Lab 6

Omar Aldawy Ibrahim Aldawy

2025-04-21

Part 1: Data Wrangling

Task 1.0: Install and load required packages

```
install_and_load <- function(pkg) {
  if (!requireNamespace(pkg, quietly = TRUE)) {
    install.packages(pkg)
  }
  library(pkg, character.only = TRUE)
}

install_and_load("tidyverse")
install_and_load("ggplot2")
install_and_load("BiocManager")
install_and_load("clValid")
install_and_load("scatterplot3d")
install_and_load("e1071")
install_and_load("gridExtra")
install_and_load("caret")

# install limma from Bioconductor
BiocManager::install("limma")

library(limma)
```

Task 1.1: Data Acquisition

- Read the dataset CSV file into R.
- By this point you should have two dataframes one with expression data only and the second one with expression data + phenotypes.

```
full_dataframe <- read.csv("Brain_GSE50161.csv")
print(dim(full_dataframe))
```

```
## [1]   130 54677
```

```
sample_ids <- full_dataframe$samples
full_dataframe <- full_dataframe %>% select(-samples)
rownames(full_dataframe) <- sample_ids

expression.data <- full_dataframe %>% select(-type)
```

Task 1.2: PCA Before QC

- Remove NAs by filling them with the means of their respective genes.
- Perform PCA on the imputed data.

```
gene_means <- colMeans(expression.data, na.rm = TRUE)

expr_imputed <- as.data.frame(
  Map(function(col, m) ifelse(is.na(col), m, col),
      expression.data,
      gene_means)
)

pca_res <- prcomp(expr_imputed, center = TRUE, scale. = FALSE)
pcs <- as.data.frame(pca_res$x)

head(pcs)
```

```
##          PC1       PC2        PC3       PC4        PC5        PC6         PC7
## 1 -119.93240 24.827291 -10.094289 62.736207 -35.900332  21.742327 -44.4305862
## 2  -68.84101  5.001261  41.672421 16.250916  32.350386 -64.411493  44.3314873
## 3  -57.25181 49.371208   1.737344 -2.646363  27.757115   1.688111  18.3426467
## 4   46.49128 58.391208  46.214517 31.314746   3.109689  -5.567812  -0.6694227
## 5  -66.54843 40.078145  26.095697 13.651545 -12.288291 -12.117699   4.2830189
## 6 -101.31301 33.126327  14.090266 54.989242 -13.783524   1.213111 -28.4467665
##          PC8       PC9      PC10       PC11      PC12        PC13       PC14
## 1   2.578779 -33.04951   2.84858  -0.545314 21.721916  -0.4533251  16.225963
## 2  41.655497 -24.10353  17.57165 -15.188171 -9.466229   7.2231154  -9.462181
## 3 -33.842216  23.90596 -11.01189   7.874218  5.069082  -6.3591703 -38.882183
## 4  33.989638 -13.21527  41.04846  25.864821 15.727538  -0.4740046   7.933242
## 5 -34.207421 -14.21203 -12.90762  23.210891 18.726082 -11.4974716  13.081504
## 6  -6.328682 -20.03668  12.02804   7.924769 26.087239  17.8073629   2.354601
##         PC15       PC16        PC17        PC18       PC19        PC20
## 1 -17.733297  28.861092 -12.7379890 -0.04086854  5.2828635 -11.6802420
## 2  -2.899471  25.420079 -10.4125127  2.36901833 -9.7992806  29.9229049
## 3   5.933572  -6.350106 -10.3000867 12.01559370 -0.8050119  -4.2787188
## 4 -17.530470 -22.600162  26.9718050 16.17343965 22.0217208   0.2474391
## 5  -1.379229  19.229522  -0.7032287 16.82180366 29.6313830   7.2459199
## 6  -7.236260  16.658568   0.8177497 10.11497142  6.5483476  12.2956208
##         PC21      PC22        PC23         PC24       PC25        PC26
## 1  10.912207 -7.952277  12.0058398 -17.16277314  20.450362  -4.9870240
## 2  10.697658 -8.822624 -17.9688065  -9.21390255  -3.085605   0.7925117
## 3 -11.437524 10.399054   0.8955279  -4.56460734  -1.898755   0.4229886
## 4 -10.137517 -8.526321   3.2398733  11.67742519 -37.929939  27.7908511
## 5 -17.340688 24.357745   4.4613648  -0.01439121   5.950370 -19.7640405
## 6  -1.317917  2.943043   3.3928980  -2.56966093  22.491375 -10.7572144
##         PC27       PC28       PC29        PC30      PC31       PC32       PC33
## 1  2.9525241  -1.213378  14.492722  -0.3014649  7.570983 -12.319356  0.5174670
## 2 10.9637774  26.540315 -13.637973 -21.1355150 -9.502167  29.476719 -9.7028655
## 3  0.7810011   7.434822   4.758866  -8.5328561  6.430716  -5.175358  6.0537150
## 4 30.8505826 -30.107687 -14.846850  31.0260516 -8.503102   4.425759  4.6794467
## 5  3.0990702  -9.803361  25.638226  -4.2817025 -2.822556  25.325940 -0.3977699
## 6 10.9438014  16.904352 -13.190625  -9.5518601 11.663195 -24.359029 -3.8735481
##          PC34       PC35       PC36        PC37        PC38        PC39
## 1  -1.7150459   4.983598  -8.085452  3.54679555  -0.2305111  -0.2473667
```

```
## 2  11.6787762 -15.256134   8.690024  3.06978463   2.5618143 -12.3823686
## 3   0.8112779  13.076281   4.704832 -1.83591401  -6.5119976  -7.8544720
## 4  11.6951795  22.811177 -21.961723 -5.89540937 -29.1207871  -2.3355639
## 5 -10.9536018  13.679179   1.569094 -1.88690977  -0.2116142   5.7023866
## 6 -11.8380868  15.150467  14.683500 -0.07288528  -2.2302969  -7.9001986
##        PC40        PC41       PC42       PC43       PC44      PC45       PC46
## 1 -1.319043   2.1778335  6.5724859  -2.791390  -4.596974 0.9879256   2.653356
## 2 14.155200   3.5955975 15.5636197  36.035278  15.366230 8.5768259 -11.370283
## 3 -5.518397   5.8905983  4.7108949   4.573744   7.152647 1.2300964   8.615083
## 4  2.778788 -15.9044002  4.4174247  10.440855  -6.207317 7.1493639   8.292508
## 5 -9.214930   8.8924003  3.9321105   2.265603 -21.745650 7.9116596  -8.190611
## 6 -3.806003  -0.9667394 -0.3269828 -14.213204  10.567662 1.9600125  -1.553188
##         PC47       PC48      PC49      PC50      PC51       PC52      PC53
## 1   7.333299 -11.089065 -5.122409  6.794641  1.115385   3.714771  4.652493
## 2  14.752465  -5.461702 -6.189480  4.114286 -1.447320  23.316752 -4.207146
## 3   3.908798   6.346896 10.571855  6.875930 10.209163  -1.003060 -4.995462
## 4  36.059320   6.917615 25.660189  3.986815 -4.353559   4.089820  8.430732
## 5   7.221954   5.635017 -4.336607 -1.936774 -7.761778  26.836764 -8.596473
## 6 -11.004656 -16.250333 11.435577  9.415488 10.575381 -20.453578 -1.837568
##        PC54        PC55      PC56        PC57      PC58        PC59       PC60
## 1 -6.931218 -14.9459900  4.732695   3.0036562 11.802773 -10.4411937 -7.1170526
## 2 16.666096 -15.1694873 23.534413   2.9383622 -1.707837   6.4315584 -8.4720414
## 3  7.482779  -0.5941631  5.913367   2.1451077 12.571519  -0.8059244 -1.0685679
## 4  6.451074   5.4690214 -3.078068   7.1495762  9.723485   2.7593208 -0.9296478
## 5 -9.930130  18.8788271  2.352520 -17.1820306 -9.886098  -2.5609620 -5.4639143
## 6  8.377588  -3.3808112  3.930010   0.3704952 -2.964980   7.3372143 11.4637287
##         PC61       PC62       PC63       PC64       PC65        PC66       PC67
## 1 -7.9155437  -2.768513   2.829537  0.9441257  3.2308976   6.7475022  -3.474957
## 2 -6.4272015  -5.103617 -18.622389  2.1226011 -5.0474752 -17.1574691   2.757306
## 3  0.2180210   5.570324   3.336870 -0.8296533 -0.8045933  13.2733910   1.451469
## 4  0.3528927 -18.312729 -12.381985  9.6373256 14.8497166   4.9069475 -11.416655
## 5 -4.4105890  18.639486   2.490264  1.2014827  6.2657592   0.0240235 -17.529961
## 6 19.0370243 -13.444373 -11.183459  4.9240477  4.1540932  -3.6712308   2.030822
##         PC68       PC69        PC70          PC71        PC72        PC73
## 1 -2.2038827  -2.733830  -1.2751170 -9.474963e-04  -0.5237163  -0.3200528
## 2  3.5631105   2.551003 -10.6238110  7.159030e+00  -9.0068837  -7.7534084
## 3  8.3955400   3.900256   2.1981157 -1.801394e+01  -1.0995887  -7.1363924
## 4 -0.3001657 -12.087313   0.6596114 -8.226976e-02  -0.3636192   2.8838148
## 5 -4.8088808  -9.914981  16.5271941  9.649029e+00 -14.9673032 -11.8520264
## 6  6.6530404  -7.841669  -0.1268335  1.188596e+01  -8.6482684  -1.8330667
##          PC74        PC75       PC76       PC77        PC78       PC79
## 1  10.1075485  10.4409113 -0.0463334 -7.3931320  -1.8211545 -10.072533
## 2   5.9355256  11.8617796  2.1889176 -0.3343739  -5.2273931  -3.611012
## 3 -12.1127563  -0.3959444 -3.0851607  0.1018559  -0.4885251  -8.794861
## 4  -0.2392998  -1.5450462  1.6139016  3.1003416   3.7701116  -1.132516
## 5 -14.5914935  13.5943389  1.9564865  3.3731116   3.5846095   7.529591
## 6   1.4388618 -14.6384632 -0.8247056 24.5493133 -13.5089376   3.767613
##         PC80          PC81       PC82        PC83       PC84       PC85
## 1  5.5281570  1.182935e+00  1.7807670   4.5511577   4.882353  6.0323015
## 2  2.4977534  2.972756e-05 -9.4225982  -4.7374685   4.972373  2.7994594
## 3  6.3242225 -5.200839e+00 -7.4317382 -10.9155743  15.937461  0.9941390
## 4  0.9810064 -5.607016e-01 -0.4264722  -0.7917195  -5.136132 -0.2483665
## 5 -6.6024913  5.238729e+00 -2.7903805   1.8433221  -3.437882 16.1589037
## 6 -0.7416930  9.968909e+00  2.0326027 -11.8690452 -16.626941  7.6932400
```

```
##          PC86      PC87      PC88      PC89       PC90        PC91        PC92
## 1   5.4366908 -9.577054 -4.187302 -1.485353 -4.1969751  -0.7471969   3.3931635
## 2   3.1280555  5.467162  1.725334 -6.783418  5.4006017  -0.4370818  -1.3552201
## 3 -11.1341235 -2.179700  7.810434 -5.221097 -4.3290368   1.8671692 -10.4442715
## 4  -0.7272288  1.297104 -5.481196 -3.888267 -2.2063459   2.7671047   0.6790765
## 5  -1.4407917  3.983407  3.223037  9.290440  1.9256848 -14.7276216   7.9317249
## 6  -4.6010797 16.617697 10.125812  7.642596  0.5458442   8.8100281   7.1846428
##          PC93       PC94       PC95      PC96       PC97       PC98       PC99
## 1   7.9537661  8.3185606 -1.0884919 -6.084530 -4.5179499 -2.3318725 -0.2928333
## 2   1.3397646 -3.8699964  2.2712827 -1.579663 -0.6650955 -0.9139238 -1.2864487
## 3  -0.3896491  4.5749364 -4.7177100 13.274123 11.2456210 -0.1181166 -0.8673114
## 4   0.7452045 -0.5762996 -0.6796637 -2.576290  2.8622492 -5.8761905 -1.5787022
## 5  -5.2657071 -1.5060604  0.1501855  6.418179 -2.5047912  2.9048479 -8.6246823
## 6 -11.6966509 -5.6739051 -1.5541965  1.195087 -0.4288195 -8.3633413 -8.3099550
##        PC100     PC101     PC102      PC103      PC104        PC105       PC106
## 1  8.1551337  4.983589  4.367098  5.0928125  0.5700310 -13.56149509 -12.2257390
## 2 -0.5223751 -2.198926 -1.902898 -1.3692409  1.5186820  -1.99018323  -1.5209899
## 3 -3.4379898 15.544213 -4.447825 13.6391105 -0.7170527  -0.34275533   9.3055093
## 4  1.3485570 -1.351270 -1.600846  0.1045436  0.7060513  -1.81218692  -2.1266497
## 5  4.6617562 -3.282393 -7.127533  5.7689584  0.5200250   0.71604675   1.0090677
## 6  2.9606312  2.800001 -5.889269  0.1419995 -0.8079020   0.08423659   0.2147539
##          PC107       PC108      PC109      PC110      PC111       PC112
## 1 -1.439820711  7.96870163  1.5162638 12.2504257  5.8177077   2.7118322
## 2  1.266611394  0.09128964 -2.0992102  0.8977650 -0.2605935  -0.4867882
## 3 -7.203929936  1.61008096 -3.7070817 15.3320283  2.4284586 -21.1323940
## 4 -1.173660080  0.44212133 -0.4132916  0.5824686  0.7260252  -1.6454786
## 5  0.007905867 -1.20157071  1.7289882 -0.1562909 -0.7481727  -1.5029046
## 6 -1.532236537  0.99499234  3.1843217  4.3253168 -5.1056222   0.4203596
##        PC113      PC114      PC115      PC116       PC117       PC118
## 1 -2.0172946 -3.8757956 -1.7072924 -3.9758299 -10.4526737  1.10989511
## 2 -0.5068230  0.6483819  0.5343626 -0.4426099  -0.4057555  1.36815882
## 3 -0.9763932 -4.5221772  8.4676890  0.3584558   4.9334485  1.84564820
## 4 -0.9704757  1.3689293 -0.2868778  0.7076164  -0.2494771 -0.02014409
## 5 -1.9199575 -0.5164828  1.6630291  1.6076865  -1.7841272 -0.48374961
## 6 -0.4475982 -0.8253509 -0.4344208  4.6689466  -4.5984338  3.27295799
##         PC119      PC120      PC121      PC122      PC123      PC124
## 1 25.02944040 -6.0129004 -1.8884602 -6.8314603  8.8351460 -2.6999500
## 2 -0.68665331 -0.3593652  0.3784540  0.6600305 -0.5475772  0.1077035
## 3  2.06229396 -2.6615028 11.5892595  3.7224567 -1.3191123 -3.2861830
## 4 -0.12143832  0.1594981 -0.1661852  0.2632925 -0.2025322  0.9721007
## 5 -3.16841893 -0.8161031  0.1129896 -0.2165675 -0.4454278  1.8375385
## 6  0.05170174 -2.0146117 -2.0170764  1.3640659 -0.2329948  2.3670492
##         PC125      PC126       PC127       PC128       PC129         PC130
## 1 -2.97903256  3.7988662 -1.72371526  0.61597437 -0.58349585 -2.235561e-13
## 2 -0.31644179  0.1072522  0.20485887 -0.01968366 -0.30481594 -4.322571e-13
## 3  0.09220287  2.5044990  0.33828898 -0.76225701  1.18996028 -1.039375e-13
## 4  0.06192446 -0.2548097  0.50360056 -0.10426067  0.05544012 -1.494752e-14
## 5  0.29736806 -0.9379206  0.06013985 -0.35960285 -0.53202148 -4.139866e-13
## 6 -1.65334300  1.6733750  0.82040446  0.79877491 -0.15781011 -8.564705e-13
```

