Mass Isotopologue Distribution Analysis (MIDA)

Van Andel Research Institute Metabolomics Core
Brejnev Muhire and Christine Isaguirre

Table of Contents

1. Description

Isotopes are naturally occurring atoms that have one or more extra neutrons in their nucleus; the added neutron(s) increase the molecular weight of the atom. When an atom with an isotope is found in a metabolite, the mass shift from the additional neutron can be seen in Gas Chromatography-Mass Spectrometry (GC-MS) analysis.

This phenomenon can be utilized to design metabolite tracing experiments in which a labeled substrate (such as ^{13}C glucose) is introduced into a biological system (cell culture, mouse, etc.), and the labelled atoms can be tracked through metabolic pathways using GC-MS analysis. As the labeled atoms move through metabolic processes, labeling may accumulate in a certain metabolite, causing the metabolite to have a mass of *M+1, M+2, M+3…to M+i* (where *i* equals the number of carbons in the metabolite). In this manner, a researcher can understand the flux of metabolites in a system.

When analyzing the results of a tracing experiment, it is crucial to account for the natural abundance of stable isotopes (Midani). Fernandez et al. describes the matrix equation which can be used to account for the natural abundance of ^{13}C:

$$\begin{bmatrix} \mathrm{M0_{obs}} \\ \mathrm{M1_{obs}} \\ \mathrm{M2_{obs}} \\ \mathrm{M3_{obs}} \end{bmatrix} = \begin{bmatrix} \mathrm{S(0)_{m0}} & 0 & 0 & 0 \\ \mathrm{S(0)_{m1}} & \mathrm{S(1)_{m0}} & 0 & 0 \\ \mathrm{S(0)_{m2}} & \mathrm{S(1)_{m1}} & \mathrm{S(2)_{m0}} & 0 \\ \mathrm{S(0)_{m3}} & \mathrm{S(1)_{m2}} & \mathrm{S(2)_{m1}} & \mathrm{S(3)_{m0}} \end{bmatrix} \times \begin{bmatrix} \mathrm{M0_{corr}} \\ \mathrm{M1_{corr}} \\ \mathrm{M2_{corr}} \\ \mathrm{M3_{corr}} \end{bmatrix}$$

$A_{observed}$ **[CM]** **$A_{corrected}$**

$$A_{observed} = [CM] \times A_{corrrected}$$

$$A_{corrected} = [CM]^{-1} \times A_{observed}$$

Where:

[CM]	=	matrix of natural abundance correction factors
$A_{observed}$	=	vector of observed abundance (uncorrected data)
$A_{corrected}$	=	vector of corrected abundance (corrected data)

The correction factors represent the of the naturally occurring abundance of each isotopologue (*M+i).* The correction factors can be determined theoretically (based on the number of carbons contained in the compound, and the naturally occurring abundance for ^{13}C). However, when performing GC-MS analysis, the samples are typically derivatized to improve chromatographic separation and peak identification. During derivatization, other chemical groups are attached to

the compounds of interest, making it difficult to identify the exact chemical formula for the molecule. Furthermore, the impacts of GC-MS noise and uncertainty cannot be quantified and applied as part of the correction factor.

It is preferred to determine the natural abundance experimentally, by analyzing unlabeled samples alongside the labeled samples. In this manner, the difficulty of assigning chemical formulas to compounds can be avoided, and GC-MS variation can be incorporated.

Although relatively straightforward, construction and manipulation of the matrices needed to perform the correction becomes a cumbersome and error-prone process when completed with the large datasets generated by GC-MS analysis. Trefely et al. published a webtool based on R code to automatically perform the correction process. However, only one metabolite can be corrected at a time, reducing the potential efficiency of the correction process. Furthermore, the corrected data is presented in a simple matrix format, which requires additional manipulation to generate the figures needed to visualize the results.

