wangjun's project result

wangjun

2023-09-15

Data Preprocessing

Import the data set and confirm whether the import is successful:

```
# import data set
data <- read.csv("BRCA_RNASeqv2_top50.csv")

# check the first 6 rows
head(data)
```

```
##        FIGF    LYVE1   CD300LG     SCARA5     PAMR1     SDPR     MYOM1
## 1  544.1270 1360.952  602.8571   882.2222 2508.8889 1425.079  195.2381
## 2  846.2623 3508.792 1736.9064  2060.7428 1572.1674 4267.795  460.7428
## 3  741.3442 3423.116 4074.3381  6321.7923  675.1527 7398.676 2019.8574
## 4  905.1546 9641.237 2786.5979 10177.3196 1591.7526 3809.794  966.4948
## 5 1093.8680 4783.355 2650.7514  3507.0129 1523.2185 5314.676 1707.1561
## 6  913.3363 3324.652 2725.6399  1906.6008 2103.7270 6595.420  724.2928
##       BTNL9    KCNIP2    SLC2A4     PDE2A        LEP    ACVR1C   ABCA10
## 1  717.7778  283.8095   68.8889  658.4127   987.3016   57.7778 164.1270
## 2 2261.9652 2579.7837  400.1881 1729.0080 16210.6253  761.2600 244.0997
## 3 4518.3299 8273.9308 1040.2240 2649.1853 37746.4358 3481.6701 244.3992
## 4 2610.3093 5992.7835  958.7629 2597.9381  7954.1237 2986.0825 321.1340
## 5 2827.3706 5088.7792 1287.5636 2153.5831 28808.9305 2052.1002 271.2715
## 6 2493.0400 6469.2411  792.0970 1814.5487 13086.6637 1602.6044 528.9627
##        AQP7    GPR146    ATP1A2    FXYD1  ARHGAP20      NPR1     ATOH8
## 1  213.9683  262.5397  123.1746 302.5397  110.4762  807.9365  628.2540
## 2 2694.4993  900.7992 1587.2120 782.3225  278.7024 3299.2948 1084.7203
## 3 9354.8880 1536.1507 5110.9980 742.8717  759.1650 5919.5519 1054.4807
## 4 3496.9072 1191.7526 4108.2474 588.1443  816.4948 4698.4536 1208.7629
## 5 4275.9406 1051.9087 1224.1368 667.9328 1170.4680 3651.9187  788.4436
## 6 4052.9861  889.5375  573.8662 724.2928 1705.8824 3189.4926 1246.5200
##       ABCA9   ALDH1L1  ADAMTS5      RDH5      GPAM       CA4   KLHL29   GPIHBP1
## 1  396.1905  175.8730 1400.952  217.2603  1293.333   70.1587 505.3968  383.8095
## 2 1042.5952 1243.0654 1906.159 1804.9760  8317.818  754.1138 509.2619 1598.4955
## 3 2514.7658 1870.6721 2656.823 3389.8371 42443.992 1114.5621 588.0855 1929.2261
## 4 2755.1546 3053.0928 8410.825 2129.8969 22096.392  428.8660 592.7835  775.7732
## 5  853.3341 1800.3447 3519.698 2639.5298 51813.823  656.2232 644.0257 1169.0043
## 6 1022.9008  885.9452 3348.002 1780.8621 23591.828  722.9457 713.9650  918.2757
##   LOC728264    MAMDC2  TMEM132C     ITIH5      HSPB7     HSPB6      DMD
## 1 4158.4127 1114.4444  194.9206  2855.238   288.8889  2822.222 1161.270
## 2 1922.7080  717.1641 1130.2304 13662.811  5728.6319  9107.287 1048.237
```

```
## 3  418.5336  938.2943 3162.9328 30184.827 10879.8371 20623.727 3536.660
## 4  455.6701 2045.2062 2252.0619  9205.155  5174.7423 14653.608 2420.619
## 5  529.3696  912.8333 1305.1279 15925.976  5655.2291 14374.947 2742.964
## 6 2588.6843  974.3556 1524.4724 11991.917  3330.0404 10604.400 2626.403
##      SPRY2   IGFBP6    CXCL2      EBF1       KLB   CLEC3B  TMEM220   IBSP
## 1 1142.540 1773.333 617.7778  538.7302   80.6349 2584.127 211.7460 2.2222
## 2 1176.117 6479.737 248.9892 1916.3141 1270.8980 6389.093 352.4213 0.0000
## 3 1567.719 3161.405 386.4562 4590.6314 3139.0020 6565.682 312.6273 0.0000
## 4 1173.196 5676.289  63.4021 3756.7010 1383.5052 7955.670 202.5773 0.0000
## 5 1301.713 1690.568 342.0167 3006.9172 4086.1482 6804.717 296.6422 0.0000
## 6 2016.614 1788.954 814.0997 2040.8621 2154.0189 3478.222 298.1590 0.0000
##      HIF3A    IGSF10      CIDEC   C2orf40      LEPR   ANGPTL1 class
## 1  70.4762  644.7619   716.1905 1760.3175  770.9333  230.7937     N
## 2 374.2360  435.5430  7336.7710  225.2938 1397.3785  459.2384     N
## 3 653.2587 1254.5825 15562.9633  207.7393 1645.2037 1347.7597     N
## 4 354.6392 1019.5876 12504.7113  224.7423 2484.2216  596.3918     N
## 5 283.9568 1367.5789 13990.4835  157.1033 1857.0726  677.6907     N
## 6 310.7319 1704.9843  7711.2708  414.9079 2087.7863  550.0674     N
```

