Objective: Find interesting genes for melanoma, CNS, and Leukemia in the NCI60 data set in ISLR. Use a false discovery rate (FDR) of 0.2.


Finding Interesting Genes

To find interesting genes for the three conditions in the NCI60 data set in ISLR, I first calculated the p-values for each gene for each of the three conditions by using t-tests.

```
calculate_p_values <- function(gene_index, cancer_indices) {
  cancer_gene_expression <- gene_data[cancer_indices, gene_index]
  other_gene_expression <- gene_data[-cancer_indices, gene_index]
  test_result <- t.test(cancer_gene_expression, other_gene_expression)
  return(test_result$p.value)
}
```

```
# Melanoma
p_values_melanoma <- sapply(1:ncol(gene_data), function(i) calculate_p_values(i, melanoma_indices))
interesting_genes_melanoma <- fdr(p_values_melanoma, 0.2)

# CNS
p_values_cns <- sapply(1:ncol(gene_data), function(i) calculate_p_values(i, cns_indices))
interesting_genes_cns <- fdr(p_values_cns, 0.2)

# Leukemia
p_values_leukemia <- sapply(1:ncol(gene_data), function(i) calculate_p_values(i, leukemia_indices))
interesting_genes_leukemia <- fdr(p_values_leukemia, 0.2)
```

The output of the fdr function given for the assignment is shown below.
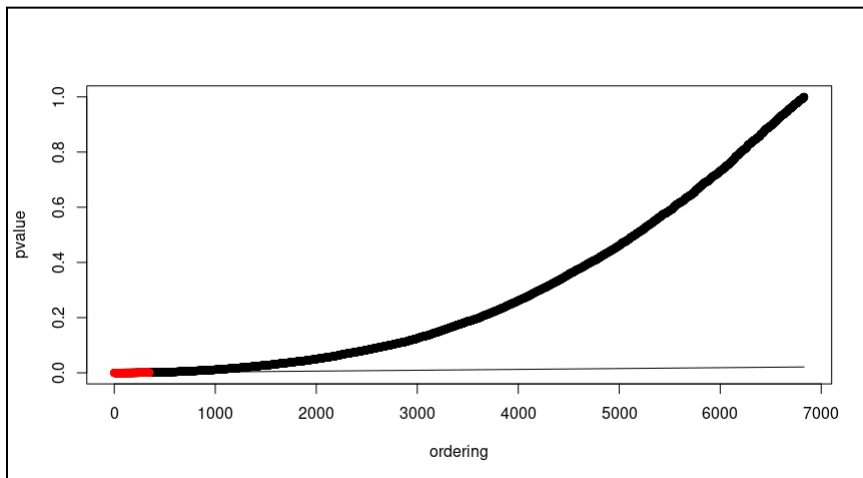
Melanoma:
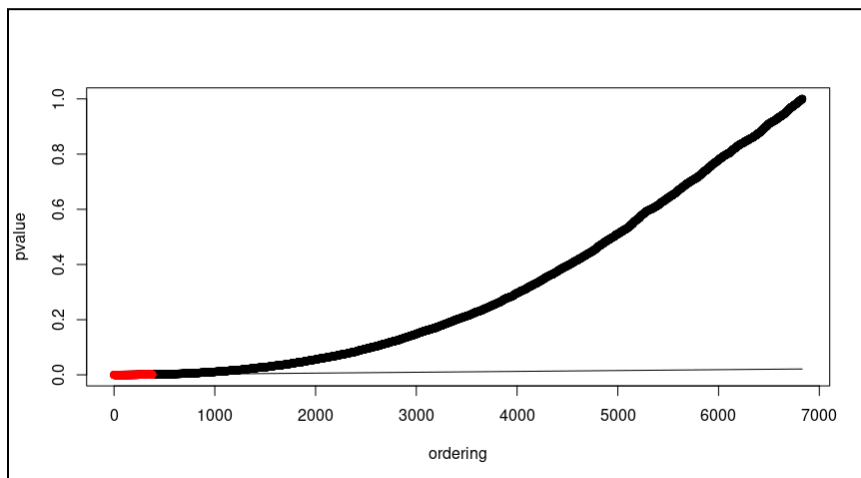
CNS:
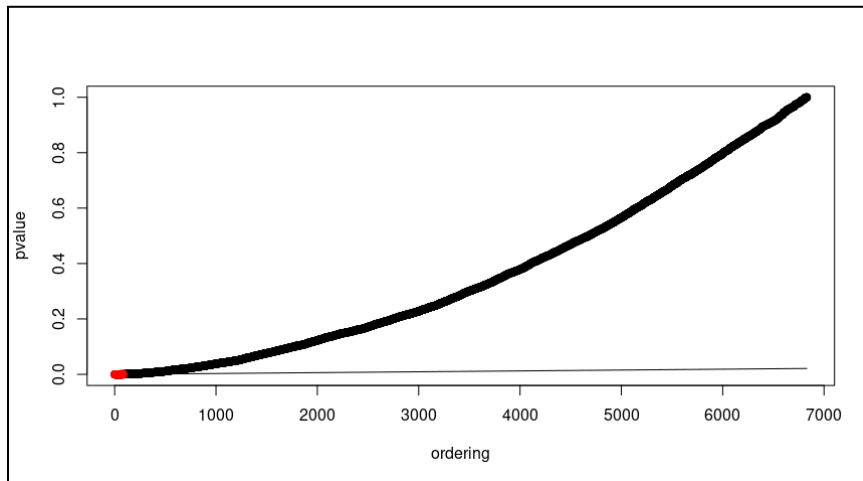
Leukemia:

