Workflow metabolite set enrichment analysis with bootstrapping with `bmetenrichr`

This R-package aims to perform metabolite set enrichment analysis (MSEA) on single-cell metabolomics datasets. In contrast to bulk-metabolomics, metabolite annotation is often more ambiguous with fully resolved molecular structures. That means, annotations are vectors of isomeric (and/or isobaric) molecules, complicating downstream MSEA. This package uses a boostrapping approach by performing enrichment analyses many times with random sampling of the isomers/isobars.

```
options(stringsAsFactors = FALSE, warn = -1)

## install devtools if not installed
if(!("devtools" %in% rownames(installed.packages()))){
  install.packages("devtools",  repos = c(CRAN = "http://cran.rstudio.com"))
  }

## install bmetenrichr if not installed
if(!("bmetenrichr" %in% rownames(installed.packages()))){
   devtools::install_github(repo = "martijnmolenaar/bmetenrichr", build_vignettes = TRUE)
}

library(bmetenrichr)
#> Loading required package: ggplot2
#> Loading required package: dplyr
#>
#> Attaching package: 'dplyr'
#> The following objects are masked from 'package:stats':
#>
#>     filter, lag
#> The following objects are masked from 'package:base':
#>
#>     intersect, setdiff, setequal, union
```

The package contains example data from Rappez et al., 2021 (https://doi.org/10.1038/s41592-021-01198-0).

```
data("Rappez_et_al")

## the main input is a single-cell metabolomics matrix with molecules as rows and cells as columns
Rappez_et_al$sc_matrix[1:10,1:10]
#>                          8        15         18         25         27
#> C10H10N4O.K     0.01145815 0.0000000 0.00000000 0.00000000 0.00000000
#> C10H10O6.K      0.00000000 0.0000000 0.00000000 0.00000000 0.00000000
#> C10H11N4O7P.Na  0.00000000 0.0000000 0.00000000 0.00000000 0.00000000
#> C10H11N5O3.H    0.00000000 0.0000000 0.00000000 0.00000000 0.00000000
#> C10H12ClN5O3.K  0.00000000 0.0000000 0.00000000 0.00000000 0.00000000
#> C10H12ClN5O3.Na 0.00000000 0.0000000 0.00000000 0.00000000 0.00000000
#> C10H12ClNO4.H   0.01707534 0.0620677 0.05203464 0.01376432 0.02417577
#> C10H12ClNO4.K   0.00000000 0.0000000 0.00000000 0.00000000 0.00000000
```

```
#> C10H12ClNO4.Na  0.17496077 0.1702377 0.17311070 0.16992632 0.01635700
#> C10H12FN5O4.H   0.00000000 0.0000000 0.00000000 0.00000000 0.00000000
#>                         29          35         38          41          47
#> C10H10N4O.K     0.008558847 0.000000000 0.00000000 0.000000000 0.000000000
#> C10H10O6.K      0.000000000 0.000000000 0.00000000 0.000000000 0.006726224
#> C10H11N4O7P.Na  0.000000000 0.005782876 0.02968958 0.000000000 0.000000000
#> C10H11N5O3.H    0.000000000 0.000000000 0.00000000 0.000000000 0.000000000
#> C10H12ClN5O3.K  0.000000000 0.000000000 0.00000000 0.000000000 0.000000000
#> C10H12ClN5O3.Na 0.000000000 0.000000000 0.00000000 0.000000000 0.000000000
#> C10H12ClNO4.H   0.028755533 0.030977567 0.03611660 0.009692678 0.031959898
#> C10H12ClNO4.K   0.000000000 0.000000000 0.00000000 0.000000000 0.000000000
#> C10H12ClNO4.Na  0.174631645 0.147417684 0.12427280 0.160955912 0.120047020
#> C10H12FN5O4.H   0.000000000 0.000000000 0.00000000 0.000000000 0.000000000

## for the molecules, a vector with molecular formulas plus adduct is required
rownames(Rappez_et_al$sc_matrix)[1:10]
#>  [1] "C10H10N4O.K"     "C10H10O6.K"      "C10H11N4O7P.Na"  "C10H11N5O3.H"
#>  [5] "C10H12ClN5O3.K"  "C10H12ClN5O3.Na" "C10H12ClNO4.H"   "C10H12ClNO4.K"
#>  [9] "C10H12ClNO4.Na"  "C10H12FN5O4.H"

## a conditions vector is required to define to which condition a given cell belongs
Rappez_et_al$conditions[1:10]
#>  [1] "F" "F" "F" "F" "F" "F" "F" "F" "F" "F"

## in this analysis, only specific annotations should be included as others are extracellular
Rappez_et_al$cellular[1:10]
#>     C10H10N4O.K      C10H10O6.K  C10H11N4O7P.Na    C10H11N5O3.H  C10H12ClN5O3.K
#>           FALSE           FALSE           FALSE            TRUE           FALSE
#> C10H12ClN5O3.Na   C10H12ClNO4.H   C10H12ClNO4.K  C10H12ClNO4.Na   C10H12FN5O4.H
#>           FALSE           FALSE           FALSE           FALSE           FALSE
```

