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 kintamųjų 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

Pavadinimai	Reikšmės
T3	3 tiriamasis mėginys
avg_K10	Kontrolinių mėginių vidurkis
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

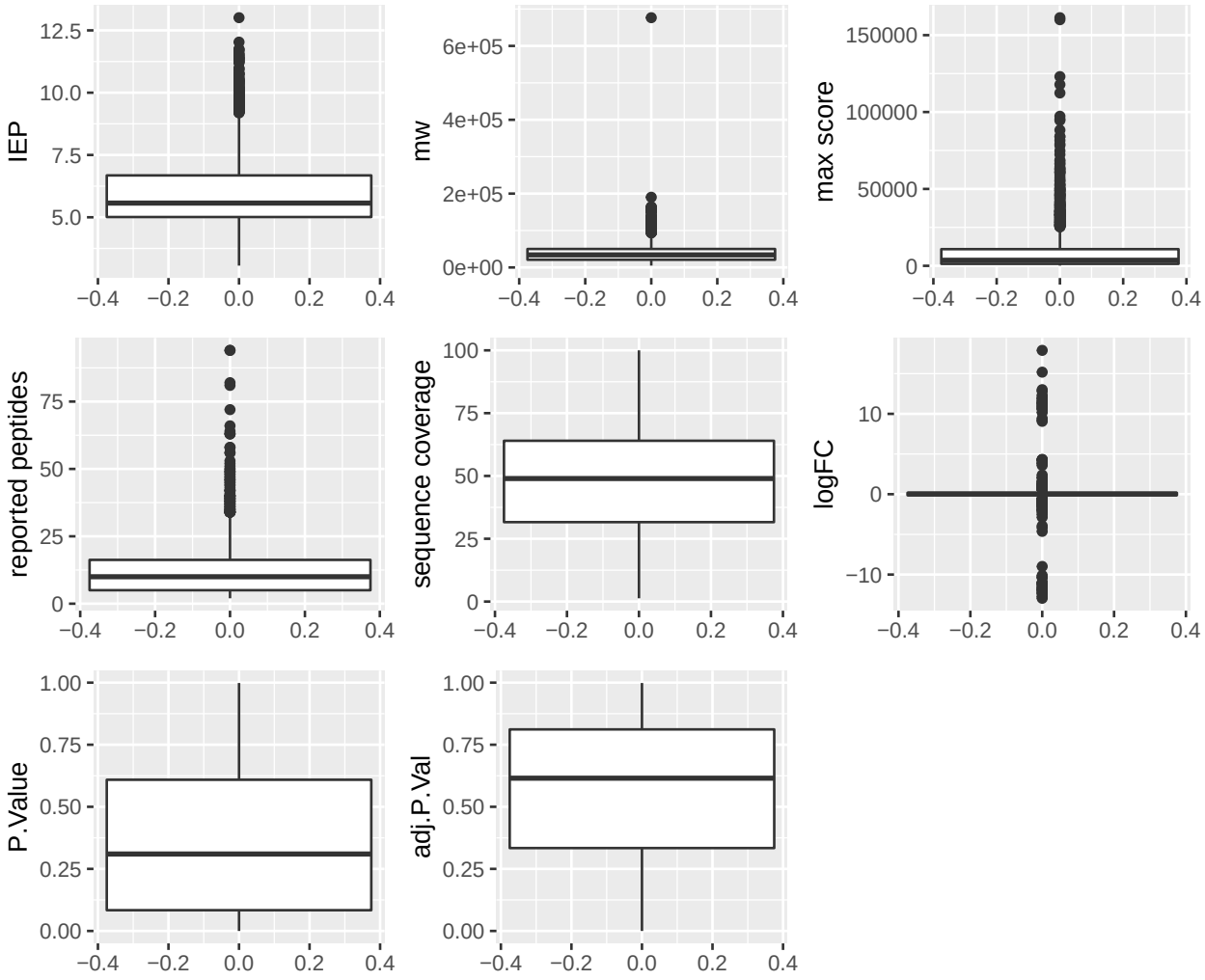
Pašalinus kategorinius kintamuosius, galima atlikti dar keletą veiksmų, kad su duomenimis būtų galima dirbti paprasčiau. Visų pirma, **FDR level** charakteristika nusako masių spektrometrijos metu nustatytų peptidų patikimumą. Kai FDR level = 0, peptidai nustatyti teisingai. Šiame darbe nagrinėjami visi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

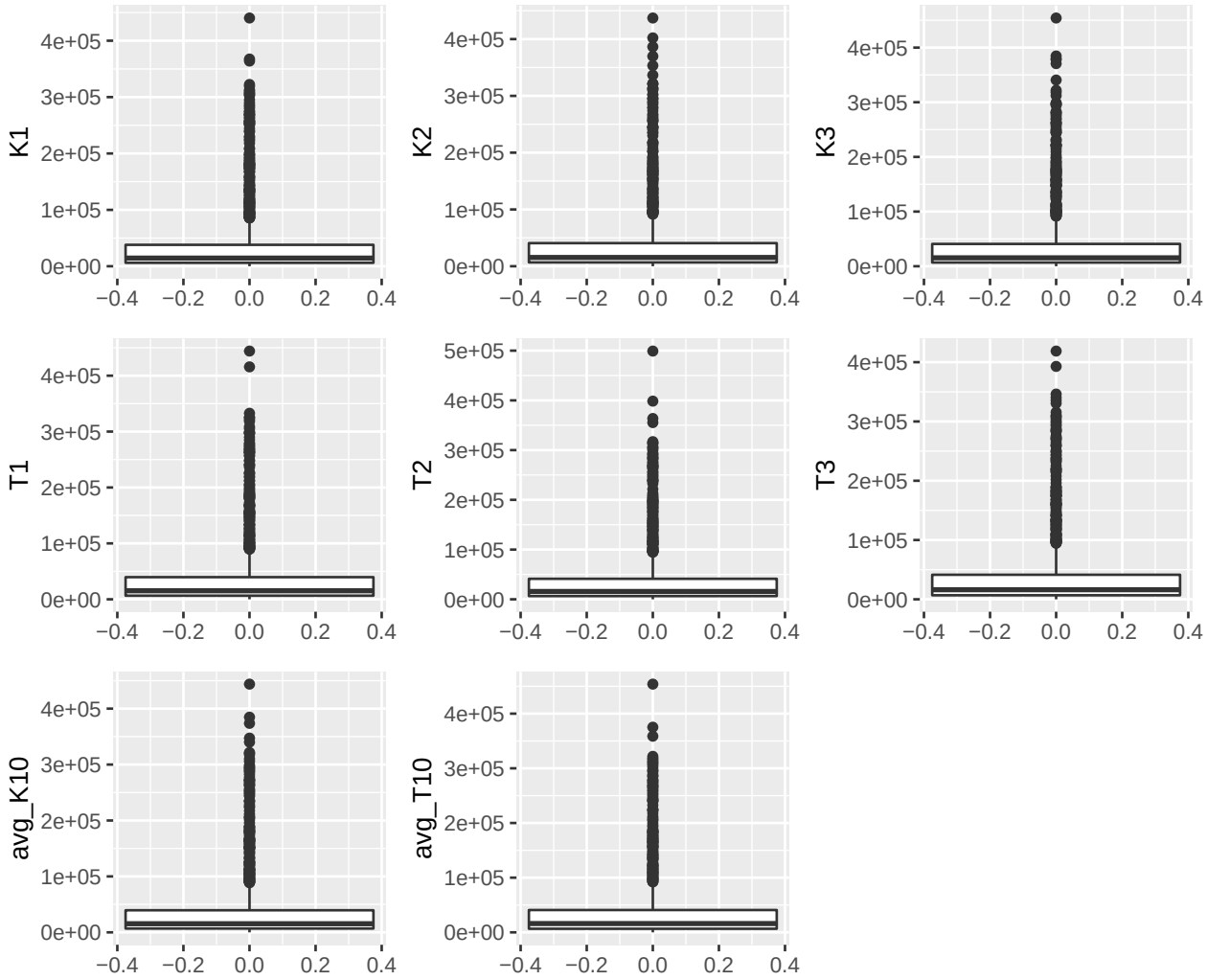
log FC charakteristika nusako nustatyrų peptidų raiškos pokytį tarp toksinu veiktų bakterijų ir kontrolinių bakterijų mėginių bei gali būti išreikšta formule 2^n, kur n = log FC vertė. Todėl pravartu šią vertę apsiskaičiuoti ir pridėti naują stulpelį į duomenų lentelę, pavadinimu “expression”.

