

EVome

Alex Cope

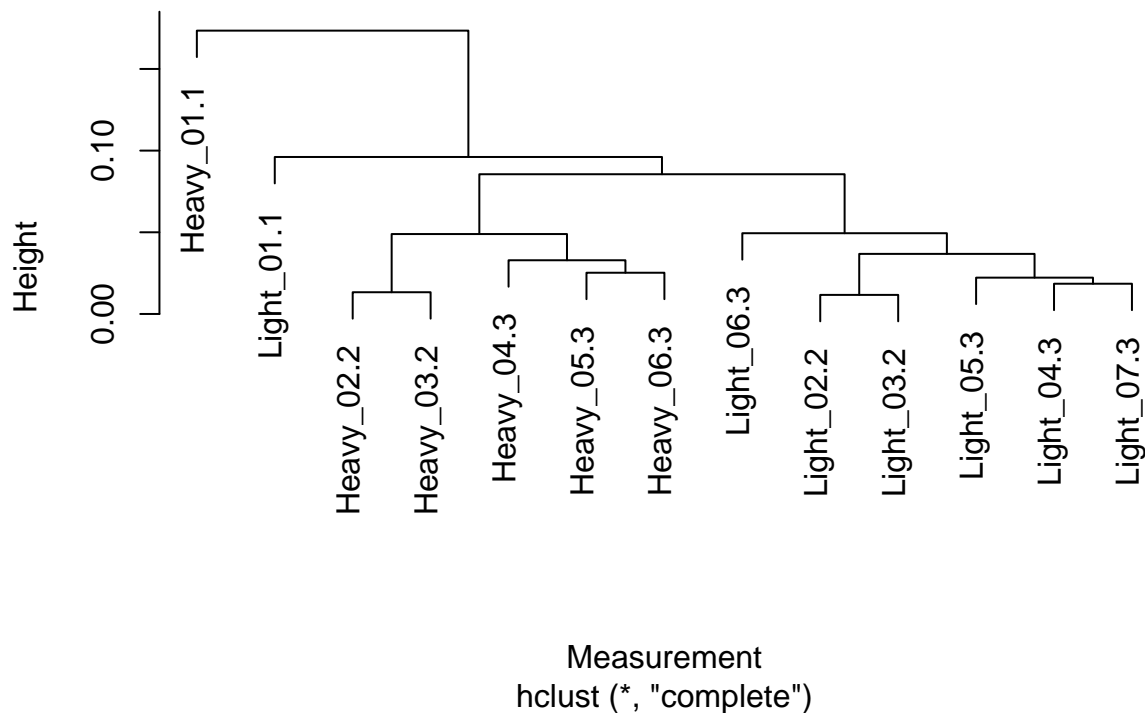
9/18/2020

I'm going to drop the Mixed sample for our data. Although Inna found that it was closer to Heavy EVs, it still could be considered its own group without any replicates. We have N=3 biological replicates for Heavy and Light EVs, which is the minimum we need to do a differential expression analysis.

The first thing I'm going to do is convert spectral counts to NSAF, which is a within sample normalization that accounts for protein length and total number of spectral counts taken in the sample.

