Sintetinės biologijos projektas

Acinetobacter baumannii bakterijų, paveiktų toksinu CheT, proteominių duomenų analizė

Beatričė Radavičiūtė

December 20, 2018

Contents

Įvadas

Mano nagrinėjami duomenys gauti toksinu CheT veiktų bakterijų *Acinetobacter baumannii* bei kontrolinių bakterijų (nepaveiktų toksinu) lizatus tyrus masių spektrometrijos metodu. Visi tyrimai atlikti Gyvybės mokslų centre. Darbe pateikiami jau apdotori duomenys, kuriuose nurodoma nustatytų bakterijose peptidų pavidinimai, jų raiškos pokytis, lyginant su kontrolinėmis bakterijomis bei kitos charakteristikos. Šio darbo tikslas yra atlikti intamųjų aprošomosios statiskos analizę. Taip pat darbo metu bus bandoma nustatyti, ar tarp kintamųjų (nustatytų baltmų charakteristikų) yra koreliacija, priežastinis ryšys ir kt.

Duomenų įsikėlimas, peržiūra, pradinė analizė

Duomenų peržiūra

Įsikeliami duomenys, nustatomas charakteristikų (variables) ir nustatytų baltymų (observations) skaičius, kintamųjų tipas.

```
## Observations: 1,344
## Variables: 20
## $ description         <chr> "Acyl-CoA dehydrogenase OS=Acinetobacter b...
## $ IEP                 <dbl> 5.54, 5.30, 4.69, 4.84, 4.53, 9.87, 5.24, ...
## $ mw                  <dbl> 65872.10, 76792.25, 37183.58, 22315.70, 36...
## $ `max score`         <dbl> 3658.1450, 779.0961, 255.2396, 1830.3930, ...
## $ accession           <chr> "A0A1G5LU08", "A0A241YAV2", "A0A1C9CQK6", ...
## $ `reported peptides` <dbl> 3, 16, 3, 7, 13, 7, 11, 6, 25, 9, 7, 11, 6...
## $ `sequence coverage` <dbl> 7.49, 38.96, 13.86, 55.72, 45.59, 55.35, 5...
## $ `FDR level`         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ entry               <chr> "A0A1G5LU08_ACIBA", "A0A241YAV2_ACIBA", "A...
## $ K1                  <dbl> 7336.7442, 4498.4030, 6098.0000, 4677.4877...
## $ K2                  <dbl> 8147.7455, 5594.2167, 6321.5000, 4949.2173...
```

```
## $ K3                  <dbl> 8346.3076, 12322.5575, 4043.0000, 4994.310...
## $ T1                  <dbl> 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0....
## $ T2                  <dbl> 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0....
## $ T3                  <dbl> 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0....
## $ avg_K10             <dbl> 7943.5991, 7471.7258, 5487.5000, 4873.6717...
## $ avg_T10             <dbl> 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0....
## $ logFC               <dbl> -12.953526, -12.724876, -12.394068, -12.25...
## $ P.Value             <dbl> 4.390731e-12, 3.946895e-08, 6.940879e-10, ...
## $ adj.P.Val           <dbl> 1.180228e-09, 1.360161e-06, 3.109514e-08, ...
```

Iš viršuje pateiktos lentelės matyti, jog nustatyta 1344 peptidai, įvertintos 21 charakteristikos, kurių dauguma - skaitinės. Svarbu paminėti, kad likusios neskaitinės (šiuo atveju kategorinės) baltymų charakteristikos yra jų pavadinai bei identifikacijos numeriai (description, accession ir entry) duomenų bazėse (šiuo atveju UniProt). Su jais tolesni veiksmai nebus daromi, taigi šie kintamieji bus pašalinti o baltymai bus išrikiuoti pagal jų pavadinimą abėcėlės tvarka ir užkoduoti skaičiais. Visi kiti kintamieji yra tolydieji, todėl bus daromos jų skaitinės suvestinės ir kitos manipuliacijos.

```
## Observations: 1,344
## Variables: 18
## $ IEP                 <dbl> 4.90, 5.44, 4.87, 4.49, 7.44, 5.80, 5.59, ...
## $ mw                  <dbl> 10129.64, 41645.43, 26205.95, 19170.71, 20...
## $ `max score`         <dbl> 35042.1500, 6539.5190, 7178.4810, 444.8390...
## $ `reported peptides` <dbl> 5, 15, 13, 3, 9, 8, 15, 14, 16, 18, 23, 18...
## $ `sequence coverage` <dbl> 71.88, 46.92, 70.37, 27.91, 78.07, 41.21, ...
## $ `FDR level`         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ K1                  <dbl> 135483.333, 4461.333, 37535.667, 4089.667,...
## $ K2                  <dbl> 149256.000, 4091.667, 43926.000, 3772.333,...
## $ K3                  <dbl> 167503.667, 10768.000, 45422.667, 4156.000...
## $ T1                  <dbl> 132090.333, 5321.667, 45934.000, 4588.333,...
## $ T2                  <dbl> 175413.333, 8561.333, 49163.000, 4381.000,...
## $ T3                  <dbl> 188138.667, 16920.333, 48734.000, 4185.667...
## $ avg_K10             <dbl> 150747.667, 6440.333, 42294.778, 4006.000,...
## $ avg_T10             <dbl> 165214.111, 10267.778, 47943.667, 4385.000...
## $ logFC               <dbl> 0.121326863, 0.657090910, 0.185100563, 0.1...
## $ P.Value             <dbl> 0.4799392164, 0.2711783419, 0.1123523767, ...
## $ adj.P.Val           <dbl> 0.73216607, 0.58152015, 0.37845011, 0.4587...
## $ id                  <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...
```