It can be found that the data has been successfully imported.

check whether there is missing values or not:

```
# View the number of missing values in each column
missing_values_per_column <- sapply(data, function(x) sum(is.na(x)))

# Output the number of missing values in each column
missing_values_per_column
```

```
##      FIGF     LYVE1   CD300LG    SCARA5     PAMR1      SDPR     MYOM1     BTNL9
##         0         0         0         0         0         0         0         0
##    KCNIP2    SLC2A4     PDE2A       LEP    ACVR1C    ABCA10      AQP7    GPR146
##         0         0         0         0         0         0         0         0
##    ATP1A2     FXYD1  ARHGAP20      NPR1     ATOH8     ABCA9   ALDH1L1   ADAMTS5
##         0         0         0         0         0         0         0         0
##      RDH5      GPAM       CA4    KLHL29   GPIHBP1 LOC728264    MAMDC2  TMEM132C
##         0         0         0         0         0         0         0         0
##     ITIH5     HSPB7     HSPB6       DMD     SPRY2    IGFBP6     CXCL2      EBF1
##         0         0         0         0         0         0         0         0
##       KLB    CLEC3B   TMEM220      IBSP     HIF3A    IGSF10     CIDEC   C2orf40
##         0         0         0         0         0         0         0         0
##      LEPR   ANGPTL1     class
##         0         0         0
```