Task 1.3: PCA Before QC [Visualization]

- Visualize the PC1 vs PC2, PC1 vs PC3, and PC2 vs PC3 plots.
- Colors are according to the phenotype of each sample

```r
pcs_plot <- pcs %>%
  mutate(phenotype= full_dataframe$type)

p1 <- ggplot(pcs_plot, aes(PC1, PC2, color = phenotype)) +
  geom_point() +
  ggtitle("PC1 vs PC2 Before QC")

p2 <- ggplot(pcs_plot, aes(PC1, PC3, color = phenotype)) +
  geom_point() +
  ggtitle("PC1 vs PC3 Before QC")

p3 <- ggplot(pcs_plot, aes(PC2, PC3, color = phenotype)) +
  geom_point() +
  ggtitle("PC2 vs PC3 Before QC")

grid.arrange(p1, p2, p3, ncol = 1)
```

Task 1.4: Data Cleaning

- Remove outliers from the data.
- Re-impute the data after removing outliers.
- Perform quantile normalization on the data (Normalizes expression intensities so that the intensities or log-ratios have similar distributions across a set of arrays).

```
# Replace outliers with NA using apply
expr_clean <- as.data.frame(
  apply(expr_imputed, 2, function(vals) {
    mu <- mean(vals)
    sdv <- sd(vals)
    outliers <- abs(vals - mu) > 3 * sdv
    vals[outliers] <- NA
```

```
    return(vals)
  })
)

# Re-impute missing after outlier removal
new_gene_means <- colMeans(expr_clean, na.rm = TRUE)
expr_clean <- as.data.frame(
  Map(function(col, m) ifelse(is.na(col), m, col),
      expr_clean,
      new_gene_means)
)
# Quantile normalization
expr_norm <- normalizeBetweenArrays(as.matrix(expr_clean), method = "quantile")
expr_norm <- as.data.frame(expr_norm)

# Explanation:
# `normalizeBetweenArrays()` adjusts the distributions of gene expression
# values across samples to be the same; the quantile method aligns
# empirical quantiles, ensuring comparability across arrays.
```

Task 1.5: Data Inspection

- Check the distribution of the data before and after normalization.

```
pca_norm <- prcomp(expr_norm, center = TRUE, scale. = FALSE)
pcs_norm <- as.data.frame(pca_norm$x)
head(pcs_norm)
```

```
##         PC1      PC2       PC3       PC4         PC5        PC6        PC7
## 1 -91.66181 18.46092 -37.83219 59.902227 -11.6308255  44.270490 -31.863292
## 2 -49.21615 12.86853  51.17655 10.759418  31.3222788 -21.248329   4.263852
## 3 -42.23751 36.59539  17.76164 -6.337331  15.3096276 -19.771474  11.278661
## 4  60.70806 18.52200  26.07607 27.492593  -0.2948381  14.241085  34.065638
## 5 -38.38705 31.18689  16.72598 31.990190  -1.1355782   5.522872 -43.355787
## 6 -72.47518 20.35510 -11.59330 56.283563  -3.3253633  20.706451 -19.413471
##          PC8        PC9        PC10       PC11       PC12       PC13       PC14
## 1 -26.269361  21.941904 -19.7820812  12.670441  -2.092730  9.6767388  -6.158493
## 2 -12.689859 -33.800869   8.1477639  16.664084 -39.748517 10.9538537 -20.321550
## 3  14.800202 -12.845988   0.1133992 -26.319503  26.577826 -5.3256252  -4.484167
## 4 -19.864696 -13.145974  -7.8721014  -3.101286 -23.257807 -3.1891688  10.728850
## 5   1.551413  -4.025139 -27.0317463   1.938334   9.762934 -9.8919824   2.229049
## 6 -17.846032   1.367711 -21.3627106  -2.224670  -4.354057  0.6385803 -19.009796
##        PC15        PC16       PC17       PC18       PC19       PC20        PC21
## 1  9.809573  -8.8303046  0.9372469  -2.192460 -12.072512 -9.230792   7.3139944
## 2  6.673205   3.8689931 -4.4599414  -2.288978  -4.645045  7.549552 -18.0559547
## 3 18.536527   4.5956411 14.5688468   5.710921  15.253255  8.803289  -2.1413740
## 4  4.643211 -24.2815524 -11.4683668  5.690325  18.379709 -7.703705  12.7356699
## 5 12.800297  11.6209587  -5.3659415 -5.654590   1.157146  1.108098  -1.3834650
## 6  1.536763   0.9796452  1.4619540 -11.115324   5.552319  8.453915  -0.5328217
##          PC22       PC23       PC24       PC25       PC26       PC27        PC28
## 1  -0.5746241   7.603374 -4.7906116  5.067100 -4.0906724  1.644949  -1.6336518
## 2   2.6872751   9.807959 -0.1542869 -3.157790 -5.2665953 15.940312  -9.6124656
## 3   5.4361476  -3.713468 -1.4347187  6.119747  0.3885904 -5.434759   2.0648362
## 4 -11.6234959   7.207767 -1.0721257 -4.065366  4.1228041 -9.054264  -0.7509123
```

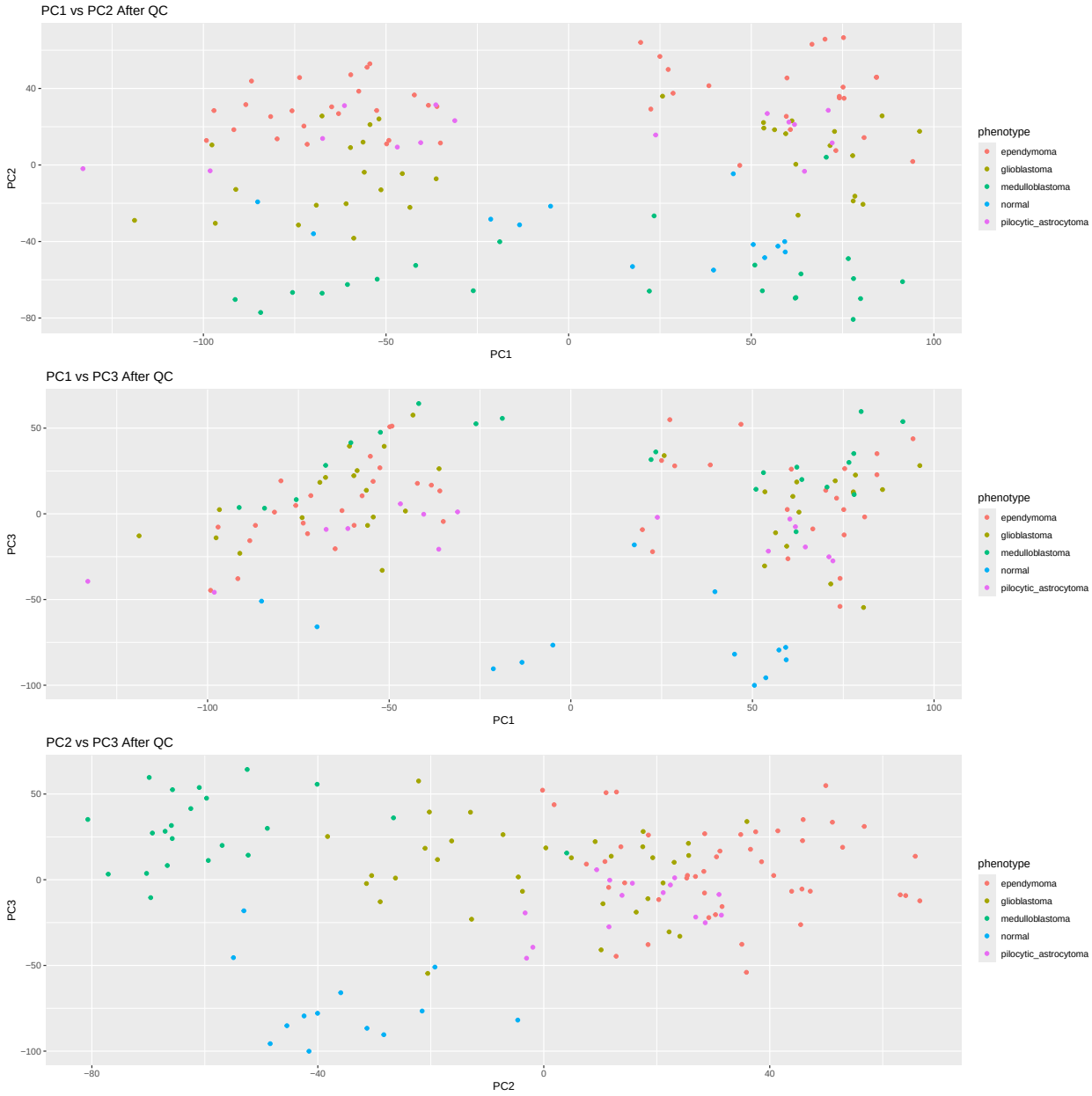
```
## 5 -18.7680053 -20.080747 -9.2402099  9.074533 -2.1129410 10.557703   9.8034174
## 6  -6.0119313   8.954091 -1.8648268  5.415404 -9.7154049  3.319007 -20.5770884
##        PC29       PC30      PC31       PC32       PC33       PC34       PC35
## 1 12.114949  -7.178393 11.754220 -20.848068  10.922862  -3.156749   4.933811
## 2 -3.396890 -10.515732  3.493692  14.329714   5.607671  15.318675 -20.829849
## 3 13.493919  -1.697055 -2.242778  -2.385621   8.589111   9.137029  -5.182252
## 4  3.886464  -7.013062 -1.216654  17.371273 -17.466998  -7.729865   5.004284
## 5 10.357448  -4.222694 -1.605271  -1.837850 -13.379107 -14.342129 -10.450935
## 6 17.750246  13.291496 -9.371251  -2.617660  -5.606546   2.936529   6.719579
##         PC36        PC37       PC38       PC39      PC40        PC41
## 1  -6.413467  -5.2524333   3.231752  -7.992833 22.055167   0.8394198
## 2  -3.166676  -1.4460244 -22.511665 -22.120329 -9.630694  -3.9713727
## 3   0.091110 -10.7441199   6.934720   2.923016  7.930806  -7.6469090
## 4 -13.627912   8.1912480  26.045077   6.538160  8.181689   4.4509212
## 5 -15.222542   0.2663931  -6.204964  10.822000  1.120565  10.7676081
## 6  16.412011  -0.6620488   8.909985   4.486773 -1.078688 -10.4995101
##          PC42       PC43        PC44       PC45        PC46       PC47
## 1   1.0524177   2.440311  -4.0102734  -7.006467   4.5889260  3.5364301
## 2 -20.8753382 -13.259880  -3.0514579  -1.410135   7.8928190 10.5165599
## 3   0.2925945  -2.368455   5.9219017  -1.786099   5.8266783 -2.0406224
## 4  -9.7354491   1.846787 -21.4880085 -24.318228 -15.8862138 -0.4665691
## 5  -3.4349197   7.818752  -0.9950213   6.774443  13.4093216 11.9064071
## 6   3.3706001   1.900727  16.1459955   6.983747   0.7669582  0.5892932
##         PC48        PC49       PC50       PC51       PC52       PC53      PC54
## 1  -2.852321  -0.2971495 -10.315050  -3.784924   6.845927   6.364289 -7.261437
## 2  -6.907769  -7.7271240  -2.814964 -11.759129   9.548651  -7.875377 -5.109196
## 3  -3.838071  -8.9275420  -6.733576  -8.859117   4.451029  -8.778108 -3.764561
## 4 -12.083608 -15.3374698   8.521742  -2.162885  -6.412029 -20.584367 20.708489
## 5 -10.349380   5.3365768   2.386373 -10.209070  10.344846   8.791504 27.737784
## 6  10.776384  -0.1327187  -1.247270   6.542193 -12.675413   5.356517 -5.144122
##          PC55        PC56      PC57       PC58      PC59      PC60        PC61
## 1  -1.0788645  -0.1586225 -2.552087   4.412386  4.913328  2.797164 -5.53905449
## 2   0.9622349 -12.1406233 -9.033134 -15.544541  9.406763 18.485204 -0.03456041
## 3  -5.8612808   7.7045956  3.567098  -3.818155 -1.445936  1.012271 -1.83164592
## 4 -10.5405024  12.2377230 -1.694580  -1.002441  1.657736  6.233066 -3.72858290
## 5   4.7113017 -10.3222591  4.106768  -4.020470 -7.691563 -2.599476  5.47840610
## 6   7.1882505   1.3025077 13.865972  -4.412213 -6.354921  8.715718 -8.95176872
##         PC62        PC63       PC64      PC65       PC66      PC67       PC68
## 1 -3.8179418  -7.4646587   3.067144  8.085455  0.6388518  3.656808  0.8663305
## 2  3.3913750  -9.6177989 -10.192862 -8.691382 -5.5068859  4.811400  5.5521377
## 3  7.6284242  -2.1887704  13.590732 -9.823012 -2.0711729  2.164336  3.9797517
## 4 -0.9022489   0.9863942  20.825249  9.212996 -3.8567702 17.017686 -5.6955195
## 5 -1.4738247   1.4525447   2.909124  5.200892 -1.6196276  5.382067  3.7924116
## 6  7.1758605 -10.7893591  -6.468697  6.971455 -0.9986690  6.476056  4.6497677
##         PC69      PC70       PC71      PC72       PC73       PC74       PC75
## 1  -9.181489 -1.882496 -1.4634980 -4.887243  7.1642665  4.1532957  1.9117199
## 2  12.575587 -5.855922 12.4207768 -2.395147 -2.8526307  0.7214526 -1.6804544
## 3  -2.136191 -1.220240 -1.0916327 -1.163674  0.5821001  4.3741150  0.1818238
## 4  16.294937  6.759186  9.2668613  2.829066 -3.5552940 -0.1444241  8.8751617
## 5 -21.618329 20.083830 -0.9981809  4.935875  3.4716174 -12.1627118 -1.0972785
## 6  21.698613 -7.319587 -9.6244483  5.522868 -1.4485206  8.3341910 -5.8031339
##          PC76       PC77        PC78       PC79      PC80      PC81      PC82
## 1   0.7947095  5.0135957   0.8191000 -4.4957293 -1.043270 -4.754723 -3.865914
## 2 -12.9710302 14.6683528   0.3096906 -5.9740855 -3.249420  4.511326  5.996589
```

```
## 3  -4.2881612 -0.9794356   8.1461826  2.0792609 -2.901278 -5.160753 -5.118885
## 4  -1.0774081 10.6499650  -3.9600063 10.8750283 -1.319446 10.280089 -5.169247
## 5 -26.2091337 -2.2965439  -9.1602591 -0.5520696 -5.716144  0.655327 -3.280247
## 6  -4.5354008  0.6635965 -12.3873661  7.7938174  8.878827 -3.433235 -8.253099
##         PC83      PC84      PC85       PC86       PC87       PC88       PC89
## 1 -7.6629411  1.960395  4.138042  3.1020685   1.810274   6.909908   4.746096
## 2  0.4772508 10.616772  9.557230 -4.1426793  -2.057030  -3.502673   7.798675
## 3 -8.3509586  2.828436 -9.030920  1.6346302 -10.449166  -3.371392   1.096440
## 4 -4.9072414 -4.951317 -4.527349 -0.2642927   8.266561  -4.953660   1.985734
## 5 -1.3092021 -1.224719 -5.052514 -2.7674427  -1.364435 -10.492675  -2.606502
## 6 13.8789543  1.025341 -6.821520 -4.1431135  -1.847044  -3.397619 -16.289359
##        PC90       PC91       PC92       PC93      PC94        PC95       PC96
## 1 -1.529516 -5.2845135  -8.174418  -1.272759 -6.025442  13.4201725  0.8010566
## 2 11.518022 -3.5711755  -1.038111   5.984955  1.186786   5.0191625  8.1411682
## 3  1.091852 -5.6478300   2.798466   4.747722  5.080899 -18.4892043  5.1647237
## 4  5.004409  0.7899476   2.500173  -3.151771  1.158003  -1.2073756 -6.0846403
## 5 -2.686007 -5.2152120 -11.672044   4.107813 -6.406675  -0.4995136 13.3422127
## 6 10.711904  4.2592968  -5.429720 -13.832931 16.788286  -1.0956164  1.0130204
##        PC97       PC98        PC99     PC100     PC101      PC102     PC103
## 1  2.367454  5.5931108 -27.5287649 -7.645025 -9.148043 -6.7685248  5.273377
## 2  8.062327 -1.2312229   2.0866429 -6.019498  3.765601  0.7079325 -4.579788
## 3 -3.019974  0.2099759  -3.9638555  2.618486 -5.366366 -3.5413111 -7.685771
## 4  3.178207  3.3857624  -0.4912187 -5.542967  4.350661 -1.5407766  4.054482
## 5  2.540261 -1.3813150   7.4476240  8.399179  2.695332  8.6201208  9.770047
## 6  6.000768 -1.2660664   5.4977848 -2.018941  3.549158 13.7709002  6.897728
##       PC104      PC105     PC106       PC107       PC108     PC109       PC110
## 1 14.496731 19.6371211  6.356148 -12.2677041 -13.9200240 -3.595968 -1.65115338
## 2  1.357059 -2.4598604  6.951901   0.3510554   0.5239934 -6.882441 -1.32268411
## 3 -1.161830 -1.0442822 -6.045696   5.7219071   2.4664381  5.377590 16.24279971
## 4  2.505302  1.7493837  3.718209  -1.1230799   0.9009402 -1.459771  0.04032209
## 5  2.130557  1.9384227  1.674512   5.0021484   7.5244488 -2.433567  0.65793478
## 6  7.628291  0.7962132 -3.393428  -5.5935848   4.4472168 -2.877378  2.24273124
##        PC111      PC112       PC113     PC114       PC115      PC116
## 1 13.9580530  1.7369893  -2.9321822 -3.669606  -5.7891983  0.8156073
## 2  3.8482340  0.2296650   3.0586164 -1.513745   0.3377609  0.1259450
## 3  8.4837213  6.9916989 -15.8910143 -5.251490 -10.2821751 19.7485020
## 4  0.6570842 -0.3191966  -0.9492123  2.666985  -0.3211794 -1.8888274
## 5 -2.4077897  2.7641046   3.0072321  1.082876  -5.8200575 -1.0544731
## 6 -0.2214633  8.1470240  -1.1848480  1.552709 -13.4913986 -3.4431561
##         PC117      PC118        PC119      PC120      PC121      PC122
## 1   0.6616463  0.6133686  0.563560564  1.9085029 -2.3935672 -0.6850051
## 2  -4.3129499  2.4268223  0.888416753  0.9438208  0.2585674 -1.6764848
## 3 -15.4746522 -6.6602657 -1.236849828 -7.2611597 -6.0410503  3.8770047
## 4  -2.7697305 -0.3371165  0.333847172  0.5645988 -0.1977469 -0.7046990
## 5   0.2192694 -2.2597183 -0.009373023 -1.9448278  4.4689427 -2.0516717
## 6  -0.3196146  0.2633060  9.592031001  0.2493069  1.4193979  0.3756665
##        PC123      PC124      PC125      PC126       PC127       PC128
## 1  0.5424589  0.3837293 -1.5011999  1.8531203  -1.1769630  3.06986464
## 2  0.6896699  0.2181854  0.2175052 -0.4007371  -0.9276218 -1.16350050
## 3 -3.1537460  0.2285349  1.0800451 11.8957053 -13.3602421 -7.21790665
## 4  0.6694846  0.4613463  0.7999281 -0.4212578   0.2891144 -0.85893361
## 5 -0.5201540  3.0635632 -1.5077222 -1.4310458  -0.2967790 -0.01557093
## 6  0.4670280 -2.9016470  0.2483461 -2.8377712  -0.4806704 -0.16832687
##        PC129         PC130
```

```
## 1 -1.1893807  8.371119e-12
## 2  2.0214683 -2.908535e-13
## 3 -0.9407434 -1.314908e-13
## 4  1.2130060 -4.977003e-14
## 5 -0.2541687 -2.670978e-13
## 6  3.9132028 -1.527958e-12
```