Next, I used ggplot to plot the p-values and FDR threshold and set the FDR for whether a gene was interesting or not to 0.2 (which was asked for by the assignment directions). I wrote the following function that used ggplot to plot any given conditions' gene array, with the indices on the x axis and their p values on the y axis. This way, we can see how many genes are below the threshold set by the FDR, which will be previewed as a red dotted line.

```
plot_genes <- function(p_values, interesting_genes, cancer_type) {
  data_frame <- data.frame(Gene = 1:length(p_values), P_Value = p_values)
  data_frame$Status <- ifelse(data_frame$Gene %in% interesting_genes, "Interesting", "Not Interesting")

  ggplot(data_frame, aes(x = Gene, y = P_Value, color = Status)) +
    geom_point(alpha = 0.5) +
    geom_hline(yintercept = 0.2, linetype = "dashed", color = "red") +
    scale_y_log10() +
    labs(title = paste(cancer_type),
         x = "Gene Index",
         y = "P-Value (log scale)",
         color = "Gene Status") +
    theme_minimal() +
    theme(legend.position = "top")
}
```

```
plot_genes(p_values_melanoma, interesting_genes_melanoma, "Melanoma")
plot_genes(p_values_cns, interesting_genes_cns, "CNS")
plot_genes(p_values_leukemia, interesting_genes_leukemia, "Leukemia")
```

The output of the plots are as follows:

P-Values and FDR for CNS

P-Values and FDR for Leukemia
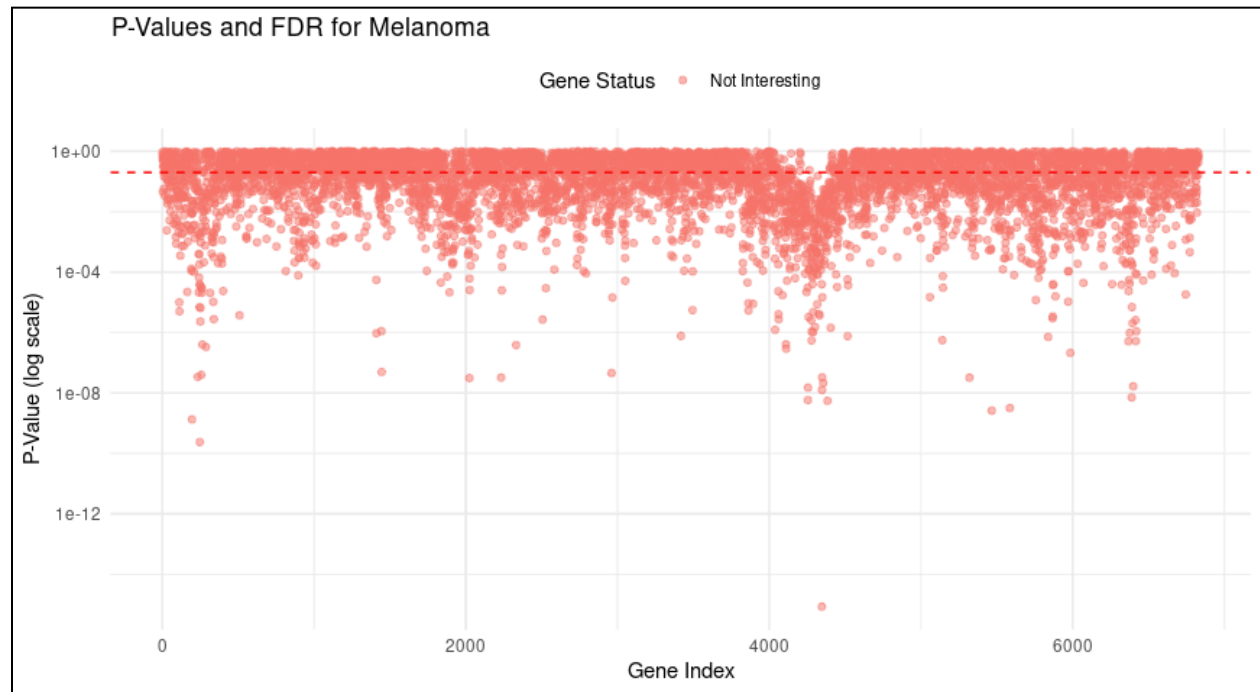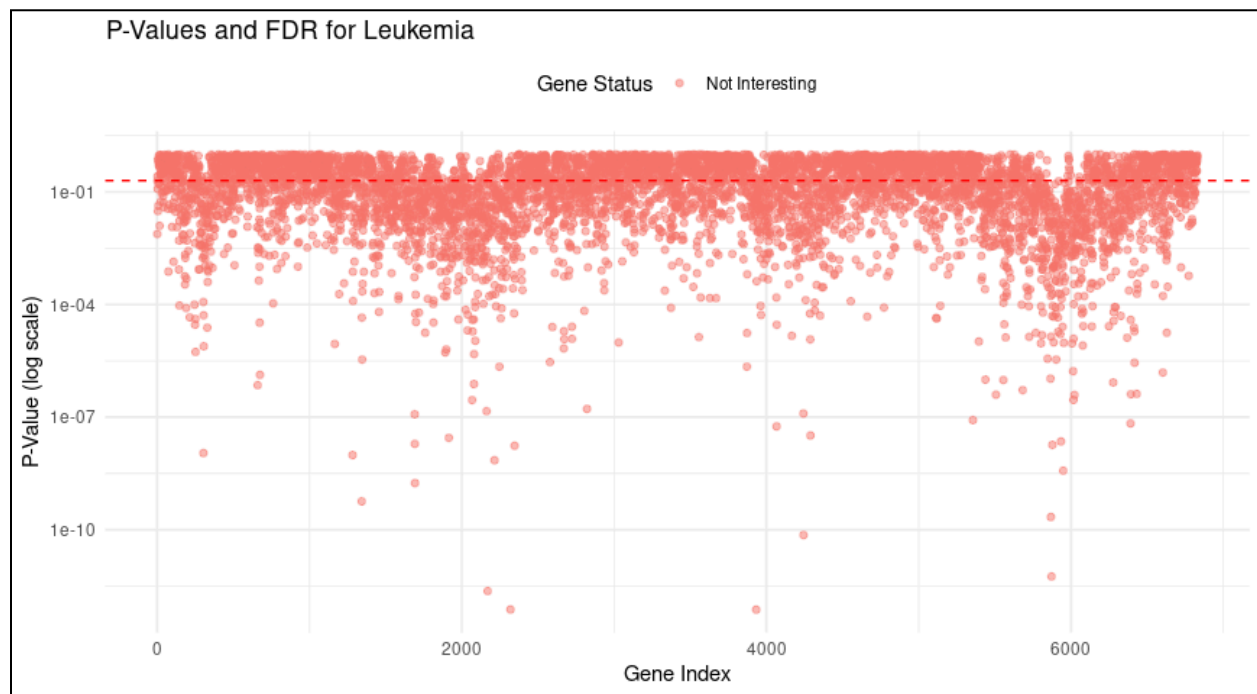