Data Clustering & Stability Analysis

Processing and clustering of S. hygroscopicus proteomic data

```
knitr::opts_chunk$set(echo = TRUE)
#setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
library('tidyr')
library('MSstats')
library('fpc')
library('ggplot2')
library('reshape2')
```

Timecourse proteomic profiles for S. hygroscopicus NRRL 30439 are clustered using k-means, and the stability of the k-means centroids are assessed using bootstrapped recalculation k-mean centroids. This is done using randomly sampled subsets of the original data, where a range of subset fractions are tested.

MSstats (Choi et al, 2014, DOI: 10.1093/bioinformatics/btu305) is used for processing of the Spectronaut output, and label free quantification.

```
if (file.exists("../data/ProcessedData-MSstats-spect.rda")) {
  load(file='../data/ProcessedData-MSstats-spect.rda')} else {
    specnaut_results <- read.table(file=params$input_file,sep=",",header = TRUE)
    rawData <- SpectronauttoMSstatsFormat(specnaut_results,
                                          removeProtein_with1Feature = TRUE,
                                          filter_with_Qvalue = TRUE)
    QuantData <- dataProcess(rawData, normalization='quantile',
                             summaryMethod="TMP", cutoffCensored='minFeature',
                             censoredInt="0", MBimpute=TRUE,
                             maxQuantileforCensored = 0.999)
    save(QuantData,file="../data/ProcessedData-MSstats-spect.rda")
    }
if (file.exists('../data/groupQuantMat-spect.rda')) {
  load(file='../data/groupQuantMat-spect.rda')} else {
  groupQuantMat <- quantification(QuantData, type="Group", format="matrix")
  groupQuantMat[is.na(groupQuantMat)] <- 0
  groupQuantMat <- as.matrix(groupQuantMat)
  gene_names <- groupQuantMat[,1]
  groupQuantMat <- groupQuantMat[,-1]
  rownames(groupQuantMat) <- gene_names
  class(groupQuantMat) <- "numeric"
  save(groupQuantMat, file = "../data/groupQuantMat-spect.rda")
  #write.csv(groupQuantMat, file = "../data/groupQuantMat-spect.csv")
}
rdata <- as.data.frame(groupQuantMat)
```

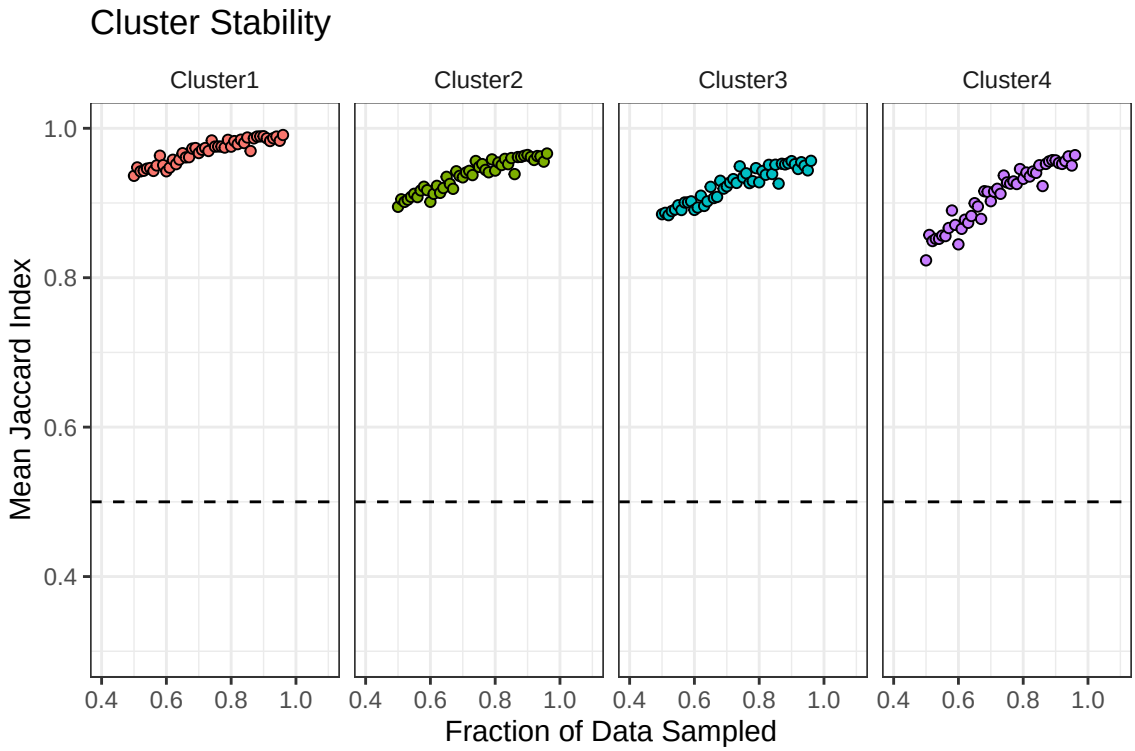
Stability of clustered protein time-courses

The clusterboot function in the `fpc` package is used to determine k-mean cluster stability. Protein expression profiles are standardized (zero unit, unit variance) across proteins in order to preserve profile shape.

