# XCMS Data Analysis

## Summary

This document will cover the basic usage of XCMS in LC-MS data analysis.

## Installing XCMS

XCMS must be installed and the library loaded. The xcms.r script will perform this operation.

Alternatively, you can install it manually by entering the following lines:

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

The library is then loaded by entering

```
library(xcms)
```

## Converting files

Files must be converted to an open format such as netCDF, mzXML or mzData files. Two programs that can do this are trapper (available from http://tiny.cc/yqn3s ) and msConvert (available from http://tiny.cc/z40om). While msConvert is available for all platforms, we have not yet tested it fully. Trapper is only available for Windows host systems.

The script d2mzXML.r will convert the proprietary Agilent format .d folders into mzXML files using trapper (this will likely be changed to msConvert once we have tested it, as trapper is no longer supported). It will scan the current folder for the .d extension and convert each folder to a single mzXML file using the command line.

## Folder structure

Once the files have been converted, they need to be assigned to groups. This is done by creating separate folders for each group (for example 'blank', 'control', 'treated') and placing the mzXML files into the appropriate folder, as presented in Figure 1.

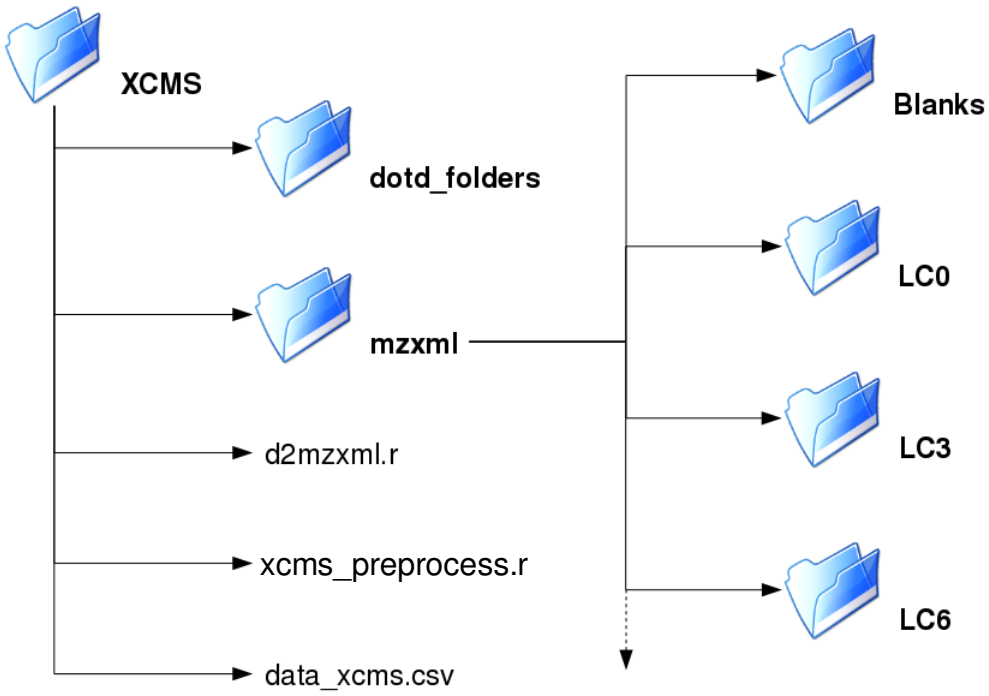


*Figure 1: Example folder structure for XCMS data processing and analysis. Proprietary format files are stored in one folder, and open format mzXML files are stored in another. The scripts that process the files are in the parent (XCMS) folder, and the output file, data_xcms.csv is also shown here. In the mzXML folder, the samples are allocated to groups by placing them in subfolders.*

## Processing

The script will process all the mzXML files listed in the output directory, and apply retention time correction and peak filling. Retention time correction accounts for the drift in retention time that occurs in chromatographic data. Peak filling is the method by which the data is integrated in the region where there was no peak found so that there is a value (however small).

It will also plot the TIC for a single file, and a single peak from this file to verify peak width is similar to the value specified for `fwhm` in the `xcmsSet` function.

`xcmsSet` takes multiple files and treats them as a single object, allowing retention time correction and peak filling to be performed on all files at once. As it stands, the script performs 3 passes of retention time correction which seems to be adequate for our test data. This may need to be extended for other data sets.

## Report generation

The final step in the script is to produce a report file in .csv format for further processing and analysis, including filtering the data matrix and statistical analysis.

## Running the scripts

To convert the .d files, if the folder structure is as it is specified within the script itself, just type

```
source("d2mzXML.r")
d2mzXML(input_dir, output_dir)
```

where input_dir and output_dir are the names of the folders, for example:

```
d2mzXML("Raw data files/", "mzxml/")
```

note that the trailing "/" is required.

To process the mzXML files, all that needs to be called is the `source` function. Change the working directory to the directory containing the scripts and the folder of mzXML files (for the example in Figure 1, the parent XCMS folder). At the R console, type in

```
source("xcms.r")
```

will run the `xcms.r` script as it is. This assumes the directory containing the mzxml files is "mzxml" (note that there is **no** trailing "/" in this script); if this is not the case, edit as necessary.