- Visualize the PC1 vs PC2, PC1 vs PC3, and PC2 vs PC3 plots after normalization.
- Colors are according to the phenotype of each sample

```
pcs_norm_plot <- pcs_norm %>%
  mutate(phenotype= full_dataframe$type)

p1 <- ggplot(pcs_norm_plot, aes(PC1, PC2, color = phenotype)) +
  geom_point() +
  ggtitle("PC1 vs PC2 After QC")

p2 <- ggplot(pcs_norm_plot, aes(PC1, PC3, color = phenotype)) +
  geom_point() +
  ggtitle("PC1 vs PC3 After QC")

p3 <- ggplot(pcs_norm_plot, aes(PC2, PC3, color = phenotype)) +
  geom_point() +
  ggtitle("PC2 vs PC3 After QC")

grid.arrange(p1, p2, p3, ncol = 1)
```



Part 2: Analysis

Task 2.1: Regression Analysis

Read the list of top 5000 genes from the file `top_5000.txt`.

- Select only these genes from the normalized data.

```
# Encode phenotype: Tumor = 1, Normal = 0
phenotype_binary <- ifelse(full_dataframe$type == "normal", 0, 1)

# Load list of top 5000 genes
top_genes <- read.table("top_5000.txt", header = FALSE, stringsAsFactors = FALSE)
```

```r
# Extract only these genes from normalized data
genes_subset <- expr_norm[, colnames(expr_norm) %in% top_genes$V1]
```

Perform logistic regression for each gene.

- Create a dataframe with the p-values and coefficients for each gene.

```r
# Perform logistic regression for each gene
results_without_pc1 <- data.frame(Gene = character(),
                      p_value = numeric(),
                      coefficient = numeric())

for(gene in colnames(genes_subset)) {
  model <- glm(phenotype_binary ~ genes_subset[,gene], family = binomial)
  results_without_pc1 <- rbind(results_without_pc1, data.frame(
    Gene = gene,
    p_value = summary(model)$coefficients[2,4],
    coefficient = summary(model)$coefficients[2,1]
  ))
}
```

Perform logistic regression for each gene with PC1 as a covariate.

```r
results_with_pc1 <- data.frame(Gene = character(),
                      p_value = numeric(),
                      coefficient = numeric())

for(gene in colnames(genes_subset)) {
  model <- glm(phenotype_binary ~ genes_subset[,gene] + pcs_norm$PC1, family = binomial)
  results_with_pc1 <- rbind(results_with_pc1, data.frame(
    Gene = gene,
    p_value = summary(model)$coefficients[2,4],
    coefficient = summary(model)$coefficients[2,1]
  ))
}
```

Filter the results to get significant genes ($p < 0.05$) and sort them by p-value.

```r
# Get significant genes (p < 0.05)
significant_genes_no_pc1 <- results_without_pc1 %>% filter(p_value < 0.05) %>% arrange(p_value)
significant_genes_pc1 <- results_with_pc1 %>% filter(p_value < 0.05) %>% arrange(p_value)
```

