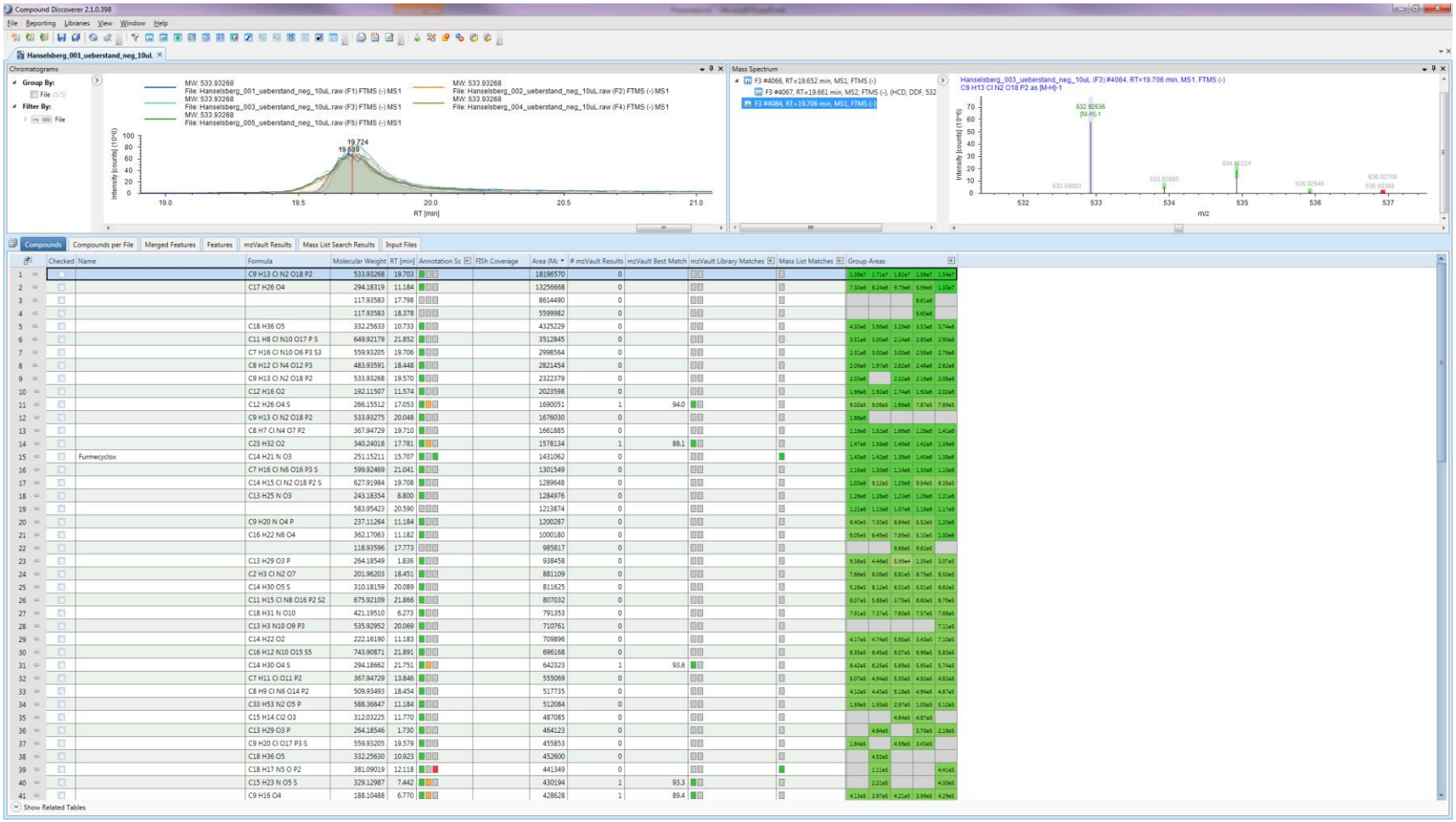
Tutorial to use R with the export data of Compound Discoverer to obtain input files for the FOR-IDENT platform

