Metabolomics spectral library building and data extraction with MetaboDIA

Gengbo Chen, Hyungwon Choi

22 May 2017

1. Introduction

The R package MetaboDIA performs a new workflow for building spectral assay libraries from mass spectrometry (MS) data acquired in both data-dependent acquisition (DDA) and data-independent acquisition (DIA) modes for metabolomics application, and performing targeted extraction of quantitative MS/MS data in DIA-MS. MetaboDIA has two major workflows as illustrated by Figure 1 below.

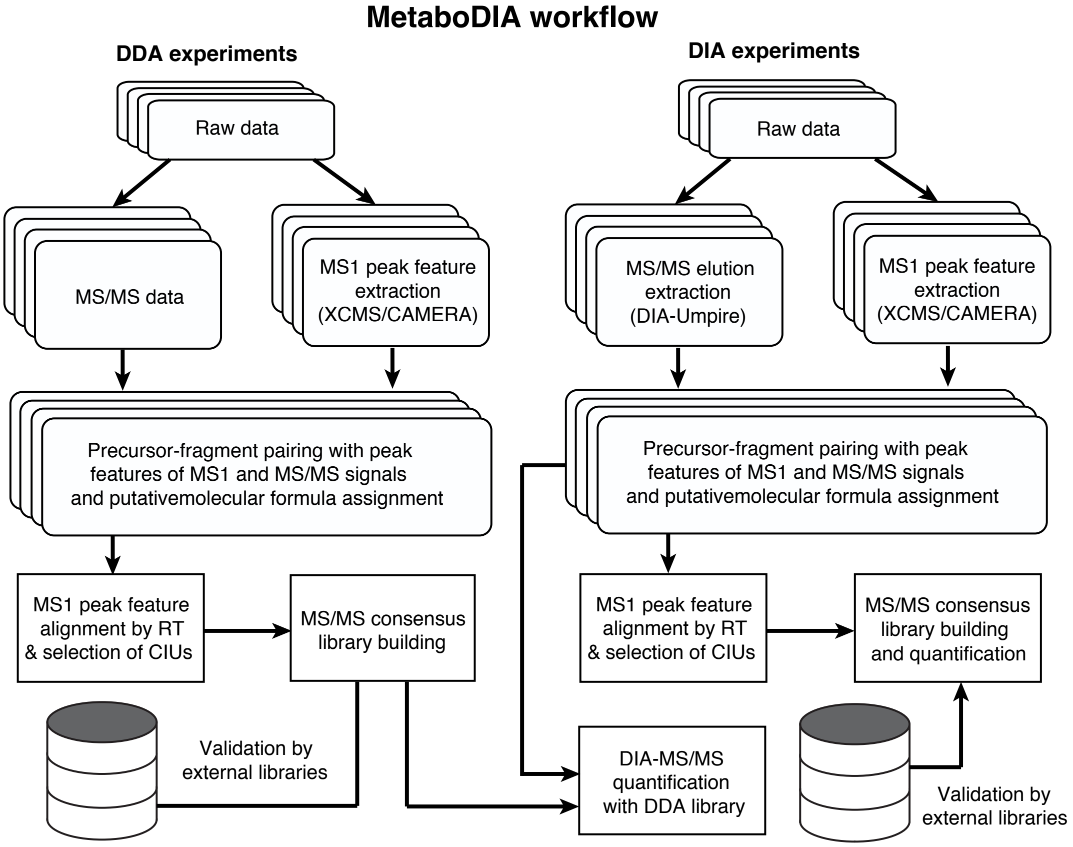
MetaboDIA workflow
DDA experiments
DIA experiments
Raw data
MS/MS data
MS1 peak feature extraction (XCMS/CAMERA)
Raw data
MS/MS elution extraction (DIA-Umpire)
MS1 peak feature extraction (XCMS/CAMERA)
Precursor-fragment pairing with peak features of MS1 and MS/MS signals and putativemolecular formula assignment
Precursor-fragment pairing with peak features of MS1 and MS/MS signals and putativemolecular formula assignment
MS1 peak feature alignment by RT & selection of CIUs
MS/MS consensus library building
MS1 peak feature alignment by RT & selection of CIUs
MS/MS consensus library building and quantification
Validation by external libraries
DIA-MS/MS quantification with DDA library
Validation by external libraries

Figure 1: Two workflows of MetaboDIA. In the default pipeline, a consensus MS/MS spectral library is built using DDA data (left) and quantification of MS/MS fragments is performed using DIA data. Alternatively, the library can be directly constructed from the DIA data (right).

