Exploring proteomics for DREAM

Attila Gabor

4/29/2019

Checking the proteomics data for the DREAM challenge.

```
proteomics_raw <- readRDS("data/proteomics/MSstat_groupComparison_selceted.rds") %>% as_tibble()
proteomics_raw
```

```
## # A tibble: 559,922 x 11
##    Protein Label   log2FC     SE Tvalue    DF    pvalue adj.pvalue issue
##    <fct>   <fct>    <dbl>  <dbl>  <dbl> <dbl>     <dbl>      <dbl> <fct>
##  1 A0A024~ norm~    2.28   0.311   7.34   129  2.14e-11   1.54e-10 <NA>
##  2 A0A087~ norm~  Inf     NA      NA       NA  NA          0.       oneC~
##  3 A0A087~ norm~    0.535  0.208   2.58   111  1.13e- 2   2.16e- 2 <NA>
##  4 A0A096~ norm~    0.666  0.447   1.49    56  1.42e- 1   2.03e- 1 <NA>
##  5 A0A0A6~ norm~    1.22   0.547   2.23    58  2.99e- 2   5.15e- 2 <NA>
##  6 A0A0B4~ norm~    1.27   0.209   6.09   129  1.21e- 8   6.22e- 8 <NA>
##  7 A0A0B4~ norm~   NA     NA      NA       NA  NA         NA        comp~
##  8 A0A0G2~ norm~   NA     NA      NA       NA  NA         NA        comp~
##  9 A0A0U1~ norm~ -Inf     NA      NA       NA  NA          0.       oneC~
## 10 A0A0U1~ norm~    0.986  0.744   1.32    84  1.89e- 1   2.59e- 1 <NA>
## # ... with 559,912 more rows, and 2 more variables:
## #   MissingPercentage <dbl>, ImputationPercentage <dbl>
```

```
str(proteomics_raw)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    559922 obs. of  11 variables:
##  $ Protein             : Factor w/ 9031 levels "A0A024RBG1","A0A087WUL8;P0DPF2;P0DPF3;Q3BBV2;Q6P3W6;
##  $ Label               : Factor w/ 62 levels "normal_vs_AU565",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ log2FC              : num  2.281 Inf 0.535 0.666 1.217 ...
##  $ SE                  : num  0.311 NA 0.208 0.447 0.547 ...
##  $ Tvalue              : num  7.34 NA 2.58 1.49 2.23 ...
##  $ DF                  : num  129 NA 111 56 58 129 NA NA NA 84 ...
##  $ pvalue              : num  2.14e-11 NA 1.13e-02 1.42e-01 2.99e-02 ...
##  $ adj.pvalue          : num  1.54e-10 0.00 2.16e-02 2.03e-01 5.15e-02 ...
##  $ issue               : Factor w/ 2 levels "oneConditionMissing",..: NA 1 NA NA NA NA 2 2 1 NA ...
##  $ MissingPercentage   : num  0 0.8333 0.1667 0.0556 0.6667 ...
##  $ ImputationPercentage: num  0 0 0 0 0 0 0 0 0 0 ...
```

```
unique(proteomics_raw$Label)
```

```
##  [1] normal_vs_AU565       normal_vs_BT20        normal_vs_BT474
##  [4] normal_vs_BT483       normal_vs_BT549       normal_vs_CAL120
##  [7] normal_vs_CAL148      normal_vs_CAL51       normal_vs_CAL851
## [10] normal_vs_CAMA1       normal_vs_DU4475      normal_vs_EFM19
## [13] normal_vs_EFM192A     normal_vs_EVSAT       normal_vs_HBL100
## [16] normal_vs_HCC1143     normal_vs_HCC1187     normal_vs_HCC1395
## [19] normal_vs_HCC1419     normal_vs_HCC1428     normal_vs_HCC1500
## [22] normal_vs_HCC1569     normal_vs_HCC1599     normal_vs_HCC1806
## [25] normal_vs_HCC1937     normal_vs_HCC1954     normal_vs_HCC202
## [28] normal_vs_HCC2157     normal_vs_HCC2185     normal_vs_HCC2218
## [31] normal_vs_HCC3153     normal_vs_HCC38       normal_vs_HCC70
## [34] normal_vs_HDQP1       normal_vs_Hs578T      normal_vs_JIMT1
## [37] normal_vs_KPL1        normal_vs_LY2         normal_vs_MACLS2
## [40] normal_vs_MCF7        normal_vs_MDAkb2      normal_vs_MDAMB134VI
## [43] normal_vs_MDAMB157    normal_vs_MDAMB175VII normal_vs_MDAMB231
## [46] normal_vs_MDAMB361    normal_vs_MDAMB415    normal_vs_MDAMB436
## [49] normal_vs_MDAMB453    normal_vs_MDAMB468    normal_vs_MFM223
## [52] normal_vs_MPE600      normal_vs_MX1         normal_vs_OCUBM
## [55] normal_vs_SKBR3       normal_vs_T47D        normal_vs_UACC3199
## [58] normal_vs_UACC812     normal_vs_UACC893     normal_vs_ZR751
## [61] normal_vs_ZR7530      normal_vs_ZR75B
## 62 Levels: normal_vs_AU565 normal_vs_BT20 ... normal_vs_ZR75B
```

