yan, J.; He, X.; Li, X.; Liang, X. *Trends Anal. Chem.*, **2020**, *124*, 115570 (https://doi.org/10.1016/j.trac.2019.06.020).
23. Saleem, S.; Sajid, M. S.; Hussain, D.; Jabeen, F.; Najam-ul-Haq, M.; Saeed, A. *Anal. Bioanal. Chem*., **2020**, *412*, pp 1509-1520 (https://doi.org/10.1007/s00216-020-02427-9).
24. Jiang, B.; Huang, J.; Yu, Z.; Wu, M.; Liu, M.; Yao, J.; Zhao, H.; Yan, G.; Ying, W.; Cao, W.; Yang, P. *Talanta*, **2019**, *199*, pp 254-261 (https://doi.org/10.1016/j.talanta.2019.02.010).

25. Sajid, M. S.; Jabeen, F.; Hussain, D.; Ashiq, M. N.; Najam-ul-Haq, M. *Anal. Bioanal. Chem*., **2017**, *409*, pp 3135-3143 (https://doi.org/10.1007/s00216-017-0254-5).
26. Napoli, A. Use of Galactose Oxidase in Hydrazide-Capturing Technique to Isolate N-linked Glycoallergens from Olive Tree Pollen. In: *IOP Conference Series: Materials Science and Engineering*, **2020**, *739* (1), 012052. IOP Publishing.
27. McDonagh, B.; Ogueta, S.; Lasarte, G.; Padilla, C. A.; Bárcena, J. A. *J. Proteomics*, **2009***, 72*, pp 677-689 (https://doi.org/10.1016/j.jprot.2009.01.023).
28. Bykova, N. V.; Rampitsch, C. *Proteomics*, **2013**, *13*, pp 579-596 (https://doi.org/10.1002/pmic.201200270).
29. Zhang, T.; Gaffrey, M. J.; Qian, W. J.; Thrall, B. D. Oxidative Stress and Redox Modifications in Nanomaterial–Cellular Interactions. In: Bonner, J. C.; Brown, J. M. (Eds.). *Interaction of Nanomaterials with the Immune System*. Springer, Cham, **2020**, pp 127-148.
30. McShane, A. J.; Shen, Y.; Castillo, M. J.; Yao, X. *J. Am. Soc. Mass Spectrom*., **2014**, *25*, pp 1694-1704 (https://doi.org/10.1007/s13361-014-0951-7).
31. Alcock, L. J.; Perkins, M. V.; Chalker, J. M. *Chem. Soc. Rev*., **2018**, *47*, pp 231-268 (https://doi.org/10.1039/C7CS00607A).