co-IP task

Dunja Petrovic

October 6, 2020

Read and preprocess data

This piece of code is used to load data file and perform folowing cleaning steps:

- filter our _RV and _CON proteins
- keep only the first instance of gene name
- remove one hit wonders
- log2 transform of LFG intensity

```
## Read and clean data
df = readandclean("data/proteinGroups_166_Bioinf.txt")
```

Apply filtering for rows without valid measurments (remove all empty rows)

```
df.F = filter_valids(df,
                     conditions = c("H0", "Hc"),
                     min_count = c(2, 2),
                     at_least_one = TRUE)
```

Remove rows where KEEP is FALSE

```
df.F = filter(df.F, KEEP)
```

Imputation and Normalization

Since missing values are associated with proteins with low levels of expression, we can substitute the missing values with numbers that are considered “small” in each sample. We can define this statistically by drawing from a normal distribution with a mean that is down-shifted from the sample mean and a standard deviation that is a fraction of the standard deviation of the sample distribution.

Reference link

- filling data gaps (zeros) if min 2/3 data per protein is valid (non zeros)
- data normalization (median centering)

```
## Normalize data
df.FN = median_centering(df.F)

## Apply imputation
df.FNI = impute_data(df.FN)
```

Statistical analysis

In this step we are ready to compare protein expression between the two sample groups, H0 and Hc. According to literarure, this is done with LIMMA R package, originally developed for RNA microarray experiments. The package also contains visualizations for exploring proteins of significance.

```
# define treatment and control groups for two group comparison, assuming 3 cases and 3 controls
ct <- c("LOG2.H0_1", "LOG2.H0_2", "LOG2.H0_3")
tr <- c("LOG2.Hc_1", "LOG2.Hc_2", "LOG2.Hc_3")


fit = MarayLM(df.FNI,tr,ct)
result = fitlimma(df.FNI,tr,ct)
signproteins = result  %>%  # print significant protein list based on p < 0.005
  filter (p.val < 0.005 & logFC >3)
```

Significant proteins (p.val < 0.005 & logFC >3)