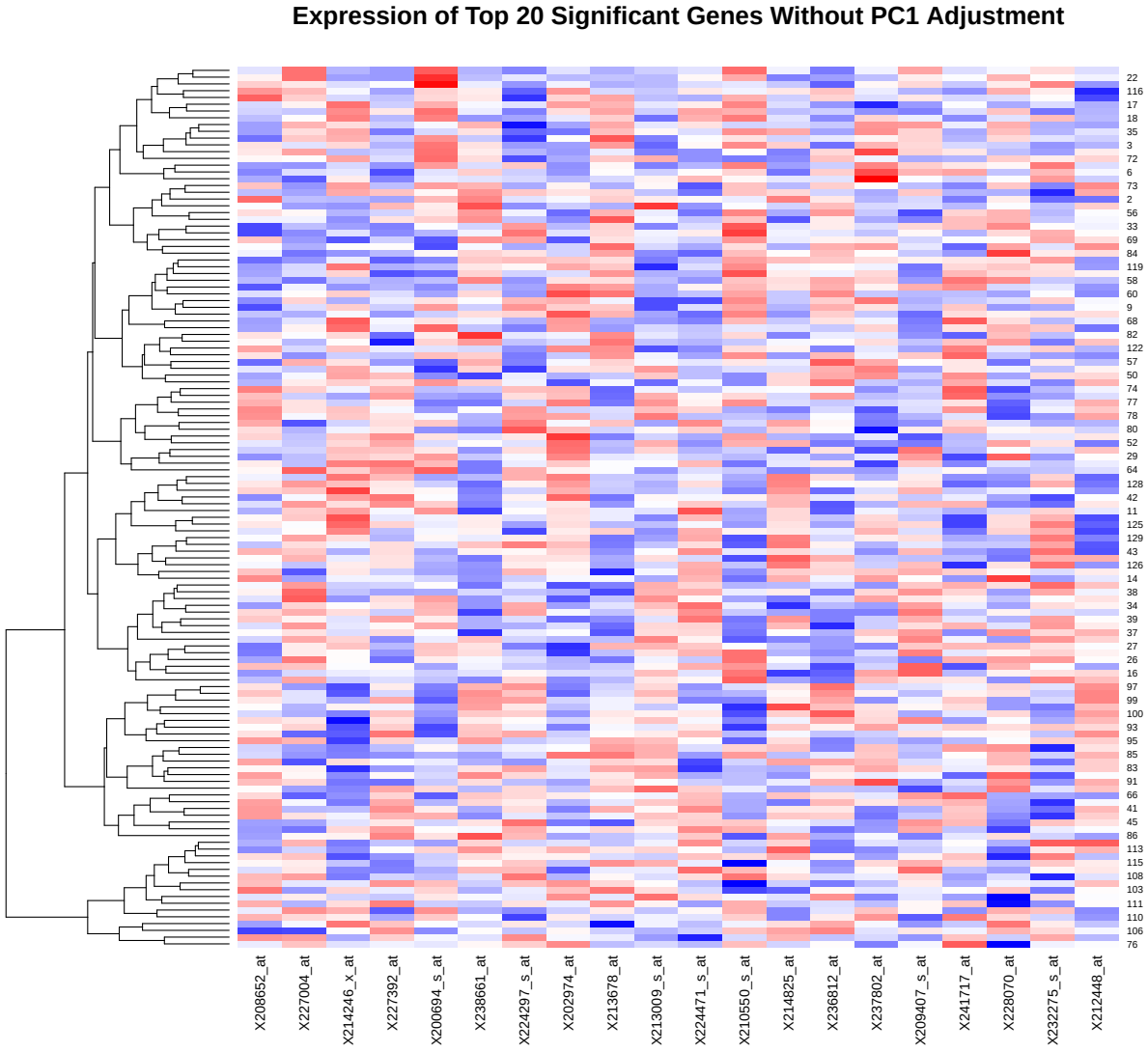
Task 2.2: Visualization

Heatmap

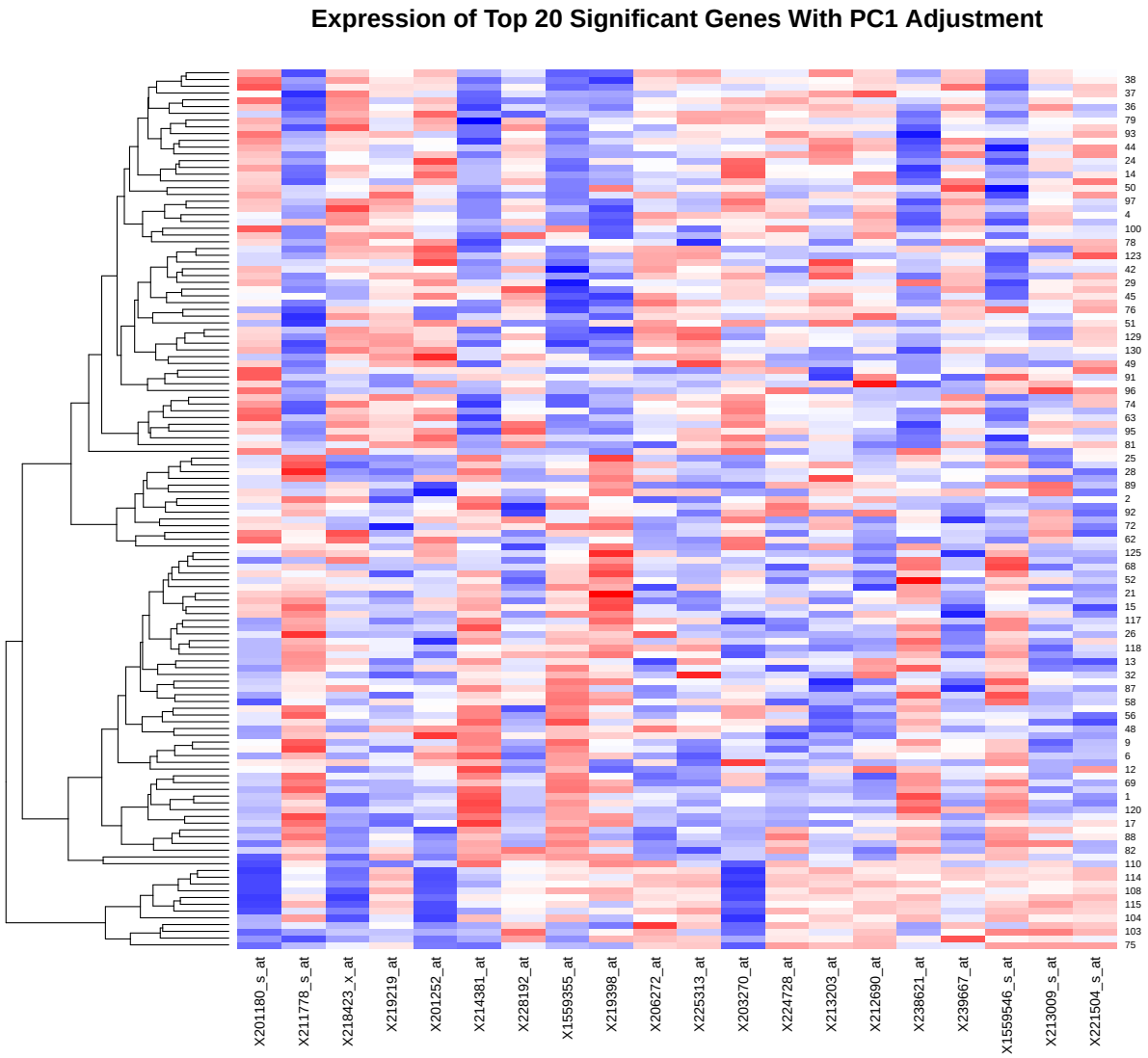
- Create a heatmap of the top 20 significant genes (without PC1 adjustment).
- Create a heatmap of the top 20 significant genes (with PC1 adjustment).

```r
# Heatmap of top 20 significant genes
top_20_genes <- significant_genes_no_pc1$Gene[1:20]
heatmap_data <- as.matrix(genes_subset[, top_20_genes])

heatmap(heatmap_data, Colv = NA, scale = "row",
```

```
        col = colorRampPalette(c("blue", "white", "red"))(100),
        main = "Expression of Top 20 Significant Genes Without PC1 Adjustment")
```




```
top_20_genes_pc1 <- significant_genes_pc1$Gene[1:20]
heatmap_data_pc1 <- as.matrix(genes_subset[, top_20_genes_pc1])

heatmap(heatmap_data_pc1, Colv = NA, scale = "row",
        col = colorRampPalette(c("blue", "white", "red"))(100),
        main = "Expression of Top 20 Significant Genes With PC1 Adjustment")
```



Volcano plot

- Calculate log2 fold changes for the whole genes.
- Add the log2 fold changes to results dataframes.

```
# Calculate log2 fold changes manually
phenotype <- full_dataframe$type

logFC <- apply(expr_norm, 2, function(x) {
  tumor_mean <- mean(x[phenotype != "normal"])
  normal_mean <- mean(x[phenotype == "normal"])
  log2(tumor_mean / normal_mean)
})
```

```r
# Add logFC to your results dataframes
results_without_pc1$logFC <- logFC[match(results_without_pc1$Gene, names(logFC))]
results_with_pc1$logFC <- logFC[match(results_with_pc1$Gene, names(logFC))]
```

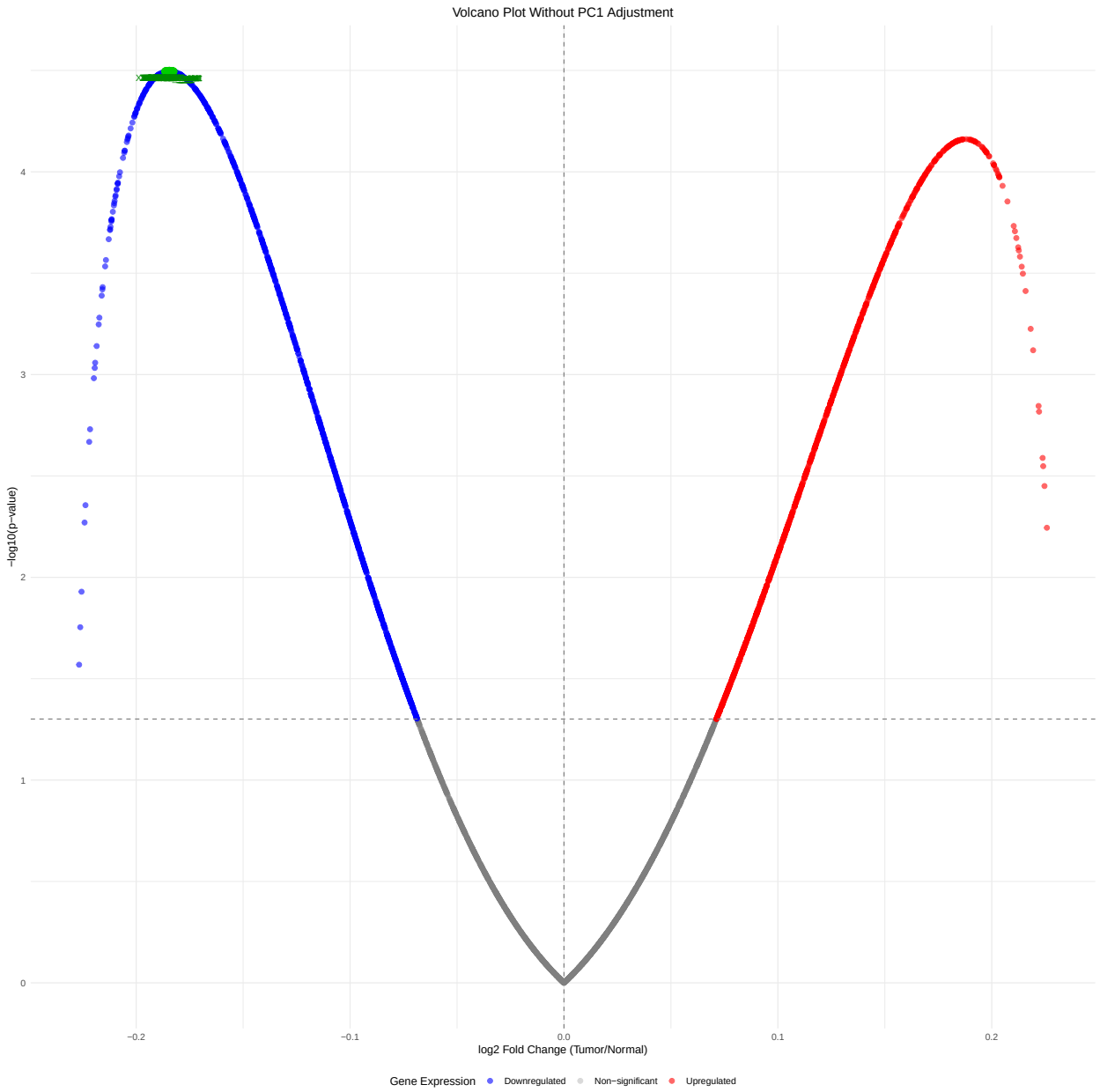
- Create volcano plots for the results without and with PC1 adjustment.

```r
create_volcano_plot_all <- function(results_df, title) {
  volcano_data <- results_df %>%
    mutate(
      neg_log_pval = -log10(p_value),
      direction = case_when(
        p_value >= 0.05 ~ "Non-sig",
        logFC > 0 ~ "Up",
        logFC <= 0 ~ "Down"
      ),
      significance = ifelse(p_value < 0.05, "Significant", "Non-significant"),
      top_20 = ifelse(p_value < 0.05 & rank(p_value) <= 20, TRUE, FALSE)
    )

  ggplot(volcano_data, aes(x = logFC, y = neg_log_pval)) +
    # Non-significant points
    geom_point(
      data = filter(volcano_data, significance == "Non-significant"),
      aes(color = direction),
      alpha = 0.3,
      size = 2
    ) +
    # Significant points (but not top 20)
    geom_point(
      data = filter(volcano_data, significance == "Significant", !top_20),
      aes(color = direction),
      alpha = 0.6,
      size = 2
    ) +
    # Top 20 points - always green
    geom_point(
      data = filter(volcano_data, top_20),
      color = "green3",  # fixed color
      size = 3,
      shape = 21,        # filled circle with border
      fill = "green3"
    ) +
    # Top 20 labels
    geom_text(
      data = filter(volcano_data, top_20),
      aes(label = Gene),
      color = "green4",  # darker green for text
      vjust = 1.5,
      hjust = 0.5,
      size = 3,
      show.legend = FALSE
    ) +
    scale_color_manual(
      values = c("Down" = "blue", "Up" = "red", "Non-sig" = "grey50"),
```

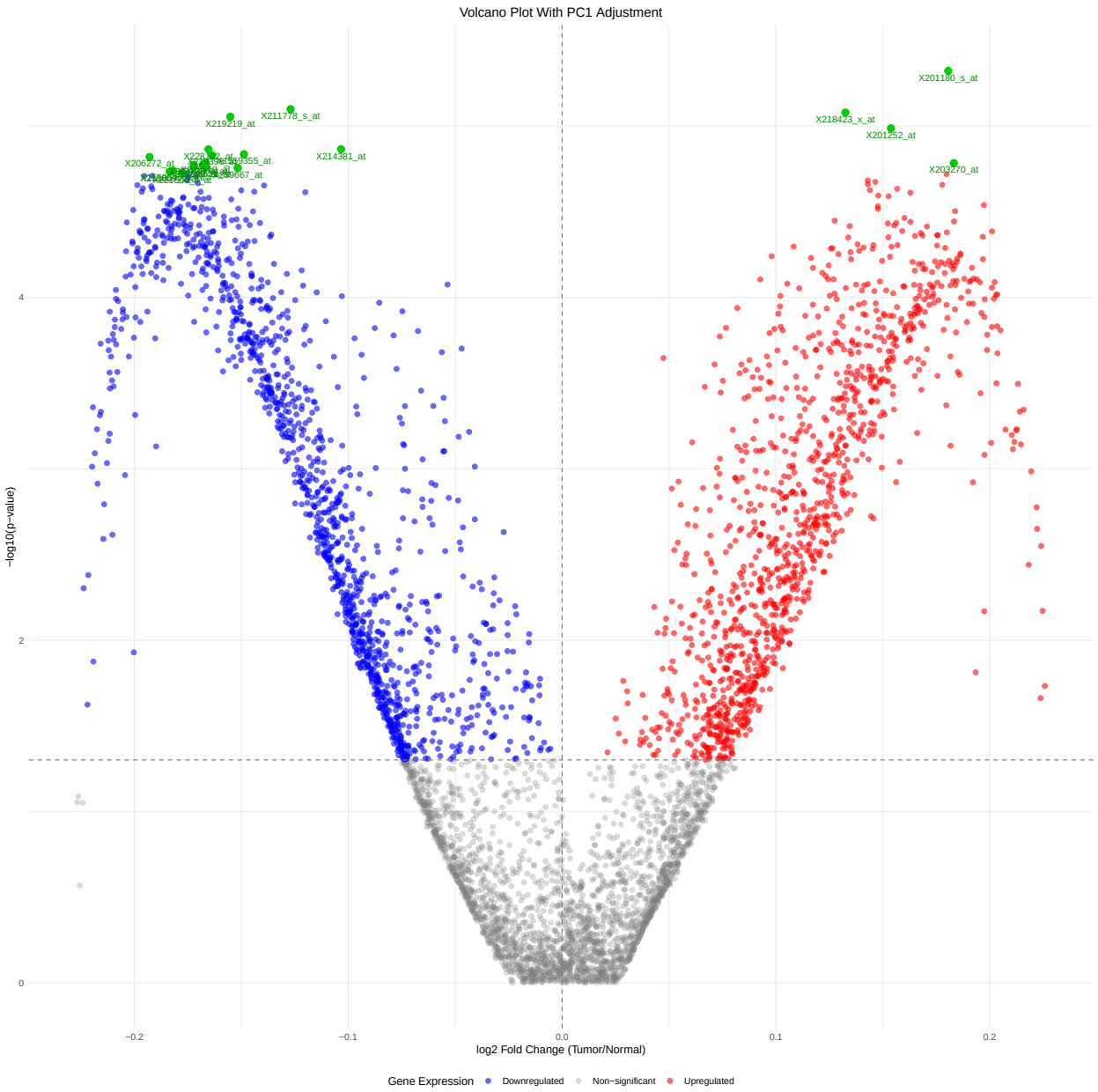
```
      labels = c("Downregulated", "Non-significant", "Upregulated")
    ) +
    geom_vline(xintercept = 0, linetype = "dashed", color = "grey50") +
    geom_hline(yintercept = -log10(0.05), linetype = "dashed", color = "grey50") +
    labs(
      x = "log2 Fold Change (Tumor/Normal)",
      y = "-log10(p-value)",
      title = title,
      color = "Gene Expression"
    ) +
    theme_minimal() +
    theme(
      legend.position = "bottom",
      plot.title = element_text(hjust = 0.5)
    )
}
```

- Draw the volcano plots.

```
# Create plots
create_volcano_plot_all(results_without_pc1, "Volcano Plot Without PC1 Adjustment")
```




```
create_volcano_plot_all(results_with_pc1, "Volcano Plot With PC1 Adjustment")
```



Part 3: Annotation

Save the top 20 genes from both analyses to text files.

```
# For top_20_genes (without PC3 adjustment)
print(top_20_genes)
```