The Mass Isotopologue Distribution Analysis (MIDA) program was created by the Van Andel Institute Metabolomics Core to streamline data analysis for metabolite tracing experiments. MIDA is stored on the VAI-Metabolism Github and is currently available for use by users within the Van Andel Institute. Shiny has been used to create an application that can be used to upload input files and view results without relying on the version of R on the user's computer and installation of necessary packages, or requiring the user to input commands into the command window.

2. MIDA Program

2.1 Input Format

Two input files are required for analysis: the data file, containing the GC-MS results, and a metadata file, containing information about the data. A third file is optional, which provides MIDA direction on which compounds to analyze. All three files should be saved together in one file folder. The files can be saved in either CSV or Excel format. It is important to ensure that no additional characters or spaces are included within these CSV or Excel files. Additionally, if Excel files are used, they should only have one tab per file.

2.1.1 Uncorrected Data File

For MIDA to identify a file as an uncorrected data file, the file name must end in "_uncorr.csv" or "_uncorr.xlsx". Additionally, the filename should be as brief as possible, and should not contain any spaces.

MIDA was designed to accept a standard data file output from MassHunter Workstation Software Quantitative Analysis by Agilent Technologies (Quant). With the standard output

procedure used by the VAI Metabolomics Core, little to no manipulation of the output data file will be necessary before uploading the data to the MIDA program.

However, Quant allows the user to change the formatting of the output file, so it is necessary to understand how the data should be formatted. Figure 2-1 shows a properly formatted MIDA input file. The sample name column should be the first column of the sheet. The next four columns contain metadata useful in Quant but are not needed for MIDA. MIDA ignores these columns, but it is important to include them in the uncorrected data file, as removing them may cause MIDA to generate an error.

The data should then be arranged in order of compound retention time (RT) from left to right. For each compound, the response for the unlabeled (M+0) isotopologue should be displayed first, and then the *M+i* isotopologues should be arranged in ascending order, moving right. It is unlikely that this ordering would be disrupted; however, if MIDA finds a compound out of order (i.e. M+0, M+2, M+1), it will generate an error. Simply using Excel to reorder the isotopologues will correct this error.

Most importantly, only the data for the compound response should be reported in the uncorrected data file. It will be common for the user to configure Quant to display other information such as RT or final compound concentration. However, this information should be excluded when exporting data for natural isotope correction.

					Pyruvate Results	Pyruvate M+1 Results	Pyruvate M+2 Results	Pyruvate M+3 Results	Lactate Results	Lactate M+1 Results	Lactate M+2 Results	Lactate M+3 Results
Name	Data File	Type	Level	Acq. Date-Time	Resp.	Resp.	Resp.	Resp.	Resp.	Resp.	Resp.	Resp.
Blank_1	Blank_1.D	Sample		7/15/2019 16:31	15380	1980	614	1216	163959	32336	377893	40380
0min_a	0min_a.D	Sample		7/15/2019 17:14	6442646	868006	324565	24364	5217190	1250544	830266	103262
0min_b	0min_b.D	Sample		7/15/2019 17:57	5020177	662908	240234	16755	5064552	1234812	821446	119660
0min_c	0min_c.D	Sample		7/15/2019 18:40	3717213	478084	174101	12140	3601465	849116	698383	75078
Blank_2	Blank_2.D	Sample		7/15/2019 19:23	29424	2561	928	1353	333694	66642	416185	18587
20min_a	20min_a.D	Sample		7/15/2019 20:06	6701832	936370	344214	26422	4922736	1189268	821469	128954
20min_b	20min_b.D	Sample		7/15/2019 20:49	8707993	1223262	457281	37342	5477819	1295587	858876	99288
20min_c	20min_c.D	Sample		7/15/2019 21:32	7724890	1075413	388054	32863	4630630	1075568	797195	84972
Blank_3	Blank_3.D	Sample		7/15/2019 22:15	35560	4736	1959	76	394830	76273	420337	28706
1h_a	1h_a.D	Sample		7/15/2019 22:58	5181952	685477	258728	22756	4536691	1088041	797703	100957
1h_b	1h_b.D	Sample		7/15/2019 23:41	7096782	953857	363127	31163	5197828	1226366	842972	
1h_c	1h_c.D	Sample		7/16/2019 0:24	7841484	1044085	398633	33100	5483322	1347829	879188	201018
Blank_4	Blank_4.D	Sample		7/16/2019 1:07	32889	3144	1115	1482	487766	109553	414711	15110
2h_a	2h_a.D	Sample		7/16/2019 1:50	9105700	1303694	483999	43042	5864504	1409711	906430	116367
2h_b	2h_b.D	Sample		7/16/2019 2:33	6556746	873036	323530	30224	5589634	1333775	877744	105710
2h_c	2h_c.D	Sample		7/16/2019 3:16	7121191	951436	365710	30754	5431881	1359129	857430	137872

