Format MESA TOPMed Multiomics Project Metabolite Data

LekkiWood (LekkiWood@Gmail.com)

2025-10-20

Table of contents

1. Overview

Last Update

- This file was last updated on Monday, October, 20, 2025

Summary

This README file describes the process for taking the raw data from the Broad lab for the metabolite assays from the MESA TOPMed multiomics project and creating the following:

- File 1: A metabolite table in wide format for the compounds (compounds are rows) and long format for the time points (one row per participant, per exam), with any data points of concern (e.g., duplicated abundances) removed. Current file name: MESA_TOPMed_Metabolite_Longform_2025-10-16.csv[1]
- File 2: A mapping file, that maps the compound identifiers from file 1 (i.e., column names) to (1) the compound name used in the X01 pilot; (2) The original metabolite name for known compounds as supplied by the lab; (3) the ID for the compound from the Human Metabolome Database (HMDB) (if available); and information on (1) Whether the compound is known or not (0= unknown; 1 = known); and (2) whether the compound was included in each of the three MESA TOPMed multiomics project sub-projects. Current file name: MESA_TOPMed_Metabolite_Mappingfile_2025-10-16.csv[1]
- File 3: A file with QC_information, including the coefficient of variation (CV) for each compound and the missingness for each compound by exam. Current file name: MESA_TOPMed_Metabolite_QCfile_2025-10-18.csv[1]
- File 4: A file with flags for duplicated *compounds*. Current file name: MESA_TOPMed_Metabolite_Dupli 10-19.csv [1]
- File 5: A final metabolite table, formatted as in file 1, and where the same compound was measured in more than one assay, only assay with the lowest CV is retained in this file. Current file name: MESA_TOPMed_Metabolite_cleanfile_2025-10-19.csv [1]

Important notes:

- You should always check you have the most recent files! I use dynamic dates, so this can be done by checking that your files match the versions on Github:
  - My profile can be found here
  - My repositories are listed here

- The most recent version of this PDF README file can be found here
- The most recent version of the source code for this README file can be found here
- The most recent source code for creating the output files can be found here
- If you want to know the all the pain I went through to create these, or check I am not trying to cheat the system, my entire R history for this project can be found here

1. Unresolved concerns

Assay questions

- The C18 assay has a smaller N than some others

MESA DCC questions

- Duplicated SHARe IDs (sidno) / exam combinations (advice from DCC / lab = exclude all).
- Some SHARe IDs are still missing from the bridging file

Metabolite names

Unresolved

- Most duplicated compounds are measured on different assays. One exception is LPC 16:0/0:0, which is measured twice on the C8 assay, with the same HMDB, but different assigned LAB IDs. For now, the compound with the Compound ID in the X01 of "TF04" is renamed as LPC 16:0/0:0_v2.
- Some lipids have the same HMDB ID, but the different common metabolite names, and these names are not listed as synonyms in the HMDB (e.g., Cer 40:1;O2 (C8 assay) is not listed as a synonym for 18:1 (HILIC), and Cer 42:2;O2_A (C8) is not listed for a synonym for Cer 18:1;O2/24:1 (HILIC). For now, these are treated as separate compounds.

Lab ID	HMDB	Metabolite	Assay
QI04610	HMDB0004952	Cer 18:1;O2/22:0	hilic
QI8023	HMDB0004952	Cer 40:1;O2	c8

Lab ID	HMDB	Metabolite	Assay
QI04540	HMDB0004953	Cer 18:1;O2/24:1	hilic
QI4265	HMDB0004953	Cer 42:2;O2_A	c8

Probably resolved, but still to be checked with lab

- Glutamic acid had different HMDBs on the original HILIC and amide assays. Per advice: The HMDB from the HILIC was retained for both assays, and these were treated as the same compound.

Lab ID	HMDB	Metabolite	Assay
QI18171	HMDB0000148	Glutamic acid	hilic
Glutamic.acid	HMDB0003339	Glutamic acid	amide

- For the amide assay, the compound with HMDB ID "HMDB0000122" had a different name to that used for the HILIC and C18 assays (but both names "correct" - using different conventions). They were treated as the same compounds when screening for duplicates.