Žemiau pateiktoje lentelėje išvardintos kitų kintamųjų pavadinimai ir jų reikšmės.

Table 1: Kintamųjų pavadinimai ir jų reikšmės.

Pavadinimai	Reikšmės
IEP	Izoelektrinis taškas
mw	Molekulinis svoris
max score	Didžiausia suminė jonų krūvio vertė atitinkamam peptidui
reported peptides	Nustatytų peptidų skaičius
sequence coverage	Sekos perdengimas
K1	1 kontrolinis mėginys
K2	2 kontrolinis mėginys
K3	3 kontrolinis mėginys
T1	1 tiriamasis mėginys
T2	2 tiriamasis mėginys
T3	3 tiriamasis mėginys
avg_K10	Kontrolinių mėginių vidurkis

Pavadinimai	Reikšmės
avg_T10	Tiriamųjų mėginių vidurkis
logFC	Pokyčio logaritmas, kurio pagrindas 2
P.Value	Pvertė, nusako statistinų duomenų patikimumą
adj.P.Val	Koreguota P vertė

Pradinė duomenų analizė

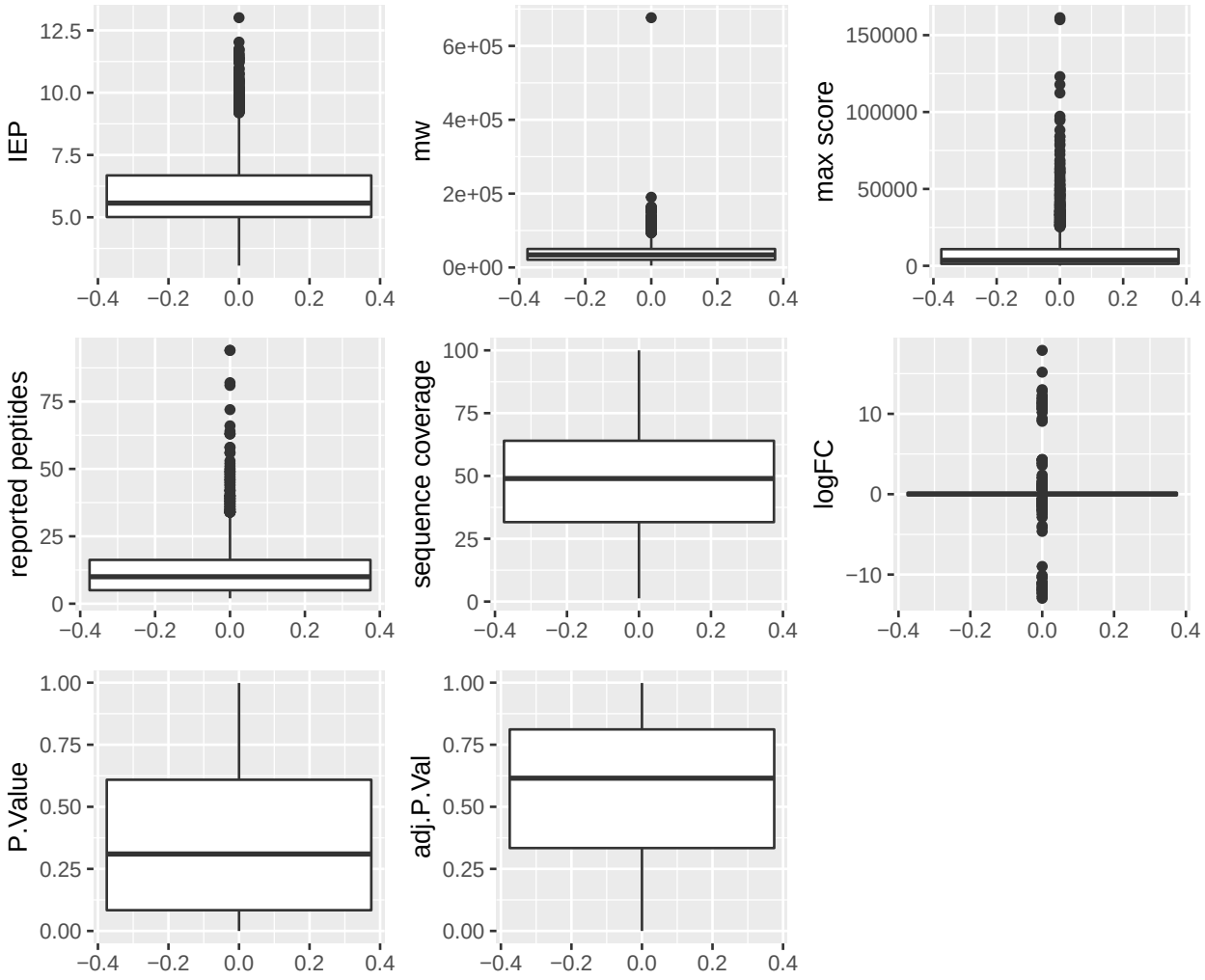
Pašalinus kategorinius kintamuosius, galima atlikti dar keletą veiksmų, kad su duomenimis būtų galima dirbti paprasčiau. Visų pirma, **FDR level** charakteristika nusako masių spektrometrijos metu nustatytų peptidų patikimumą. Kai FDR level = 0, peptidai nustatyti teisingai. Šiame darbe nagrinėjami visi peptidai, kurių FDR level = 0, taigi šią charakteristiką galima pašalinti, kadangi tai konstanta.