enrichment analysis with isomers

The main enrichment object can be generated as follows. By default, `bmetenrichr` uses LION (Molenaar et al., 2019, GigaScience, https://doi.org/10.1093/gigascience/giz061) as metabolite sets.

```
myTestRun <-
  initEnrichment(scmatrix = Rappez_et_al$sc_matrix,
                 annotations = rownames(Rappez_et_al$sc_matrix),
                 conditions = Rappez_et_al$conditions,
                 include = Rappez_et_al$cellular,
                 condition.x = "U",
                 condition.y = "F"                    )
#>
#> Parsing isomers...
#> single-cell metabolomics matrix of 3385 metabolites and 8807 cells
#> active pathway: LION
#>
#> conditions: F, FI, FIT, U
#>
#> condition.x: U
#> condition.y: F
```

First, metabolites are ranked by `rankScore()`

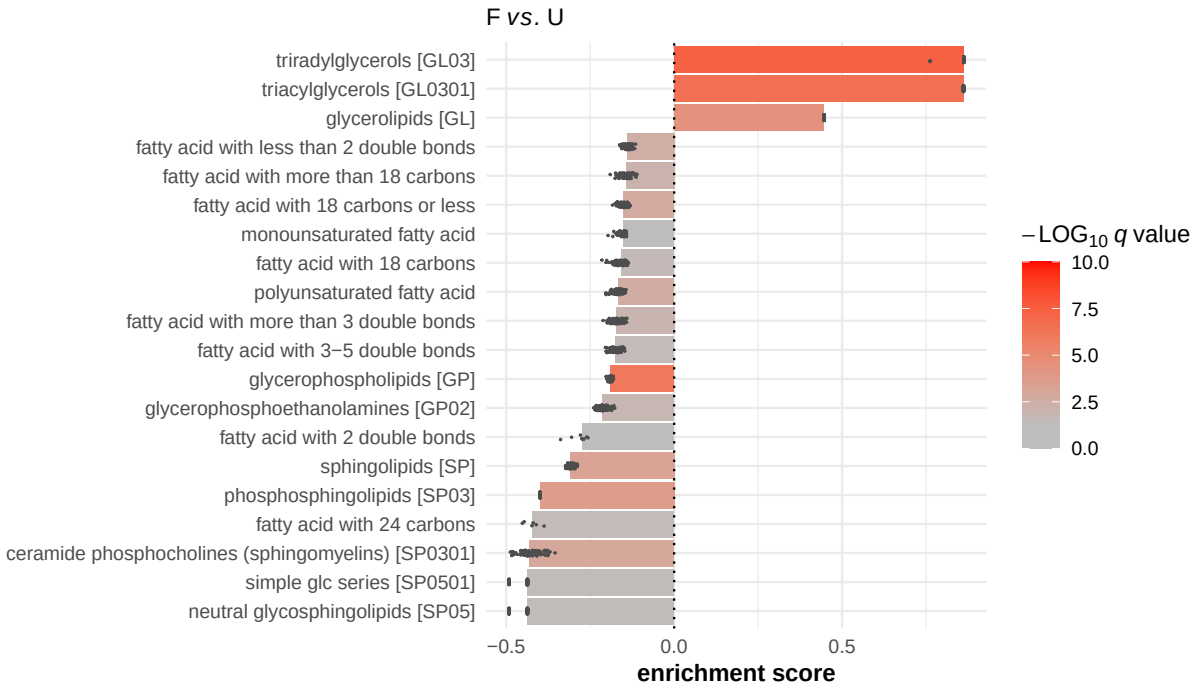
```
## rank metabolites, in this case by t.test statistic

myTestRun <- rankScore(myTestRun, ranking.by = 't.test')
#> number of ties: 19 (1.73%)
```