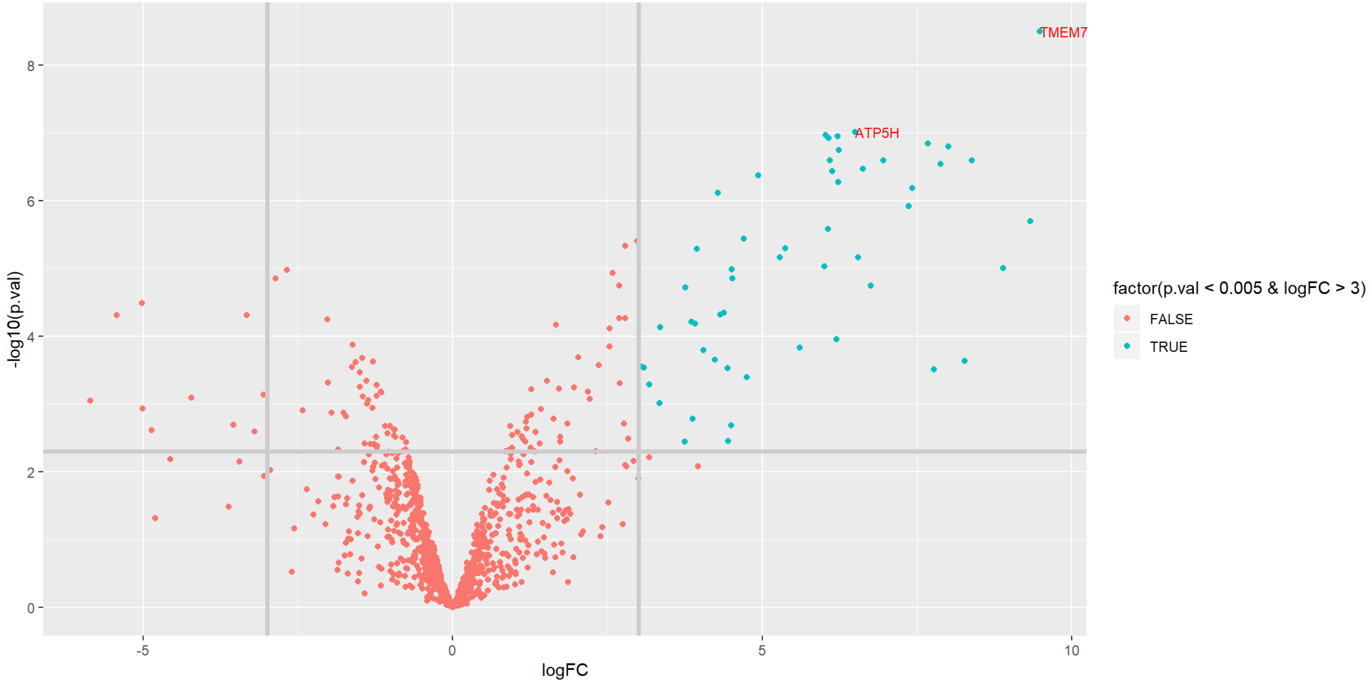
This is ordered list of proteins fulfilling criteria: p value < 0.005 and log fold change > 3

```
##     Gene.names   LOG2.H0_1   LOG2.H0_2    LOG2.H0_3   LOG2.Hc_1   LOG2.Hc_2
## 1       TMEM70 -4.76560556 -4.49466939 -4.489492836  4.77215310  4.84944497
## 2        ATP5H  1.87333469  2.44218846  2.272694683  8.82534190  8.46228064
## 3       ATP5A1  5.21058197  5.74067262  5.485702058 11.58283930 11.31247735
## 4        ATP5L  1.39201043  1.61368864  1.168898748  7.80534404  7.32126278
## 5       ATP5C1  1.86002052  2.12560155  1.830639724  8.17969977  7.68101383
## 6      MT-ATP8 -1.86431657 -2.12340162 -2.590277340  5.86702424  5.12903705
## 7         Cxxx -3.57443853 -4.62760862 -4.444144788  3.76811398  3.73952615
## 8       ATP5F1  1.40075608  2.00716310  1.766036352  8.11465644  7.67144914
## 9      LRRC10B -5.00391993 -4.58999429 -4.208669947  3.28479593  3.73389084
## 10       ATP5B  5.35073032  6.15922974  5.745055973 11.80513873 11.76236708
## 11       ATP5I -0.13615631  0.52404046 -0.338295706  7.24969819  6.74610213
## 12      PDGFRB -4.16366819 -3.71321640 -3.895899916  4.11483508  3.30215135
## 13       ATP5J  0.43422701  1.29246155  1.247029467  7.73784385  7.40495586
## 14       ATP5D  0.02990964  0.54082636  0.804602519  6.40587256  6.49563911
## 15      TIMM13 -3.41914548 -3.68771140 -3.780010605  1.00911577  1.38501013
## 16       ATP5O  2.51672101  3.30296916  2.916508771  9.42904438  8.75344633
## 17     NDUFB11 -3.24856390 -3.14993056 -3.100810524  4.43874398  3.49202698
## 18     FAM162A -1.30236689 -0.91785811 -1.417052430  2.90411926  3.16518720
## 19       ETFDH -4.49693109 -5.47669193 -4.910501818  1.70661913  2.79779868
## 20       TMxxx -4.82945414 -2.63208753 -4.087741242  5.34830083  5.47481519
## 21      TIMM8A -3.19034823 -4.59480836 -4.019553391  2.26178263  1.98139296
## 22        POLG -3.43225026 -4.37354793 -3.877161046  0.48145088  0.88762679
## 23      ATP5G1 -0.17882641 -0.30695703 -0.757031071  4.24823854  5.49879231
## 24      TIMM21 -4.42345042 -4.13901551 -4.003711667 -0.66966264 -0.22570899
## 25       USMG5 -0.32805242  1.48704515  0.290573210  7.22214781  6.76996082
## 26      AFG3L2 -2.90183459 -1.92215807 -1.882682173  2.66931506  2.92503533
## 27      ATP5J2 -0.60885110 -0.08687797 -0.251531315  4.68529661  6.12635110
## 28     C14orf2 -3.93279103 -4.39677971 -4.751137919  3.90816788  3.60025809
## 29     DNAJC19 -3.76572114 -4.15465864 -4.702451655  0.24105110  0.78618348
## 30       ATP5B -5.39671232 -4.18994430 -4.189078445 -0.10542924 -0.17155948
## 31     ATP5EP2 -2.92950304 -0.86247024 -1.006477619  5.22822224  4.87875960
## 32       GHITM -2.80241571 -2.45610888 -2.759731211  0.44348685  1.32708365
## 33    TMEM126A -4.28243298 -4.26330079 -5.283044281 -0.92629815 -0.07624287
## 34       ABCB7 -4.76322801 -3.84986245 -4.304815919 -0.42712107  0.85930801
## 35     C2orf47 -4.10578547 -3.18570685 -4.727766129 -0.08588635 -0.03918496
## 36      TIMM44 -5.20223105 -5.52364016 -4.194804032 -1.41945284 -0.57591345
## 37       HADHB -0.66761181  0.39427593  0.001095025  3.26174933  3.70093324
## 38       LETM1 -3.05392111 -0.14372198 -1.650365248  4.34871247  5.02364339
## 39       HADHA -2.49192112  0.05135940 -0.591196487  4.64800324  