**User manual for “HDPairFinder.r”**

(Version 1.0, December 21st, 2022)

HDPairFinder is a bioinformatics software to extract natural and deuterium methyl group labeled amino-containing compounds in LC-MS untargeted analysis of source water. This software is composed of four modules:

1. Extraction of hydrogen- and deuterium-labeled pairs
2. Alignment across multiple samples
3. Missing value retrieval
4. Tentative annotation

The program is written in R (ver 4.2.1). The R script “HDPairFinder.r” is the main program. An AMINES library, which archives the accurate mass, molecular formula and SMILES of amine compounds, is also curated to facilitate the identify the extracted H-/D-labeled pairs. Both the source code and the library are freely available on GitHub (<https://github.com/HuanLab/HDPairFiner>).

**Prerequisite**

To run the above R scripts, the user needs to prepare their computer with R software, RStudio software and R packages as below:

* install the R language ([www.r-project.org](http://www.r-project.org))
* install the RStudio ([www.rstudio.com](http://www.rstudio.com))
* install the R package of “xcms” using the following R scripts:

if (!require("BiocManager", quietly = TRUE))

install.packages("BiocManager")

BiocManager::install("xcms")

* install the R package of “ISFrag” using the following R scripts:

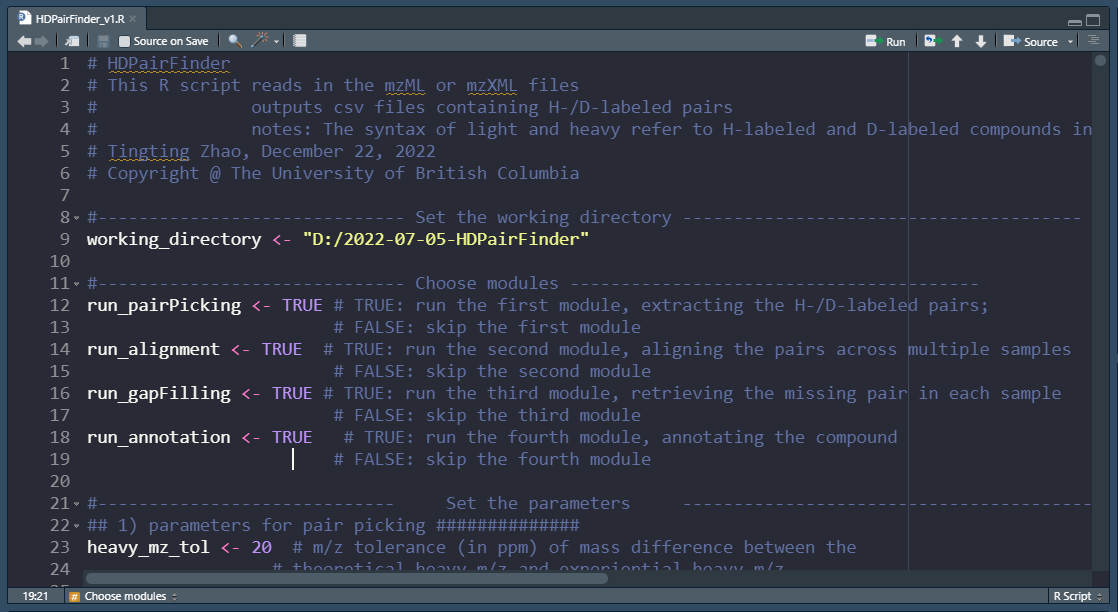
if (!requireNamespace("devtools", quietly = TRUE)){install.packages("devtools")}

library(devtools)

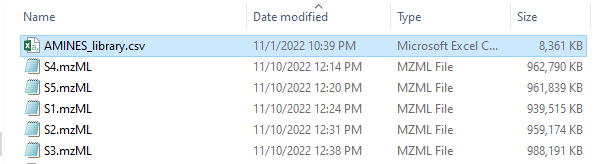
if (!requireNamespace("ISFrag", quietly = TRUE)){install\_github("HuanLab/ISFrag")

**Instructions for “HDPairFinder.r”**

1. Download and open the R script of “HDPairFinder.r” in RStudio.



1. Create a folder to all the sample files (in either mzML or mzXML format) and the AMINES library (“AMINES\_library”), as shown below.