Then, perform enrichment analysis with n = 100 bootstraps

```
myTestRun <- calcEnrichment(myTestRun, n = 100)
#>
#> Bootstrapping...
#>
#> Match to pathway...
#> 35.34% of annotations were matched to pathway
#>
#> Perform enrichment analysis...
```

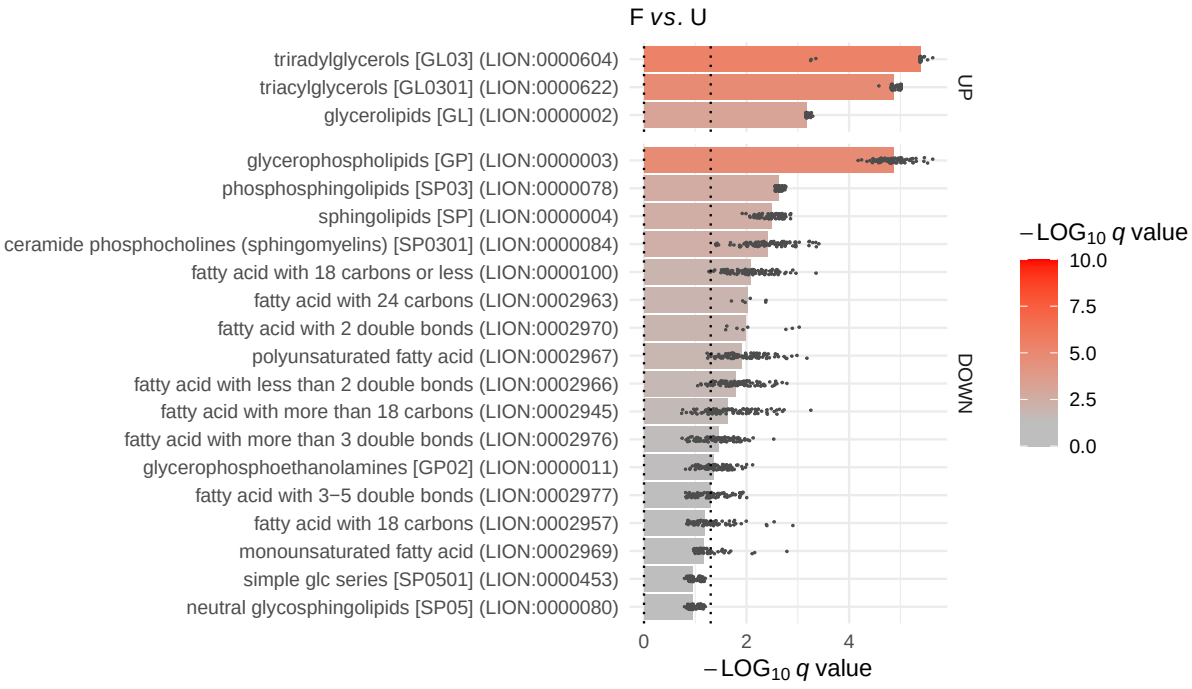
The enrichment analysis can be visualized with `plotEnrichment()`, here with enrichment score (ES) on x-axis

```
plotEnrichment(myTestRun, min.annotations = 5, q.value.cutoff = .05, by.statistic = "ES")
```



Plots can also be arranged with q.values on x-axis, and with LION IDs

```
plotEnrichment(myTestRun, min.annotations = 5, q.value.cutoff = .05, plotIDs = T,
               by.statistic = "q.value")
```