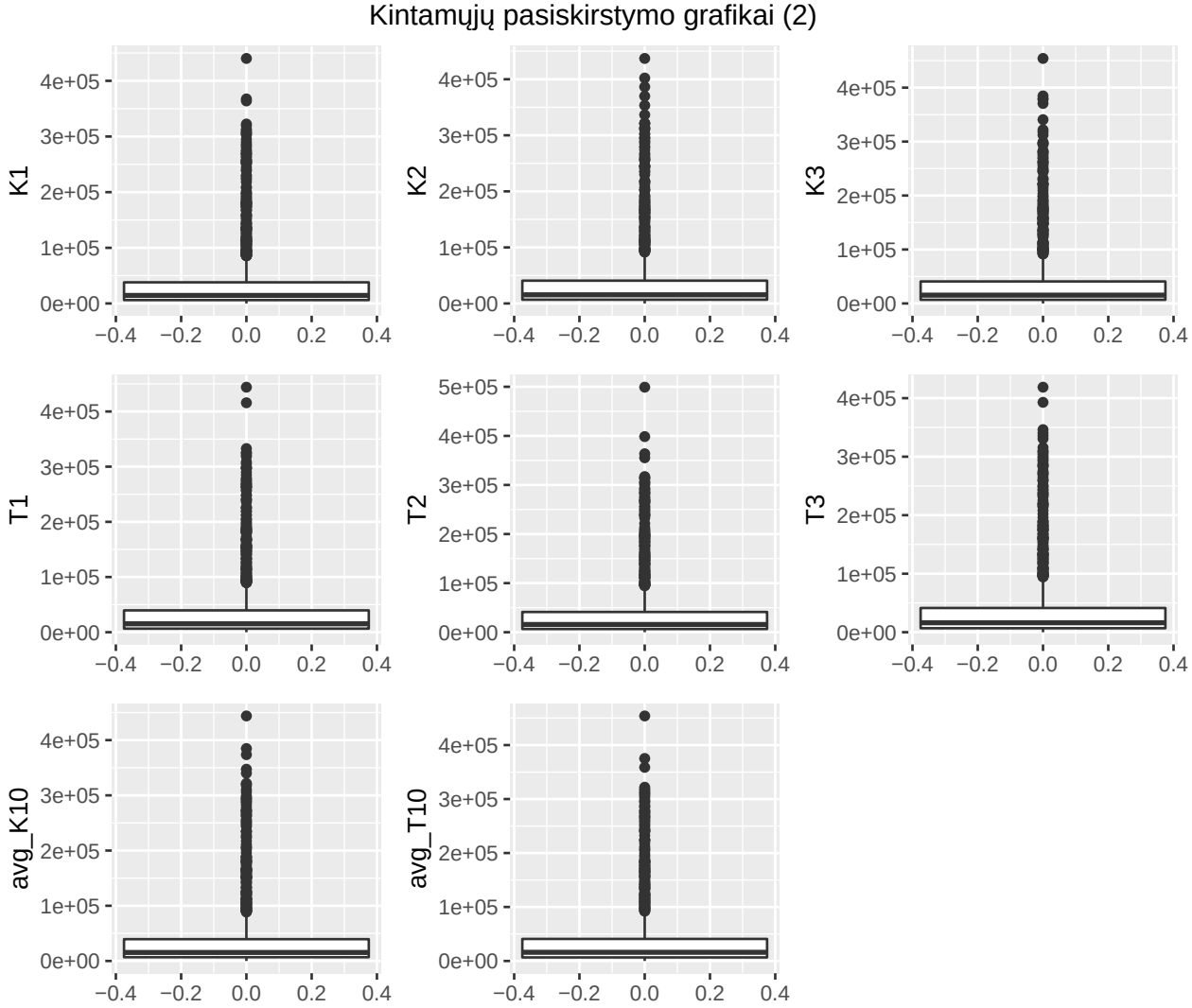
log FC charakteristika nusako nustatyrų peptidų raiškos pokytų tarp toksinu veiktų bakterijų ir kontrolinių bakterijų mėginių bei gali būti išreikšta formule 2^n, kur n = log FC vertė. Todėl pravartu šią vertę apsiskaičiuoti ir pridėti naują stulpelį į duomenų lentelę, pavadinimu “expression”.