Figure 2-1 Proper Formatting for Uncorrected Data File

2.1.2 Metadata File

A metadata file must be uploaded alongside the uncorrected data file. The metadata file name should be the same as the uncorrected data file name, except it should end with "_mtd.csv" or "_mtd.xlsx". Refer to Figure 2-2 for the proper file format. The first column of the metadata file should exactly match the first column of the uncorrected data file, displaying the identical sample name, with identical row order. The second column provides MIDA information on which samples are blanks, unlabeled, and biological or technical replicates. Unlabeled samples should simply be denoted as "unlabeled" and blanks should be denoted as "blank". Labeled

samples should be labeled “labeled_sampleX” where the X provides a numerical representation for a group of sample replicates. All samples from the same replicate group should be designated with the same number.

Name	Labeling
Blank_1	blank1
0min_a	unlabeled
0min_b	unlabeled
0min_c	unlabeled
Blank_2	blank2
20min_a	labeled_sample1
20min_b	labeled_sample1
20min_c	labeled_sample1
Blank_3	blank3
1h_a	labeled_sample2
1h_b	labeled_sample2
1h_c	labeled_sample2
Blank_4	blank4
2h_a	labeled_sample3
2h_b	labeled_sample3
2h_c	labeled_sample3

Figure 2-2 Proper Formatting for Metadata File

2.1.3 Compound File

Frequently, specific compounds are not found within a sample; response values associated with these compounds do not provide reliable data. Such values are often generated by signal noise and are unrelated to the expected isotope distribution of the compound. Including this erroneous data requires unnecessary analysis and may generate errors due to the irregular data quality. For such compounds, it is best to exclude them from the analysis.

A third file, which includes the compound names and a corresponding binary input (1=Analyze, 0=Ignore), can be used to tell MIDA to ignore compounds that are not of interest in the current study. The compound name should exactly match the compound name in the uncorrected data file. It is not necessary to include each isotopologue of each compound. Figure 2-3 provides an example of a properly formatted compound file. The compound file should be a CSV or Excel file named with the same file name as the uncorrected and metadata files but ending with “_cmpd.csv” or “_cmpd.xlsx”.

Compound	Included
Pyruvate	1
Lactate	1
L.Alanine	1
L.Glycine	1
X3.Hydroxybutyrate	0
L.2.Aminobutanoic.Acid	0
Beta.Alanine	0
N.Acetyl.glutamate	0
L.Valine	0
L.Leucine	0
L.Isoleucine	0
Succinate	1
L.Proline	1
Fumarate	0
Pyroglutamate	0
L.Methionine	1
Serine	1
alpha.Ketoglutarate	0
L.Threonine	0
L.Phenylalanine	0
Malate	1
Aspartic.Acid	1
L.Cysteine	0
L.Glutamate	1
Palmitate	0
L.Asparagine	0

Figure 2-3 Proper Formatting for Compound File

2.2 Workflow

The following section describes how to run MIDA and its computational processes.

2.2.1 Starting MIDA

Place the input *_uncorr files and the associated *_metadata and *_cmpd files in one folder. Within the MIDA app use the “Browse Files” icon to select the folder.

