**Overview**

This program creates graphic representations of mass spectrometry data, which include statistic tests to conclude over the significance of differences interpreted visually. It provides two **quality control graphics** to verify stability of two analytic parameters, and graphics with the **statistic comparisons of isotopologue fraction in relation to different administration methods** **of the tracer**.

**Author**

Name: Ana Díaz García-Guijas

Email: [ana.diaz-garcia-guijas@univ-tlse3.fr](mailto:ana.diaz-garcia-guijas@univ-tlse3.fr)

**Dates:**

Created: June 2025

Last update: 17.06.2025

**Program version**

Requires RStudio 4.5.0

**Dependencies:**

This program requires the following R packages:

* ‘tidyverse’
* ‘ggsignif’
* ´mgcv´

To install required packages:

install.packages(c("tidyverse", "ggsignif", "mgcv"))

**Objective**

This program was developed as part of a study on the reprogramming of metabolism during aging in the African Turquoise Killifish (*Nothobranchius furzeri*). It assists in the analysis and interpretation of mass spectrometry data obtained from fluxomic experiments using stable isotope tracing. This technique relies on characterizing the fate of a [U-13C] tracer administered to the model organism, with a HPLC-HRMS method (High-Performance Liquid Chromatography – High-Resolution Mass Spectrometry). The objective of the program is to **automate data processing**, to reduce human errors and save time.

**Program functions:**

The program provides tools to:

* Identify **drift in two parameters** of the analytic method (retention time and exact mass)
* **Compare isotopologue percentages** across different **administration methods** as well as across multiple **injections**
* Characterise **intra-** and **inter- experimental** variations
* **Detect potential** **errors in the experimental conditions** using a color-coded representation of studied injections
* Determine **statistical significance of the quantity of isotope tracing** in the tissue following tracer administration via **different methods**

**Usage**

Working directory: when the program is opened from the R project directory the **working directory is automatically set** to the project folder, no additional configuration is needed. If running the script outside RStudio or from a compressed .zip file, make sure to **first extract the project** and **manually set the working directory** to the project folder before executing the script.

Interactions with user

RStudio doesn’t include a graphic interface, so user interaction is handled through **command-line prompts**. At the beginning of the script the user is prompted to paste the names of the documents to analyse. They are also asked to select the metabolites to study, the files shown in the graphics and to correct the injection date if samples of the same injection contain different injection dates (explained in more detail in the “Comments” section).

Annotations

The program contains explanatory annotations to facilitate understanding of the functions being applied.

Graph export

All graphs are **automatically saved** in folders created during the script’s execution.

*Note: references like (1.1) of (2.3.1) correspond to number sections and subsections within the R script*

**Functioning**

The program is structured in **three parts**: the first one imports the document in a database and transforms data, the second one generates the analytic controls graphs and the third one the statistic comparisons graphs.

*#1. Import document and create database with pertinent modifications*

In the first part the user is prompted to **enter the name of the .tsv document** to charge. *(1.1)* The file is read and organised in a database. Some variables are not explicit to the document; they are exported from sample names. *(1.2)* The user selects the metabolites to study and the documents to analyse by a prompt (*1.3* and *1.4*). The necessary **directories are automatically created**, and all the graphs will be saved here in .png format *(1.6).*

*#2. Analytic controls*

In the second part the exact mass graph and the retention time graph are structured with a **for loop**. For each metabolite studied the **data is reduced** to the concerning samples *(2.1.1 /2.2.1).*  Two mathematical methods are applied: **linear method** and **GAM** *(2.1.2 / 2.2.2)* and the graph is generated with ggplot. The functions used mainly are **“geom\_point()”** and **“geom\_smooth()” *(****2.1.4/2.2.4*. and *2.1.5./2.2.5.*) Finally, the graphs are saved in the corresponding directories *(2.1.6).*