Lab ID	HMDB	Metabolite	Assay
Glucose.Fructose. Galactose_waterloss	HMDB0000122	Glucose/Fructose/ Galactose_waterloss	amide
QI3656	HMDB0000122	Hexose	c18
QI14125	HMDB0000122	Hexose	hilic

For Lekki

- Table hyperlinks are still not rendering in Quarto and no one can fix it.

2. File 1: Formatted Metabolite Assay Tables

Input file names

Raw metabolite tables

- A table of metabolite abundances from the amide assay: 25_0107_TOPMed_MESA_Amide-neg_rev031325.csv
- A table of metabolite abundances from the C8-pos assay: 24_1210_TOPMed_MESA_C8-pos_checksums_rev031325.csv
- A table of metabolite abundances from the C18-neg assay: 24_1210_TOPMed_MESA_C18-neg_checksums_rev031325.csv
- A table of metabolite abundances from the HILIC assay: 24_1210_TOPMed_MESA_HILIC-pos_checksums_rev031325.csv

Sample info files

- Sample information for the amide assay: MesaMetabolomics_PilotX01_AmideNeg_SampleInfo_2025032
- Sample information for the C8-pos assay: MesaMetabolomics_PilotX01_C8Pos_SampleInfo_20250329.tx
- Sample information for the C18-neg assay: MesaMetabolomics_PilotX01_C18Neg_SampleInfo_20250329
- Sample information for the HILIC assay: MesaMetabolomics_PilotX01_HILIC-Pos_SampleInfo_20250329.txt

Bridging file

- A file to bridge SHARe ids (sidno) with MESA IDs (idno) MESA-SHARE_IDList_Labeled.csv

Formatting metabolite tables

Amide

- The amide sample info file contained information on N=13727 TOM IDs (each corresponding to one participant's abundances at one exam).
- Metabolites were renamed according to the following rules:
  - For known compounds:

        * The compound name (supplied by the lab) was transformed using the "make.names" command with unique=TRUE, and prefixed with the string "Amide_".

    - The amide assay only has known compounds.

- When the TOMIDs in the sample info file were extracted from the original metabolite table, there were N=NULL TOM IDs remaining.[2]
- We searched for duplicated TOMIDs, and N= were found.[2]
- One value for TOM148122 had an unusual value - likely an incorrect rounding problem (1+2559:2584.03935869994933). This was re-coded to missing.
- The table was transposed. The following duplicates were identified:

sidno	exam	n
10509	5	2
14687	6	2
25200	5	2

    - The duplicates were removed leaving a final table of N= 155 metabolites across 13721 observations.

C8

- The C8 sample info file contained information on N=13722 TOM IDs (each corresponding to one participant's abundances at one exam).
- Metabolites were renamed according to the following rules:
    - For known compounds:
        * The compound name (supplied by the lab) was transformed using the "make.names" command with unique=TRUE, and prefixed with the string "C8_".
    - For unknown compounds:
        * The lab ID (arbitrarily assigned by The Broad) assigned within the X01 study (variable: Compound_ID_X01; arbitrarily chosen from the MESA2 and MESA PILOT names by Lekki) was prefixed with the string "C8_".
- When the TOMIDs in the sample info file were extracted from the original metabolite table, there were N=NULL TOMIDs remaining.[2]
- We searched for duplicated TOMIDs, and N= were found.[2]

- The the table was transposed. The following duplicates were identified:

sidno	exam	n
10509	5	2
14687	6	2
25200	5	2

- The duplicates were removed leaving a final table of N=869 metabolites across N=13716 observations.

C18

- The C18 sample info file contained information on N=13684 TOM IDs (each corresponding to one participant's abundances at one exam).
- Metabolites were renamed according to the following rules:
  - For known compounds:
    - The compound name (supplied by the lab) was transformed using the "make.names" command with unique=TRUE, and prefixed with the string "C18_".
  - For unknown compounds:
    - The lab ID (arbitrarily assigned by The Broad) assigned within the X01 study (variable: Compound_ID_X01; arbitrarily chosen from the MESA2 and MESA PILOT names by Lekki) was prefixed with the string "C18_".
- When the TOMIDs in the sample info file were extracted from the original metabolite table, there were N=NULL TOMIDs remaining.[2]
- We searched for duplicated TOMIDs, and N= were found.[2]
- The the table was transposed. The following duplicates were identified:

sidno	exam	n
10509	5	2
14687	6	2

- The duplicates were removed leaving a final table of N=2655 metabolites across N=13680 observations.