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.

```
## data_1
##
##  18  Variables      1344  Observations
## ---------------------------------------------------------------------------
## IEP
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1344        0      476        1    6.269     1.89    4.521    4.680
##      .25      .50      .75      .90      .95
##    5.010    5.570    6.680    9.567   10.008
##
## lowest :  3.06  3.55  3.60  3.65  3.70, highest: 11.71 11.73 11.75 12.03 13.01
## ---------------------------------------------------------------------------
## mw
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1344        0     1344        1    40698    28655    11216    13961
##      .25      .50      .75      .90      .95
##    20938    34053    49988    77244    96916
##
## lowest :   4923.749   4938.595   5175.187   5677.590   6071.789
```

```
## highest: 162706.079 163962.616 164006.641 190225.138 676512.999
## ---------------------------------------------------------------------------
## max score
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1344        0     1342        1     9978    13203    437.2    612.3
##      .25      .50      .75      .90      .95
##   1278.8   3618.5  10833.2  29585.2  40716.3
##
## lowest :    123.1164    126.1948    153.1083    177.5567    185.3649
## highest: 112399.7000 117884.4000 123060.9000 160039.5000 161357.7000
## ---------------------------------------------------------------------------
## reported peptides
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1344        0       61    0.998    12.83    10.97     2.00     3.00
##      .25      .50      .75      .90      .95
##     5.00    10.00    16.25    27.00    35.00
##
## lowest :  2  3  4  5  6, highest: 66 72 81 82 94
## ---------------------------------------------------------------------------
## sequence coverage
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1344        0     1206        1    47.76    24.36    12.17    17.15
##      .25      .50      .75      .90      .95
##    31.62    48.97    64.00    74.86    81.15
##
## lowest :   1.30   2.21   2.24   3.09   3.32, highest:  94.12  94.96  96.30  97.12 100.00
## ---------------------------------------------------------------------------
## K1
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1344        0     1315        1    36832    47339     2117     3288
##      .25      .50      .75      .90      .95
##     6088    14447    37995    94228   172179
##
## lowest :      0.0000    514.0500    563.7069    771.7846    855.0794
## highest: 319235.0000 322413.6667 363908.6667 367498.3333 440279.3333
## ---------------------------------------------------------------------------
## K2
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1344        0     1323        1    39039    49796     2451     3694
##      .25      .50      .75      .90      .95
##     6777    15446    40685   100311   169998
##
## lowest :      0.0000    419.4030    503.5893    868.0000   1040.2580
## highest: 353331.3333 369784.0000 386452.3333 402142.6667 437125.6667
## ---------------------------------------------------------------------------
## K3
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1344        0     1316        1    38314    48642     2407     3670
##      .25      .50      .75      .90      .95
##     6795    15412    40817    99150   169974
##
## lowest :      0.0000    504.9175    671.6584    730.1390    929.4852
## highest: 370667.3333 378044.0000 379694.6667 384685.6667 454108.3333
## ---------------------------------------------------------------------------
```

```
## T1
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1344        0     1320        1    37705    48122     2208     3511
##      .25      .50      .75      .90      .95
##     6473    15254    39609    94718   168296
##
## lowest :      0.0000    461.2726    463.7317    616.4161    735.1477
## highest: 325049.6667 325694.5735 332798.6667 415763.3333 444025.9147
## ------------------------------------------------------------------------
## T2
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1344        0     1321        1    38956    49871     2303     3566
##      .25      .50      .75      .90      .95
##     6525    15742    41226    97160   177162
##
## lowest :      0.0000    512.9397    602.1514    685.9047    698.1047
## highest: 317028.3333 355357.4613 363581.6667 398519.6667 499187.5919
## ------------------------------------------------------------------------
## T3
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1344        0     1322        1    38879    49261     2301     3765
##      .25      .50      .75      .90      .95
##     6789    16178    41378    97893   174003
##
## lowest :      0.0000    475.7963    511.7064    573.4898    591.1595
## highest: 335638.1498 340479.0000 346246.3333 392910.3333 418801.1395
## ------------------------------------------------------------------------
## avg_K10
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1344        0     1324        1    38070    48371     2467     3718
##      .25      .50      .75      .90      .95
##     6817    15394    39486    96390   165850
##
## lowest :      0.0000    507.5189    620.9487   1012.6096   1115.7963
## highest: 340083.8889 347345.2222 373676.1111 384775.5556 443837.7778
## ------------------------------------------------------------------------
## avg_T10
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1344        0     1326        1    38520    48900     2472     3747
##      .25      .50      .75      .90      .95
##     6590    16051    40826    96684   170925
##
## lowest :      0.0000    545.0579    586.7127    614.0581    677.6285
## highest: 318306.2363 321911.2612 358739.0000 375197.1111 454004.8820
## ------------------------------------------------------------------------
## logFC
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1344        0     1344        1   0.0375    1.059 -0.66829 -0.37435
##      .25      .50      .75      .90      .95
## -0.12711  0.02584  0.17426  0.38341  0.65602
##
## lowest : -12.95353 -12.72488 -12.39407 -12.25049 -12.07386
## highest:  12.30189  12.81200  13.00953  15.18731  17.90021
## ------------------------------------------------------------------------
```

```
## P.Value
##         n   missing  distinct      Info      Mean       Gmd       .05
##      1344         0      1344         1    0.3651    0.3453 0.0007278
##       .10       .25       .50       .75       .90       .95
## 0.0102103 0.0836285 0.3100024 0.6089421 0.8408583 0.9315692
##
## lowest : 2.512858e-12 2.668965e-12 2.699179e-12 4.001109e-12 4.390731e-12
## highest: 9.963631e-01 9.965712e-01 9.974734e-01 9.975385e-01 9.984760e-01
## ------------------------------------------------------------------------
## adj.P.Val
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1344        0      626        1   0.5616   0.3353  0.01427  0.10142
##      .25      .50      .75      .90      .95
##  0.33377  0.61519  0.81172  0.93390  0.97990
##
## lowest : 1.180228e-09 1.727541e-09 2.079298e-09 5.257910e-09 5.837051e-09
## highest: 9.976058e-01 9.979769e-01 9.980984e-01 9.982813e-01 9.984760e-01
## ------------------------------------------------------------------------
## id
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1344        0     1344        1    672.5    448.3    68.15   135.30
##      .25      .50      .75      .90      .95
##   336.75   672.50  1008.25  1209.70  1276.85
##
## lowest :    1    2    3    4    5, highest: 1340 1341 1342 1343 1344
## ------------------------------------------------------------------------
## expression
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1344        0     1344        1    249.3    495.9   0.6293   0.7715
##      .25      .50      .75      .90      .95
##   0.9157   1.0181   1.1284   1.3044   1.5757
##
## Value           0   2000   4000   6000   8000  38000 244000
## Frequency    1327      9      3      1      2      1      1
## Proportion  0.987  0.007  0.002  0.001  0.001  0.001  0.001
## ------------------------------------------------------------------------
```