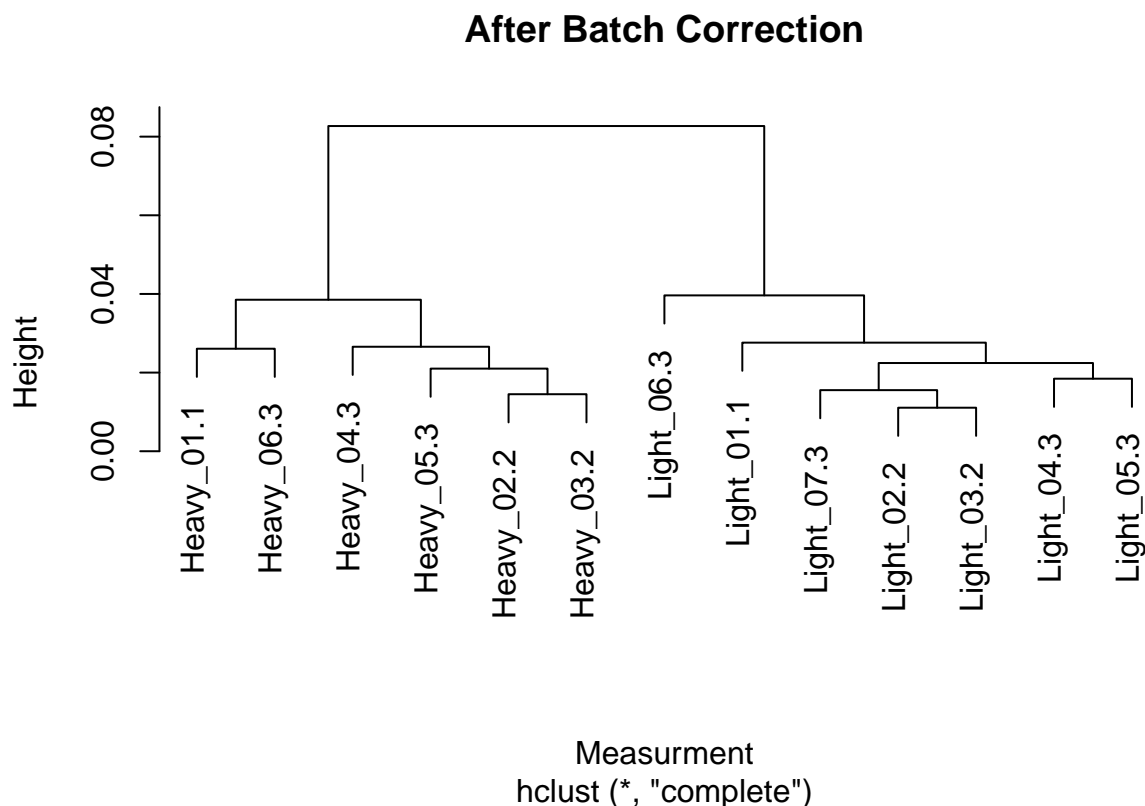
Our data consists of multiple measurements taken on different instruments (Orbitrap, Eclipse) and on different days. This is begging for possible batch effects, which can reduce our signal of differential expression between the heavy and light fractions. Using the NSAF values, let's cluster the different measurements to assess if we have a possible batch effect.

Before Batch Correction



From the hierarchical clustering, it is Heavy_01 and Light_01 appear to be outliers, which is unsurprising given that they were measured on a different instrument. I'm going to apply the mean-centering correction method used in the Bioconductor package msmsEDA to reduce the batch effects. Note that the vignette

for this tool states to treat missing values as 0. Question to investigate further: should we correct for batches before filtering (as done here) or after filtering of proteins with too many missing data points?



After applying the batch correction, Light_01 and Heavy_01 cluster with the other Light and Heavy EVomes, respectively.

Now that we've applied our batch corrections, let's filter proteins. For our differential expression analysis, we ideally want proteins to be truly measured in 2 of the 3 biological replicates. We can impute

```
## 'summarise()' regrouping output by 'Protein' (override with '.groups' argument)
```

```
## 'summarise()' ungrouping output (override with '.groups' argument)
## 'summarise()' ungrouping output (override with '.groups' argument)
## 'summarise()' ungrouping output (override with '.groups' argument)
## 'summarise()' ungrouping output (override with '.groups' argument)
## 'summarise()' ungrouping output (override with '.groups' argument)
## 'summarise()' ungrouping output (override with '.groups' argument)
## 'summarise()' ungrouping output (override with '.groups' argument)
## 'summarise()' ungrouping output (override with '.groups' argument)
## 'summarise()' ungrouping output (override with '.groups' argument)
## 'summarise()' ungrouping output (override with '.groups' argument)
## 'summarise()' ungrouping output (override with '.groups' argument)
## 'summarise()' ungrouping output (override with '.groups' argument)
## 'summarise()' ungrouping output (override with '.groups' argument)
## 'summarise()' ungrouping output (override with '.groups' argument)
## 'summarise()' ungrouping output (override with '.groups' argument)
## 'summarise()' ungrouping output (override with '.groups' argument)
## 'summarise()' ungrouping output (override with '.groups' argument)
## 'summarise()' ungrouping output (override with '.groups' argument)
```