To run the command using the command prompt:
Execute the following commands:
MIDA.r : Use all compounds provided in the table
MIDA.r List or *MIDA.r L* : Use only a selection of compound indicated in the compounds.csv file

2.2.2 Data Preprocessing

The following data manipulation and filtering is completed before correction

1. Reads input of table containing labeled & unlabeled samples
2. Selects only compounds indicated by compound table
3. Checks that the ordering of isotopes is correct (M+0, M+1, M+2…M+i)

4. Excludes compounds below baseline detection (2/3 of unlabeled samples M+0 are below 1000 or all labeled sample entries are < 1000)
5. For compounds which meet the proceeding criteria but have missing values (due to low detection), substitutes the missing values with zeros.
    a. For each isotopologue, if at least one sample in a sample group, contains a value, all missing data is replaced with the minimum value (corresponding to the lowest detection)

2.2.3 Isotopologue Correction for Labeled Samples

The correction process itself can be described in three steps:

1. For each compound, create the correction matrix from unlabeled data (if multiple unlabeled data samples are available, the mean of each isotopologue is used)
2. For each sample, take the inverse of the data matrix ($A_{observed}$)
3. For each sample, multiply the correction matrix by the data vector to obtain the corrected data vector ($A_{corrected}$)

Refer to Section 1 for a mathematical explanation of the correction process.

2.2.4 Data Visualization

After correction, MIDA prepares multiple figures to help the user visualize and understand the data:

1. Total Enrichment Bar Plot
    a. Allows user to visualize the total pool size of each compound for each sample group (or individual sample), and the labeling present within the compound.
        i. Averages corrected isotopologue results for each compound across each sample group
        ii. Calculates standard error between XXXXX
        iii. Uses ggBar function to plot data
2. Percent Enrichment Bar Plot
    a. Normalizes the total pool size of each compound for each sample group (or individual sample), which allows the user to more easily compare differences in labeling between sample groups.
        i. Averages corrected isotopologue results for each compound across each sample group
        ii. Performs data normalization (sets total pool size of each compound equal to 100%)
        iii. Uses ggBar function to plot normalized data
3. TCA Cycle Figure
    a. Summary figure which displays isotope labeling for key TCA cycle metabolites.
        i. Arranges total enrichment graphs into a TCA cycle figure

4. Heatmap
    a. Heatmap comparing total sample labeling for heavily labeled compounds
        i. Sums (M+1, M+2, . . . , M+i) for each sample
        ii. Excludes analytes with < 5% labeling
        iii. For each analyte, divides the values obtained for each sample by the mean of all values
        iv. Plots heatmap using pheatmap function

2.2.5 Data Quality Control

MIDA also calculates the coefficient of variation across sample groups for each isotopologue for uncorrected and corrected data. The coefficient of variation provides a measure of precision, indicating how closely replicate data is grouped. The coefficient of variation is the standard deviation of the sample group divided by the mean, then multiplied by 100 to express it as a percent.

[Describe coefficient of variation figure]

Remove compounds with fractional enrichment of <95 or >105 in unlabeled samples

2.3 Output Format

When analysis is complete, a file folder is generated with all the outputs prepared by MIDA. This folder will be found in the location where the input files were saved, and it will be named with the same file name as the input files, with "_out" appended to the end of the file name.

Table 2-1 displays the file name or file name format, file type, and description of the files contained within the main directory of the output folder. Examples of the generated figures are shown on the following pages.

Table 2-1 Contents of MIDA Output Folder

	File Name/Name Format	File Type	Description
1	InputFileName**_TE_corr**	CSV or Excel	Corrected total enrichment matrix (contains corrected isotopologue data of the total response for each isotopologue)
2	InputFileName**_PRCNT_corr**	CSV or Excel	Corrected percent enrichment matrix (contains corrected isotopologue data as a percentage of each compound)
3	CompoundName**_TE.repset**	PDF	Total enrichment plots for each analyzed compound showing the average values of replicates (Figure 2-4)
4	CompoundName**_PRCNT.repset**	PDF	Percent enrichment plots for each analyzed compound showing the average values of replicates (Figure 2-5)

5	**TCA**	PDF	Figure visualizing isotope labeling for key TCA cycle metabolites (Figure 2-6)
6	**Heatmap**	PDF	Heatmap comparing total sample labeling for heavily labeled compounds (Figure 2-7)
7	**CV**	PDF	Bar plot for coefficients of variation (Figure 2-8)
8	**extra**	File Folder	Folder containing additional figures and analysis (Table 2-2)
9	**Formatted Data**	File Folder	Folder containing corrected data files for each compound which is formatted for easy transfer to XXXX to generate publication-quality figures.