1. Change the working directory to the above directory in the R script (line 9). Use “/” instead of “\” in the directory as shown below:



1. Choose the modules in lines 12 – 20.

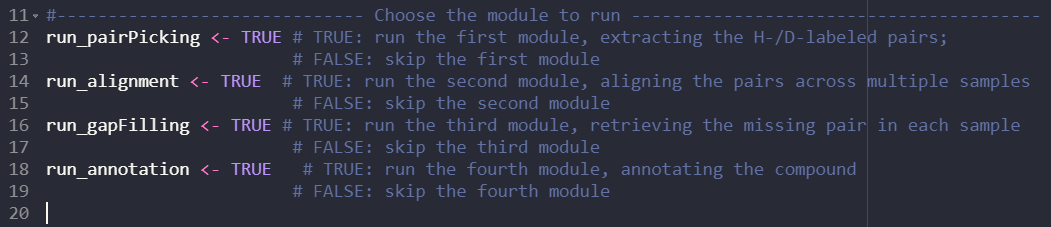


Table. Options of modules in HDPairFinder.

|  |  |
| --- | --- |
| **Parameters** | **Function** |
| run\_pairPicking | Logical, TRUE if extraction of H-/D-labeled pairs is performed, FALSE if not performed. |
| run\_alignment | Logical, TRUE if the alignment of pairs across samples is performed, FALSE if not performed. |
| run\_gapFilling | Logical, TRUE if the missing value retrieval is performed, FALSE if not performed. |
| run\_annotation | Logical, TRUE if annotation of the extracted H-/D-labeled pair is performed, FALSE if not performed. |

