Bachelor's Degree Final Project

COMPARATIVE ANALYSIS BETWEEN TWO R-BASED SOFTWARE FOR PEAK ANNOTATION IN UNTARGETED METABOLOMICS DATA

Degree in Biochemistry and Molecular Biology

Author

Júlia Vall Carabasa

Tutors

Sra. Sara Martínez de Cripán (Eurecat)
Dr. Jorge Ricardo Soliz Rueda (URV)

UNIVERSITAT ROVIRA i VIRGILI Tarragona, 2023

This project is the result of the external internship carried in the Computational Metabolomics for System Biology (Metsyslab) group at the EURECAT's Centre for Omic Sciences, Reus, (EURECAT) under the supervision of Mrs. Sara Martinez de Cripan and Dr. Xavier Domingo-Almenara.

ABSTRACT

The analysis of small molecules, known as metabolites, is applied in almost all facets of life sciences, but the identification of metabolites remains a significant challenge in untargeted liquid chromatography-mass spectrometry-based metabolomics (LC/MS). The term annotation is referred to the process which consist on associating ion features (adducts, isotopes and in-source fragments) to molecules derived from the same compound which provide valuable chemical information for later achieving the identification of metabolites. It is a crucial step when performing a metabolomics LC/MS-based assay and its completion is achieved through bioinformatic tools like software and computational data analysis. In the present study, two widely used software programs, CAMERA and CliqueMS, for peak annotation using the R programming language were compared. Different functions from these R packages were applied to annotate adducts, neutral losses, and isotopes in 29 samples. Additionally, the pseudospectra associated with each feature group with reference metabolites spectra from NIST database were evaluated. It was perceived that the software selection would be contingent upon the properties of the samples and the experimental objectives. Overall, it was concluded that while CliqueMS provides more detailed annotation, CAMERA pseudospectra generally aligned more accurately with reference spectrums, resulting in improved metabolite identity association.

Keywords:
Mass spectrometry, metabolomics, adducts, peaks, metabolites, pseudospectrum, CAMERA, CliqueMS.

INDEX

1. INTRODUCTION

1.1. Fundamentals and workflow of mass spectrometry

Mass spectrometry (MS) is an analytical technique which precisely measures the molecular masses of individual compounds and atoms by converting them into charged ions (1,2). MS is one of most precise, fast and reliable methods available to determine molecular and atomic masses of analytes with high accuracy in a single measurement (3). It is efficiently used in many fields to obtain analytical information, mainly molecular weight and chemical structure from molecules such as carbohydrates, amino acids, proteins, DNA and other biologically relevant compounds. This instrumental technique was reported for the first time in 1912 as a parabola spectrogram. Since then, many improvements have been made and nowadays has become an irreplaceable tool in considerable scientific approaches (4). This rich informational technique is based on ions production which are thereafter separated in relation to their mass-to-charge ratio (m/z) and finally detected (5).

The mass spectrometer is a highly sophisticated and computerized instrument which basically consist of five parts (Figure 1): sample introduction, ionization, mass analysis, ion detection, and computational analysis (6).

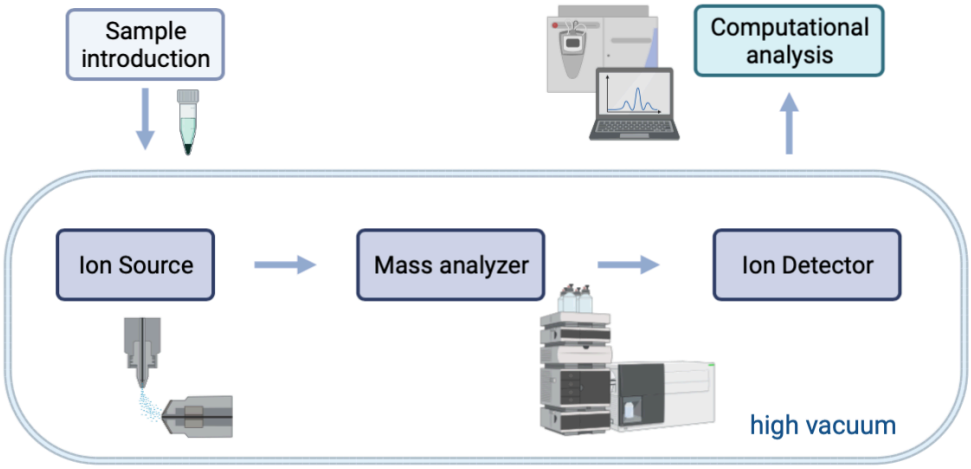


Figure 1 *- General workflow of a mass spectrometry analysis. Samples are introduced, a ion source is applied on them, in the mass analyzer they are sorted according to mass-to-charge ratio and finally detected to allow the computational analysis.*

The experimental part of the process that occur in a mass spectrometer starts with **ionization**, this first step converts analyte molecules or atoms into gas-phase ionic species. It requires the addition or removal of an electron or proton (1). A broad diversity of ionization techniques are available for MS and can be classified as 'soft' or 'hard', depending on the fragmentation degree that takes place during the ionization process (7). An example of a soft ionization method is electrospray ionization (ESI) in liquid chromatography (LC), one of the most widely used mass spectrometry techniques that allows the generation of positively and negatively charged ions. It uses electrical energy to transfer ions from a solution to the gas phase for analysis. ESI is a

powerful technique that allows for the analysis of a wide range of ionic species in solution as well as neutral compounds that can be converted to their ionic form with high sensitivity and resolution (8–10). Another example of a soft ionization method that can also be performed in LC is Matrix-assisted laser desorption/ionization (MALDI). Whereas in gas chromatography, mainly electron ionization (EI) is carried out (11,12). EI causes extensive fragmentation, which provides structural information for interpreting unknown spectra. However, molecular ion (M+) is not observed for many compounds in EI. In this technique when energetic electrons bombard vapor, they can elastically scatter or interact with molecules to cause electron excitation. The ions generated during EI have varying internal energies, which can lead to unique unimolecular dissociation reactions and the formation of fragment ions. These reactions may result in the loss of either a radical or a neutral (13).

After the ionization process, every chemical compounds produce one or multiple ion species such as isotopologue ions, fragment ions, and particularly in ESI, adduct and cluster ions. The emitted ions are then sampled and accelerated into the mass analyser for subsequent analysis of molecular mass and measurement of ion intensity (8,14). Then, the process is followed by the chromatographic separation and sorting of the obtained ions according to mass-to-charge ratio (m/z), this process takes place in the **mass analyzer**, a component of the mass spectrometer. A large amount of mass analysers uses magnetic or electric fields to control the motion of ions (2). The most usually utilized mass analysers are Quadrupole Mass, Time of Flight, Magnetic/Electrostatic Double Sector, Quadrupole Ion Trap and Ion Cyclotrol Resonance (2). Mass analysers can be classified depending on the physical properties by which ions having different m/z are separated such as TOF of magnetic sector. Additionally, they can also be grouped in relation to the operation mode with reference to the generation of the ion beam. Therefore, some instruments have a continuous mode of operation such as quadrupoles and magnetic sectors which are scanning instruments. Some others, like TOF, use a pulsed-based operation mode. And finally mass analysers including QITs, ICRcells and orbitraps use an ion trapping mode (15). All mass analyzers have advantages and disadvantages, and there is not a single instrument that is ideal for all applications and experiments (16). Nonetheless, Quadrupole Time of Flight (q-TOF) is an analytical technique that has gained considerable attention over the last decade. It consists of an instrument that uses the hight compound fragmentation efficiency of quadrupole Technology in combination with the fast-analysing process and high mass resolution that characterizes TOF (17,18).

When ions have completed their way through the mass analyzer, the procedure is finalized in a detector. The process starts when the electron beam hits a cathode, releasing electrons, then a series of dynodes placed at increasingly higher potentials amplify the electron current to finally be detected as the **detection** system turns the electron beam into an electric signal and simultaneously measures the ion current while information about m/z ratios and relative abundance is converted into operable data that is stored in a computer displayed in the form of a mass spectrum and further can be processed and used for comparison with suitable reference standards (2,3,19). As it is explained before, most mass analyzers produce a beam of ions that can be detected as an electrical current. The result of changing the external magnetic field strength, ions of different mass will be detected because they will focus distinctively on the detector. There are some detectors that count ions and others that measure ion current.

Different kind of devices like electron multipliers, multichannel plates and photomultipliers are used to detect the ion beam. The most common is electron multiplier which consist of a vacuum tube that multiplies incident charges (15).

These procedures, which are represented in Figure 2 (ionization, separation and detection), are carried out under a high vacuum. Ions have a short live and are immensely reactive, this environmental condition allows ions to move freely in space avoiding the interaction between other species and collision production (1).

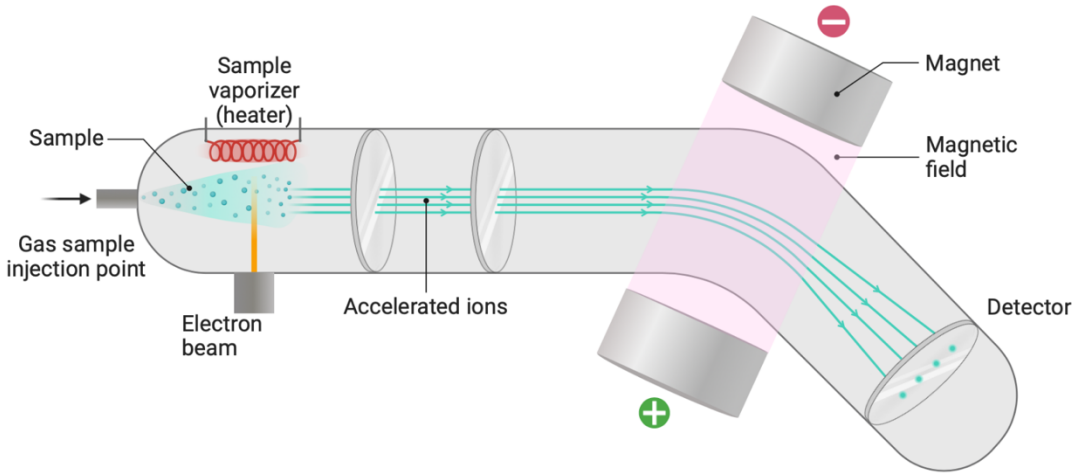


Figure 2 *- Detailed mass spectrometry procedure from ionization to detection. Sample is introduced and ionized by the electron beam, ions are accelerated into the mass analyzer device. Then, thanks to the magnetic field applied, ions are separated according to mass-to-charge ratio and finally, the detection system turns the electron beam into an electric signal to store the obtained information in a computer device. (Biorender)*

To conclude the process, the acquired data is processed through a **computational analysis**. Many different types of data can be collected to perform quantitative analysis, determination of sequencies or molecules and definition of chemical structures. Nevertheless, the resulting mass spectrum is a plot where the mass-to-charge ratio is represented against its intensity or relative abundance.

If the mass spectrum reveals the m/z of the molecular ion, the analyte's molecular mass can be calculated. A variety of adduct ions can be seen in the mass spectrum using the soft ionization, the protonated molecule in the positive-ion mode or the deprotonated molecule in the negative-ion mode. The use of the positive ionization mode, which results in the m/z of the [M+H] + ion, and the negative ionization mode, which results in the m/z of the [M-H]- ion, usually leads to an unambiguously molecular mass determination for an unknown compound (6).

Although it represents the lowest level of annotation possible, annotating measured m/z values with potential metabolites is typically one of the initial steps in metabolite identification workflows. Various tools have been proposed for this task to accomplish MS annotation (22,23). This project is particularly focused in MS-based annotation, which involves comparing the measured m/z values and/or retention times of LC-MS features with reference values (24).

1.2. Metabolomics, a powerful omic science

Biological entities are systems, the collection of simple parts that work together as a single unity. Systems biology is an integrative discipline that connects molecular components within a single or multiple biological scales to physiological functions and phenotypes (25). This biology-based field is subject to computational and mathematical analysis from experimental data which provide the understanding of complex interactions and dynamics at various levels, within cells, tissues, organs and organisms (26). Systems biology is sustained by the integration of the whole structural and functional information acquired from omics sciences, which are genomics (containing metagenomics and epigenomics), transcriptomics, proteomics and metabolomics (Figure 3) (26). Increased availability of high-throughput technologies has generated an ever-growing number of omics data that seek to portray many different but complementary biological layers including all the omics (27). On the one hand, genomics, transcriptomics and metabolomics are increasingly generating key insights for the development of precision approaches in health and disease. Omic technologies have unlocked the high-throughput discovery of diagnostic biomarker and the systems-level evaluation of the efficacy and toxicity of novel therapies (28). On the other hand, metabolomics studies focus on the investigation of the complex and dynamic biochemical interactions of metabolites with other biochemicals and their environment (29).

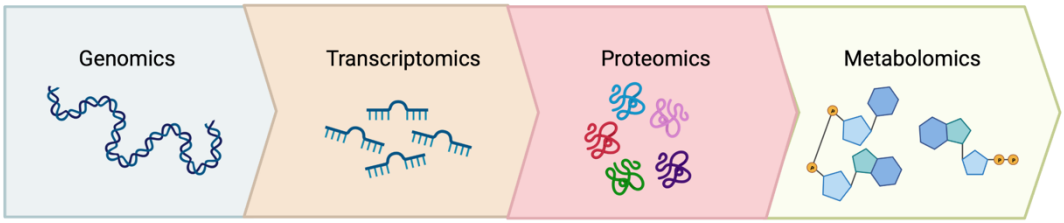


***Figure 3** - The omics cascade which starts with genomics followed by transcriptomics, then proteomics and finally metabolomics.*

This project is only focused on metabolomics, which is a spreading omics field arising after genomics, transcriptomics, and proteomics, and it is a vital part of systems biology (30). Because it deals with small molecule products, environmental interactions and as a result it provides additional information in comparison to other omics (31). Metabolites are the end products of complex cellular regulation networks (32), they can also influence or even alter metabolic pathway regulation (33). The term metabolite refers to a large number of compounds that belong to multiple categories, such as amino acids, lipids, nucleotides, carbohydrates and organic acids. Thus, metabolomic research is a challenge in analytical chemistry because there is no an universal method for metabolome analysis, chemical properties differ a lot between metabolites, there is a dynamic range of the metabolome (34) and compounds are frequently distributed over a broad extent of concentrations (35). Moreover, a concurrent quantification of many metabolites contained in complex samples can be complicated as well, owing to ion suppression, fragmentation and the presence of isomers. Therefore, it is a sophisticated job to get a reliable quantification and a proper compound identification as it is said before (34).

One of the most common analytical techniques used for acquiring metabolomic data are liquid and gas chromatography coupled to mass spectrometry (36,37). In fact, mass spectrometry-based metabolomics is one of the key technologies to detect and identify the small molecules due to its high sensitivity, throughput and speed (30,38). MS-based approaches have permitted the analysis of a huge amount of chemicals with diverse structures. MS has been used for detecting and quantifying metabolites from different type of samples (39,40). Furthermore, metabolomic studies can help us to enhance the understanding of disease mechanisms and drug effects, as well as to improve the ability to predict personal disease progression or variation in drug response phenotypes. In the end, the discovery of significant metabolic changes provides insights for fundamental understanding of biological mechanisms (41). Application of metabolomics have been described for the fields of plant biology and agronomy, disease diagnosis, cancer research, metabolic mechanisms, drug efficacy and screening and nutritional studies (42).

There are two methods commonly used to conduct a metabolomic experiment: targeted and untargeted. The targeted approach involves identifying compounds before quantifying them to detect differences. Meanwhile, untargeted metabolomics utilizes chemometric processing of spectral features from two or more sample sets to determine significant differences (31). That is why the present project is particularly focused on untargeted metabolomics because this kind of studies allow the research of thousands of diverse metabolites in samples without prior chemical identification of metabolites, the goal is to detect the maximum number of metabolites possible, in order to increase the chances of identifying compounds that may be dysregulated in a specific biological state (31,43,44).

The major seps which constitute a metabolomic data processing pipeline are explained next (Figure 4). The first step is based on **metadata organization**. In order to address pertinent scientific inquiries, the field of metabolomics must adopt technologies and workflows that facilitate the generation of compatible bioscience data (45–47). Afterwards, the **data acquisition and quality assessment** stage is managed by the instrument control software, which is in charge of executing the sequential analysis of samples included in a list. This list might be imported from an external file or rather manually generated within the instrument software. In order to reduce human intervention, the first option is recommended, since the data analysis pipeline can generate the sequence file. It is crucial to continuously monitor the analysis's quality and to randomize the sample list to eliminate possible perturbations (45,48). Later, **data is converted, stored and organized**. Mass spectrometers save raw data in restricted formats, which creates challenges for data exchange and comparison when multiple instruments from different vendors and models are used. Ideally, raw data could be converted into open formats to facilitate analysis and submission to public repositories. However, the conversion process is complex and certain essential analytical information cannot be extracted easily from the mass spectrometer and chromatograph (45,48). Subsequently**, data processing** involves summarizing data into a matrix containing experimental variable intensities, which are commonly mass-to-charge ratio and retention time in GC/MS and LC/MS. Chromatographic alignment is necessary to ensure variable comparison across all samples, feature grouping based on chromatographic or chemical information and feature correlation can increase grouping selectivity. Finally, the final data matrix can be generated by using the most significant features of each metabolite and

determine its relative concentration (45,49). The next step consists on the association of experimentally obtained features to specific metabolites (50), in Figure 4 this process is named **annotation**, nonetheless later in the project this term will be adapted. In order to identify certain compounds, it is necessary to analyse a pure chemical standard under the same chromatography conditions and MS methods (51). Another option is to use pure MS databases containing full scan information and MS/MS spectra, which can be freely available online (52). Annotation can be improved by using external information and biological relationships between annotated metabolites (45). Then, the **statistical analysis** is required to be performed. It is a complex research field as the datasets themselves often exhibit variability due to biological and technical factors, which must be accounted for through normalization and quality control (pool of samples utilized to reduce the size of analytical errors or calibration) (45,53–56). Finally an essential step is to **submit** data to public repositories because it must be accessible to the research community, enabling independent verification of the results and contributing to incremental scientific progress (45,48).

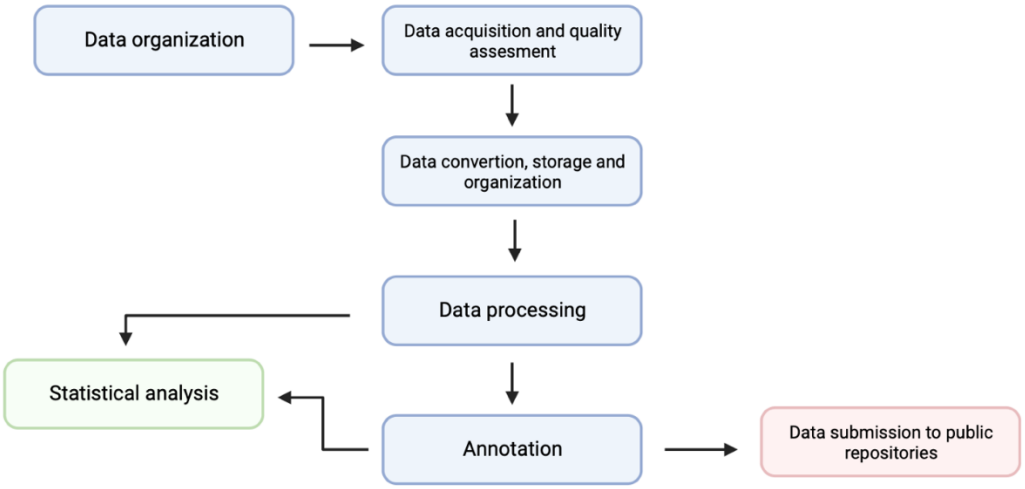


Figure 4 *- Metabolomics workflow for a mass spectrometry dataset. Initially data is organized, then it is acquired and a quality assessment is performed, after that, data is converted, stored and organized to finally get it processed and annotated. A statistical analysis can be performed and data should be submitted to public repositories. Adapted from (45).*

The following part in the introduction is entirely focused on the annotation step as it represents a big challenge for untargeted metabolomic studies and take the major interest in this project.

1.3. Annotation, a key bioinformatic step in computational workflow

Mass spectrometry (MS) is the highly preferred technology for metabolomic studies due to its enhanced accuracy, sensitivity, and coverage. To achieve greater separation of the sample, chromatography is frequently paired with mass spectrometry. Both, gas chromatography (GC) and LC have been utilized in metabolomics research (57,58). The employment of liquid chromatography coupled with mass spectrometry (LC-MS) has become more prevalent in untargeted metabolomic studies because it enables the separation of compounds without

requiring derivatization. ESI is a widely used technique in LC/MS that helps create intact molecular ions and aids in the preliminary identification of metabolites (59).

When conducting large metabolic biomarker discovery studies with hundreds of samples, it's necessary to perform multiple experiments with subsets of samples to avoid long analytical runs or preparing a large number of samples simultaneously. The selected ions from each experiment are compared to identify overlapping ions. However, one challenge is to identify significant number of derivative ions, including isotopes, adducts, and fragments, which are not usually acknowledged. This lack of recognition can lead to incorrect metabolite identification when using a mass-based search, as databases assume each derivative is a distinct molecular ion. To improve metabolite identification accuracy, it's essential to recognize ions derived from the same metabolite (31). The computational workflow for analysing untargeted data obtained from LC/MS includes the appliance of peak-picking algorithms, followed by the alignment of those peaks across multiple samples to obtain peak features, defined as a peak or a group of aligned peaks across samples with unique m/z and retention time values and the process is finalized by peak annotation (60,61). After that, features can be recognized through fragmentation, applying MS or MS/MS. Finally, metabolites can be identified by comparing the obtained features patterns to spectrums from reference libraries (61,62).

Grouping and annotation[1] of computational features are essential steps to reduce the number of putative identities. In this project the term annotation is referred to the process which consist on associating **ion features** (adducts, isotopes and in-source fragments) to molecules derived from the same compound which provide valuable chemical information for later achieving the identification of metabolites and determination of the monoisotopic or neutral molecular mass for each putative metabolite.

To clearly understand the annotation process it is necessary to comment on the fundamentals of ion annotation (12). (Figure 5) A single metabolite may be represented by multiple peaks that have distinct m/z values but similar retention times. Those peaks are generated by three types of ions in LC/MS data: adducts, isotopes, and in-source fragments (31). In the first place an **adduct** ion refers to an ion that is created through the interaction of two species, typically an ion and a molecule, usually formed within the ion source. The resulting ion contains all the atoms of one specie and one or more additional atoms. Metabolomic measurements obtained from LC/MS produce protonated and deprotonated ion species under normal conditions. However, the addition of ionic species like Na+, K+, Cl−, F−, acetate, ammonium, and other additives commonly found in solvents and samples or added to enhance chromatographic and ionization conditions can result in the formation of different molecular adducts (31,63,64). The term **isotope** refers to different variants of atoms of the same chemical element, which possess an identical number of protons but varying numbers of neutrons. Consequently, atoms of the same element can have diverse masses due to the number of neutrons they contain. The majority of metabolites have at least one stable, naturally-occurring isotope, resulting in metabolite

[1] Annotation can be defined as the process of notifying each observed feature with a putative identity but also to the assignment of formed adducts, neutral losses, isotopes and in-source fragments to each obtained feature which is achieved by comparing the mass differences between experimental peaks with the differences between combinations of established adducts, neutral losses, molecular multimers, or ions with multiple charges.(61)

samples being a blend of various isotopic species. During mass spectrometry analysis, distinct isotopic species are isolated, generating a sequence of peaks differentiated by a difference of approximately one Da in m/z. The peak with the lowest m/z among them is known as the monoisotopic peak (31). **In-source fragments** cause the third type of ions. While ESI is commonly considered as a soft-ionization method that predominantly produces intact molecular ions, fragmentation can still occur during the process. Common in-source fragments mainly arise from neutral losses like H2O or CO2 which are lost molecules during the ionization process and can also be detected in mass spectrums.

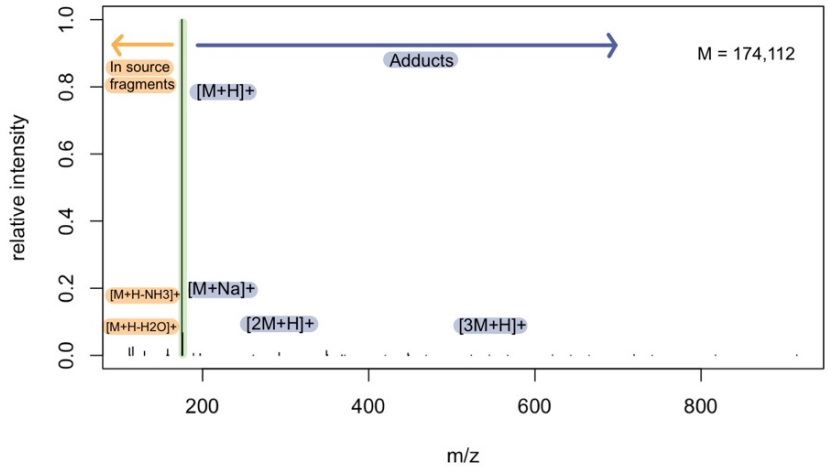


Figure 5 *– Example of a mass spectrum obtained from the annotation process. On the left hand side of the protonated mass (peak in green) we can find in source fragments (highlighted in orange), which have lower molecular weight than the metabolite and on the right hand side of the protonated mass there are adducts (highlighted in blue), molecules with higher molecular mass than the metabolite.*

A strategy for annotating metabolomics data is to extract pseudospectra[2] by determining which peaks belong to each metabolite or compound in study. Each metabolite should have a pseudospectrum that includes its adducts, isotopes, common neutral losses, in-source fragments, and any other peaks that are deemed to originate from the same molecule based on a specific metric. Searching for mass relationships within a defined group of peaks, helps to reduce the number of false positive features that may result in inaccurate annotations (63).

Even though the use of chromatography with MS or MS/MS provides specificity in annotating and identifying compounds, there are common problems that contribute to peak **misidentification**. In the first place, the presence of **isomers** which are compounds with an identical molecular formula but present different structures. The use of high-resolution MS in isolation may not be adequate to differentiate between sets of isomers, particularly when their fragmentation patterns are alike (38,65). The second issue refers to the presence of **overlapping compounds** that can hinder the detection of certain metabolites. Despite the improving resolution of mass spectrometers, the ability of current instruments to differentiate between ions with a mass difference of less than 5 ppm is limited. Nevertheless, this problem is exacerbated only when chromatography fails to separate analytes that cannot be distinguished

[2] The term pseudospectra refers to an artificially generated spectra that simulate the fragmentation patterns of metabolites. These spectra are constructed based on known fragmentation rules and can be used as references for identification and annotation of metabolites in mass spectrometry-based metabolomics experiments. They serve as valuable tools which allow to match experimental spectra with theoretical or simulated spectra to identify and confirm the presence of specific metabolites in a samples.

based on mass (38). Finally, in LC/MS, a significant obstacle is the generation of **in-source degradation products**, which is not as common in GC-MS. These degradation products are often observed in low energy spectra with varying relative intensities and decrease the parent ion's signal intensity of the metabolite, and the resulting fragment ions can complicate the analysis of other compounds that co-elute. This is particularly problematic if they share the same molecular formula as another metabolite's molecular ion. Moreover, the exact mass of in-source fragments is specific to each metabolite, unless it could serve as an identity indicator it is difficult to be predicted. (38,63,66,67)

In the past few years, many new algorithms and computational tools for improving this annotating step in MS-based metabolomics have been introduced (38). However, a proper metabolite identification is still a big struggle in untargeted metabolomics since only a small fraction of the thousands of metabolites in samples can be annotated and identified at a satisfactory confidence level (68). This can be associated to different facts: during the sample process there is a high redundancy of features which are linked to the same metabolite owing to the existence of many in-source fragments, isotopes and adducts, also without a previous knowledge of monoisotopic masses the search in libraries of significant features may lead to miss annotations (61,69). Additionally, searches for specific masses, considering expected adducts, can lead to a large number of potential molecular formulas and therefore possible molecular entities.

Regarding to metabolomic research and just in order to summarize the project justification it is clear that non-targeted studies aim to analyse tens to thousands of metabolites in a single sample without prior knowledge of their chemical identity. When conducting targeted studies with pre-existing knowledge of the chemical identity, there is no issue of bottleneck. However, in non-targeted studies, it is crucial to perform rigorous annotation and identification of metabolites to ensure their maximum interpretation and impact (43). Furthermore, this project will only be focused in LC/MS as it allows the detection of a broader range of metabolites than technologies as gas chromatography-MS and capillary electrophoresis-MS (70).

In metabolomic studies, the identification of putative metabolites is a significant challenge due to shortcomings in data obtained during or after the study (44). This presents a bottleneck in analytical chemistry, as metabolomic workflows output complex molecular signatures, and there is no a universal method for metabolome analysis, as the chemical properties of metabolites differ significantly and their concentrations vary widely (35). Additionally, the quantification of many metabolites contained in complex samples is complicated by ion suppression, fragmentation, and the presence of isomers that can result in peak misidentification (34). Moreover, untargeted metabolomics workflows suffer from widely recognized hurdle across various stages, including insufficient standardization in data production, limited number of identified metabolites, lack of automatic feature detection in data processing, which impede the inter-operability and reusability of metabolomics data (30,35).

What it is reachable in the field of bioinformatics and specifically in metabolomics is the development of improved annotation and metabolite identification software tools to minimize.