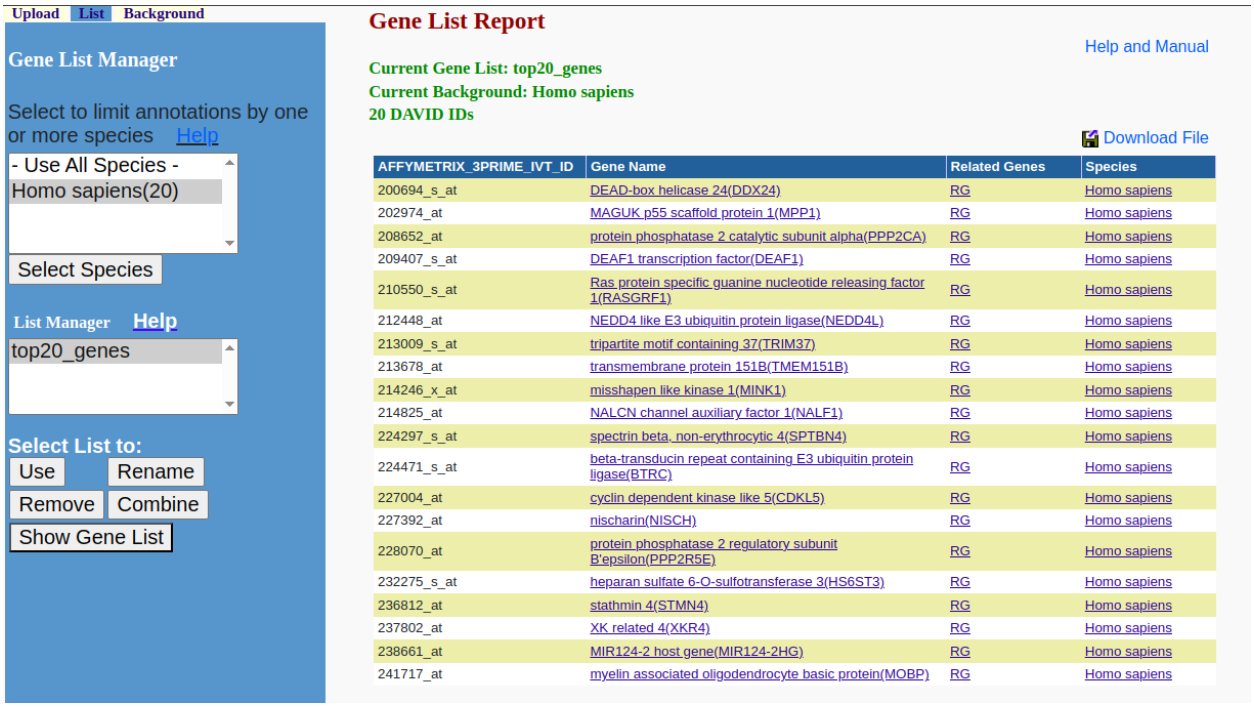
```
##  [1] "X208652_at"   "X227004_at"   "X214246_x_at" "X227392_at"   "X200694_s_at"
##  [6] "X238661_at"   "X224297_s_at" "X202974_at"   "X213678_at"   "X213009_s_at"
## [11] "X224471_s_at" "X210550_s_at" "X214825_at"   "X236812_at"   "X237802_at"
## [16] "X209407_s_at" "X241717_at"   "X228070_at"   "X232275_s_at" "X212448_at"
```

```r
# Remove first character from each gene name
top_20_genes_trimmed <- sub("^.", "", top_20_genes)  # ^. matches first character

write.table(top_20_genes_trimmed,
            file = "top20_genes.txt",
            quote = FALSE,
            row.names = FALSE,
            col.names = FALSE)

# For top_20_genes_pc3 (with PC3 adjustment)
print(top_20_genes_pc1)
```

```
##  [1] "X201180_s_at"  "X211778_s_at"  "X218423_x_at"  "X219219_at"
##  [5] "X201252_at"    "X214381_at"    "X228192_at"    "X1559355_at"
##  [9] "X219398_at"    "X206272_at"    "X225313_at"    "X203270_at"
## [13] "X224728_at"    "X213203_at"    "X212690_at"    "X238621_at"
## [17] "X239667_at"    "X1559546_s_at" "X213009_s_at"  "X221504_s_at"
```

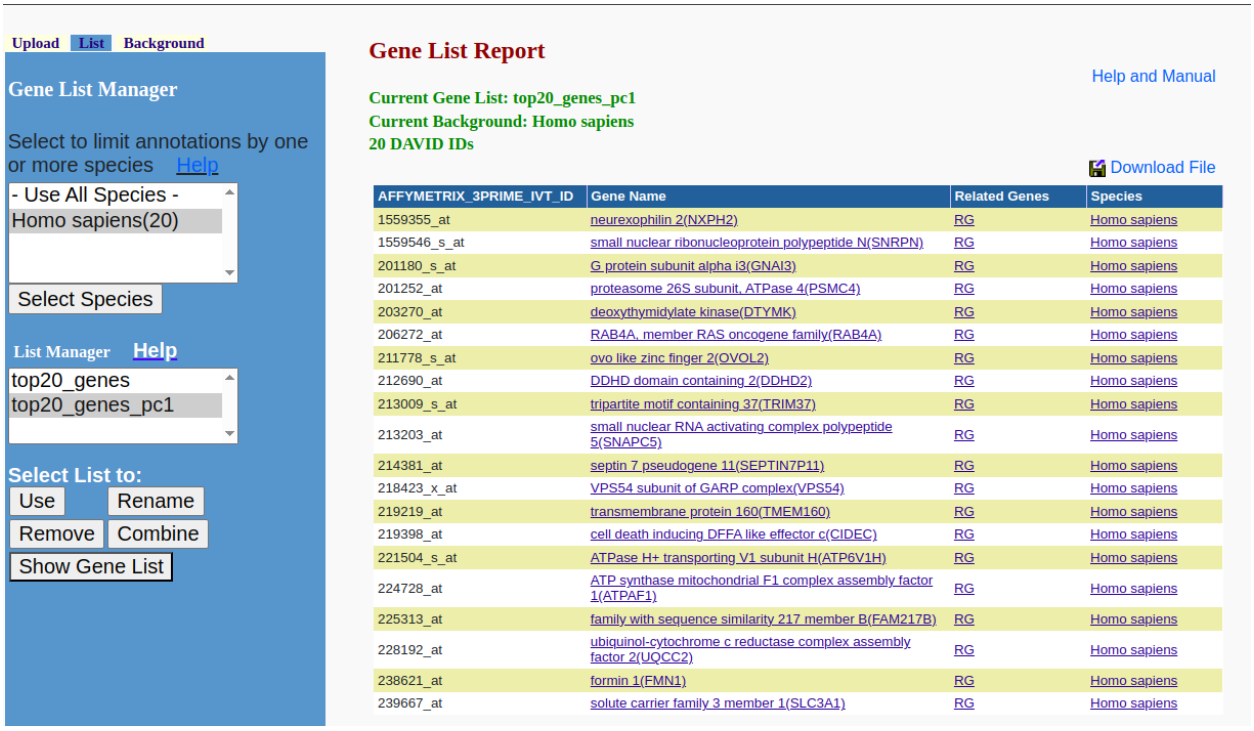
```r
# Remove first character from each gene name
top_20_genes_pc1_trimmed <- sub("^.", "", top_20_genes_pc1)

write.table(data.frame(Gene = top_20_genes_pc1_trimmed),
            file = "top20_genes_pc1.txt",
            quote = FALSE,
            row.names = FALSE,
            col.names = FALSE)
```

Use david tools to convert the top 20 significant gene names to normal gene names.



AFFYMETRIX_3PRIME_IVT_ID	Gene Name	Related Genes	Species
200694_s_at	DEAD-box helicase 24(DDX24)	RG	Homo sapiens
202974_at	MAGUK p55 scaffold protein 1(MPP1)	RG	Homo sapiens
208652_at	protein phosphatase 2 catalytic subunit alpha(PPP2CA)	RG	Homo sapiens
209407_s_at	DEAF1 transcription factor(DEAF1)	RG	Homo sapiens
210550_s_at	Ras protein specific guanine nucleotide releasing factor 1(RASGRF1)	RG	Homo sapiens
212448_at	NEDD4 like E3 ubiquitin protein ligase(NEDD4L)	RG	Homo sapiens
213009_s_at	tripartite motif containing 37(TRIM37)	RG	Homo sapiens
213678_at	transmembrane protein 151B(TMEM151B)	RG	Homo sapiens
214246_x_at	misshapen like kinase 1(MINK1)	RG	Homo sapiens
214825_at	NALCN channel auxiliary factor 1(NALF1)	RG	Homo sapiens
224297_s_at	spectrin beta, non-erythrocytic 4(SPTBN4)	RG	Homo sapiens
224471_s_at	beta-transducin repeat containing E3 ubiquitin protein ligase(BTRC)	RG	Homo sapiens
227004_at	cyclin dependent kinase like 5(CDKL5)	RG	Homo sapiens
227392_at	nischarin(NISCH)	RG	Homo sapiens
228070_at	protein phosphatase 2 regulatory subunit B'epsilon(PPP2R5E)	RG	Homo sapiens
232275_s_at	heparan sulfate 6-O-sulfotransferase 3(HS6ST3)	RG	Homo sapiens
236812_at	stathmin 4(STMN4)	RG	Homo sapiens
237802_at	XK related 4(XKR4)	RG	Homo sapiens
238661_at	MIR124-2 host gene(MIR124-2HG)	RG	Homo sapiens
241717_at	myelin associated oligodendrocyte basic protein(MOBP)	RG	Homo sapiens

Figure 1: Normal gene names(no PC1)



AFFYMETRIX_3PRIME_IVT_ID	Gene Name	Related Genes	Species
1559355_at	neurexophilin 2(NXPH2)	RG	Homo sapiens
1559546_s_at	small nuclear ribonucleoprotein polypeptide N(SNRPN)	RG	Homo sapiens
201180_s_at	G protein subunit alpha i3(GNAI3)	RG	Homo sapiens
201252_at	proteasome 26S subunit, ATPase 4(PSMC4)	RG	Homo sapiens
203270_at	deoxythymidylate kinase(DTYMK)	RG	Homo sapiens
206272_at	RAB4A, member RAS oncogene family(RAB4A)	RG	Homo sapiens
211778_s_at	ovo like zinc finger 2(OVOL2)	RG	Homo sapiens
212690_at	DDHD domain containing 2(DDHD2)	RG	Homo sapiens
213009_s_at	tripartite motif containing 37(TRIM37)	RG	Homo sapiens
213203_at	small nuclear RNA activating complex polypeptide 5(SNAPC5)	RG	Homo sapiens
214381_at	septin 7 pseudogene 11(SEPTIN7P11)	RG	Homo sapiens
218423_x_at	VPS54 subunit of GARP complex(VPS54)	RG	Homo sapiens
219219_at	transmembrane protein 160(TMEM160)	RG	Homo sapiens
219398_at	cell death inducing DFFA like effector c(CIDEC)	RG	Homo sapiens
221504_s_at	ATPase H+ transporting V1 subunit H(ATP6V1H)	RG	Homo sapiens
224728_at	ATP synthase mitochondrial F1 complex assembly factor 1(ATPAF1)	RG	Homo sapiens
225313_at	family with sequence similarity 217 member B(FAM217B)	RG	Homo sapiens
228192_at	ubiquinol-cytochrome c reductase complex assembly factor 2(UQCC2)	RG	Homo sapiens
238621_at	formin 1(FMN1)	RG	Homo sapiens
239667_at	solute carrier family 3 member 1(SLC3A1)	RG	Homo sapiens

Figure 2: Normal gene names(with PC1)

Install and load the `hgu133plus2.db` package.

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("hgu133plus2.db")

library(hgu133plus2.db)
```

Map the probe IDs to gene symbols using the `mapIds` function.

```
gene_symbols <- mapIds(hgu133plus2.db,
                       keys = top_20_genes_trimmed,
                       column = "SYMBOL",
                       keytype = "PROBEID")
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
print(gene_symbols)
```

```
##    208652_at    227004_at  214246_x_at    227392_at  200694_s_at    238661_at
##     "PPP2CA"     "CDKL5"      "MINK1"      "NISCH"      "DDX24" "MIR124-2HG"
##  224297_s_at    202974_at    213678_at  213009_s_at  224471_s_at  210550_s_at
##     "SPTBN4"      "MPP1"   "TMEM151B"     "TRIM37"      "BTRC"    "RASGRF1"
##    214825_at    236812_at    237802_at  209407_s_at    241717_at    228070_at
##     "NALF1"     "STMN4"      "XKR4"      "DEAF1"      "MOBP"    "PPP2R5E"
##  232275_s_at    212448_at
##     "HS6ST3"     "NEDD4L"
```

```
gene_symbols_pc1 <- mapIds(hgu133plus2.db,
                      keys = top_20_genes_pc1_trimmed,
                      column = "SYMBOL",
                      keytype = "PROBEID")
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
print(gene_symbols_pc1)
```

```
##  201180_s_at  211778_s_at  218423_x_at    219219_at    201252_at    214381_at
##      "GNAI3"      "OVOL2"      "VPS54"    "TMEM160"      "PSMC4" "SEPTIN7P11"
##    228192_at   1559355_at    219398_at    206272_at    225313_at    203270_at
##      "UQCC2"      "NXPH2"      "CIDEC"      "RAB4A"    "FAM217B"      "DTYMK"
##    224728_at    213203_at    212690_at    238621_at    239667_at 1559546_s_at
##     "ATPAF1"     "SNAPC5"      "DDHD2"       "FMN1"     "SLC3A1"      "SNRPN"
##  213009_s_at  221504_s_at
##     "TRIM37"    "ATP6V1H"
```

Extracting Kegg pathways annotation

- Install and load the `enrichR` package.
- Perform KEGG pathway enrichment analysis using the `enrichR` function.
- Specify the KEGG database.

```
install.packages("enrichR")
library(enrichR)
```

```
dbs <- c("KEGG_2021_Human") # Specify KEGG database
```

- Extract the KEGG results.
- Save the results to a CSV file.