1. Set the parameters in lines 24 - 41.

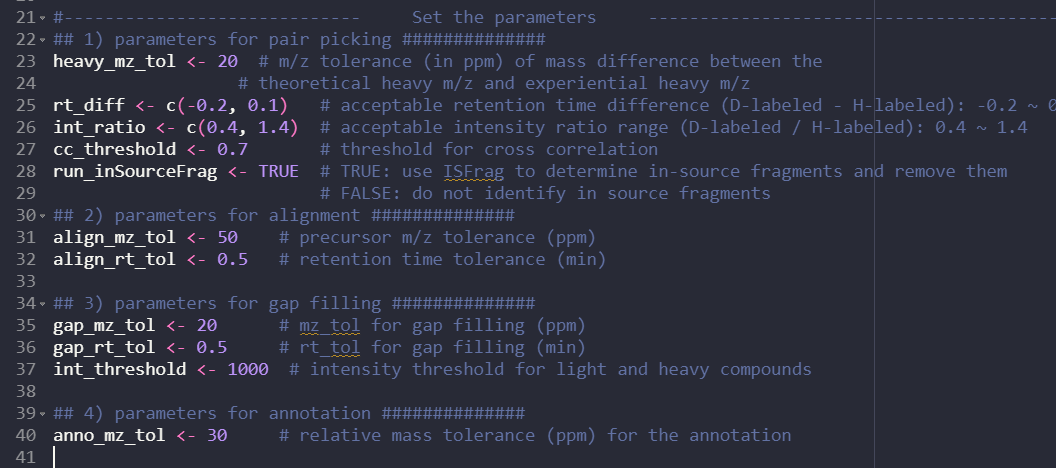
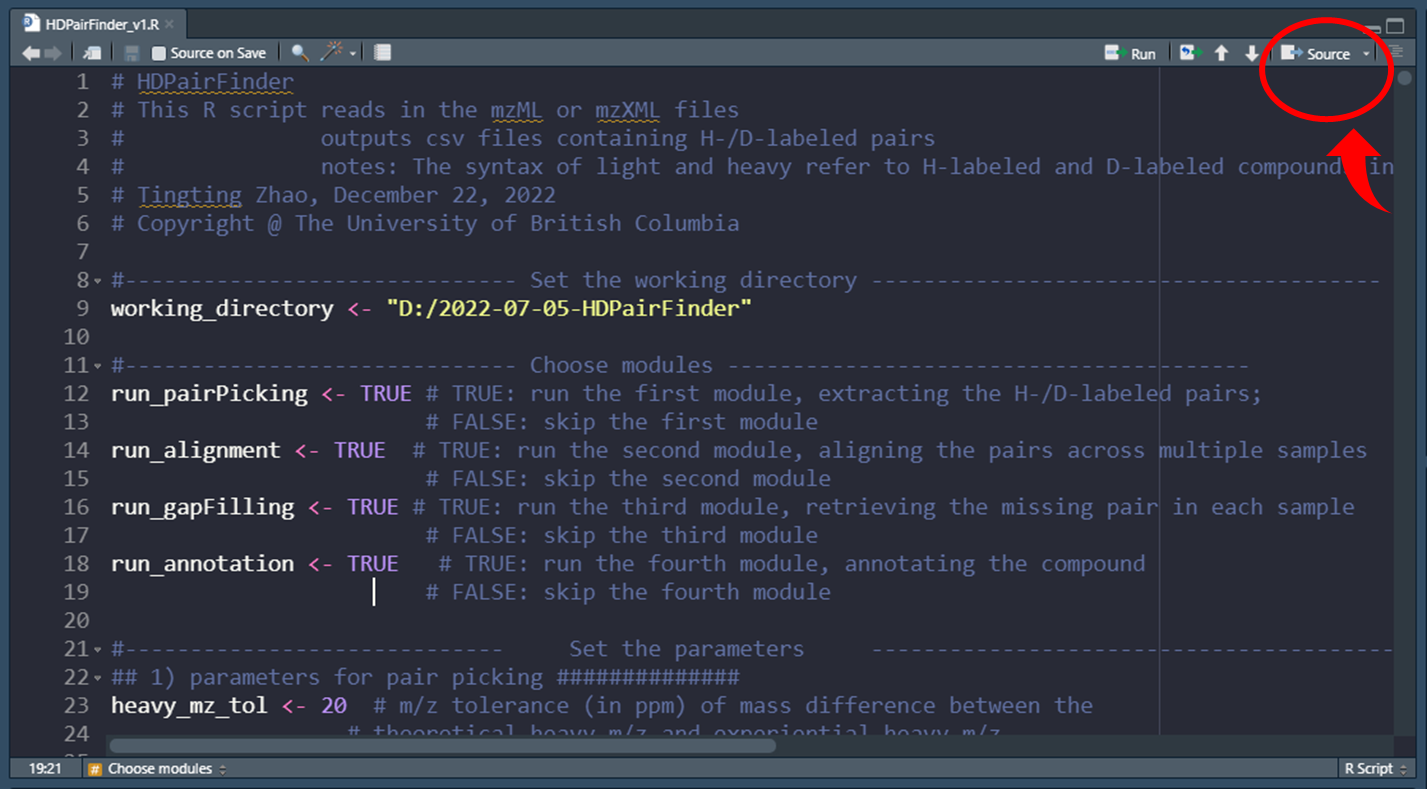


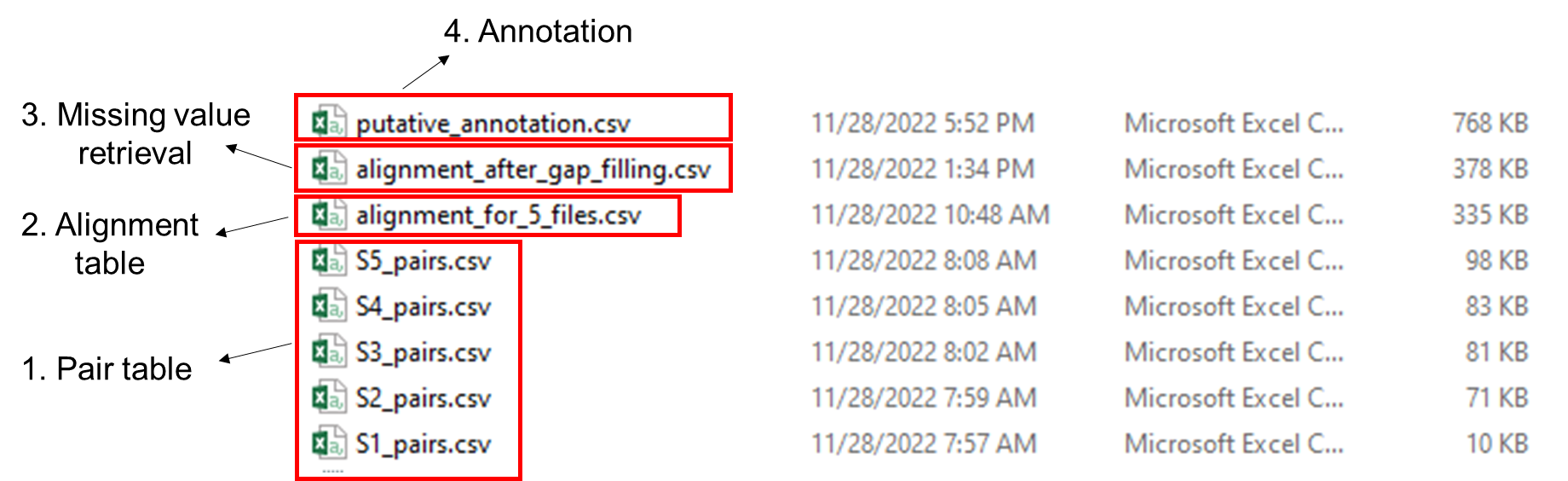
Table. Parameters of HDPairFinder.