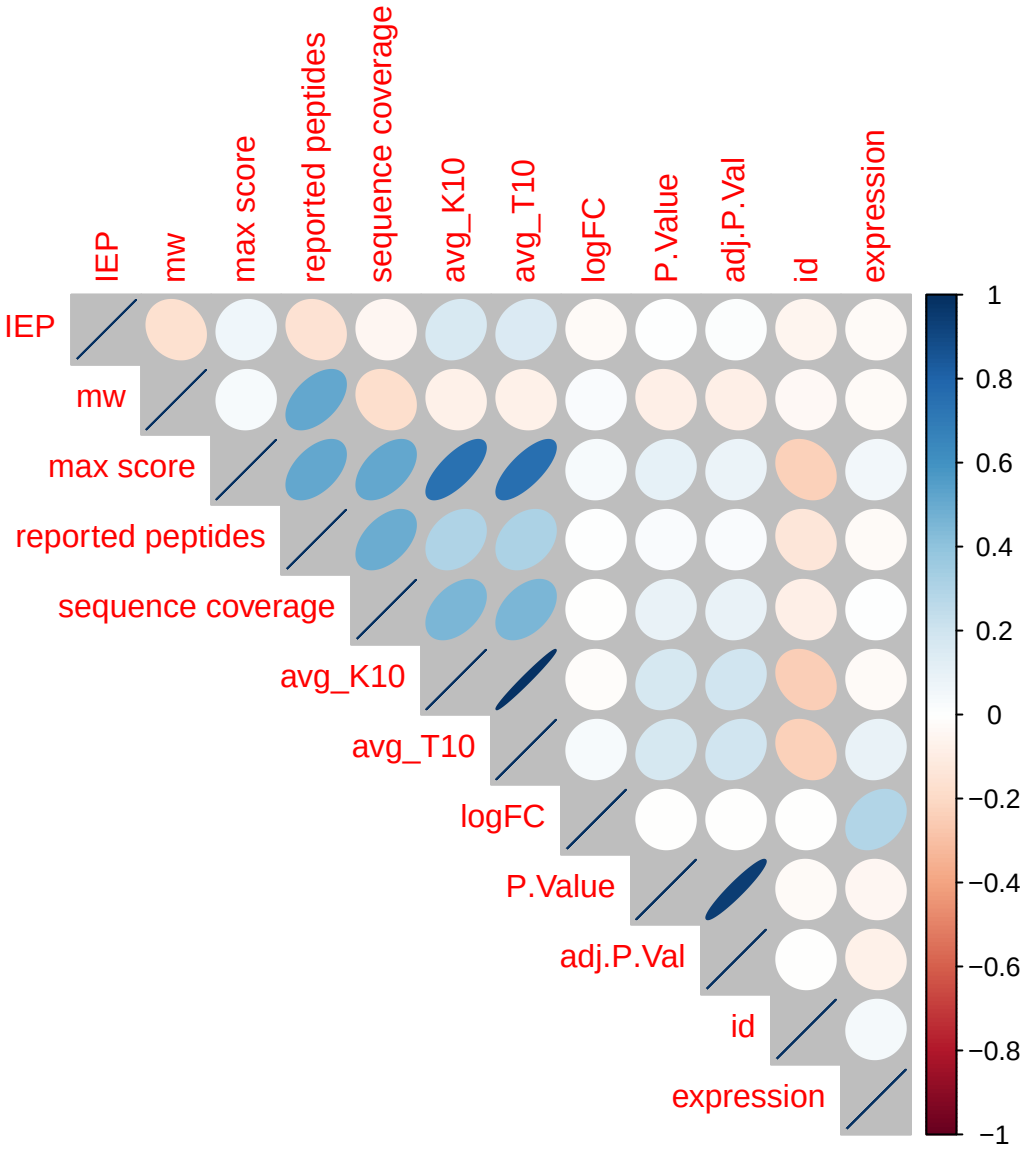
Ryškiausiai matosi tai, kad biologinėse mėginių replikose (K1, K2, K3 ir T1, T2, T3) nėra itin didelių skirtumų, todėl toliau bus naudojami jų vidurkiai (avg_K10 ir avg_T10).

Duomenų gavyba (data mining)

Pagrindinis šio etapo tikslas - nustatyti, ar tarp kintamųjų yra sąsajų, kurias būtų galima aptikti statistiniais metodais. Šiam tikslui pasiekti bus pasitelkiama koreliacijos analizė.

