# Analyzing LC-MS data with xcms

Xue Chen

April 10, 2016

**Introduction**

Xcms is an R package designed to process chromatography-mass spectrometry data generated from paired experimental conditions for comparative and statistical analysis. Xcms is useful to identify metabolites present at significantly different levels across experimental conditions. In this study, xcms will be used to analyze opium poppy latex metabolites-benzylisoqunoline alkaloids (BIAs) data generated from liquid chromatography- mass spectrometry system (LC-MS).

**Background**

To study the function of a gene encoding an *O*-methyltransferase (OMT2) in opium poppy noscapine (a type of BIAs) biosynthesis pathway, two groups of opium poppy plants were treated with Virus-Induced Gene silencing (VIGs). The first group of plants would have the endogenous OMT2 gene being down-regulated using VIGs, and thus was designated as *omt2* plants. The second group was used as a negative control, in which no gene will be affected after VIGs treatment and designated as EV (empty vector) plants. Quantitative real-time PCR was used to confirm the OMT2 gene expression difference between the two groups of plants. LC-MS was used to obtain the alkaloid profiles of these plants. The conventional way to analyze the LC-MS results is to hand pick and integrate the peaks corresponding to each desired alkaloid species based on their m/z ratios. Although this is time consuming, this approach allows absolute quantification if having appropriate standard curves and then relative comparison of alkaloids accumulation levels between the two types of samples. Xcms does not provide the power of absolute quantification; rather it performs relative comparison and statistical analysis. Thus xcms offers a robust way to identify potential BIA metabolites (including both known and unknown ones) that are present at significantly different levels between the *omt2* and EV plants.

**Processing LC-MS data with xcms**

***Data file preparation***

To be able to perform a comparative analysis across the sample classes using xcms, LC-MS data have to be acquired using the same LC-MS method and separation column. Ideally, the samples should be acquired within one run. Xcms accepts LC-MS data in the format form of AIA/ANDI, including NetCDF, mzXML and mzData files. In the present study, all LC-MS data were first converted to mzData using the software of Agilent Masshunter Qualitative Analysis.

Another important consideration is the directory structure in which the files are organized. Xcms automatically assign separate classes to samples based on their storage locations. For example, *omt2* data stored into a subfolder named OMT2 will be classified as OMT2, and EV data in a subfolder named EV will be considered as EV class. Xcms uses sample class information during preprocessing to help decide which groups of peaks (metabolites) are significantly different between the two classes. In addition, within each of those directories, samples could be further classified by another experimental factor, such as the day the samples were collected. In this case, the data will be classified into “OMT2/Day1”, “OMT2/Day2”, etc. In the present study, the data were classified by one experimental factor, i.e. VIGs; thus, two classes of data (OMT2 vs EV) were compared. All LC-MS data were stored under two folders, i.e., “/vigs mzDATA/OMT2” and “/vigs mzDATA/EV”. Two subfolders-“OMT2” and “EV” were set within this folder, each containing 17 and 23 mzDATA files, respectively.

***Installing xcms***

>source("http://bioconductor.org/biocLite.R")

>biocLite("xcms", dep=T)

>library(xcms)

***Importing data***

> setwd("/Users/XC/Desktop/LCMS data/vigs mzDATA")

> files<-dir(recursive=T, full.names = TRUE)

> files ### 41 mzDATA files will be analyzed ###

[1] "./13.tsv" "./20.tsv" "./EV/ev-10.mzdata.xml"

[4] "./EV/ev-11.mzdata.xml" "./EV/ev-12.mzdata.xml" "./EV/ev-13.mzdata.xml"

[7] "./EV/ev-14.mzdata.xml" "./EV/ev-16.mzdata.xml" "./EV/ev-18.mzdata.xml"

[10] "./EV/ev-19.mzdata.xml" "./EV/ev-1b.mzdata.xml" "./EV/ev-21.mzdata.xml"

[13] "./EV/ev-22.mzdata.xml" "./EV/ev-23.mzdata.xml" "./EV/ev-24.mzdata.xml"

[16] "./EV/ev-25.mzdata.xml" "./EV/ev-27.mzdata.xml" "./EV/ev-28.mzdata.xml"

[19] "./EV/ev-29.mzdata.xml" "./EV/ev-3.mzdata.xml" "./EV/ev-30.mzdata.xml"

[22] "./EV/ev-4.mzdata.xml" "./EV/ev-7.mzdata.xml" "./EV/ev-8.mzdata.xml"

[25] "./EV/ev-9.mzdata.xml" "./OMT2/omt2-1.mzdata.xml" "./OMT2/omt2-10.mzdata.xml"

[28] "./OMT2/omt2-12.mzdata.xml" "./OMT2/omt2-14.mzdata.xml" "./OMT2/omt2-17.mzdata.xml"