Uwe Kunkel

2018/01/10

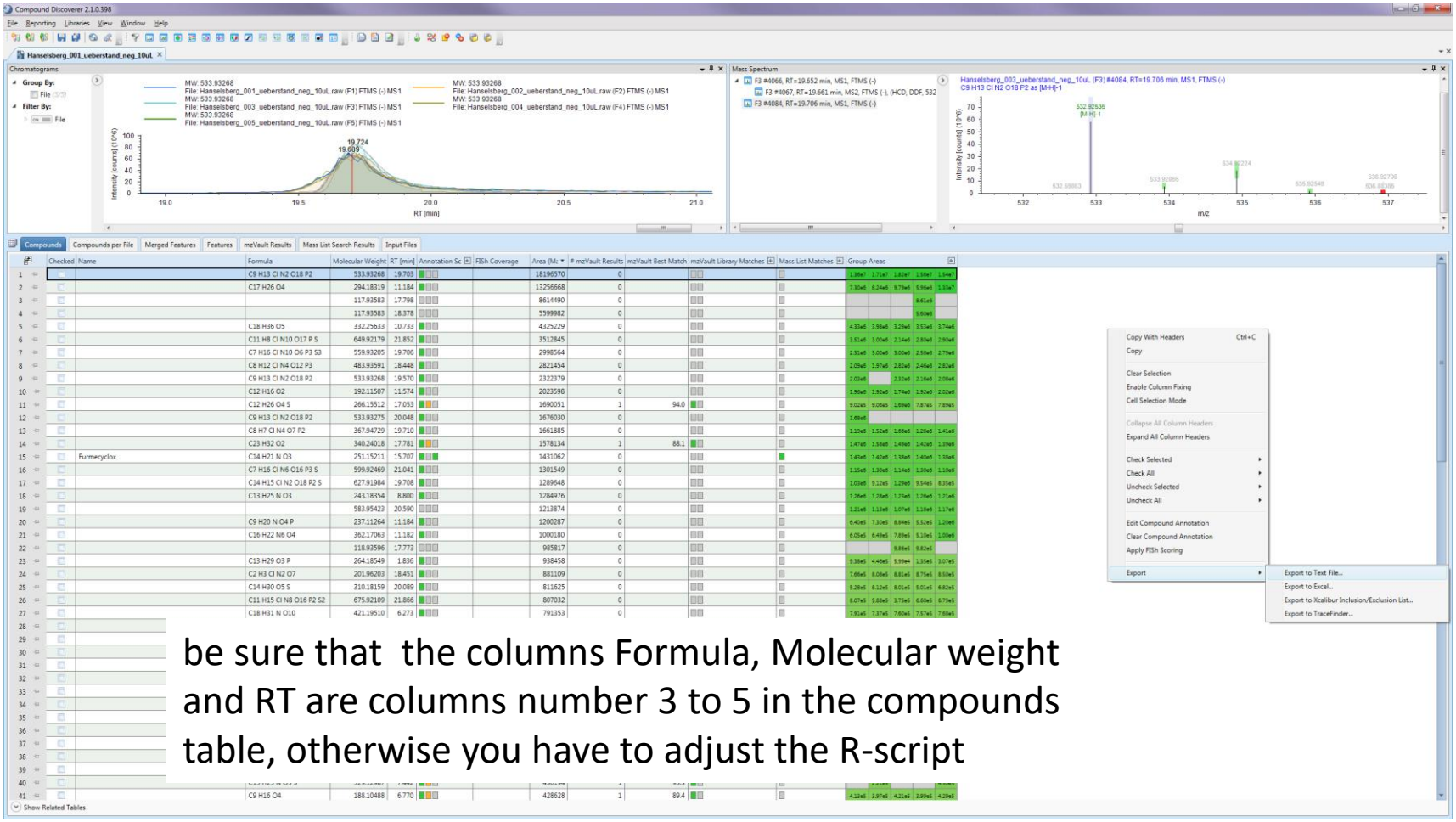
1. Step: export data from CD

- Go to the results page of your CD study



1. Step: export data from CD

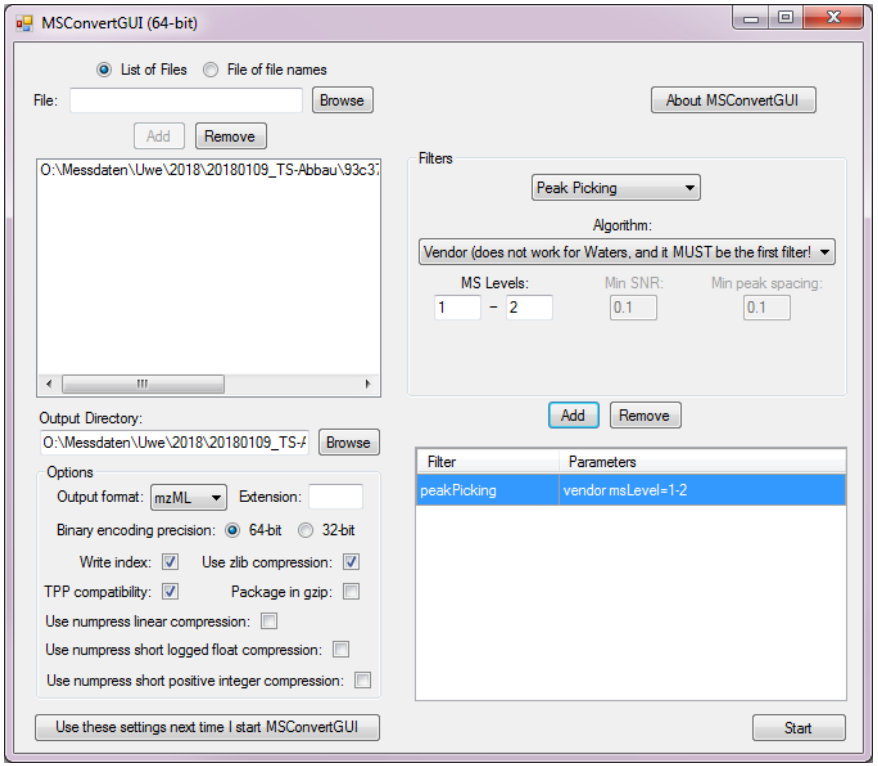
- Export the compounds table to a text-file (e.g. Compounds_CD_neg.csv)



be sure that the columns Formula, Molecular weight and RT are columns number 3 to 5 in the compounds table, otherwise you have to adjust the R-script

2. Step: Convert raw-file to mzML-formate

- This can e.g. be done with Proteowizard (http://proteowizard.sourceforge.net/downloads.shtml)



Activate the pick picking algorithm for MS levels 1 and 2 during the file conversion

This will convert your data from “profile” to “centroid”

This centroidization is not absolutely necessary but recommend.

3. Step: Install R (and R-studio)

- Install the latest version of R
  - https://cran.r-project.org/bin/windows/base/

- You may optionally also install R-Studio, by this you get a nicer GUI for R
  - https://www.rstudio.com/products/rstudio/download/

4. Preparing R for the work

- At first you have to install an extra package of R that can handle the mzML-data
- Thereto paste in the R console