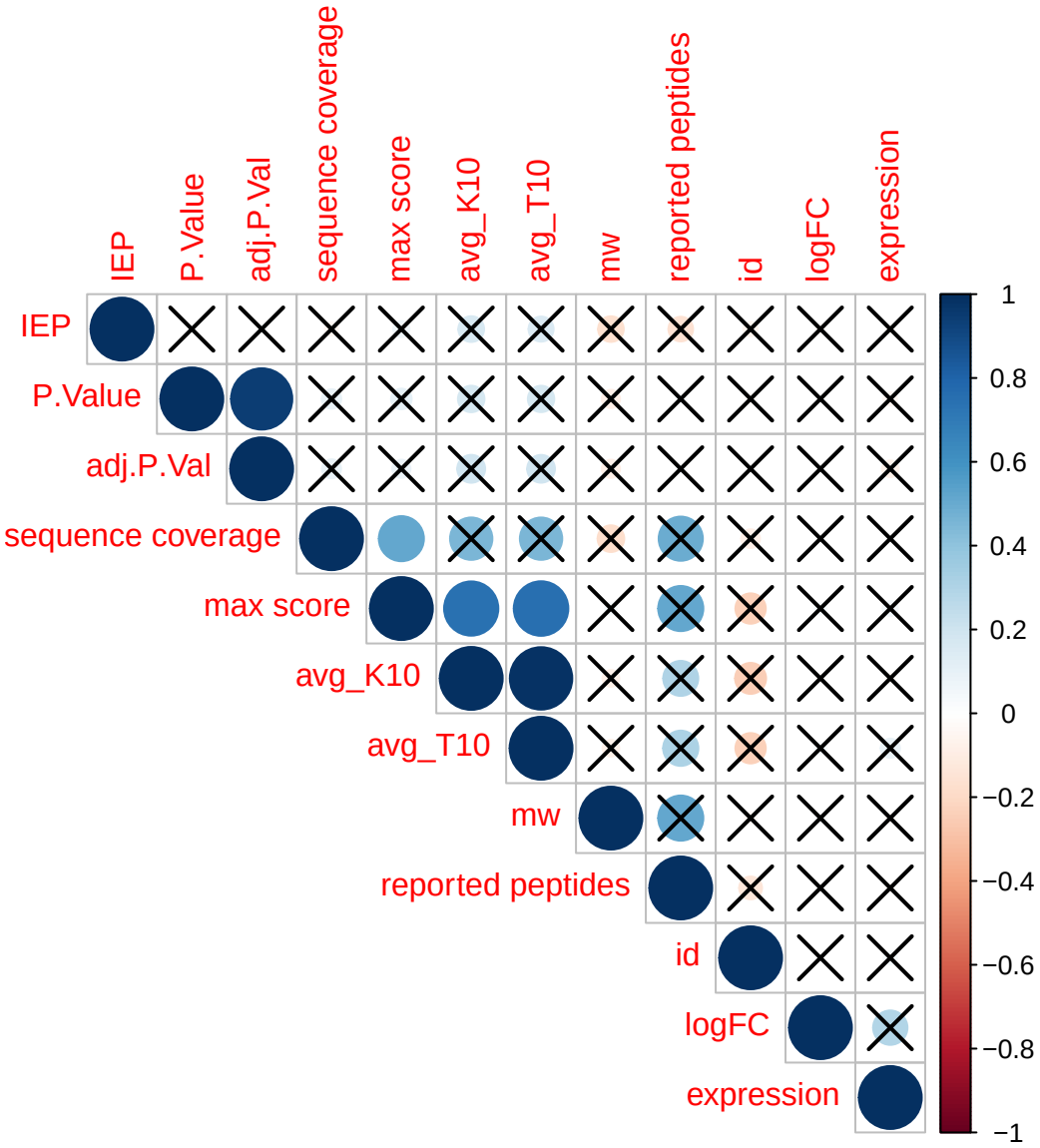
Koreliacijos nustatymas

Sudaroma koreliacijos lentelė



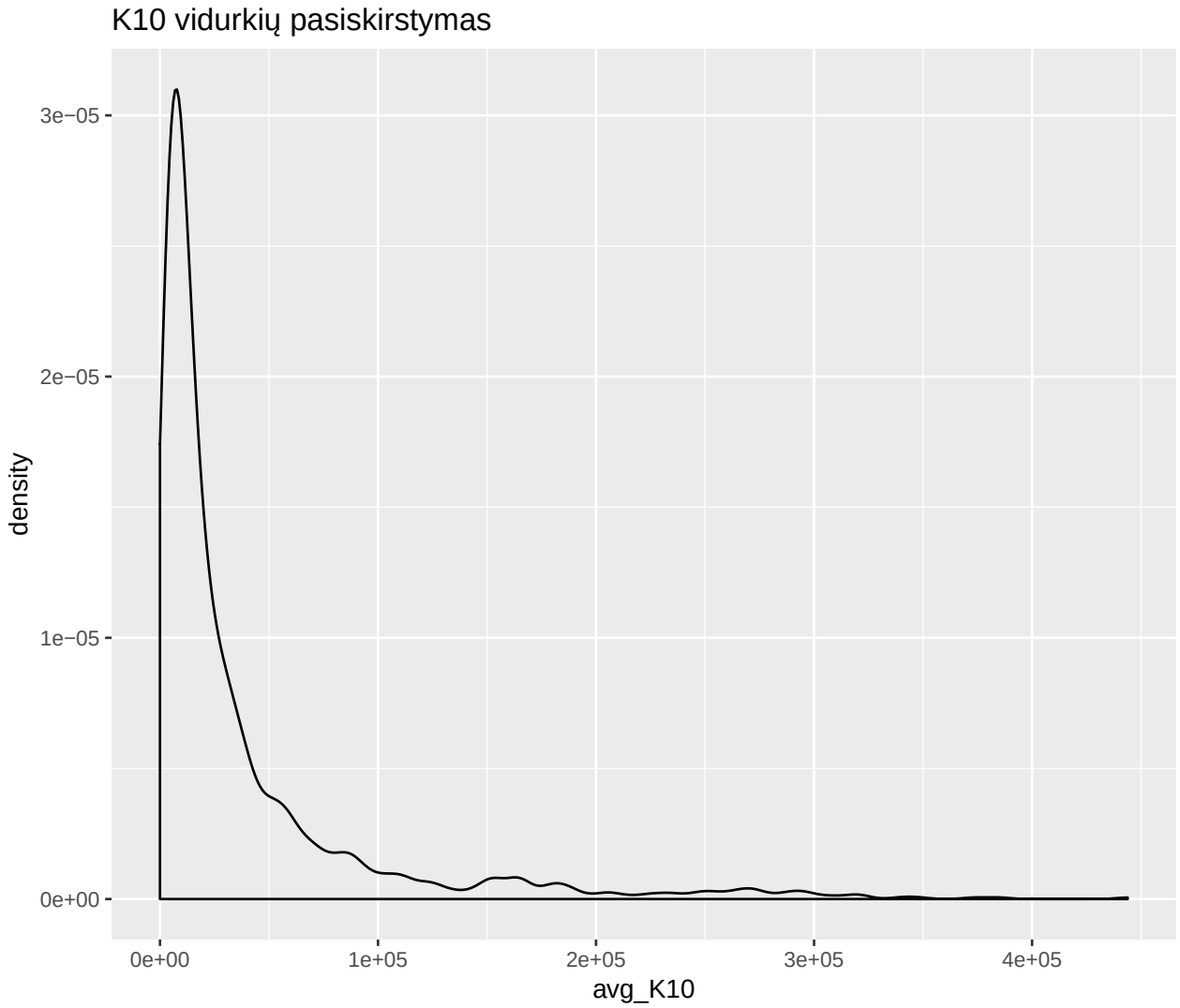
Elipsių spalvos intensyvumas bei forma nurodo koreliacijos stiprumą.