HILIC

- The HILIC sample info file contained information on N=13726 TOM IDs (each corresponding to one participant's abundances at one exam).
- Metabolites were renamed according to the following rules:
  - For known compounds:
    - The compound name (supplied by the lab) was transformed using the "make.names" command with unique=TRUE, and prefixed with the string "HILIC_".
  - For unknown compounds:
    - The lab ID (arbitrarily assigned by The Broad) assigned within the X01 study (variable: Compound_ID_X01; arbitrarily chosen from the MESA2 and MESA PILOT names by Lekki) was prefixed with the string "HILIC_".
- When the TOMIDs in the sample info file were extracted from the original metabolite table, there were N=NULL TOMIDs remaining.[2]
- We searched for duplicated TOMIDs, and N= were found.[2]
- The the table was transposed. The following duplicates were identified:

sidno	exam	n
10509	5	2
14687	6	2
25200	5	2

- The duplicates were removed leaving a final table of N=906 metabolites across N=13720 observations.

Binding formatted metabolite assay tables into one metabolite table

- The data from the 4 assays were merged, yielding information on N= 4585 compounds, across N= 13726 observations.
- Those observations represented N=6431 unique individuals, with data on N=6375 individuals at exam 1, N=4341 at exam 5, and N=3010 at exam 6.
- MESA SHARe IDs were merged in to the final metabolite table using the bridging file above. The following SHARe IDs were missing from the bridging table: 10185, 10185, 26389, 26389, 26389 .

- The final metabolite table for analysis is in long form, and has 13726 observations on 4590 variables (5 variables are not metabolites: sidno, subject_id, TOM_ID, exam, idno).

Output files

Metabolite table

The metabolite table was saved in long form with the file name MESA_TOPMed_Metabolite_Longform_2025-10-16.csv[1]

File 2: Mapping file

Input file names

Raw metabolite tables

- A table of metabolite abundances from the amide assay: 25_0107_TOPMed_MESA_Amide-neg_rev031325.csv
- A table of metabolite abundances from the C8-pos assay: 24_1210_TOPMed_MESA_C8-pos_checksums_rev031325.csv
- A table of metabolite abundances from the C18-neg assay: 24_1210_TOPMed_MESA_C18-neg_checksums_rev031325.csv
- A table of metabolite abundances from the HILIC assay: 24_1210_TOPMed_MESA_HILIC-pos_checksums_rev031325.csv

Formatting

Amide

- Variables were renamed as follows:
  - "Presence MESA X01" renamed to: "Included_X01" (coded 1=yes, 0 = no)
  - "Presence MESA PILOT" renamed to: "Included_Pilot" (coded 1=yes, 0 = no)
  - "Presence MESA MESA2" renamed to: "Included_MESA2" (coded 1=yes, 0 = no)
  - "DB_ID" renamed to "HMDB_ID".
- The original metabolite name (where known) from the lab was preserved as the variable 'Original_Metabolite_Name'.
- Metabolites were renamed according to the following rules:
  - For known compounds:
    - The compound name (supplied by the lab) was transformed using the "make.names" command with unique=TRUE, and prefixed with the string "Amide_".
- A variable "Known_Compound" was created to indicate whether the compound had been assigned a name as of Monday, October, 20, 2025[2] (coding 1= yes, 0=no) [amide is currently all known compounds).

  - The amide assay has N= 155 known compounds and no unknown compounds.

- A variable "Compound_ID_X01" was created allow this to map on to other assays which have unknown compounds included.

C8

- The following variables were created:
  - "Presence MESA X01" (coded 1)
  - "Presence MESA PILOT" (coded 1)
  - "Presence MESA MESA2" (coded 1).
- The original metabolite name (where known) from the lab was preserved as the variable 'Original_Metabolite_Name'
- Metabolites were renamed according to the following rules:
  - For known compounds:
    * The compound name (supplied by the lab) was transformed using the "make.names" command with unique=TRUE, and prefixed with the string "C8_".
  - For unknown compounds:
    * The lab ID (arbitrarily assigned by The Broad) assigned within the X01 study (variable: Compound_ID_X01; arbitrarily chosen from the MESA2 and MESA PILOT names by Lekki) was prefixed with the string "C8_".
- A variable "Known_Compound" was created to indicate whether the compound had been assigned a name as of Monday, October, 20, 2025[2] (coding 1= yes, 0=no).
  - The C8 assay has N= 348 known compounds and 521 unknown compounds.