```
enriched <- enrichr(gene_symbols, dbs)
```

```
## Uploading data to Enrichr... Done.
##   Querying KEGG_2021_Human... Done.
## Parsing results... Done.
```

```
# Extract KEGG results
kegg_results <- enriched[["KEGG_2021_Human"]]

# Save results
write.csv(kegg_results, "kegg_pathways.csv", row.names = FALSE)
print(kegg_results)
```

```
##                                       Term Overlap      P.value
## 1                            Oocyte meiosis   3/129 0.0002757874
## 2             Ubiquitin mediated proteolysis   3/140 0.0003507100
## 3                  mRNA surveillance pathway    2/98 0.0042631106
## 4             Sphingolipid signaling pathway   2/119 0.0062188008
## 5                     AMPK signaling pathway   2/120 0.0063204270
## 6                       Dopaminergic synapse   2/132 0.0075988978
## 7     Adrenergic signaling in cardiomyocytes   2/150 0.0097166465
## 8                     Hippo signaling pathway   2/163 0.0113913007
## 9                              Tight junction   2/169 0.0122042960
## 10                          Circadian rhythm    1/31 0.0305618601
## 11 Aldosterone-regulated sodium reabsorption    1/37 0.0363736593
```

```
## 12            Human papillomavirus infection   2/331 0.0426431405
## 13                PI3K-Akt signaling pathway   2/354 0.0481252626
## 14            Glycosaminoglycan biosynthesis    1/53 0.0517104132
## 15                Hedgehog signaling pathway    1/56 0.0545601179
## 16                      Long-term depression    1/60 0.0583470713
## 17                TGF-beta signaling pathway    1/94 0.0899594741
## 18                           Chagas disease   1/102 0.0972497716
## 19                               Autophagy   1/137 0.1284974931
## 20                     Cellular senescence   1/156 0.1450274169
## 21                             Hepatitis C   1/157 0.1458891094
## 22                   Wnt signaling pathway   1/166 0.1536073032
## 23                           Focal adhesion   1/201 0.1829972755
## 24  Human immunodeficiency virus 1 infection   1/212 0.1920320863
## 25                   Ras signaling pathway   1/232 0.2082163693
## 26                               Shigellosis   1/246 0.2193616046
## 27                               Endocytosis   1/252 0.2240923673
## 28                     MAPK signaling pathway   1/294 0.2564531517
##    Adjusted.P.value Old.P.value Old.Adjusted.P.value Odds.Ratio Combined.Score
## 1        0.00490994           0                   0  27.806723     227.900566
## 2        0.00490994           0                   0  25.559897     203.343059
## 3        0.03539439           0                   0  23.013889     125.604195
## 4        0.03539439           0                   0  18.863248      95.828660
## 5        0.03539439           0                   0  18.702448      94.708609
## 6        0.03546152           0                   0  16.965812      82.788956
## 7        0.03796892           0                   0  14.888889      68.993842
## 8        0.03796892           0                   0  13.677709      61.206452
## 9        0.03796892           0                   0  13.182302      58.080791
## 10       0.08557321           0                   0  35.000000     122.080086
## 11       0.09258750           0                   0  29.157895      96.626651
## 12       0.09950066           0                   0   6.636609      20.937765
## 13       0.10184555           0                   0   6.195707      18.797453
## 14       0.10184555           0                   0  20.170040      59.745598
## 15       0.10184555           0                   0  19.066986      55.455415
## 16       0.10210737           0                   0  17.770740      50.492824
## 17       0.14816855           0                   0  11.254669      27.105700
## 18       0.15127742           0                   0  10.359041      24.141462
## 19       0.18936473           0                   0   7.679567      15.757287
## 20       0.19451881           0                   0   6.731749      12.997879
## 21       0.19451881           0                   0   6.688259      12.874287
## 22       0.19550020           0                   0   6.320574      11.840685
## 23       0.22277929           0                   0   5.205263       8.840015
## 24       0.22403743           0                   0   4.931155       8.136863
## 25       0.23239208           0                   0   4.499658       7.060762
## 26       0.23239208           0                   0   4.239527       6.431506
## 27       0.23239208           0                   0   4.136926       6.187588
## 28       0.25645315           0                   0   3.536375       4.812332
##                 Genes
## 1  PPP2CA;PPP2R5E;BTRC
## 2   NEDD4L;TRIM37;BTRC
## 3       PPP2CA;PPP2R5E
## 4       PPP2CA;PPP2R5E
## 5       PPP2CA;PPP2R5E
## 6       PPP2CA;PPP2R5E
## 7       PPP2CA;PPP2R5E
```

```
## 8          PPP2CA;BTRC
## 9        PPP2CA;NEDD4L
## 10               BTRC
## 11             NEDD4L
## 12      PPP2CA;PPP2R5E
## 13      PPP2CA;PPP2R5E
## 14             HS6ST3
## 15               BTRC
## 16             PPP2CA
## 17             PPP2CA
## 18             PPP2CA
## 19             PPP2CA
## 20               BTRC
## 21             PPP2CA
## 22               BTRC
## 23            RASGRF1
## 24               BTRC
## 25            RASGRF1
## 26               BTRC
## 27             NEDD4L
## 28            RASGRF1
```

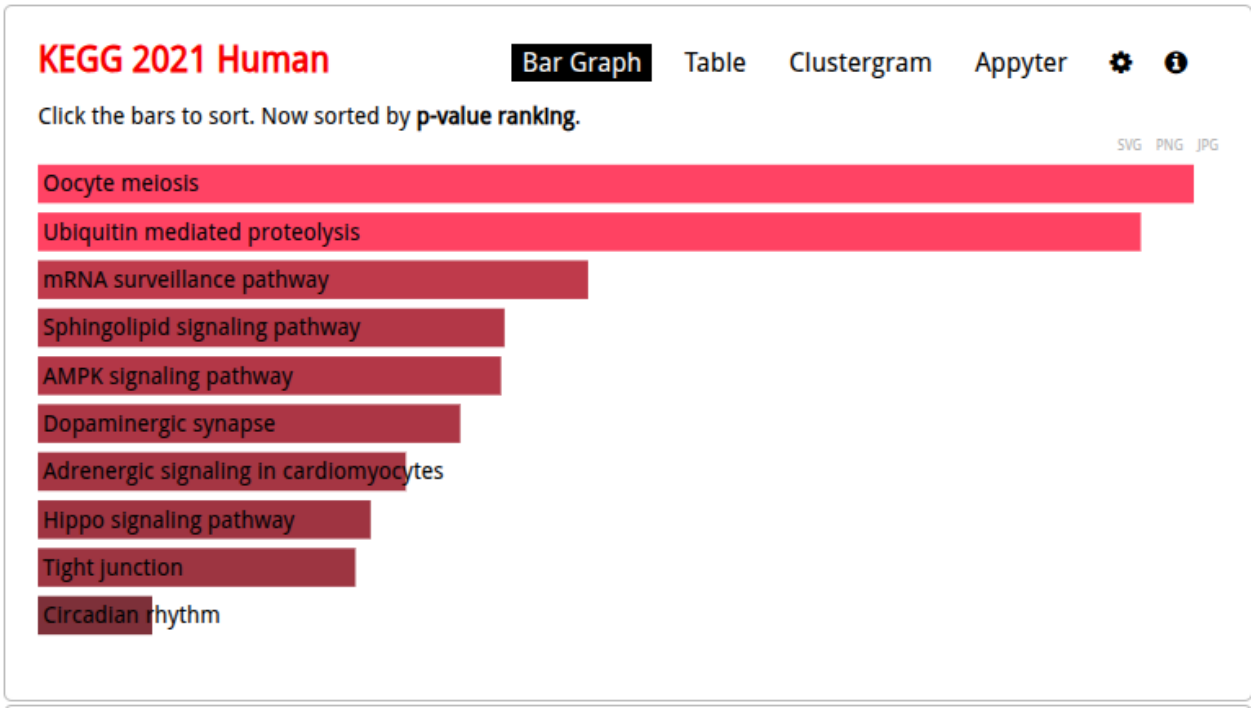


Figure 3: NO_PC1

```
enriched_pc1 <- enrichr(gene_symbols_pc1, dbs)
```

```
## Uploading data to Enrichr... Done.
##   Querying KEGG_2021_Human... Done.
## Parsing results... Done.
```

```
# Extract KEGG results
kegg_results_pc1 <- enriched_pc1[["KEGG_2021_Human"]]
```

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Oocyte meiosis	0.0002758	0.004910	27.81	227.90
2	Ubiquitin mediated proteolysis	0.0003507	0.004910	25.56	203.34
3	mRNA surveillance pathway	0.004263	0.03539	23.01	125.60
4	Sphingolipid signaling pathway	0.006219	0.03539	18.86	95.83
5	AMPK signaling pathway	0.006320	0.03539	18.70	94.71
6	Dopaminergic synapse	0.007599	0.03546	16.97	82.79
7	Adrenergic signaling in cardiomyocytes	0.009717	0.03797	14.89	68.99
8	Hippo signaling pathway	0.01139	0.03797	13.68	61.21
9	Tight junction	0.01220	0.03797	13.18	58.08
10	Circadian rhythm	0.03056	0.08557	35.00	122.08
11	Aldosterone-regulated sodium reabsorption	0.03637	0.09259	29.16	96.63
12	Human papillomavirus infection	0.04264	0.09950	6.64	20.94
13	PI3K-Akt signaling pathway	0.04813	0.1018	6.20	18.80
14	Glycosaminoglycan biosynthesis	0.05171	0.1018	20.17	59.75
15	Hedgehog signaling pathway	0.05456	0.1018	19.07	55.46
16	Long-term depression	0.05835	0.1021	17.77	50.49
17	TGF-beta signaling pathway	0.08996	0.1482	11.25	27.11
18	Chagas disease	0.09725	0.1513	10.36	24.14
19	Autophagy	0.1285	0.1894	7.68	15.76
20	Cellular senescence	0.1450	0.1945	6.73	13.00
21	Hepatitis C	0.1459	0.1945	6.69	12.87
22	Wnt signaling pathway	0.1536	0.1955	6.32	11.84
23	Focal adhesion	0.1830	0.2228	5.21	8.84
24	Human immunodeficiency virus 1 infection	0.1920	0.2240	4.93	8.14
25	Ras signaling pathway	0.2082	0.2324	4.50	7.06
26	Shigellosis	0.2194	0.2324	4.24	6.43
27	Endocytosis	0.2241	0.2324	4.14	6.19
28	MAPK signaling pathway	0.2565	0.2565	3.54	4.81

Showing 1 to 28 of 28 entries | Export entries to table

Previous Next

Terms marked with an * have an overlap of less than 5

Figure 4: NO_PC1

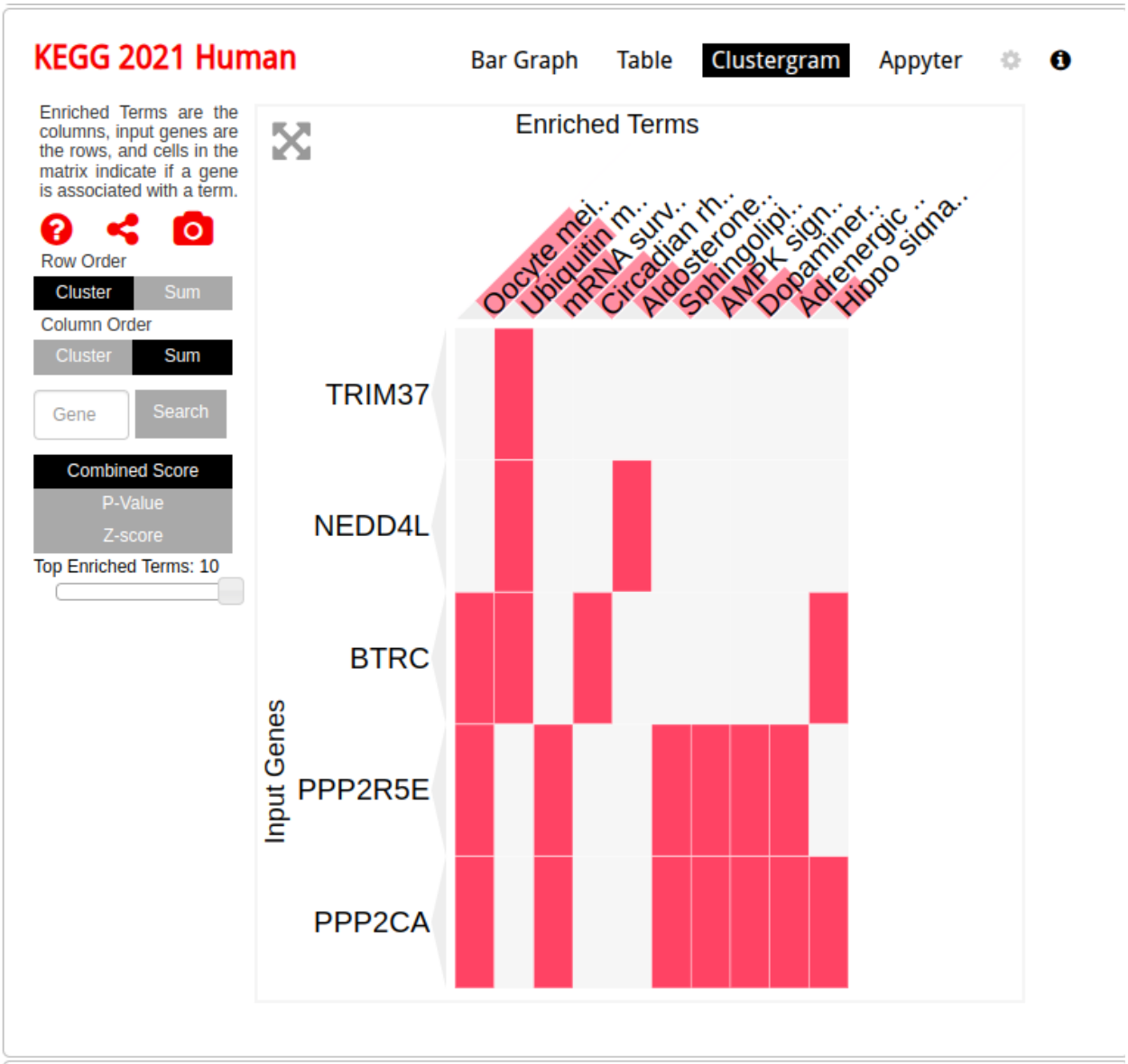


Figure 5: NO_PC1

```r
# Save results
write.csv(kegg_results_pc1, "kegg_pathways_pc3.csv", row.names = FALSE)
print(kegg_results_pc1)
```

```
##                                                    Term Overlap
## 1                                    Parkinson disease   2/249
## 2                                           Proteasome    1/46
## 3                                   Cocaine addiction    1/49
## 4                             Vibrio cholerae infection    1/50
## 5                 Regulation of lipolysis in adipocytes    1/55
## 6                              Pyrimidine metabolism    1/56
## 7                               Long-term depression    1/60
## 8                                    Renin secretion    1/69
## 9  Epithelial cell signaling in Helicobacter pylori infection    1/70
## 10                                          Pertussis    1/76
## 11                           Gastric acid secretion    1/76
## 12                           Synaptic vesicle cycle    1/78
## 13                                     Gap junction    1/88
## 14                                GABAergic synapse    1/89
## 15                               Morphine addiction    1/91
## 16                             Rheumatoid arthritis    1/93
## 17                             Circadian entrainment    1/97
## 18              Progesterone-mediated oocyte maturation   1/100
## 19                                      Melanogenesis   1/101
## 20                                    Chagas disease   1/102
## 21                 Protein digestion and absorption   1/103
## 22 Parathyroid hormone synthesis, secretion and action   1/106
## 23                                      Toxoplasmosis   1/112
## 24                             Serotonergic synapse   1/113
## 25                             Cholinergic synapse   1/113
## 26                 Leukocyte transendothelial migration   1/114
## 27                           Glutamatergic synapse   1/114
## 28                       Sphingolipid signaling pathway   1/119
## 29      Growth hormone synthesis, secretion and action   1/119
## 30                           Platelet activation   1/124
## 31                                      Lysosome   1/128
## 32                       Relaxin signaling pathway   1/129
## 33                            Dopaminergic synapse   1/132
## 34                       Oxidative phosphorylation   1/133
## 35                        Apelin signaling pathway   1/137
## 36                      Estrogen signaling pathway   1/137
## 37                    Ubiquitin mediated proteolysis   1/140
## 38                        Spinocerebellar ataxia   1/143
## 39                 Retrograde endocannabinoid signaling   1/148
## 40               Adrenergic signaling in cardiomyocytes   1/150
## 41                                      Phagosome   1/152
## 42                          mTOR signaling pathway   1/154
## 43                      Oxytocin signaling pathway   1/154
## 44                           Cushing syndrome   1/155
## 45                      cGMP-PKG signaling pathway   1/167
## 46                                    Tuberculosis   1/180
## 47                                  Axon guidance   1/182
## 48                                     Alcoholism   1/186
```