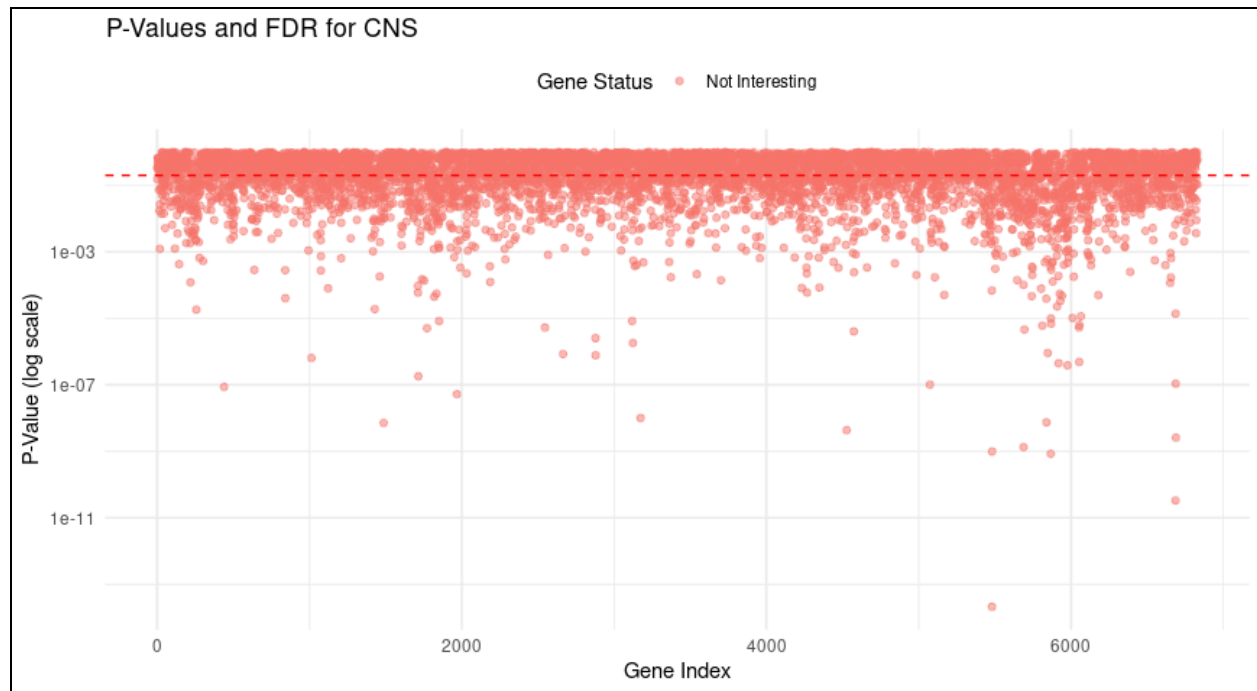
Are genes from the same cell statistically independent?