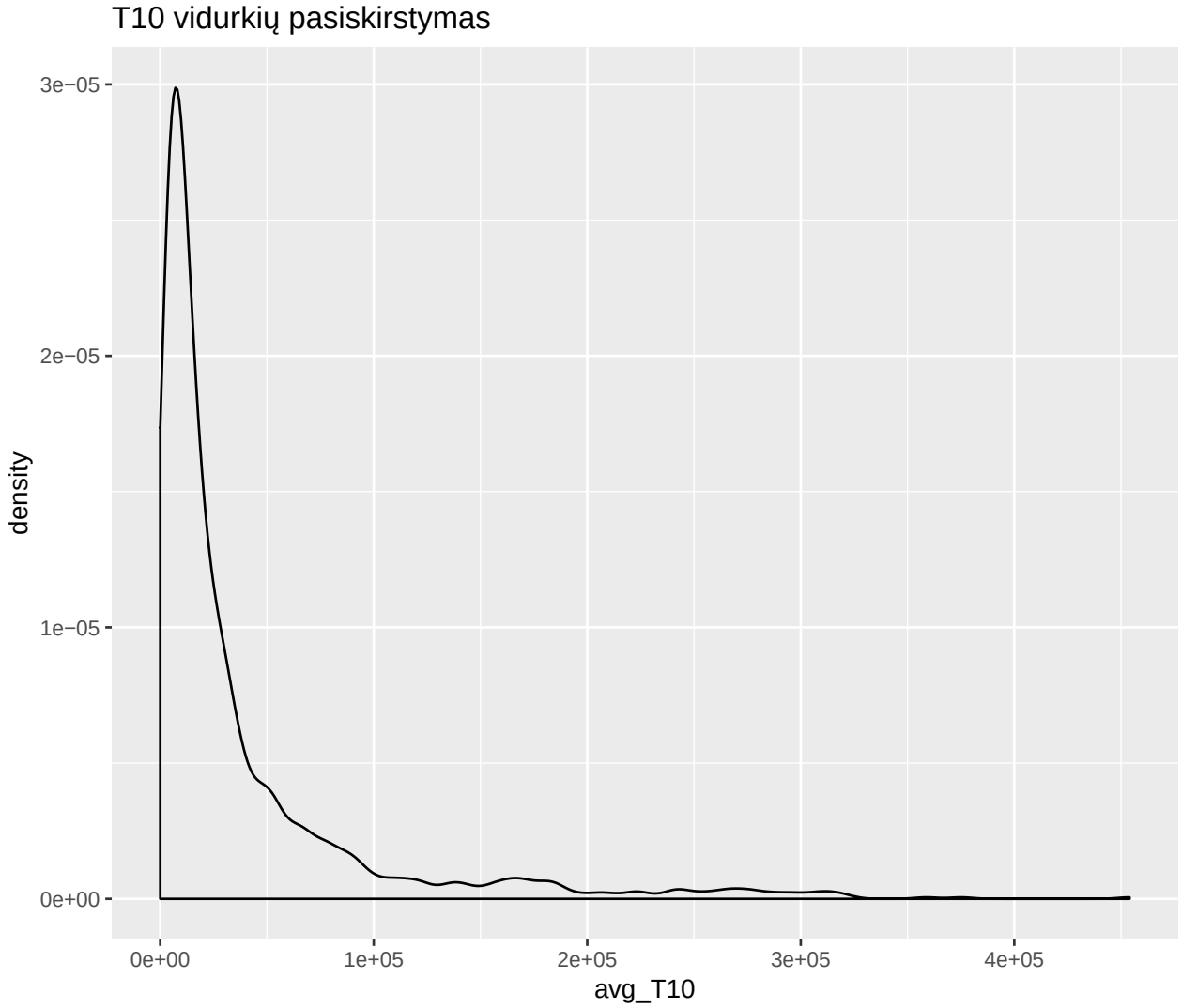
Patikrinamos koreliacijos p reikšmės.



Užbraukti langeliai rodo, jog koreliacija nėra statistiškai patikima (p reikšmė > 0.01).

Tikrinamas sąryšis tarp K10 ir T10 vidurkių. Dar kartą nustatoma, ar duomenų pasiskirstymas yra normalusis.


K10 vidurkių pasiskirstymas
density
3e−05
2e−05
1e−05
0e+00
0e+00
1e+05
2e+05
3e+05
4e+05
avg_K10



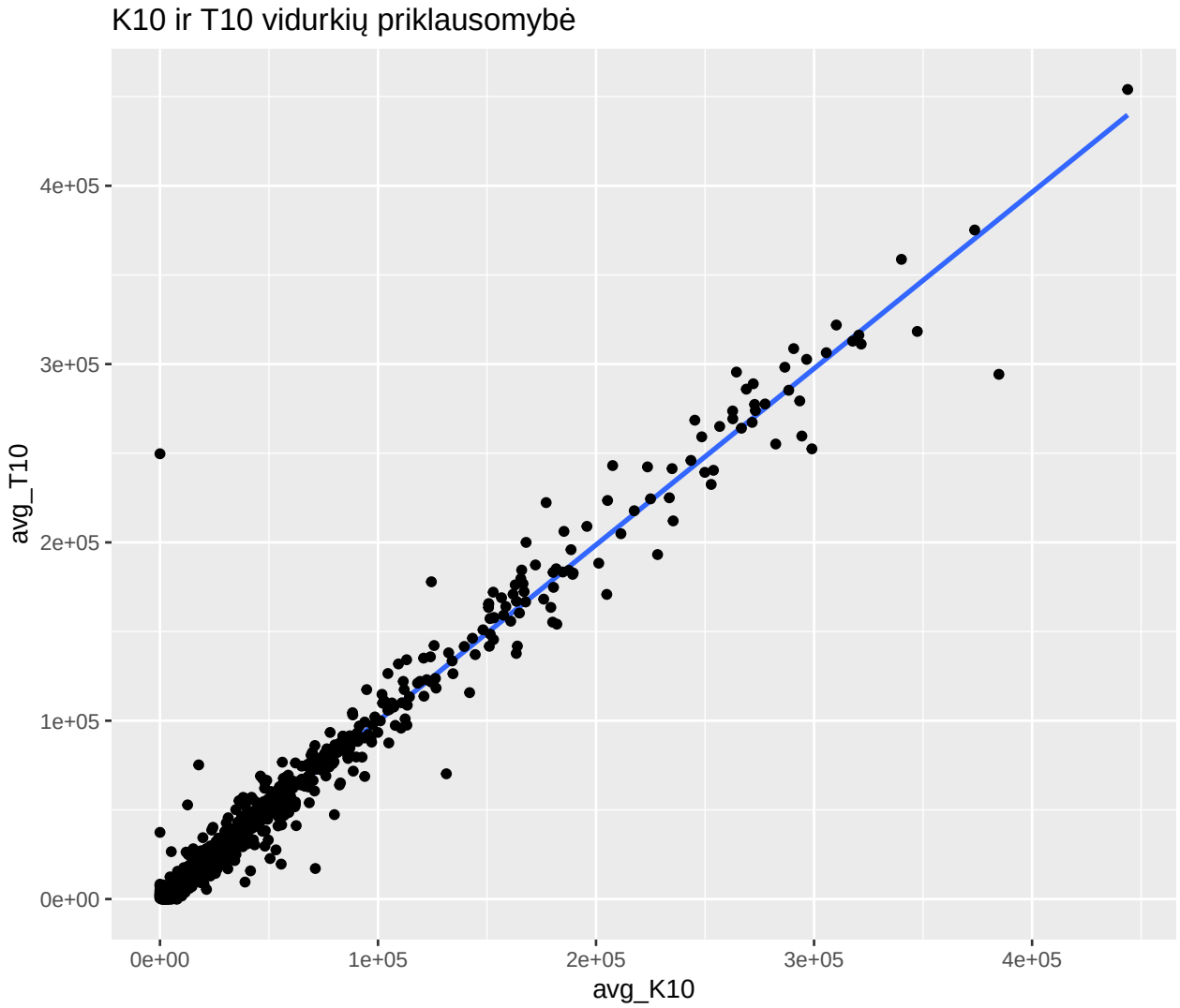
Kadangi skirstiniai n4ra normalieji, T testas negali būti taikomas šiems duomenims.