```
## 49                                Chemokine signaling pathway   1/192
## 50                               Epstein-Barr virus infection   1/202
## 51                                    Rap1 signaling pathway   1/210
## 52                    Human immunodeficiency virus 1 infection   1/212
## 53                                    cAMP signaling pathway   1/216
## 54                              Human cytomegalovirus infection   1/225
## 55                                    Chemical carcinogenesis   1/239
## 56                                                Endocytosis   1/252
## 57                                              Prion disease   1/273
## 58                                         Huntington disease   1/306
## 59                               Human papillomavirus infection   1/331
## 60                                Amyotrophic lateral sclerosis   1/364
## 61                                         Alzheimer disease   1/369
## 62                                Pathways of neurodegeneration   1/475
## 63                                         Pathways in cancer   1/531
##       P.value Adjusted.P.value Old.P.value Old.Adjusted.P.value Odds.Ratio
## 1  0.02530959        0.2064179           0                   0   8.876743
## 2  0.04502934        0.2064179           0                   0  23.315789
## 3  0.04789811        0.2064179           0                   0  21.855263
## 4  0.04885255        0.2064179           0                   0  21.408163
## 5  0.05361112        0.2064179           0                   0  19.421053
## 6  0.05456012        0.2064179           0                   0  19.066986
## 7  0.05834707        0.2064179           0                   0  17.770740
## 8  0.06681510        0.2064179           0                   0  15.411765
## 9  0.06775151        0.2064179           0                   0  15.187643
## 10 0.07335128        0.2064179           0                   0  13.968421
## 11 0.07335128        0.2064179           0                   0  13.968421
## 12 0.07521075        0.2064179           0                   0  13.604238
## 13 0.08445508        0.2064179           0                   0  12.034483
## 14 0.08537467        0.2064179           0                   0  11.897129
## 15 0.08721122        0.2064179           0                   0  11.631579
## 16 0.08904426        0.2064179           0                   0  11.377574
## 17 0.09269987        0.2064179           0                   0  10.901316
## 18 0.09543242        0.2064179           0                   0  10.569378
## 19 0.09634153        0.2064179           0                   0  10.463158
## 20 0.09724977        0.2064179           0                   0  10.359041
## 21 0.09815715        0.2064179           0                   0  10.256966
## 22 0.10087408        0.2064179           0                   0   9.962406
## 23 0.10628464        0.2064179           0                   0   9.421053
## 24 0.10718338        0.2064179           0                   0   9.336466
## 25 0.10718338        0.2064179           0                   0   9.336466
## 26 0.10808127        0.2064179           0                   0   9.253377
## 27 0.10808127        0.2064179           0                   0   9.253377
## 28 0.11255786        0.2064179           0                   0   8.859054
## 29 0.11255786        0.2064179           0                   0   8.859054
## 30 0.11701310        0.2064179           0                   0   8.496791
## 31 0.12056198        0.2064179           0                   0   8.227518
## 32 0.12144709        0.2064179           0                   0   8.162829
## 33 0.12409731        0.2064179           0                   0   7.974689
## 34 0.12497903        0.2064179           0                   0   7.913876
## 35 0.12849749        0.2064179           0                   0   7.679567
## 36 0.12849749        0.2064179           0                   0   7.679567
## 37 0.13112751        0.2064179           0                   0   7.512685
## 38 0.13374999        0.2064179           0                   0   7.352854
```

```
## 39 0.13810408       0.2064179          0                   0   7.100967
## 40 0.13983989       0.2064179          0                   0   7.004945
## 41 0.14157238       0.2064179          0                   0   6.911467
## 42 0.14330155       0.2064179          0                   0   6.820433
## 43 0.14330155       0.2064179          0                   0   6.820433
## 44 0.14416490       0.2064179          0                   0   6.775803
## 45 0.15446078       0.2162451          0                   0   6.282181
## 46 0.16548182       0.2238105          0                   0   5.822111
## 47 0.16716521       0.2238105          0                   0   5.757197
## 48 0.17052230       0.2238105          0                   0   5.631579
## 49 0.17553385       0.2256864          0                   0   5.453017
## 50 0.18382257       0.2316164          0                   0   5.179104
## 51 0.19039649       0.2321418          0                   0   4.978847
## 52 0.19203209       0.2321418          0                   0   4.931155
## 53 0.19529388       0.2321418          0                   0   4.838433
## 54 0.20258723       0.2363518          0                   0   4.641917
## 55 0.21380775       0.2449071          0                   0   4.365767
## 56 0.22409237       0.2521039          0                   0   4.136926
## 57 0.24043647       0.2657456          0                   0   3.813467
## 58 0.26546094       0.2883455          0                   0   3.395168
## 59 0.28389531       0.3031425          0                   0   3.133971
## 60 0.30755650       0.3212753          0                   0   2.844280
## 61 0.31107607       0.3212753          0                   0   2.804920
## 62 0.38181153       0.3879698          0                   0   2.165889
## 63 0.41633879       0.4163388          0                   0   1.931480
##    Combined.Score       Genes
## 1       32.635986 PSMC4;GNAI3
## 2       72.289232       PSMC4
## 3       66.411133       GNAI3
## 4       64.630147     ATP6V1H
## 5       56.825976       GNAI3
## 6       55.455415       DTYMK
## 7       50.492824       GNAI3
## 8       41.701557       GNAI3
## 9       40.883745     ATP6V1H
## 10      36.492436       GNAI3
## 11      36.492436       GNAI3
## 12      35.200436     ATP6V1H
## 13      29.743651       GNAI3
## 14      29.275335       GNAI3
## 15      28.374333       GNAI3
## 16      27.518048     ATP6V1H
## 17      25.927561       GNAI3
## 18      24.831030       GNAI3
## 19      24.482281       GNAI3
## 20      24.141462       GNAI3
## 21      23.808321      SLC3A1
## 22      22.852586       GNAI3
## 23      21.118557       GNAI3
## 24      20.850327       GNAI3
## 25      20.850327       GNAI3
## 26      20.587577       GNAI3
## 27      20.587577       GNAI3
## 28      19.350725       GNAI3
```

```
## 29      19.350725       GNAI3
## 30      18.229604       GNAI3
## 31      17.406064     ATP6V1H
## 32      17.209501       GNAI3
## 33      16.640697       GNAI3
## 34      16.457769     ATP6V1H
## 35      15.757287       GNAI3
## 36      15.757287       GNAI3
## 37      15.262658      TRIM37
## 38      14.792346       PSMC4
## 39      14.058122       GNAI3
## 40      13.780529       GNAI3
## 41      13.511533     ATP6V1H
## 42      13.250766     ATP6V1H
## 43      13.250766       GNAI3
## 44      13.123359       GNAI3
## 45      11.733953       GNAI3
## 46      10.473360     ATP6V1H
## 47      10.298317       GNAI3
## 48       9.961639       GNAI3
## 49       9.487832       GNAI3
## 50       8.772286       PSMC4
## 51       8.258147       GNAI3
## 52       8.136863       GNAI3
## 53       7.902370       GNAI3
## 54       7.411214       GNAI3
## 55       6.734973       GNAI3
## 56       6.187588       RAB4A
## 57       5.435333       PSMC4
## 58       4.502969       PSMC4
## 59       3.946139     ATP6V1H
## 60       3.353681       PSMC4
## 61       3.275355       PSMC4
## 62       2.085379       PSMC4
## 63       1.692471       GNAI3
```

- Visualize the KEGG pathway enrichment results.
- Filter significant pathways ($p < 0.05$) and create a bar plot.

```r
# Filter significant pathways (p < 0.05)
sig_pathways <- kegg_results[kegg_results$Adjusted.P.value < 0.05, ]