For all three of these plots, most of the genes are above the 0.2 FDR threshold which means that they are not significantly associated with their condition. Across the three conditions, if we were to find common genes, we would look for genes that appear below the threshold in multiple graphs. The existence of such genes could mean that there are common trends in genes present between the conditions.

To find if any of the genes have statistical inference, we would need to find the correlation between genes within the same cell type. However, independence is really rare, since genes work together in complex networks.

Common Interesting Genes Between Any of the Two Cancers

To determine if there are any common interesting genes between any two cancers, I analyzed the lists of genes that fall below the FDR threshold for each condition and find any overlaps. To do this, I made lists for each of the interesting genes for each condition. For my own observation, I also found the percentage of genes for each condition that was marked as interesting based on the FDR rate (percentage of interesting genes out of all genes for that condition).

```
interesting_genes_melanoma <- which(p_melanoma < 0.2)
interesting_genes_cns <- which(p_cns < 0.2)
interesting_genes_leukemia <- which(p_leukemia < 0.2)

percent_interesting_melanoma <- length(interesting_genes_melanoma) / ncol(gene_data) * 100
percent_interesting_cns <- length(interesting_genes_cns) / ncol(gene_data) * 100
percent_interesting_leukemia <- length(interesting_genes_leukemia) / ncol(gene_data) * 100
```

The output was as follows:

```
Percent for Melanoma: 30.497803806735
Percent for CNS: 9.92679355783309
Percent for Leukemia: 32.913616398243
```

Next, I used intersect(x, y) to compare all the combinations of the conditions' datasets to find similarities.

```
common_genes_melanoma_cns <- intersect(interesting_genes_melanoma, interesting_genes_cns)
common_genes_melanoma_leukemia <- intersect(interesting_genes_melanoma, interesting_genes_leukemia)
common_genes_cns_leukemia <- intersect(interesting_genes_cns, interesting_genes_leukemia)

overlap_melanoma_cns <- length(common_genes_melanoma_cns) > 0
overlap_melanoma_leukemia <- length(common_genes_melanoma_leukemia) > 0
overlap_cns_leukemia <- length(common_genes_cns_leukemia) > 0
```

I found out that there were a substantial amount of gene objects shared between the lists. They are seen on the below.

Melanoma and CNS:

```
[1] "Common genes between Melanoma and CNS:"
  [1]   15  117  130  134  152  189  191  202  219  224  239  242  243  245  249  251  252  256  257  260  273  328  355  752  984 1013 1019 1122 1271 1429
 [31] 1430 1462 1723 1726 1730 1834 1837 1873 1877 1923 1924 1968 1989 1990 2026 2031 2032 2033 2039 2237 2241 2257 2286 2345 2371 2510 2512 2521 2525 2541
 [61] 2542 2774 2915 2916 2955 3033 3046 3072 3130 3148 3261 3366 3372 3373 3438 3496 3506 3590 3619 3670 3863 3864 3922 3933 3934 3955 3956 3957 4037 4101
 [91] 4115 4146 4173 4186 4220 4221 4227 4229 4232 4240 4262 4263 4265 4279 4283 4293 4312 4313 4322 4329 4343 4344 4346 4348 4367 4379 4381 4382 4426 4427
[121] 4472 4522 4527 4574 4586 4587 4642 4709 4726 4844 4897 5093 5106 5149 5308 5336 5337 5346 5431 5432 5461 5465 5470 5472 5508 5521 5522 5523 5591 5614
[151] 5619 5648 5665 5667 5688 5689 5694 5695 5696 5716 5729 5730 5734 5738 5742 5743 5744 5746 5750 5759 5804 5805 5808 5835 5836 5837 5839 5846 5866 5867
[181] 5868 5869 5870 5871 5872 5885 5899 5930 5941 5957 5961 5970 5971 5972 5973 5975 5978 5984 5985 6007 6053 6054 6055 6108 6109 6110 6111 6114 6140 6288
[211] 6289 6353 6354 6356 6357 6358 6388 6537 6540 6742 6743 6772 6819
```

