Module2 Quiz

mindan

2021/10/9

```
rm(list=ls())
```

Dependencies

```
  library(devtools)
```

```
## Loading required package: usethis
```

```
  library(Biobase)
```

```
## Loading required package: BiocGenerics
```

```
## Loading required package: parallel
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which.max, which.min
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
  library(broom)
  library(limma)
```

```
##
## Attaching package: 'limma'

## The following object is masked from 'package:BiocGenerics':
##
##     plotMA
```

```
  library(sva)
```

```
## Loading required package: mgcv

## Loading required package: nlme

## This is mgcv 1.8-37. For overview type 'help("mgcv-package")'.

## Loading required package: genefilter

## Loading required package: BiocParallel
```

Load the Montgomery and Pickrell eSet:

```
con =url("http://bowtie-bio.sourceforge.net/recount/ExpressionSets/montpick_eset.RData")
load(file=con)
close(con)
mp = montpick.eset
pdata=pData(mp)
edata=as.data.frame(exprs(mp))
fdata = fData(mp)
ls()
```

```
## [1] "con"           "edata"         "fdata"         "montpick.eset"
## [5] "mp"            "pdata"
```

Questions and Answers

1.What percentage of variation is explained by the 1st principal component in the data set if you:

- Do no transformations?
- log2(data + 1) transform?
- log2(data + 1) transform and subtract row means?

```r
edata1 = edata
edata2 = log2(edata + 1)
edata3 = edata2 - rowMeans(edata2)

pc1 = prcomp(edata1,center=F, scale=F)
pc2 = prcomp(edata2,center=F, scale=F)
pc3 = prcomp(edata3,center=F, scale=F)

summary(pc1)
```

```
## Importance of components:
##                             PC1       PC2       PC3       PC4       PC5
## Standard deviation     3873.7493 810.62834 536.45091 449.61345 349.43577
## Proportion of Variance    0.8873   0.03886   0.01702   0.01195   0.00722
## Cumulative Proportion     0.8873   0.92620   0.94322   0.95517   0.96239
##                              PC6       PC7       PC8       PC9      PC10      PC11
## Standard deviation     2.80e+02 2.59e+02 241.92322 212.31559 197.3879 174.05075
## Proportion of Variance 4.63e-03 3.97e-03   0.00346   0.00267   0.0023   0.00179
## Cumulative Proportion  9.67e-01 9.71e-01   0.97445   0.97712   0.9794   0.98121
##                             PC12      PC13      PC14      PC15      PC16
## Standard deviation     152.17269 149.73198 139.46644 134.77771 124.57325
## Proportion of Variance   0.00137   0.00133   0.00115   0.00107   0.00092
## Cumulative Proportion    0.98258   0.98391   0.98506   0.98613   0.98705
##                             PC17      PC18     PC19      PC20      PC21
## Standard deviation     121.48271 114.90131 108.9518 103.17290 100.24950
## Proportion of Variance   0.00087   0.00078   0.0007   0.00063   0.00059
## Cumulative Proportion    0.98792   0.98870   0.9894   0.99004   0.99063
##                             PC22     PC23     PC24     PC25     PC26     PC27
## Standard deviation     94.07503 90.56096 84.88036 83.21744 80.09393 77.62831
## Proportion of Variance  0.00052  0.00048  0.00043  0.00041  0.00038  0.00036
## Cumulative Proportion   0.99115  0.99164  0.99206  0.99247  0.99285  0.99321
##                             PC28    PC29     PC30     PC31     PC32     PC33
## Standard deviation     74.81886 70.6504 69.82351 66.44161 65.43414 61.88733
## Proportion of Variance  0.00033  0.0003  0.00029  0.00026  0.00025  0.00023
## Cumulative Proportion   0.99354  0.9938  0.99412  0.99438  0.99464  0.99486
##                             PC34    PC35    PC36     PC37     PC38     PC39
## Standard deviation     60.25671 58.8526 58.6074 55.82487 