|  |  |
| --- | --- |
| **Parameters** | **Function** |
| heavy\_mz\_tol | Numeric, relative mass tolerance (ppm) for the D-labeled compound in the extraction of H-/D-labeled pairs. Default: 20 |
| rt\_diff | Numeric, acceptable range of ΔRT (min) between the D-labeled compound and H-labeled compound. Default: -0.2 ~ 0.1 |
| int\_ratio | Numeric, acceptable range of intensity ratio between the D-labeled compound and H-labeled compound. Default: 0.4 ~ 1.4 |
| cc\_threshold | Numeric, cross-correlation threshold in the extraction of H-/D-labeled pairs. Default: 0.7 |
| run\_inSourceFrag | Logical, TRUE if data cleaning of the in-source fragments is performed, FALSE if not performed. |
| align\_mz\_tol | Numeric, relative mass tolerance (ppm) for the pairs in the alignment. Default: 50 |
| align\_rt\_tol | Numeric, RT tolerance (min) for the pairs in the alignment. Default: 0.5 |
| gap\_mz\_tol | Numeric, relative mass tolerance (ppm) for the pairs in the gap-filling. Default: 20 |
| gap\_rt\_tol | Numeric, RT tolerance (min) for the pairs in the gap-filling. Default: 0.5 |
| int\_threshold | Numeric, absolute intensity threshold of the H-labeled and D-labeled compounds in the gap-filling. Default: 1000 |
| anno\_mz\_tol | Numeric, relative mass tolerance (ppm) of parent compounds (before labeling) in the annotation. Default: 30 |

