Data-Dependent-Assisted Data-Independent Acquisition (DaDIA.R) User Manual

(Version 1, 2020-11-23)

Jian Guo[1], Sam Shen[1], Shipei Xing[1], Tao Huan[1,*]

[1] Department of Chemistry, Faculty of Science, University of British Columbia, Vancouver Campus, 2036 Main Mall, Vancouver, V6T 1Z1, BC, Canada

[*] Author to whom correspondence should be addressed:

Dr. Tao Huan
Tel: (+1)-604-822-4891
E-mail: thuan@chem.ubc.ca
Internet: https://huan.chem.ubc.ca/

- DaDIA.R is an R script for performing DaDIA workflow of metabolic feature extraction and annotation.
- The program is written in the language 'R' and is publicly available at https://github.com/HuanLab/DaDIA.git
- Please see below for detailed instructions on using the DaDIA.R code:

1) **File preparation.** User needs to create two folders to store DDA and DIA files separately. All mzXML files from DDA analyses need to be put in the DDA folder. All mzXML files from DIA analyses need to be put in the DIA folder. The library file in the format of .msp should also be put in the DIA folder. In addition, a .txt file containing the information about the DIA *m/z* range should also be put into the DIA folder. **Figure 1** illustrates the details of how the files should be organized in the corresponding folders. The values in the .txt file are separated by tab. If it is DIA(SWATH) data, the txt file should contain the information about the *m/z* range for the survey scan and SWATH windows. If it is DIA(AIF) data, the txt file should contain the information about the *m/z* range for the survey scan and AIF window. The *m/z* range file examples for both DIA(SWATH) and DIA(AIF) are illustrated in **Figure 2**. Note that the column headers should be kept the same as the examples shown.

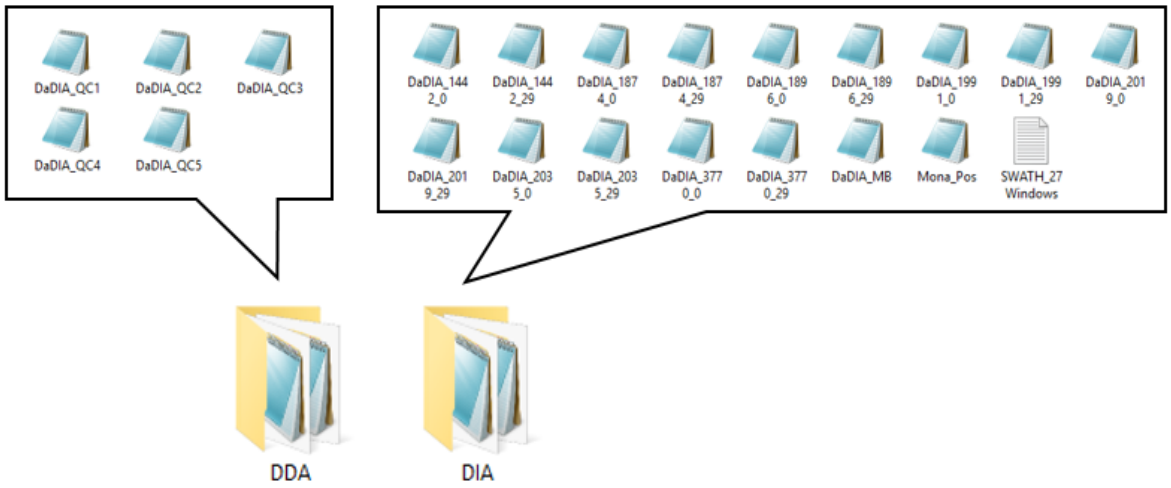

Figure 1. *m/z* range file sample format for DIA(SWATH) (left) and DIA(AIF) (right) files.

DIA_SWATH_pocket.... — Edited

Experiment	MSType	Minmz	Maxmz
0	SCAN	100	1000
1	MSMS	100	126
2	MSMS	125	151
3	MSMS	150	176
4	MSMS	175	201
5	MSMS	200	226
6	MSMS	225	251
7	MSMS	250	276
8	MSMS	275	301
9	MSMS	300	326
10	MSMS	325	351
11	MSMS	350	376
12	MSMS	375	401
13	MSMS	400	426
14	MSMS	425	451
15	MSMS	450	476
16	MSMS	475	501
17	MSMS	500	526
18	MSMS	525	551
19	MSMS	550	576
20	MSMS	575	601
21	MSMS	600	626
22	MSMS	625	651
23	MSMS	650	676
24	MSMS	675	701
25	MSMS	700	801
26	MSMS	800	901
27	MSMS	900	1000

DIA_AIF_pocket.txt — Edited

Experiment	MSType	Minmz	Maxmz
0	SCAN	100	1000
1	DIA	100	1000

Figure 2. *m/z* range file sample format for DIA(SWATH) (left) and DIA(AIF) (right) files.