55.08763 54.39750
## Proportion of Variance  0.00021  0.0002  0.0002  0.00018  0.00018  0.00017
## Cumulative Proportion   0.99508  0.9953  0.9955  0.99567  0.99585  0.99603
##                             PC40     PC41     PC42     PC43     PC44     PC45
## Standard deviation     51.72409 51.39219 50.05970 48.58951 46.76889 46.26263
## Proportion of Variance  0.00016  0.00016  0.00015  0.00014  0.00013  0.00013
## Cumulative Proportion   0.99618  0.99634  0.99649  0.99663  0.99676  0.99688
##                             PC46     PC47     PC48    PC49    PC50    PC51
## Standard deviation     45.00010 44.76527 42.34228 41.7974 41.4754 40.5577
## Proportion of Variance  0.00012  0.00012  0.00011  0.0001  0.0001  0.0001
## Cumulative Proportion   0.99700  0.99712  0.99723  0.9973  0.9974  0.9975
##                             PC52     PC53     PC54     PC55     PC56     PC57
## Standard deviation     39.67859 39.05412 38.26626 36.98885 36.25703 35.49016
## Proportion of Variance  0.00009  0.00009  0.00009  0.00008  0.00008  0.00007
## Cumulative Proportion   0.99762  0.99771  0.99780  0.99788  0.99796  0.99803
##                             PC58     PC59     PC60     PC61     PC62     PC63
## Standard deviation     35.11572 34.48071 33.54100 33.07237 32.83210 32.15031
```

```
## Proportion of Variance  0.00007  0.00007  0.00007  0.00006  0.00006  0.00006
## Cumulative Proportion   0.99811  0.99818  0.99824  0.99831  0.99837  0.99843
##                            PC64     PC65     PC66     PC67     PC68     PC69
## Standard deviation     31.60368 31.17992 30.80950 30.19975 29.50094 28.73286
## Proportion of Variance  0.00006  0.00006  0.00006  0.00005  0.00005  0.00005
## Cumulative Proportion   0.99849  0.99855  0.99861  0.99866  0.99871  0.99876
##                            PC70     PC71     PC72     PC73     PC74     PC75
## Standard deviation     28.70925 28.20686 27.70565 27.59380 26.54453 26.29041
## Proportion of Variance  0.00005  0.00005  0.00005  0.00005  0.00004  0.00004
## Cumulative Proportion   0.99881  0.99886  0.99890  0.99895  0.99899  0.99903
##                            PC76     PC77     PC78     PC79     PC80     PC81
## Standard deviation     25.97327 25.47720 25.16383 24.18564 24.04465 23.59967
## Proportion of Variance  0.00004  0.00004  0.00004  0.00003  0.00003  0.00003
## Cumulative Proportion   0.99907  0.99911  0.99914  0.99918  0.99921  0.99925
##                            PC82     PC83     PC84     PC85     PC86     PC87
## Standard deviation     23.54567 23.03415 22.53007 22.08676 21.96347 21.64444
## Proportion of Variance  0.00003  0.00003  0.00003  0.00003  0.00003  0.00003
## Cumulative Proportion   0.99928  0.99931  0.99934  0.99937  0.99940  0.99943
##                            PC88     PC89     PC90     PC91     PC92     PC93
## Standard deviation     20.99670 20.68782 20.49515 20.20433 19.98489 19.79378
## Proportion of Variance  0.00003  0.00003  0.00002  0.00002  0.00002  0.00002
## Cumulative Proportion   0.99945  0.99948  0.99950  0.99953  0.99955  0.99957
##                            PC94     PC95     PC96     PC97     PC98     PC99
## Standard deviation     19.36796 18.98478 18.48887 18.25852 17.96342 17.71434
## Proportion of Variance  0.00002  0.00002  0.00002  0.00002  0.00002  0.00002
## Cumulative Proportion   0.99959  0.99962  0.99964  0.99966  0.99967  0.99969
##                           PC100    PC101    PC102    PC103    PC104    PC105
## Standard deviation     17.61242 17.30981 17.13379 16.79936 16.39976 15.92666
## Proportion of Variance  0.00002  0.00002  0.00002  0.00002  0.00002  0.00001
## Cumulative Proportion   0.99971  0.99973  0.99975  0.99976  