[31] "./OMT2/omt2-2.mzdata.xml" "./OMT2/omt2-21.mzdata.xml" "./OMT2/omt2-22.mzdata.xml"

[34] "./OMT2/omt2-23.mzdata.xml" "./OMT2/omt2-24.mzdata.xml" "./OMT2/omt2-26.mzdata.xml"

[37] "./OMT2/omt2-27.mzdata.xml" "./OMT2/omt2-28.mzdata.xml" "./OMT2/omt2-29.mzdata.xml"

[40] "./OMT2/omt2-4.mzdata.xml" "./OMT2/omt2-6.mzdata.xml" "./OMT2/omt2-7.mzdata.xml"

[43] "./OMT2/omt2-8.mzdata.xml"

***Filtration and peak identification***

The “xcmsSet” function identifies and stores the peaks. It also provides methods for grouping and aligning the peaks. Part of the results from this function was shown as below:

> xset<-xcmsSet()

ev-10: 250:36 300:97 350:212 400:373 450:568 500:739 550:959 600:1153 650:1352

ev-11:250:32 300:89 350:208 400:361 450:557 500:715 550:930 600:1125 650:1359

ev-12:250:30 300:91 350:204 400:360 450:568 500:730 550:936 600:1113 650:1311

ev-13:250:39 300:94 350:215 400:371 450:569 500:735 550:954 600:1142 650:1357

The output is composed of pairs of numbers separated by a colon. The first number is the m/z it processed. The second number is the number of peaks.

***Matching peaks across samples***

After peak identification, peaks between samples that were identified as the same analyte (with the same mass/charge (m/z) value and similar retention time (rt)) will be matched into a group.

>xset<-group(xset)

263 325 388 450 513 575 638

***Retention time correction***

After matching peaks into groups, xcms uses those groups to identify and correct correlated drifts in retention time from sample to sample. Then a second round of peak grouping has to be performed with the newly aligned peaks. This round of grouping will provide a more accurate result than the first time. To be noted that this process can be repeated in an iterative fashion.

>xset2 <- retcor(xset)

Retention Time Correction Groups: 92

> xset2<-group(xset2)

263 325 388 450 513 575 638

***Filling in missing peaks***

Missing peaks can happen when peaks were missed during peak identification or certain analytes were not present in certain samples. In any case, those missing data points was set to NA and now will be filled with random noise based on the raw data files.

> xset3 <- fillPeaks(xset2)

***Analyzing and visualizing results***

The method *diffreport* will generate a report showing all the identified peaks with the corresponding m/z ratio and retention time and their statistically significant differences as represented by p-values. It also provides a “metlin” link for each analyte that lead to a metabolomics database to provide some clues for identification of the analyte.

>reporttab <- diffreport(xset3, "EV","OMT2", metlin = 0.15) ### There are 580 analytes being identified in this report. ###

> colnames(reporttab)

[1] "name" "fold" "tstat" "pvalue" "mzmed"

[6] "mzmin" "mzmax" "rtmed" "rtmin" "rtmax"

[11] "npeaks" "EV" "OMT2" "metlin" "ev-10.mzdata"

[16] "ev-11.mzdata" "ev-12.mzdata" "ev-13.mzdata" "ev-14.mzdata" "ev-16.mzdata"

[21] "ev-18.mzdata" "ev-19.mzdata" "ev-1b.mzdata" "ev-21.mzdata" "ev-22.mzdata"

[26] "ev-23.mzdata" "ev-24.mzdata" "ev-25.mzdata" "ev-27.mzdata" "ev-28.mzdata"

[31] "ev-29.mzdata" "ev-3.mzdata" "ev-30.mzdata" "ev-4.mzdata" "ev-7.mzdata"

[36] "ev-8.mzdata" "ev-9.mzdata" "omt2-1.mzdata" "omt2-10.mzdata" "omt2-12.mzdata"

[41] "omt2-14.mzdata" "omt2-17.mzdata" "omt2-2.mzdata" "omt2-21.mzdata" "omt2-22.mzdata"

[46] "omt2-23.mzdata" "omt2-24.mzdata" "omt2-26.mzdata" "omt2-27.mzdata" "omt2-28.mzdata"

[51] "omt2-29.mzdata" "omt2-4.mzdata" "omt2-6.mzdata" "omt2-7.mzdata" "omt2-8.mzdata"

***Selection of key BIA metabolites***

Table 1 shows a list of the metabolites that are involved in biosynthesis pathways of noscapine and other closely related BIAs. Their m/z ratios and retention times on LC-MS were also included in Table 1. Such targeted metabolites can be picked out from “reporttab”.