C18

- The following variables were created:
  - "Presence MESA X01" (coded 1)
  - "Presence MESA PILOT" (coded 1)
  - "Presence MESA MESA2" (coded 1).

- The original metabolite name (where known) from the lab was preserved as the variable 'Original_Metabolite_Name'
- Metabolites were renamed according to the following rules:
  - For known compounds:
    * The compound name (supplied by the lab) was transformed using the "make.names" command with unique=TRUE, and prefixed with the string "C18_".
  - For unknown compounds:
    * The lab ID (arbitrarily assigned by The Broad) assigned within the X01 study (variable: Compound_ID_X01; arbitrarily chosen from the MESA2 and MESA PILOT names by Lekki) was prefixed with the string "C18_".
- A variable "Known_Compound" was created to indicate whether the compound had been assigned a name as of Monday, October, 20, 2025[2] (coding 1= yes, 0=no).
  - The C18 assay has N= 130 known compounds and 2525 unknown compounds.

HILIC

- The following variables were created:
  - "Presence MESA X01" (coded 1)
  - "Presence MESA PILOT" (coded 1)
  - "Presence MESA MESA2" (coded 1).
- The original metabolite name (where known) from the lab was preserved as the variable 'Original_Metabolite_Name'
- Metabolites were renamed according to the following rules:
  - For known compounds:
    * The compound name (supplied by the lab) was transformed using the "make.names" command with unique=TRUE, and prefixed with the string "HILIC_".
  - For unknown compounds:
    * The lab ID (arbitrarily assigned by The Broad) assigned within the X01 study (variable: Compound_ID_X01; arbitrarily chosen from the MESA2 and MESA PILOT names by Lekki) was prefixed with the string "HILIC_".

- A variable "Known_Compound" was created to indicate whether the compound had been assigned a name as of Monday, October, 20, 2025[2] (coding 1= yes, 0=no).
  - The HILIC assay has N= 302 known compounds and 604 unknown compounds.

Binding formatted mapping files into one mapping file

- The mapping information from the 4 assays were row bound.
- The final mapping file has information on 935 known compounds and 3650 unknown compounds

Output files

Mapping file

The metabolite table was saved in long form with the file name MESA_TOPMed_Metabolite_Mappingfile_2025-10-16.csv[1]

File 3: QC file

Input files

Raw metabolite tables

- A table of metabolite abundances from the amide assay: 25_0107_TOPMed_MESA_Amide-neg_rev031325.csv
- A table of metabolite abundances from the C8-pos assay: 24_1210_TOPMed_MESA_C8-pos_checksums_rev031325.csv
- A table of metabolite abundances from the C18-neg assay: 24_1210_TOPMed_MESA_C18-neg_checksums_rev031325.csv
- A table of metabolite abundances from the HILIC assay: 24_1210_TOPMed_MESA_HILIC-pos_checksums_rev031325.csv

Sample info files

- Sample information for the amide assay: MesaMetabolomics_PilotX01_AmideNeg_SampleInfo_2025032
- Sample information for the C8-pos assay: MesaMetabolomics_PilotX01_C8Pos_SampleInfo_20250329.tx
- Sample information for the C18-neg assay: MesaMetabolomics_PilotX01_C18Neg_SampleInfo_20250329
- Sample information for the HILIC assay: MesaMetabolomics_PilotX01_HILIC-Pos_SampleInfo_20250329.txt

Formatting

- For each of the four metabolite tables above, the QC samples were selected by selecting Sample_Type=="QC-pooled_ref" from the sample info files.
- The number of QC samples used for calculating coefficients of variation (CVs) is included in the final mapping file.
- Metabolites were renamed according to the following rules:
  - For known compounds:
    - The compound name (supplied by the lab) was transformed using the "make.names" command with unique=TRUE, and prefixed with the string "[assay name]_".
  - For unknown compounds:

    - The lab ID (arbitrarily assigned by The Broad) assigned within the X01 study (variable: Compound_ID_X01; arbitrarily chosen from the MESA2 and MESA PILOT names by Lekki) was prefixed with the string "[assay name]_".