Aprašomoji statistika

Apačioje pateiktuose paveiksluose pateikti kintamųjų grafikai, kuriuose atsispindi jų pasiskirstymas, pasiskirstymo simetrija, didžiausios ir mažiausios reikšmės, mediana, 1 ir 3 kvartiliai.

Kintamųjų pasiskirstymo grafikai (1)





Toliau pateiktoje suvestinėje yra dauguma statistinių kiekvieno kintamojo įverčių.

```
##                     vars    n     mean       sd   median  trimmed      mad
## IEP                    1 1344     6.27     1.82     5.57     6.00     1.01
## mw                     2 1344 40698.31 31960.06 34053.22 36284.75 20694.65
## max score              3 1344  9977.77 16193.56  3618.51  6285.33  4201.69
## reported peptides      4 1344    12.83    11.29    10.00    10.85     7.41
## sequence coverage      5 1344    47.76    21.25    48.97    47.92    24.20
## K1                     6 1344 36831.65 57564.76 14446.76 22856.82 15477.83
## K2                     7 1344 39038.52 60326.82 15446.17 24437.39 16412.88
## K3                     8 1344 38313.73 58893.20 15412.00 24094.14 16178.38
## T1                     9 1344 37704.87 58323.33 15254.00 23501.87 16070.18
## T2                    10 1344 38956.18 60193.04 15742.48 24383.68 16843.32
## T3                    11 1344 38879.06 59120.72 16177.50 24593.49 17054.35
## avg_K10               12 1344 38069.56 58661.10 15393.99 23864.15 16243.86
## avg_T10               13 1344 38519.83 58915.90 16051.44 24201.01 16891.47
## logFC                 14 1344     0.04     2.05     0.03     0.02     0.22
## P.Value               15 1344     0.37     0.31     0.31     0.34     0.37
## adj.P.Val             16 1344     0.56     0.29     0.62     0.58     0.35
## id                    17 1344   672.50   388.12   672.50   672.50   498.15
## expression            18 1344   249.29  6760.28     1.02     1.02     0.16
```