*#3. Statistical comparisons*

In the third part *(3)* a database is created with the selected information *(3.1),* the **mean** and **variance** are calculated (*3.3)*, as well as the **biological replica** *(3.4)*, the **t.test** is used to compare conditions *(3.5)* and the graph is generated with **“geom\_bar()”** *(3.6).*

**Input**

Documents ***.tsv*** obtained with a Q Exactive Orbitrap Mass Spectrometers (MS) after data treatment using **skyline**:

* For fluxomic data,
  + after adapting the format of the output of skyline to IsoCor using Skyline2Isocor tool
  + Followed by correction of natural abundance and purity of the injected/infused U-13C labelled nutrient tracer using IsoCor

Input examples:

« Raw\_20250221\_titration\_20250127\_plasma\_Flux\_wo\_sky2iso.tsv » « Raw\_28032025\_titration2\_IP\_13C\_Glc\_20250303\_20250312\_plasma\_Flux\_sky2iso\_res.tsv »

**Output**

Three plots are generated:

* Exact mass or retention time variation graphs:

**Due to the unavailability of experimental data, the program simulates data in this section. For now, this part serves only to demonstrate the application of the tool. To interpret real data, some minor modifications would be necessary, as described in the "Adjustments" section.**

Plot showing exact mass/retention time variation over injection number for all isotopologues of a given metabolite (red linear regression line and yellow GAM smoother). Two graphs are generated: one including all isotopologues combined, and a second one organised in facets.

**To interpret the mathematical models results:** a **low R²** value and a **non-significant p-value** indicate that there is no meaningful linear drift in exact mass/time retention across the injection sequence. This suggests that the **analytical conditions remained stable** throughout the experiment.

* Isotopologue fraction comparaison bar plot:

Bar plot with jittered points showing isotopologue fraction (%) of a metabolite in a tissu compared between different administration condition of the tracer

Includes error bars (SD) and significance brackets between conditions based on t-tests.

Directories:

All graphic outputs are systematically saved with a .png format within a directory named “**Plasma**”. Exact mass variation figures are stored in the “Analytical control > Exact mass variation” subdirectory, retention time variation figures in “Analytical control > Retention time variation,” and isotopologue fraction comparison plots in “Comparisions > Young (12 weeks) > 5mM 10mM 20mM IP-Glc.” Each file’s name includes date corresponding to the injection and the metabolite for efficient data organisation.

Examples of filenames:

* Statistic comparisons: 250210\_Glucose.png
* Mass exact variation (all isotopologues combined): 250210\_Glucose\_exact\_mass\_all.png
* Mass exact variation (each isotopologue in a separated graph): 250210\_Glucose\_exact\_mass\_separated.png

Examples output plots for glucose are available in the “Plasma” folder within the R project directory

**Comments**

Some errors were found in the sample names in the **injection dates**. Some documents had two injection dates (when each document corresponds to one injection). To solve this, the program automatically checks that each document has a single injection date, and if not, it asks the user to select the correct one *(1.5)*.

**Adjustments**

* Show real data in the analytic control graphs

When analytic parameters are available in the document (exact mass, retention time or others) to create the analytic control graphs *(2.)* the programs requires some small adjustments in the *2.1.1 /2.2.1* section. **Keep only lines 105** and **108** for exact mass or **225** and **228** for retention time. In both cases **select the pertinent column** from “raw data” with “select()”. Additionally, the scale should be reviewed (change the set scale to a **free scale**). The **text** should also be checked to ensure it doesn’t overlap the points.

* Change directory names

The directories used to save graphs are currently named “Young (12 weeks)” and “5mM 10mM 20mM IP-Glc.” This naming was chosen for simplicity during initial development, when other samples were not yet available. To ensure flexibility as more data is added, a supplementary function is now needed to generate directory names based on the **age studied** and the **experimental conditions compared** (and eventually the tissue studied)**.** To create pertinent directories for the graphs generated create new variables (example names: “age” and “conditions\_compared”) extracting this information from the sample name with “str\_extract()” function.