2) Download the R-scrip "DaDIA.R" from Github (https://github.com/HuanLab/DaDIA.git).
3) **R package installation.** In R-studio, user needs to first install libraries "xcms", "MSnbase", "dplyr", "doParallel", "foreach", "metaMS", and "CAMERA" if they are previously not installed. (R Version 4.0 or above, XCMS Version 3.11.4, and metaMS Version 1.25.1 are required; all other packages should be updated to the newest available version)
4) **Parameter setting.** After all the required libraries are successfully installed. User needs to set the parameters to their desired values. All the parameters available for customized setting are in line 17 – 60, as shown in **Figure 3**. The function of each parameter is described in **Table 1**.

```
16 ############################################################
17 #Part 1: Parameters for feature extraction
18 DDA.directory <- "E:/DaDIA_DDA"
19 DIA.directory <- "E:/DaDIA_DIA"
20 cwpDDA <- CentWaveParam(ppm=10,
21                         peakwidth=c(5,60),
22                         mzdiff = 0.01,
23                         snthresh = 6,
24                         integrate = 1,
25                         prefilter = c(3,100),
26                         noise = 100) #XCMS parameters for DDA feature extraction
27 cwpDIA <- CentWaveParam(ppm=10,
28                         peakwidth=c(5,60),
29                         mzdiff = 0.01,
30                         snthresh = 6,
31                         integrate = 1,
32                         prefilter = c(3,100),
33                         noise = 100) #XCMS parameters for DIA feature extraction
34 mass.tol <- 10 #mz tolerance in ppm: used in feature dereplication and MS2 matching
35 mass.const.tol <- 0.05 #mz tolerance in constant value: used in feature rescue
36 rt.tol <- 60 #rt tolerance in seconds
37 num.samples <- 3 #enter how many DIA samples here
38 plot.DaDIA <- TRUE #plot DaDIA features
39 plot.DaDIA.mztol <- 0.5 #DaDIA feature plotting mz window width
40 plot.DaDIA.rttol <- 30 #DaDIA feature plotting rt window width
41 #Parameters for alignment
42 bw <- 5
43 minfrac <- 0.5
44 mzwid <- 0.015
45 max <- 100
46 quantitative.method <- "maxo"
47 # "maxo" = peak height
48 # "into" = peak area
49 ############################################################
50 #Part 2: Parameters for database search (dot product)
51 feature.annotation <- TRUE #annotate DaDIA features
52 db.name <- "convertedLibraryPos.msp" #annotation library name
53 ms1.tol <- 0.01 #dot product calculation ms1 tolerance
54 ms2.tol <- 0.02 #dot product calculation ms2 tolerance
55 dot.product.threshold <- 0.1 #dot product annotation threshold
56 match.number.threshold <- 1 #annotation match number threshold
57 adduct_isotope.annotation <- TRUE #perform CAMERA annotation
58 export.mgf <- TRUE #export individual MS2 spectra as .mgf
59 combine.mgf <- TRUE #combine all exported .mgf files
60 MS2mirrorplot <- TRUE #plot mirror plots for features with dot product larger than dot product threshold
61 ############################################################
```

Figure 3. Parameter settings of DaDIA.R.

Table 1. The functions of all DaDIA parameters.

Line #	Parameter Name	Parameter Function
18	*DDA.directory*	Set the directory containing all DDA .mzxml files
19	*DIA.directory*	Set the directory containing all DIA .mzxml files, *m/z* window .txt file (no specific name is required as the program recognizes it by its file type), and annotation library .msp file
20	*cwpDDA*	Set XCMS parameters for DDA feature extraction
27	*cwpDIA*	Set XCMS parameters for DIA feature extraction