So as to deal with the above commented shortcomings and as a result, improve the feature annotation process and avoid peak misidentification. Particularly, in simple mass spectrometry which is the process where the project is focused and the first step to proceed with a further computational or statistical analysis.

1.4. Computational metabolomics annotation tools utilized

In order to get data processed and achieve significant metabolomics information R-packages have to be used. The most notable software programs on which the project is based are described below.

1.4.1. XCMS:

Once raw data from the samples is obtained, it is necessary to **process** it previous to perform annotation. This step is frequently achieved through software packages. **XCMS** is a powerful R-based package, freely available, that has gained popularity in untargeted metabolomic studies, it uses nonlinear retention time correction, matched filtration, peak detection, and peak matching to extract relevant information from raw LC/MS data (49,71,72).

The most significant steps in raw data processing with XCMS are the following ones (see the flowchart in Figure 5):

1. Peak detection where the software identifies peaks corresponding to metabolites and noise artifacts.
2. The peak matching algorithm bins peaks by mass and identifies groups of peaks with similar retention times. It eliminates insignificant groups and resolves cases where a sample has multiple peaks in a group.
3. Retention time correction is crucial for accurate data processing. Well-behaved peak groups are identified, and the algorithm calculates the median retention time and deviation for each sample in the group.
4. Once peak groups are established, XCMS identifies absent samples in each group. By utilizing statistical data from peak detection, it aligns retention times for all samples. The raw LC/MS data is integrated in the m/z-rt area of a feature to fill in intensity values for samples with missing chromatographic peaks.

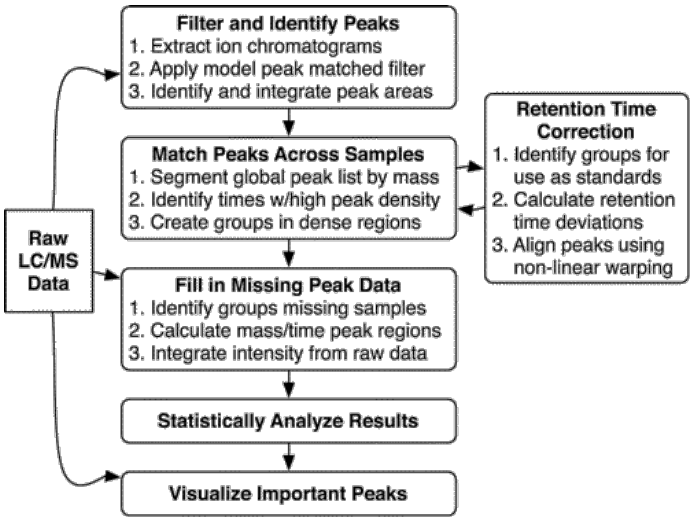


Figure 5 *- Flowchart of the general strategy for pre-processing LC/MS data in untargeted metabolomic analysis. Peaks are filtered and identified, they are match across the samples, a retention time correction is applied and missing peaks are filled to then analyze statistically and visualize the obtained results. (71)*

1.4.2. CAMERA:

The R-package CAMERA consist of a **C**ollection of **A**lgorithms for **ME**tabolite p**R**ofile **A**nnotation. It is specifically designed to post-process and directly interact with processed peak data (feature list) from the XCMS R-package. The package's functionality includes collecting all features related to a metabolite into a compound spectrum (71). The primary objective of the software is to annotate isotope peaks, adducts, and fragments in peak lists, as well as to evaluate LC/MS data. To accomplish this annotation process, a collection of algorithms has been implemented, including the fast retention time-based grouping and a graph-based algorithm. The purpose of these methods is to cluster mass signals originated from a single metabolite by integrating peak shape analysis, isotopic information, and intensity correlation across samples (73). Annotation results are used to calculate molecular composition if the mass spectrometer has accurate mass and isotope pattern intensities. Additionally, automatic sample selection avoids unsatisfactory results for low-intensity compounds or absent samples. Also, ion species annotation combines spectral information from positive and negative ion modes for more accurate results (14,74). As it is affirmed along the project, during the ionization process many different kind of ions such as adducts or in-source fragments as well as protonated molecular ions are produced for a single molecule. Moreover, if a molecule has an intrinsic charge it might also be observed (75). Overall, CAMERA is essential to conduct a discovery process, separate different substances and identify their ion species (74).

The annotation process performed in CAMERA are represented by the next stages (Figure 6):

1. Feature grouping steps integrate Retention Time.
2. Features are identified as isotopic peaks.
3. Feature grouping steps also integrate Chromatographic Peak Shape.
4. Adducts are annotated by using a dynamic rule table.
5. The annotation process can be verified with LC/MS data acquired in opposite ion mode.

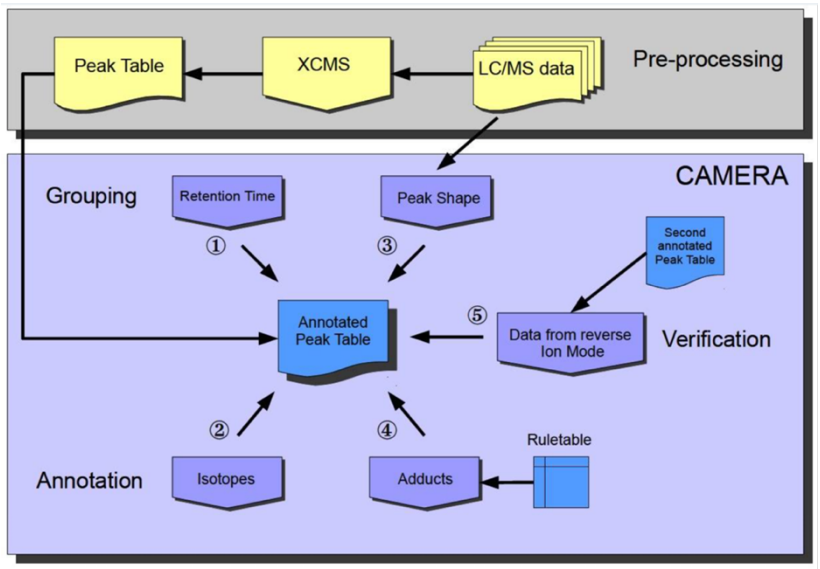


Figure 6 *- CAMERA workflow for LC/MS data analysis. The above stages from 1 to 5 are represented.(14)*

1.4.3. CliqueMS:

CliqueMS is a network-based algorithm that annotates isotopes, adducts and in-source fragments from LC/MS data. This software can perform the reduction of multiple features to single metabolites, a crucial step for getting a correct annotation from LC/MS experiments. It groups the signals that are likely to be generated from the same metabolite. CliqueMS transforms the spectral data into a network, within this network it finds groups according to a probabilistic algorithm. Each group is a clique, and represents all the signals such as isotopes, adducts and fragments belonging to the same metabolite. The algorithm provides a flexible annotation in which for each group it provides up to five different annotations from all possible annotations within that group. Those annotations are ranked through a score, which is computed by an adduct list that contains the intensity of the putative adducts. It can use a build-in adduct list or a customized list. CliqueMS is able to annotate samples one by one and reduces the complexity of the data and even in large samples. It is available as a web application or as an R Package.(76)

The annotation process can be summarized in three steps (a, b and c in Figure 7):

1. Divide features into clique groups.
2. Annotate isotopes.
3. Annotate adducts and fragments by isotope and parental mass identification.

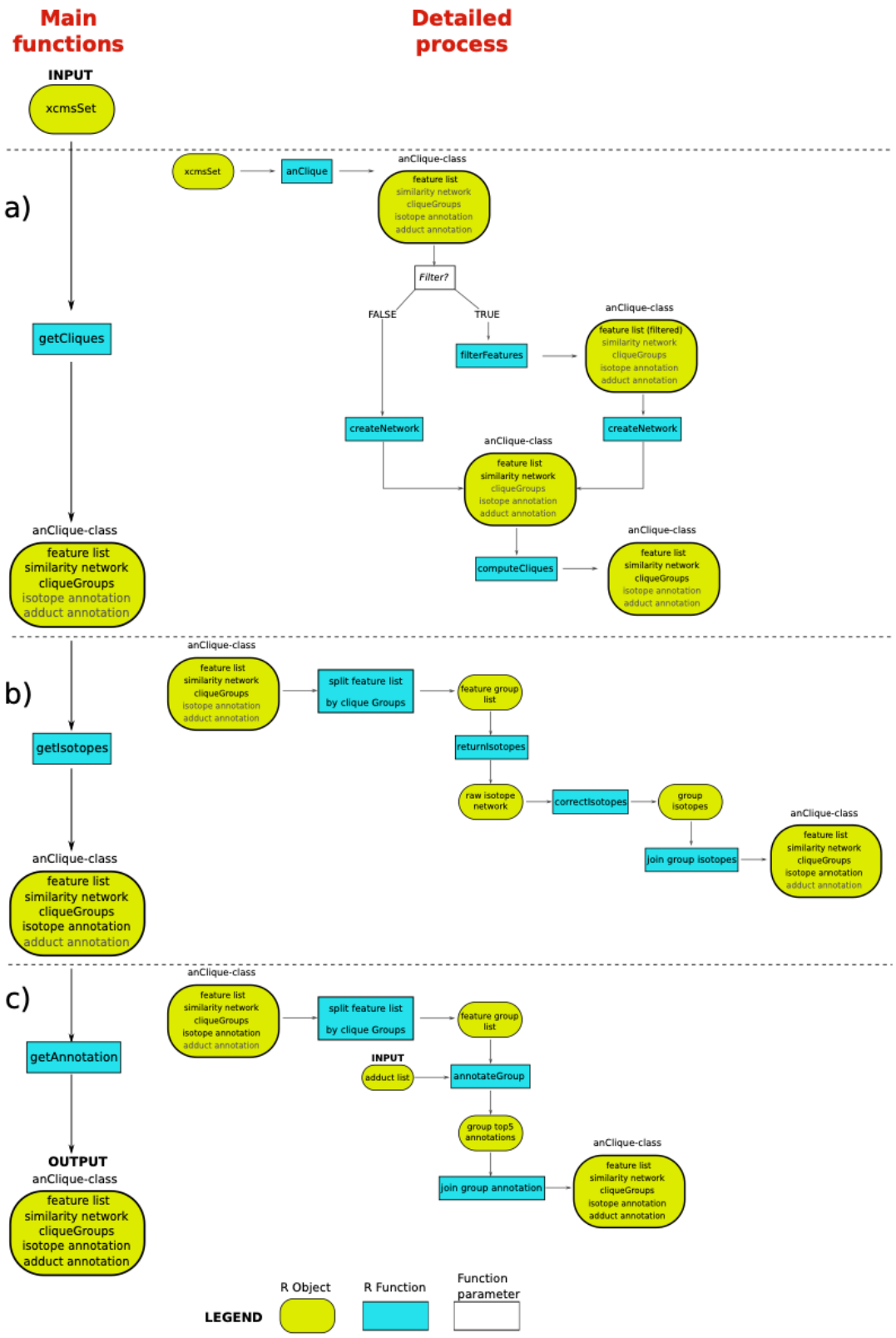


Figure 7 *– Shows the three stages in CliqueMS R-package annotation workflow (a, b and c). (76)*

2. HYPOTHESIS AND OBJECTIVES

It has been observed that peak annotation and though metabolite identification continues to be a big struggle in metabolomics research, particularly in untargeted metabolomics assays. If the annotation process is not achieved correctly, it can significantly hinder the future interpretation and identification of metabolites, as everything that is annotated (adducts, isotopes, and fragments in source) is used to identify the molecules. There are several known tools that complement the annotation process, but the majority of them do not carry out annotation as we have described previously. Consequently, we selected two software programs that perform this process as we comprehend it, in order to understand this bottleneck step is addressed in untargeted metabolomics assays.

We hypothesize that both software will provide significant different outputs. While CAMERA process all the samples in a single analysis providing a simpler output it will be suitable for getting a general overview of the most abundant adducts and further metabolites, CliqueMS will be appropriate in cases where samples need to be analyzed one by one providing a complete description for each one.

The aim of this project is to compare the annotation process and output between CAMERA and CliqueMS, two different R-software, in a metabolomics LC-MS-based experiment.

This project will facilitate the comparison of CAMERA and CliqueMS, two common tools used so far for peak annotation to identify the particularities of each process as well as their limitations and advantages. To direct the path towards potential future advancements in untargeted metabolomics bioinformatic annotation tools taking in account both packages characteristics.

In the way to complete it, the following subobjectives will be also achieved:

- Get to know better the shortcomings in the final step for a typical qualitative metabolomic study MS-based.
- To perform the computational procedure and data analysis from raw data obtained in a LC/MS assay from biological samples.
- To get familiarized with existing software tools for executing feature annotation performance while understanding how the annotation process is developed.
- Determine which software is better to choose according to their employment, the adduct annotation result and its reliability or validation so as to decide under which circumstances is better to choose one or another package for data processing.

3. MATERIALS AND METHODS

This project has been developed through the comparison of two different software which are implemented in RStudio, which is an integrated development environment for the programming language R that provides an interface for writing and executing R code, it is possible to visualize data and manage packages as well. The software, also called R-packages in which the project is focused are CAMERA and CliqueMS. However, the XCMS package is also essential to be used. The workflow (Figure 8) to achieve the comparison has been focused on determining which of the employed software can provide a better feature annotation, referring to the identification of what each peak stands for and the further matching to a certain metabolite to verify the previous process.

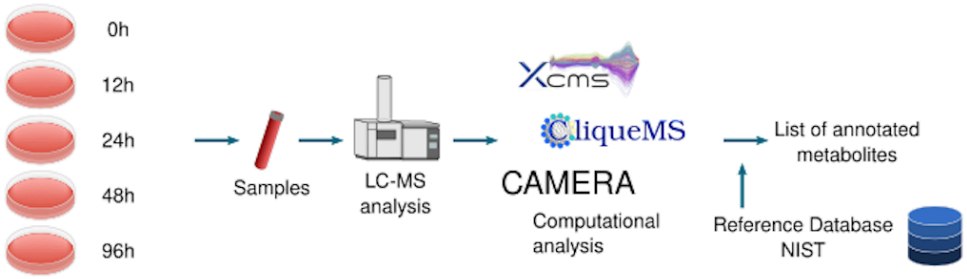


Figure 8 *- The workflow to complete the objectives starts with the obtention of a metabolomics dataset from experimental samples, this process is followed by the data processing with the R-packages and finally the list of annotated results is explained and validated through peak correlation and the alignment with reference mass spectrums.*

3.1. Experimental Data

In this project the dataset that has been used to carry out CAMERA and CliqueMS algorithms comes from 25 samples and 4 quality controls from RAW264.7 murine macrophages (Figure 9) subjected to different serum reduction conditions in the culture medium.

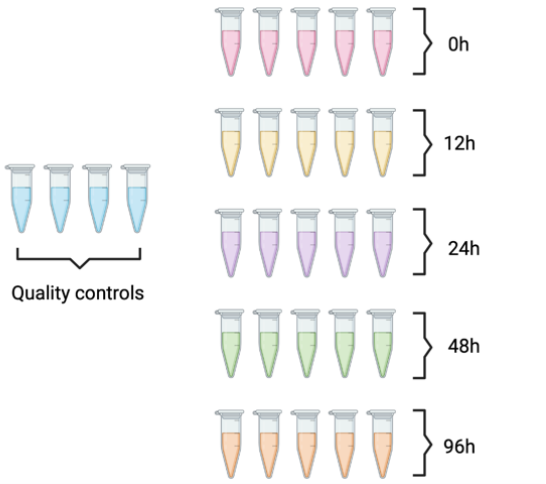


Figure 9 *- Sample distribution. Four quality controls which contained a pool of different samples and five samples in each experimental time group (0h, 12h, 24h, 48h and 96h).*

The cells were obtained from the American Type Culture Collection (Rockville, MD) and cultured in Dulbecco's modified Eagle's medium, which contained 10% (v/v) fetal bovine serum, 100 U/ml penicillin, 100 µg/ml streptomycin, and 2 mM 1-glutamine. They were maintained at 37 °C in a humidified atmosphere of C02/air (1:19). Subsequently, the cells were washed with PBS, harvested in ice-cold water, and stored at −80 °C until metabolite extraction and LC-MS/MS analysis.

Cell extracts were analyzed using ultra-high performance liquid chromatography (UPLC) (Bruker Elute UHPLC, Bruker Corp., Billerica, MA) with a hydrophilic interaction liquid chromatography (HILIC) ACQUITY BEAH column (2.1 x 100 mm, 1.7 µm, Waters Corp., Milford, MA). The gradient consisted of 99% B for 1 minute, 65% B for 13 minutes, 40% B for 3 minutes, and held at 40% B for 1 minute, with a flow rate of 150 µL/min. The mobile phase compositions A and B were composed of water + 0.1% formic acid and acetonitrile + 0.1% formic acid, respectively. The UPLC was coupled to a quadrupole time-of-flight (qToF) mass spectrometer (Bruker Corp.) of the Bruker Impact II system. The mass spectrometer was set to acquire ions within the m/z range of 50-1000 with an acquisition rate of 4 spectra/second.

Data was acquired in positive ionization mode as the differences in number of metabolites and features annotated are specially remarkable for the positive ionization mode spectrum, in which the number of adducts is larger mainly due to the influence of mobile phase additives and organic solvents, and therefore more features can coelute. In the negative ionization mode, the number adducts is much smaller and therefore the differences between algorithms are not as rigid. (76)

3.2. Software selection

Before deciding which R software could be compared in this project a browsing task was done consisting on reading the documentation and trying as well as understanding their scripts. Some of the candidate software or packages were RAMClustR, CAMERA, MetaboAnnotation, CliqueMS, IDL.CSA, Mz.unity, xMSannotator, xMSanalyzer, CPVA, McSearch, IP4M and MassFlowR.

The selection criteria we created in order to choose which packages to compare was based on the following conditions: the software should be implemented in R-Studio, it should be able to be use in LC/MS, should also have mistakes updated and actually being utilized and according to their Reference Manuals, guarantee the completion of feature annotation as the concept of accomplishing the matching of a feature to an adduct, isotope or in source fragment. Regarding to these conditions CAMERA and CliqueMS where the ones which fulfilled all the requirements as it can be seen in Table 1.

***Table 1**. **Selection criteria of evaluated software**. The symbol (✓) represents the completion of the requirement and on the contrary, the (X) symbol indicates the lack of accomplishment of the studied characteristic.*

Packages	R-Studio	LC/MS	Updated	Annotation
RAMClustR (77)	✓	✓	✓	X
CAMERA (14)	✓	✓	✓	✓
MetaboAnnotation (24)	✓	✓	X	X
CliqueMS (76)	✓	✓	✓	✓
IDL.CSA (78)	✓	✓	X	✓
Mz.unity (79)	✓	✓	X	X
xMSannotator (80)	✓	✓	✓	X
xMSanalyzer (81)	✓	✓	✓	X
CPVA (82,83)	X	✓	✓	✓
McSearch (83,84)	✓	X	✓	✓
IP4M (83,85)	X	✓	✓	✓
MassFlowR (86)	✓	✓	✓	X

3.3. Result obtention through script development in R-Studio

In order to accomplish the stated objectives, a series of procedures were implemented and are described in general terms below. Detailed steps for each section of this workflow, along with corresponding scripts and explanations of the functions or parameters utilized, are extensively described in the S1 and S2 Supporting Material of this project.

For conducting the bioinformatics analysis, RStudio version 2022.07.2 was utilized. The data processing was primarily executed on a conventional MacBook Air 2019 laptop equipped with 16 GB of RAM, except for the processing of individual CliqueMS samples, which required processing them through a web server, with Ubuntu 16.04.7 LTS (95 GB of RAM, 24 CPU).

Initially, the **data processing** was executed using CAMERA and CliqueMS. The workflow followed in each R package which is comparable is illustrated in Figure 10. Regarding to data pre-processing both packages employ the XCMS R package which was used to identify chromatographic features, utilizing the optimized parameters for data obtained through UPLC/QuadrupoleTOF (method = centWave, ppm = 10, mzdiff = 0.01, peakwidth = c(2, 20), and noise=100). The retention times of the samples were aligned using the Obiwarp method, and the peaks were clustered based on retention times and m/z values. The peaks were filled to integrate signals in the m/z-RT area of a feature. Subsequently, feature grouping was performed to associate groups with putative metabolites.

CAMERA performs chromatographic deconvolution through retention time-based feature grouping using the default parameters of the *groupFWHM* function, followed by a group verification process through the *groupCorr* function, which computes the Pearson correlation. The final outcome is the grouping of features into *pcgroups*, ideally related to one metabolite.

CliqueMS generates groups based on the *getCliques* function, utilizing cosine metric to establish a similarity network where features represent nodes and edges correspond to the cosine similarity between these features. The final output in this case is the *cliqueGroups* formation. In the end of this first step, the molecules resulting from sample ionization were annotated, mainly comprising of adducts, isotopes, and in-source fragments.

Initially, features satisfying the adduct conditions were identified, and then isotope annotation was generated. (Figure 10) A default CliqueMS adduct table from positive ionization mode was used in both R-packages to perform the annotation of adduct peaks and in-source fragments/neutral losses. On the one hand, CAMERA employs as the main functions *findIsotopes* and *findAdducts* and finally provides a peaklist (*getPeaklist)* containing all the annotation information. On the other hand, CliqueMS first annotates isotopes with *getIsotopes* function, afterwards the function *getAnnotation* produced the whole adduct annotation. In the end a list of cliques is provided in a peaklist (*getPeaklistanClique*) where all the obtained information is summarized. The output from CAMERA provides the best annotation associated with a particular feature, while CliqueMS provides the top 5 annotations with the best score.

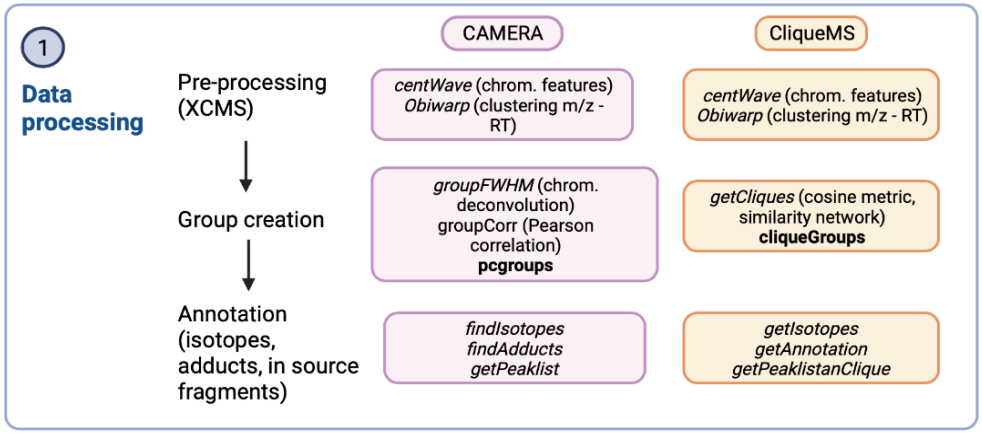


Figure 10 - *Applied workflow showing the functions for each R-package. It shows the 1st step which is based on data processing. The first column shows the general process, the second one the CAMERA functions and the last one the corresponding CliqueMS functions.*

The procedure was followed by **data filtering** (Figure 11) which was carried out to extract pertinent information from the extensive data files generated earlier. A script (refer to S3 Supporting Material) was created using *for* loops, which first added the proton mass (1,007242) to the mass of reference metabolites obtained with positive ionization mode. An error in ppm was then imposed to relate the reference mass to that of the features obtained with the software, to obtain the association of these identified features with the 44 reference metabolites. As a consequence of this process, a unique list is obtained in CAMERA, and 29 distinct lists are obtained from CliqueMS. Subsequently, to specifically obtain the annotation of adducts associated with each metabolite, the *cliqueGroup* or *pcgroup* was assumed to be associated with one putative metabolite and was determined by checking if the mass of the identified adduct exactly matched the neutral mass of the metabolite itself. As a result, the

obtained *pcgroups* and *cliqueGroups* contained at least a feature with the protonated mass of a metabolite from the reference list. Different tables were synthesized, indicating the metabolites with the adducts belonging to the same group associated with them according to each software (see Results).

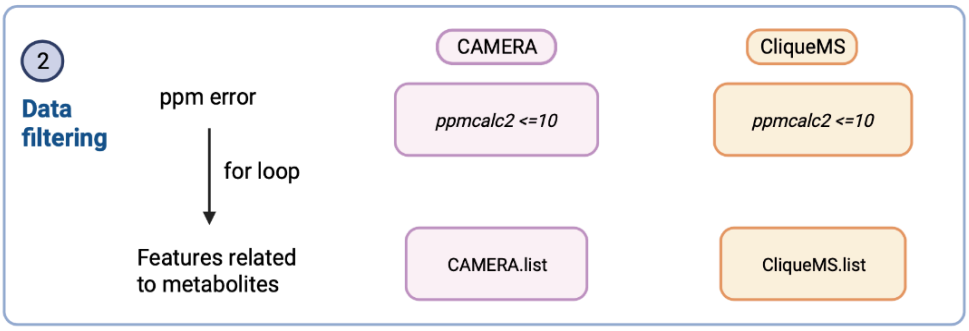


Figure 11 *- Applied workflow showing the functions for each R-package. It shows the 2nd step which is based on data filtering. The first column shows the general process, the second one the CAMERA functions and the last one the corresponding CliqueMS functions.*

Furthermore, a graphical representation of the filtered results was obtained by creating mass **pseudospectrum** (Figure 12) for the formed groups of certain metabolites to visually compare the software. In addition, pseudospectra were generated using a function in RStudio to show the relationship between the m/z values and the relative intensity of the identified peaks. Each representation aimed to depict the peaks from a particular *pcgroup* or *cliqueGroup*, which were ideally associated with a specific metabolite. The most representative samples, based on their score, are displayed in the Results section in this project.

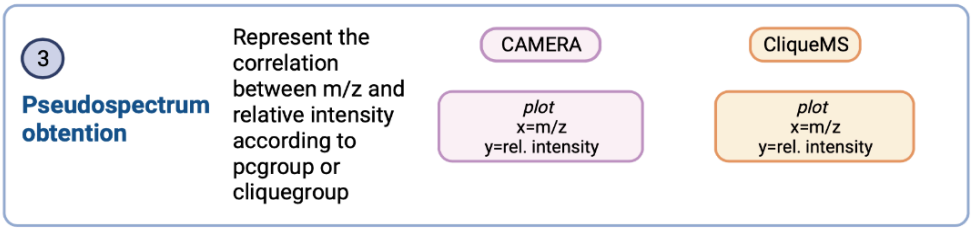


Figure 12 *- Applied workflow showing the functions for each R-package. It shows the 3rd step which is based on pseudospectrum obtention. The first column shows the general process, the second one the CAMERA functions and the last one the corresponding CliqueMS functions.*

The **correlation** of variables between the intensity values of the peaks from different CAMERA samples was then checked (Figure 13). The 7 previously selected metabolites were chosen to compare the annotation of CAMERA and CliqueMS, and then the intensities of the peaks from different samples associated with a corresponding *pcgroup* for a metabolite were examined for correlation. To do this, the *cor* function was used, which correlates the columns (intensities) and rows (features of a specific *pcgroup*) of a provided matrix (result of the CAMERA processing). After trying different parameters for this function, the *use= pairwise.complete.obs* was chosen as it ignores NAs (not available intensity values) and as a result. For each pair of variables being

correlated, any pair of observations with missing values for either feature was excluded from the computation. Then, the *ggplot2* and *reshape2* libraries were used to create a default *ggplot2* from the provided data, which was then drawn and finally represented using a heatmap (see the complete script in S5 Supporting Material).

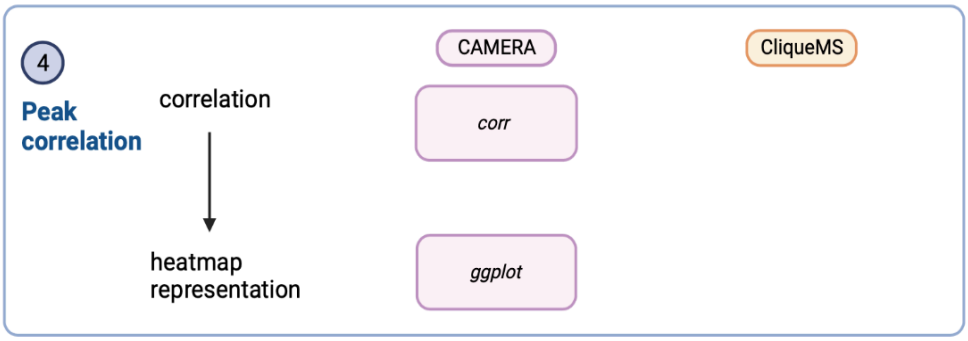


Figure 13 *- Applied workflow showing the functions for each R-package. It shows the 4th step which is based on peak correlation. The first column shows the general process, the second one the CAMERA functions, for CliqueMS it was not possible to perform this step.*

After that, we carried out the **validation** of the annotation by comparing pseudospectras of the putative metabolites obtained experimentally and processed by CAMERA and CliqueMS with **reference** spectra of metabolites deposited in the NIST database (Figure 14). This comparison is important for metabolite identification and helps to increase confidence in the accuracy of the process. The 7 pseudospectra selected from CAMERA and CliqueMS, respectively, where compared to metabolites present in the NIST data base. In the first place, a reduced reference library containing information about Glutamine, Creatine, Arginine, Histidine, Lysine, Phenylalanine and Carnitine was created with the information provided by NIST. The function to generate the plot isprovided in Supporting Material (see the complete script in S6 part in the Supporting Material). Finally, for each putative metabolite, a reference spectrum was selected based on its NIST ID number in order to obtain pseudospectra that were symmetrically represented according to their *pcgroup* or *cliqueGroup* against the reference spectrum. The purpose of this was to show the patterns of peak intensity and m/z values.