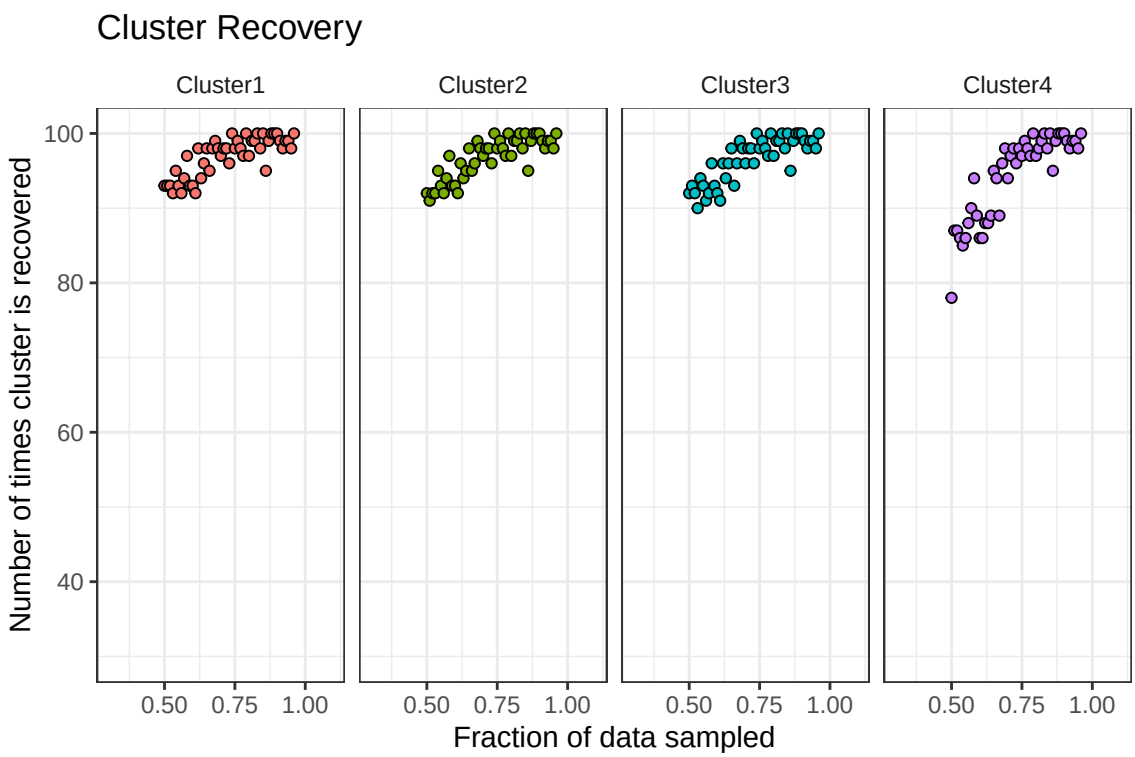
Cluster stability is determined by boostrapped calculation of a Jaccard coefficient using a randomly sampled subset of the data. K-means centroids are first calculated for the complete dataset, and subsequent centroids (using subset data) are assigned to a 'true cluster' for calculation of the Jaccard coefficient. Various subset sizes were tested for their ability to recover original centroids, and the mean Jaccard coefficient (Jaccard, 1901) was used for comparison of botstrapped calculations. Hennig 2007 suggests that a mean Jaccard coefficient of >0.80 is a reliable indicator of cluster stability, and a value of 0.5 is shown to be a critical value for cluster "dissolution" (Hennig, 2006).

```
data <- as.data.frame(groupQuantMat)
# Data at standardized across timecourses, to preserve profile shape
scaled_data <- t(scale(t(data)))

jacc_mean <- data.frame(matrix(ncol = 5, nrow = 0, dimnames=list(NULL, c("data_frac",
                                                                         "Cluster1",
                                                                         "Cluster2",
                                                                         "Cluster3",
                                                                         "Cluster4"))))
numb_recv <- jacc_mean
for (sample in seq(from=0.5, to=0.96, by=0.01)) {
  subset = floor(sample*nrow(scaled_data))
  kclusboot <- clusterboot(scaled_data, clustermethod = kmeansCBI,
                       bootmethod = "subset", subtuning = subset,
                       B=100, iter.max=100,
                       krange=4, seed=42)
  jacc_mean[nrow(jacc_mean) + 1, ] <- c(sample, kclusboot$subsetmean)
  numb_recv[nrow(numb_recv) + 1, ] <- c(sample, kclusboot$subsetrecover)
}
```