2 Download instructions and other requirements

The Package is written in R programming language. Please download and install R program at https://www.r-project.org according to the instructions. The MetaboDIA package available at https://sourceforge.

net/projects/metabodia/. User can install the package using the following command:

```
install.packages("PATH_TO_THE_PACKAGE", repos = NULL, type = "source")
```

In addition to the R package, the molecular formula database and the example adduct files are also available in .csv format. We also provide the consensus libraries described in the manuscript.

The package requires the R package *xcms*[1] and *CAMERA*[2] for data pre-processing. Please download and install it at https://bioconductor.org/packages/release/bioc/html/CAMERA.html before installing the package, or just use the following R command to install:

```
## Try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("CAMERA")
```

Parallel computing is a very powerful technique to reduce the run time especially for large data set. The package uses *mclapply* function from *parallel* package for parallel computing. The user can use the following command in R to know how many cores are available in the computer and set the parameter *n_core* accordingly.

```
detectCores(all.tests = FALSE, logical = TRUE)
```

The raw MS files are converted to mzXML format in the centroid mode using MSconvert, which is a part of ProteoWizard (http://proteowizard.sourceforge.net). Wiff files from AB_SCIEX instruments need to be converted to mzML format in the centroid mode first using the AB_SCIEX data converter software (http://sciex.com/software-downloads-x2110).

The Age-Related Macular Degeneration (AMD) date set (converted to .mzXML format) described in the manuscript is available in MetaboLights repository. http://www.ebi.ac.uk/metabolights/MTBLS417.

The extraction module from *DIA_Umpire*[3] is used for library-free MS/MS data extraction of DIA samples. User can download the install it at http://diaumpire.sourceforge.net. Please follow its user manual for the signal extraction part. Please also be reminded that the example parameters need to be properly adjusted based on the instrument type and experimental settings.

3 Input data preparation for metaboDIA

3.1 MS/MS data extraction form DIA samples

The signal extraction module from *DIA-Umpire* is used with a set of proper parameters (see Supplementary Information in the paper). In the extraction results, we use Q1 (more than two isotopes in precursor) and Q2 (only two isotopic peaks in precursor) results while we discard Q3 (without precursor peaks). At this point, the intensity value for each MS/MS peak is peak area of the XIC.

3.2 Input files preperation

To begin with the pipeline, several files should be provided for molecular formula identification:

Database file: The database of molecular formula and its molecular weight. An example Database file is available at sourceforge (*Database.txt*).

Adduct file: The file indicating which adduct/adducts should be considered in the molecular identification step. Example adduct files in both positive and negative modes are also available at sourceforge. (*Adducts_pos.csv and Adducts_neg.csv*)

The package also requires a simple file structure for the ease of accessing the files:

DDA samples: Move all the .mzXML file from DDA sample into one folder together with Database file and Adduct file.

DIA samples: Move all the .mzXML and the output results (Q1.mgf and Q2.mgf) from *DIA-Umpire* into one folder together with Database file and Adduct file.

All the intermediate and end results files are stored in the corresponding folders.