1. Run the R script by clicking “Source” on the top right of the RStudio panel.

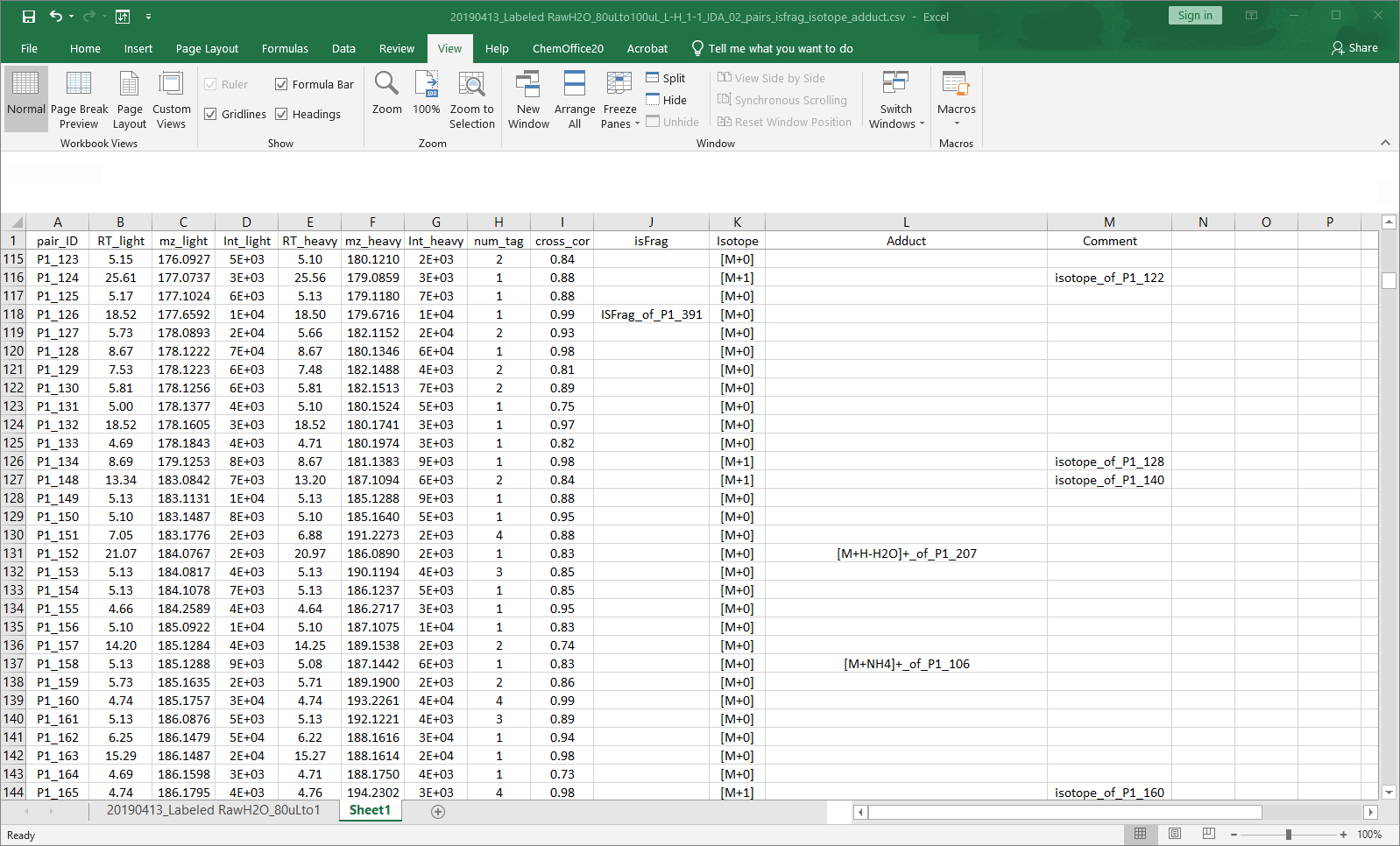


1. Four CSV files will be output in the working directory as shown below:



Detailed information about the above output files:

7.1 “S1.csv” to “S5.csv” represent the individual pair table for each sample. An example is shown below:

****

“pair\_ID”: refer the index of the pair.

“RT\_light”, “RT\_heavy”: retention times for H-labeled and D-labeled compounds in min.

“mz\_light”, “mz\_heavy”: *m/z* values for H-labeled and D-labeled compounds.

“Int\_light”, “Int\_heavy”: peak intensities for H-labeled and D-labeled compounds.

“num\_tag”: number of the methyl groups in the H-labeled and D-labeled compounds.

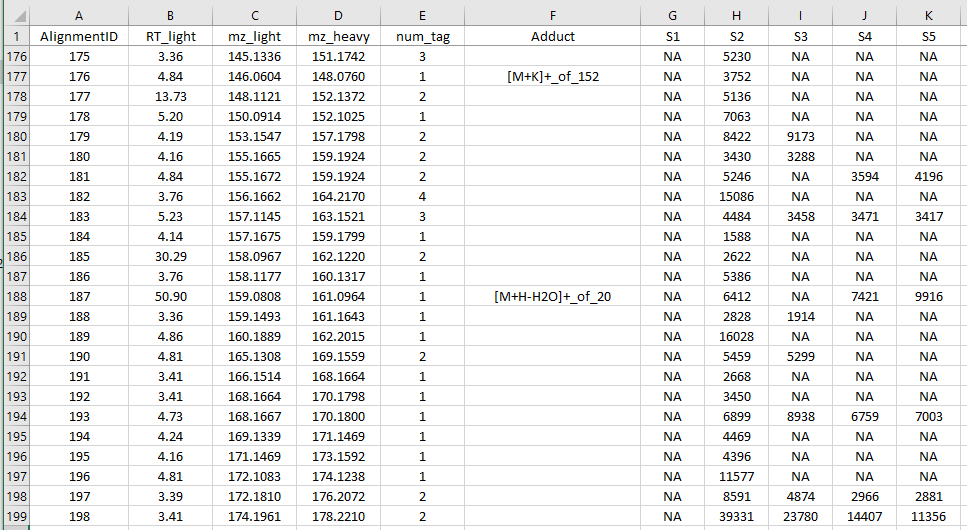
“cross\_cor”: peak-peak cross correlation between H-labeled and D-labeled compounds.