[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

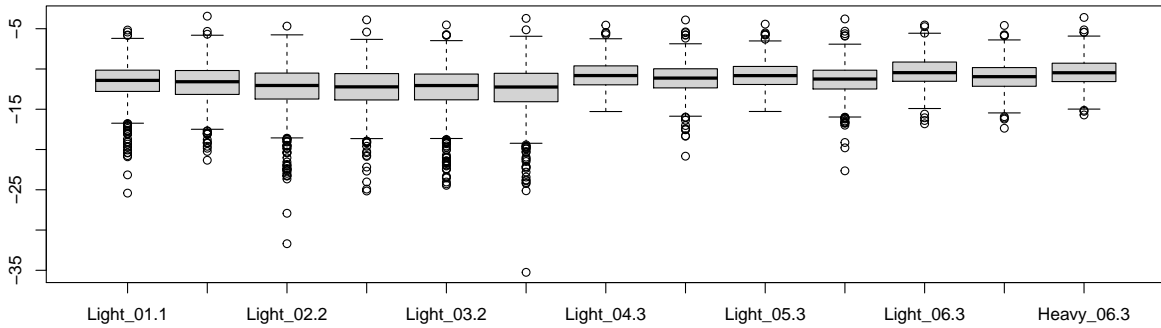
[illegible]

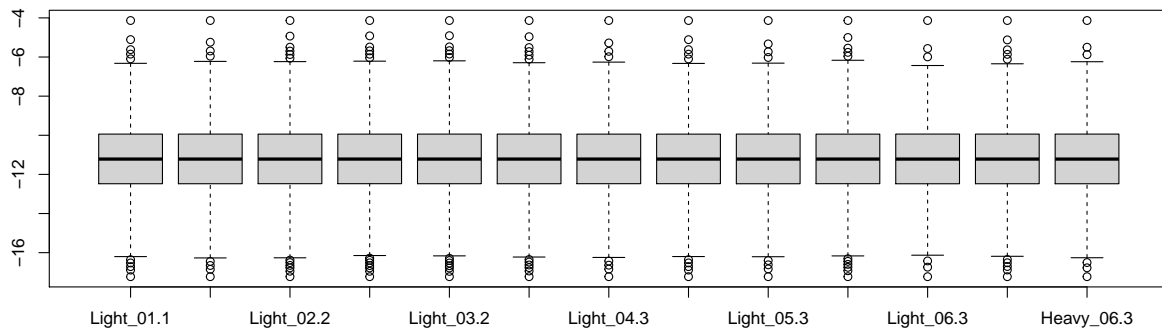
[illegible]

[illegible]