3.3 MS1 peak feature extraction with isotope/ion annotation

Next we perform peak feature extraction in each sample using *xcms* [1] and feed the result into CAMERA [2] package to annotate the isotopic peaks. Note that Patti *et al* suggest adjusting the *xcms* parameters for different instruments [4]. It is also necessary to export the annotation result in csv format with the extension ".cam.csv" (e.g. **SAMPLE_NAME.cam.csv**) for downstream analysis.

```
## Example for isotope annotation
xs <- xcmsSet("DDA1.mzXML",method="centwave",ppm=30,peakwidth=c(5,60),prefilter=c(2,50))
an <- xsAnnotate(xs,polarity = "positive")
anF <- groupFWHM(an)
anI <- findIsotopes(anF,ppm=30)
anIC <- groupCorr(anI, cor_eic_th = 0.6)
peaklist <- getPeaklist(anIC)
write.csv(peaklist, file="DDA1.cam.csv")
```

3.4 Molecular formula identification

In this workflow, the charge state of the precursor ios is deconvoluted by either the SM/MS information or the isotopic peaks clusters from *CAMERA*. With the consideration of different adduct types from the input file *Adduct file* Then we searched charge and adduct-deconvoluted MS1 features against the database *Database file* with ±30 ppm (user defined) threshold in each sample and recorded all the putative molecular formulae. Then the results are exported for downstream analysis (e.g. **SAMPLE_NAME.cam.metab.csv**). The unidentified elution profiles are discarded. In **Section 3**, the molecular formula identification is based on the isotope peaks information from *CAMERA* package.

We also provide a wrapper function to perform the isotope annotation and molecular formula identification for multiple files in a directory. The function processes multiple samples in parallel and run the *CAMERA* and molecular formula identification automatically.

```
## Wrapper function to perform isotope annotation and molecular formula identificaiton
runCAMERA.DBsearch(dir="PATH_TO_YOUR_FILES", DB.file="Database.txt",
                   adduct.file="Adduct.csv",n_core=4,prefilter=c(2,50),
                   mode="positive", ppm=30, method="centWave",
                   peakwidth=c(5,60), cor_eic_th = 0.6)
```

If the user has already run the isotope annotation using *CAMERA*, the user can also run the molecular formula identification with the following wrapper function:

```
## Wrapper function to perform formula identification
runDBsearch(dir="PATH_TO_YOUR_FILES", DB.file="Database.txt",
            adduct.file="Adduct.csv",mode="positive", n_core=4)
```

These wrapper functions were written with *mclapply* function for parallel processing of *n_core* samples.

3.5 Files generated in the above steps

After the input data preparation steps for **Sample1.mzXML**, the following files are generated for downstream processing:

- DIA MS/MS extraction: **Sample1_Q1.mgf** and **Sample1_Q2.mgf**
- CAMERA output: **Sample1.cam.csv**
- MS1 peak features with isotope annotation: **Sample1.cam.iso.csv**
- MS1 peak features without isotope annotation: **Sample1.cam.mono.csv**
- Molecular formula identification result: **Sample1.cam.metab.csv**

4 Consensus library building and data extraction

Once the data processing is finished for each sample, the program will generate a tab delimited linkage file (file name: "file_matching.txt") to link the outputs from the above data processing steps when we run the different workflows. As a result, the file name of the outputs generated in the above data processing steps **SHOULD NOT** be changed and the linkage file is generated in the following format:

DDA samples

Sample	Elution	DBsearch	MS2
DDA1	DDA1.cam.csv	DDA1.cam.metab.csv	DDA1.mzXML
DDA2	DDA2.cam.csv	DDA2.cam.metab.csv	DDA2.mzXML
DDA3	DDA3.cam.csv	DDA3.cam.metab.csv	DDA3.mzXML
DDA4	DDA4.cam.csv	DDA4.cam.metab.csv	DDA4.mzXML
DDA5	DDA5.cam.csv	DDA5.cam.metab.csv	DDA5.mzXML
DDA6	DDA6.cam.csv	DDA6.cam.metab.csv	DDA6.mzXML
DDA7	DDA7.cam.csv	DDA7.cam.metab.csv	DDA7.mzXML
DDA8	DDA8.cam.csv	DDA8.cam.metab.csv	DDA8.mzXML
DDA9	DDA9.cam.csv	DDA9.cam.metab.csv	DDA9.mzXML
DDA10	DDA10.cam.csv	DDA10.cam.metab.csv	DDA10.mzXML