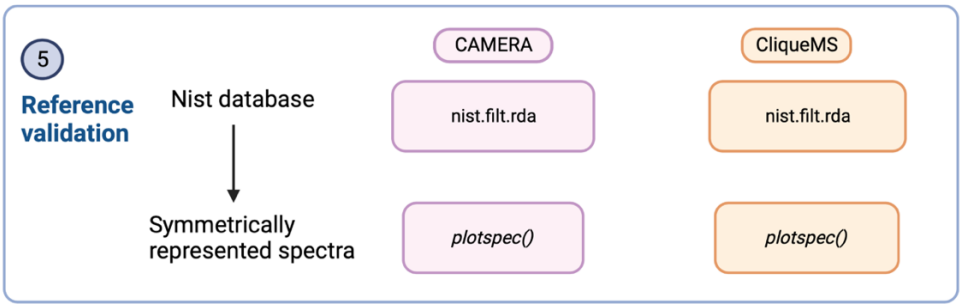


Figure 14 *- Applied workflow showing the functions for each R-package. It shows the 5th step which is based on reference validation. The first column shows the general process, the second one the CAMERA functions and the last one the corresponding CliqueMS functions.*

3. RESULTS

We focused the present study on the evaluation of two bioinformatic tools (CAMERA and CliqueMS) for annotating metabolites, with a specific emphasis on comparing the output generated by each tool. We based this comparison on several factors, including the amount of information provided by each software, their reliability, the complexity of the procedure and data treatment.

3.1. Analytical results

In order to compare both R-packages, we selected the *pcgroups* or *cliqueGroups* which contained a feature with the protonated mass from the 44 metabolites that have been previously reported to be present in macrophage cells. We selected the groups of features with at least one feature with an error less than 10 ppm (parts per million) with respect to the protonated mass of the putative metabolites. According to this, we determined the number of created groups of features in each software which were matched to the studying metabolites.

Table 2 depicts the adduct and isotope CAMERA annotation for each of the 44 metabolites under study, we can see that 36 metabolites achieved a complete annotation. We sourced the information from the CAMERA result data frame, to obtain the present information in Table 2, where figures the metabolite name, the column mass refers to the molecular mass value of each metabolite. Ions are protonated molecules, which can take the form of isotopes or adducts, are listed using letters in alphabetical order. We also indicated their corresponding m/z and retention time (RT) values.

***Table 2**. Adduct annotation for all studying metabolites in CAMERA.*

Metabolite name	Mass	ions	m/z	RT	Isotopes	Adducts
D-Alanine	89,048					
L-Glutamine	146,069	A	147,077	539,292		[M+H]+ 146,069
		B	293,146	537,766		[2M+H]+ 146,069
Taurine	125,015					
Thiamine	265,112					
Pyridoxine (Vitamin B6)	205,64					
D-Pipecolic acid	129,079	A	130,086	655,324	[20][M]+	
Creatine	131,07	A	170,033	447,469		[M+K]+ 131,069
		B	176,040	446,462		[M-H+2Na]+ 131,069
L-Arginine	174,112	A	175,119	639,157	[47][M]+	[M+H]+ 174,112
		B	157,108	638,663		[M+H-H2O]+ 174,112
		C	158,092	639,165	[37][M]+	[M+H-NH3]+ 174,112
		D	197,101	639,151		[M+Na]+ 174,112
		E	349,230	639,153		[2M+H]+ 174,112
		F	523,342	638,667		[3M+H]+ 174,112
L-Asparagine	132,053493	A	116,0345	558,466		[M+H-NH3]+ 132,053
		B	155,043	558,963		[M+Na]+ 132,053
		C	265,113	558,964		[2M+H]+ 132.053
Citrulline	175,096	A	176,102	472,686	[46][M+1]+	
L-Glutamate	169,035104	A	207,9984	498,9348		[M+K]+ 169.035
		B	361,059	498,934		[2M+Na]+ 169.035
L-Histidine	155,069	A	156,077	649,758	[36][M]+	
		B	157,080	649,758	[36][M+1]+	
		C	158,082	649,261	[36][M+2]+	
L-Leucine	131,094629	A	154,0839	418,4468	[34][M]+	[M+Na]+ 131.095
		B	155,087	418,700	[34][M+1]+	
		C	170,058	418,697		[M+K]+ 131.095
		D	285,179	418,697	[106][M]+	[2M+Na]+ 131.095
L-Lysine	146,106	A	129,102	655,318		[M+H-H2O]+ 146,105

		B	130,086	655,324	[20][M]+	[M+H-NH3]+ 146,105
		C	147,113	655,322	[29][M]+	[M+H]+ 146,105
		D	293,218	654,829	[112][M]+	[2M+H]+ 146.105
L-Methionine	149,051051					
L-Phenylalanine	165,079	A	148,076	417,197		[M+H-H2O]+ 165,079
		B	149,060	416,186	[32][M]+	[M+H-NH3]+ 165,079
		C	166,086	416,682	[42][M]+	[M+H]+ 165,079
		D	188,070	418,208	[51][M]+	[M+Na]+ 165.079
		E	204,043	418,202		[M+K]+ 165,079
		F	210,050	416,681	[62][M]+	[M-H+2Na]+ 165,079
		G	353,147	416,187	[159][M]+	[2M+Na]+ 165,079
L-Tryptophan	204,089878					
Glutathione, oxidized	612,152	A	298,571	801,671		[M+2H-NH3]2+ 612,15
		B	613,159	801,673	[343][M]+	[M+H]+ 612,15
		C	651,106	801,159		[M+K]+ 612,15
L-Carnitine	161,105193	A	162,113	337,728	[40][M]+	[M+H]+ 161,105
		B	163,116	337,480	[40][M+1]+	
		C	200,068	337,465	[55][M]+	[M+K]+ 161,105
Hypoxanthine	136,038511					
Inosine	268,080771	A	269,0879	435,617		[M+H]+ 268,079
		B	268,081	435,859		[Cat]+ 268,079
		C	313,047	435,877		[M-H+2Na]+ 268,079
Adenosine	267,096755			325,368		
Guanidylic acid (guanosine monophosphate)	363,058003			816,794		
Nicotinamide adenine dinucleotide (NAD)	663,109	A	664,116	886,946		[M+H]+ 663,109
		B	332,562	886,952	[137][M]2+	[M+2H]2+ 663.109
Pantothenic Acid	219,11	A	220,118	133,032		[M+H]+ 219,11
		B	242,100	135,708		[M+Na]+ 219.11
Glycerophosphocholine	257,103	A	240,099	644,225		[M+H-H2O]+ 257.103
		B	258,110	645,211	[89][M]+	[M+H]+ 257,103
		C	280,092	645,232	[100][M]+	[M+Na]+ 257.103
		D	296,066	645,236	[115][M]+	[M+K]+ 257.103
		E	515,213	644,738	[265][M]+	[2M+H]+ 257.103
		F	537,195	645,214	[285][M]+	[2M+Na]+ 257.103
		G	553,169	644,731		[2M+K]+ 257.103
		H	772,315	645,210	[527][M]+	[3M+H]+ 257.103
Phosphocholine	184,073872					
N-Acetyl-D-glucosamine	221,090	A	222,097	616,476	[70][M]+	
		B	223,100	616,969	[70][M+1]+	
Cytidine	243,085522					
5'-CMP	323,052	A	324,059	749,172	[128][M]+	
		B	325,062	749,183	[128][M+1]+	
		C	326,063	748,159	[128][M+2]+	
Uridine monophosphate (UMP)	324,035871		325,044			
Deoxyguanosine diphosphate (dGDP)	427,029					
Adenosine monophosphate	347,063088	A	348,070	749,677	[152][M]+	
		B	350,075	748,680	[152][M+2]+	
ADP	427,029					
L-Cystine	240,023852					
L-Isoleucine	130,095					
L-Methionine S-oxide	165,045966	A	148,043	569,566		[M+H-H2O]+ 165,046
		B	149,027	569,551	[31][M]+	[M+H-NH3]+ 165.046
		C	166,053	569,554	[41][M]+	[M+H]+ 165.046
		D	188,035	569,578		[M+Na]+ 165,046
		E	331,099	569,557		[2M+H]+ 165,046
(±)-Propionylcarnitine	217,131			99,882		
Acetylcarnitine	203,115759		204,123	101,356		
D-Mannitol	178,048					
m-Coumaric acid	164,047345	A	165,055	433,344		[M+H]+ 164,044
		B	148,043	433,836		[M+H-OH]+ 164,044
		C	182,081	433,344		[M+NH4]+ 164,044
Coumarin	146,037		147,044			
3-Amino-3-(4-hydroxyphenyl)propanoate	181,073894	A	182,081	433,344		[M+H]+ 181.073
		B	137,078	433,589		[M-CO2H+H]+ 181,073
		C	165,055	433,344		[M+H-NH3]+ 181,073
		D	226,045	434,590		[M-H+2Na]+ 181,073
N2-Acetyl-L-ornithine	174,100	A	175,108	483,786		[M+H]+ 174,101
			197,090	484,794		[M+Na]+ 174.101

In Figure 15, we represented the overall number of metabolites with at least one adduct annotation. The graph presents 36 metabolites with annotation for CAMERA and an average of 39 for CliqueMS. For CAMERA, we took all samples into account, while for CliqueMS, we represented the average of 25 processed samples and their corresponding standard deviation

in order to handle the data in a general manner. Figure 15 shows a visual result of the list we obtained after applying the loop which considered the ppm error.

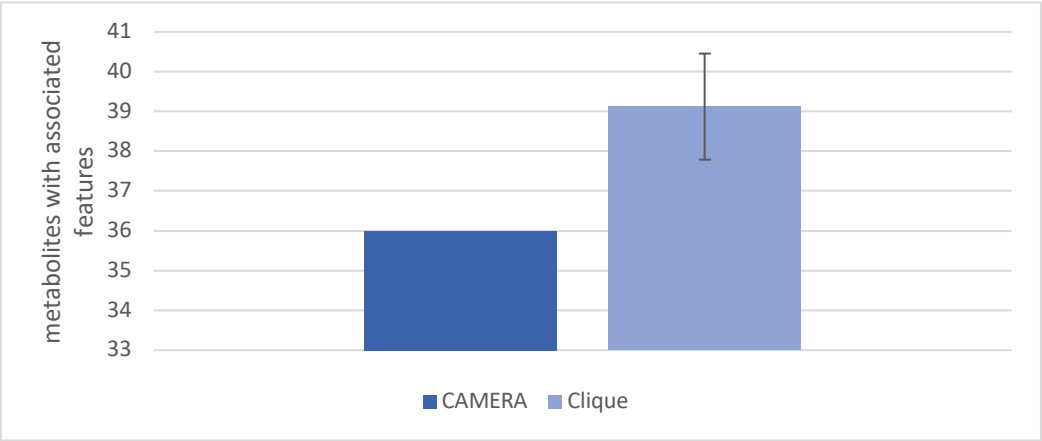


Figure 15 *- Number of metabolites with associated features. For CAMERA (dark blue) total number of features associated to a putative metabolite across all samples are represented. For CliqueMS (light blue) the mean of features across all samples and the standard deviation are represented.*

Furthermore, in addition to the observation that the number of metabolites with associated features was slightly higher in CliqueMS (three more metabolites presented annotation) throughout the lists, the number of features contained on CliqueMS lists for each metabolite was significantly greater compared to the peaks which CAMERA had obtained (see Figure S3 in Supporting Material).

We evaluated the adduct and neutral losses annotation in 10 aleatory metabolites: Alanine, Glutamine, Pyridoxine or vitamin B6, creatine, Arginine, Histidine, Lysine, Phenylalanine, Tryptophan and carnitine. This number of samples was enough for us to see the differences between both software. From these selected metabolites, 3 were excluded (Arginine, vitamin B6 and Tryptophan) as CAMERA did not produce any annotation for them and neither CliqueMS did for vitamin B6.

With the following table (Table 3) it can be seen the resultant annotation for the 7 selected metabolites. Unless CliqueMS needed the samples to be processed one by one, here we presented all the samples together as a group or replicates, like CAMERA did, in order to present an overview of the whole result.

Table 3. Annotation comparison between CAMERA and CliqueMS processing.

	CliqueMS		CAMERA	
Metabolite	**Annotation**	**Isotopes**	**Annotation**	**Isotopes**
L-Glutamine	[M-H+2Na]+			
	[3M+H]+			
	[M+Na]+	3		
	[M+Na-H2O]+			
	[2M+H]+	3	[2M+H]+	
	[M+H]+	3	[M+H]+	
	[M+H-H2O]+			
	[M+H-NH3]+	2		
	[M+K]+			
Creatine	[M+K]+		[M+K]+	
	[M+K-H2O]+			
	[3M+H]+			
	[2M+H]+	2		
	[M+Na-H2O]+	2		
	[M+H-NH3]+	0		
	[M+H]+	3		
	[M+H-H2O]+	2		
	[M-H+2Na]+	3	[M-H+2Na]+	
	[2M+Na]+	3		
	[M+Na]+	3		
	[Cat]+			
L-Arginine	[M-H+2Na]+			
	[M+Na-H2O]+			
	[3M+H]+		[3M+H]+	1
	[2M+Na]+		[2M+Na]+	
	[M+K]+			1
	[M+H-NH3]+	2	[M+H-NH3]+	
	[2M+H]+	2	[2M+H]+	1
	[M+H-H2O]+		[M+H-H2O]+	
	[M+Na]+		[M+Na]+	
	[M+H]+	2	[M+H]+	
L-Histidine	[M+H-H2O]+	0	[36][M]+	3
	[M+H]+	4	[36][M+1]+	
	[M-H+2Na]+		[36][M+2]+	
	[2M+Na]+			
	[M+H-NH3]+			
	[2M+H]+	2		
	[M+Na]+	2		
	[M+K]+			
L-Lysine	[2M+Na]+			
	[M-H+2Na]+			1
	[M+Na]+			1
	[2M+H]+	2	[2M+H]+	1
	[M+H]+	3	[M+H]+	0
	[M+H-NH3]+	2	[M+H-NH3]+	
	[M+H-H2O]+		[M+H-H2O]+	
L-Phenylalanine	[M-H+2Na]+	3	[M-H+2Na]+	1
	[2M+Na]+	2	[2M+Na]+	1
	[M+H-H2O]+		[M+H-H2O]+	
	[M+K]+		[M+K]+	
	[M+H]+	3	[M+H]+	1
	[M+H-NH3]+	2	[M+H-NH3]+	1
	[2M+H]+	0		1
	[M+Na-H2O]+			
	[M+Na]+	2	[M+Na]+	
L-Carnitine	[M+H]+	2	[M+H]+	1
	[M+K]+		[M+K]+	1
	[Cat-H]+			

Table 3 represents, for each metabolite, an associated list containing all the adducts and neutral losses which CliqueMS (in blue) and CAMERA (in orange) had identified and annotated. Next to the adduct column, we presented the number of isotopes for each adduct. Additionally, while CAMERA exclusively displayed the best annotation found, CliqueMS presented every one of the five annotation results, some of them with low annotation scores. Overall, despite the number of adducts and neutral losses annotated by CliqueMS is considerably higher than in CAMERA, it can be perceived that all the molecules identified in CAMERA are also present in CliqueMS. It is

important to note that all the adducts linked to the metabolites belong to the same *pcgroup* (CAMERA) or *cliqueGroup* (CliqueMS) and we only selected the groups which contained at least a feature with the protonated mass of the metabolite as we mentioned before in data filtering in the Materials and Methods part.

Table 4. Sample percentage of adducts for each experimental time groups.

		CliqueMS					
Metabolite	**Annotation**	**QC (1-4)**	**0h (5-9)**	**12h (10-14)**	**24h (15-19)**	**48h (20-24)**	**96h (25-29)**
L-Glutamine	[M-H+2Na]+	0%	0%	0%	0%	25%	0%
	[3M+H]+	0%	0%	20%	25%	0%	0%
	[M+Na]+	25%	0%	0%	25%	75%	100%
	[M+Na-H2O]+	50%	0%	0%	0%	100%	100%
	[2M+H]+	75%	20%	20%	25%	50%	67%
	[M+H]+	75%	100%	60%	100%	100%	100%
	[M+H-H2O]+	75%	40%	0%	50%	0%	0%
	[M+H-NH3]+	75%	100%	60%	100%	75%	100%
	[M+K]+	0%	0%	0%	0%	50%	0%
Creatine	[M+K]+	0%	0%	0%	0%	25%	0%
	[M+K-H2O]+	0%	0%	20%	0%	0%	0%
	[3M+H]+	0%	0%	20%	50%	25%	0%
	[2M+H]+	0%	0%	40%	75%	50%	66,67%
	[M+Na-H2O]+	25%	0%	100%	50%	50%	100%
	[M+H-NH3]+	50%	60%	60%	75%	50%	66,67%
	[M+H]+	75%	100%	80%	100%	100%	33,33%
	[M+H-H2O]+	50%	40%	40%	75%	75%	100%
	[M-H+2Na]+	25%	40%	80%	50%	50%	100%
	[2M+Na]+	25%	0%	60%	50%	75%	66,67%
	[M+Na]+	25%	0%	80%	25%	100%	66,67%
	[Cat]+	0%	0%	20%	0%	0%	0%
L-Arginine	[M-H+2Na]+	25%	80%	100%	100%	75%	100%
	[M+Na-H2O]+	100%	100%	100%	100%	75%	100%
	[3M+H]+	100%	0%	100%	100%	100%	100%
	[2M+Na]+	100%	100%	100%	100%	100%	100%
	[M+K]+	100%	100%	80%	100%	100%	100%
	[M+H-NH3]+	100%	100%	100%	100%	100%	100%
	[2M+H]+	100%	100%	100%	100%	100%	100%
	[M+H-H2O]+	100%	100%	100%	100%	100%	33,33%
	[M+Na]+	75%	100%	100%	75%	50%	66,67%
	[M+H]+	100%	100%	100%	100%	100%	66,67%
L-Histidine	[M+H-H2O]+	100%	100%	100%	75%	50%	100%
	[M+H]+	100%	100%	100%	100%	100%	100%
	[M-H+2Na]+	50%	20%	100%	50%	50%	100%
	[2M+Na]+	50%	20%	100%	75%	75%	66,67%
	[M+H-NH3]+	50%	0%	40%	25%	75%	33,33%
	[2M+H]+	50%	20%	100%	75%	75%	100%
	[M+Na]+	50%	20%	100%	75%	75%	100%
	[M+K]+	0%	20%	40%	0%	25%	0%
L-Lysine	[2M+Na]+	25%	20%	20%	0%	25%	0%
	[M-H+2Na]+	25%	80%	20%	0%	25%	0%
	[M+Na]+	25%	80%	60%	0%	25%	0%
	[2M+H]+	25%	100%	80%	0%	75%	33,33%
	[M+H]+	100%	100%	100%	100%	100%	100%
	[M+H-NH3]+	100%	60%	100%	50%	100%	100%
	[M+H-H2O]+	100%	80%	100%	100%	100%	100%
L-Phenylalanine	[M-H+2Na]+	100%	100%	80%	100%	75%	100%
	[2M+Na]+	100%	100%	80%	100%	75%	100%
	[M+H-H2O]+	100%	100%	80%	100%	75%	100%
	[M+K]+	100%	100%	80%	100%	100%	100%
	[M+H]+	100%	100%	100%	100%	100%	100%
	[M+H-NH3]+	100%	80%	60%	75%	25%	100%
	[2M+H]+	75%	0%	0%	0%	25%	33,33%
	[M+Na-H2O]+	0%	40%	40%	0%	0%	33,33%
	[M+Na]+	0%	20%	20%	25%	0%	0%
L-Carnitine	[M+H]+	0%	60%	100%	100%	100%	100%
	[M+K]+	0%	60%	100%	100%	100%	66,67%
	[Cat-H]+	0%	0%	0%	25%	25%	0%

As a result of CliqueMS requiring an individual sample processing, we created Table 4 above showing, for each metabolite, the percentage of samples in which different adducts appear for each study time (0h, 12h, 24h, 48h and 96h), also distinguishing quality controls (QC).

3.2. Obtained pseudospectra

The obtention of the pseudospectra allowed us to obtain a graphical representation of the results. We compared pseudospectrums for the 7 metabolites we selected previously. In the pseudospectra we get from CAMERA, we showed the sample that the software itself considers the best annotated and most representative. In CliqueMS, on the other hand, we chose the spectrum of the sample that presented adducts with higher score values, and therefore we represented the sample that had a better annotation globally.

This comparison provided us the perception of the visual differences between the obtained pseudospectrum as well as it allowed us to associate the main peaks to annotated adducts in each example. Figures 16 to 22 represent the pseudospectra generated by each software, in order to show how each software created the groups of features we added tags to the adducts that were annotated by each of them.

In Figure 16 we observe that the most prominent peak, corresponding to [M+H]+, was represented in both cases. However, the annotation produced by CliqueMS presents a higher number of peaks than CAMERA.

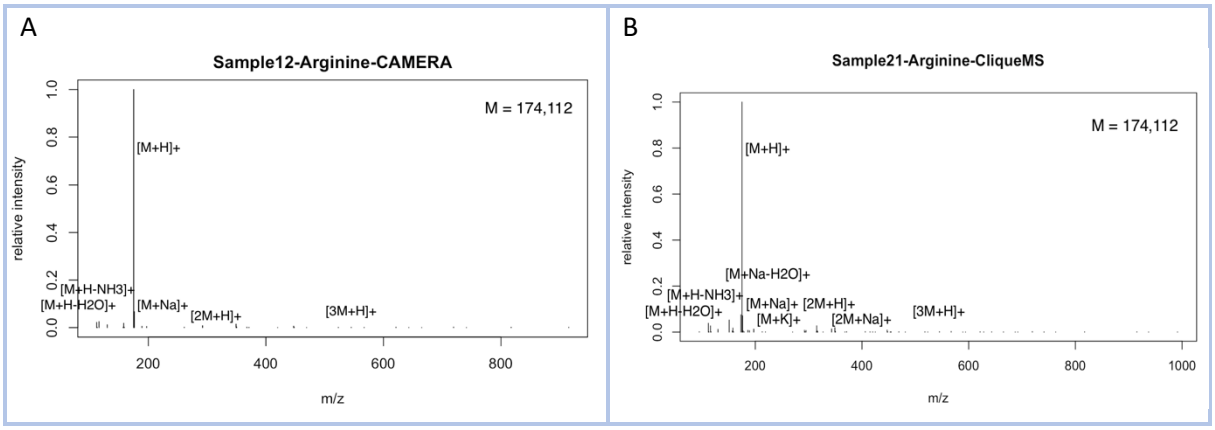


Figure 16 *- Representation of* ***Arginine*** *CAMERA pseudospectrum (A) and CliqueMS pseudospectrum (B)*

Figure 17 represented that the groups of peaks which appear are equivalent. It can be seen that the identification of the two main adducts found for this metabolite [M+H]+ and [M+K]+ were produced in both cases.

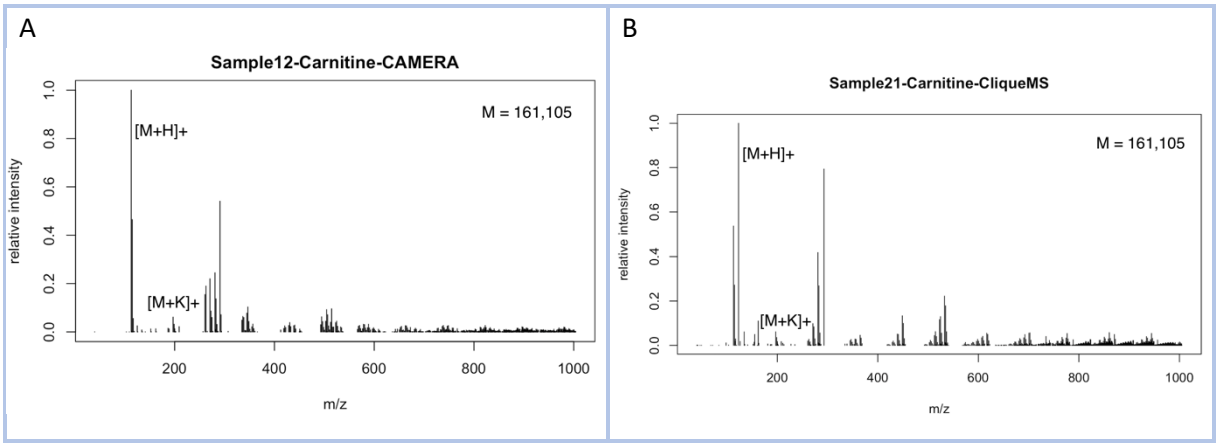


Figure 17 *- Representation of* ***carnitine*** *CAMERA pseudospectrum (A) and CliqueMS pseudospectrum (B)*

Figure 18 presented that CAMERA does not annotate the peak corresponding to the protonated mass, which is important in mass spectrometry because it is the most common and stable ionized species and serves as a reference standard for comparing the mass of other ions, also it can provide information about the chemical structure and molecular mass of a compound. This peak only appears in CliqueMS's pseudospectrum.

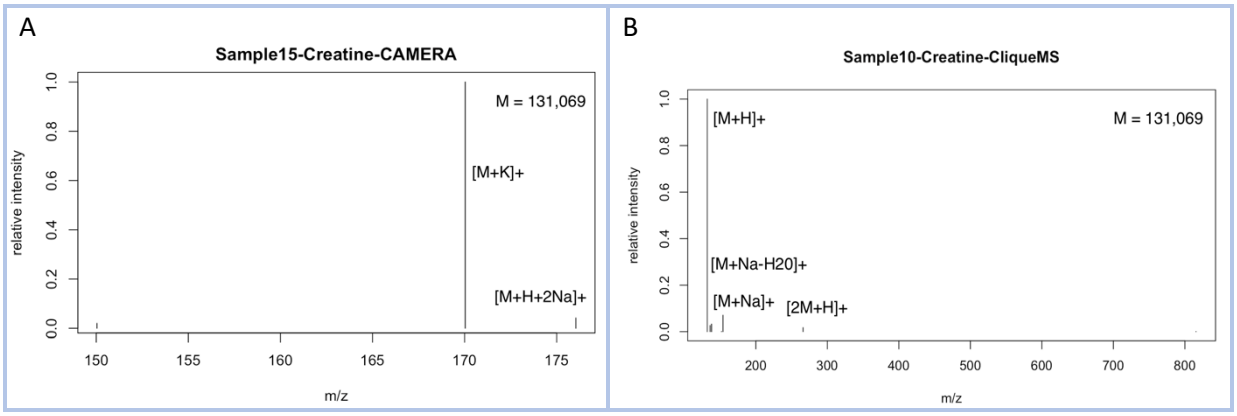


Figure 18 *- Representation of* ***creatine*** *CAMERA pseudospectrum (A) and CliqueMS pseudospectrum (B).*

For Glutamine it is possible for both methods to have annotated this metabolite as the [M+H]+ feature is present in both pseudospectra (Figure 19A and 19B) even though CliqueMS had selected a group which contained more features.

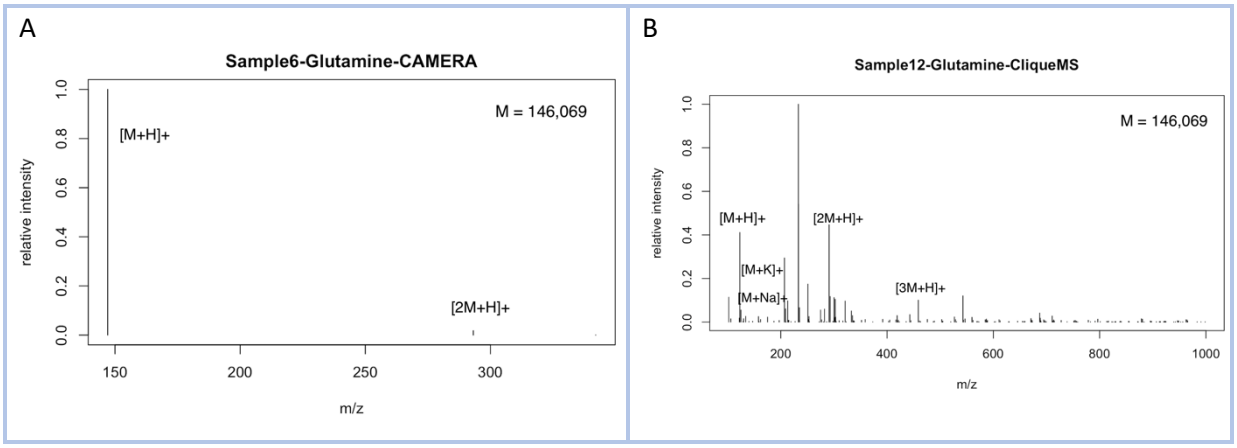


Figure 19 *- Representation of* ***Glutamine*** *CAMERA pseudospectrum (A) and CliqueMS pseudospectrum (B)*

As we exposed in Figure 20, CAMERA had only annotated isotopes, whereas CliqueMS, despite presenting a spectrum with a lot of background noise, presented relevant adducts that such as [M+Na]+ and [M+H]+.