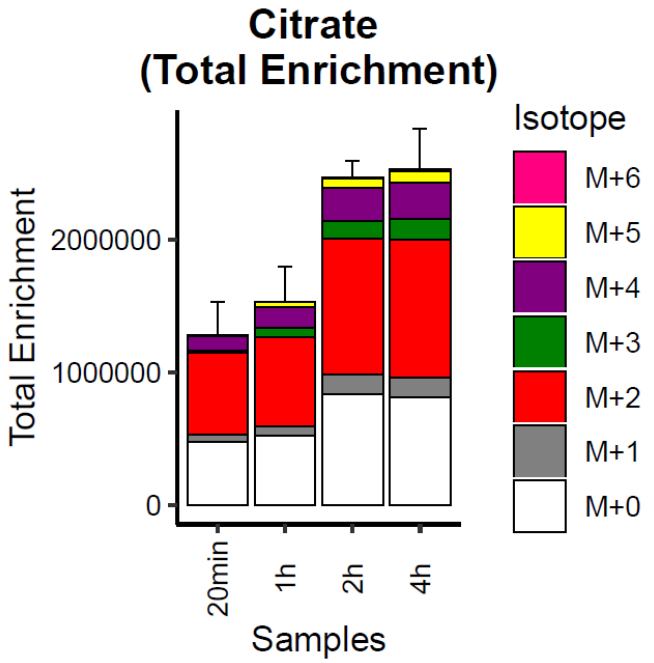

Figure 2-4 Example of Total Enrichment Figure Generated by MIDA

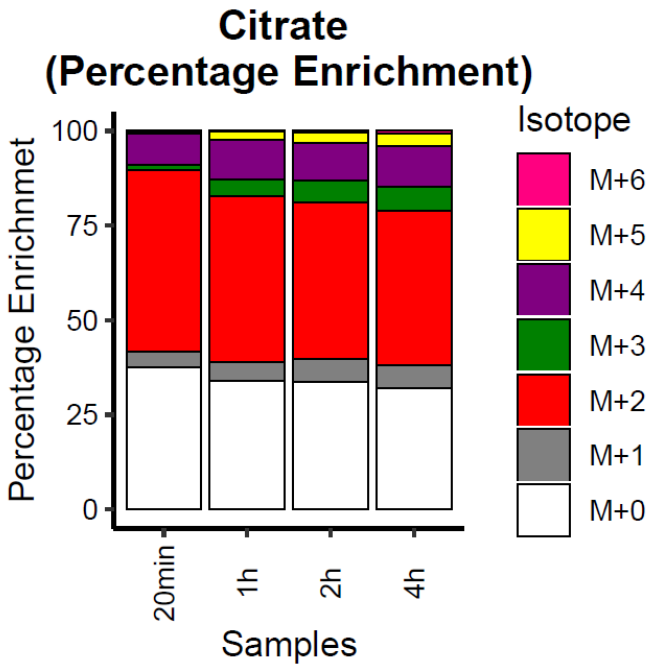

Figure 2-5 Example of Percent Enrichment Figure Generated by MIDA

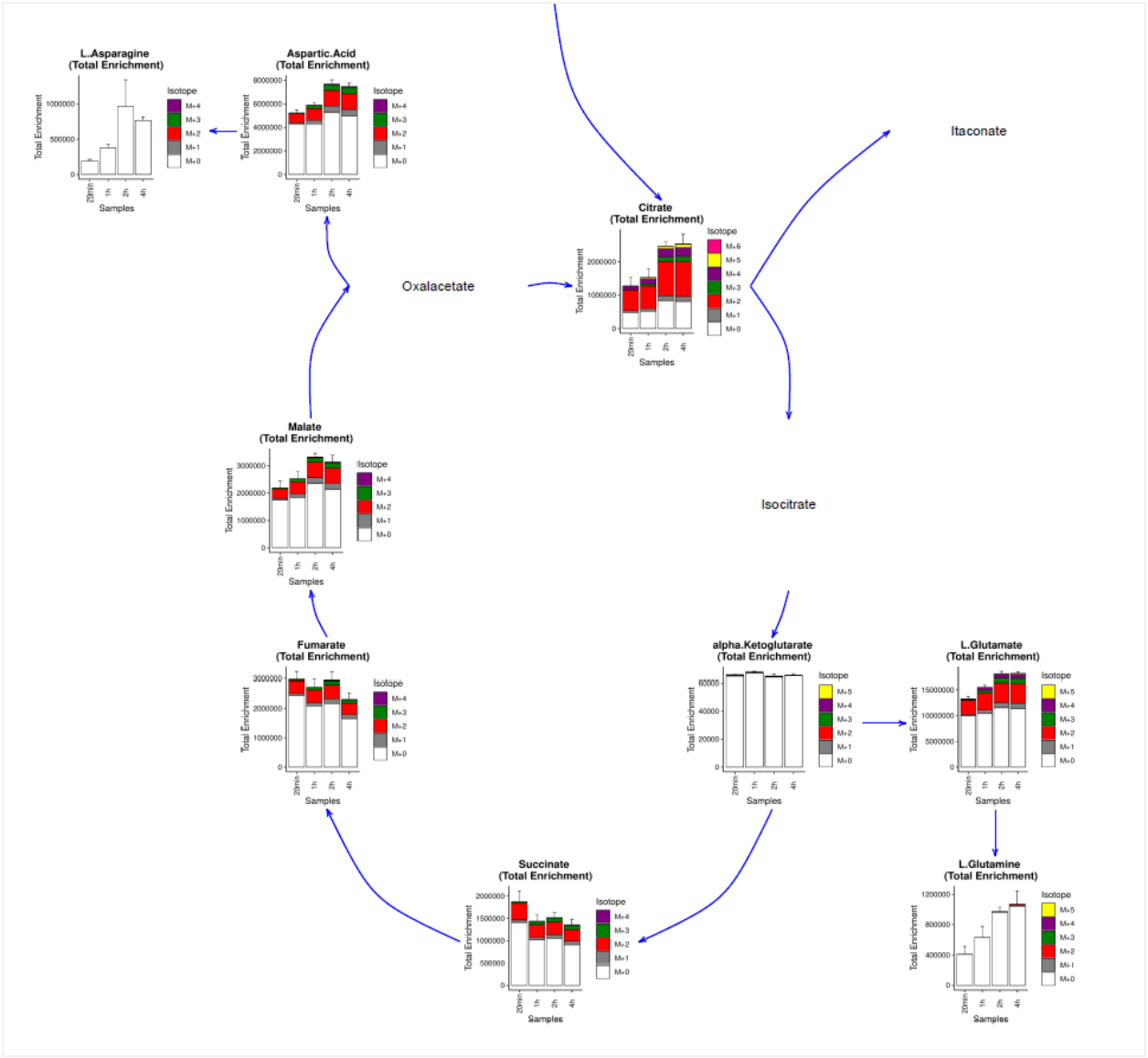

Figure 2-6 Example of TCA Figure Generated by MIDA

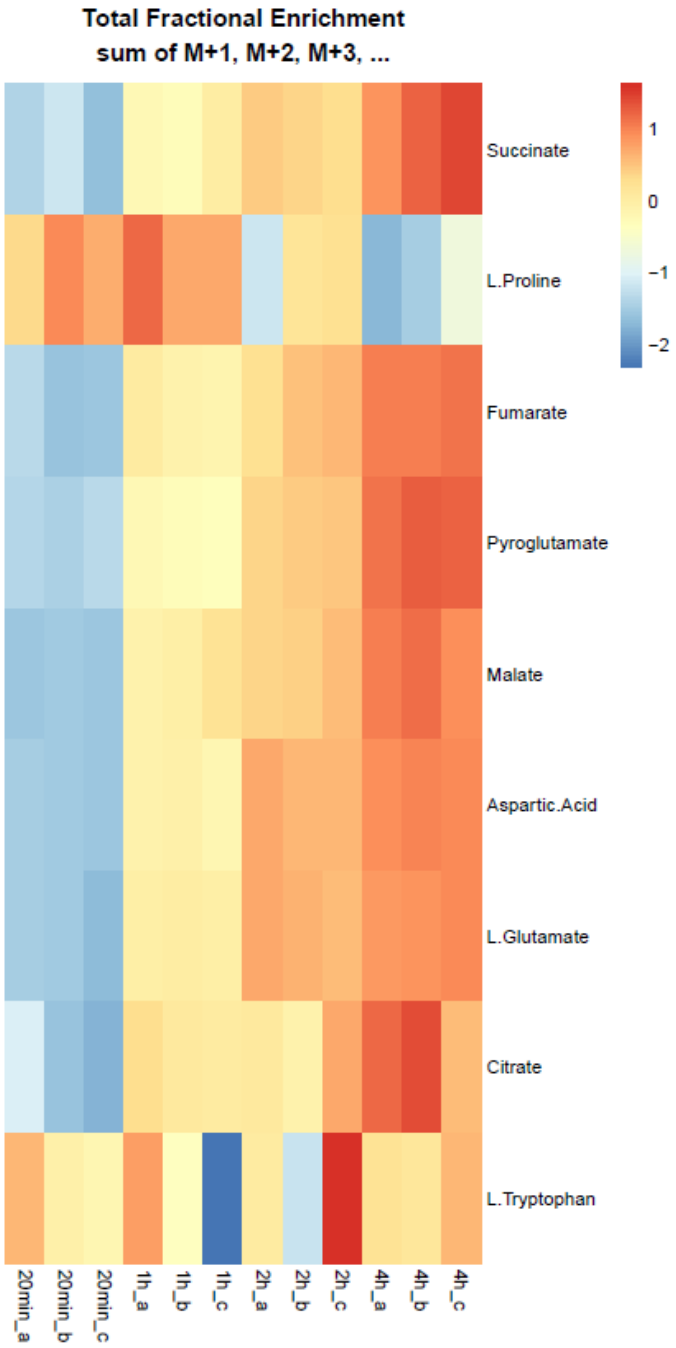

Figure 2-7 Example of the Heatmap for Heavily Labeled Compounds

