Antibase Parser for LCMS Data with R Programming | Instruction Guide 2015

Introduction:

This guide is designed to help you use the AntiBase Parser R Script to analyze LCMS data. The XCMS Bioconductor package was used to analyze LCMS data (PDF attached). The purpose of this script is to streamline the process of analyzing LCMS files and comparing resulting masses against AntiBase in order to prioritize hits.

File Input:

Files containing LCMS data must be in either netCDF, mzXML, or mzData files, (However only netCDF files have been used in testing this script thus far).

AntiBase Input:

The Antibase output file containing all valuable information for each compound found in Antibase 2013 is included in *antibase_tableout.csv*. Store this in the same location as the LCMS files to be analyzed along with the *Antibase_parser.R* script.

The files to be analyzed need to be in a format following this example:

Ex: WMMC370_A_F401.CDF

The most important part of naming is that the F4 (or F1, F2, F3) is separated by an underscore separating it from the strain name (so that wells can be correctly identified).

Running The Script:

General Notes: This script is designed to run on R version 3.2.0, but any current version of R should work. You will also need to install the package for **xcms** via the R console (once R is installed) by using the command *install.packages("xcms")* see figure 1 to the right for further information.

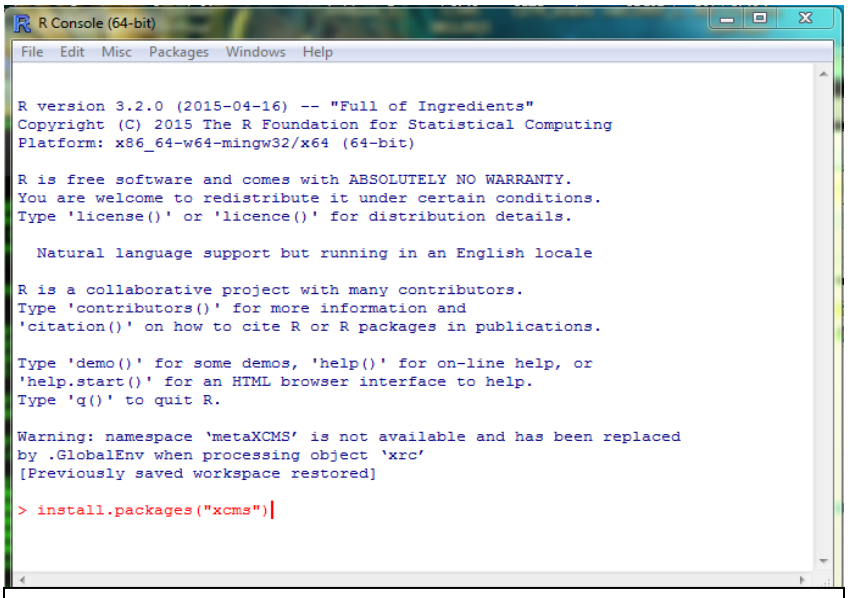

Figure 1: R Console view for installing package xcms

Different Computing Environments:

Option 1. to run the script on windows open the command prompt, navigate to directory with LCMS data to analyze, and enter:

R <Antibase_parser.R> AB_OUT.out

- Where the pink highlighted part is the path to the Antibase_parser.R Script (recommended to be in the same directory as the LCMS files to be analyzed).

Option 2. If you have Cygwin installed or are using Mac OS or Linux (any flavor) use this following command in terminal:

R CMD BATCH Antibase_mass_parser.R

- Where the pink highlighted part is the path to the Antibase_parser.R script

IMPORTANT: You will also need to put the R.bat script in the same directory as the LCMS files, along with Antibase_parser.R script, and antibase_tableout.csv (Directions on creating the R.bat script file are below).

Creating the R.bat script file:

The purpose of this file is to tell the R program where the R.exe file is. Which on windows is in the Program Files Directory:

Example for Windows: "C:\Program Files\R\R-3.2.0\bin\R.exe"

To create the R.bat file, open your favorite text editor (Vim, emacs, Notepad++, Nano) and write this line at the top of your script:

"C:\Program Files\R\R-3.2.0\bin\R.exe" CMD BATCH Antibase_parser.R

Where the path in quotes is the file path to the R.exe script described above. You must put this file in the same directory as the LCMS file, the *antibase_tableout.csv*, and *Antibase_parser.R* files.