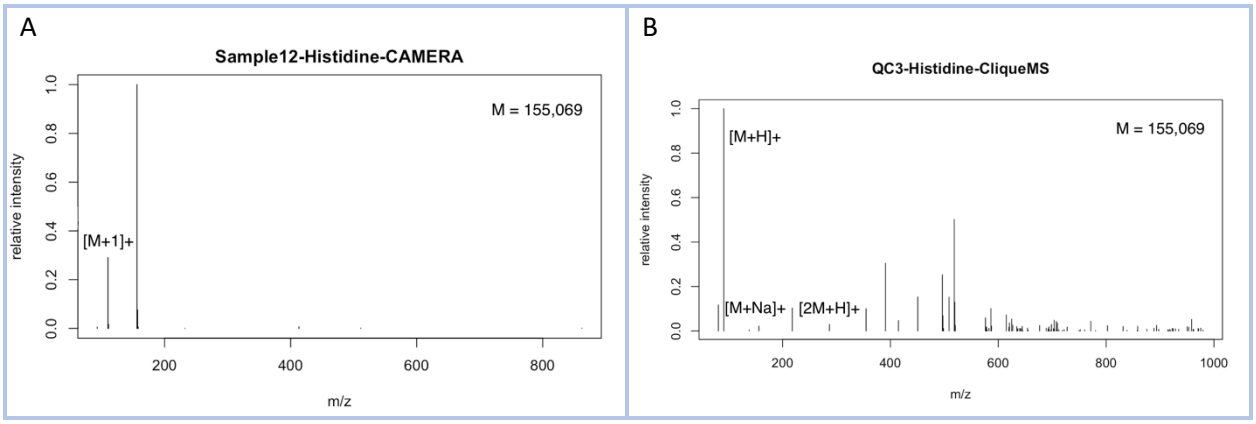


Figure 20 *- Representation of* ***Histidine*** *CAMERA pseudospectrum (A) and CliqueMS pseudospectrum (B)*

In Lysine, CAMERA pseudospectrum allowed the identification of a larger number of adducts compared to CliqueMS, see Figure 21. Additionally, almost in both graphical representations we realized there was the appearance of the same adducts and background noise.

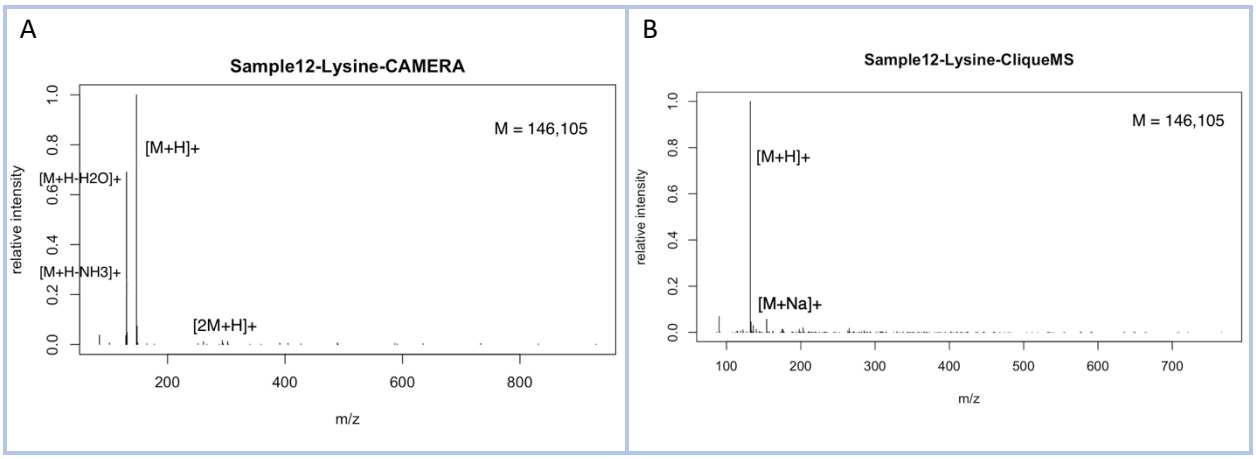


Figure 21 *- Representation of* ***Lysine*** *CAMERA pseudospectrum (A) and CliqueMS pseudospectrum (B).*

As we presented in Figure 22, almost both graphical representations showed the presence of the same adducts and background noise.

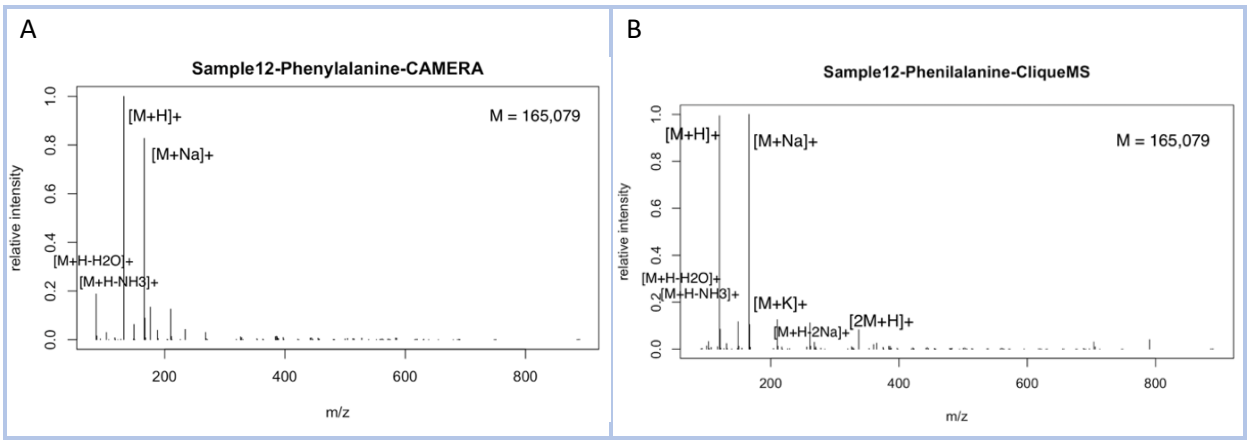


Figure 22 *- Representation of* ***Phenylalanine*** *CAMERA pseudospectrum (A) and CliqueMS pseudospectrum (B).*

3.3. Peak correlation on CAMERA samples

Figures 23 to 29 show the correlation between the peaks in each *pcgroup* associated with the evaluated metabolites. As a result of performing the peak correlation, we obtained seven heatmaps according to *pcgroups* associated to metabolites we expose next.
The correlation values range between 0 and 1 or -1. As we indicated in the legend located on the right side of each heatmap, the cells with shades of red correspond to a correlation value of around 1, if cells were blue, they mean that there was an inverse correlation between those peaks and finally, cells that appear yellow-white indicated low correlation with values close to 0. These heatmaps allowed us to obtain correlation coefficients based on the available data without filling in missing values. A greater positive correlation between peaks supports us that they were generated from the same compound.

In Figure 23, we demonstrated that apart from the [M+Na]+ adduct, which did not seem to exhibit correlation with the others. However, [M+H-NH3]+, [M+H]+, [2M+H]+, and [3M+H]+

presented a noticeable correlation, connections points between them presented a strong reddish coloration.

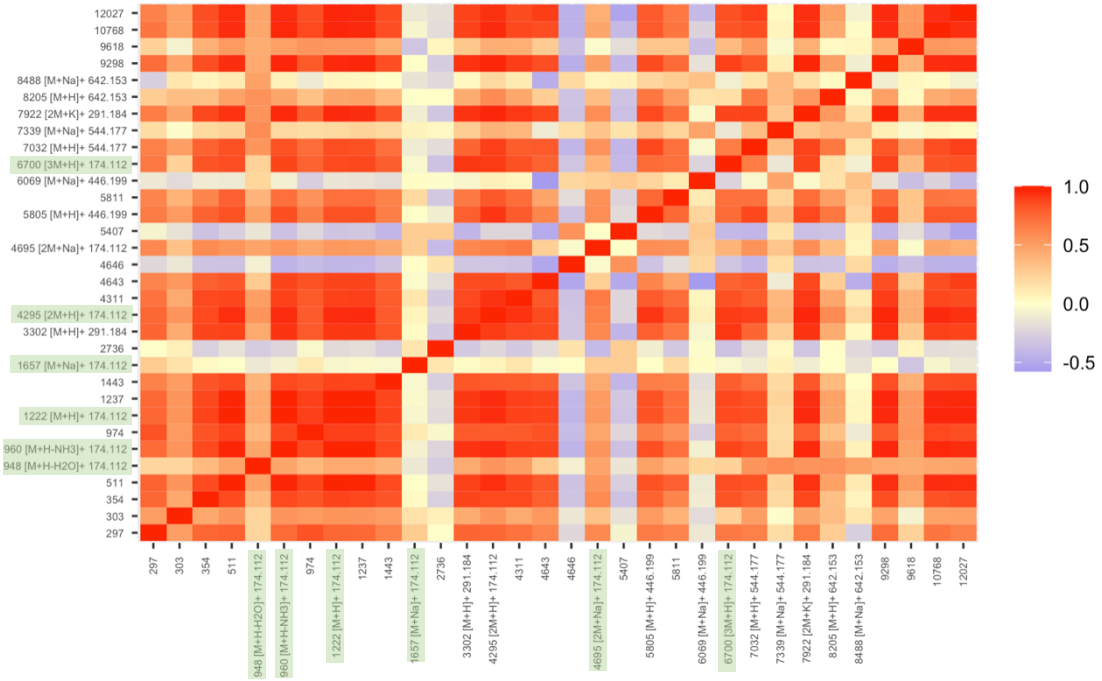


Figure 23 *- **Arginine** sample correlation heatmap. The correlated adducts of this metabolite are highlighted in green. The first value in each axis entrance refers to the CAMERA ID number, followed by the adduct formula and finally the molecular mass.*

Although we obtained a large number of peaks for Carnitine, we demonstrated by zooming in on the heatmap (Figure 24) that the previously annotated adducts (Figure 19) for this metabolite, [M+H]+ and [M+K]+, showed correlation, indicated by the green circle and arrow of the same colour.

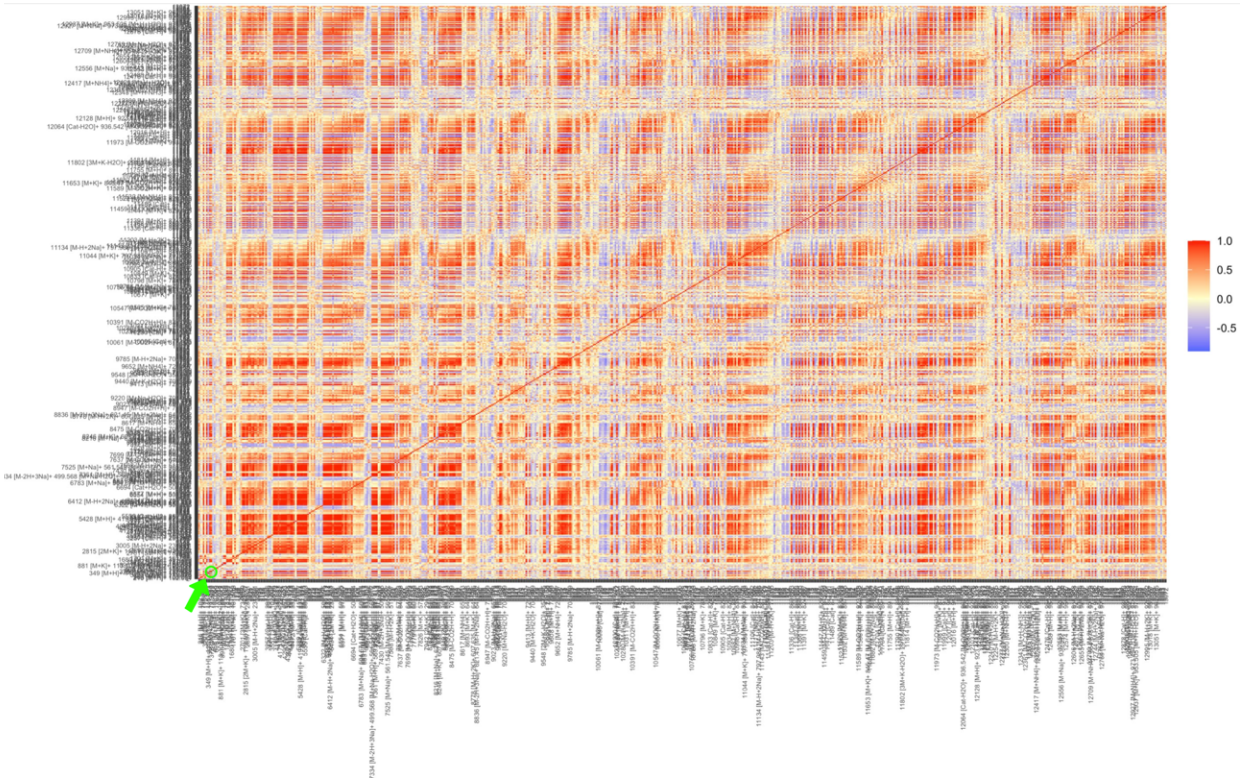


Figure 24 *- **Carnitine** sample correlation heatmap. The related adducts to this metabolite are highlighted in green. The first value in each axis entrance refers to the CAMERA ID number, then it figures the adduct formula and finally the molecular mass of the metabolite appears.*

We illustrated with Figure 25 that there was not any correlation between features apparently annotated for Creatine (Figure 18).

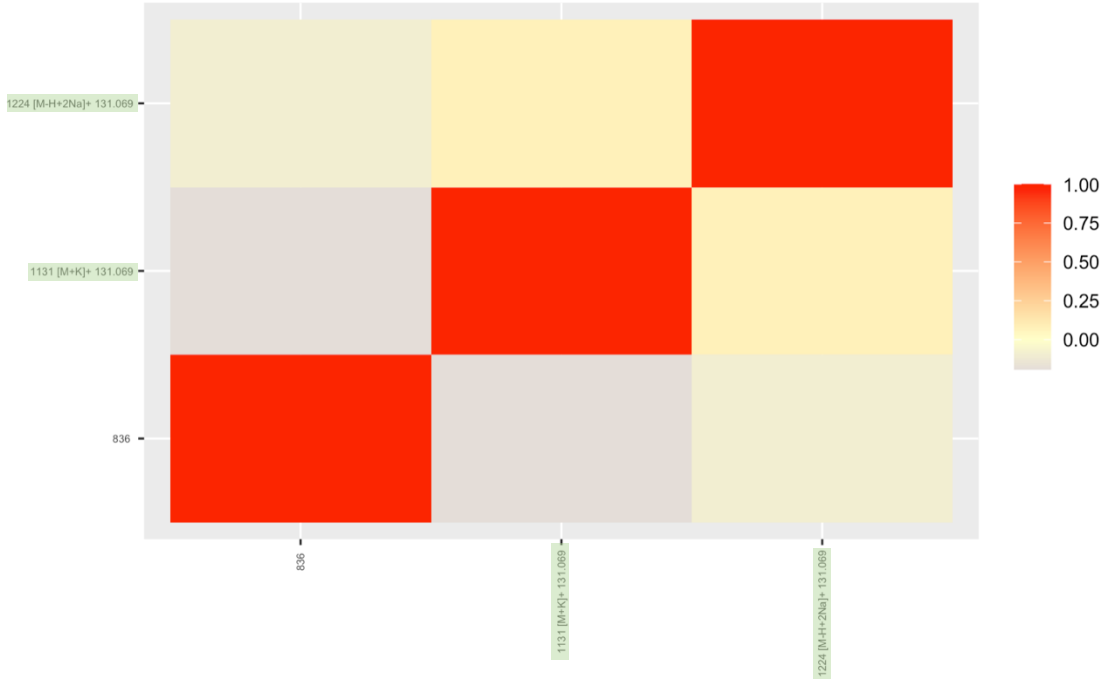


Figure 25 *- **Creatine** sample correlation heatmap. The related adducts to this metabolite are highlighted in green. The first value in each axis entrance refers to the CAMERA ID number, then it figures the adduct formula and finally the molecular mass of the metabolite appears.*

For Glutamine, the only two annotated adducts, [M+H]+ and [2M+H]+, as we represented in Figure 26, exhibited a significant correlation.

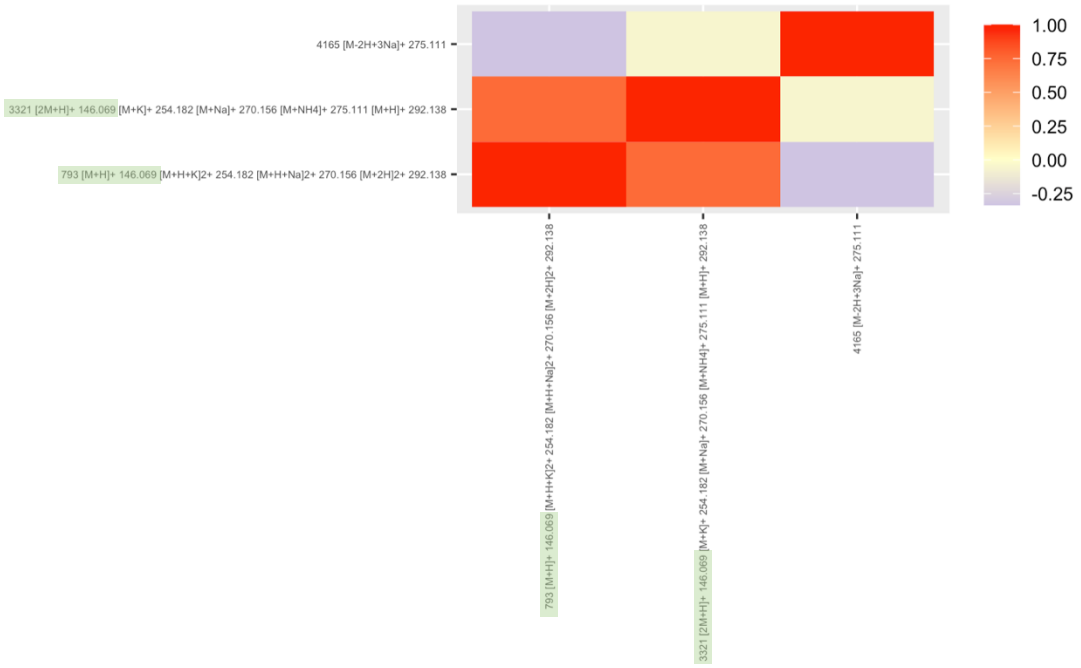


Figure 26 *- **Glutamine** sample correlation heatmap. The related adducts to this metabolite are highlighted in green. The first value in each axis entrance refers to the CAMERA ID number, then it figures the adduct formula and finally the molecular mass of the metabolite appears.*

Although Histidine did not present annotation for any adducts, isotopes for this metabolite were indeed annotated. As we demonstrated in Figure 27, the three isotopes, corresponding to peaks 943, 949, and 959, express a remarkable correlation.

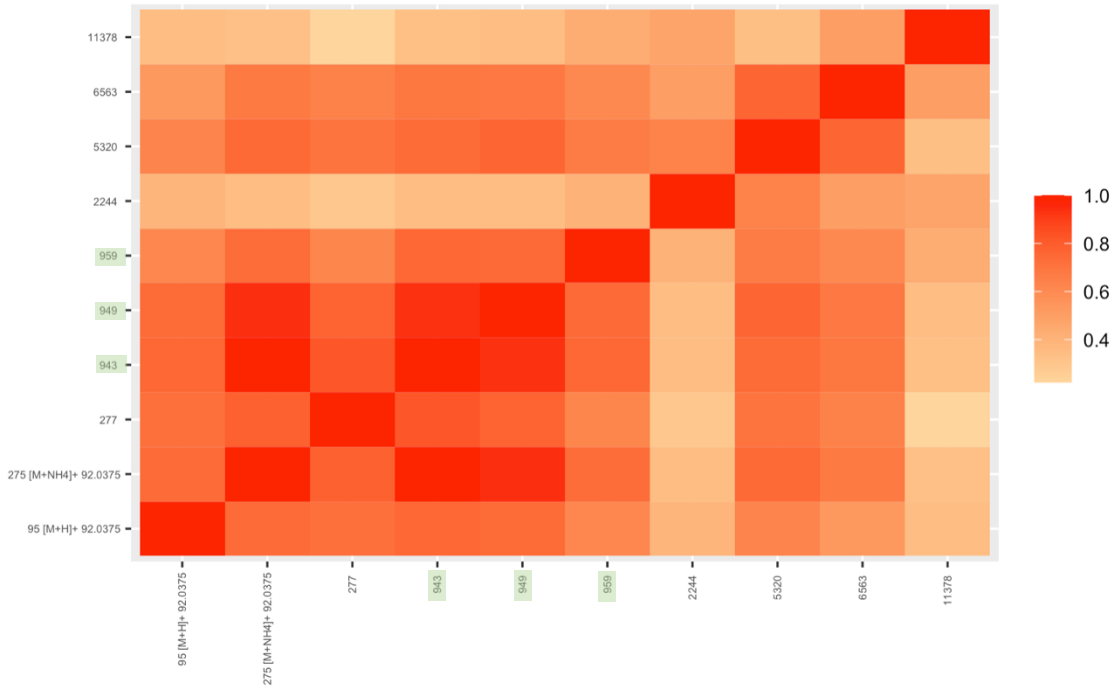


Figure 27 *- **Histidine** sample correlation heatmap. The related adducts to this metabolite are highlighted in green. The first value in each axis entrance refers to the CAMERA ID number, then it figures the adduct formula and finally the molecular mass of the metabolite appears.*

With Figure 28 we presented that there was a clear correlation between the annotated peaks for Lysine through CAMERA which are [M+H-H2O]+, [M+H-NH3]+, [M+H]+ and [2M+H]+.

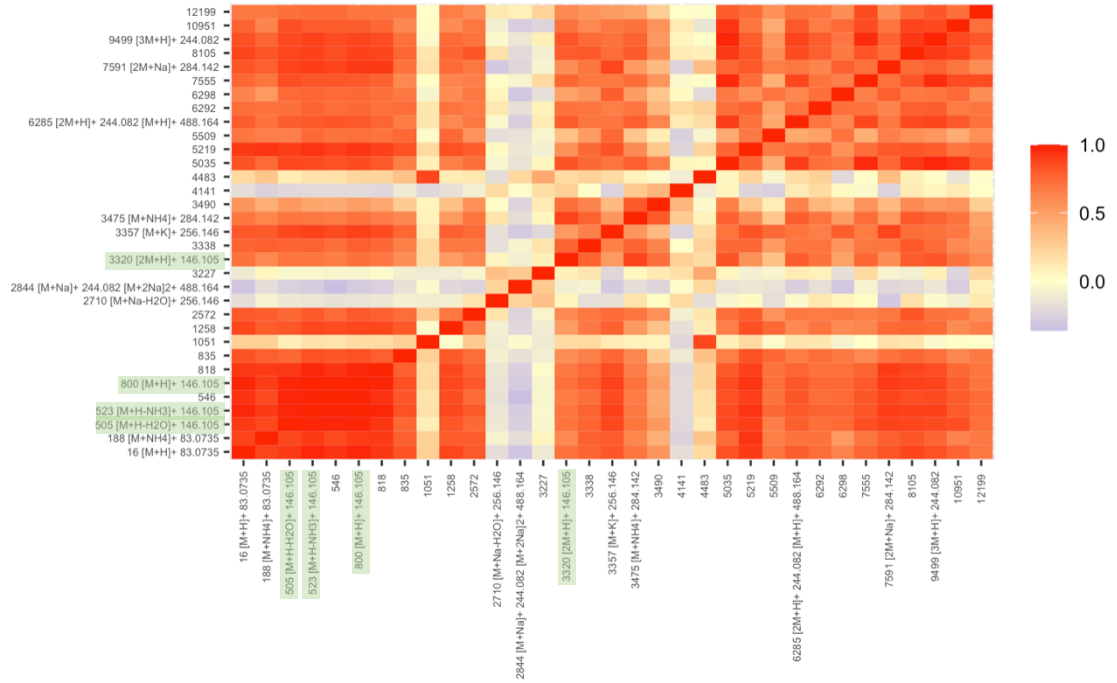


Figure 28 *- **Lysine** sample correlation heatmap. The related adducts to this metabolite are highlighted in green. The first value in each axis entrance refers to the CAMERA ID number, then it figures the adduct formula and finally the molecular mass of the metabolite appears.*

The heatmap for the metabolite Phenylalanine (Figure 29) did not present a significant correlation between the previously annotated peaks (Figure 22) which we highlighted in green in the present heatmap.

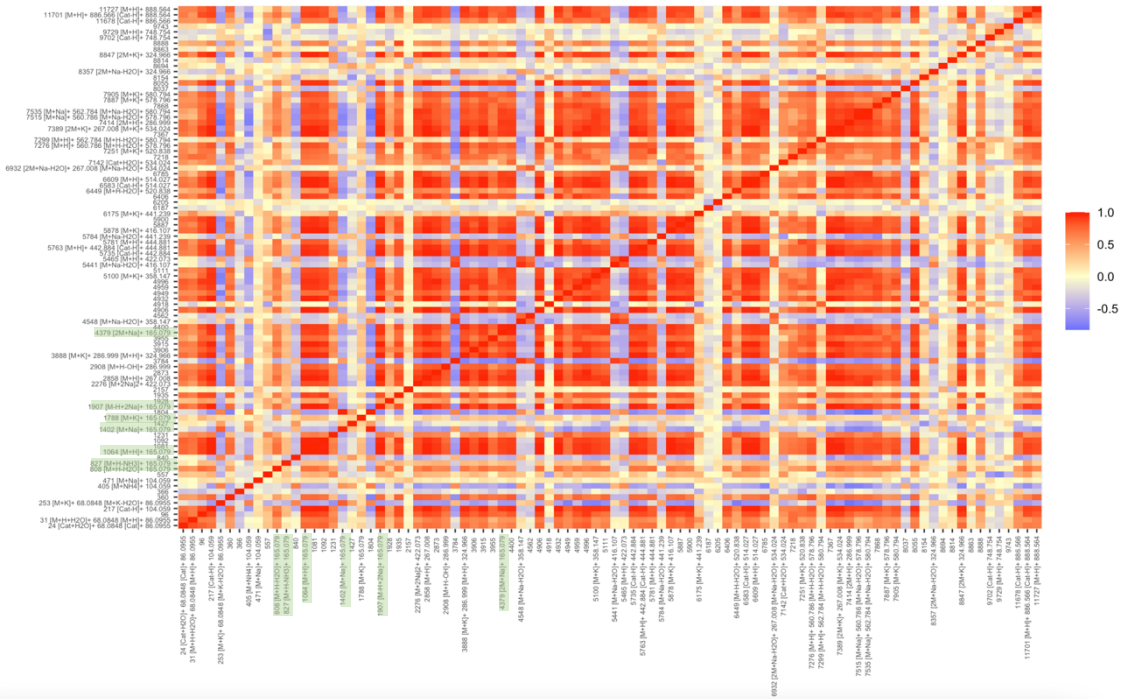


Figure 29 *- **Phenylalanine** sample correlation heatmap. The related adducts to this metabolite are highlighted in green. The first value in each axis entrance refers to the CAMERA ID number, then it figures the adduct formula and finally the molecular mass of the metabolite appears.*

3.4. Validation through comparison to reference spectra

Figures 30 to 36 show the comparison of the pseudospectra we obtained through the data processing by Camera and CliqueMS (in black in the upper part) with respect to the reference metabolite spectrum from the NIST database (in orange in the lower part). In all cases, we showed the pseudospectra that we represented previously and which correspond to the best example in terms of representation and score in each case.

Throughout the spectra, we realized that adducts located at the right of the corresponding peak of the protonated mass (having higher m/z values) do not normally appear in the reference spectra in the lower part (orange), but instead they appeared in the obtained pseudospectra from our CAMERA or CliqueMS processing. This happens because of ionization, which allows the obtention of mass spectra and triggers the addition of different molecules to the metabolite. Therefore, as it is a variable factor, we cannot take it into account when comparing with reference spectra unless it can be ensured that they have been obtained in the same way.

Conversely, what does appear in reference libraries are in-source fragments, which are derived from the fragmentation of the metabolite and have a smaller m/z value than the molecule in question. Some of these fragments may also be present in the pseudospectra obtained through experimental data and therefore may help in the identification of the metabolite.

To enable the alignment between peaks, we assigned a maximum mass difference of 50 ppm to consider that the features annotated.

As Figure 30 shows, in both pseudospectra, we corroborated that the metabolite was correctly identified, as the peak corresponding to the protonated mass, which is typically the most significant, is present in both cases. Both software tools had successfully identified Arginine because CAMERA and CliqueMS pseudospectra matched with 4 peaks with the reference spectra.