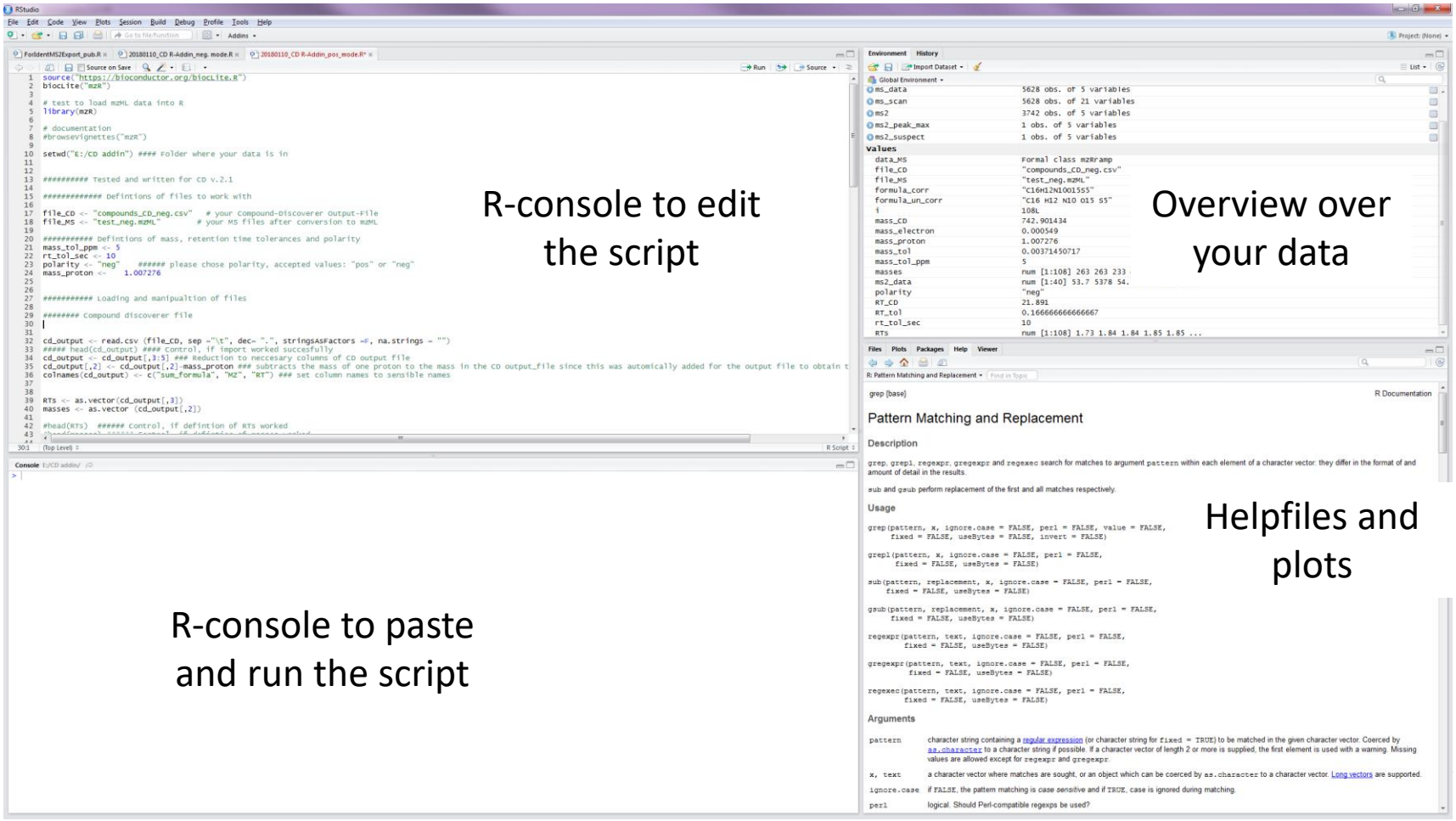
```
source("https://bioconductor.org/biocLite.R")
biocLite("mzR")
```

This only works if you have an internet connection, otherwise there are also options the install the mzR package if you do not have internet access on the

- You may optionally also install R-Studio, by this you get a nicer GUI for R and also directly an editor which you need to write and adapt R-code
  - In the following I used R-Studio and there workflow is described using R-Studio

4. Preparing R for the work

- Open the respective R-script (20180110_CD R-Addin_neg. mode.R or 20180110_CD R-Addin_pos. mode.R) for pos or neg. mode



5. Adjust the R-script to your data

- Change the necessary things in the R-code to import and run your data
  - The path where your data in located

    In script for neg mode: line 10: setwd("E:/CD addin")
  - The name of your Compound discover output file

    In script for neg mode: line 19: file_MS <- "test_neg.mzML"
  - The name of the mzML-file which you want to process and look for MS/MS data

    In script for neg mode: line 18: file_CD <- "compounds_CD_neg.csv“

    The name of the mzML-file which you want to process and look for MS/MS data
  - The info on the polarity of your data

    In script for neg. mode: line 25: polarity <- “neg"
  - The RT (in seconds) and mass tolerances (in ppm) which the script uses for finding MS2-information based on the informtation in the CD export file. I think 5 ppm and 10 s are a good value