Judging from the results, there is no missing data in the data set, so there is no need to perform any processing on missing values.

```
summary(data)
```

```
##       FIGF              LYVE1             CD300LG             SCARA5
##  Min.   :   0.000   Min.   :    0.00   Min.   :   0.000   Min.   :    0.000
##  1st Qu.:   2.197   1st Qu.:   17.45   1st Qu.:   3.102   1st Qu.:    3.314
##  Median :   6.385   Median :   32.65   Median :  15.084   Median :   18.669
```

```
##  Mean   :  84.983   Mean   :  257.07   Mean   : 211.590   Mean   :  246.129
##  3rd Qu.:  30.594   3rd Qu.:   89.26   3rd Qu.:  85.894   3rd Qu.:   92.496
##  Max.   :2284.338   Max.   :11111.93   Max.   :6292.725   Max.   :11533.028
##      PAMR1              SDPR              MYOM1              BTNL9
##  Min.   :   0.467   Min.   :    2.806   Min.   :    0.88   Min.   :   0.833
##  1st Qu.:  35.774   1st Qu.:   46.249   1st Qu.:   15.79   1st Qu.:  33.151
##  Median :  76.714   Median :  106.540   Median :   30.60   Median :  78.672
##  Mean   : 243.117   Mean   :  484.364   Mean   :  163.42   Mean   : 297.891
##  3rd Qu.: 191.350   3rd Qu.:  348.140   3rd Qu.:   63.54   3rd Qu.: 207.829
##  Max.   :4442.947   Max.   :11292.082   Max.   :76348.50   Max.   :8577.470
##      KCNIP2             SLC2A4             PDE2A              LEP
##  Min.   :    0.00   Min.   :    0.000   Min.   :   8.834   Min.   :    0.00
##  1st Qu.:   13.08   1st Qu.:    7.591   1st Qu.:  65.868   1st Qu.:    1.73
##  Median :   31.46   Median :   17.038   Median : 121.833   Median :   14.51
##  Mean   :  330.79   Mean   :   96.034   Mean   : 287.571   Mean   :  786.65
##  3rd Qu.:  101.47   3rd Qu.:   42.093   3rd Qu.: 274.830   3rd Qu.:  101.01
##  Max.   :17898.33   Max.   :10459.152   Max.   :4116.836   Max.   :47936.24
##      ACVR1C             ABCA10             AQP7              GPR146
##  Min.   :    0.000   Min.   :  0.000   Min.   :    0.00   Min.   :   7.44
##  1st Qu.:    8.844   1st Qu.:  4.824   1st Qu.:    2.85   1st Qu.:  53.51
##  Median :   21.187   Median : 12.093   Median :   15.29   Median :  83.88
##  Mean   :  197.546   Mean   : 55.358   Mean   :  235.92   Mean   : 160.22
##  3rd Qu.:   59.182   3rd Qu.: 42.140   3rd Qu.:   87.05   3rd Qu.: 151.82
##  Max.   :15144.774   Max.   :861.533   Max.   :11613.34   Max.   :2849.61
##      ATP1A2             FXYD1            ARHGAP20             NPR1
##  Min.   :    0.00   Min.   :   0.00   Min.   :   0.4965   Min.   :   5.192
##  1st Qu.:    7.53   1st Qu.:   4.24   1st Qu.:  20.5644   1st Qu.:  78.937
##  Median :   28.63   Median :  16.15   Median :  41.1099   Median : 145.770
##  Mean   :  284.64   Mean   :  85.89   Mean   :  87.1257   Mean   : 388.295
##  3rd Qu.:   92.08   3rd Qu.:  64.75   3rd Qu.:  83.1719   3rd Qu.: 298.952
##  Max.   :58904.34   Max.   :4385.21   Max.   :1705.8824   Max.   :9271.230
##      ATOH8              ABCA9             ALDH1L1            ADAMTS5
##  Min.   :   0.5924   Min.   :   0.00   Min.   :    0.000   Min.   :   3.692
##  1st Qu.:  13.8957   1st Qu.:  18.83   1st Qu.:    3.215   1st Qu.: 116.083
##  Median :  30.1735   Median :  53.81   Median :   12.918   Median : 231.296
##  Mean   : 103.0018   Mean   : 195.25   Mean   :  168.590   Mean   : 500.670
##  3rd Qu.:  82.2975   3rd Qu.: 154.94   3rd Qu.:   64.642   3rd Qu.: 457.266
##  Max.   :2357.1137   Max.   :5814.01   Max.   :18530.408   Max.   :8410.825
##       RDH5               GPAM               CA4               KLHL29
##  Min.   :   1.609   Min.   :   77.66   Min.   :   0.0000   Min.   :   3.671
##  1st Qu.:  12.637   1st Qu.:  297.87   1st Qu.:   0.0000   1st Qu.:  32.181
##  Median :  26.371   Median :  448.50   Median :   0.8031   Median :  61.969
##  Mean   : 141.050   Mean   : 1629.26   Mean   :  49.4441   Mean   : 133.980
##  3rd Qu.:  72.489   3rd Qu.:  694.04   3rd Qu.:   9.0870   3rd Qu.: 130.499
##  Max.   :4722.437   Max.   :86428.08   Max.   :2345.7160   Max.   :1682.311
##     GPIHBP1           LOC728264            MAMDC2            TMEM132C
##  Min.   :   0.00   Min.   :   4.409   Min.   :   0.00   Min.   :   0.000
##  1st Qu.:  15.04   1st Qu.:  62.645   1st Qu.:  30.23   1st Qu.:   2.149
##  Median :  39.37   Median : 122.047   Median :  74.32   Median :  14.865
##  Mean   : 133.88   Mean   : 393.760   Mean   : 198.83   Mean   : 139.027
##  3rd Qu.: 114.95   3rd Qu.: 274.893   3rd Qu.: 187.22   3rd Qu.:  70.364
##  Max.   :3996.41   Max.   :7221.175   Max.   :4074.26   Max.   :6837.168
##      ITIH5              HSPB7              HSPB6               DMD
##  Min.   :    3.993   Min.   :    0.00   Min.   :    0.00   Min.   :    1.538
```

```
##  1st Qu.:  126.200   1st Qu.:   18.77   1st Qu.:   70.86   1st Qu.:   85.545
##  Median :  257.909   Median :   40.28   Median :  172.99   Median :  178.876
##  Mean   :  994.826   Mean   :  309.46   Mean   : 1076.23   Mean   :  527.454
##  3rd Qu.:  670.645   3rd Qu.:   93.77   3rd Qu.:  539.07   3rd Qu.:  465.328
##  Max.   :30476.583   Max.   :44881.59   Max.   :81041.37   Max.   :15983.971
##      SPRY2             IGFBP6             CXCL2              EBF1
##  Min.   :  13.63   Min.   :    4.438   Min.   :   0.000   Min.   :   6.39
##  1st Qu.: 111.79   1st Qu.:   75.485   1st Qu.:   5.498   1st Qu.: 100.18
##  Median : 198.26   Median :  147.285   Median :  17.584   Median : 173.87
##  Mean   : 360.58   Mean   :  351.459   Mean   :  91.627   Mean   : 358.76
##  3rd Qu.: 387.96   3rd Qu.:  287.143   3rd Qu.:  61.339   3rd Qu.: 314.75
##  Max.   :3602.64   Max.   :12035.628   Max.   :3290.953   Max.   :6512.31
##       KLB              CLEC3B            TMEM220             IBSP
##  Min.   :   0.00   Min.   :    2.174   Min.   :  3.618   Min.   :    0.00
##  1st Qu.:  11.49   1st Qu.:   76.331   1st Qu.: 26.863   1st Qu.:    3.99
##  Median :  22.29   Median :  146.489   Median : 49.115   Median :   18.06
##  Mean   : 123.70   Mean   :  463.011   Mean   : 78.556   Mean   :  126.85
##  3rd Qu.:  47.91   3rd Qu.:  322.470   3rd Qu.: 89.731   3rd Qu.:   58.11
##  Max.   :5352.45   Max.   :14579.682   Max.   :736.621   Max.   :88358.72
##      HIF3A              IGSF10             CIDEC             C2orf40
##  Min.   :   0.000   Min.   :   0.00   Min.   :    0.00   Min.   :   0.00
##  1st Qu.:   1.998   1st Qu.:  24.45   1st Qu.:    3.34   1st Qu.:   3.93
##  Median :   5.845   Median :  67.64   Median :   33.22   Median :  19.62
##  Mean   :  48.872   Mean   : 197.19   Mean   :  715.63   Mean   : 142.42
##  3rd Qu.:  24.567   3rd Qu.: 170.28   3rd Qu.:  221.10   3rd Qu.:  95.86
##  Max.   :2282.147   Max.   :6666.12   Max.   :32690.62   Max.   :6933.38
##       LEPR             ANGPTL1            class
##  Min.   :   8.036   Min.   :   0.00   Length:1212
##  1st Qu.:  95.140   1st Qu.:  16.13   Class :character
##  Median : 173.602   Median :  37.75   Mode  :character
##  Mean   : 391.079   Mean   : 110.21
##  3rd Qu.: 373.135   3rd Qu.:  89.81
##  Max.   :8709.717   Max.   :1919.26
```

Question 1

```
library(pcalg)

data_no_class <- data[, !names(data) %in% "class"]

suffStat <- list(C = cor(data_no_class), n = nrow(data_no_class))
pc.fit <- pc(suffStat, indepTest = gaussCItest, alpha = 0.05, labels = colnames(data_no_class))
```

```
library(qgraph)

qgraph(pc.fit@graph,  title = "Gene Regulatory Network")
```

Gene Regulatory Network