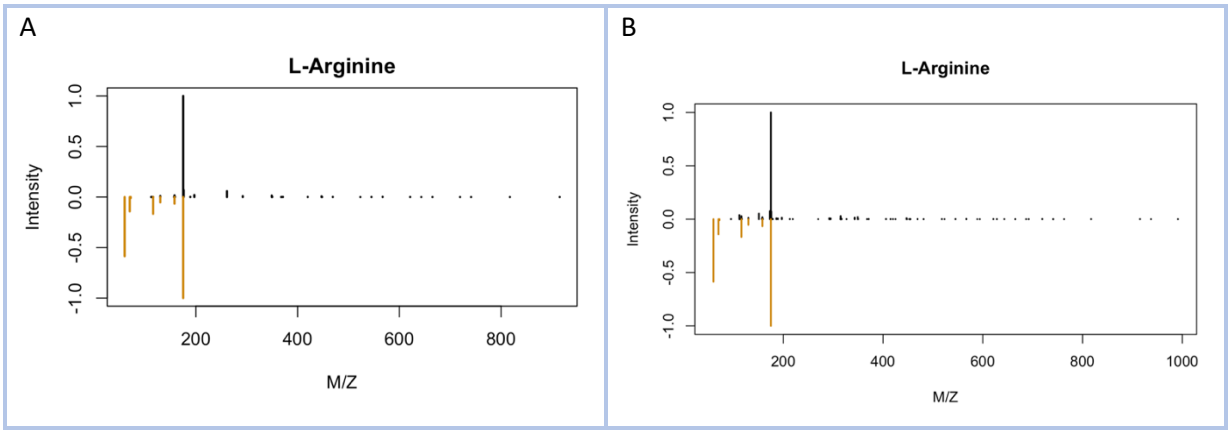


Figure 30 *- **Arginine** representation of CAMERA pseudospectrum (A) and CliqueMS pseudospectrum (B) against the metabolite from NIST database (in orange).*

For carnitine (Figure 31) the feature corresponding to the protonated mass appears in both pseudospectra (A and B). However, we could consider that the most properly annotated and identified metabolite would be through CAMERA as there are to peaks which matched with the reference, differently from CliqueMS, which aligned one single peak.

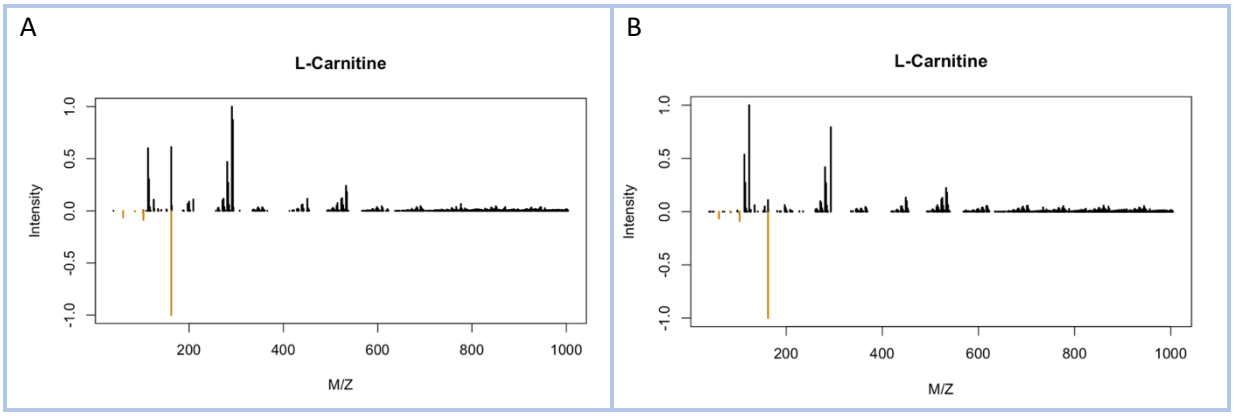


Figure 31 *- **Carnitine** representation of CAMERA pseudospectrum (A) and CliqueMS pseudospectrum (B) against the metabolite from NIST database (in orange).*

For creatine, we considered that only CliqueMS successfully annotated the metabolite. As we could see in Figure 32, CAMERA pseudospectrum did not allow the alignment with any of the reference peaks. However, there is a peak in CliqueMS which aligned with a reference peak.

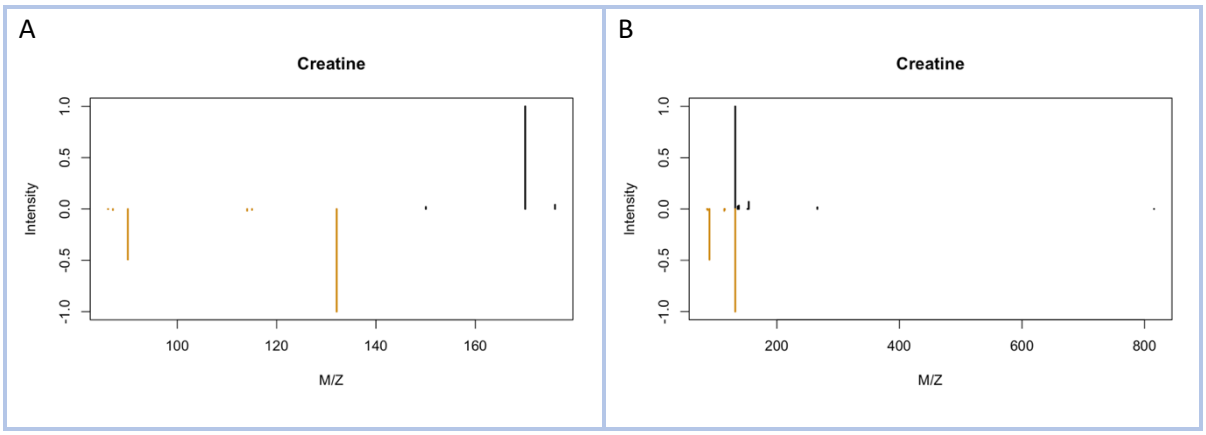


Figure 32 *- **Creatine** representation of CAMERA pseudospectrum (A) and CliqueMS pseudospectrum (B) against the metabolite from NIST database (in orange).*

As we presented in Figure 33, the peak with the highest intensity in the reference spectrum does not correspond to the protonated mass of the metabolite. However, a peak with a lower m/z value corresponds to M+H (around m/z=147) and it is only identified in CAMERA pseudospectrum, even though CliqueMS aligns one low intensity peak in m/z=130. Overall we considered CAMERA to have accomplished Glutamine annotation more effectively.

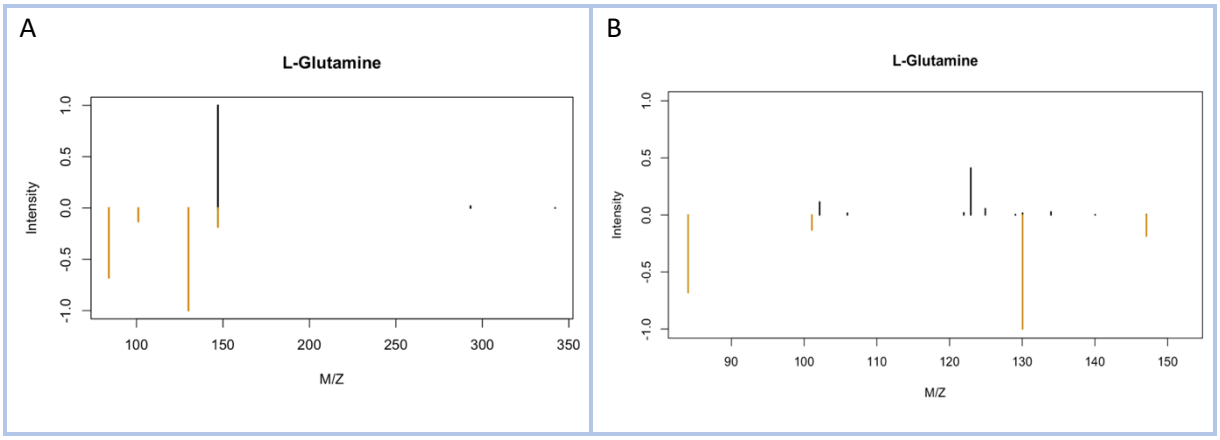


Figure 33 *- **Glutamine** representation of CAMERA pseudospectrum (A) and CliqueMS pseudospectrum (B) against the metabolite from NIST database (in orange).*

Although CAMERA only annotated isotopes, we could associate the peak with the highest intensity in the reference spectrum to the M+1 peak. However, for CliqueMS, there was not any peak that we could be associated with M+H, questioning the identification of histidine through data processing using this software. Overall, as we appreciated in Figure 34, we assigned a better accomplished annotation through CAMERA as its pseudospectrum allowed the alignment of two peaks compared to the only feature which was aligned by CliqueMS.

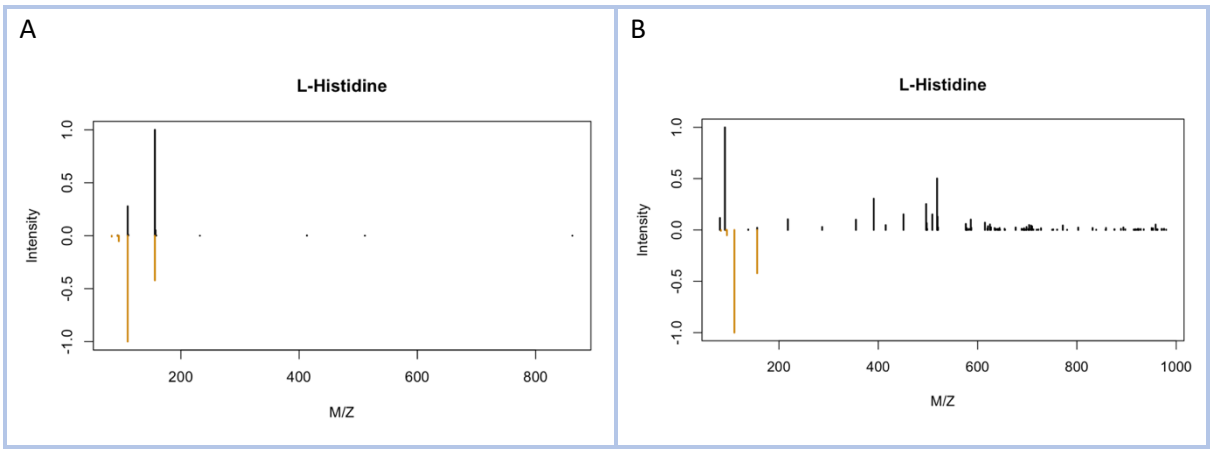


Figure 34 *– **Histidine** representation of CAMERA pseudospectrum (A) and CliqueMS pseudospectrum (B) against the metabolite from NIST database (in orange).*

The most intense peak in the reference spectrum corresponds to [M+H]+was annotated in both CAMERA and CliqueMS (Figure 35). We assumed a better completion of the metabolite annotation and identification through CAMERA, as three peaks aligned with the reference spectrum and only one CliqueMS peak matched to the reference.

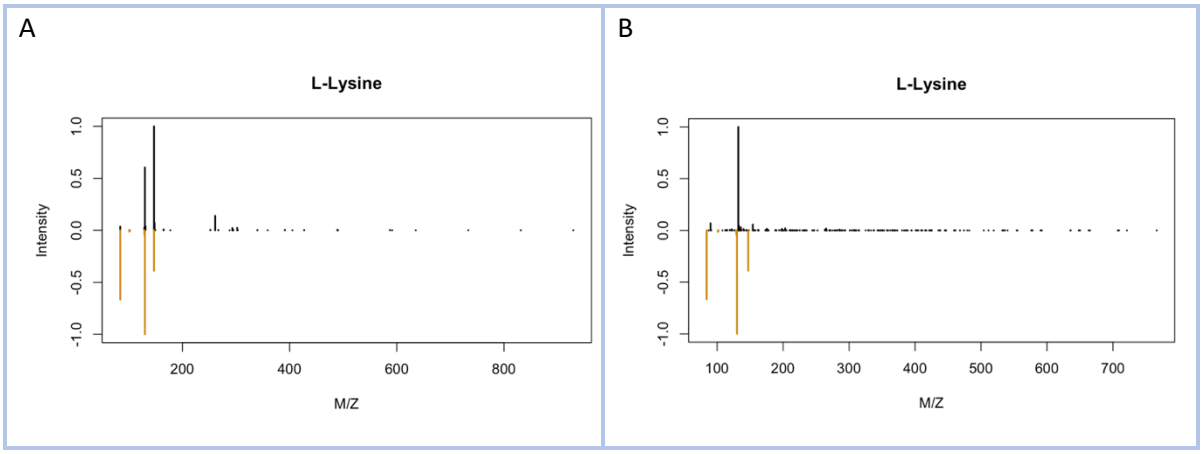


Figure 35 *- **Lysine** representation of CAMERA pseudospectrum (left) and CliqueMS pseudospectrum (right) against the metabolite from NIST database (in orange).*

For Phenylalanine, the most intense peak in the reference spectrum corresponding to the protonated amino acid mass only aligned with the pseudospectrum we obtained from CliqueMS processed data. However, it should be noted that in CAMERA, a less intense peak with the same m/z value had also been identified, so it cannot be concluded with certainty that the metabolite has not been identified or annotated through data processed by CAMERA. Considering what we showed in Figure 36, we postulated that CliqueMS better performed the annotation process for Phenylalanine as it aligned 7 features with the reference spectrum and moreover, the highest intensity peak corresponded to the protonated mass. Different from CAMERA which aligned only 4 peaks, instead.

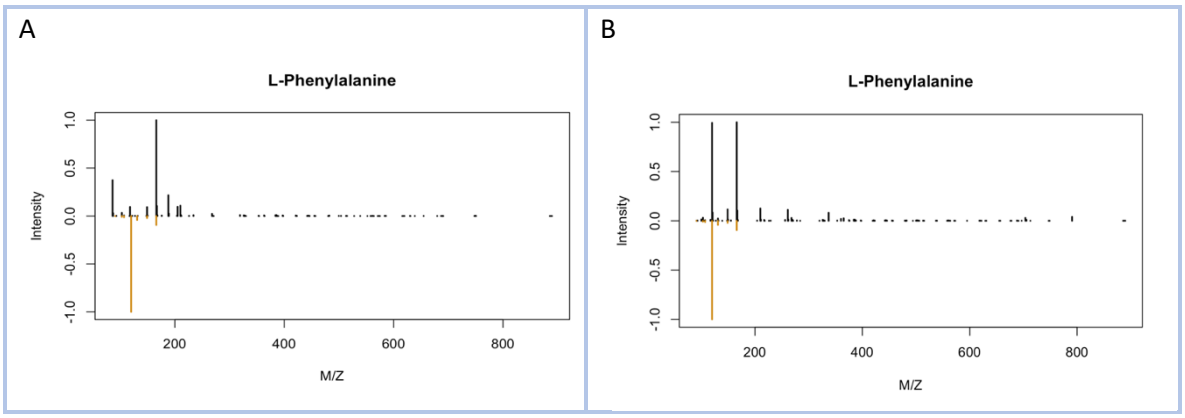


Figure 36 *- **Phenylalanine** representation of CAMERA pseudospectrum (A) and CliqueMS pseudospectrum (B) against the metabolite from NIST database (in orange).*

4. DISCUSSION

The two software we compared, CAMERA and CliqueMS, have a large number of dissimilarities in terms of parameters and execution, as well as in the obtained results at the final data processing and filtering.

Regarding to the utilized **parameters and execution** in the first place we noticed that feature grouping was distinct in the two R-packages. It consists on a process which groups similar features together considering their attributes or characteristics. It is often used to simplify feature selection or to reduce the dimensionality of high-dimensional data (87). In the context of metabolomics, **feature grouping** consist on the construction of a similarity network between features corresponding to the same metabolite. CAMERA uses the Pearson Correlation coefficient, which detects the similarity of features that are linearly growing or decreasing and also considers the relative importance of different features. Nonetheless, it is not an optimal option when features are non-linear. Additionally, CAMERA implements the fast retention time-based grouping as well a graph-based algorithm to integrate the peak shape analysis, isotopic information and intensity correlation across samples (14). Besides, CliqueMS employs Cosine similarity, which assesses the alignment between intensity vectors but does not consider the relative importance of different features. It is demonstrated that the cosine similarity has a superior discriminatory power than the Pearson correlation (14,76,88).

Furthermore, as it is said, CliqueMS produces a smaller number of **groups** than CAMERA (76). And we corroborated it in our project. According to this, we apparently assigned to CliqueMS a more completed annotation performance because groups contained more associated features due to the generation of a smaller number of them. However, each *cliqueGroup* created by CliqueMS was not necessarily associated to a single metabolite and as we presented in Figures 16 to 22, CliqueMS pseudospectra contained considerably more background noise than the ones obtained from CAMERA processing as we showed in Histidine (Figure 20) and Glutamine (Figure 19) examples. With reference to *pcgroups* created by CAMERA, ideally each one corresponded to a certain metabolite although pseudospectra may remain incomplete as features corresponding to the same metabolite can be assigned to different *pcgroups* or they can even be lost, resulting in a difficult association to a certain metabolite, that is the case of Histidine (Figure 20A) and creatine (Figure 18A).

Moreover, CAMERA is capable of handling a wide range of sample types and **data formats** like GC-MS, while CliqueMS is specifically designed for processing LC/MS data, which is a disadvantage unless it did not affect our project procedure.

Additionally, CliqueMS related bibliography mentioned that normally around 200 features were detected per sample. However, when samples are much larger, like in our project, data processing could not be performed on a standard computer and instead, we required to use a more powerful machine to run the analysis (see Methods). In contrast, we performed CAMERA sample processing on a conventional computer using RStudio, regardless of their size or the number of detected features for each processed sample (76). We demonstrated that CAMERA

has a greater **availability and comfort** than CliqueMS as it was interesting for us to perform the whole analysis in RStudio.

Although in methodology and results part we assumed that *pcgroups* and *cliqueGroups* are associated with a single metabolite in order to carry out the comparison, it should be taken into account that *cliqueGroups* are a concept in CliqueMS that refer to groups of mass spectral features that show similar correlation patterns (at least one peak in common), consequently ***cliqueGroups*** do not necessarily correspond to a single metabolite because these features corresponding to different metabolites that coelute sometimes can be assigned wrongly to the same metabolite. However, their developers point that it has to be considered that CliqueMS's priority is to annotate large amounts of adducts or fragments associated to the same parental mass rather than annotating the same features with two different parental masses and more common adducts (76). In the context of CAMERA package, the concept of *pcgroup* refers to "peak groups", which are clusters of mass spectral peaks that represent a single metabolite. As we commented before, those groups are constructed based on the similarity of their mass spectral patterns across multiple samples or experimental conditions. Even though both software required tools to process the obtained data so as to achieve further annotation and identification, they needed the use of additional information such as retention time, accurate mass, isotopic patterns, fragmentation spectra and reference data bases which helped to identify putative metabolites. The difference is that when CAMERA generates the ***pcgroups*** these typically represent a single metabolite or at least a group of peaks that are likely to be originated from the same compound (14,74).

Regarding to XCMS **raw data file processing**, we corroborated that in CAMERA more processing steps were required (see S1.1.1 in Supporting Material) as it is a designed package to post-process feature list generated by XCMS as well as to collect peaks related to a metabolite into a compound spectrum. On the other hand, even though CliqueMS also required XCMS library to pre-process raw data, less steps were needed because functions are optimized to be more easily used in this package in an efficient manner, consequently, as we can see in S1.1 and S1.2 in Supporting Material, data pre-processing was easier achieved in CliqueMS.

An advantage of CAMERA we perceived is that it allowed the generation of **pseudospectra** with a function included in the software, which selects the most significant sample for the pcgroup in question, this software also provided another **visualization tool** based on heatmaps to show metabolite abundance data. In contrast, if we want to obtain metabolite pseudospectra in CliqueMS, we need to generate them manually by representing the m/z ratio with respect to the intensity. In addition, CAMERA has a function that allows specific **neutral losses** to be found by searching their exact molecular mass using a formula included in the R package (*FindNeutralLoss*) (74).

In terms of **obtained results**, one of the main differences we considered between the two software was the **number** of **samples** which could be processed at a time. With reference to this, CAMERA was capable of processing both individual and multiple samples simultaneously, while CliqueMS could only process one sample at a time (14,74,76). As a consequence, CAMERA annotation process resulted to be much shorter in time and more efficient so it was possible for

us to determine the annotation of adducts for all 44 metabolites under study, as we represented in Table 2 which allowed us to have a general idea of the present metabolites in the samples. This overall view was not possible to be achieved in CliqueMS as it required considerably more time and effort for processing all the samples for all the metabolites, additionally, this software did not permit a whole sample homogenization of the created feature *cliqueGroups* because every sample contains different *cliqueGroup* numbers for the same metabolite and definitely, it represented a significant disadvantage for this R-package.

The information regarding the number of **metabolites that were associated** with at least one feature suggested us that CAMERA may identify fewer metabolites in general, specifically three less according to the graph presented in Figure 15. It is important to note, however, that the value we associated with CliqueMS represents an average of all samples. Additionally, as it is commented previously, we cannot rule out that CliqueMS may mistakenly associate features that correspond to the metabolite, which would lead to possible confusion in the molecule identification process.

Upon a closer **examination of the individually processed samples by CliqueMS**, we appreciated that for each time group (0h, 12h, 24h, 48h and 96h), the percentage of identified samples remained consistent (high percentages) within that group if the metabolite was present in that condition, as we indicated in Table 4. This consistency is logical because it is believed that the presence of metabolites varies over time due to the effect of reducing conditions on macrophages, but not within a specific time group. As a result, we took advantage of this time consuming CliqueMS procedure to obtain a more detailed and personalized result which allowed to represent data according to what happened along the study time and also we got information about samples reliability or preservation when they did not annotate an adduct found in the rest of samples in the time group.

Regarding to Table 3, we noted a significant difference in the **number of adducts** associated with the 7 selected metabolites. CliqueMS identified a much larger number of adducts and neutral losses and therefore provided a more comprehensive annotation compared to CAMERA. With respect to the number of isotopes, CliqueMS also identified more of them which can apparently contribute to a more accurate identification of the putative metabolite. However, it should be noted that the annotation result from CliqueMS corresponded to all 5 annotations found, and some of them had low scores, hence, they might be unreliable or not significant when being taken into account. Furthermore, despite the fact that CliqueMS assigned rankings to annotations based on their likelihood and generated the top five probable annotations for each clique in contrast to CAMERA which provided a single annotation, it should be emphasized that the score assigned to an annotation is contingent upon the size of the clique or group of features. As a result, a direct comparison of annotation scores for different groups of features was not possible.(76)

Moreover, in the Table 3, we perceived that the adducts and in source fragments associated by CAMERA with the metabolite were present in all cases in the annotation produced by CliqueMS, except for Histidine (Figure 20) for which CAMERA only found isotopes. We noted that in most examples, adducts annotated by CAMERA were present in high percentages along time in the

samples we evaluated with CliqueMS, it is perceivable for [M+H]+ in Glutamine; [3M+H]+, [2M+Na]+, [M+H-NH3]+, [2M+H]+, [M+H-H2O]+, [M+Na]+, [M+H]+ in Arginine; [M+H]+ and [M+H-H2O]+ in Lysine; [M-H+2Na]+, [2M+Na]+, [M+H-H2O]+, [M+K]+ and [M+H]+ in Phenylalanine and finally adduct [M+H]+ in carnitine, we collected these examples in Table 4. Therefore, we concluded that these adducts are likely to have a greater significance or importance. Additionally, it was remarkable that all the metabolites presented the [M+H]+ adduct, which is the most representative and informative.

In terms of **pseudospectra visualization**, we presented in Figures 16-22, which displayed the different pseudospectra generated by the two software, with them we represented the associated peaks with each group (*cliqueGroup* or *pcgroup*) and thus, to a specific metabolite. There was a notable difference in Creatine (Figure 18), Glutamine (Figure 19), and Histidine (Figure 20). While CAMERA produced pseudospectra with reduced background noise, CliqueMS generated them for these same metabolites which exhibited a significantly higher number of peaks. Likewise, we did not see such a considerable difference in Lysine (Figure 21) and Arginine (Figure 16). Finally, and regarding to Carnitine (Figure 17) and Phenylalanine (Figure 22) we considered their pseudospectra comparable in terms of the amount of peaks they presented and their resultant annotation.

Although CliqueMS generated pseudospectra contained a higher **number of peaks** than those obtained from CAMERA processed data, they did not always present a more accurate association with reference metabolites, as we saw in the case of Glutamine (Figure 19), Histidine (Figure 20) or Lysine (Figure 21). Moreover, we observed in the CliqueMS examples (from Figure 16 to 22, side A) that there might be the presence of peaks that could correspond to other metabolites, but their identification was not possible due to the lack of sufficient data.

To validate the obtained pseudospectra and determine which ones closely match **the reference spectrum** in terms of the relevant peaks, including common features like in-source fragments, it was important to consider that the intensity of peaks in the reference spectrum may differ from the pseudospectrum we obtained experimentally. This happens due to the use of different methods and equipment to obtain them, resulting in variations in the fragmentation patterns. Furthermore, in the pseudospectra we generated from CAMERA and CliqueMS data, a higher number of lower-intensity peaks were observed compared to the reference spectra as we proved in Figures 30-36. Although there aren't many fragments that match the reference, this might be because of the different amounts of energy that each fragment has been exposed to. When the molecule is subjected to a higher collision energy, it gets more fragmented and as a result, more peaks appear in the spectra. Additionally, it has to be noted that while NIST is representing a spectrum obtained from experiments with a collision energy (CE) of about 10eV and our experiment was undertaken with CE=0eV. However, the represented groups of features in CAMERA and CliqueMS pseudospectra contained more peaks due to the presence of a larger number of molecules present in the samples.

It seemed evident to us that neither of the two software programs provided an exact alignment with the reference spectrum. However, we were able to determine which of the two software programs had **effectively annotated metabolites**, thus enabling an optimal identification of

them. For example, CliqueMS seemed to perform better in identifying Creatine (Figure 32) and Phenylalanine (Figure 36), while CAMERA showed closer resemblance to the pseudospectrum for Glutamine (Figure 33), Histidine (Figure 34), and Lysine (Figure 35). Conversely, we considered that both software programs successfully identified Arginine (Figure 30) and Carnitine (Figure 31) based on the obtained annotations, even though carnitine seems to be finer characterized by CAMERA in terms of peak intensity matching, we considered that CliqueMS also had identified this metabolite properly. Considering what we observed, it appeared to us that CAMERA generally exhibited a more precise alignment with the reference metabolites as it had achieved an accurate identification of 55% of metabolites based on the generated pseudospectra from the feature annotation of the 7 analyzed samples. Consequently, CliqueMS successfully identified the remaining 45% of those metabolites.

Moreover, we could only perform the **correlation analysis** in CAMERA since all samples were processed together and had common pcgroups. This represents a clear disadvantage for CliqueMS because due to its lack of peak grouping homogenization, intensity feature correlation could not be performed. Correlation analysis are an important step for evaluating peaks in a mass spectrometry experiment as it enables the identification of peaks or samples that are highly correlated with each other and consequently, are more likely to represent the same metabolite. (89)

In this project, we undertook the correlation between peaks, where a strong correlation (around 1) between them provided stronger evidence that those features may have been originated from the same metabolite (Figures 23-29). Also, we could have conducted the correlation between samples, where a strong correlation between them would have indicated that their pseudospectra would be related, suggesting the presence of the same compound in different samples and allowing the comparison against metabolites changes along time, an analogy to what we presented in Table 4. However, we contemplated that it was much more valuable to perform the correlation between peaks in terms of evaluating the annotation process, and that is the reason why we chose to only perform that correlation.

With respect to peak or **feature correlation** we noticed that there were two metabolites, creatine (Figure 25) and Phenylalanine (Figure 29) which did not present a considerable correlation or any at all between the adducts we assumed to have been annotated previously (Figures 18 and 22). Hence, this observation gave rise to uncertainties regarding the true association between these adducts and the respective metabolites (creatine and Phenylalanine) through CAMERA. Furthermore, the aforementioned conclusion we obtained from the correlation analysis, remained consistent and was strengthened by the outcome of the comparison with reference spectrum. We perceived that for these two metabolite in particular, CAMERA was not capable to facilitate a precise annotation and identification.

On the top of that, we also emphasized that samples included 4 **quality controls** which consisted of a pooled samples formed by the combination of equal amounts of each experimental samples in study. Those 4 pool samples we utilized where analysed in the same way as individual samples and are essential to ensure the accuracy, reproducibility and reliability of experimental results. On the one hand, it allows the detection of systematic errors in the experimental process, such

as variability in the sample processing, differences in instrument performance or batch effects. Further, the obtained results from the pooled simples compared to expected values can be detected and be used to identify the source of the error. On the other hand, pooled samples are also utilized as a reference sample for normalization or to correct technical variability in the data. It is said that by comparing the results obtained from the individual simples to those obtained from the pooled sample, it is possible to correct any systematic differences in data (90–92).

Another aspect to mention is that when we associated features to metabolites, we had been quite restrictive by setting a tolerance of 10 parts per million (ppm) as the **error threshold**. Therefore, only features with a mass difference within this narrow range are considered for association. It is possible that by relaxing this restriction, more features could have been detected for the same metabolite, particularly in the cases of Histidine and Glutamine in CAMERA, as well as Creatine in both CAMERA and CliqueMS. Furthermore, the error value of 50 ppm we chose so as to align pseudospectra with reference mass spectra could also have been less restrictive. However, it is important to note that a higher ppm error can lead to reduced confidence in the identification of the metabolite and may require further validation or confirmation using additional analytical methods.

5. CONCLUSION

It was concluded that, although the protocol or workflow followed by both software tools was nearly equivalent, it has been observed that they exhibit certain peculiarities that lead to significant differences in their final results and the interpretation of the processed data.

CAMERA and CliqueMS are two distinct software tools used for the analysis of metabolomics data, and they present notable differences in their methodologies and outcomes. On the one hand, it is true that while CliqueMS requires an extensive and individual sample processing, it yields a more comprehensive annotation of adducts and in-source fragments and generates fewer groups as it manifests a superior feature grouping performance compared to CAMERA, owing to the cosine similarity metric utilized, which is capable of identifying non-linear relationships, which are often prevalent in metabolomics data. However, these groups are not always associated to a single metabolite. In contrast, CAMERA utilizes Pearson correlation, which assumes a linear relationship between feature abundances. On the other hand, CAMERA did not provide as comprehensive quantity of identified adducts as CliqueMS. Nonetheless, it allowed us the processing of multiple samples, significantly reducing the time invested in data analysis. Additionally, its outcome in terms of the created groups and the spectra generated by the program itself were better designed to be associated with a single metabolite, even if it results in more groups compared to CliqueMS. Nevertheless, this may occasionally lead to missing of peaks in pseudospectra because they will be incorrectly linked to another *pcgroup*. These observations allowed us to confirm our hypothesis.