“IsFrag”: the in-source fragment form of the pair.

“Isotope”: the isotopic form of carbon. [M+0] for 13C0, [M+1] for 13C1, [M+2] for 13C2.

“Adduct”: the adduct form of the pair.

7.2 “alignment\_for\_5\_files.csv” represents the aligned pair table after. An example is shown below:



“AlignmentID”: index of the pair in the aligned pair table.

“RT\_light”: retention time of H-labeled peak for each pair.

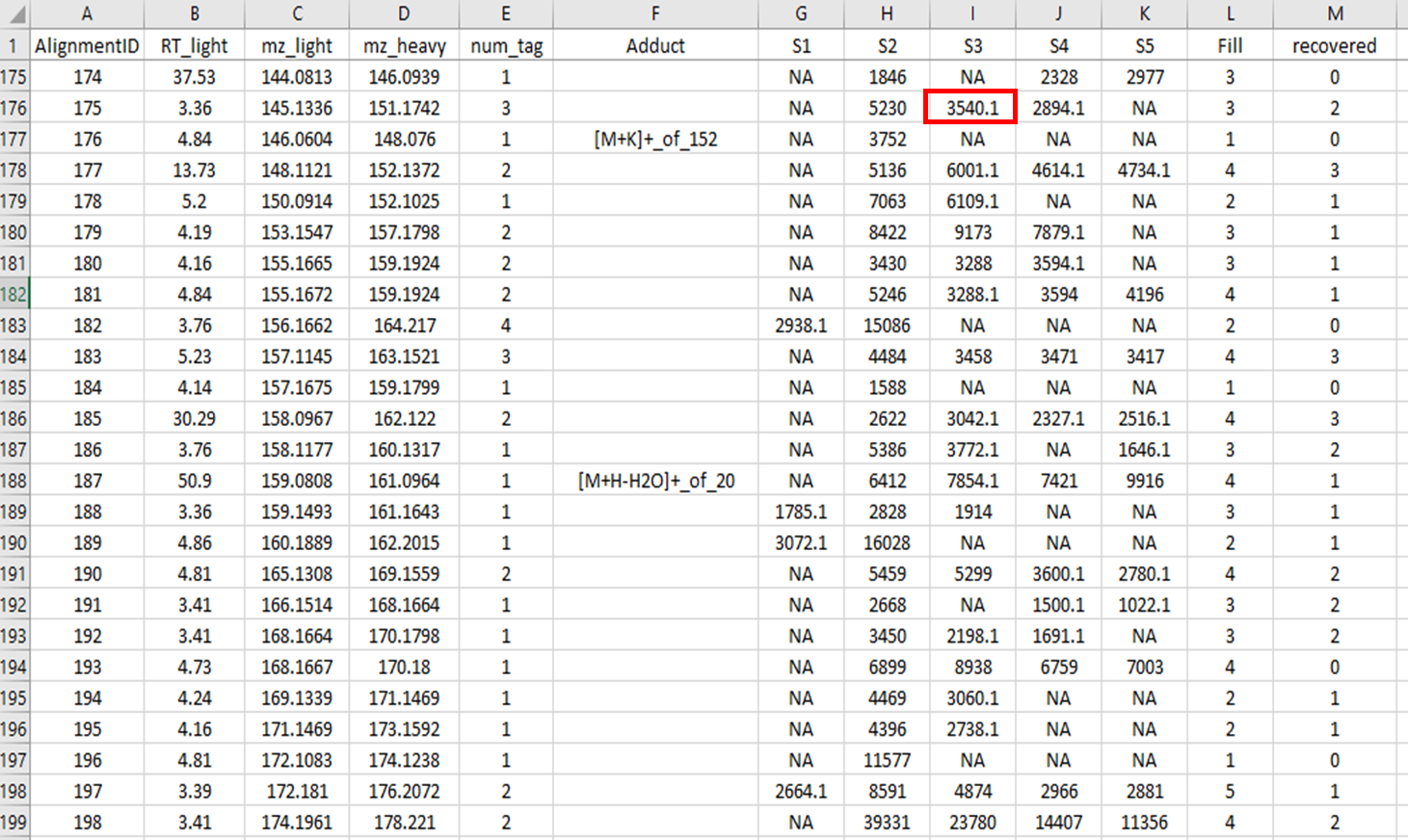
“mz\_light” and “mz\_heavy”: m/z value for H-labeled and D-labeled compounds