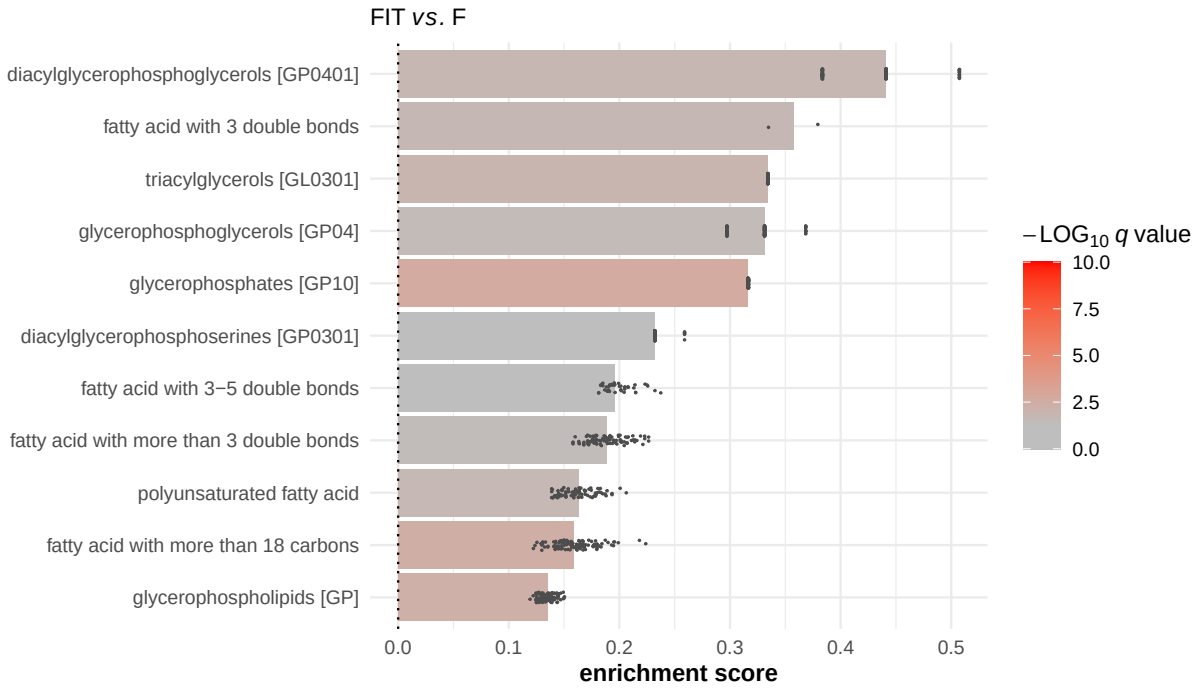
The enrichment analysis can also be exported as `data.frame`:

```
enrichmentTable(myTestRun)[1:10,]
#>         LION_ID                                                LION_name     n
#> 1  LION:0000002                                    glycerolipids [GL]  28.0
#> 2  LION:0000003                               glycerophospholipids [GP] 290.5
#> 3  LION:0000004                                    sphingolipids [SP]  49.0
#> 4  LION:0000010                          glycerophosphocholines [GP01]  54.0
#> 5  LION:0000011                     glycerophosphoethanolamines [GP02]  68.5
#> 6  LION:0000012                         glycerophosphoinositols [GP06]  50.0
#> 7  LION:0000013                           glycerophosphoserines [GP03]  49.0
#> 8  LION:0000014                         glycerophosphoglycerols [GP04]  26.0
#> 9  LION:0000030                  diacylglycerophosphocholines [GP0101]  38.0
#> 10 LION:0000032 1-(1z-alkenyl),2-acylglycerophosphocholines [GP0103]   4.0
#>     ES_median       ES_sd p.value_median   p.value_sd q.value_median
#> 1   0.4463952 0.000000000   3.787071e-05 0.000000e+00   6.740987e-04
#> 2  -0.1891438 0.004319113   4.516093e-07 3.854567e-07   1.378359e-05
#> 3  -0.3102381 0.008198835   2.908494e-04 1.888309e-04   3.199343e-03
#> 4  -0.1774058 0.022117846   6.841570e-02 5.347221e-02   2.034424e-01
#> 5  -0.2123365 0.013985159   7.654728e-03 8.713201e-03   4.445395e-02
#> 6  -0.1737786 0.000000000   1.119827e-01 0.000000e+00   2.723902e-01
#> 7  -0.2061049 0.009185895   3.747630e-02 1.567312e-02   1.452873e-01
#> 8  -0.2118250 0.018913394   2.428458e-01 1.083380e-01   4.740169e-01
#> 9  -0.2188399 0.027117832   6.351120e-02 6.792230e-02   1.857073e-01
#> 10  0.4373445 0.267906826   2.968817e-01 3.091787e-01   4.963781e-01
#>       q.value_sd
#> 1   3.345144e-05
#> 2   1.070033e-05
#> 3   1.931130e-03
#> 4   9.334177e-02
#> 5   3.332223e-02
```

```
#> 6  1.757933e-02
#> 7  3.696830e-02
#> 8  1.076772e-01
#> 9  1.104868e-01
#> 10 3.028986e-01
```