DIA samples

Sample	Elution	DBsearch	MS2_Q1	MS2_Q2
DIA1	DIA1.cam.csv	DIA1.cam.metab.csv	DIA1_Q1.mgf	DIA1_Q2.mgf
DIA2	DIA2.cam.csv	DIA2.cam.metab.csv	DIA2_Q1.mgf	DIA2_Q2.mgf
DIA3	DIA3.cam.csv	DIA3.cam.metab.csv	DIA3_Q1.mgf	DIA3_Q2.mgf
DIA4	DIA4.cam.csv	DIA4.cam.metab.csv	DIA4_Q1.mgf	DIA4_Q2.mgf
DIA5	DIA5.cam.csv	DIA5.cam.metab.csv	DIA5_Q1.mgf	DIA5_Q2.mgf
DIA6	DIA6.cam.csv	DIA6.cam.metab.csv	DIA6_Q1.mgf	DIA6_Q2.mgf
DIA7	DIA7.cam.csv	DIA7.cam.metab.csv	DIA7_Q1.mgf	DIA7_Q2.mgf
DIA8	DIA8.cam.csv	DIA8.cam.metab.csv	DIA8_Q1.mgf	DIA8_Q2.mgf
DIA9	DIA9.cam.csv	DIA9.cam.metab.csv	DIA9_Q1.mgf	DIA9_Q2.mgf
DIA10	DIA10.cam.csv	DIA10.cam.metab.csv	DIA10_Q1.mgf	DIA10_Q2.mgf

4.1 MS1 quantification for DDA and DIA data

This workflow is part of the default workflow (**Section 4.2** below). It is used when the MS/MS data is not available. It process and exports not only the MS1 quantification results of identified CIUs but also the number of isotope peaks in the peak feature of each sample.

```
## MS1 level quantification
MS1_extraction(file_dir="PATH_TO_FILES",n_core=4)
```

4.2 DDA-based library building followed by DIA MS/MS data extraction

In this default workflow, we build a consensus spectral library using DDA samples and use the library to extract MS/MS quantification data from DIA samples. In this workflow, the DIA MS/MS quantification result, the DDA MS1 quantification result and DDA-based library are exported to separate files.

```
## DDA-based library and DIA extraction workflow (default)
DDA_DIA_workflow(DIA_dir="PATH_TO_DIA_FILES",DDA_dir="PATH_TO_DDA_FILES",
                 n_core=4,DB.file="Database.txt",adduct.file="Adduct.csv",
                 mode="positive or negative")
```

If the DIA runs are not available, we also provide another function to build DDA-based consensus library and extract DDA MS1 quantification result, without the DIA MS/MS data extraction step.

```
## DDA consensus library building
DDA_library_building(DDA_dir="PATH_TO_DDA_FILES",DB.file="Database.txt",
              adduct.file="Adduct.csv",n_core=4,mode="positive or negative")
```

4.3 Direct DIA-based library building followed by DIA MS/MS data extraction

This alternative workflow is designed for DIA-MS only analysis without the paired DDA samples. In this workflow, we build a consensus spectral library using DIA samples and extract MS/MS quantification data from DIA samples again – this is similar to the library-free extraction of DIA-Umpire. Here, the MS/MS quantification results are generated simultaneously with the library construction. MS/MS quantification results and the DIA-based library are exported to separate files.

```
## DIA-based library and DIA extraction workflow (alternative)
DIA_DIA_workflow(DIA_dir="PATH_TO_DIA_FILES",DB.file="Database.txt",
                 adduct.file="Adduct.csv",n_core=4,mode="positive or negative")
```

References

[1] Smith, C.A., et al., XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem, 2006. 78(3): p. 779-87.

[2] Kuhl, C., et al., CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Anal Chem, 2012. 84(1): p. 283-9.

[3] Tsou, C.C., et al., DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat Methods, 2015. 12(3): p. 258-64, 7 p following 264

[4] Patti, G.J., R. Tautenhahn, and G. Siuzdak, Meta-analysis of untargeted metabolomic data from multiple profiling experiments. Nat Protoc, 2012. 7(3): p. 508-16.