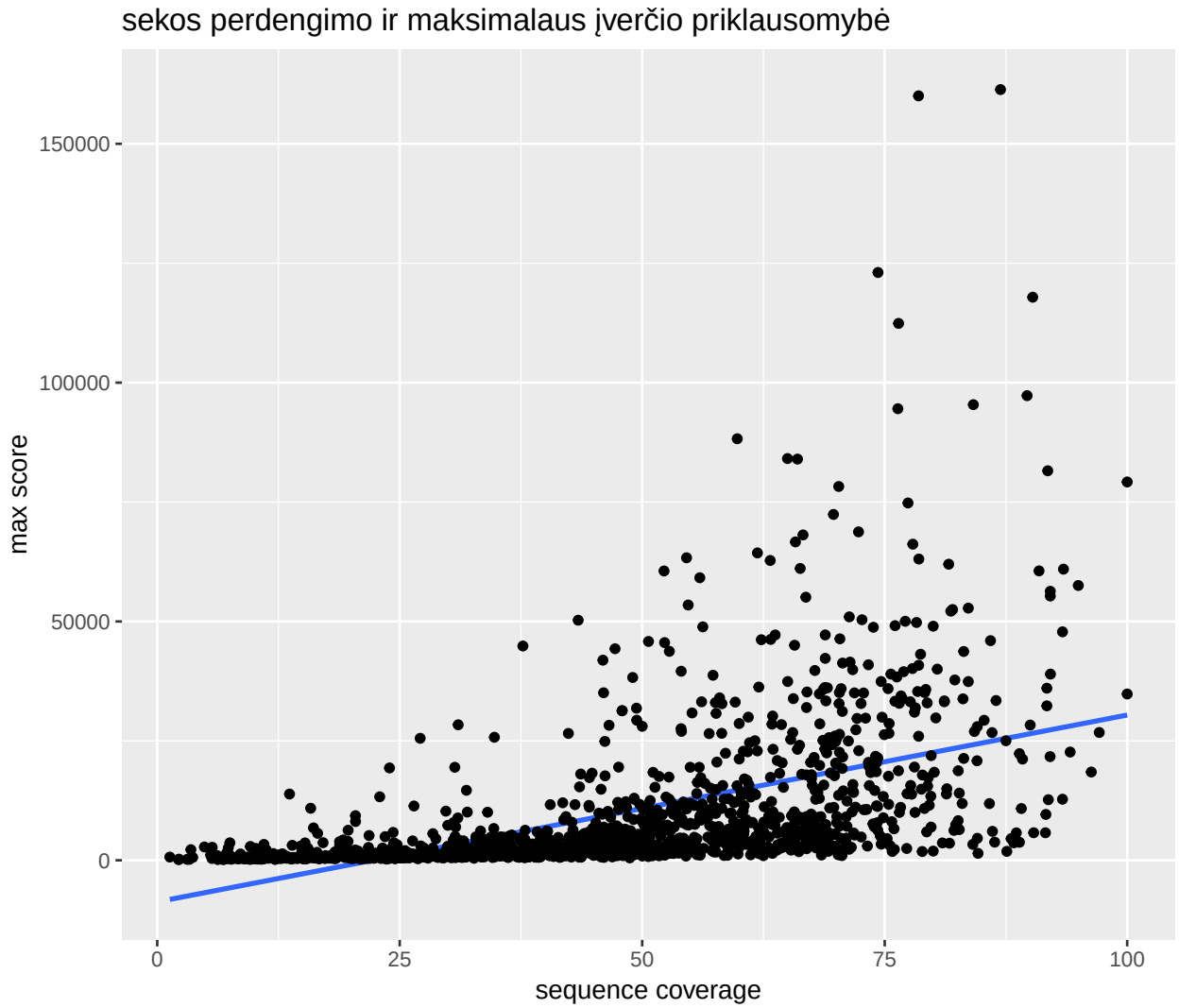
Nenormiai pasiskirsčiusiems duomenims galima taikyti Mano-Vitnio-Vilkoksono rangų sumų testą.

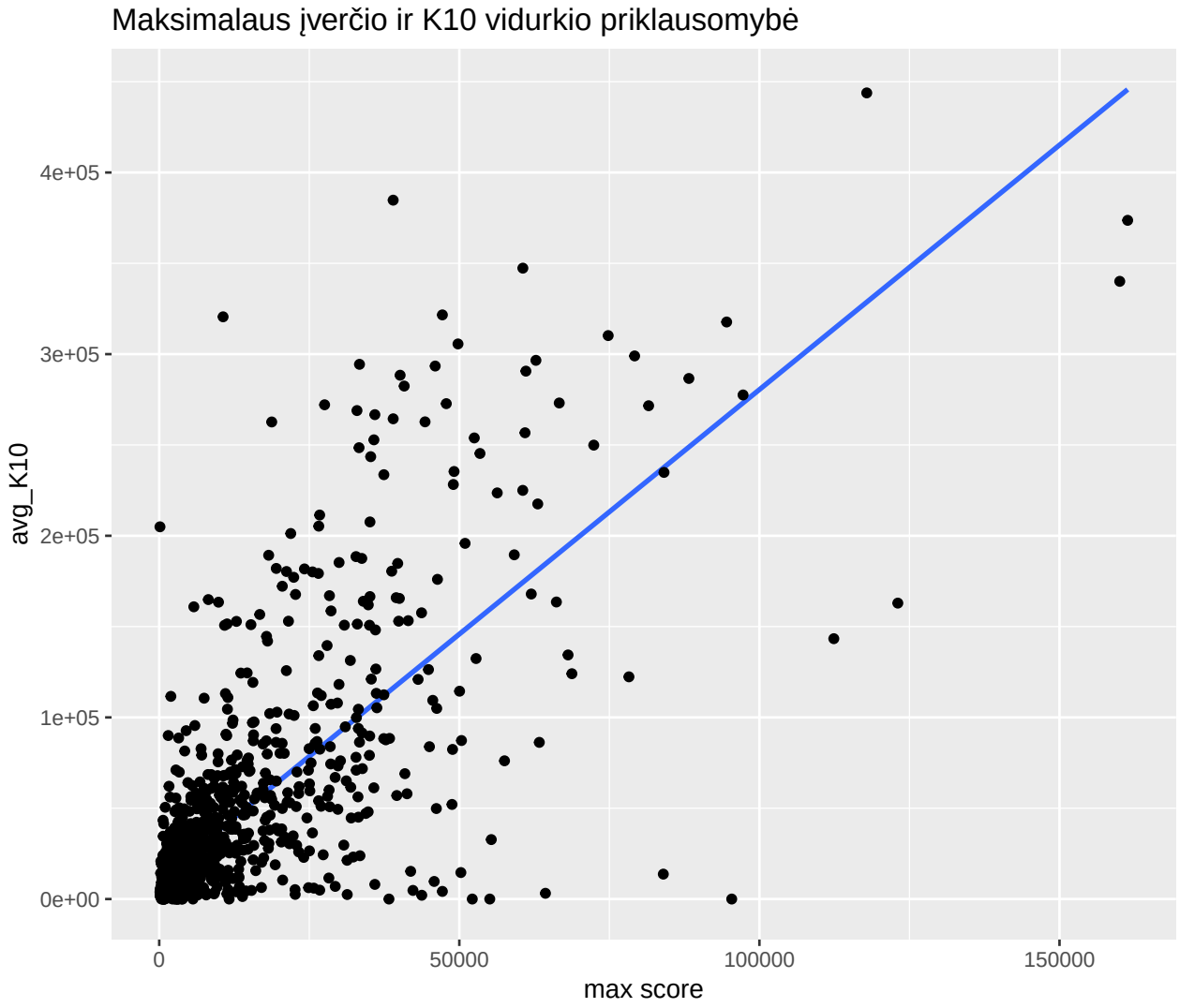
```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  data_1$avg_K10 and data_1$avg_T10
## V = 396860, p-value = 0.0001094
## alternative hypothesis: true location shift is not equal to 0
```