Melanoma and Leukemia:

```
[1] "Common genes between Melanoma and Leukemia:"
  [1]   74   78  106  119  130  142  143  152  158  159  165  167  181  187  190  191  193  195  196  198  214  224  245  246  248  249  251  252  269  273
 [31]  278  281  283  285  287  288  293  295  297  307  315  316  321  325  326  327  332  333  335  336  338  342  348  349  373  391  392  424  438  461
 [61]  489  660  668  673  687  734  761  762  804  814  828  897  926  927  930  986 1016 1107 1128 1157 1192 1211 1223 1247 1271 1283 1293 1297 1346 1347
 [91] 1430 1431 1435 1442 1443 1444 1459 1460 1462 1499 1625 1629 1644 1648 1654 1657 1658 1660 1667 1691 1692 1693 1694 1695 1696 1697 1699 1701 1702 1704
[121] 1707 1708 1709 1710 1723 1728 1729 1730 1732 1733 1734 1735 1737 1738 1744 1745 1746 1762 1767 1793 1794 1809 1810 1811 1812 1832 1841 1844 1858 1859
[151] 1860 1861 1862 1867 1868 1869 1870 1871 1873 1878 1881 1882 1884 1885 1886 1887 1890 1891 1892 1893 1894 1895 1896 1897 1898 1899 1900 1901 1902 1904
[181] 1909 1911 1912 1913 1914 1915 1918 1919 1920 1923 1924 1927 1929 1934 1936 1940 1946 1947 1956 1958 1959 1961 1968 1976 1980 1981 1988 2010 2015 2022
[211] 2023 2024 2025 2029 2030 2032 2033 2034 2036 2039 2042 2053 2058 2078 2080 2084 2085 2114 2126 2142 2163 2164 2171 2180 2182 2195 2211 2215 2221 2222
[241] 2233 2235 2236 2237 2239 2241 2246 2247 2257 2262 2266 2270 2284 2286 2287 2301 2303 2312 2320 2321 2326 2333 2334 2335 2336 2337 2338 2339 2340 2341
[271] 2343 2345 2346 2347 2348 2367 2368 2394 2395 2431 2501 2503 2506 2532 2534 2559 2581 2583 2590 2598 2599 2600 2601 2604 2605 2606 2607 2670 2716 2723
[301] 2724 2726 2732 2733 2742 2756 2772 2773 2807 2843 2855 2918 2925 2939 2941 2942 2943 2955 2961 3030 3035 3042 3054 3060 3083 3130 3188 3189 3233 3344
[331] 3372 3373 3377 3408 3444 3445 3449 3456 3458 3459 3464 3511 3519 3520 3562 3568 3602 3606 3607 3670 3761 3841 3842 3854 3860 3872 3873 3874 3875 3895
[361] 3900 3902 3903 3920 3923 3933 3935 3946 3952 3953 3956 3957 3959 3960 3962 3963 3964 3965 3970 3971 3976 3977 4004 4005 4009 4011 4013 4016 4028 4037
[391] 4042 4056 4057 4066 4067 4072 4079 4095 4098 4099 4100 4101 4103 4106 4107 4109 4111 4112 4139 4142 4143 4146 4153 4154 4156 4158 4163 4164 4166 4167
[421] 4168 4191 4192 4193 4195 4198 4201 4214 4215 4216 4217 4218 4234 4235 4238 4241 4242 4243 4244 4245 4246 4247 4248 4251 4253 4254 4257 4258 4263 4264
[451] 4274 4277 4278 4282 4284 4285 4286 4287 4288 4289 4290 4291 4292 4293 4294 4299 4300 4302 4303 4308 4310 4311 4315 4316 4317 4318 4330 4331 4337 4338
[481] 4339 4340 4341 4342 4343 4345 4349 4350 4351 4352 4353 4354 4359 4360 4362 4363 4364 4365 4373 4374 4379 4380 4381 4387 4388 4389 4413 4420 4422 4432
[511] 4444 4445 4446 4447 4448 4451 4454 4455 4479 4480 4532 4533 4586 4587 4588 4591 4620 4655 4726 4747 4839 4846 5047 5099 5104 5108 5136 5142 5214 5346
[541] 5356 5360 5406 5411 5417 5419 5420 5431 5432 5437 5439 5445 5449 5450 5451 5452 5453 5463 5465 5470 5487 5491 5492 5508 5524 5526 5537 5555 5556 5557
[571] 5558 5559 5563 5566 5567 5578 5579 5582 5583 5584 5586 5589 5595 5614 5660 5661 5662 5663 5667 5686 5687 5688 5689 5695 5718 5724 5728 5739 5742 5743
[601] 5744 5745 5746 5755 5758 5759 5760 5761 5762 5769 5771 5779 5785 5794 5802 5804 5805 5806 5807 5808 5809 5810 5816 5824 5825 5837 5839 5841 5846 5863
[631] 5866 5867 5868 5869 5871 5872 5876 5877 5878 5882 5883 5884 5885 5886 5887 5888 5899 5900 5930 5931 5932 5941 5944 5945 5949 5961 5965 5970 5971 5972
[661] 5973 5976 5984 6003 6004 6005 6006 6007 6015 6016 6024 6031 6033 6039 6051 6053 6054 6055 6058 6059 6061 6066 6068 6070 6071 6072 6080 6081 6124 6125
[691] 6126 6138 6140 6141 6142 6182 6189 6190 6192 6196 6198 6208 6209 6217 6221 6237 6238 6245 6254 6259 6262 6275 6281 6288 6289 6291 6294 6297 6312 6322
[721] 6324 6328 6329 6332 6351 6355 6356 6357 6369 6370 6371 6380 6381 6382 6383 6384 6385 6388 6389 6390 6391 6392 6393 6394 6398 6399 6414 6415 6416 6418
[751] 6432 6433 6475 6505 6520 6534 6535 6536 6554 6564 6582 6585 6598 6604 6617 6627 6647 6695 6720 6773 6774 6796 6810
```