Despite the project's limitations in terms of time and lack of experience in bioinformatics, we achieved a successful comparison between the two software tools. Firstly, both packages were properly utilized, and we accomplished the understanding of their annotation processes and final results. It is important to highlight the progress made from having raw data samples to annotate some of the metabolites in study and the extent to which the goals were achieved.

Moreover, it is clear that metabolomics is playing an increasingly significant role in systems biology. Particularly, in the data analysis stage of a mass spectrometry experiment, there is a lack of well-equipped and validated bioinformatics tools that can successfully provide results with sufficient information and validity for the annotation and association of different ion peaks derived from molecules. This is crucial for the interpretation and determination of which metabolite corresponds to the different sets of peaks. Consequently, it is also evident that there is a need to develop a tool which integrates the identified shortcomings in the feature annotation aspect of data analysis in order to improve and optimize the peak annotation and the subsequent metabolite identification process. Based on our experience from the completion of this project an ideal bioinformatic tool could involve a software which permitted the processing of large-scale sample sets collectively (as in CAMERA) to obtain a more comprehensive annotation result (as in CliqueMS), while maintaining an integrated workflow through RStudio without using any powerful web server , as well as employing a grouping algorithm which created a suitable number of groups in order to get a complete and precise adduct annotation and further metabolite identification.

In conclusion, the existing tools and methodologies for metabolite annotation represent a significant advancement towards the wider utilization of metabolomics. Nevertheless, achieving a complete annotation and metabolite interpretation from complex mass spectrometry data still relies on the development of novel computational resources, algorithms, and instrumental advancements. Sustained efforts in these area are needed to drive a further progress in the field and reveal the complete potential of metabolomics in comprehensive metabolite annotation.

Even though each software tool has its own set of strengths and limitations such as the sample size, whether multiple or individual sample processing is more suitable, the presence of replicates, the desire to study samples in a general manner, or the need for more accurate annotation results. The software selection will depend on the specific research question being addressed and the characteristics of the studying data. After verifying the obtained pseudospectra from the CAMERA and CliqueMS data through the annotation result, performing the comparison with reference spectra and even obtaining the feature correlation in CAMERA, we concluded that metabolite annotation and identification was in general more accurately achieved by CAMERA regarding to the studied molecules although this software provided a more summarized annotation result than CliqueMS.

6. BIBLIOGRAPHY

1. Chhabil Dass. Tandem Mass Spectrometry. In: Fundamentals of Contemporary Mass Spectrometry. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2006. p. 119–50.
2. Chhabil Dass. Fundamentals of Contemporary Mass Spectrometry. 2006.
3. Rubakhin SS, Sweedler J V. A mass spectrometry primer for mass spectrometry imaging.
4. Smith RW. Mass Spectrometry. Encyclopedia of Forensic Sciences: Second Edition. 2013 Jan 1;603–8.
5. J.R. Chapmasn. Practical Organic Mass Spectrometry. 2n Edition. London; 1993.
6. Niessen WMA (Wilfried MA),. Liquid chromatography--mass spectrometry. Third. 2006. 23–45 p.
7. E. de Hoffmann JCVS. Mass spectrometry, Principles and Applications. 1996.
8. Ho CS, Lam CWK, Chan MHM, Cheung RCK, Law LK, Lit LCW, et al. Electrospray ionisation mass spectrometry: principles and clinical applications. Clin Biochem Rev. 2003;24(1):3–12.
9. Banerjee S, Mazumdar S. Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. Int J Anal Chem. 2012;2012:1–40.
10. Haag AM. Mass Analyzers and Mass Spectrometers. In 2016. p. 157–69.
11. El-Aneed A, Cohen A, Banoub J. Mass Spectrometry, Review of the Basics: Electrospray, MALDI, and Commonly Used Mass Analyzers. Appl Spectrosc Rev. 2009 Apr;44(3):210–30.
12. Varghese RS, Zhou B, Nezami Ranjbar MR, Zhao Y, Ressom HW. Ion annotation-assisted analysis of LC-MS based metabolomic experiment [Internet]. 2012. Available from: http://www.proteomesci.com/content/10/S1/S8
13. Hocart CH. Mass Spectrometry: An Essential Tool for Trace Identification and Quantification. In: Comprehensive Natural Products II. Elsevier; 2010. p. 327–88.
14. Kuhl C, Tautenhahn R, Böttcher C, Larson TR, Neumann S. CAMERA: An integrated strategy for compound spectra extraction and annotation of LC/MS data sets. Available from: http://pubs.acs.org/.
15. Faull KF, Dooley AN, Halgand F, Shoemaker LD, Norris AJ, Ryan CM, et al. Chapter 1 An Introduction to the Basic Principles and Concepts of Mass Spectrometry. In 2008. p. 1–46.
16. Brunnée C. The ideal mass analyzer: Fact or fiction? Int J Mass Spectrom Ion Process. 1987 Jun;76(2):125–237.
17. Allen DR, Mcwhinney BC. Quadrupole Time-of-Flight Mass Spectrometry: A Paradigm Shift in Toxicology Screening Applications. Clin Biochem Rev. 40(3):2019–135.
18. Ens W, Standing KG. Hybrid Quadrupole/Time-of-Flight Mass Spectrometers for Analysis of Biomolecules. In 2005. p. 49–78.
19. Mellon FA. MASS SPECTROMETRY | Principles and Instrumentation. In: Encyclopedia of Food Sciences and Nutrition. Elsevier; 2003. p. 3739–49.
20. Ahuja S 1933 ; JND. Modern instrumental analysis. 6th ed. Vol. 47. 2006. 319–396 p.
21. van Agthoven MA, Lam YPY, O'Connor PB, Rolando C, Delsuc MA. Two-dimensional mass spectrometry: new perspectives for tandem mass spectrometry. European Biophysics Journal. 2019 Apr 13;48(3):213–29.

22. Wägele B, Witting M, Schmitt-Kopplin P, Suhre K. MassTRIX Reloaded: Combined Analysis and Visualization of Transcriptome and Metabolome Data. PLoS One. 2012 Jul 6;7(7):e39860.
23. Wang X, Shen S, Rasam SS, Qu J. MS1 ion current-based quantitative proteomics: A promising solution for reliable analysis of large biological cohorts. Mass Spectrom Rev. 2019 Nov 28;38(6):461–82.
24. Domingo-Almenara X, Rainer J, Vicini A, Salzer L, Stanstrup J, Badia JM, et al. A Modular and Expandable Ecosystem for Metabolomics Data Annotation in R. Metabolites 2022. 2022;12:173.
25. Tavassoly I, Goldfarb J, Iyengar R. Systems biology primer: the basic methods and approaches. Essays Biochem. 2018 Oct 26;62(4):487–500.
26. Voit EO. A First Course in Systems Biology. Second edition. | New York : Garland Science, 2017.: Garland Science; 2017.
27. Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021;19:3735–46.
28. Elrayess MA, Botrè F, Palermo A, Weckwerth W. Editorial: OMICS-Based Approaches in Sports Research. Article [Internet]. 2022;9:1. Available from: www.frontiersin.org
29. Dunn WB, Broadhurst DI, Atherton HJ, Goodacre R, Griffin JL. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. Chem Soc Rev. 2011;40(1):387–426.
30. Ren JL, Zhang AH, Kong L, Wang XJ. Advances in mass spectrometry-based metabolomics for investigation of metabolites. 2018;
31. Varghese RS, Zhou B, Nezami Ranjbar MR, Zhao Y, Ressom HW. Ion annotation-assisted analysis of LC-MS based metabolomic experiment [Internet]. 2012. Available from: http://www.proteomesci.com/content/10/S1/S8
32. Fiehn O. Metabolomics – the link between genotypes and phenotypes. Plant Mol Biol. 2002;48(1/2):155–71.
33. Grüning NM, Rinnerthaler M, Bluemlein K, Mülleder M, Wamelink MMC, Lehrach H, et al. Pyruvate Kinase Triggers a Metabolic Feedback Loop that Controls Redox Metabolism in Respiring Cells. Cell Metab. 2011 Sep;14(3):415–27.
34. Alseekh S, Aharoni A, Brotman Y, Contrepois K, D'auria J, Ewald J, et al. Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. Available from: https://doi.org/10.1038/s41592-021-01197-1
35. Castelli FA, Rosati G, Moguet C, Fuentes C, Marrugo-Ramírez J, Lefebvre T, et al. Metabolomics for personalized medicine: the input of analytical chemistry from biomarker discovery to point-of-care tests. Available from: https://doi.org/10.1007/s00216-021-03586-z
36. Yao L, Sheflin AM, Broeckling CD, Prenni JE. Data processing for GC-MS- and LC-MS-based untargeted metabolomics. In: Methods in Molecular Biology. Humana Press Inc.; 2019. p. 287–99.
37. Dunn WB, Bailey NJC, Johnson HE. Measuring the metabolome: current analytical technologies. Analyst. 2005;130(5):606.
38. Bauermeister A, Mannochio-Russo H, Costa-Lotufo L V., Jarmusch AK, Dorrestein PC. Mass spectrometry-based metabolomics in microbiome investigations. Nat Rev Microbiol. 2022 Mar 22;20(3):143–60.

39. Gowda GAN, Djukovic D. Overview of Mass Spectrometry-Based Metabolomics: Opportunities and Challenges. Methods Mol Biol. 2014;1198:3–12.
40. Bennett BD, Yuan J, Kimball EH, Rabinowitz JD. Absolute quantitation of intracellular metabolite concentrations by an isotope ratio-based approach. Nat Protoc. 2008 Aug 17;3(8):1299–311.
41. Newgard CB. Metabolomics and Metabolic Diseases: Where Do We Stand? Cell Metab. 2017 Jan;25(1):43–56.
42. Yang Q, Zhang A hua, Miao J hua, Sun H, Han Y, Yan G li, et al. Metabolomics biotechnology, applications, and future trends: a systematic review. RSC Adv. 2019;9(64):37245–57.
43. Salek RM, Steinbeck C, Viant MR, Goodacre R, Dunn WB. The role of reporting standards for metabolite annotation and identification in metabolomic studies. 2013;2:1. Available from: http://www.metabolomicsworkbench.org.
44. Feng X, Liu X, Luo Q, Liu BF. Mass spectrometry in systems biology: An overview. Mass Spectrom Rev. 2008 Nov;27(6):635–60.
45. Franceschi P, Mylonas R, Shahaf N, Scholz M, Arapitsas P, Masuero D, et al. BIOENGINEERING AND BIOTECHNOLOGY TECHNOLOGY REPORT MetaDB a data processing workflow in untargeted MS-based metabolomics experiments. 2014; Available from: http://www.isa-tools.org/format.html
46. Sumner LW, Alexander AE, Ae A, Ae DB, Ae MHB, Beger R, et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics [Internet]. 2007;3:211–21. Available from: http://msi-workgroups.sourceforge.net/http://www.metabolomicssociety.org/Reference:http://msi-workgroups.sourceforge.net/bio-metadata/reporting/pbc/http://msi-workgroups.sourceforge.net/chemical-analysis/
47. Sansone SA, Rocca-Serra P, Field D, Maguire E, Taylor C, Hofmann O, et al. Toward interoperable bioscience data. Nat Genet. 2012 Feb 27;44(2):121–6.
48. Chang HY, Colby SM, Du X, Gomez JD, Helf MJ, Kechris K, et al. A Practical Guide to Metabolomics Software Development. 2023;16:15. Available from: https://dx.doi.org/10.1021/acs.analchem.0c03581
49. Grace SC, Hudson DA. Processing and Visualization of Metabolomics Data Using R. In: Metabolomics - Fundamentals and Applications. InTech; 2016.
50. Dunn WB, Erban A, Weber RJM, Creek DJ, Brown M, Breitling R, et al. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. Metabolomics. 2013 Mar 26;9(S1):44–66.
51. Stanstrup J, Gerlich M, Dragsted LO, Neumann S. Metabolite profiling and beyond: approaches for the rapid processing and annotation of human blood serum mass spectrometry data. Anal Bioanal Chem. 2013 Jun 25;405(15):5037–48.
52. CREEK DJ, BARRETT MP. Determination of antiprotozoal drug mechanisms by metabolomics approaches. Parasitology. 2014 Jan 5;141(1):83–92.
53. Hendriks MMWB, Eeuwijk FA van, Jellema RH, Westerhuis JA, Reijmers TH, Hoefsloot HCJ, et al. Data-processing strategies for metabolomics studies. TrAC Trends in Analytical Chemistry. 2011 Nov;30(10):1685–98.

54. Van Mechelen I, Smilde AK. A generic linked-mode decomposition model for data fusion. Chemometrics and Intelligent Laboratory Systems. 2010 Nov;104(1):83–94.
55. Scholz M, Gatzek S, Sterling A, Fiehn O, Selbig J. Metabolite fingerprinting: detecting biological features by independent component analysis. Bioinformatics. 2004 Oct 12;20(15):2447–54.
56. Zou H, Hastie T, Tibshirani R. Sparse Principal Component Analysis. Journal of Computational and Graphical Statistics. 2006 Jun;15(2):265–86.
57. Chen C, Gonzalez FJ, Idle JR. LC-MS-Based Metabolomics in Drug Metabolism. Drug Metab Rev. 2007 Jan 9;39(2–3):581–97.
58. Jonsson P, Gullberg J, Nordström A, Kusano M, Kowalczyk M, Sjöström M, et al. A Strategy for Identifying Differences in Large Series of Metabolomic Samples Analyzed by GC/MS. Anal Chem. 2004 Mar 1;76(6):1738–45.
59. Laaniste A, Leito I, Kruve A. ESI outcompetes other ion sources in LC/MS trace analysis. Anal Bioanal Chem. 2019 Jun 26;411(16):3533–42.
60. Smith CA, Maille GO, Want EJ, Qin C, Trauger SA, Brandon TR, et al. METLIN. Ther Drug Monit. 2005 Dec;27(6):747–51.
61. Domingo-Almenara X, Montenegro-Burke JR, Benton HP, Siuzdak G. Annotation: A Computational Solution for Streamlining Metabolomics Analysis. Vol. 90, Analytical Chemistry. American Chemical Society; 2018. p. 480–9.
62. Patti GJ, Yanes O, Siuzdak G. Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol. 2012 Apr 22;13(4):263–9.
63. Domingo-Almenara X, Montenegro-Burke JR, Benton HP, Siuzdak G. Annotation: A Computational Solution for Streamlining Metabolomics Analysis. Vol. 90, Analytical Chemistry. American Chemical Society; 2018. p. 480–9.
64. Keller BO, Sui J, Young AB, Whittal RM. Interferences and contaminants encountered in modern mass spectrometry. Anal Chim Acta. 2008 Oct;627(1):71–81.
65. Lu W, Su X, Klein MS, Lewis IA, Fiehn O, Rabinowitz JD. Metabolite Measurement: Pitfalls to Avoid and Practices to Follow. Annu Rev Biochem. 2017 Jun 20;86(1):277–304.
66. Tohge T, Fernie AR. Combining genetic diversity, informatics and metabolomics to facilitate annotation of plant gene function. Nat Protoc. 2010 Jun 10;5(6):1210–27.
67. Schiffman C, Petrick L, Perttula K, Yano Y, Carlsson H, Whitehead T, et al. Filtering procedures for untargeted LC-MS metabolomics data. BMC Bioinformatics. 2019 Dec 14;20(1):334.
68. Beniddir MA, Kang K Bin, Genta-Jouve G, Huber F, Rogers S, van der Hooft JJJ. Advances in decomposing complex metabolite mixtures using substructure- and network-based computational metabolomics approaches. Nat Prod Rep. 2021;38(11):1967–93.
69. Kind T, Fiehn O. Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. BMC Bioinformatics. 2006 Dec 28;7(1):234.
70. Verhoeven HA, Ric de Vos CH, Bino RJ, Hall RD. Plant Metabolomics Strategies Based upon Quadrupole Time of Flight Mass Spectrometry (QTOF-MS). In: Plant Metabolomics. Berlin/Heidelberg: Springer-Verlag; p. 33–48.
71. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. Anal Chem. 2006 Feb 1;78(3):779–87.

72. Albóniga OE, González O, Alonso RM, Xu Y, Goodacre R. Optimization of XCMS parameters for LC–MS metabolomics: an assessment of automated versus manual tuning and its effect on the final results. Metabolomics. 2020 Jan 10;16(1):14.
73. Tautenhahn R, Böttcher C, Neumann S. Annotation of LC/ESI-MS Mass Signals. In: Bioinformatics Research and Development. Berlin, Heidelberg: Springer Berlin Heidelberg; p. 371–80.
74. Kuhl C, Tautenhahn R, Neumann S. LC-MS Peak Annotation and Identication with CAMERA. 2022.
75. Laaniste A, Leito I, Kruve A. ESI outcompetes other ion sources in LC/MS trace analysis. Anal Bioanal Chem. 2019 Jun 26;411(16):3533–42.
76. Senan O, Aguilar-Mogas A, Navarro M, Capellades J, Noon L, Burks D, et al. CliqueMS: a computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network. Available from: https://academic.oup.com/bioinformatics/article/35/20/4089/5418951
77. Broeckling CD, Afsar FA, Neumann S, Ben-Hur A, Prenni JE. RAMClust: A Novel Feature Clustering Method Enables Spectral-Matching-Based Annotation for Metabolomics Data. Anal Chem. 2014 Jul 15;86(14):6812–7.
78. Fakouri Baygi S, Kumar Y, Kumar Barupal D. IDSL.CSA: Composite Spectra Analysis for Chemical Annotation of Untargeted 1 Metabolomics Datasets 2. Available from: https://doi.org/10.1101/2023.02.09.527886
79. Mahieu NG, Spalding JL, Gelman SJ, Patti GJ. Defining and Detecting Complex Peak Relationships in Mass Spectral Data: The Mz.unity Algorithm. Anal Chem. 2016 Sep 20;88(18):9037–46.
80. Uppal K, Walker DI, Jones DP. xMSannotator: an R package for network-based annotation of high-resolution metabolomics data Graphical abstract HHS Public Access. Anal Chem [Internet]. 2017;89(2):1063–7. Available from: http://pubs.acs.org.
81. Uppal K, Soltow QA, Strobel FH, Pittard WS, Gernert KM, Yu T, et al. xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. BMC Bioinformatics. 2013 Dec 16;14(1):15.
82. Luan H, Jiang X, Ji F, Lan Z, Cai Z, Zhang W. CPVA: a web-based metabolomic tool for chromatographic peak visualization and annotation. Bioinformatics. 2020 Jun 1;36(12):3913–5.
83. Misra BB. New software tools, databases, and resources in metabolomics: updates from 2020. Metabolomics. 2021 May 11;17(5):49.
84. Xing S, Hu Y, Yin Z, Liu M, Tang X, Fang M, et al. Retrieving and Utilizing Hypothetical Neutral Losses from Tandem Mass Spectra for Spectral Similarity Analysis and Unknown Metabolite Annotation. Anal Chem. 2020 Nov 3;92(21):14476–83.
85. Liang D, Liu Q, Zhou K, Jia W, Xie G, Chen T. IP4M: an integrated platform for mass spectrometry-based metabolomics data mining. BMC Bioinformatics. 2020 Dec 7;21(1):444.
86. Elzbieta Lauzikaite. LC-MS data processing with massFlowR. 2020.
87. Zheng L, Chao F, Parthaláin N Mac, Zhang D, Shen Q. Feature grouping and selection: A graph-based approach. Inf Sci (N Y). 2021 Feb;546:1256–72.
88. Bobadilla-Suarez S, Ahlheim · C, Mehrotra · A, Panos · A, Love · B C. Measures of Neural Similarity. Available from: https://doi.org/10.1007/s42113-019-00068-5

89. Bauer C, Cramer R, Schuchhardt J. Evaluation of Peak-Picking Algorithms for Protein Mass Spectrometry. In 2011. p. 341–52.
90. Kirwan JA, Gika H, Beger RD, Bearden D, Dunn WB, Goodacre R, et al. Quality assurance and quality control reporting in untargeted metabolic phenotyping: mQACC recommendations for analytical quality management. Metabolomics. 2022 Aug 27;18(9):70.
91. Theodorsson E. Quality Assurance in Clinical Chemistry: A Touch of Statistics and A Lot of Common Sense. J Med Biochem. 2016 Apr 1;35(2):103–12.
92. Galli C, Plebani M. Quality controls for serology: an unfinished agenda. Clin Chem Lab Med [Internet]. 2020;58(8):1169–70. Available from: https://doi.org/10.1515/cclm-2020-0304
93. Tautenhahn R, Böttcher C, Neumann S. Highly sensitive feature detection for high resolution LC/MS. BMC Bioinformatics. 2008 Dec 28;9(1):504.

7. SUPPORTING MATERIAL

S1. Description of software employment

Here we detail the employment of CAMERA and CliqueMS R-packages. In both software we assigned default parameters for all the functions.

S1.1. CAMERA

Camera requires us to provide an input in the format of an xcmsSet object. It can be obtained after the processing of raw data files. In this project we have utilized XCMS but other options like MetAlign or MZMine could have been chosen.

S1.1.1. Pre-processing with XCMS

The following instructions show how to process raw data before CAMERA processing.

1. Open libraries (XCMS and BiocParallel), adjust raw file format (.mzML) and assign to `ourFiles` the path to find all raw data files in the computer.

```r
library(xcms)
library(BiocParallel)
register(bpstart(MulticoreParam(2)))

filePattern <- '.mzML'

ourFiles <- "TFG/Dataset exemple(CAMERA)/Immunometabolism(RAW DATA)/"
```

2. Denote file directory and parameter arrangement

```r
mzML <- dir(ourFiles, pattern = filePattern, full.names = TRUE,recursive = TRUE)
```

3. Generation of a dataframe

```r
pd <- data.frame(sample_name = sub(basename(mzML), pattern = filePattern,
replacement = "", fixed = TRUE), sample_group = c(rep("RAW", length(mzML))),
stringsAsFactors = FALSE)
```

4. Raw data reading

```r
raw_data <- readMSData(files = mzML, pdata = new("NAnnotatedDataFrame", pd), mode =
"onDisk")
```

5. Peak inspection instructions

```r
rtr <- c(461, 500)
mzr <- c(100, 130.1)
```

6. Suitable function to extract the chromatogram and plot generation.

```r
chr_raw <- chromatogram(raw_data, mz = mzr, rt = rtr)
plot(chr_raw)
```

7. The first function consists of adjust parameters from *centWave* algorithm which performs peak density and weavelet based chromatographic peak detection for high resolution LC/MS data in centroid mode[3]. The following instructions are used to first align the detected features across different samples and then adjust retention time (RT) by correcting small variations in RT values that can occur due to differences in instrument performance or sample preparation.

```{r}
cwp <- CentWaveParam(ppm=10, peakwidth = c(2, 20), mzdiff=0.01, noise = 100)
xdata <- findChromPeaks(raw_data, param = cwp)

xdata <- adjustRtime(xdata, param = ObiwarpParam(binSize = 0.6))

save(xdata, file='xdata.rda')

load('xdata.rda')
```

8. In the first place, the *PeakDensityParam* permits the specification of all settings for the peak grouping. The *groupChromPeaks* function then executes the correspondence among overlapping MS data slices in the m/z dimension based on the density distribution of the recognized chromatographic peaks in the slice along the time axis. The final result involves clustering peaks based on their m/z and retention time values. (71)

```{r}
pdp <- PeakDensityParam(sampleGroups = xdata$sample_group, minFraction = 0.4, bw =
3)
xdata <- groupChromPeaks(xdata, param = pdp)
```

9. The first function is used to fill in the missing peaks and integrate signals in the m/z-RT area of a feature (chromatographic peak group) for samples in which no chromatographic peak for this feature has been found. Later, the file with the completion of raw data processing is stored as xdata_fillpeaks.rda

   The *featureValues* is an optional function that extracts a feature values matrix with rows denoting features and columns denoting samples. In this function, the parameter value enables the user to specify which column from the *chromPeaks* matrix to be returned.

```{r}
xdata <- fillChromPeaks(xdata)

save(xdata, file='xdata_fillpeaks.rda')

featureValues(xdata, value="into")

head(featureValues(xdata, value = "into"))

write.csv(xdata, file='xcmsresults.csv')
```

S1.1.2. Data processing with CAMERA.

In the following instructions we show how to process multiple samples with the R-package CAMERA.

[3] MS data collected off an instrument can be presented as either profile or centroid mode. In centroid mode, signals are shown as zero line-width discrete m/z signals. Because there is less information characterizing a sign, centroid data has the Benefit of having a file size that is substantially smaller.

1. Xdata is the resulting file from preprocessing raw data with XCMS (xdata_fillpeaks.rda) and has the format of a XCMSnExp. The generation of a CAMERA object is accomplished through the use of the *xsAnnotate* function. In order to determine which peaks are derived from the same molecule, the *groupFWHM* function is used to group the peaks after the retention time, every peak that falls into a defined window are considered the same *pcgroup*. CAMERA uses the peak group verification process to determine the exact mass of the peaks and to annotate the ion species, accomplished through the use of the *groupCorr* function, which calculates the Pearson correlation coefficient based on the peak shapes of every peak in the pseudospectrum to separate co-eluted substances, in short, performs peak shape correlation.

   Both methods divide the peaks into different groups, which are referred to as "pseudospectra." The size of these pseudospectra ranges from one to one hundred ions, depending on the number of molecules and their ionizability. Afterwards the *findIsotope* function permits the annotation of possible isotopes.

```{r}
xset <- as(xdata, 'xcmsSet')

xsa <- xsAnnotate(xset)

xsaF <- groupFWHM(xsa, perfwhm = 0.6)

xsaC <- groupCorr(xsaF)

xsaFI <- findIsotopes(xsaC)
```

2. Here the rules parameter is modified to provide the same rule table of adducts as in CliqueMS

```{r}
rules <- read.csv("TFG/ref.adducts.+.edit.csv", sep=";")
```

3. The *findAdducts* function generates a matrix and performs the annotation of adduct peaks and calculates hypothetical masses for the group.

```{r}
xsaFA <- findAdducts(xsaFI, polarity = "positive", rules = rules)
```

4. This workflow results in a peaklist containing all information from an xsAnnotate object and it is a data-frame file that can be stored in a .csv format.

```{r}
write.csv (getPeaklist (xsaFA), file="result_CAMERA.csv")
```

5. A graphical representation of the annotation result can be achieved in CAMERA with the plotPsSpectrum which plots the spectrum of a pseudospectrum, with labeling the most intense peaks. The following instructions will plot the spectrum of pseudospectrum 2082 and highlight the annotation and m/z labels of the 10 strongest peaks.

```{r}
plotPsSpectrum(xsaFA, pspec=2082, maxlabel=5)
```

S1.2. CliqueMS

The 'cliqueMS' algorithm initially separates features in the dataset into distinct groups. It accomplishes this by computing a similarity weighted network from the data and subsequently searching for clique groups. These cliques are fully connected components that exhibit greater similarity in inner edges than edges located outside of cliques. Therefore, the computation of clique groups, isotopic annotation is carried out. After completing isotopic annotation, adducts are annotated within each group.

1. Open library (cliqueMS) and mention the full path of the file directory. CliqueMS also preprocess raw data through xcms. As in CAMERA, *CentWaveParam* facilitates the identification of chromatographic peaks in high-resolution LC/MS in centroid mode by employing peak density and wavelet-based methods. This is accomplished by enabling the specification of all the necessary settings for chromatographic peak detection. The approach utilizes two distinct techniques for chromatographic peak identification, namely density-based detection of m/z regions of interest and a Continuous Wavelet Transform (CWT)-based method for peak resolution. (93) Within the xcms software, the 'centWave' algorithm serves as the peak detection method that is applied to establish the intensity vector corresponding to each feature. Upon specifying the relevant parameters, the *findChromPeaks* function executes the chromatographic peak detection process on LC/GC-MS data as part of the updated xcms user interface. The raw data is designated as the object, while the *CentWaveParam* parameters, previously defined, are utilized in the process.

```
library(cliqueMS)
mzfile <- "~/TFG/muestras
(cliqueMS)/samples/05082019_Exp40_HILICpos_Sample1_60_01_1414_0h1.mzML"
library(xcms)
mzraw <- readMSData(files = mzfile, mode = "onDisk")
cpw <- CentWaveParam(ppm = 10, peakwidth = c(2,20), mzdiff = 0.01,noise=100)
mzData <- findChromPeaks(object = mzraw, param = cpw)
```

2. The *createanClique* function generates an *anClique* object using processed m/z data, which facilitates the identification of isotope and adduct annotations for features within each group.

```
```{r}
ex.anClique <- createanClique(mzData)
show(ex.anClique)
```
```

3. Feature grouping is achieved by utilizing a cosine metric to construct a similarity network where features represent nodes, and edges represent the cosine similarity between these features, and it is useful to discriminate pairs of features that come from the same metabolite from pairs of features that come from different metabolites. The cosine similarity is computed using the profile mode of data, having each feature a m/z value and vector intensities. The clique groups, which are fully connected components with high similarity in inner edges and lower similarity in edges outside the clique, are then identified within this network. The **cliqueMS** function operates under the assumption that the similarity score for all features originating from the same metabolite is greater than 0. Furthermore, it posits that the similarity values between features belonging to the same metabolite tend to be higher than those between features originating from different metabolites. Utilizing this information, **cliqueMS** employs a probabilistic model to identify

groups of features. Additionally, nodes are subsequently moved to different groups until the groups with the maximum log-likelihood are determined. The resulting *anClique* object includes the computed clique groups and adds a *cliqueGroup* column to the peaklist. Ideally, each group should include all features produced by a single metabolite. Notably, this process occurs before any annotation has taken place, and at this point, and isotopes and adducts are yet to be annotated.