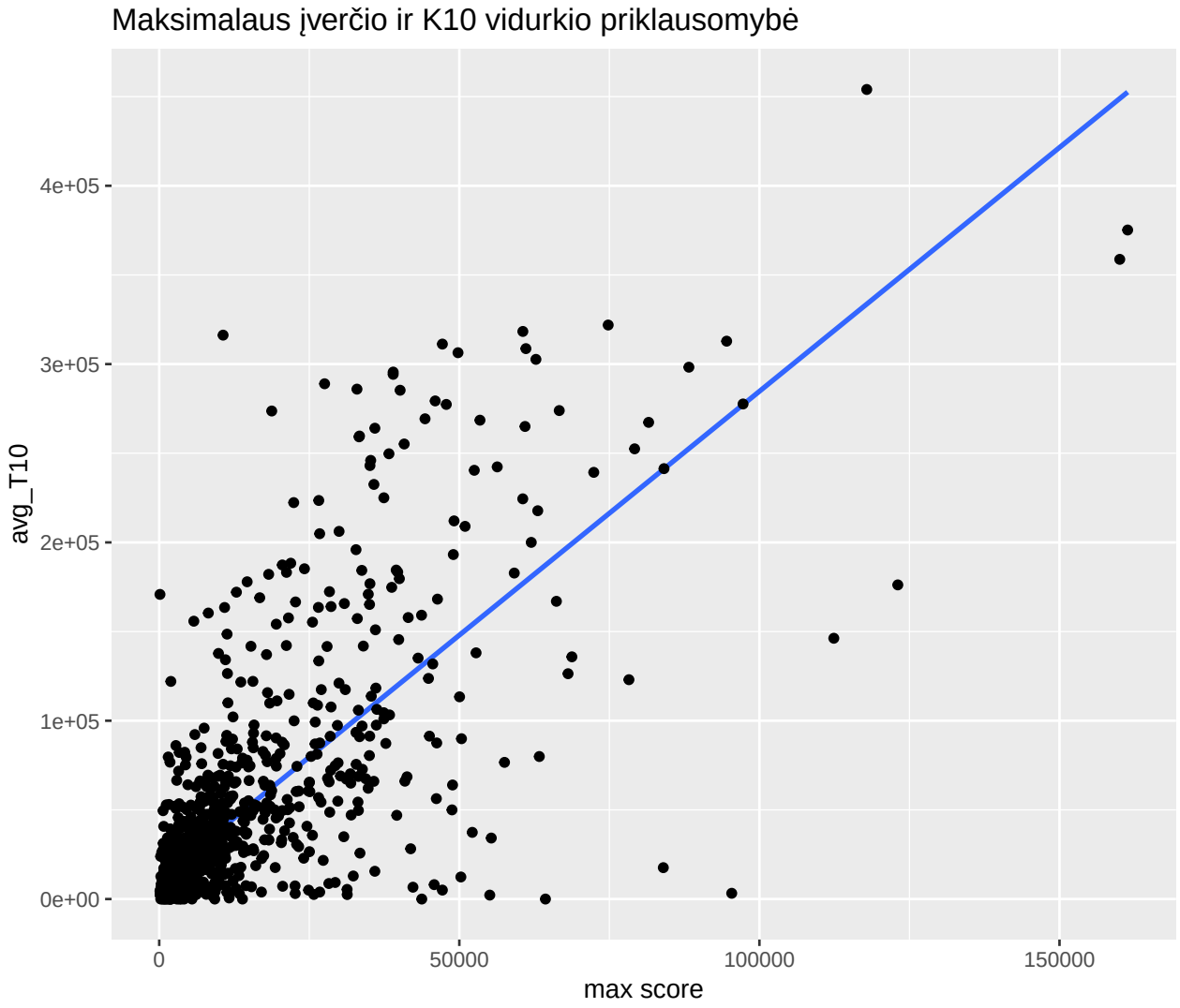
Iš gautų statistikų matome, kad populiacijos yra neidentiškos, o duomenys yra statistiškai patikimi ($p < 0.001$).

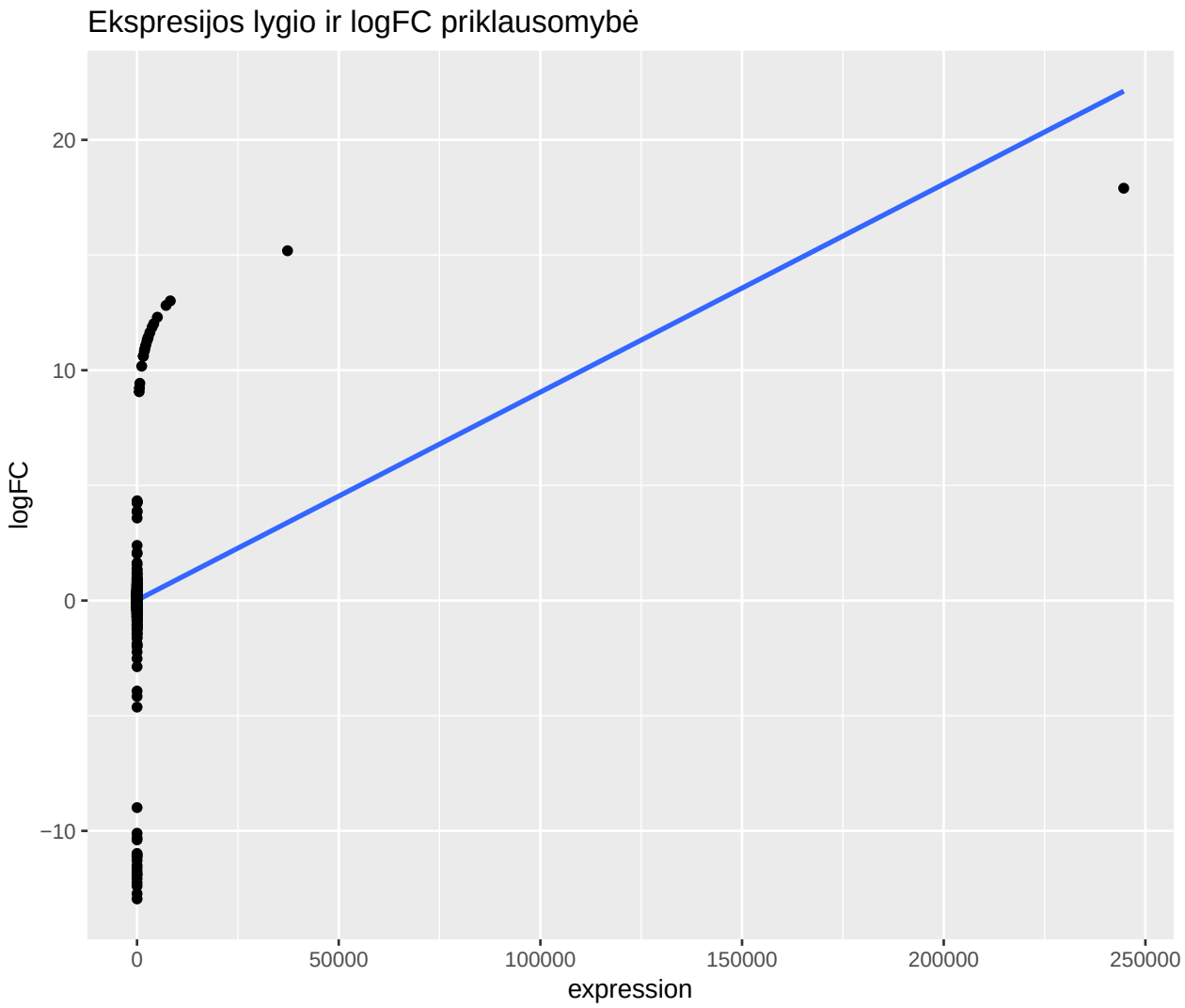
Toliau grafiškai atvaizduojama kintamųjų poros, bandoma nustatyti, ar yra tiesinė koreliacija tarp jų.