Columns: - Protein: uniprotIDs divided by ";". - Why are multiple proteins in the same row? - in total 9031 protein families measured – how does it distribute across cell-lines? - Label: cell-line info

How many proteins measured per cell-line?

```
protein_cell_line_table <- table(proteomics_raw[,1:2])
all(protein_cell_line_table==1) == TRUE
```

```
## [1] TRUE
```

apparently all the proteins are measured across all cell-lines? – appears in the table, maybe some values are NA/Inf.

```
print(nrow(proteomics_raw))
```

```
## [1] 559922
```

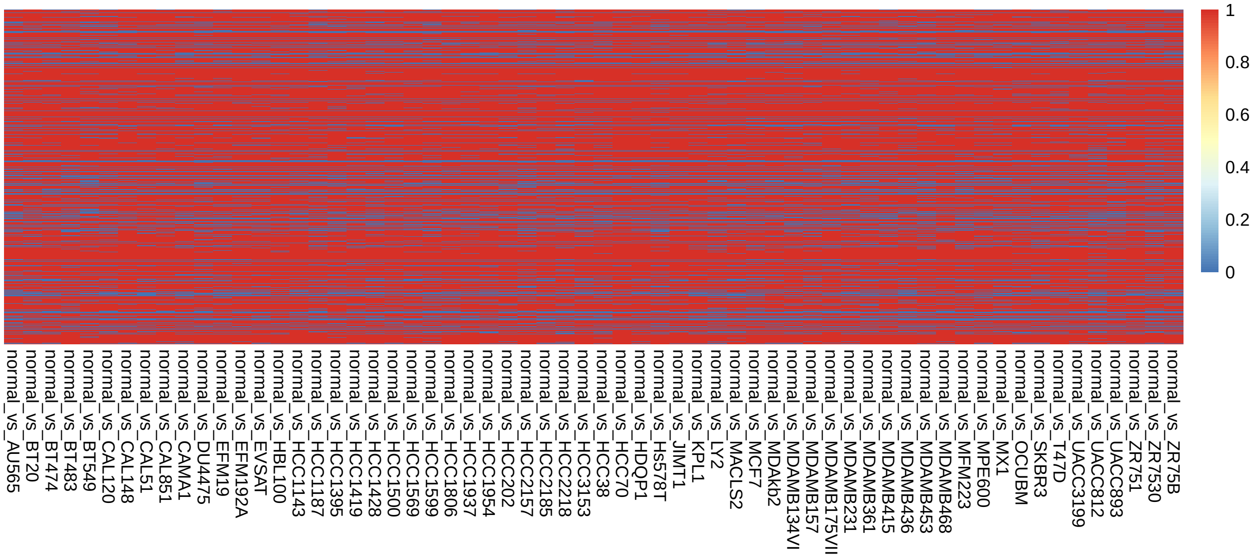
```
print(table(proteomics_raw$issue))
```

```
##
## oneConditionMissing     completeMissing
##               87274               48702
```

Print a heatmap to show missing elements:

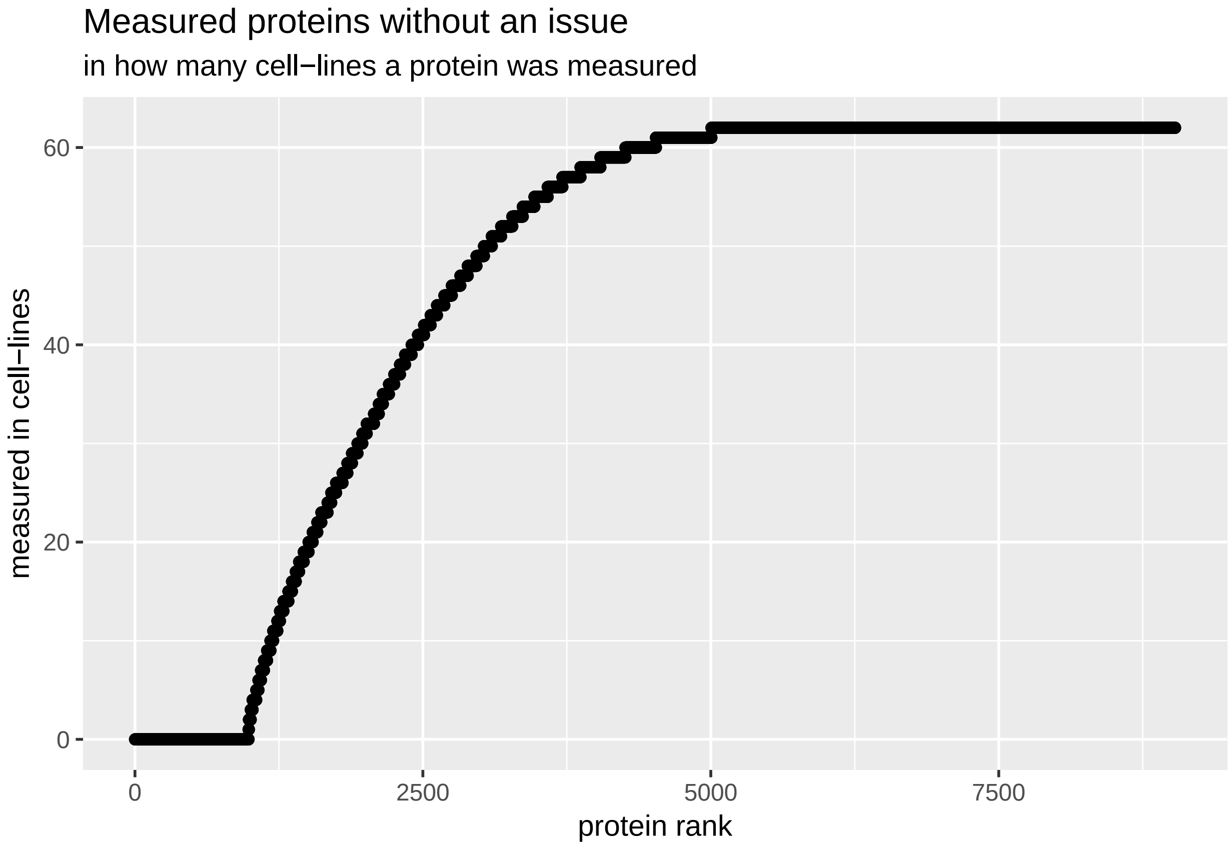
```
measured_conds <- proteomics_raw[,1:2]
measured_conds <- measured_conds[is.na(proteomics_raw$issue),]
real_proteomics_data <- table(measured_conds)
pheatmap::pheatmap(real_proteomics_data,cluster_cols = F,cluster_rows = F,labels_row = "")
```

```
## Warning: partial match of 'just' to 'justification'
```



Some proteins seems to be missing from the majority of the cell-lines:

```
rowSums(real_proteomics_data) %>% enframe() %>% arrange(value) %>% mutate(name=factor(name,levels = .$na
    ggplot() + geom_point(aes(1:length(name),value)) +
    xlab("protein rank") + ylab("measured in cell-lines") +
    ggtitle("Measured proteins without an issue",subtitle = "in how many cell-lines a protein was measu
```



```
any(proteomics_raw$ImputationPercentage>0)
```

```
## [1] FALSE
```