5.08940868
## 40       SQRDL -4.67796550 -3.10418544 -4.127169901  0.80429396 -0.21077941
## 41      HIGD1A -5.87745942 -4.31391139 -3.800032515 -0.64697432 -0.02605781
## 42     MT-ATP6 -4.29351450 -0.08905011 -0.551206247  6.80015764  6.45036124
## 43       COX15 -4.05408146 -4.71518942 -5.198788836 -1.86989209 -0.95312854
## 44     FAM210A -4.40371039 -4.47711060 -4.689846672 -2.38277316 -0.94905879
## 45      YME1L1 -5.04404088 -3.35261328 -3.420886092 -0.31786068  0.49332005
## 46        HCCS -6.29665064 -2.45292754 -6.475514360  2.36420640  2.89599833
## 47        AGPS -1.68046149 -4.22166068 -3.828653365  2.03859342  1.23350632
## 48     SLC35A4 -4.04766724 -2.47616420 -3.170535447 -0.68127008  0.45598432
## 49       ADPGK -5.31782941 -4.26292114 -3.442899483 -1.80488121 -0.64655359
## 50      ZNF608 -5.70063061 -4.54013300 -5.115950109 -2.03773871 -2.08493158
## 51     TOMM70A -3.86817979 -4.81691045 -5.722585065 -2.16810122  0.42315307
## 52       ROMO1 -5.29980308 -1.38348271 -4.408472579  0.61583079  0.60104507
## 53      TMEM65 -3.24967826 -3.41814362 -5.943891214 -1.03757475  0.52547961
##       LOG2.Hc_3    logFC        p.val
## 1    5.08025782 9.483875 3.193924e-09
## 2    8.80010506 6.499837 9.817697e-08
## 3   11.61334481 6.023902 1.080430e-07
## 4    7.70802874 6.220013 1.116212e-07
## 5    8.18149859 6.075317 1.202421e-07
## 6    5.44968913 7.674582 1.438284e-07
## 7    3.87911418 8.010982 1.585906e-07
## 8    8.10785588 6.240002 1.805624e-07
## 9    4.35535974 8.392210 2.539581e-07
## 10 11.98041594 6.097635 2.539669e-07
## 11   6.92358329 6.956598 2.552303e-07
## 12   4.45323947 7.881003 2.889903e-07
## 13   7.72112603 6.630069 3.418742e-07
## 14   6.88576651 6.137313 3.679586e-07
## 15   1.54134444 4.940779 4.285123e-07
## 16   9.25209494 6.232796 5.339760e-07
## 17   4.83820350 7.422760 6.494331e-07
## 18  3.14562717 4.284070 7.754614e-07
## 19  2.70956032 7.366034 1.204783e-06
## 20  5.62831826 9.333572 2.002817e-06
## 21  2.15710176 6.068329 2.618946e-06
## 22  1.04892309 4.700320 3.689212e-06
## 23  5.12286500 5.370903 5.108347e-06
## 24  0.15010036 3.940302 5.156825e-06
## 25  7.11785211 6.553465 6.848665e-06
## 26  3.57073262 5.290586 6.892189e-06
## 27  6.27432229 6.011077 9.443176e-06
## 28  6.10550755 8.898214 9.997421e-06
## 29 -0.13182646 4.506080 1.033344e-05
## 30  0.06378553 4.520844 1.412773e-05
## 31  5.35879478 6.754743 1.810507e-05
## 32  1.49021591 3.759681 1.901161e-05
## 33  0.32835451 4.384864 4.573958e-05
## 34 -0.37981493 4.323426 4.809598e-05
## 35 -0.32547243 3.856238 6.186320e-05
## 36 -1.16486261 3.920149 6.596964e-05
## 37  2.82789880 3.354274 7.402391e-05
## 38  4.38911585 6.203160 1.123229e-04
## 39  4.05497213 5.608047 1.472832e-04
## 40 -0.34599737 4.052279 1.625309e-04
## 41 -0.61732421 4.233682 2.254625e-04
## 42  6.62545276 8.269914 2.323387e-04
## 43 -2.01971527 3.041775 2.782854e-04
## 44 -0.96135978 3.092492 2.892234e-04
## 45  1.32570990 4.439570 2.955621e-04
## 46  2.83992794 7.775075 3.127556e-04
## 47  1.26046985 4.754448 4.052940e-04
## 48  0.06138369 3.176822 5.219740e-04
## 49 -0.54140719 3.343603 9.673566e-04
## 50  0.39201685 3.875353 1.652397e-03
## 51  0.82292170 4.495216 2.065726e-03
## 52  1.04915772 4.452597 3.543980e-03
## 53 -0.85778176 3.747279 3.587418e-03
```