- A variable: "Known" was created to indicate identified compounds (coded 1) and unidentified compaoiunds (coded 0).

Coefficients of Variations

Calculation

- CVs were calculated across all exams (since batches were not startified by exam) as the mean of all QC samples, divided by the standard deviation of all QC samples, multiplied by 100, and saved as the variable "cv_percent".

Information

- Across, all assays, the mean CV was 13.7398974, with a range of 1.4705788 - 666.0002291.
- The CVs, per assay, are:

Assay	Mean_CV
Amide	11.48314
C18	14.49480
C8	11.50480
HILIC	14.04263

- The CVs per assay, stratified into known and unknown metabolites are:

Assay	Known	Mean_CV
Amide	Known Metabolites	11.483137
C18	Unknown Compounds	14.900993
C18	Known Metabolites	6.605369
C8	Unknown Compounds	15.601759
C8	Known Metabolites	5.371135
HILIC	Known Metabolites	14.042632

Missingness

Calculation

- Missingness was calculated per exam, as people often conduct exam-specific analysis. The proportion of missingness for each assay, by exam is below (note, this is for people with assay data at each exam, missingness does not include, for example, people who have data at exam 1, but have no data at exam 5):

Assay	Overall	E1	E5	E6
Amide	18.831274	19.2825806	22.044571	13.2412389
C18	6.181399	7.1368650	5.248741	5.5028499
C8	0.893234	0.7809743	1.017541	0.9517183
HILIC	9.420818	10.2592391	7.988617	9.7106041

Output file

- The file with CV and missing ness information was saved as MESA_TOPMed_Metabolite_QCfile_2025-10-18.csv [1]

File 4: Duplicate flags

Formatting

Cleaning / Harmonizing names:

- Glutamic acid had different HMDBs on the original HILIC and amide assays. Per advice: for the amide assay, the metabolite Glutamic acid had the HMDB ID changed from HMDB0003339 -> HMDB0000148.

Flagging duplicates

- After the file cleaning (see above), only the HMDB_ID variable was used to flag compounds measured on more than one assay. To flag these, a variable was created "Retain", with the following coding:
  - 0, labelled "Unique or missing HMDB ID": an unknown compound, or a unique HMDB ID
  - 1, labelled "Duplicated HMDB ID with lowest CV": A duplicated HMDB ID with the lowest coefficient of variation (CV) across all compounds sharing that HMDB ID
  - 2, labelled "Internal standard": A compound used as an internal standard (generally excluded from standard analyses)
  - 3, labelled "Duplicated HMDB_ID and not lowest CV": A duplicated HMDB ID that does *not* have the lowest coefficient of variation (CV) across all compounds sharing that HMDB ID (generally excluded from analyses where you do not want the same compound included more than once)
  - The variable Retain had the following frequencies:

Var1	Freq
Unique or missing HMDB ID	4343
Duplicated HMDB ID with lowest CV	116
Internal standard	5
Duplicated HMDB_ID and not lowest CV	121

Output

- The final file flagging potential duplicates was saved as MESA_TOPMed_Metabolite_Duplicateflag_2025-10-19.csv. [1]

File 5: Cleaned Metabolite File

Formatting

- The duplicate flagging file was used to select from file 1 created above, all metabolites that had either a unique HMDB ID, or that had a duplicated HMDB ID but had the lowest CV from all those metabolites with the same HMDB ID.
- The final metabolite file contained information on N=4458 metabolites across N=13726 observations.

Output file name

- The final metabolite file without duplicated compounds was saved as MESA_TOPMed_Metabolite_cleanf 10-19.csv [1]

Notes

Footnotes

[1]Dates are dynamic

[2]Missing, or "NULL" values here arise from using a standard script for all QC to flag potential issues. Missing, or "NULL" values typically indicate a lack of potentially concerning issue.