    In script for neg. mode: line 23: mass_tol_ppm <- 5

    In script for neg. mode: line 24: rt_tol_sec <- 10

5. Adjust the R-script to your data(2)

- Change the necessary things in the R-code to import and run your data
  - When importing the CD export file into R, you have to specify the formate of the csv-file, i.e. what is used to separate the columns and what is used for decimal points

    If you use English language settings, you normally have a “,” as column separator and “.” as decimal point

    In script for neg. mode: line 33: cd_output <- read.csv (file_CD, sep =“.", dec= ".", stringsAsFactors =F, na.strings = "")
  - To check the correct importing of the data from the CD output file you can use line 34: head(cd_output)
  - The output should look as follows (if not, you must change the code in line 33)

```
35:12  (Top Level) ÷                                                    R Script ÷

Console  E:/CD addin/
>
> file_CD <- "compounds_CD_neg.csv"   # your Compound-Discoverer Output-File
> file_MS <- "test_neg.mzML"        # your MS files after conversion to mzML
>
>
> ########### Defintions of mass, retention time tolerances and polarity
> mass_tol_ppm <- 5
> rt_tol_sec <- 10
> polarity <- "neg"    ###### please chose polarity, accepted values: "pos" or "neg"
> mass_proton <-    1.007276
>
> ########### Loading and manipualtion of files
>
> ######## Compound discoverer file
>
>
> cd_output <- read.csv (file_CD, sep ="\t", dec= ".", stringsAsFactors =F, na.strings = "")
> head(cd_output)
  Checked Name             Formula Molecular.Weight RT..min. Annotation.Source..Predicted.Compositions Annotation.Source..mzVault.Search Annotation.Source..MassList.Match
1   False <NA>        C13 H29 O3 P          264.1855    1.730                              Full match                      No results                       No results
2   False <NA>        C13 H29 O3 P          264.1855    1.836                              Full match                      No results                       No results
3   False <NA>          C15 H22 O2          234.1620    1.840                              Full match                      No results                       No results
4   False <NA>   C10 H14 N2 O17 P2          495.9772    1.847                              Full match                      No results                       No results
5   False <NA> C13 H15 N4 O3 P3 S2          431.9809    1.847                              Full match                      No results                       No results
6   False <NA>                <NA>          196.8184    1.917                              No results                      No results                       No results
  FISh.Coverage Area..Max.. X..mzVault.Results mzVault.Best.Match mzVault.Library.Match..mzVault.February.2017 mzVault.Library.Match..Lfu_20170811
1            NA   464123.30                  0                 NA                          No matches found                     No matches found
2            NA   938457.82                  0                 NA                          No matches found                     No matches found
3            NA   123458.45                  0                 NA                          No matches found                     No matches found
4            NA   145444.70                  0                 NA                          No matches found                     No matches found
5            NA    76996.93                  0                 NA                          No matches found                     No matches found
6            NA   191369.78                  0                 NA                          No matches found                     No matches found
  Mass.List.Match..EFS.HRAM.Compound.Database Group.Area..F1 Group.Area..F2 Group.Area..F3 Group.Area..F4 Group.Area..F5
1                          No matches found            NA       464123.3            NA       369849.74       215593.62
2                          No matches found      938457.82       446312.5        59944.03      135160.34       307033.00
3                          No matches found       91576.34       117357.7        38242.17       57823.58       123458.45
4                          No matches found             NA             NA       48081.21       82310.20       145444.70
5                          No matches found             NA             NA             NA             NA        76996.93
6                          No matches found      191369.78             NA             NA             NA              NA
> |
```

6. Work in R-Studio and run the script

- To work in R-Studio, just copy and paste parts of the script from the editor window to the console window. The script then runs automaically
- In the editor, you can make some comment to explain the code and data manipulation steps. This is done by using “#”. All that is written in a line in the editor after the “#” sign is not interpreted in the R-console.

- If all the adjustments worked, you can copy and paste all things from the editor window and you will automatically get a txt-file in the FOR-IDENT format. The file will be saved in the folder where all your data is located

- The file name will be automatically chosen. It will be:

  “Name_of_the_mzML file”_”polarity”_”export_CD”.txt

  So if you file is “test.mzmL” and you measured in neg. mode, the file will get the name “test_neg_pos_export_CD.txt”