Results visualization

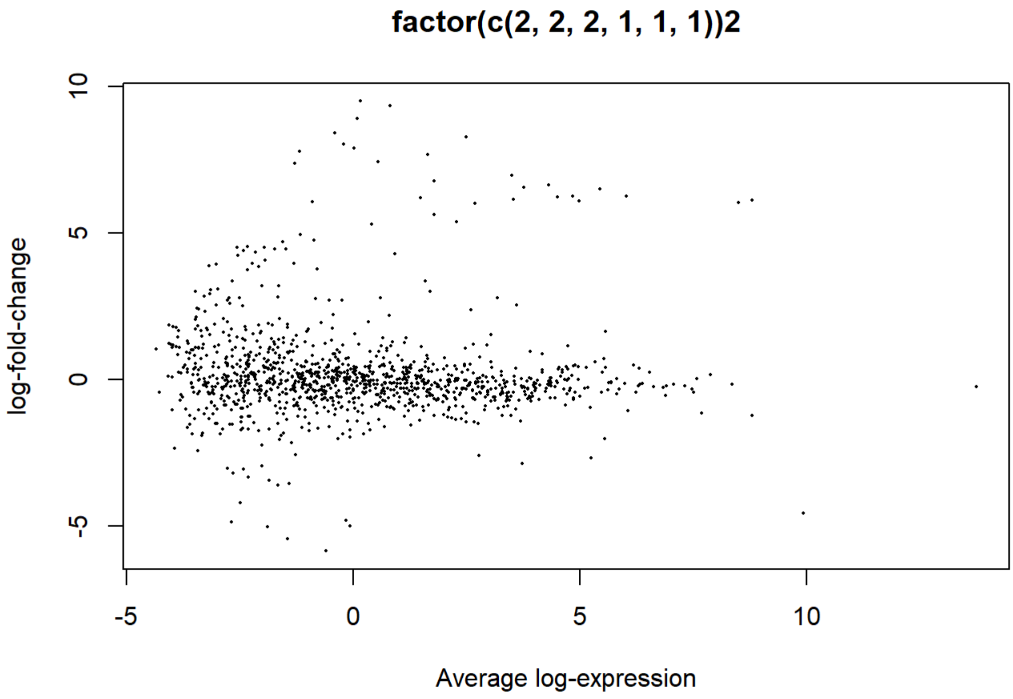
Volcano plot (logFC~-log10(p.value))

Significant proteins are marked blue and top two proteins are annotated with their names