0.99978  0.99979
##                           PC106    PC107    PC108    PC109    PC110    PC111
## Standard deviation     15.39662 15.11251 14.84553 14.69385 14.21491 14.15865
## Proportion of Variance  0.00001  0.00001  0.00001  0.00001  0.00001  0.00001
## Cumulative Proportion   0.99981  0.99982  0.99983  0.99985  0.99986  0.99987
##                           PC112    PC113    PC114    PC115    PC116    PC117
## Standard deviation     13.56818 13.29272 12.88818 12.82644 12.45120 12.15145
## Proportion of Variance  0.00001  0.00001  0.00001  0.00001  0.00001  0.00001
## Cumulative Proportion   0.99988  0.99989  0.99990  0.99991  0.99992  0.99993
##                           PC118    PC119    PC120    PC121    PC122   PC123
## Standard deviation     11.96481 11.37602 11.07883 11.02159 10.26807 9.99854
## Proportion of Variance  0.00001  0.00001  0.00001  0.00001  0.00001 0.00001
## Cumulative Proportion   0.99994  0.99995  0.99995  0.99996  0.99997 0.99997
##                          PC124 PC125 PC126 PC127 PC128 PC129
## Standard deviation     9.43563 9.148 9.085 8.751 8.365 7.538
## Proportion of Variance 0.00001 0.000 0.000 0.000 0.000 0.000
## Cumulative Proportion  0.99998 1.000 1.000 1.000 1.000 1.000
```

```
summary(pc2)
```

```
## Importance of components:
##                            PC1     PC2     PC3    PC4     PC5     PC6     PC7
## Standard deviation     26.5395 1.75044 1.06819 1.0057 0.87056 0.80102 0.77007
## Proportion of Variance  0.9738 0.00424 0.00158 0.0014 0.00105 0.00089 0.00082
## Cumulative Proportion   0.9738 0.97801 0.97959 0.9810 0.98204 0.98293 0.98374
```

```
##                             PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.70552 0.65430 0.61274 0.59750 0.57926 0.53438 0.51586
## Proportion of Variance 0.00069 0.00059 0.00052 0.00049 0.00046 0.00039 0.00037
## Cumulative Proportion  0.98443 0.98502 0.98554 0.98604 0.98650 0.98690 0.98726
##                            PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.48309 0.47055 0.45301 0.44017 0.41963 0.41087 0.40881
## Proportion of Variance 0.00032 0.00031 0.00028 0.00027 0.00024 0.00023 0.00023
## Cumulative Proportion  0.98759 0.98789 0.98818 0.98844 0.98869 0.98892 0.98915
##                            PC22    PC23    PC24   PC25   PC26    PC27    PC28
## Standard deviation     0.39820 0.39333 0.39274 0.3838 0.3776 0.36627 0.35970
## Proportion of Variance 0.00022 0.00021 0.00021 0.0002 0.0002 0.00019 0.00018
## Cumulative Proportion  0.98937 0.98959 0.98980 0.9900 0.9902 0.99039 0.99056
##                            PC29    PC30    PC31    PC32    PC33    PC34    PC35
## Standard deviation     0.35097 0.34734 0.34341 0.34123 0.33557 0.33375 0.33073
## Proportion of Variance 0.00017 0.00017 0.00016 0.00016 0.00016 0.00015 0.00015
## Cumulative Proportion  0.99073 0.99090 0.99106 0.99123 0.99138 0.99153 0.99169
##                            PC36    PC37    PC38    PC39    PC40    PC41    PC42
## Standard deviation     0.32379 0.32220 0.31797 0.31652 0.31350 0.31305 0.30851
## Proportion of Variance 0.00014 0.00014 0.00014 0.00014 0.00014 0.00014 0.00013
## Cumulative Proportion  0.99183 0.99197 0.99211 0.99225 0.99239 0.99252 0.99266
##                            PC43    PC44    PC45    PC46    PC47    PC48    PC49
## Standard deviation     0.30601 0.30277 0.29895 0.29605 0.29523 0.29301 0.29179
## Proportion of Variance 0.00013 0.00013 0.00012 0.00012 0.00012 0.00012 0.00012
## Cumulative Proportion  0.99279 0.99291 0.99304 0.99316 0.99328 0.99340 0.99351
##                            PC50    PC51    PC52    PC53    PC54    PC55    PC56
## Standard deviation     0.28679 0.28357 0.28229 0.28166 