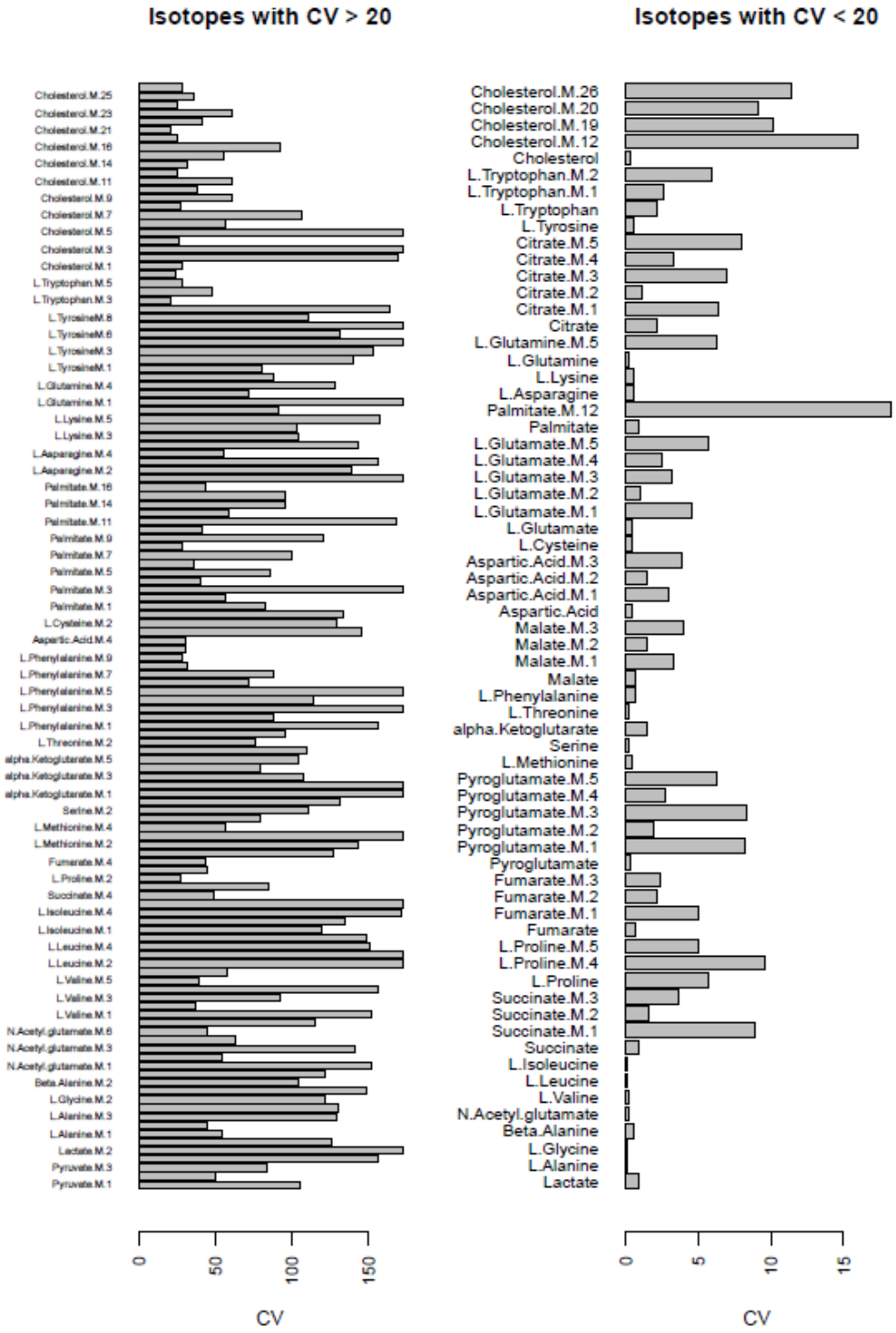

Figure 2-8 Example of the Coefficient of Variation Bar Plot

Table 2-2 Contents of MIDA Extra Folder

	File Name/Name Format	File Type	Description
3	CompoundName**_TE**	PDF	Total enrichment plots for each compound showing results for each replicate
4	CompoundName**_PRCNT**	PDF	enrichment plots for each analyzed compound showing results for each replicate
5	**CV_corr.FE**	CSV or Excel	Coefficient of variation results for fractional enrichment
6	**CV_uncorr**	CSV or Excel	Coefficient of variation for uncorrected data

3. References

Fernandez, C. A., Des Rosiers, C., Previs, S. F., David, F., & Brunengraber, H. (1996). Correction of 13C mass isotopomer distributions for natural stable isotope abundance. *Journal of Mass Spectrometry*, *31*(3), 255–262. https://doi.org/10.1002/(SICI)1096-9888(199603)31:3<255::AID-JMS290>3.0.CO;2-3

Midani, F. S., Wynn, M. L., & Schnell, S. (2017). The importance of accurately correcting for the natural abundance of stable isotopes. *Analytical Biochemistry*, *520*, 27–43. https://doi.org/10.1016/j.ab.2016.12.011

Trefely, S., Ashwell, P., & Snyder, N. W. (2016). FluxFix: Automatic isotopologue normalization for metabolic tracer analysis. *BMC Bioinformatics*, *17*(1), 1–8. https://doi.org/10.1186/s12859-016-1360-7