“num\_tag”: number of methyl groups attached to the labeled compounds.

“Adduct”: adduct form of the pair.

“S1” to “S5”: the intensity of the H-labeled compound for each sample.

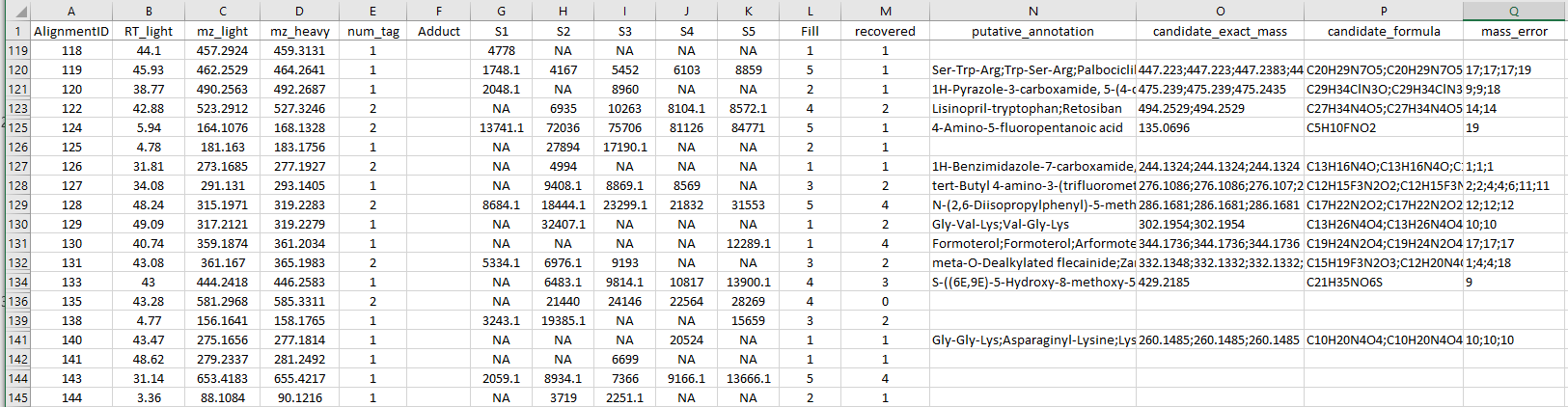
7.3 “alignment\_after\_gap\_filling.csv” represents the pair table after gap filling. An example is shown below:



“S1” to “S5”: the intensity of the H-labeled compound for each sample. Particularly, values with 1 decimal place (highlighted in red rectangle) represent the pairs recovered from the gap filling.

“Fill”: the number of each pair detected among all samples after gap-filling. Here, “3” means that this pair is detected in 3 samples among 5 samples.

7.4 “putative\_annotation.csv” represents the annotation results with the amino-containing database. An example is shown below:



“putative annotation”: the chemical names of the candidate based on the user-defined mass tolerance. If there are multiple hits in the library, they will be ranked by mass and the top 10 will be reported.

“candidate\_exact\_mass”: exact mass of the candidates.

“candidate\_formula”: the molecular formula of the candidates.

“mass\_error”: the relative mass error between the amino-containing compounds and the candidates.