Understanding Script Output:

This script will output data in 2 formats, one will be all the peaks detected with the LCMS prior to processing (file_LCMS_RawData.csv) and the other file will contain the processed data once matched against Antibase, with all masses that matched one or less times against Antibase (file_two_or_less_ABMatch.csv). The file prefix will be the netCDF file name prefix to easily identify which output file matches to a certain netCDF file. Two examples of each output file are provided below:

Table 1: LCMS_RawData file output

	mz	mzmin	mzmax	rt	rtmin	rtmax	into	intf	maxo	maxf	i	sn
1	197.3189	197.2828	197.3703	1474.44	1461.23	1494.24	482.4723	1021.088	29.47272	54.35557	1	10.28716
2	234.3149	234.2891	234.3347	1476.84	1463.03	1499.65	2379.187	4778.662	112.7288	225.3937	1	16.87176
3	349.4304	349.3981	349.4568	372.86	356.05	387.87	2784.642	5910.577	184.0682	312.4605	1	12.15501
4	398.4722	398.4338	398.5211	1114.85	1101.64	1128.06	3759.179	15579.23	551.2461	957.8634	1	12.85585

Column Names:

m/z : weighted (by intensity) mean of peak m/z

mzmin : m/z of minimum step

mzmax : m/z of maximum step

rt : retention time of peak midpoint

rtmin : leading edge of peak retention time

rtmax : trailing edge of peak retention time

into : integrated area of original (raw) peak

intf : integrated area of filtered peak

maxo : maximum intensity of original (raw) peak

maxf : maximum intensity of filtered peak

i : rank of peak identified in merged EIC

sn : signal to noise ratio of the peak

Table 2: One_or_less_ABMatch output file

	masses	retention	fractions	wells	times_matched	names	sources	type_match
1	197.3189	24.574	21	B 1	0	no matched names	no matched sources	
2	234.3149	24.614	21	B 1	0	no matched names	no matched sources	
3	349.4304	6.214334	4	A 5	0	no matched names	no matched sources	
4	398.4722	18.58083	15	B 7	0	no matched names	no matched sources	
5	399.4773	18.58083	15	B 7	0	no matched names	no matched sources	
6	400.4831	18.58083	15	B 7	0	no matched names	no matched sources	
7	401.4849	18.58083	15	B 7	0	no matched names	no matched sources	
8	416.4657	18.58083	15	B 7	0	no matched names	no matched sources	
9	417.4696	18.58083	15	B 7	0	no matched names	no matched sources	
10	496.4283	17.13017	13	B 9	1	E-24-Ethylidene-5a-lan	[F] Mucor rouxii	m
11	520.4001	16.44967	13	B 9	0	no matched names	no matched sources	
12	531.45	22.66283	19	B 3	0	no matched names	no matched sources	
13	590.1903	5.083833	2	A 3	1	10-Dihydrosteffimycin	[F] [B] Actinoplanes uta	m
14	661.1144	6.5945	4	A 5	0	no matched names	no matched sources	
15	677.0889	6.4945	4	A 5	1	U-78608	[synthetic]	m-Na

Column Names:

Masses : Masses processed from raw LCMS data that had one or fewer matches against AntiBase.

Retention : Retention time of masses matched one or less time against AntiBase.

Fractions : Which fraction mass falls into based on its retention time.

Wells : Well ID formulated from the retention time, fractions number, and input file name

Times_matched : How many times compound was matched against Antibase

Names : Name of any matched natural product (if any exist).

Sources : which source the natural product matched originates from

Type_match : Whether the match of the mass to antibase was a result from the mass alone, the mass minus a hydrogen ion (m –H) or mass minus a sodium ion (m – Na)

Further Questions/Concerns:

I am expecting some issues to occur when operating this script on a new computer (because script has only been tested on one computer/ only one OS thus far). Please feel free to email me at Phoenix@PhoenixLogan.net if you have any questions / trouble shooting help.

Useful Websites:

Information regarding R programming Setup:

https://cran.r-project.org/doc/manuals/r-release/R-admin.html

Information about Bioconductor package XCMS:

http://bioconductor.org/packages/release/bioc/html/xcms.html