To compare other conditions, use `setConditions()`:

```
## now, let's test FIT vs F

myTestRun <-  setConditions(object = myTestRun, condition.x = 'F', condition.y = 'FIT')
myTestRun
#> single-cell metabolomics matrix of 3385 metabolites and 8807 cells
#> active pathway: LION
#>
#> conditions: F, FI, FIT, U
#>
#> condition.x: F
#> condition.y: FIT

## rank metabolites, in this case by t.test statistic

myTestRun <- rankScore(myTestRun, ranking.by = 't.test')
#> number of ties: 24 (2.19%)

## and perform enrichment analysis

myTestRun <- calcEnrichment(myTestRun, n = 100)
#>
#> Bootstrapping...
#>
#> Match to pathway...
#> 34.52% of annotations were matched to pathway
#>
#> Perform enrichment analysis...

plotEnrichment(myTestRun, min.annotations = 5, q.value.cutoff = .05)
```



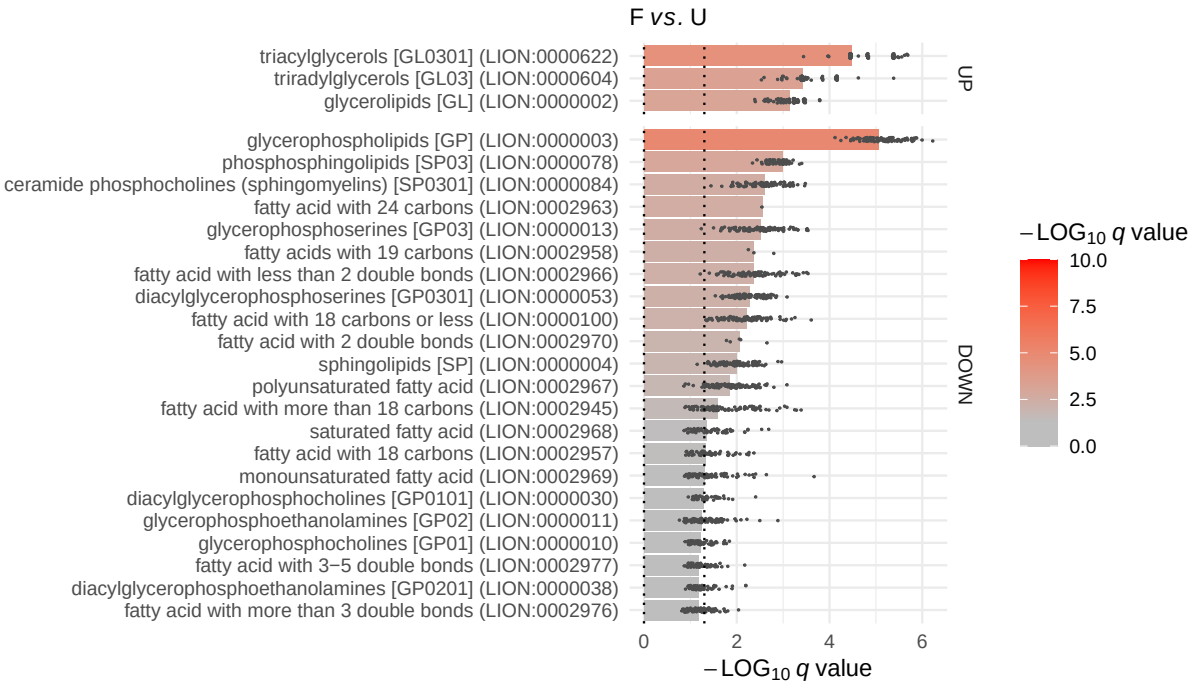
enrichment analysis with isomers and isobars

By default, only isomers are included. With isobars = TRUE, it's also possible to include isobars within a set m/z range:

```
## create object

myTestRun <-
  initEnrichment(scmatrix = Rappez_et_al$sc_matrix,
                 isobars = TRUE,                      ## to include isobars (default is FALSE)
                 mass_range_ppm = 3,                  ## mass range to define isobars
                 polarization_mode = "positive",      ## mode is important to include the right adducts
                 annotations = rownames(Rappez_et_al$sc_matrix),
                 conditions = Rappez_et_al$conditions,
                 include = Rappez_et_al$cellular,
                 condition.x = "U",
                 condition.y = "F"                    )
#> polarization_mode is: positive
#>
#> Parsing isomers...
#> Parsing potential isobars...
#>
#> single-cell metabolomics matrix of 3385 metabolites and 8807 cells
#> active pathway: LION
#>
#> conditions: F, FI, FIT, U
#>
#> condition.x: U
#> condition.y: F
```