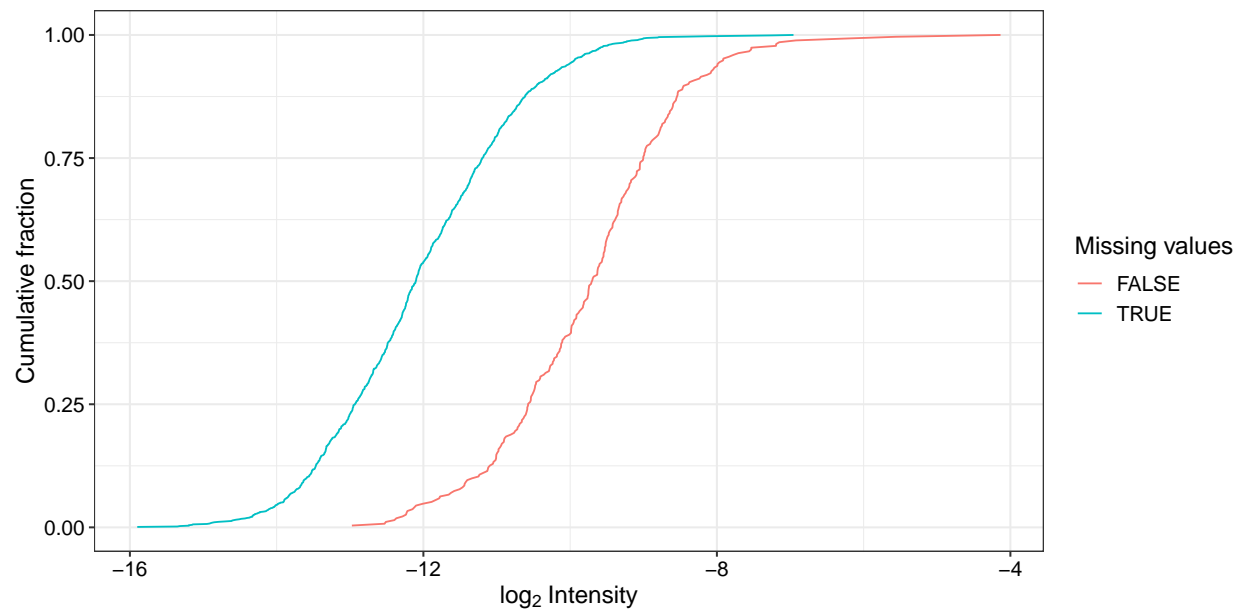
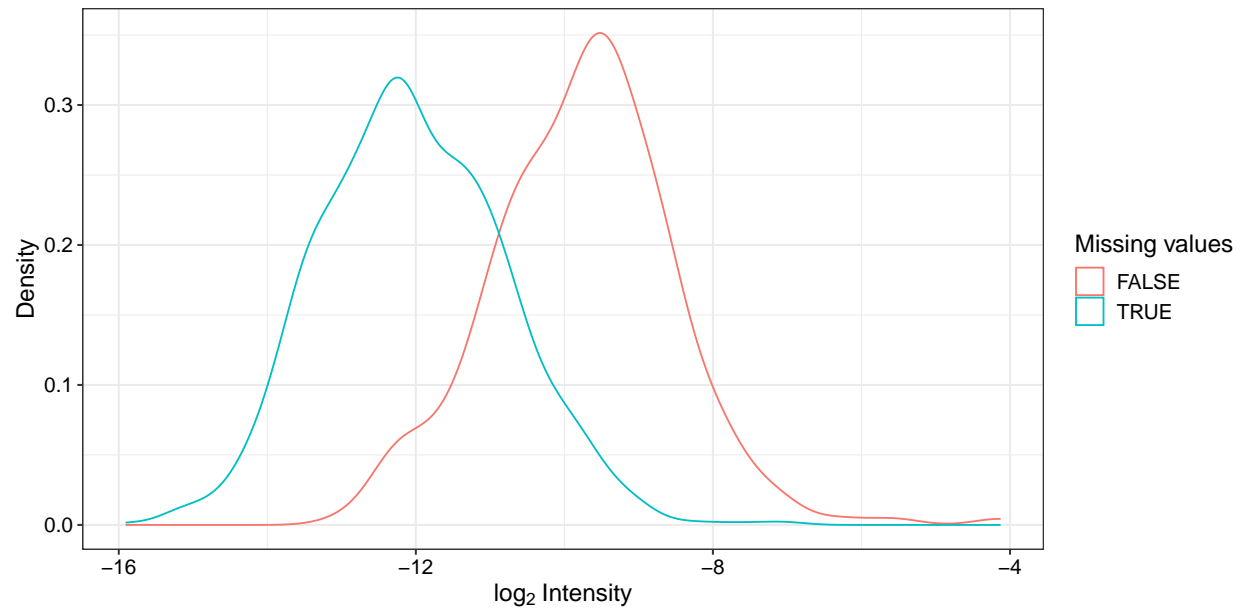
[illegible]