Note that the *getCliques* function is utilized to ensure reproducibility when generating variables that rely on random values. By incorporating the *set*.seed() function, the same random values are consistently produced each time the code is executed.

```
```{r}
set.seed(2)
ex.cliqueGroups <- getCliques(mzData, filter = TRUE)
show(ex.cliqueGroups)
```
```

Table S1 *– getCliques main parameters*

Parameter	Value	Default	Usage
filter	TRUE/FALSE	TRUE	If «TRUE», filter features that have cosine similarity > 0.99 and equal m/z, retention time and intensity value
mzerror	numeric	$5*10^{-6}$	If m/z relative error is below this value features are considered with the same m/z value
intdiff	numeric	$1*10^{-4}$	If intensity relative error is below this value features are considered with the same m/z value
rtdiff	numeric	$1*10^{-4}$	If retention time relative error is below this value features are considered with the same m/z value
tol	numeric	$1*10^{-5}$	Minimum relative increase in log-likelihood to do a new round of log-likelihood maximisation

4. Once clique groups have been computed, the subsequent task is to obtain the isotope annotation. A specific intensity pattern must be observed among features corresponding to isotopic variants of the same metabolite. Specifically, the monoisotopic feature should exhibit greater abundance than its isotope, while the isotope feature should possess a mass value higher than the monoisotopic feature within the relative error range specified by the user. This function performs the annotation of features that are carbon isotopes based on their m/z and intensity data. Isotopes are annotated within each clique group, with the aid of the CliqueMS function. This function searches for pairs of features that can be regarded as carbon isotopes, based on two rules: the monoisotopic feature should be more intense than the isotopic feature, and that the mass difference between the features should correspond to the difference of a carbon isotope, as determined by a ppm relative error.

Isotopes are annotated using the *getIsotopes* function, which identifies pairs of features satisfying the isotope conditions, and subsequently generates the isotope annotation. Any incoherencies, such as two monoisotopic masses for one isotope, or two second isotopes for one first isotope, are removed, with the less similar pair being discarded. Finally, the isotope annotations for all groupings are combined and stored in the 'anClique-class' object as the final step.

```{r}
ex.Isotopes <- getIsotopes(ex.cliqueGroups, ppm = 10)
show(ex.Isotopes)
```

***Table S2** - getIsotopes main parameters*

Parameter	Value	Default	Usage
maxCharge	numeric	3	Maximum charge considered when comparing pairs of features
maxGrade	numeric	2	Maximum number of isotopes in an isotope cluster, without counting the monoisotopic mass
ppm	numeric	10	Relative error in ppm to consider that two features have the mass difference of an isotope
isom	numeric	1,003355	Mass difference of an isotope

The final step of CliqueMS is to perform adduct annotation. A feature is characterized by a specific m/z value, which represents the mass of the metabolite and the ion adduct (or fragmented ion adduct. Although the neutral mass is unknown, the ion adduct mass is recognizable since several ion adducts are already known.

5. The list of possible adducts has to be given as an input by the user or either use one of the default adduct lists. In this essay the reference or default adduct table from positive ionization mode has been chosen. This command show how to visualize it in a summarized manner.

```{r}
data(positive.adinfo)
head(positive.adinfo)
```

***Table S3** - Firsts rows from the utilized reference adduct data frame.*

	adduct <fctr>	log10freq <dbl>	massdiff <dbl>	nummol <int>	charge <int>
1	[M+2H-NH3]2+	-3.512904	-15.012016600	1	2
2	[Cat]3+	-3.512904	-0.001645737	1	3
3	[Cat]2+	-3.512904	-0.001040400	1	2
4	[Cat+H]2+	-3.336813	1.006178842	1	2
5	[M+2H]2+	-1.813934	2.014552000	1	2
6	[M+H+Na]2+	-2.699991	23.996494000	1	2

The column labeled "adduct" contains a string that describes the adduct. The "log10freq" column indicates the log10 frequency of observation of the adduct or the log-score linked to it, which will be used to calculate the annotation score. The "massdiff" column shows the mass associated with the adduct or in-source fraction. The "nummol" column has a value of 1 if the adduct requires only one molecule of metabolite to form, 2 for dimerization, 3 for trimerization, and so on. The "charge" column represents the charge of the adduct.

6. In each clique, the function *getAnnotation* annotates pre-defined adducts and in-source fragments and obtains molecular neutral masses of metabolites (mass in the final output). To accomplish it, CliqueMS looks for clusters of two or more features that match a neutral mass and two or more adducts from the provided adduct list. Any neutral mass that have only one adduct are not considered for scoring. After obtaining all potential neutral masses and their associated adducts, the algorithm tests various combinations of neutral masses

and adducts to determine the most likely annotation. The selected masses have the highest frequencies and the greatest number of adducts.

Each combination is assigned a score, and the top five annotations with the highest scores are reported. The scoring algorithm considers the log-frequency of the adduct, along with the minimum number of empty features and neutral masses.
The computed score, which is a logarithmic score, is the sum of the adducts log-frequencies plus the number of empty features (with a log-frequency smaller than the least frequent adduct) and the number of neutral/parental masses. For larger clique groups, it is possible that the annotation of some features is independent of others due to a lack of a neutral mass with adducts in both feature groups. In such cases, the clique group is divided into non-overlapping annotation groups. The reported scores correspond to these annotation groups.

The score is designed to compare the likelihood of annotations within a specific annotation group, from the first to the fifth. The utilized score is normalized and scaled and their values range from 0 (minimum score, which represent features with empty annotations), up to 100 (is the score value of the theoretical maximum annotation which are all the adducts of the list with the minimum number of neutral masses). However, it is not suitable for comparing annotations across different groups because a larger number of features within a group will result in a smaller score. (76)

```
ex.Adducts <- getAnnotation(ex.Isotopes, ppm = 10,
    adinfo = positive.adinfo, polarity = "positive",
    normalizeScore = TRUE)
show(ex.Adducts)
```

***Table S4** - getAnnotation main parameters.*

Parameter	Value	Default	Usage
polarity	«positive/negative»		Polarity of the adducts
ppm	numeric	10	Relative error in ppm under which we consider two or more features compatible with a neutral mass and two or more adducts from the adduct list
emptyS	numeric	$1*10^{-6}$	Score given to non annotated features, used to compute the group score

7. To get the list of annotated adducts from clique groups the algorithm *getlistofCliques* has to be supplied according to the following indication. A comparable function is *getPeaklistanClique* where an anClique-class object has to be provided to get a list of features with their current annotation.

```
```{r}
features.clique6 <- getlistofCliques(ex.Adducts)[[6]]

head(getPeaklistanClique(ex.Adducts)[features.clique6,
    c("an1","mass1","an2", "mass2", "an3", "mass3", "an4", "mass4", "an5",
    "mass5")], n = 10)
```
```

The final annotations are stored in a data frame like the one shown and explained in *Result obtention* below.

S2. Result obtention

S2.1. CAMERA

The result of processing data in RStudio as it is explained in the previous part is a huge data frame containing 13.082 rows of detected features and 42 columns where a large amount of information is classified.

...1	mz	mzmin	mzmax	rt	rtmin	rtmax	npeaks	RAW	ms_level	X1	X2	X3	X4	X5	X6	X7	X8
1	37.90938	37.73908	37.98278	332.430969	329.899994	339.99609	19	15	1	NA	312.9760	341.1788	434.3043	NA	4.092480e+02	4.065856e+02	NA
2	39.71818	39.64413	39.85352	336.458130	331.415985	339.99609	17	13	1	5.528855e+02	360.1480	512.4768	NA	278.4908	4.909821e+02	2.962670e+02	NA
3	43.62904	43.49005	43.73467	332.428009	323.342926	341.02097	27	17	1	NA	369.7892	NA	289.2686	NA	4.122096e+02	NA	2.957376e+02
4	49.88838	49.74541	49.95584	332.433044	330.422607	337.47525	21	17	1	4.425342e+02	447.9449	356.5890	401.3432	293.5212	4.840547e+02	3.968670e+02	NA
5	79.07618	79.07598	79.07663	574.605316	572.569763	579.14148	18	15	1	8.056000e+02	316.3681	530.2880	883.2167	680.3697	2.345675e+02	9.046828e+01	NA
6	80.94774	80.94751	80.94786	396.756210	390.450531	403.07196	28	20	1	4.449132e+04	38234.5560	33739.5866	28672.0628	25873.6318	1.558558e+04	2.363400e+03	5.172072e+04
7	80.94781	80.94754	80.94812	427.032516	421.224426	434.35654	26	18	1	3.097576e+03	3574.6313	12842.2922	11534.6192	312.6472	9.581614e+03	1.758975e+03	3.450917e+03
8	80.94762	80.94727	80.94804	410.638306	403.567963	418.20731	25	13	1	2.255991e+03	3494.8853	4524.5224	1029.3583	283.0873	2.334069e+03	1.338276e+03	3.183355e+02
9	80.94778	80.94723	80.94804	367.745804	362.689514	374.31912	20	12	1	2.449142e+03	3394.9946	1902.6899	421.3391	1276.3111	2.904786e+03	1.469099e+03	8.689421e+02
10	82.02171	81.93724	82.02260	782.985504	771.894836	796.10474	26	14	1	3.191576e+03	1196.4718	166.8704	1090.5846	666.9139	2.803333e+03	4.665765e+02	1.016547e+03
11	82.94484	82.94457	82.94515	382.384094	374.816467	389.44269	47	24	1	7.661316e+03	8366.9607	7026.1856	6225.1010	6524.0352	8.029972e+03	4.125829e+03	1.192454e+04
12	82.94475	82.94426	82.94526	366.489563	361.687500	373.82028	42	25	1	9.733127e+02	4631.0494	1237.0995	3711.2800	1268.0862	4.130328e+03	1.404511e+03	4.247683e+03
13	82.94483	82.94441	82.94516	411.149139	404.578003	418.71423	39	25	1	3.856720e+03	6601.3600	6815.0187	6303.4697	2741.9613	4.003272e+03	1.080855e+03	4.240134e+03
14	82.94481	82.94461	82.94495	396.759323	390.446000	403.07196	34	28	1	4.625454e+04	48487.0700	48624.4494	42061.6167	50288.6160	4.387418e+04	4.682140e+04	2.200215e+04
15	82.94493	82.94453	83.06061	429.819946	420.212189	439.91100	71	29	1	3.870524e+03	2225.1044	9689.8442	9647.9024	5476.7752	5.492412e+03	6.363836e+03	5.962605e+03
16	84.08074	84.08056	84.08106	655.809021	653.294495	658.33405	29	29	1	2.823175e+04	28118.7039	29283.7509	32881.2133	14522.8132	1.993972e+04	1.576617e+04	1.116629e+04
17	84.04440	84.04427	84.08081	493.365000	487.318085	499.93948	31	26	1	1.791187e+05	61489.0764	85847.6611	221.7249	390696.3485	2.681343e+05	3.559335e+05	1.904891e+05
18	84.04446	84.04423	84.08111	540.792236	527.675842	554.92468	31	25	1	1.179738e+05	111672.1650	121142.6792	118688.3474	117614.9217	1.107941e+05	2.221646e+05	2.319311e+05
19	84.04438	84.04421	84.04455	203.796982	187.147766	217.88351	28	12	1	2.665289e+03	10246.3376	8455.3586	8314.1584	672.5165	7.268551e+04	NA	3.867222e+03
20	85.02868	85.02802	85.04788	492.366943	482.266235	499.93170	63	22	1	1.385049e+03	1857.6020	1093.9674	401.3940	9910.4213	4.165142e+03	3.552690e+03	4.246086e+04
21	85.04741	85.02799	85.04851	540.286316	531.708679	548.37341	38	13	1	4.659553e+03	6095.5913	398.1294	1245.7817	1218.9244	4.595664e+03	5.882208e+03	4.838877e+03
22	85.08398	85.08335	85.08471	656.571136	651.791199	662.36517	26	18	1	1.330429e+02	1343.1135	662.7903	3048.9370	1432.8890	9.420485e+02	1.445433e+03	7.929985e+02
23	85.03832	85.02817	85.04812	505.494385	500.937073	518.12201	28	14	1	4.935259e+03	11628.6350	24133.5783	22197.2230	NA	NA	5.708090e+03	4.151490e+03
24	86.09642	86.09630	86.09660	418.196808	412.645935	423.24182	29	29	1	4.649662e+05	420929.3852	438078.9160	471763.6256	416693.7565	4.226912e+05	4.375109e+05	4.302458e+05
25	86.05999	86.04874	86.09652	501.943665	491.342000	507.00256	27	21	1	6.751313e+03	5530.3631	4554.3860	837.8257	22844.3926	3.676262e+03	5.387206e+03	5.740237e+03
26	86.09640	86.06026	86.09651	407.098999	398.027954	410.12872	24	22	1	3.402380e+05	319316.8815	339592.4591	350294.4865	297774.9143	3.331443e+05	3.372226e+05	1.413455e+05
27	87.05526	87.02590	87.10003	432.316498	424.760803	448.48364	45	27	1	1.143926e+03	2037.2089	3790.0250	1679.7324	2804.4399	1.458199e+03	1.725448e+03	1.572296e+03
28	87.05532	87.05513	87.05583	558.962524	556.939148	573.10461	35	29	1	7.013017e+04	73246.5322	78741.7999	85561.1486	18765.9741	2.089094e+04	2.129539e+04	1.853599e+04
29	87.09965	87.04382	87.09988	406.595184	401.042450	410.12872	27	26	1	2.724920e+04	27819.7401	28705.8001	26236.1394	25258.5629	2.788639e+04	2.834504e+04	2.496584e+04
30	87.04408	87.04329	87.04428	394.735525	389.956116	397.52045	26	26	1	2.621382e+03	2173.8218	2613.9669	3170.1662	6894.9176	2.784548e+03	1.350703e+04	1.345603e+03

X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25
3.793500e+02	NA	4.599520e+02	4.482525e+02	NA	NA	NA	NA	NA	4.006800e+02	788.1099	NA	NA	NA	NA	NA	5.880140e+02
NA	NA	NA	4.448554e+02	NA	NA	316.8858	304.3332	NA	NA	530.5587	NA	NA	NA	3.143502e+02	NA	NA
4.324875e+02	NA	NA	6.695945e+02	NA	341.3434	279.4862	652.5360	3.561660e+02	NA	NA	657.6547	NA	NA	4.672360e+02	4.692240e+02	9.146573e+02
2.770008e+02	NA	NA	4.086088e+02	4.206752e+02	NA	NA	541.3600	NA	NA	NA	271.3672	NA	NA	3.986400e+02	3.600048e+02	NA
NA	NA	4.179542e+02	3.184038e+03	1.072618e+02	1155.0015	166.8517	843.9573	6.127595e+02	1.755193e+02	685.8254	844.4120	3.642887e+02	539.3691	4.105977e+02	3.261192e+02	6.825751e+01
3.764633e+04	3.535545e+04	2.377521e+04	8.349545e+03	2.528632e+04	53437.4585	37156.3689	8057.0216	2.279802e+04	1.468865e+04	13712.7041	18271.5811	2.261229e+04	14794.5821	4.256806e+04	2.424933e+04	3.897127e+04
2.893372e+03	3.949836e+03	3.326501e+03	2.212623e+03	4.763659e+03	3540.5777	3880.6613	1385.1932	1.562700e+03	3.038460e+03	7104.3977	3790.1786	3.168837e+03	3008.4915	6.052865e+03	3.568182e+03	4.776701e+03
7.749102e+02	4.117433e+03	1.899549e+03	6.824063e+04	1.502071e+03	3194.3520	2469.0280	194.3701	2.244596e+03	6.113364e+02	1605.4465	8596.9500	2.770290e+02	1174.7022	2.880415e+03	5.876982e+02	1.755954e+03
1.215960e+03	7.113590e+02	1.005849e+03	3.999531e+03	3.702141e+03	1881.3577	1638.9756	1884.2911	1.163991e+03	1.147227e+03	1314.3098	1543.4199	1.295869e+03	1058.3770	1.153163e+03	4.401505e+03	1.597276e+03
1.298464e+03	9.954590e+02	3.279775e+02	9.210615e+02	2.259246e+03	370.5760	234.9807	425.6897	9.158223e+02	9.921056e+02	1250.1500	796.1464	1.322788e+03	922.1094	1.034301e+03	1.206601e+03	3.495030e+02
5.683392e+03	5.232600e+03	7.020407e+03	1.155743e+04	6.461707e+03	8809.0605	15960.4302	14824.1388	7.175910e+03	2.267691e+04	28381.2536	25583.1450	7.859941e+03	7124.8290	7.504916e+03	2.606887e+03	5.128305e+03
1.581788e+03	2.124252e+03	1.917998e+03	5.959830e+03	1.741659e+03	1213.1942	2024.4090	5231.5311	1.793634e+03	1.335979e+03	637.3500	6947.2871	4.186955e+03	4897.3438	6.556347e+03	1.165733e+03	3.115637e+03
2.870023e+03	4.002087e+03	2.647190e+03	8.050801e+04	3.310081e+03	11234.1072	3915.1914	1698.5893	3.926967e+03	5.663511e+03	3985.8675	4451.3024	1.211968e+04	4104.3687	1.000405e+04	1.231298e+03	1.688400e+03
4.311876e+04	5.008731e+04	4.299758e+04	1.730405e+04	7.213324e+03	40811.8492	25482.9504	32984.6998	6.074960e+04	5.041817e+04	61771.3927	42937.5792	5.505118e+04	36885.7677	4.159810e+04	3.910655e+04	3.037103e+04
3.753903e+03	4.472201e+03	5.362617e+03	9.277110e+03	1.044056e+04	5640.6040	10463.9191	2528.7173	1.485574e+04	4.044105e+03	9918.7552	12131.4866	4.431423e+03	4637.3040	2.767443e+03	2.369210e+03	2.784144e+03
1.779807e+04	2.476125e+04	1.864146e+04	6.496203e+04	2.558311e+04	28172.9071	25730.8927	14191.0050	2.158195e+04	2.353462e+04	32743.3128	29700.0803	2.668631e+04	22613.0755	1.495111e+04	1.077511e+04	1.678051e+04
4.500473e+05	3.513363e+05	4.159164e+05	6.891715e+03	9.342842e+04	118834.6432	114451.6900	335204.8213	7.398496e+04	6.078779e+04	NA	157619.6610	1.698083e+05	141442.0951	3.723597e+05	2.926228e+05	3.298492e+05
2.160968e+05	2.761548e+05	2.368634e+05	1.196966e+04	4.255983e+02	1696.6525	1345.0038	120147.0567	1.108641e+03	1.930622e+03	15038.4450	7462.0978	9.657981e+03	5706.9040	1.216139e+05	6.417815e+04	7.876500e+04
3.807280e+03	4.395123e+03	4.378537e+03	2.315723e+04	4.639453e+03	3011.6602	21340.2163	3409.1022	1.438172e+04	6.885675e+03	16020.7355	7271.3585	4.862933e+04	5402.0126	8.859014e+02	8.729769e+03	4.245266e+03
4.212923e+04	NA	2.392707e+03	NA	5.146335e+03	2295.1157	1691.5770	6576.8981	4.621123e+03	1.134591e+04	535.6407	5242.9465	4.882888e+03	6343.7678	3.664355e+03	8.788300e+03	8.225891e+03
5.006099e+03	3.806102e+03	4.270275e+03	1.555552e+02	NA	NA	NA	1545.5518	NA	NA	NA	NA	NA	NA	1.225228e+03	4.397218e+03	3.691658e+03
8.853633e+02	1.644357e+03	6.557010e+02	2.411166e+03	2.750254e+02	976.5800	2164.9886	357.0597	5.821282e+02	9.929111e+02	4361.2575	1623.3071	7.488262e+02	517.9732	3.170942e+02	NA	5.833947e+02
1.395403e+03	4.722909e+03	1.639901e+04	1.491318e+04	NA	NA	NA	NA	NA	NA	1545.5691	NA	NA	NA	NA	2.396515e+03	2.430885e+03
4.156650e+05	4.764533e+05	4.250257e+05	5.686700e+05	4.165514e+05	468372.3430	453365.5998	384869.6570	3.685689e+05	4.144840e+05	438200.5503	427116.6307	3.897379e+05	408746.2888	3.908026e+05	3.234904e+05	3.388152e+05
1.601806e+02	1.356588e+03	NA	NA	2.293036e+02	3160.3545	998.6400	15999.6403	8.156393e+02	9.251482e+02	NA	534.1158	6.361640e+02	NA	1.801784e+04	1.601886e+04	1.905371e+04
9.670694e+04	3.545516e+05	3.281210e+05	4.030632e+02	2.112216e+05	308273.6794	4822.4180	59356.6413	2.627349e+05	2.710802e+05	285.8300	305549.3034	2.552127e+05	246064.3701	2.957501e+05	2.423132e+05	2.603941e+05
9.351750e+02	1.086231e+03	1.213586e+03	6.246154e+04	1.412598e+03	2201.0051	1034.5667	2122.2347	1.351159e+03	3.346471e+02	41808.6053	1147.3430	1.133107e+03	552.7493	1.529746e+03	6.370179e+02	1.898587e+03
1.920441e+04	2.098079e+04	1.934418e+04	1.863603e+04	8.988893e+03	10099.0587	8617.8625	13070.4945	7.182124e+03	2.386134e+04	38368.4462	23248.0080	2.437705e+04	20231.7867	1.502119e+04	1.470913e+04	1.793123e+04
2.771093e+04	2.823437e+04	2.953485e+04	3.688868e+03	2.295919e+04	27281.1228	17191.4652	20955.4802	2.257446e+04	2.131918e+04	820.5424	24891.7378	2.390011e+04	23816.7340	2.704299e+04	2.101558e+04	1.868905e+04
1.449663e+03	1.354671e+03	1.737630e+03	NA	2.256073e+03	2347.1175	2171.7069	6338.6873	1.713531e+03	6.182980e+02	NA	893.4695	5.277429e+02	740.0951	6.608415e+03	5.624883e+03	6.496480e+03

X26	X27	X28	X29	isotopes	adduct	pcgroup
5.935610e+02	5.068224e+02	4.258800e+02	5.194268e+02	NA	NA	82
3.395612e+02	3.864312e+02	NA	NA	NA	NA	44
5.326220e+02	4.639998e+02	4.737472e+02	NA	NA	NA	82
7.038500e+02	1.024119e+03	4.554334e+02	NA	NA	NA	82
2.799482e+02	NA	NA	3.329676e+02	NA	NA	4593
8.991786e+03	1.856660e+04	1.532124e+04	1.070963e+04	NA	NA	93
2.596822e+03	3.472368e+03	6.406678e+03	3.306659e+03	NA	NA	3696
7.690317e+03	8.657845e+02	2.466382e+03	2.924501e+02	NA	NA	70
4.185124e+03	2.032933e+03	1.866394e+03	5.973628e+02	NA	NA	159
7.350242e+02	4.292708e+03	6.368318e+02	2.415298e+02	NA	[M+Na]+ 59.0328 [M+H]+ 81.0148	312
4.170814e+03	6.408531e+03	5.663220e+03	1.585055e+04	NA	NA	283
1.606656e+03	7.954296e+02	5.054429e+03	2.038685e+03	NA	[Cat-H]+ 83.9605	159
6.801330e+03	1.389735e+04	3.055106e+03	6.574595e+03	NA	NA	70
4.665661e+04	2.319197e+04	3.458143e+04	3.454454e+04	NA	[M+2Na]2+ 119.911	93
3.914416e+03	6.033363e+03	4.756752e+03	3.540016e+03	NA	NA	1977
1.703976e+04	2.936962e+04	1.384320e+04	1.314940e+04	NA	[M+H]+ 83.0735	58
2.550920e+05	2.617428e+05	1.924462e+05	2.914313e+05	NA	[Cat+H]2+ 167.088	7
8.197559e+04	1.305702e+05	8.400764e+04	1.074931e+05	[1][M]+	[M+H]+ 83.0372	1
6.546670e+03	1.432336e+04	9.488881e+03	8.613773e+03	NA	NA	69
4.309387e+03	7.749369e+03	1.894353e+04	1.804163e+04	NA	NA	843
5.152950e+03	1.716947e+03	4.613591e+03	2.182034e+03	[1][M+1]+	NA	1
1.394624e+03	2.350574e+03	5.828453e+02	1.103852e+03	NA	NA	1819
NA	3.298409e+03	NA	NA	NA	NA	864
3.982583e+05	5.087912e+05	4.407296e+05	4.426769e+05	NA	[Cat+H2O]+ 68.0848 [Cat]+ 86.0955	24
5.949167e+02	2.548820e+02	1.085641e+03	4.121357e+02	NA	NA	3172
3.070598e+05	3.686044e+05	1.297659e+05	3.490006e+05	NA	NA	33
1.268568e+03	1.166504e+03	4.713182e+02	1.469101e+03	NA	NA	72
2.257603e+04	2.042583e+04	1.879004e+04	2.075628e+04	NA	[M+Na]+ 64.0599	46
1.448560e+04	2.989839e+04	2.816524e+04	2.769228e+04	NA	NA	33
2.837630e+03	2.013940e+03	2.457257e+03	1.892614e+03	NA	[Cat]3+ 261.131	106

Figure S1 *- Part of CAMERA final output shown in RStudio. It shows the first 30 rows.*

Regarding to the information that the data frame provides in the first place the most relevant information are mz and rt values. After that the column *ms_level* show the MS level according to each feature. The values from the X1 column until the X29 represent the concentration of the samples from the Quality control 1 to sample number 25. The subsequent column contains the annotated isotope for a monoisotopic peak, the values between the first square brackets indicates the isotope-group-id and the second one the isotope annotation, after that the charge of the isotope is denoted. After that, the column *adduct* displays the annotation proposition for the ion species, while the value within the brackets represents the approximated molecular weight and must correspond to the exact mass of the metabolite. Lastly, the "pcgroup" column shows the outcome of the peak correlation-based annotation that is independent of isotope and adduct annotation. Peaks categorized under the same group are supposed to be part of the same spectrum and therefore, linked to the same metabolite.

S2.2. CliqueMS

The process which is performed for obtaining results in CliqueMS is quite similar to the previous one. First of all it is needed to comment that from the processing of the samples in RStudio, the result is output based on a data frame for every one of the samples formed by between 20.000 and 35.000 rows each, which represent detected features. Additionally, from the 4 quality controls and the 25 samples for which the raw data was provided, samples 11, 18, 22 and 24 had processing mistakes because of their dimensions and they were discarded.