Session info

The exact session information used for this analysis

```
- Session info ---------------------------------------------------------------
 setting  value
 version  R version 4.5.1 (2025-06-13)
 os       Linux Mint 21
 system   x86_64, linux-gnu
 ui       X11
 language (EN)
 collate  en_US.UTF-8
 ctype    en_US.UTF-8
 tz       America/Chicago
 date     2025-10-20
 pandoc   3.2 @ /usr/lib/rstudio-server/bin/quarto/bin/tools/x86_64/ (via rmarkdown)
 quarto   1.8.25 @ /usr/local/bin/quarto

- Packages -------------------------------------------------------------------
 package     * version date (UTC) lib source
 backports     1.5.0   2024-05-23 [1] CRAN (R 4.5.0)
 base64url     1.4     2018-05-14 [1] CRAN (R 4.5.1)
 callr         3.7.6   2024-03-25 [1] CRAN (R 4.5.1)
 cli           3.6.5   2025-04-23 [1] CRAN (R 4.5.1)
 codetools     0.2-20  2024-03-31 [4] CRAN (R 4.5.0)
 data.table    1.17.8  2025-07-10 [1] CRAN (R 4.5.1)
 digest        0.6.37  2024-08-19 [1] CRAN (R 4.5.0)
 dplyr         1.1.4   2023-11-17 [1] CRAN (R 4.5.0)
 evaluate      1.0.5   2025-08-27 [1] CRAN (R 4.5.1)
 fastmap       1.2.0   2024-05-15 [1] CRAN (R 4.5.0)
 generics      0.1.4   2025-05-09 [1] CRAN (R 4.5.1)
 glue          1.8.0   2024-09-30 [1] CRAN (R 4.5.0)
 htmltools     0.5.8.1 2024-04-04 [1] CRAN (R 4.5.0)
```

```
 igraph        2.1.4   2025-01-23 [1] CRAN (R 4.5.0)
 jsonlite      2.0.0   2025-03-27 [1] CRAN (R 4.5.0)
 knitr         1.50    2025-03-16 [1] CRAN (R 4.5.0)
 later         1.4.2   2025-04-08 [1] CRAN (R 4.5.0)
 lifecycle     1.0.4   2023-11-07 [1] CRAN (R 4.5.0)
 magrittr      2.0.4   2025-09-12 [1] CRAN (R 4.5.1)
 pillar        1.11.1  2025-09-17 [1] CRAN (R 4.5.1)
 pkgconfig     2.0.3   2019-09-22 [1] CRAN (R 4.5.0)
 prettyunits   1.2.0   2023-09-24 [1] CRAN (R 4.5.0)
 processx      3.8.6   2025-02-21 [2] CRAN (R 4.5.1)
 ps            1.9.1   2025-04-12 [1] CRAN (R 4.5.0)
 quarto        1.5.1   2025-09-04 [1] CRAN (R 4.5.1)
 R6            2.6.1   2025-02-15 [1] CRAN (R 4.5.0)
 Rcpp          1.0.14  2025-01-12 [1] CRAN (R 4.5.0)
 rlang         1.1.6   2025-04-11 [1] CRAN (R 4.5.0)
 rmarkdown     2.29    2024-11-04 [1] CRAN (R 4.5.0)
 rstudioapi    0.17.1  2024-10-22 [1] CRAN (R 4.5.0)
 secretbase    1.0.5   2025-03-04 [1] CRAN (R 4.5.1)
 sessioninfo   1.2.3   2025-02-05 [1] CRAN (R 4.5.1)
 targets       1.11.4  2025-09-13 [1] CRAN (R 4.5.1)
 tibble        3.3.0   2025-06-08 [1] CRAN (R 4.5.1)
 tidyselect    1.2.1   2024-03-11 [1] CRAN (R 4.5.0)
 vctrs         0.6.5   2023-12-01 [1] CRAN (R 4.5.0)
 withr         3.0.2   2024-10-28 [1] CRAN (R 4.5.0)
 xfun          0.53    2025-08-19 [1] CRAN (R 4.5.1)
 yaml          2.3.10  2024-07-26 [1] CRAN (R 4.5.0)

 [1] /home/awood/R/x86_64-pc-linux-gnu-library/4.5
 [2] /usr/local/lib/R/site-library
 [3] /usr/lib/R/site-library
 [4] /usr/lib/R/library

------------------------------------------------------------------------------
```