[illegible]

[illegible]



Next, we will impute missing values. Results seems to suggest that most of the missing values come from proteins that are low abundances, sometimes referred to as missing not at random. Often times these values will be imputed We will impute these using the imputation function available in the Bioconductor package DEP.



```
## Loading required package: imputeLCMD
## Loading required package: tmvtnorm
## Loading required package: mvtnorm
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##   expand, pack, unpack
```

```

## Loading required package: stats4

## Loading required package: gmm

## Loading required package: sandwich

## Loading required package: norm

## Loading required package: pcaMethods

## Loading required package: Biobase

## Loading required package: BiocGenerics

## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##   clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##   clusterExport, clusterMap, parApply, parCapply, parLapply,
##   parLapplyLB, parRapply, parSapply, parSapplyLB

## The following object is masked from 'package:Matrix':
##
##   which

## The following object is masked from 'package:limma':
##
##   plotMA

## The following objects are masked from 'package:dplyr':
##
##   combine, intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##   IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##   anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##   dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##   grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##   order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##   rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##   union, unique, unsplit, which, which.max, which.min

```

```
## Welcome to Bioconductor
##
##   Vignettes contain introductory material; view with
##   'browseVignettes()'. To cite Bioconductor, see
##   'citation("Biobase)"', and for packages 'citation("pkgname)".

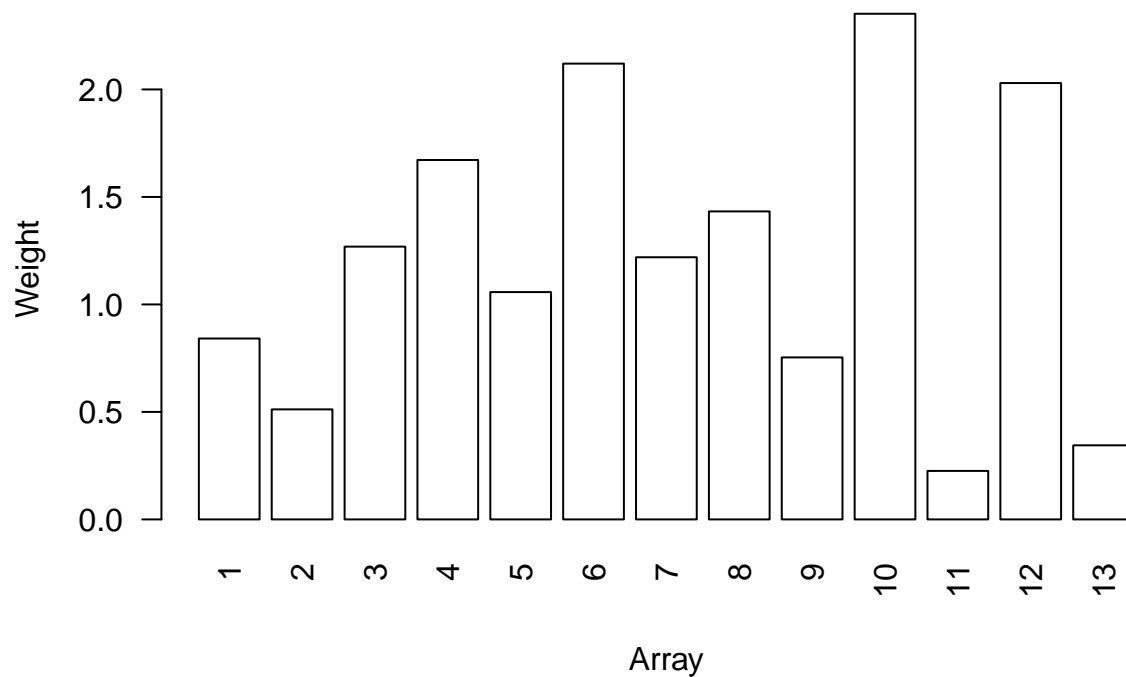
##
## Attaching package: 'pcaMethods'

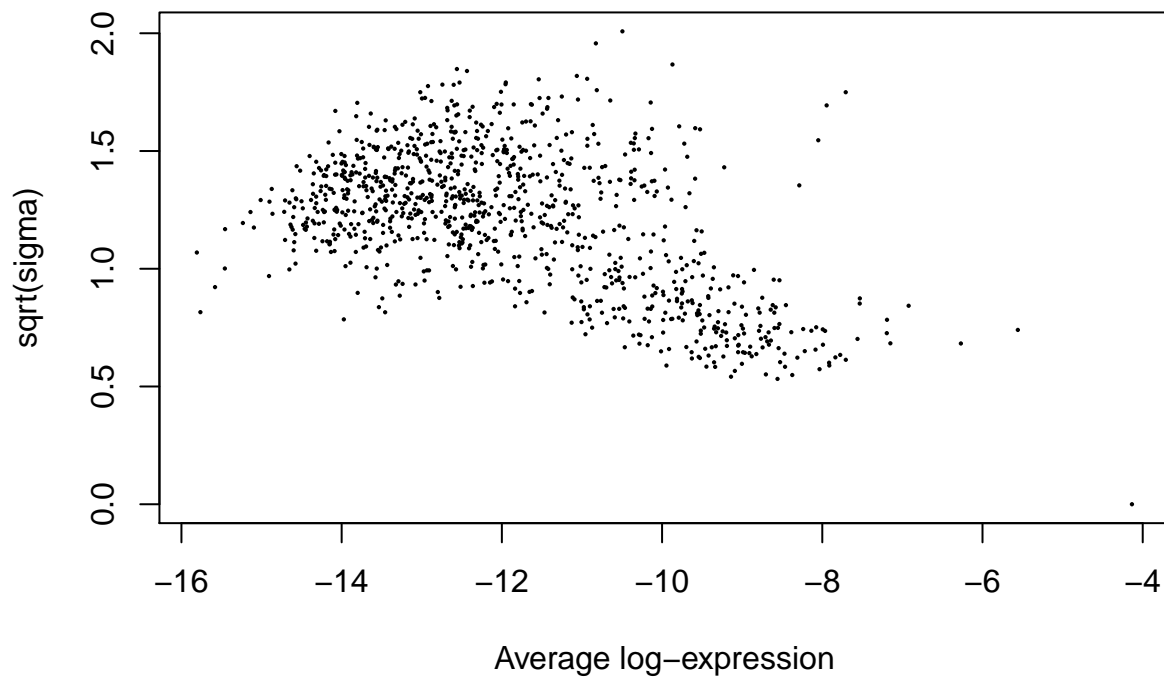
## The following object is masked from 'package:stats':
##
##   loadings

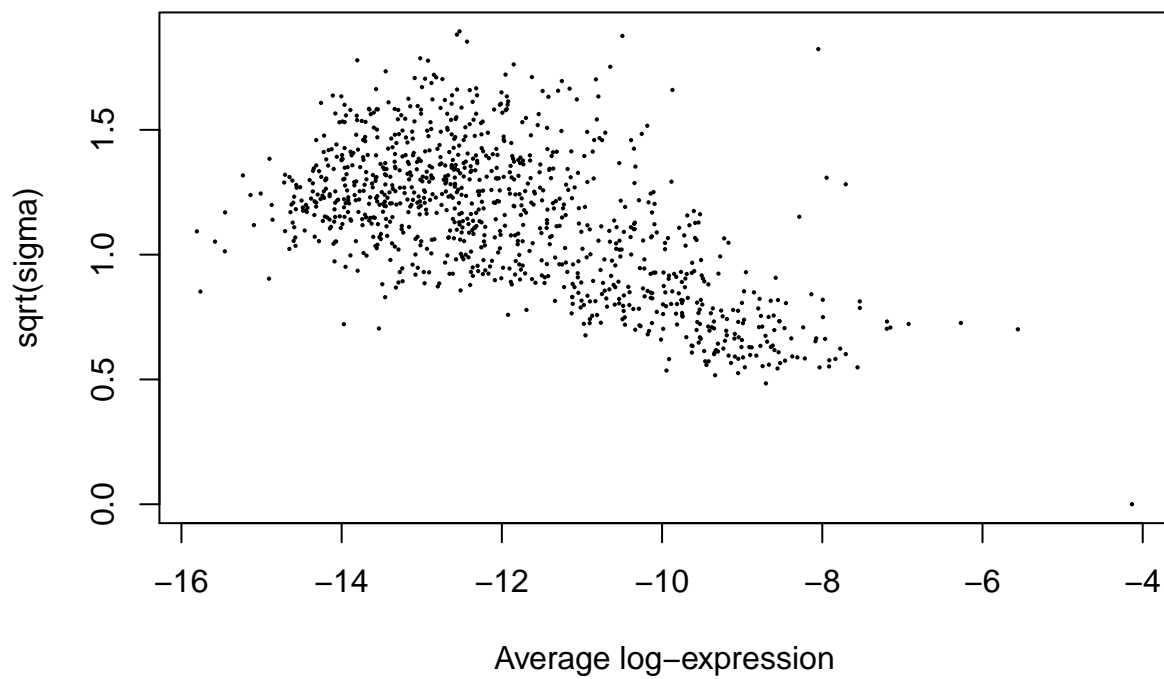
## Loading required package: impute

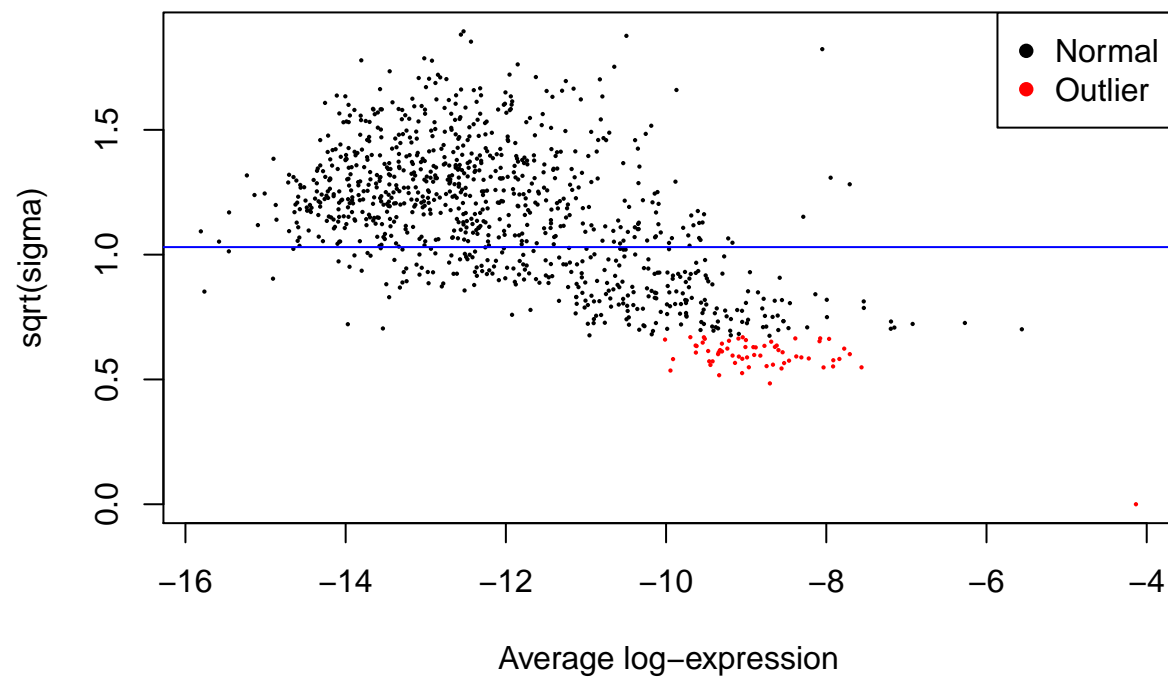
## [1] 0.9085871
```