0.28041 0.27831 0.27795
## Proportion of Variance 0.00011 0.00011 0.00011 0.00011 0.00011 0.00011 0.00011
## Cumulative Proportion  0.99363 0.99374 0.99385 0.99396 0.99407 0.99417 0.99428
##                           PC57   PC58   PC59   PC60   PC61   PC62   PC63   PC64
## Standard deviation     0.2751 0.2734 0.2722 0.2708 0.2696 0.2669 0.2663 0.2654
## Proportion of Variance 0.0001 0.0001 0.0001 0.0001 0.0001 0.0001 0.0001 0.0001
## Cumulative Proportion  0.9944 0.9945 0.9946 0.9947 0.9948 0.9949 0.9950 0.9951
##                           PC65   PC66    PC67    PC68    PC69    PC70    PC71
## Standard deviation     0.2649 0.2627 0.26171 0.26125 0.25987 0.25963 0.25834
## Proportion of Variance 0.0001 0.0001 0.00009 0.00009 0.00009 0.00009 0.00009
## Cumulative Proportion  0.9952 0.9953 0.99537 0.99547 0.99556 0.99566 0.99575
##                            PC72    PC73    PC74    PC75    PC76    PC77    PC78
## Standard deviation     0.25742 0.25611 0.25505 0.25485 0.25316 0.25130 0.25087
## Proportion of Variance 0.00009 0.00009 0.00009 0.00009 0.00009 0.00009 0.00009
## Cumulative Proportion  0.99584 0.99593 0.99602 0.99611 0.99620 0.99629 0.99637
##                            PC79    PC80    PC81    PC82    PC83    PC84    PC85
## Standard deviation     0.25055 0.24895 0.24803 0.24760 0.24603 0.24528 0.24408
## Proportion of Variance 0.00009 0.00009 0.00009 0.00008 0.00008 0.00008 0.00008
## Cumulative Proportion  0.99646 0.99654 0.99663 0.99671 0.99680 0.99688 0.99696
##                            PC86    PC87    PC88    PC89    PC90    PC91    PC92
## Standard deviation     0.24377 0.24264 0.24102 0.24034 0.23939 0.23924 0.23631
## Proportion of Variance 0.00008 0.00008 0.00008 0.00008 0.00008 0.00008 0.00008
## Cumulative Proportion  0.99705 0.99713 0.99721 0.99729 0.99737 0.99745 0.99752
##                            PC93    PC94    PC95    PC96    PC97    PC98    PC99
## Standard deviation     0.23589 0.23565 0.23440 0.23299 0.23282 0.23168 0.23120
## Proportion of Variance 0.00008 0.00008 0.00008 0.00008 0.00007 0.00007 0.00007
## Cumulative Proportion  0.99760 0.99768 0.99775 0.99783 0.99790 0.99798 0.99805
##                           PC100   PC101   PC102   PC103   PC104   PC105   PC106
## Standard deviation     0.23091 0.22966 0.22899 0.22809 0.22681 0.22623 0.22515
```

```
## Proportion of Variance 0.00007 0.00007 0.00007 0.00007 0.00007 0.00007 0.00007
## Cumulative Proportion  0.99812 0.99820 0.99827 0.99834 0.99841 0.99848 0.99855
##                           PC107   PC108   PC109   PC110   PC111   PC112   PC113
## Standard deviation     0.22372 0.22352 0.22256 0.22097 0.22065 0.21802 0.21743
## Proportion of Variance 0.00007 0.00007 0.00007 0.00007 0.00007 0.00007 0.00007
## Cumulative Proportion  0.99862 0.99869 0.99876 0.99883 0.99890 0.99896 0.99903
##                           PC114   PC115   PC116   PC117   PC118   PC119   PC120
## Standard deviation     0.21694 0.21663 0.21612 0.21499 0.21340 0.21262 0.21209
## Proportion of Variance 0.00007 0.00006 0.00006 0.00006 0.00006 0.00006 0.00006
## Cumulative Proportion  0.99909 0.99916 0.99922 0.99928 0.99935 0.99941 0.99947
##                           PC121   PC122   PC123   PC124   PC125   PC126   PC127
## Standard deviation     0.21108 0.20893 0.20847 0.20763 0.20589 0.20431 0.20336
## Proportion of Variance 0.00006 0.00006 0.00006 0.00006 0.00006 0.00006 0.00006
## Cumulative Proportion  0.99953 0.99959 0.99965 0.99971 0.99977 0.99983 0.99989
##                           PC128   PC129
## Standard deviation     0.20256 0.20041
## Proportion of Variance 0.00006 0.00006
## Cumulative Proportion  0.99994 1.00000
```

```
summary(pc3)