Downstream, the same workflow can be used:

```r
## rank metabolites, in this case by t.test statistic

myTestRun <- rankScore(myTestRun, ranking.by = 't.test')
#> number of ties: 19 (1.73%)

## perform enrichment analysis with n = 100 bootstraps

myTestRun <- calcEnrichment(myTestRun, n = 100)
#>
#> Bootstrapping...
#>
#> Match to pathway...
#> 35.43% of annotations were matched to pathway
#>
#> Perform enrichment analysis...

## example of the annotations, that now also include isobars

myTestRun$annotations[[sample(which(sapply(myTestRun$isobars_list, length) > 1), size = 1)]][1:10]
#>                                          isomer1
#> "PA(O-20:1(11Z)/28:6(10Z,13Z,16Z,19Z,22Z,25Z))"
#>                                          isomer2
#> "PA(O-18:1(11Z)/30:6(12Z,15Z,18Z,21Z,24Z,27Z))"
#>                                          isomer3
#>  "PA(O-22:1(13Z)/26:6(8Z,11Z,14Z,17Z,20Z,23Z))"
#>                                          isomer4
#> "PA(O-18:1(13Z)/30:6(12Z,15Z,18Z,21Z,24Z,27Z))"
#>                                          isomer5
#>  "PA(O-18:2(9Z,12Z)/30:5(12Z,15Z,18Z,21Z,24Z))"
#>                                          isomer6
#>  "PA(O-18:2(9Z,12Z)/30:5(15Z,18Z,21Z,24Z,27Z))"
#>                                          isomer7
#>  "PA(O-18:1(9Z)/30:6(12Z,15Z,18Z,21Z,24Z,27Z))"
#>                                          isomer8
#>     "PA(P-20:1(11Z)/28:5(10Z,13Z,16Z,19Z,22Z))"
#>                                          isomer9
#>     "PA(P-20:1(11Z)/28:5(13Z,16Z,19Z,22Z,25Z))"
#>                                         isomer10
#>     "PA(P-18:1(11Z)/30:5(12Z,15Z,18Z,21Z,24Z))"

## plot enrichment analysis, with q.values on x-axis, and with LION IDs

plotEnrichment(myTestRun, min.annotations = 5, q.value.cutoff = .05, plotIDs = T, by.statistic = "q.valu
```


F vs. U
triacylglycerols [GL0301] (LION:0000622)
triradylglycerols [GL03] (LION:0000604)
glycerolipids [GL] (LION:0000002)
glycerophospholipids [GP] (LION:0000003)
phosphosphingolipids [SP03] (LION:0000078)
ceramide phosphocholines (sphingomyelins) [SP0301] (LION:0000084)
fatty acid with 24 carbons (LION:0002963)
glycerophosphoserines [GP03] (LION:0000013)
fatty acids with 19 carbons (LION:0002958)
fatty acid with less than 2 double bonds (LION:0002966)
diacylglycerophosphoserines [GP0301] (LION:0000053)
fatty acid with 18 carbons or less (LION:0000100)
fatty acid with 2 double bonds (LION:0002970)
sphingolipids [SP] (LION:0000004)
polyunsaturated fatty acid (LION:0002967)
fatty acid with more than 18 carbons (LION:0002945)
saturated fatty acid (LION:0002968)
fatty acid with 18 carbons (LION:0002957)
monounsaturated fatty acid (LION:0002969)
diacylglycerophosphocholines [GP0101] (LION:0000030)
glycerophosphoethanolamines [GP02] (LION:0000011)
glycerophosphocholines [GP01] (LION:0000010)
fatty acid with 3−5 double bonds (LION:0002977)
diacylglycerophosphoethanolamines [GP0201] (LION:0000038)
fatty acid with more than 3 double bonds (LION:0002976)
UP
DOWN
0
2
4
6
−LOG10 q value
−LOG10 q value
10.0
7.5
5.0
2.5
0.0