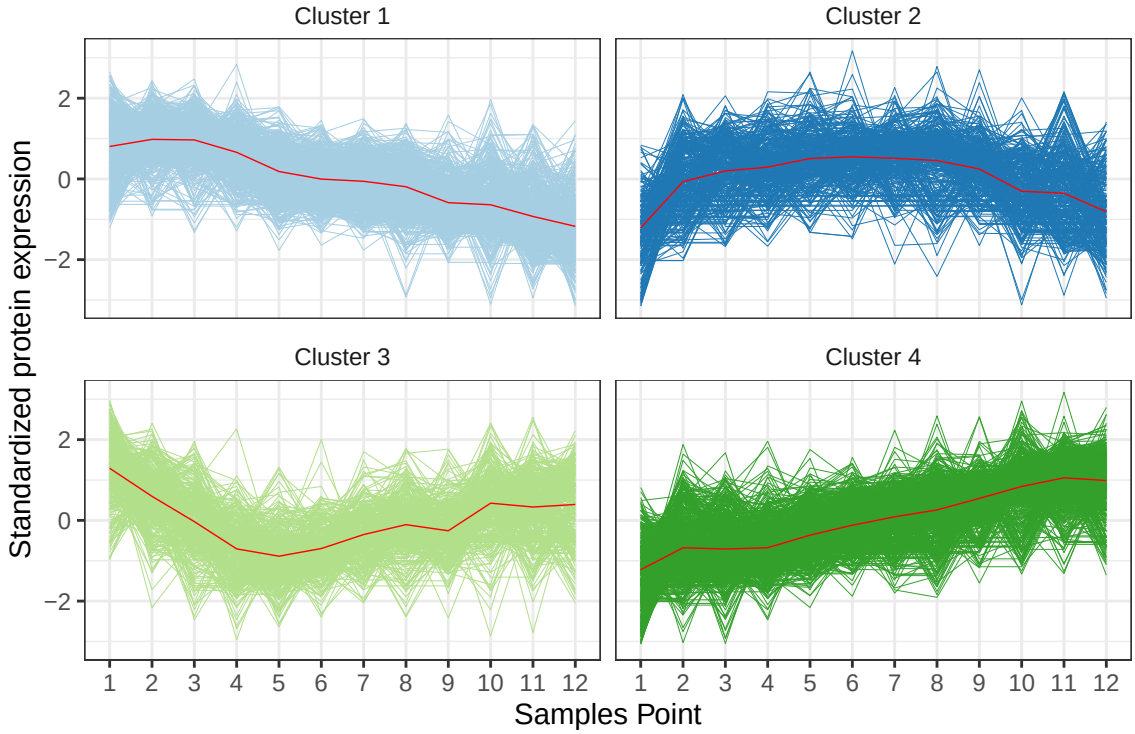


Bootstrapped Jaccard means indicate reliable cluster stability (>0.80), where centroids were found to be robust to subsetting 50% of the data.



High recovery rates are observed for the all 4 k-mean clusters, where the first 3 clusters appear marginally more stable than cluster 4 for lower fractions of randomly sampled data.

Given acceptable stability genes are clustered into 4 k-mean clustered for further analysis. The centroid for each cluster is plotted as a red line over the standardized and grouped protein profiles.



References

Meena Choi, Ching-Yun Chang, Timothy Clough, Daniel Broudy, Trevor Killeen, Brendan MacLean, Olga Vitek, MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments, Bioinformatics, Volume 30, Issue 17, 1 September 2014, Pages 2524–2526, (URL: https://doi.org/10.1093/bioinformatics/btu305)

Hennig, C., 2006. Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. Research Report no. 272, Department of Statistical Science, University College London, submitted for publication. (URL: http://www.ucl.ac.uk/Stats/research/Resrprts/psfiles/rr272.pdf)

Hennig, C., 2007. Cluster-wise assessment of cluster stability. Computtional Statistics & Data Analysis. 52, 258-271 (DOI: https://doi.org/10.1016/j.csda.2006.11.025)

Jaccard, P., 1901. Distribution de la florine alpine dans la Bassin de Dranses et dans quelques regiones voisines. Bull. Soc. Vaud. Sci. Nat. 37,241–272