**Future improvements**

* Implement an option to select specific conditions for comparison.

Currently, to generate the different plots the program creates reduced databases from the main dataset (“raw\_data”). These reduced datasets are named “raw\_metabolite” for analytical control graphs and “metabolite\_summary” for statistical comparison graphs. To implement an option to select the conditions to compare create a section after 1.4 with a **“menu()”** function for the user to choose. Further in the program the only main change would be to **filter the reduced databases** with the selected options. Other parts of the program might require small adjustments such as naming conventions or file paths, which must be carefully verified.

*(This “reduced databases” design facilitates finding errors because they can be traced more easily within these smaller datasets. It also facilitates modifications as filtering for the desired parameters, such as conditions to compare or eventually tissues to compare, as they can be done directly on these reduced databases without altering “raw\_data".)*

* Expand the data analysis to compare more tissues

Following the same principle as the first improvement, an **additional column named "tissue"** could be added to the “raw\_data” dataset extracting the information from the name with “str\_extract()”. The only requiredmodifications (though it would still be necessary to check for potential issues in other parts of the program) would be to **filter the selected tissues within the** **reduced datasets** and **define a plotting aesthetic** (e.g. colour or shape) to visually differentiate them in the graphs. Names and file paths should also be verified.

* Add a third option for metabolite selection: importing predefined metabolite lists.

These preestablished lists (e.g., lipids, proteins) imported from an Excel file would be provided within the program, eliminating the need for the user to select a large number of metabolites for analysis manually. The program can load such lists using RStudio’s **“Import Dataset > From Excel”** feature or a similar method to read Excel files into the R environment.

This option could be especially useful to focus data analysis on specific pathways, such as carbohydrate metabolism.

* Expand analytic controls (QC and Standards)

In the R project directory, there is a file named “QC (in process)”. This small script is designed to compare log2 peak areas between two QC injections (number 6 and 19). The program follows the same logic used in the other sections: it selects QC rows and creates a reduced dataset from “raw\_data” named “raw\_data\_QC” and creates a plot from it.

This part is not yet integrated into the program because some errors were encountered during its execution. The problem arises from the fact that, unlike the **t-tests** used to compare isotopologue fractions *(3)*, QC comparisons **require paired data**, as each isotopologue is compared across the two QC injections. The problem arises from the fact that **“log2 0” returns “-Inf”**, which cannot be compared to another value. The t-test for paired samples in R pairs data automatically, when “-Inf”, the value is ignored, which results in the comparison of **two vectors of different length**, which returns error. Several approaches have already been attempted to handle this, but they still produce errors. These attempted solutions are marked in the script with **#**, and they currently prevent this section from running correctly. These lines represent the main part of the script that needs to be revised for the module to function properly.

While the creation of the reduced dataset and the graphing logic appear to work, output file naming and folder saving have not yet been finalized. A new output directory should be added (*section 1.6*) of the main program to handle these graphs.

Proposed solutions:

* **Use areas instead of log 2-areas**

This avoids mathematical error of log2 0.

* **Assign NA to both values if one of them is missing and exclude the pair**

However, this approach reduces the amount of data being analysed and, more importantly, excludes values not because they lack scientific relevance, but because they cannot be properly compared. From a scientific standpoint, this may be questionable: molecules detected in only one injection carry meaningful information, and their exclusion could lead to **biased interpretations**.

**\*Compare combined isotopologue log2-areas** was also considered as an approach to avoid mismatches of data pairing. However, from a mathematic perspective it has a conceptual inaccuracy because comparing a zero to a summed value is not the equivalent to comparing a zero to log transformed areas. Nevertheless, this limitation may be less critical in practice because a value of zero implies that the molecule was not detected. In this context, treating undetected signals as “-Inf” after log transformation could still carry meaningful biological information. Therefore, while not currently implemented, this approach remains a possible alternative