ggplot(sig_pathways,
       aes(x = reorder(Term, -log10(Adjusted.P.value)),
              y = -log10(Adjusted.P.value))) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(title = "KEGG Pathway Enrichment",
       x = "Pathway",
       y = "-log10(Adjusted P-value)") +
  theme_minimal()
```

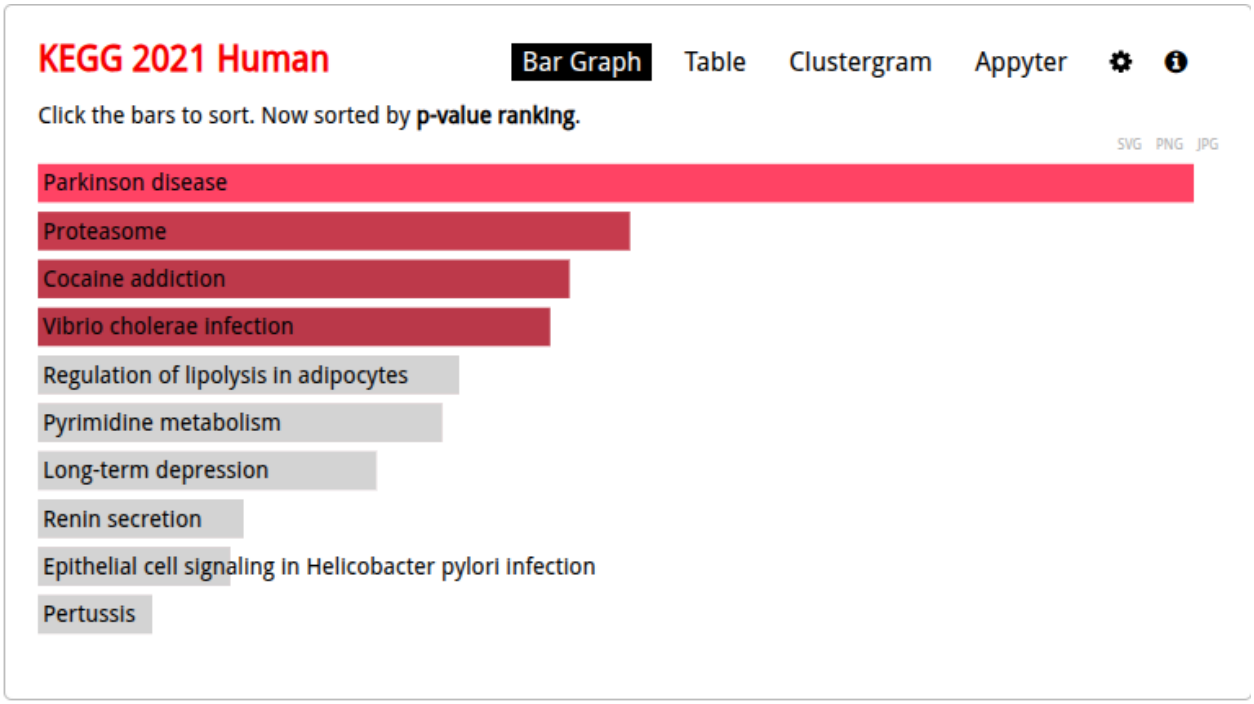
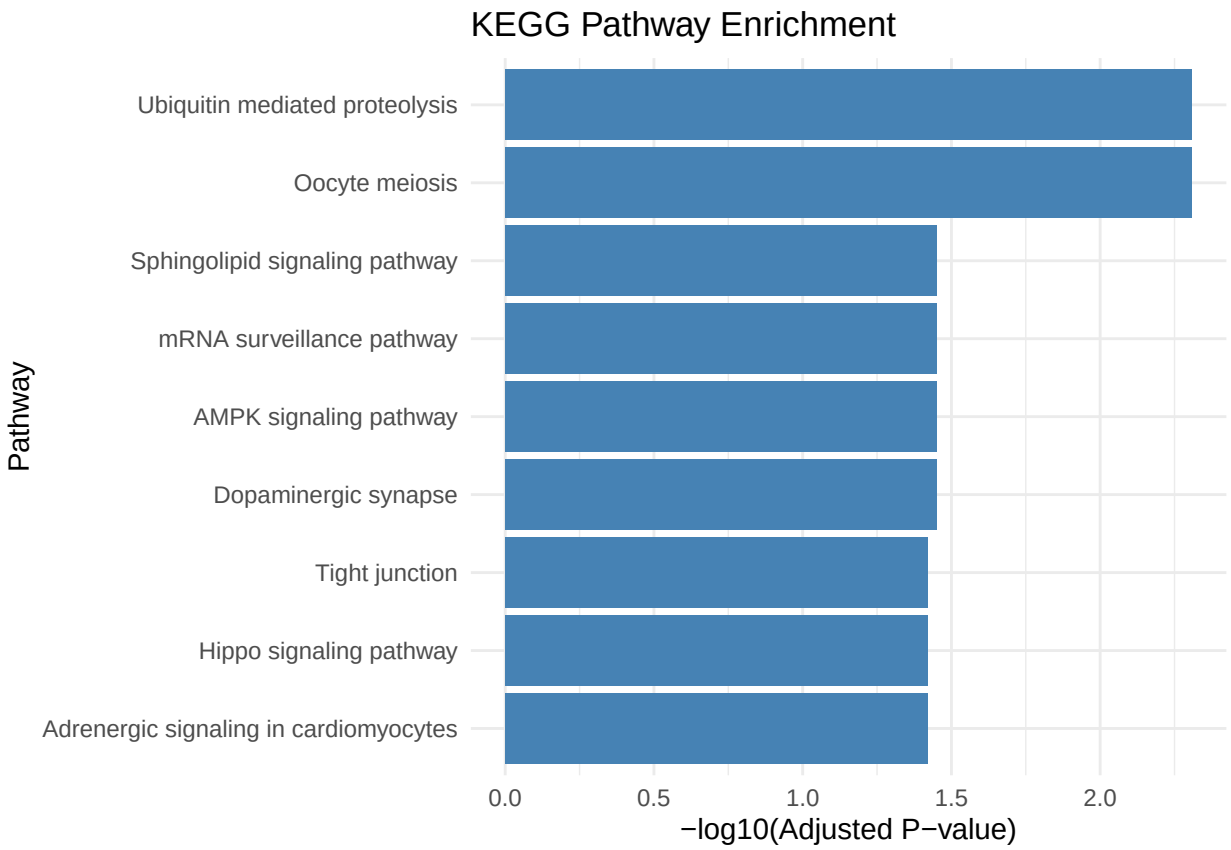

KEGG 2021 Human
Bar Graph
Table
Clustergram
Appyter
Click the bars to sort. Now sorted by p-value ranking.
SVG PNG JPG
Parkinson disease
Proteasome
Cocaine addiction
Vibrio cholerae infection
Regulation of lipolysis in adipocytes
Pyrimidine metabolism
Long-term depression
Renin secretion
Epithelial cell signaling in Helicobacter pylori infection
Pertussis

Figure 6: PC1


KEGG Pathway Enrichment
Ubiquitin mediated proteolysis
Oocyte meiosis
Sphingolipid signaling pathway
mRNA surveillance pathway
AMPK signaling pathway
Dopaminergic synapse
Tight junction
Hippo signaling pathway
Adrenergic signaling in cardiomyocytes
Pathway
0.0
0.5
1.0
1.5
2.0
−log10(Adjusted P−value)

KEGG 2021 Human Bar Graph **Table** Clustergram Appyter

Hover each row to see the overlapping genes.

100 entries per page Search:

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Parkinson disease	0.02531	0.2064	8.88	32.64
2	Proteasome	0.04503	0.2064	23.32	72.29
3	Cocaine addiction	0.04790	0.2064	21.86	66.41
4	Vibrio cholerae infection	0.04885	0.2064	21.41	64.63
5	Regulation of lipolysis in adipocytes	0.05361	0.2064	19.42	56.83
6	Pyrimidine metabolism	0.05456	0.2064	19.07	55.46
7	Long-term depression	0.05835	0.2064	17.77	50.49
8	Renin secretion	0.06682	0.2064	15.41	41.70
9	Epithelial cell signaling in Helicobacter pylori infection	0.06775	0.2064	15.19	40.88
10	Pertussis	0.07335	0.2064	13.97	36.49
11	Gastric acid secretion	0.07335	0.2064	13.97	36.49
12	Synaptic vesicle cycle	0.07521	0.2064	13.60	35.20
13	Gap junction	0.08446	0.2064	12.03	29.74
14	GABAergic synapse	0.08537	0.2064	11.90	29.28
15	Morphine addiction	0.08721	0.2064	11.63	28.37
16	Rheumatoid arthritis	0.08904	0.2064	11.38	27.52
17	Circadian entrainment	0.09270	0.2064	10.90	25.93
18	Progesterone-mediated oocyte maturation	0.09543	0.2064	10.57	24.83
19	Melanogenesis	0.09634	0.2064	10.46	24.48
20	Chagas disease	0.09725	0.2064	10.36	24.14
21	Protein digestion and absorption	0.09816	0.2064	10.26	23.81
22	Parathyroid hormone synthesis, secretion and action	0.1009	0.2064	9.96	22.85
23	Toxoplasmosis	0.1063	0.2064	9.42	21.12
24	Serotonergic synapse	0.1072	0.2064	9.34	20.85
25	Cholinergic synapse	0.1072	0.2064	9.34	20.85
26	Leukocyte transendothelial migration	0.1081	0.2064	9.25	20.59
27	Glutamatergic synapse	0.1081	0.2064	9.25	20.59
28	Sphingolipid signaling pathway	0.1126	0.2064	8.86	19.35
29	Growth hormone synthesis, secretion and action	0.1126	0.2064	8.86	19.35
30	Platelet activation	0.1170	0.2064	8.50	18.23
31	Lysosome	0.1206	0.2064	8.23	17.41
32	Relaxin signaling pathway	0.1214	0.2064	8.16	17.21
33	Dopaminergic synapse	0.1241	0.2064	7.97	16.64
34	Oxidative phosphorylation	0.1250	0.2064	7.91	16.46
35	Apelin signaling pathway	0.1285	0.2064	7.68	15.76

Figure 7: PC1

36	Estrogen signaling pathway	0.1285	0.2064	7.68	15.76
37	Ubiquitin mediated proteolysis	0.1311	0.2064	7.51	15.26
38	Spinocerebellar ataxia	0.1337	0.2064	7.35	14.79
39	Retrograde endocannabinoid signaling	0.1381	0.2064	7.10	14.06
40	Adrenergic signaling in cardiomyocytes	0.1398	0.2064	7.00	13.78
41	Phagosome	0.1416	0.2064	6.91	13.51
42	mTOR signaling pathway	0.1433	0.2064	6.82	13.25
43	Oxytocin signaling pathway	0.1433	0.2064	6.82	13.25
44	Cushing syndrome	0.1442	0.2064	6.78	13.12
45	cGMP-PKG signaling pathway	0.1545	0.2162	6.28	11.73
46	Tuberculosis	0.1655	0.2238	5.82	10.47
47	Axon guidance	0.1672	0.2238	5.76	10.30
48	Alcoholism	0.1705	0.2238	5.63	9.96
49	Chemokine signaling pathway	0.1755	0.2257	5.45	9.49
50	Epstein-Barr virus infection	0.1838	0.2316	5.18	8.77
51	Rap1 signaling pathway	0.1904	0.2321	4.98	8.26
52	Human immunodeficiency virus 1 infection	0.1920	0.2321	4.93	8.14
53	cAMP signaling pathway	0.1953	0.2321	4.84	7.90
54	Human cytomegalovirus infection	0.2026	0.2364	4.64	7.41
55	Chemical carcinogenesis	0.2138	0.2449	4.37	6.73
56	Endocytosis	0.2241	0.2521	4.14	6.19
57	Prion disease	0.2404	0.2657	3.81	5.44
58	Huntington disease	0.2655	0.2883	3.40	4.50
59	Human papillomavirus infection	0.2839	0.3031	3.13	3.95
60	Amyotrophic lateral sclerosis	0.3076	0.3213	2.84	3.35
61	Alzheimer disease	0.3111	0.3213	2.80	3.28
62	Pathways of neurodegeneration	0.3818	0.3880	2.17	2.09
63	Pathways in cancer	0.4163	0.4163	1.93	1.69

Showing 1 to 63 of 63 entries | Export entries to table

Previous Next

Terms marked with an * have an overlap of less than 5

Figure 8: PC1

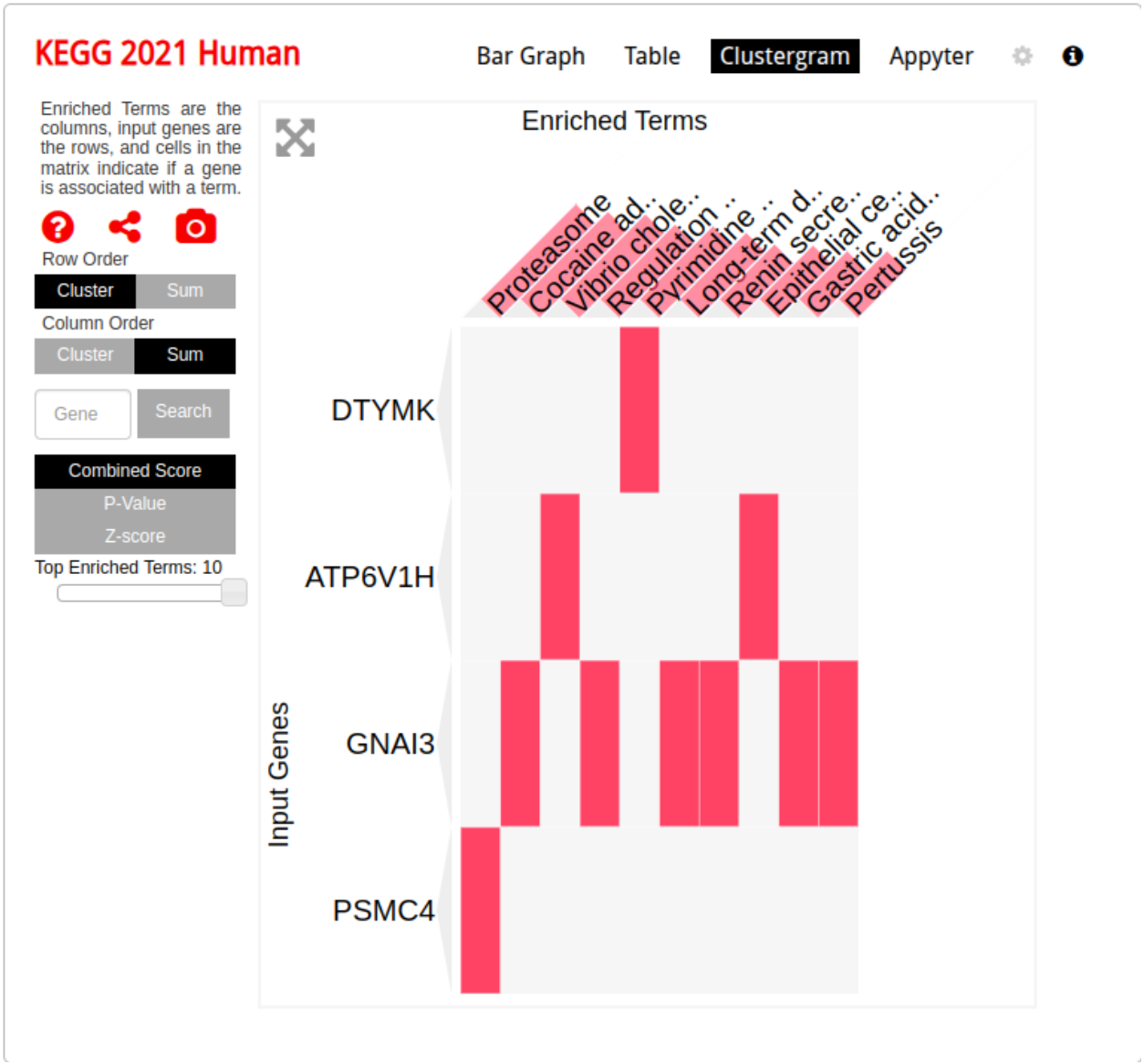
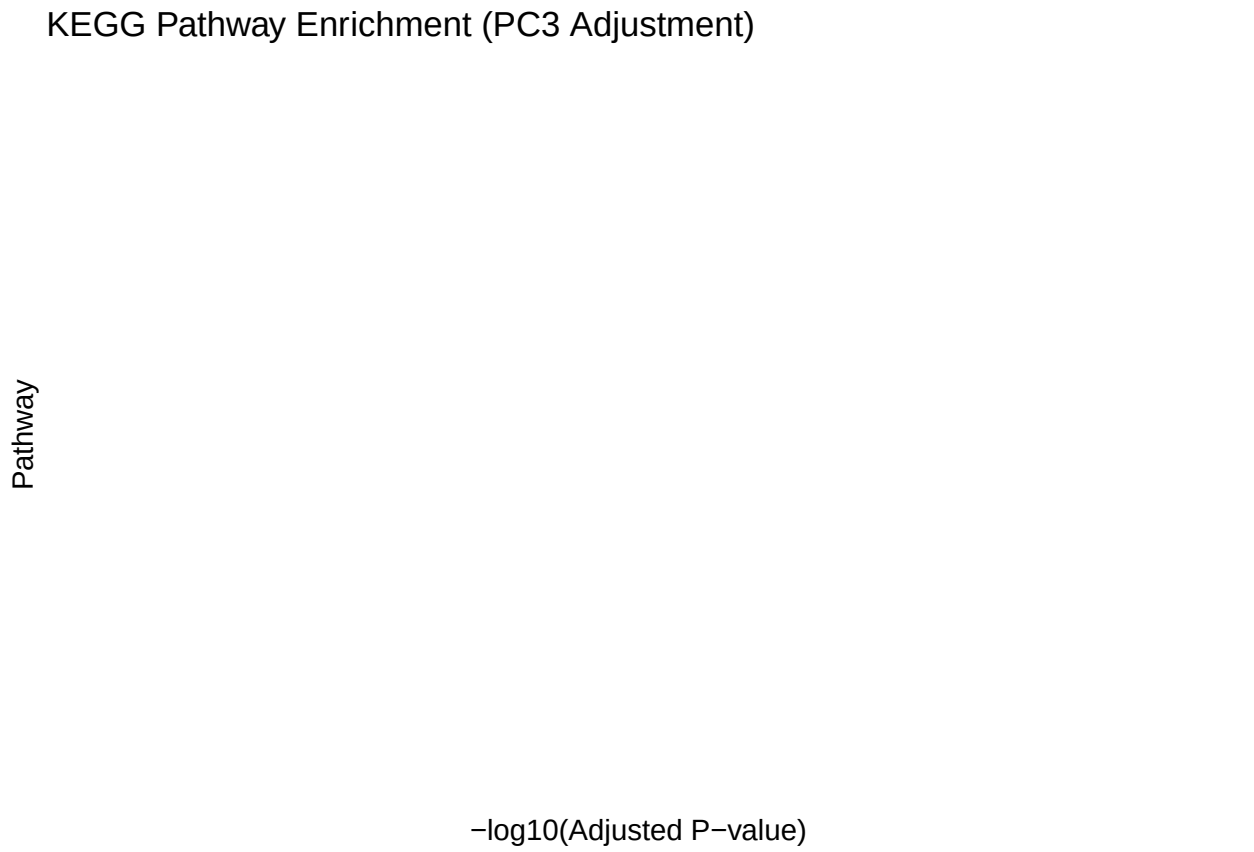

KEGG 2021 Human
Bar Graph
Table
Clustergram
Appyter
Enriched Terms are the columns, input genes are the rows, and cells in the matrix indicate if a gene is associated with a term.
Row Order
Cluster
Sum
Column Order
Cluster
Sum
Gene
Search
Combined Score
P-Value
Z-score
Top Enriched Terms: 10
Enriched Terms
Proteasome
Cocaine ad..
Vibrio chole..
Regulation ..
Pyrimidine ..
Long-term d..
Renin secre..
Epithelial ce..
Gastric acid..
Pertussis
DTYMK
ATP6V1H
GNAI3
PSMC4
Input Genes

Figure 9: PC1

```r
# Filter significant pathways (p < 0.05)
sig_pathways_pc1 <- kegg_results_pc1[kegg_results_pc1$Adjusted.P.value < 0.05, ]
ggplot(sig_pathways_pc1,
       aes(x = reorder(Term, -log10(Adjusted.P.value)),
              y = -log10(Adjusted.P.value))) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(title = "KEGG Pathway Enrichment (PC3 Adjustment)",
       x = "Pathway",
       y = "-log10(Adjusted P-value)") +
  theme_minimal()
```



Comments on the version with PC1

- Statistically, none of the pathways were significant after adjusting for multiple testing (all adjusted p-value > 0.05).
- Some pathways still show strong signals, with high odds ratios and low raw p-values, suggesting they might be biologically relevant.
- If we're doing exploratory analysis, it might make sense to relax the threshold a bit (looking at adjusted p-values below 0.15) to catch pathways that could still be worth investigating.