K10 ir T10 vidurkių priklausomybė
avg_T10
0e+00
1e+05
2e+05
3e+05
4e+05
avg_K10
0e+00
1e+05
2e+05
3e+05
4e+05


sekos perdengimo ir maksimalaus įverčio priklausomybė
150000
100000
50000
0
max score
0
25
50
75
100
sequence coverage

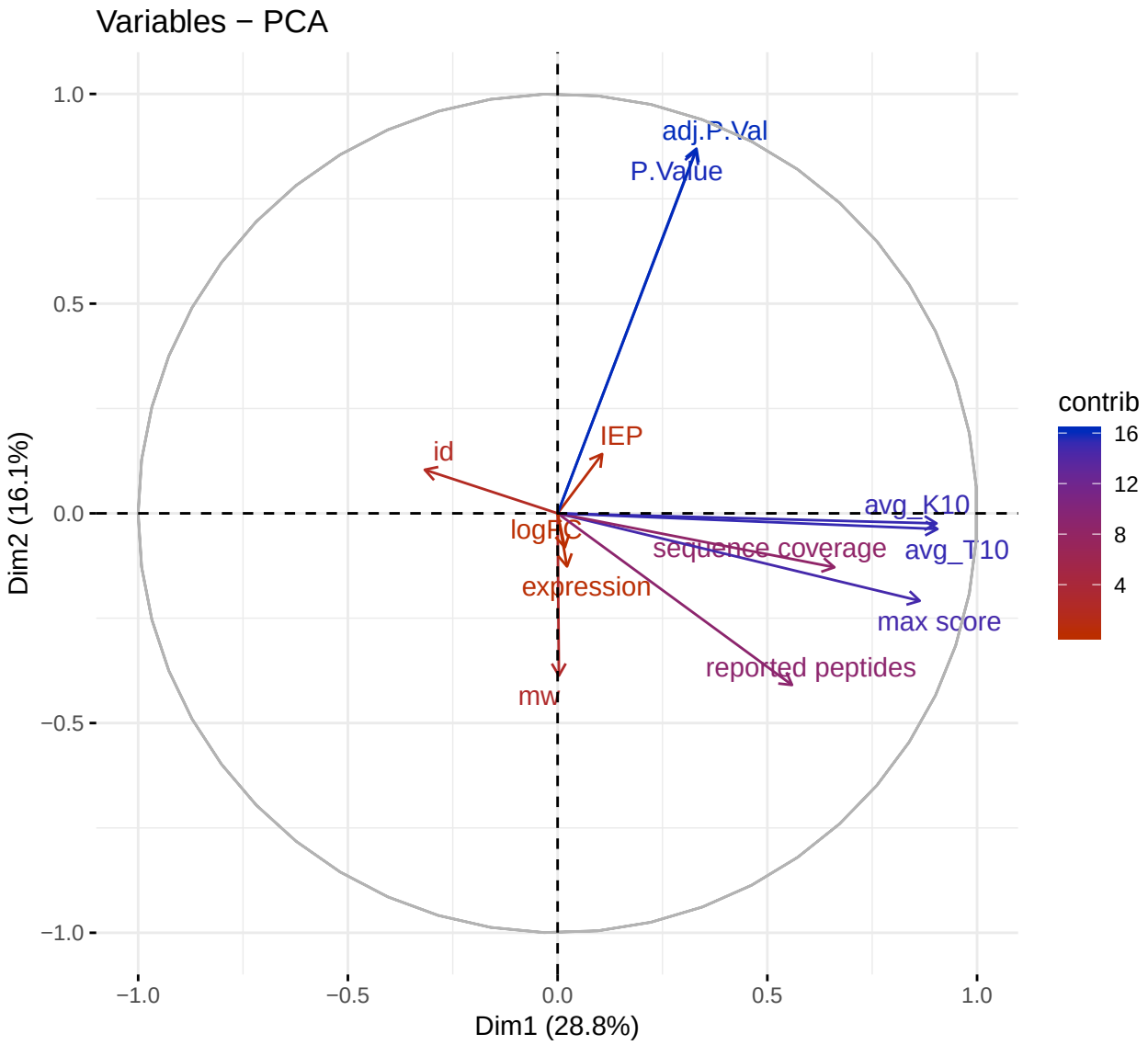

Maksimalaus įverčio ir K10 vidurkio priklausomybė
avg_K10
4e+05
3e+05
2e+05
1e+05
0e+00
0
50000
100000
150000
max score


Maksimalaus įverčio ir K10 vidurkio priklausomybė
avg_T10
4e+05
3e+05
2e+05
1e+05
0e+00
0
50000
100000
150000
max score



Principinių komponenčių analizė

Kadangi tarp kai kurių kintamųjų galima koreliacija, sąsajoms įžvelgti galima naudoti ir principinių komponenčių analizę. Ši analizė itin pravarti, kai duomenų lentelėje yra daug kintamųjų (dimensijų), o taip ir yra šiuo atveju.



Aptarimas

Darbo metu buvo bandyta įžvelgti sąsajas tarp kintamųjų. Nustatytos silpnos koreliacijos tarp maksimalaus įverčio bei T10 bei K10 mėginių vidurkių bei sekos perdengimo procento. Visdėl to šios koreliacijos biologine prasme yra abejotinos. Taip pat nustatyta stipri koreliacija tarp K10 ir T10 vidurkių, kas yra gana akivaizdu ir lengvai paaiškinama - didžiosios daugumos nustatytų peptidų raiška ženkliai nepakito. Biologine prasme būtų pravartu patyrinėti peptidus, kurių raiška T10 ir K10 mėginiuose skrėsi - buvo stebima neigiama koreliacija. Taip pat įdomu tai, kad metodais, naudotais šiame darbe nebuvo rasta ryšio tarp ekspresijos pokyčio ir kintamojo logFC, nors pirmasis ir buvo išvestas iš logFC. Praeitame skyriuje pateiktame grafike įvertinus vizualiai reikšmių pasiskirstymą galima spėti, kad ryšys tarp kintamųjų visgi egzistuoja, tačiau netiesinis. Apibendrinant galima pasakyti, kad naudoti darbe metodai gali padėti aptikti ryšių tarp kintamųjų, tačiau jei rezultatai gaunami neigiami, tai dar nereiškia, kad ryšio nėra.