CNS and Leukemia:

```
[1] "Common genes between CNS and Leukemia:"
  [1]   98  130  144  152  191  210  224  225  245  249  251  252  267  273  279  301  439  465  481  483  497  669  674  786  839  841  842 1144 1145 1159
 [31] 1195 1208 1209 1271 1417 1423 1424 1426 1430 1448 1461 1462 1488 1533 1603 1604 1605 1649 1653 1703 1711 1712 1715 1719 1723 1730 1736 1739 1753 1778
 [61] 1815 1818 1822 1828 1835 1847 1849 1850 1851 1852 1866 1873 1923 1924 1926 1928 1939 1968 1970 1982 2032 2033 2039 2068 2081 2083 2087 2090 2157 2184
 [91] 2187 2190 2237 2241 2257 2261 2282 2283 2286 2291 2295 2345 2422 2643 2665 2675 2676 2812 2878 2879 2955 3066 3127 3130 3297 3372 3373 3501 3536 3585
[121] 3670 3933 3956 3957 3968 4002 4003 4024 4037 4101 4146 4147 4263 4267 4293 4343 4379 4381 4586 4587 4648 4726 4845 4847 4876 4976 5081 5086 5139 5165
[151] 5346 5368 5429 5431 5432 5433 5434 5444 5464 5465 5470 5477 5478 5481 5482 5508 5534 5536 5614 5667 5688 5689 5691 5693 5695 5701 5725 5732 5742 5743
[181] 5744 5746 5748 5759 5788 5803 5804 5805 5808 5811 5812 5814 5837 5839 5845 5846 5847 5850 5851 5859 5866 5867 5868 5869 5871 5872 5874 5880 5881 5885
[211] 5892 5898 5899 5908 5909 5910 5911 5914 5915 5916 5917 5918 5919 5925 5926 5929 5930 5936 5937 5938 5939 5941 5946 5947 5953 5960 5961 5966 5967 5968
[241] 5970 5971 5972 5973 5977 5980 5984 6007 6010 6011 6017 6018 6020 6053 6054 6055 6060 6065 6067 6074 6094 6123 6131 6132 6140 6179 6252 6253 6273 6274
[271] 6288 6289 6356 6357 6388 6550
```

Based on these results, it is seen that Melanoma and Leukemia share the highest amount of common interesting genes.