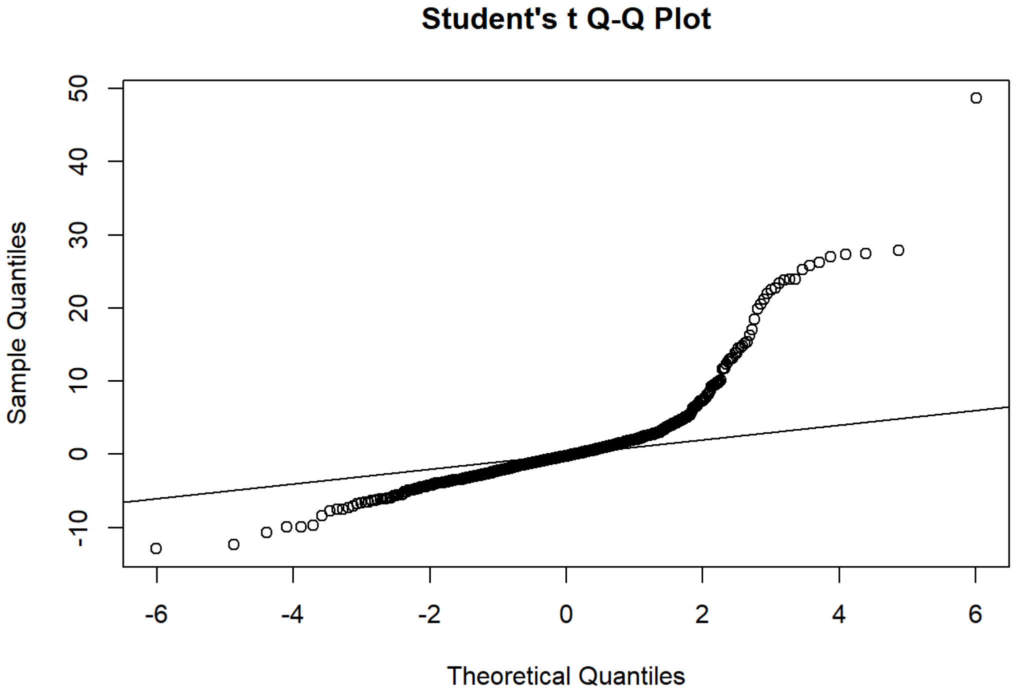


MD PLot

A mean-difference plot (MD-plot) is a plot of log-intensity ratios (differences) versus log-intensity averages (means) between samples



QQ Plot

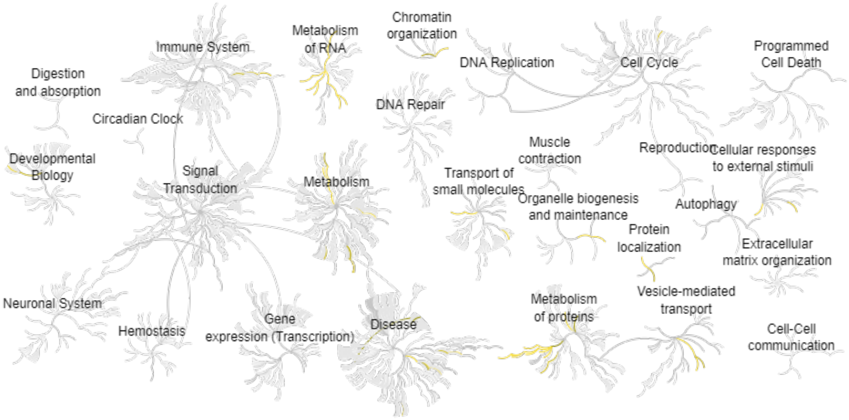


Enrichment analysis

Due to missing required R packages, pathway analysis done online via [link] (https://reactome.org/PathwayBrowser/#/)

Pathway enrichment is performed on significant list of genes from signproteins table

Full report available in attached file: reactome_report.pdf


Immune System
Metabolism of RNA
Chromatin organization
DNA Replication
Cell Cycle
Programmed Cell Death
Digestion and absorption
Circadian Clock
DNA Repair
Developmental Biology
Signal Transduction
Metabolism
Muscle contraction
Reproduction
Cellular responses to external stimuli
Transport of small molecules
Organelle biogenesis and maintenance
Autophagy
Protein localization
Extracellular matrix organization
Neuronal System
Hemostasis
Gene expression (Transcription)
Disease
Metabolism of proteins
Vesicle-mediated transport
Cell-Cell communication