34	*mass.tol*	Set *m/z* tolerance (± ppm) for MS^1 feature dereplication and MS^2 matching
35	*mass.const.tol*	Set *m/z* tolerance (± constant value) for rescuing DIA features using DDA data
36	*rt.tol*	Set retention time tolerance (± sec) for identifying the same features
37	*num.samples*	Set number of DIA samples to run
38	*plot.DaDIA*	Set whether to plot EIC for DaDIA features
39	*plot.DaDIA.mztol*	Set *m/z* window width for DaDIA feature EIC plotting
40	*plot.DaDIA.rttol*	Set RT window width for DaDIA feature EIC plotting
42	*bw*	Set XCMS feature alignment bandwidth
43	*minfrac*	Set XCMS feature alignment minimum sample fraction
44	*mzwid*	Set XCMS feature alignment *m/z* slice width
45	*max*	Set XCMS feature alignment maximum # of groups / slice
46	*quantitative.method*	Set whether to use peak height or peak area for quantitative calculations
51	*feature.annotation*	Set whether to perform MS^2 extraction and DaDIA feature annotation
52	*db.name*	Set the name of the library used for metabolite annotation
53	*ms1.tol*	Set MS^1 tolerance in dot product calculation for metabolite annotation
54	*ms2.tol*	Set MS^2 tolerance in dot product calculation for metabolite annotation
55	*dot.product.threshold*	Set annotation dot product score threshold
56	*match.number.threshold*	Set annotation match number threshold
57	*adduct_isotope.annotation*	Set whether to perform CAMERA adduct and isotope annotation
58	*export.mgf*	Set whether to export MS^2 spectra as individual .mgf files
59	*combine.mgf*	Set whether to concatenate all exported .mgf files into a single .mgf file
60	*MS2mirrorplot*	Set whether to plot MS^2 mirror plot

5) Note: user needs to set the directory in the user's computer that contains all DDA samples in line 18. User needs to set the directory in the user's computer that contains all DIA samples, *m/z* range definitions in .txt format (for DIA(SWATH) or DIA(AIF)), and the annotation library in .msp format in line 19.
6) In R-studio, click on "→Source" in the top right corner of the R-studio interface to begin the DaDIA data processing.
7) After running the scrip for single DDA and single DIA sample, one csv file "DaDIAtable.csv" containing all metabolic features extracted and one csv file

"annotated_output.csv" containing all feature annotation results will be generated in the DIA folder. After running the scrip for multiple DDA and DIA samples, multiple csv files "n_DaDIAtable.csv" (n is the number of DIA samples) containing all metabolic features extracted for each sample, one csv file "alignedDaDIAtable.csv" containing aligned features, and one csv file "annotated_output.csv" containing the annotation results for the aligned features will be generated in the DIA folder.

8) Notably, in "annotated_output.csv" file, the columns with the header "MS2-Available" contains either TRUE or FALSE values. TRUE means there are MS^2 spectra assigned to the features, while FALSE means there are no MS^2 spectra assigned. If the user performs "CAMERA", there will be three additional columns shown up in the file showing the isotopic, adduct, and pcgroup information. The features with the same number in "pcgroup" are actually the same metabolite as they are highly correlated peaks.

Specific Notes

a) Note: if users wish to use their own in-house library in .csv format for annotation, they must first convert their library from .csv file to an .msp file using the R script *"*convertMSP.R*"* at the provided website on GitHub (https://github.com/HuanLab/DaDIA.git).

b) Note: user can choose to plot the EIC of all the metabolic features by switching on the plot function in line 38 and setting the *m/z* and retention time tolerance in line 39, 40. If the user chooses to plot DaDIA features, a folder named "DaDIA_EIC" will be generated in the DIA folder containing all the EIC plots. The name of the EIC plots is composed of feature retention time and *m/z* values.

c) Note: user can choose to output individual .mgf files for each feature by switching on the function in line 58. Two folders named "DDAmgf" and "DIAmgf" will be generated in the DIA folder containing all individual mgf files from either DDA or DIA data. If the user wishes to output one additional mgf file combining all the MS^2 information into one file, they can switch on the function in line 59. An mgf file named "combined_mgf" will be generated in the DIA folder. The names of all the mgf files are composed of precursor mass, retention time and the source of the MS^2 spectrum. (whether it is from DDA or DIA data)

d) Note: user can choose to generate the MS^2 mirror plot by switching on the function in line 60. A folder named "MS2mirrorplot" will be generated in the DIA folder containing all the mirror plots in .png file for the features with dot products larger than dot product threshold set by user. In each mirror plot, the corresponding metabolite name is shown on the top of the plot. The name of each plot file is composed of the feature ID and the dot product of the feature.