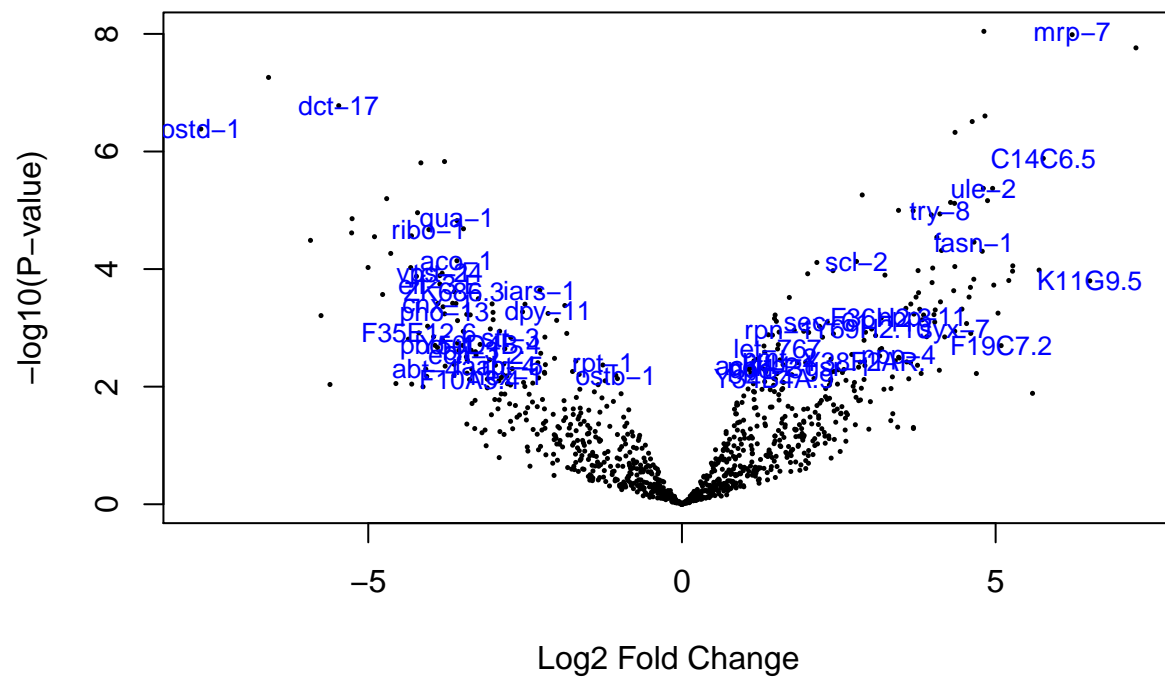
DEP log2 transforms the values and then imputes values. We will use this matrix as input into limma for comparing protein abundances between the Heavy and Light fractions of EVs.











Based on the limma analysis, we detect 235 genes with a log2 fold change of at least 1.5 and adjusted p-value < 0.05 out of the 1016 proteins remaining in the dataset after filtering.