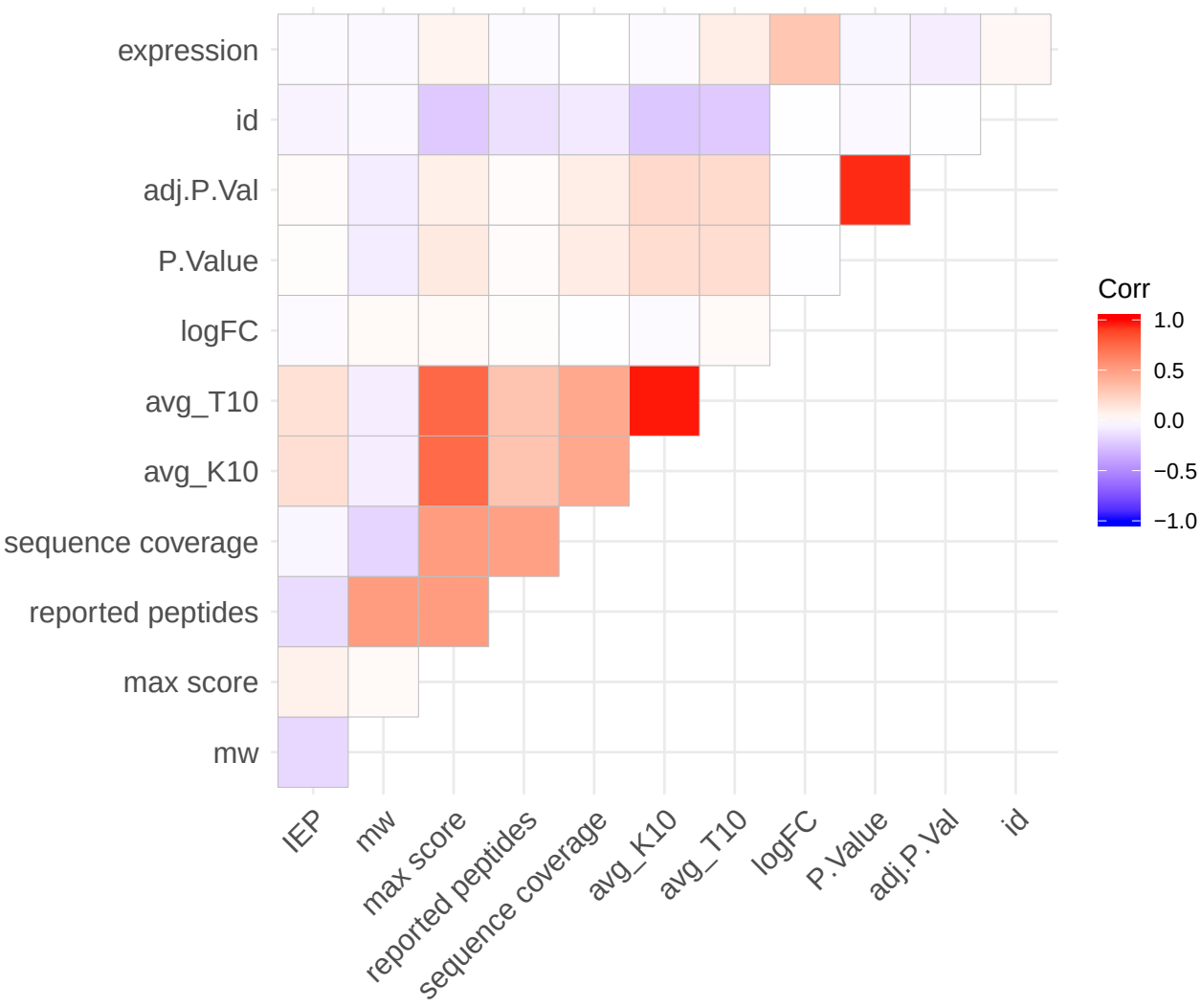
```
##                        min       max     range  skew kurtosis      se
## IEP                   3.06     13.01      9.95  1.23     0.38    0.05
## mw                 4923.75 676513.00 671589.25  6.79   116.70  871.78
## max score           123.12 161357.70 161234.58  3.59    19.07  441.72
## reported peptides     2.00     94.00     92.00  2.34     8.45    0.31
## sequence coverage     1.30    100.00     98.70 -0.07    -0.76    0.58
## K1                    0.00 440279.33 440279.33  2.97     9.86 1570.21
## K2                    0.00 437125.67 437125.67  2.97     9.87 1645.55
## K3                    0.00 454108.33 454108.33  2.99    10.28 1606.44
## T1                    0.00 444025.91 444025.91  2.95     9.78 1590.90
## T2                    0.00 499187.59 499187.59  2.95    10.00 1641.90
## T3                    0.00 418801.14 418801.14  2.89     9.16 1612.65
## avg_K10               0.00 443837.78 443837.78  2.98     9.96 1600.11
## avg_T10               0.00 454004.88 454004.88  2.91     9.42 1607.06
## logFC               -12.95     17.90     30.85  0.48    29.89    0.06
## P.Value               0.00      1.00      1.00  0.51    -1.01    0.01
## adj.P.Val             0.00      1.00      1.00 -0.34    -0.96    0.01
## id                    1.00   1344.00   1343.00  0.00    -1.20   10.59
## expression            0.00 244624.67 244624.67 35.27  1268.17  184.40
##                         IQR
## IEP                    1.67
## mw                 29049.84
## max score           9554.40
## reported peptides     11.25
## sequence coverage     32.38
## K1                 31907.76
## K2                 33907.41
## K3                 34021.84
## T1                 33135.64
## T2                 34700.10
## T3                 34588.28
## avg_K10            32668.78
## avg_T10            34235.38
## logFC                  0.30
## P.Value                0.53
## adj.P.Val              0.48
## id                   671.50
## expression             0.21
```