05082019_Exp40_HILICpos_QC4_85_01_1429_QC4

	mz	mzmin	mzmax	rt	rtmin	rtmax	into	intb	maxo	sn	sample	cliqueGroup	isotope	mass1	an1	score1	mass2	an2	score2	mass3	an3	score3	mass4	an4	score4	mass5	an5	score5
CP00001	132.986767780612	132.986543680298	132.987092899322	1.461	958	2.974	371.448	366.939490909091	290	17	1	2	M0	NA	NA	0	NA	NA	0	NA	NA	0	NA	NA	0	NA	NA	0
CP00002	178.113888527929	178.113082292676	178.114693502617	1.461	958	3.987	495.2415	488.770454545455	348	18	1	2	M0	356.2288	[Cat]2+	59.8745	NA	NA	0	NA	NA	0	NA	NA	0	NA	NA	0
CP00003	423.180620494662	423.179843395047	423.182629876117	1.461	958	2.974	548.352	546.336	586	585	1	2	M0	356.2288	[M-2H+3Na]+	59.8745	NA	NA	0	NA	NA	0	NA	NA	0	NA	NA	0
CP00004	467.203678162566	467.201852001025	467.205054598847	1.461	958	2.974	434.952	432.936	416	415	1	2	M0	233.0986	[2M+H]+	66.1281	NA	NA	0	NA	NA	0	NA	NA	0	NA	NA	0
CP00005	238.076697797764	238.07565153349	238.078048913426	1.461	958	3.481	577.767	575.244	436	435	1	2	M0	233.0986	[M+Na-H2O]+	66.1281	NA	NA	0	NA	NA	0	NA	NA	0	NA	NA	0
CP00006	400.286213655834	400.285446947723	400.287212383029	1.461	958	3.481	583.3176	580.7946	500	499	1	2	M0	NA	NA	0	NA	NA	0	NA	NA	0	NA	NA	0	NA	NA	0
CP00007	709.291542963298	709.288571335983	709.293709136739	1.461	958	3.481	507.6276	505.1046	450	449	1	2	M0	NA	NA	0	NA	NA	0	NA	NA	0	NA	NA	0	NA	NA	0
CP00478	301.205129234951	301.203994843009	301.205849866942	2.468	1.967	2.974	874.076	872.5655	619	1087	1	2	M0	NA	NA	0	NA	NA	0	NA	NA	0	NA	NA	0	NA	NA	0
CP00008	236.074633027765	236.073693029036	236.075853148444	1.967	958	6.01	911.3808	898.704872727273	312	17	1	7646	M0	235.0674	[M+H]+	69.0115	117.5337	[2M+H]+	48.0534	253.0779	[M+H-H2O]+	47.7279	213.0854	[M+Na]+	44.1285	191.1035	[M-H+2Na]+	39.5721
CP00041	192.944094772895	192.943179661141	192.944534686303	3.987	958	10.535	2906.36747368421	2896.28642105263	442	441	1	7646	M0	NA		73.2245	169.9549	[M+Na]+	58.6875	84.9778	[2M+Na]+	52.8334	84.9778	[2M+Na]+	34.2369	84.9778	[2M+Na]+	32.1627
CP00044	208.917807883091	208.916686653619	208.91841711609	2.468	958	9.021	4172.0985625	4163.531625	1190	1189	1	7646	M0	207.9114	[M+H]+	73.2245	169.9549	[M+K]+	58.6875	84.9778	[2M+K]+	52.8334	225.9211	[M+H-H2O]+	34.2369	185.9286	[M+Na]+	32.1627
CP00047	208.91863807569	208.917920045593	208.91945221154	1.967	958	7.514	3806.51446153846	3799.45415384615	1700	1699	1	7646	M0	207.9114	[M+H]+	73.2245	169.9549	[M+K]+	58.6875	84.9778	[2M+K]+	52.8334	225.9211	[M+H-H2O]+	34.2369	185.9286	[M+Na]+	32.1627
CP00049	216.020544316146	216.019484424027	216.021293655379	2.974	958	8.015	1796.51057142857	1788.9495	309	308	1	7646	M0	215.0133	[M+H]+	104.3576	233.0238	[M+H-H2O]+	95.6424	193.0313	[M+Na]+	91.715	107.5066	[2M+H]+	86.9679	232.0398	[M+H-NH3]+	80.3281
CP00051	216.020962689253	216.020324719901	216.022605722044	2.974	958	8.015	1882.20271428571	1874.64164285714	414	413	1	7646	M0	215.0133	[M+H]+	104.3576	233.0238	[M+H-H2O]+	95.6424	193.0313	[M+Na]+	91.715	107.5066	[2M+H]+	86.9679	232.0398	[M+H-NH3]+	80.3281
CP00055	226.091825492011	226.090684580536	226.09306965859	1.967	958	6.01	2078.3928	2072.8356	1076	1075	1	7646	M0	203.1026	[M+Na]+	78.4887	NA	NA	0	NA	NA	0	NA	NA	0	NA	NA	0
CP00056	301.141684923925	301.141192086274	301.142231800462	2.468	958	6.01	1727.2788	1721.7216	742	741	1	7646	M0	NA		67.2138	278.1536	[M+Na]+	58.9764	NA	NA	0	NA	NA	0	NA	NA	0
CP00064	442.299718864376	442.298311719934	442.301082336362	3.987	958	12.057	3323.1415	3311.538	806	805	1	7646	M0	419.3163	[M+Na]+	30.6817	419.3163	[M+Na]+	29.2224	419.3163	[M+Na]+	27.6546	419.3163	[M+Na]+	27.1447	419.3163	[M+Na]+	24.9953
CP00066	574.176531930942	574.175218032926	574.177573168747	2.468	958	9.021	2396.72675	2388.1598125	1068	1067	1	7646	M1 [7]	NA		30.6817	NA		29.2224	NA		27.6546	NA		27.1447	NA		24.9953
CP00069	147.015994169739	147.015245979292	147.016709903447	2.468	958	7.514	1870.98153846154	1863.92123076923	594	593	1	7646	M0	146.0071	[M+H]+	48.7705	128.9805	[M+NH4]+	44.4279	108.0537	[M+K]+	34.8211	NA	NA	0	NA	NA	0
CP00077	601.204197233791	601.202896164367	601.204981434599	1.967	958	6.01	1830.8448	1825.2876	1000	999	1	7646	M0	NA	NA	0	NA	NA	0	NA	NA	0	NA	NA	0	NA	NA	0
CP00080	242.065280288778	242.064476561314	242.066189663697	3.987	958	11.043	3197.9535	3187.36425	586	585	1	7646	M0	203.1026	[M+K]+	78.4887	NA	NA	0	NA	NA	0	NA	NA	0	NA	NA	0
CP00083	258.056211667559	258.055903906716	258.056766787829	2.468	958	10.03	3552.696	3543.12	826	825	1	7646	M0	235.0674	[M+Na]+	69.0115	117.5337	[2M+Na]+	48.0534	253.0779	[M+Na-H2O]+	47.7279	213.0854	[M-H+2Na]+	44.1285	191.1035	[M-2H+3Na]+	39.5721
CP00085	274.030333206662	274.029969845711	274.030734764306	2.468	958	12.562	4312.14730434783	4300.0387826087	926	925	1	7646	M0	235.0674	[M+K]+	69.0115	117.5337	[2M+K]+	48.0534	253.0779	[M+K-H2O]+	47.7279	NA		44.1285	NA		39.5721
CP00088	339.110745619822	339.110211310805	339.111195317408	1.967	958	7.514	12492.2058461538	12462.9376615385	5062	165	1	7646	M0	355.13	[M+H-NH3]+	33.7222	632.2538	[M+2Na]2+	17.6138	632.2538	[M+2Na]2+	11.1454	NA	NA	0	NA	NA	0
CP00090	708.283710205131	708.282390701532	708.285574456442	1.967	958	10.535	3354.97431578947	3344.8932631579	1374	1373	1	7646	M1 [6]	342.1467	[2M+Na]+	83.223	684.2934	[M+Na]+	60.3717	684.2934	[M+Na]+	40.9273	NA	NA	0	NA	NA	0
CP00092	131.042908293802	131.042649409523	131.043387561066	2.468	958	6.512	3339.97363636364	3333.91472727273	950	949	1	7646	M0	NA		48.7705	NA		44.4279	108.0537	[M+Na]+	34.8211	NA	NA	0	NA	NA	0
CP00096	438.056168510899	438.055694193947	438.057482695151	2.468	958	7.514	3783.31630769231	3776.256	1312	1311	1	7646	M0	NA		33.7222	399.0923	[M+K]+	17.6138	NA		11.1454	NA	NA	0	NA	NA	0
CP00099	681.260455295822	681.257754197551	681.262082759044	1.967	958	7.514	3254.80184615385	3247.74153846154	1450	1449	1	7646	M2 [4]	NA		69.0115	NA		48.0534	NA		47.7279	NA		44.1285	NA		39.5721
CP00103	262.07625881836	262.075683558269	262.076945288534	2.468	958	10.03	3542.616	3519.6336	1010	57	1	7646	M0	NA		30.6817	NA		29.2224	NA		27.6546	195.1232	[M-2H+3Na]+	27.1447	NA		24.9953
CP00105	273.109930336641	273.109405246735	273.110148953828	2.468	958	9.527	5905.04911764706	5879.64455294118	1712	82	1	7646	M0	NA		51.7775	NA		40.6491	134.066	[2M+Na-H2O]+	27.3818	NA	NA	0	NA	NA	0
CP00107	422.08156577221	422.080781831337	422.082385532641	2.468	958	6.512	3032.484	3026.42509090909	1344	1343	1	7646	M0	355.13	[M-2H+3Na]+	33.7222	399.0923	[M+Na]+	17.6138	345.1607	[M-H+2K]+	11.1454	NA	NA	0	NA	NA	0
CP00110	755.412159193826	755.40990382727	755.414298483272	2.468	958	7.514	2993.57046153846	2934.52061538462	1366	38	1	7646	M0	NA		63.3103	NA		57.6731	688.4569	[M-2H+3Na]+	41.05	NA	NA	0	NA	NA	0
CP00114	129.987791582211	129.987340567468	129.988162576862	2.974	958	6.512	2575.03636363636	2568.97745454545	580	579	1	7646	M0	146.0071	[M+H-NH3]+	48.7705	128.9805	[M+H]+	44.4279	106.9986	[M+Na]+	34.8211	NA	NA	0	NA	NA	0
CP00116	145.961196829357	145.960463963132	145.961780163923	4.998	958	9.527	2639.75605882353	2630.683	430	429	1	7646	M0	NA		48.7705	NA		44.4279	106.9986	[M+K]+	34.8211	NA	NA	0	NA	NA	0
CP00120	653.22658649929	653.223279392308	653.227651129221	1.967	958	6.512	3266.76181818182	3260.70290909091	1770	1769	1	7646	M2 [3]	650.2129	[M+H]+	56.373	NA		43.8533	NA	NA	0	NA	NA	0	NA	NA	0
CP00122	707.281390455969	707.280278581752	707.282670370889	1.967	958	6.512	5132.90581818182	5126.84690909091	2908	2907	1	7646	M0 [6]	342.1467	[2M+Na]+	83.223	684.2934	[M+Na]+	60.3717	684.2934	[M+Na]+	40.9273	NA	NA	0	NA	NA	0

Figure S2 *- Example of part of the CliqueMS final output from quality control 4.*

In this data frame, the columns are as follows, using our example: The initial 11 columns relate to data extracted directly from the 'xcms- Set' object. These columns are labeled as "mz", "mzmin", "mzmax", "rt", "rt", "rtmin", "rtmax", "into", "intb", "maxo", "sn", and "sample". The remaining columns, starting from column 12 ("cliqueGroup") and extending to the final column, correspond to the output of 'cliqueMS' and contain the following information. The initial column, "cliqueGroup", corresponds to the clique group of the feature. The column "isotope" show the isotopic state of the feature. If a feature is a monoisotopic mass, it will be labeled as

"M0". Nonetheless, if a feature is an isotope of another feature, the column will include both the isotope number and the isotope cluster. Features that belong to the same isotope cluster are isotopes of the same feature. Therefore, if an entry in the isotope column reads "Mi [X]", it indicates that the feature is in the isotope cluster X.
Moreover, the columns labeled as "massA anA scoreA" (where A = 1,..,5) correspond to adduct annotations in the top 5 annotations. For each annotation A, the "massA" column displays the neutral molecular mass of the feature, the "anA" column represents the adduct annotation of the feature, and the "scoreA" column displays the logarithmic score of the clique group or the score of the subdivision of the clique group. It's crucial to note that the annotation scores are dependent on the size of the clique/group of features. Therefore, annotation scores for different groups of features cannot be compared. The mass columns signify the molecular mass when matched with the reference mass.

At present, we have acquired the neutral mass and adduct annotation for our features. By utilizing the neutral mass, in combination with the retention time and MS/MS data, we may confidently annotate some of these metabolites. Additionally, we are now aware of the quantity of features in the dataset that are isotopes. Ultimately, we have successfully reduced the complexity of our dataset, progressing from numerous features to a significantly smaller number of annotated neutral masses possessing distinct adducts and isotopes. (76)

S3. CAMERA and CliqueMS data filtering

The first step to obtain any further results is to filter and extract the relevant information from the whole data frame obtained after data processing according to our metabolite reference ID data base. In order to obtain filtered results and for being able to get meaningful information from the different data frames obtained after processing the samples through CAMERA and Clique, respectively, the following script was created to obtain a list which matches our 44 reference metabolites contained in a file named ID_RAW_Immunometabolism with different features of a certain output.

The following script was exactly utilized in both software unless only the CAMERA example is presented. Nonetheless, as with CliqueMS we obtained multiple outputs from CliqueMS, the script was performed 25 times compare to the only one run in CAMERA. Firstly, we assign ID to the file which contains the 44 reference metabolites and their exact mass value and retention time. Afterwards, the file with processed data information is read. The name ppm.calc2 designs the following function to calculate the parts per million error.

$$ppm.difference = \frac{|Mass.Reference - Mass.Feature|}{Mass.Reference} \cdot 10^6$$

Then, a for loop was created inside of another one. The first operation results in adding to the exact mass value of the metabolite the mass of a proton (1,007242) as the samples were ionized in the positive mode. This reference numbers were one of the variables (Ref) in the ppm.calc2 function which also include the m/z values (Acc) from the processed data in study. Finally, 10 units was the value assigned to the ppm error.

```
library(openxlsx)
read.xlsx("TFG/CAMERA/ID_RAW_Immunometabolism.xlsx")
ID <- read.xlsx("TFG/CAMERA/ID_RAW_Immunometabolism.xlsx",rowNames =FALSE,rows = c(1:45))

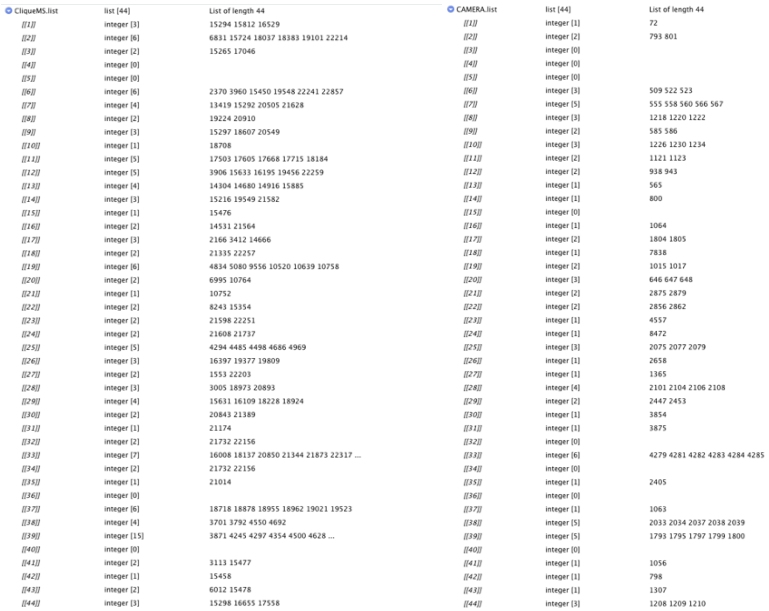
cameraResult <- read.csv("TFG/CAMERA/result_CAMERA.csv", stringsAsFactors = FALSE,
row.names = 1)

ppm.calc2 <- function(Acc, Ref){(abs(Ref - Acc) * 10 ^ 6) / Ref}

CAMERA.list <- NULL
for(r in 1:nrow(ID)) {
  RefMassProt <- 1.007242+ID[r,4]

  final.res <- NULL
  for(i in cameraResult$mz){
  result<- ppm.calc2(Acc=i, Ref = RefMassProt)
  final.res<- c(final.res,result)
  }
  CAMERA.list[[r]] <- which(final.res <= 10)
}
```

The output that was received after processing the instructions above is a list which matches each metabolite to none, one or more than one peaks from the CAMERA or CliqueMS result data frame, each peak corresponds to a particular m/z value and it is associate to a *pcgroup* in CAMERA or *CliqueGroup* in CliqueMS which are ideally related to a particular metabolite. (Figure S3)



CliqueMS.list	list [44]	List of length 44
[[1]]	integer [3]	15294 15812 16529
[[2]]	integer [6]	6831 15724 18037 18383 19101 22214
[[3]]	integer [2]	15265 17046
[[4]]	integer [0]	
[[5]]	integer [0]	
[[6]]	integer [6]	2370 3960 15450 19548 22241 22857
[[7]]	integer [4]	13419 15292 20505 21628
[[8]]	integer [2]	19224 20910
[[9]]	integer [3]	15297 18607 20549
[[10]]	integer [1]	18708
[[11]]	integer [5]	17503 17605 17668 17715 18184
[[12]]	integer [5]	3906 15633 16195 19456 22259
[[13]]	integer [4]	14304 14680 14916 15885
[[14]]	integer [3]	15216 19549 21582
[[15]]	integer [1]	15476
[[16]]	integer [2]	14531 21564
[[17]]	integer [3]	2166 3412 14666
[[18]]	integer [2]	21335 22257
[[19]]	integer [6]	4834 5080 9556 10520 10639 10758
[[20]]	integer [2]	6995 10764
[[21]]	integer [1]	10752
[[22]]	integer [2]	8243 15354
[[23]]	integer [2]	21598 22251
[[24]]	integer [2]	21608 21737
[[25]]	integer [5]	4294 4485 4498 4686 4969
[[26]]	integer [3]	16397 19377 19809
[[27]]	integer [2]	1553 22203
[[28]]	integer [3]	3005 18973 20893
[[29]]	integer [4]	15631 16109 18228 18924
[[30]]	integer [2]	20843 21389
[[31]]	integer [1]	21174
[[32]]	integer [2]	21732 22156
[[33]]	integer [7]	16008 18137 20850 21344 21873 22317 ...
[[34]]	integer [2]	21732 22156
[[35]]	integer [1]	21014
[[36]]	integer [0]	
[[37]]	integer [6]	18718 18878 18955 18962 19021 19523
[[38]]	integer [4]	3701 3792 4550 4692
[[39]]	integer [15]	3871 4245 4297 4354 4500 4628 ...
[[40]]	integer [0]	
[[41]]	integer [2]	3113 15477
[[42]]	integer [1]	15458
[[43]]	integer [2]	6012 15478
[[44]]	integer [3]	15298 16655 17558

CAMERA.list	list [44]	List of length 44
[[1]]	integer [1]	72
[[2]]	integer [2]	793 801
[[3]]	integer [0]	
[[4]]	integer [0]	
[[5]]	integer [0]	
[[6]]	integer [3]	509 522 523
[[7]]	integer [5]	555 558 560 566 567
[[8]]	integer [3]	1218 1220 1222
[[9]]	integer [2]	585 586
[[10]]	integer [3]	1226 1230 1234
[[11]]	integer [2]	1121 1123
[[12]]	integer [2]	938 943
[[13]]	integer [1]	565
[[14]]	integer [1]	800
[[15]]	integer [0]	
[[16]]	integer [1]	1064
[[17]]	integer [2]	1804 1805
[[18]]	integer [1]	7838
[[19]]	integer [2]	1015 1017
[[20]]	integer [3]	646 647 648
[[21]]	integer [2]	2875 2879
[[22]]	integer [2]	2856 2862
[[23]]	integer [1]	4557
[[24]]	integer [1]	8472
[[25]]	integer [3]	2075 2077 2079
[[26]]	integer [1]	2658
[[27]]	integer [1]	1365
[[28]]	integer [4]	2101 2104 2106 2108
[[29]]	integer [2]	2447 2453
[[30]]	integer [1]	3854
[[31]]	integer [1]	3875
[[32]]	integer [0]	
[[33]]	integer [6]	4279 4281 4282 4283 4284 4285
[[34]]	integer [0]	
[[35]]	integer [1]	2405
[[36]]	integer [0]	
[[37]]	integer [1]	1063
[[38]]	integer [5]	2033 2034 2037 2038 2039
[[39]]	integer [5]	1793 1795 1797 1799 1800
[[40]]	integer [0]	
[[41]]	integer [1]	1056
[[42]]	integer [1]	798
[[43]]	integer [1]	1307
[[44]]	integer [3]	1208 1209 1210

Figure S3 *- Example of the list from the 10th CliqueMS sample (left) and CAMERA list (right).*

Afterwards, in case that there is more than one putative *pcgroups* or *CliqueGroups* for one metabolite it is necessary to decide which one is the most appropriate. The amount of ions annotated (isotopes and adducts) and the matching of the adduct to the metabolite neutral mass are the two main parameters to assign the group to the metabolite.

In order to not to leave any feature behind, a mass search for each metabolite is executed through R program. Features are analyzed searching the values in the third column of the list

(see Figure S3) which represent the row in the output file of the certain sample, from now on a CliqueMS example is shown.

```
cliqueMSRes <- select(cliqueMSResult,mz,rt,isotope,cliqueGroup,mass1,an1,m
ass2,an2,mass3,an3,mass4,an4,mass5,an5)

CliqueMS.list[[1]]

a<-cliqueMSRes[c(3542, 15211, 15888, 16398),]
```

The first instruction in the above script shows how to create a reduced data frame from the main one (cliqueMSResult). The second function allows us to select the features for the first metabolite which is D-Alanine, the output are the values: 3542, 25211, 15888, 16398. And the final instruction provides the generation of the next dataframe.

Table S5 *- Summarized table for the first metabolite for sample 10.*

	mz	rt	isotope	cliqueGroup	mass1	an1	mass2	an2	mass3	an3	mass4	an4	mass5	an5
CP03654	204.0610	59.508	M0	4616	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CP15007	291.1610	435.288	M1 [2199]	14763	267.1671	[M+Na]+	267.1671	[M+Na]+	267.1671	[M+Na]+	267.1671	[M+Na]+	267.1671	[M+Na]+
CP15669	417.2140	441.336	M0	15594	NA		434.2133	[M+H-H2O]+	NA	NA	NA	NA	NA	NA
CP16145	263.1294	456.986	M0	16160	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Afterwards *cliqueGroups*, which group peaks which are highly correlated, are confirmed to be related to the metabolite in study (D-Alanine). In the previous created data frame (cliqueMSRes) a search of the putative *cliqueGroups* will be performed. The term putative refers to the candidates of *cliqueGroups* which have annotations for the features, in this present example comprise 14763 and 15594.

```
df<- cliqueMSRes[cliqueMSRes$cliqueGroup==14763,]

df<- cliqueMSRes[cliqueMSRes$cliqueGroup==15594,]
```

Both data frames are simplified with the above instruction for each putative *CliqueGroup* and finally the molecular mass for D-alanine (89,047679) is searched in those dataframes to see if the features match the M values of the adducts to D-Alanine. In that case for sample 10 neither of the *cliqueGroups* contained the mass value of Alanine and as a result it can be affirmed that this metabolite is not annotated for CliqueMS in this sample. In other examples, the *cliqueGroup* with higher score value is the chosen to be matched to the metabolite. In CAMERA we only get a possible annotation for each feature, and the neutral mass for the metabolite which we have to look for, appears next to the adduct chemical formula.

S4. Pseudospectra obtention

The obtention of a graphical representation of the filtered results was achieved through the mass spectrums generation for the formed groups for certain metabolites. Pseudospectrums were obtained according to the following indications. The objective is to create a graph which represents the correlation between the m/z values and the relative intensity of the identified peaks based on the pc group in CAMERA or *cliqueGroup* in CliqueMS.

```
View(result_CAMERA)

Met <- subset(result_CAMERA, pcgroup == 22)

plot(x = Met1$mz, y = Met1$X12/max(Met1$X12), type = 'h', main= "Sample12-Arginina-CAMERA", xlab="m/z",
ylab="relative intensity" )
```

S5. Peak correlation (CAMERA)

The following script and functions show how to get the peak intensity correlation in CAMERA peak intensity processed data according to *pcgroups*.

The correlation function was applied and the parameter *pairwise*.complete.obs was chosen to ignore NA (not available intensity values). It meals that the correlation coefficient is calculated using only the pairwise complete observations. For each pair of variables being correlated, any pair of observations with missing values for either feature was excluded from the computation. After that, ggplot2 and reshape2 libraries were charged and the function melt transforms the wide-format data frame into a long-format data frame in order to obtain variables and values. Finally a default heatmap was created and drown and in the end values were added to each square correlating the two variable values, we choose the colour gradient in the end.

```
camera.res <- read.csv("TFG/CAMERA/result_CAMERA.csv")
camera.pseudo <- subset(camera.res, pcgroup == 24)
cor1 <- cor(t(camera.pseudo[,11:39]), use = "pairwise.complete.obs")
head(colnames(cor1))
colnames(cor1) <- paste(camera.pseudo$X, camera.pseudo$adduct)
rownames(cor1) <- paste(camera.pseudo$X, camera.pseudo$adduct)

data1 <- melt(cor1)

ggp <- ggplot(data1, aes(Var1, Var2)) + # Create default ggplot2 heatmap
 geom_tile(aes(fill = value))

ggp +
 theme(axis.text.x = element_text(size = 5,angle = 90, vjust = 0.5, hjust=1),
 axis.text.y = element_text(size=5),
 axis.title.x = element_blank(),
 axis.title.y = element_blank(), legend.title = element_blank()) +
 scale_fill_gradient2(low = "#075AFF",
 mid = "#FFFFCC",
 high = "#FF0000")
```

S6. Validation through comparison to reference spectrums

This script indicates how it is possible to compare the experimentally obtained pseudospectra (CAMERA and CliqueMS) to the reference NIST data base spectrums.

In the first place CAMERA and CliqueMS are charged as well as NIST data base. After that, we provided the function that allows the graph generation which is presented below. It is important to note that we imposed a ppm error of less than 50 to permit the alignment between peaks from the pseudospectrums with the reference spectrums.

```r
```{r}

##Charge nist data base and CAMERA and CliqueMS results
load("~/nist.filt.rda")
camera.res <- read.csv("TFG/CAMERA/result_CAMERA.csv")

#Graph generating function
normalize <- osd::normalize
ppm.calc2 <- function(Acc, Ref){(abs(Ref - Acc) * 10 ^ 6) / Ref}

plotspec <- function(pseudo.mat, db = db.filt, nistID, isolated = FALSE, fullSpectrum = TRUE){
  nist.spectra <- as.data.frame(do.call(rbind, sapply(strsplit(db.filt[[nistID]]$Spectra, " "),
function(x) data.frame(strsplit(x, ":")))))

  empMSMS <- as.numeric(nist.spectra[,1])
  empMSMSInt <- as.numeric(nist.spectra[,2])

  toShow <- 1:nrow(pseudo.mat)
  maxMZ <- max(empMSMS)+5
  #if(isolated) toShow <- unlist(sapply(empMSMS, function(x) which(abs(x-pseudo.mat[,1])<0.01)))
  if(isolated) toShow <- unlist(sapply(empMSMS, function(x) which(ppm.calc2(Ref = x,Acc =
pseudo.mat[,1])<50)))
  if(fullSpectrum) maxMZ <- max(pseudo.mat[,1])

  plot(pseudo.mat[toShow,1],normalize(pseudo.mat[toShow,2]), type='h',
       ylim=c(-1,1), xlim=c(min(min(pseudo.mat[,1]), min(empMSMS)), maxMZ),
       xlab = "M/Z", ylab = "Intensity",
       lwd=2, main = paste(db.filt[[nistID]]$Name))
  #text(maxMZ, -1, paste("Reference", db.filt[[nistID]]$CE), cex = 1, pos = 2)
  lines(empMSMS ,-1*normalize(empMSMSInt),type='h', lwd=2, col='orange3')
}
```

For each metabolite, a reference spectrum had to be selected by indicating the ID number of each metabolite in the NIST reference database. Next, we show how to subset by the corresponding *pcgroup* for each metabolite. In this example, we show the case of arginine which corresponds to *pcgroup22*. For each metabolite, the lowest collision energy value (approximately 10eV) found in the database is selected. The following script shows to do so.

```r
#L-Arginine
##We save in a data.frame pseudospectrums from pcgroups
camera.pseudo <- subset(camera.res, pcgroup == 22)

subset(db.info.filt, Name == "L-Arginine")
grepl("10", subset(db.info.filt, Name == "L-Arginine")$CE)
subset(db.info.filt, Name == "L-Arginine")[grepl("10eV", subset(db.info.filt, Name ==
"L-Arginine")$CE),]
nistID <- which(db.info.filt$nid == 14680)

#Creatine
grepl("10", subset(db.info.filt, Name == "Creatine")$CE)
subset(db.info.filt, Name == "Creatine")[grepl("10eV", subset(db.info.filt, Name ==
"Creatine")$CE),]
nistID <- which(db.info.filt$nid == 52863)

#L-Glutamine
grepl("10", subset(db.info.filt, Name == "L-Glutamine")$CE)
subset(db.info.filt, Name == "L-Glutamine")[grepl("10eV", subset(db.info.filt, Name ==
"L-Glutamine")$CE),]
nistID <- which(db.info.filt$nid == 14226)

#L-Histidine
grepl("10", subset(db.info.filt, Name == "L-Histidine")$CE)
subset(db.info.filt, Name == "L-Histidine")[grepl("10eV", subset(db.info.filt, Name ==
"L-Histidine")$CE),]
nistID <- which(db.info.filt$nid == 14819)
```
```

```
#L-Lysine
grepl("10", subset(db.info.filt, Name == "L-Lysine")$CE)
subset(db.info.filt, Name == "L-Lysine")[grepl("10eV", subset(db.info.filt, Name ==
"L-Lysine")$CE),]
nistID <- which(db.info.filt$nid == 31656)

#L-Phenylalanine
grepl("10", subset(db.info.filt, Name == "L-Phenylalanine")$CE)
subset(db.info.filt, Name == "L-Phenylalanine")
subset(db.info.filt, Name == "L-Phenylalanine")[grepl("11eV", subset(db.info.filt, Name ==
"L-Phenylalanine")$CE),]
nistID <- which(db.info.filt$nid == 31609)

#L-Carnitine
grepl("10", subset(db.info.filt, Name == "L-Carnitine")$CE)
subset(db.info.filt, Name == "L-Carnitine")
subset(db.info.filt, Name == "L-Carnitine")[grepl("12eV", subset(db.info.filt, Name ==
"L-Carnitine")$CE),]
nistID <- which(db.info.filt$nid == 48418)
```

Afterwards, it is shown how to select the column of m/z and intensity respectively. The last part of the script shows how to obtain the representation of the reference spectrum from the NIST database compared to the pseudospectrum obtained from our data, in this case from CAMERA.

```
#_____________________________________________PLOTS

#Two column dataframe: 1-mz and 2-Intensity from one sample
pseudo.mat <- camera.pseudo[,c("mz", "X1")]

plotspec(pseudo.mat = pseudo.mat, #Pseudospectrum we want to compare
         db = db.filt, #NIST data base
         nistID = nistID, #Raw position in db.info.filt reference spectrum
         fullSpectrum = TRUE)
```

For CliqueMS, the procedure is exactly the same but the columns that are selected when representing the pseudospectrum are changed, as well as the directory of the file containing the processed data.