Table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| BIAs | m/z | RT(min) | RT(sec) |
| 4-*O*-desmethylpapaveroxine-3-*O*-acetate | 444.2 | 5.03 | 301.8 |
| noscapine | 414 | 7.18 | 430.8 |
| papaverine | 340 | 5.86 | 351.6 |
| morphine | 286 | 0.67 | 40.2 |
| codeine | 300 | 2.25 | 135 |
| thebaine | 312 | 4.27 | 256.2 |
| reticuline | 330 | 3.47 | 208.2 |
| (*S*)-1,3-dihydroxy-*N*-methylcanadine | 386 | 3.83 | 229.8 |
| narcotine hemiacetal | 416 | 5.8 | 348 |
| 3-*O*-acetyl-papaveroxine | 458 | 6.03 | 361.8 |
| narcotoline | 400 | 5.62 | 337.2 |
| narcotoline hemiacetal | 402 | 4.75 | 285 |
| (*S*)-1-hydroxy-3-*O*-acetyl-N-methylcanadine | 428 | 4.44 | 266.4 |

BIAs<-function(mz, rt)

{

mypick<-reporttab[round(reporttab$mzmed)==mz

&abs(reporttab$rtmed-rt)<10,]

}

BIA\_list<-rbind(BIAs(400,337.2),BIAs(414,430.8),BIAs(286,40.2),BIAs(444,301.8)

,BIAs(340,351.6),BIAs(300,135),BIAs(312,256.2),BIAs(330,208.2)

,BIAs(386,229.8),BIAs(416,348),BIAs(458,361.8),BIAs(402,285)

,BIAs(428,266.4))

> BIA\_list[,c("name","tstat","fold","pvalue","EV","OMT2")]

name tstat fold pvalue EV OMT2

7 M400T337 4.6451977 8.224326 0.0002194656 0 12

109 M414T430 -2.0617299 1.400250 0.0459739654 23 18

517 M286T41 -0.2671444 1.018919 0.7907912775 23 18

363 M340T352 -0.7540356 1.139529 0.4553568691 23 18

488 M300T137 -0.3724562 1.098252 0.7118501478 22 18

443 M312T257 0.5274691 1.171428 0.6022509804 23 18

398 M330T209 0.6526849 1.141399 0.5177901383 12 14

136 M428T266 -1.8828685 1.710358 0.0672241323 12 3

Only metabolites of m/z ratios of 400, 414, 286, 340, 300, 312, 330 and 428 were detected in this set of plant samples. And narcotoline(m/z 400) content in the omt2 plants were significantly higher than those in the EV plants. Noscapine(m/z 414) content in the omt2 plants were significantly lower than those in the EV plants. Therefore, down-regulation of *OMT2* gene in plants resulted in increased accumulation of narcotoline and reduced accumulation of noscapine.

In addition to the listed BIAs, other metabolites that were significantly different between the two sets of samples can also be picked out using the following code.

>p0.05<-reporttab[reporttab$pvalue<0.05,]

***Plotting peaks***

>gt<-groups(xset3)

>plotmz<-function(mz,rt)

{

groupidxmz<-which(round(gt[,"mzmed"])==mz & abs(gt[,"rtmed"]-rt)<10)

eiccormz<-getEIC(xset3,groupidx=groupidxmz)

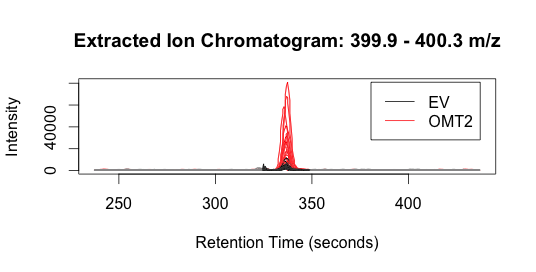
plot(eiccormz,xset3,groupidx=1)

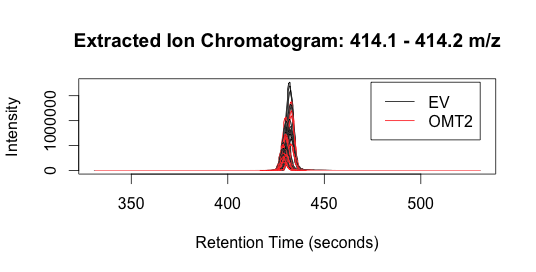
}

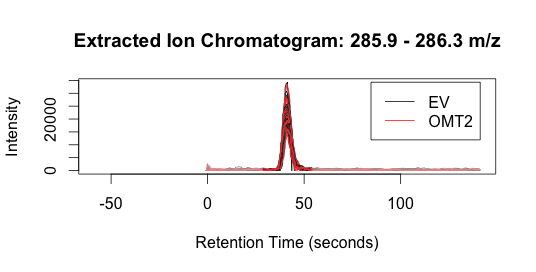
plotmz(400,337.2) ### narcotoline

plotmz(414,430.8) ### noscapine

plotmz(286, 40.2) ###morphine







**Reference**

Smith. R. (2005) LC/MS preprocessing and analysis with xcms. http://​www.​bioconductor.​org