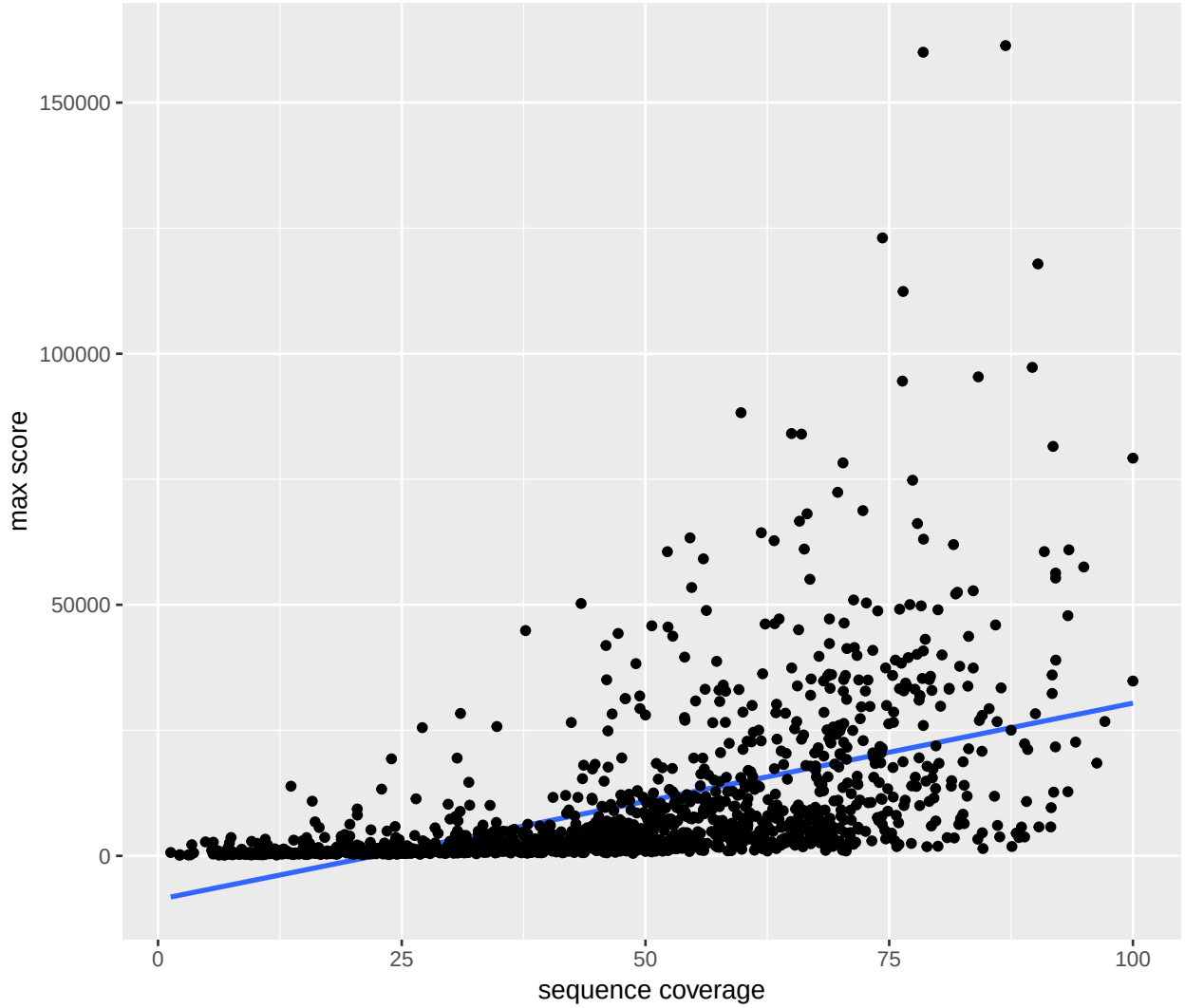
Ryškiausiai matosi tai, kad biologinėse mėginių replikose (K1, K2, K3 ir T1, T2, T3) nėra itin didelių skirtumų, todėl toliau bus naudojami jų vidurkiai (avg_K10 ir avg_T10).