```

```
## Importance of components:
##                            PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.9700 1.15225 1.05896 0.99179 0.82029 0.80098 0.71735
## Proportion of Variance 0.3464 0.05213 0.04403 0.03862 0.02642 0.02519 0.02021
## Cumulative Proportion  0.3464 0.39851 0.44254 0.48116 0.50759 0.53278 0.55298
##                            PC8     PC9   PC10   PC11    PC12    PC13    PC14
## Standard deviation     0.66245 0.64187 0.6013 0.5906 0.53527 0.51607 0.50736
## Proportion of Variance 0.01723 0.01618 0.0142 0.0137 0.01125 0.01046 0.01011
## Cumulative Proportion  0.57022 0.58639 0.6006 0.6143 0.62554 0.63600 0.64610
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.4706 0.45654 0.44885 0.42277 0.41102 0.40884 0.39880
## Proportion of Variance 0.0087 0.00818 0.00791 0.00702 0.00663 0.00656 0.00624
## Cumulative Proportion  0.6548 0.66299 0.67090 0.67792 0.68455 0.69111 0.69736
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.3942 0.39278 0.38487 0.38012 0.36830 0.35984 0.35231
## Proportion of Variance 0.0061 0.00606 0.00582 0.00567 0.00533 0.00508 0.00487
## Cumulative Proportion  0.7035 0.70952 0.71533 0.72101 0.72633 0.73142 0.73629
##                           PC29    PC30   PC31    PC32    PC33    PC34    PC35
## Standard deviation     0.34884 0.34457 0.3421 0.33577 0.33532 0.33238 0.32379
## Proportion of Variance 0.00478 0.00466 0.0046 0.00443 0.00442 0.00434 0.00412
## Cumulative Proportion  0.74107 0.74573 0.7503 0.75476 0.75917 0.76351 0.76763
##                           PC36    PC37    PC38    PC39    PC40    PC41    PC42
## Standard deviation     0.32245 0.31946 0.31676 0.31643 0.31306 0.30910 0.30601
## Proportion of Variance 0.00408 0.00401 0.00394 0.00393 0.00385 0.00375 0.00368
## Cumulative Proportion  0.77171 0.77572 0.77966 0.78359 0.78744 0.79119 0.79486
##                           PC43    PC44    PC45    PC46    PC47    PC48    PC49
## Standard deviation     0.30336 0.29909 0.29634 0.29554 0.29345 0.29191 0.28690
## Proportion of Variance 0.00361 0.00351 0.00345 0.00343 0.00338 0.00335 0.00323
## Cumulative Proportion  0.79848 0.80199 0.80544 0.80887 0.81225 0.81560 0.81883
##                           PC50    PC51    PC52    PC53    PC54    PC55    PC56
## Standard deviation     0.28406 0.28287 0.28170 0.28045 0.27867 0.27809 0.27508
## Proportion of Variance 0.00317 0.00314 0.00312 0.00309 0.00305 0.00304 0.00297
## Cumulative Proportion  0.82200 0.82514 0.82825 0.83134 0.83439 0.83743 0.84040
```

```
##                            PC57    PC58    PC59    PC60   PC61    PC62    PC63
## Standard deviation     0.27345 0.27240 0.27140 0.26962 0.2669 0.26665 0.26550
## Proportion of Variance 0.00294 0.00291 0.00289 0.00285 0.0028 0.00279 0.00277
## Cumulative Proportion  0.84334 0.84625 0.84914 0.85200 0.8548 0.85759 0.86035
##                            PC64    PC65    PC66    PC67    PC68    PC69    PC70
## Standard deviation     0.26494 0.26284 0.26176 0.26132 0.26033 0.25973 0.25835
## Proportion of Variance 0.00276 0.00271 0.00269 0.00268 0.00266 0.00265 0.00262
## Cumulative Proportion  0.86311 0.86582 0.86851 0.87119 0.87386 0.87651 0.87913
##                            PC71    PC72    PC73    PC74    PC75    PC76    PC77
## Standard deviation     0.2574 0.25619 0.25536 0.25495 0.25346 0.25133 0.25089
## Proportion of Variance 0.0026 0.00258 0.00256 0.00255 0.00252 0.00248 0.00247
## Cumulative Proportion  0.8817 0.88431 0.88687 0.88942 0.89194 0.89442 0.89689
##                            PC78    PC79    PC80    PC81    PC82    PC83    PC84
## Standard deviation     0.25066 0.24932 0.24822 0.24770 0.24604 0.24532 0.24409
## Proportion of