Data mining

Koreliacijos nustatymas

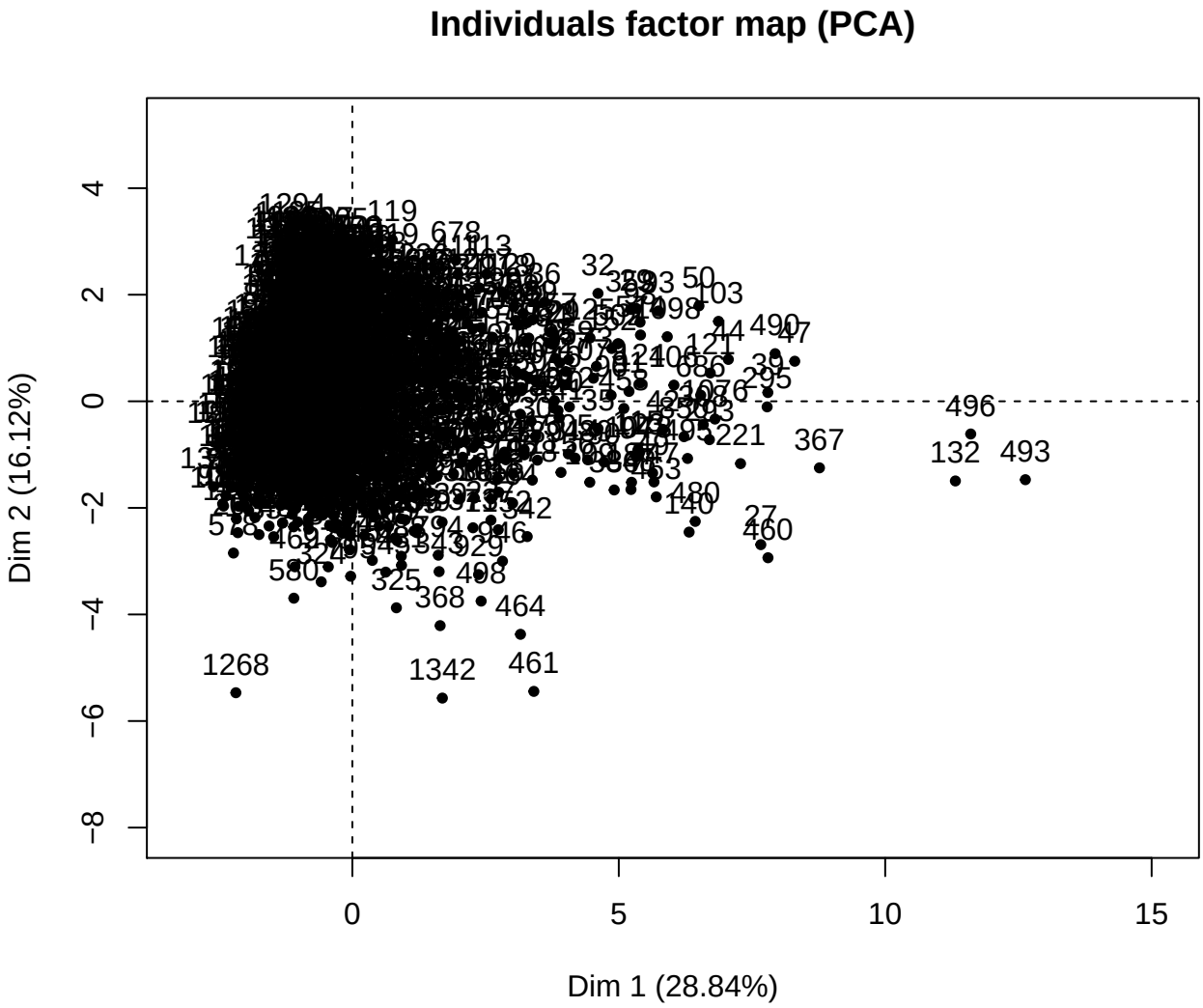
Sudaroma koreliacijos lentelė

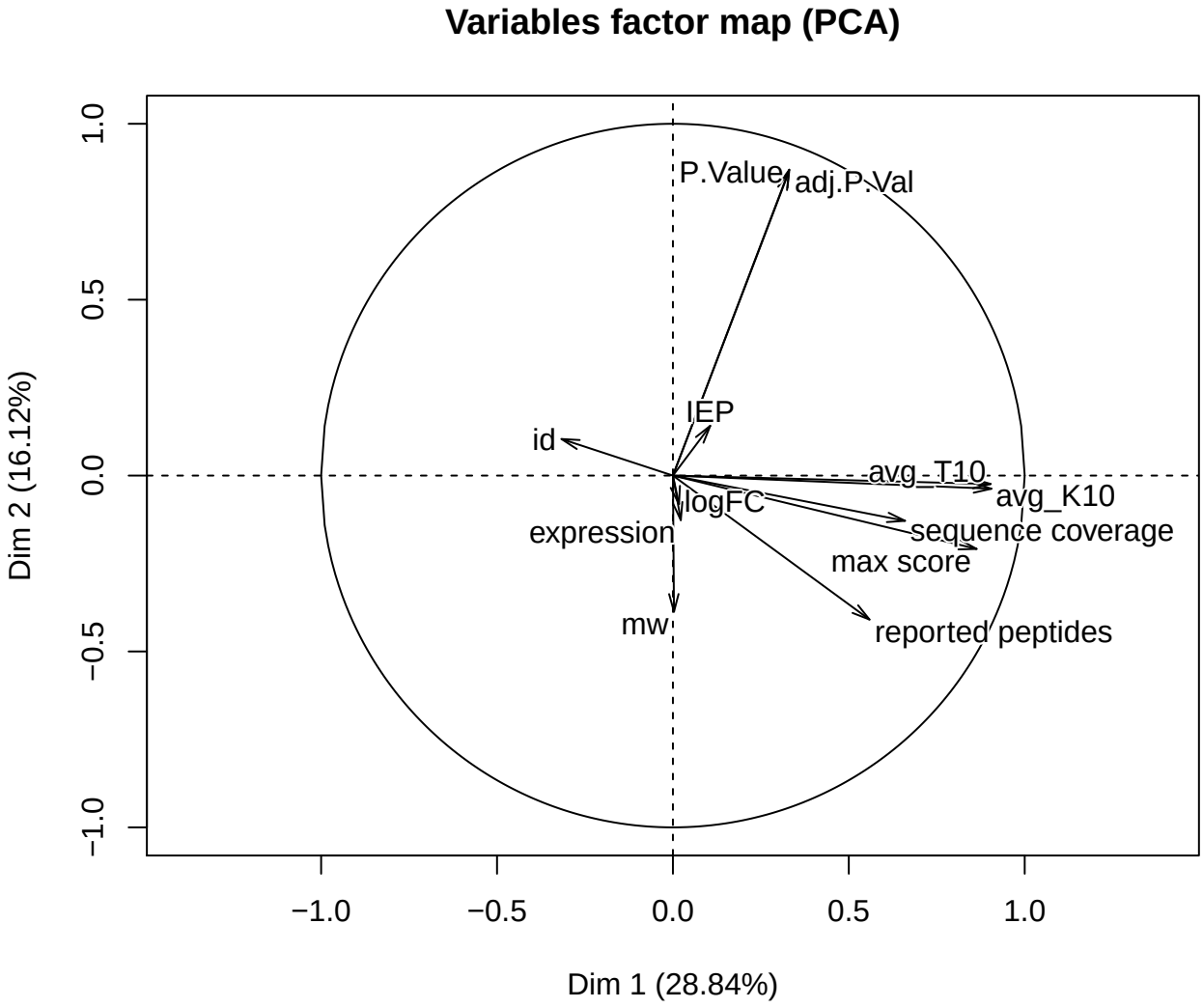

expression
id
adj.P.Val
P.Value
logFC
avg_T10
avg_K10
sequence coverage
reported peptides
max score
mw
IEP
mw
max score
reported peptides
sequence coverage
avg_K10
avg_T10
logFC
P.Value
adj.P.Val
id
Corr
1.0
0.5
0.0
−0.5
−1.0




```
##
##  Paired t-test
##
## data:  data_1$avg_K10 and data_1$avg_T10
## t = -1.5905, df = 1343, p-value = 0.112
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1005.6145   105.0835
## sample estimates:
## mean of the differences
##               -450.2655
```

Principinių komponenčių analizė




Variables factor map (PCA)
P.Value
adj.P.Val
IEP
id
avg_T10
avg_K10
logFC
expression
sequence coverage
max score
mw
reported peptides
Dim 2 (16.12%)
Dim 1 (28.84%)
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0