Variance 0.00247 0.00244 0.00242 0.00241 0.00238 0.00236 0.00234
## Cumulative Proportion  0.89936 0.90180 0.90422 0.90663 0.90901 0.91137 0.91371
##                            PC85    PC86    PC87    PC88    PC89    PC90   PC91
## Standard deviation     0.24379 0.24280 0.24103 0.24035 0.23942 0.23938 0.2366
## Proportion of Variance 0.00233 0.00231 0.00228 0.00227 0.00225 0.00225 0.0022
## Cumulative Proportion  0.91604 0.91836 0.92064 0.92291 0.92516 0.92741 0.9296
##                            PC92    PC93    PC94    PC95    PC96    PC97   PC98
## Standard deviation     0.23591 0.23570 0.23441 0.23354 0.23289 0.23169 0.2313
## Proportion of Variance 0.00219 0.00218 0.00216 0.00214 0.00213 0.00211 0.0021
## Cumulative Proportion  0.93179 0.93397 0.93613 0.93827 0.94040 0.94251 0.9446
##                            PC99   PC100   PC101   PC102   PC103   PC104   PC105
## Standard deviation     0.23091 0.22966 0.22918 0.22810 0.22681 0.22625 0.22516
## Proportion of Variance 0.00209 0.00207 0.00206 0.00204 0.00202 0.00201 0.00199
## Cumulative Proportion  0.94670 0.94878 0.95084 0.95288 0.95490 0.95691 0.95890
##                           PC106   PC107   PC108   PC109   PC110   PC111   PC112
## Standard deviation     0.22389 0.22353 0.22260 0.22097 0.22074 0.21810 0.21759
## Proportion of Variance 0.00197 0.00196 0.00195 0.00192 0.00191 0.00187 0.00186
## Cumulative Proportion  0.96087 0.96283 0.96478 0.96670 0.96861 0.97048 0.97234
##                           PC113   PC114   PC115   PC116   PC117   PC118   PC119
## Standard deviation     0.21694 0.21682 0.21631 0.21510 0.21342 0.21270 0.21210
## Proportion of Variance 0.00185 0.00185 0.00184 0.00182 0.00179 0.00178 0.00177
## Cumulative Proportion  0.97418 0.97603 0.97787 0.97968 0.98147 0.98325 0.98501
##                           PC120   PC121   PC122   PC123   PC124   PC125   PC126
## Standard deviation     0.21109 0.20900 0.20847 0.20771 0.20591 0.20432 0.20346
## Proportion of Variance 0.00175 0.00172 0.00171 0.00169 0.00166 0.00164 0.00163
## Cumulative Proportion  0.98676 0.98848 0.99019 0.99188 0.99355 0.99518 0.99681
##                           PC127   PC128     PC129
## Standard deviation     0.20261 0.20047 2.253e-14
## Proportion of Variance 0.00161 0.00158 0.000e+00
## Cumulative Proportion  0.99842 1.00000 1.000e+00
```

2.Perform the log2(data + 1) transform and subtract row means from the samples. Set the seed to 333 and use k-means to cluster the samples into two clusters. Use `svd` to calculate the singular vectors. What is the correlation between the first singular vector and the sample clustering indicator?

```
edata_centered = edata2 - rowMeans(edata2)
set.seed(333)
kmeans1 = kmeans(t(edata_centered),centers=2)
names(kmeans1)
```

```
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
table(kmeans1$cluster)
```

```
##
##  1  2
## 52 77
```

```
svd3 = svd(edata_centered)
names(svd3)
```

```
## [1] "d" "u" "v"
```

```
length(svd3$v[,1])
```

```
## [1] 129
```

```
cor(svd3$v[,1],kmeans1$cluster)
```

```
## [1] -0.8678247
```

5.Perform the log2(data + 1) transform. Then fit a regression model to each sample using population as the outcome. Do this using the `lm.fit` function (hint: don't forget the intercept). What is the dimension of the residual matrix, the effects matrix and the coefficients matrix?

```
edata = as.matrix(edata2)

mod = model.matrix(~ pdata$population)
fit = lm.fit(mod,t(edata))
names(fit)
```

```
## [1] "coefficients"  "residuals"     "effects"       "rank"
## [5] "fitted.values" "assign"        "qr"            "df.residual"
```

```
nrow(fit$coefficients)
```

```
## [1] 2
```

```
nrow(fit$residuals)
```

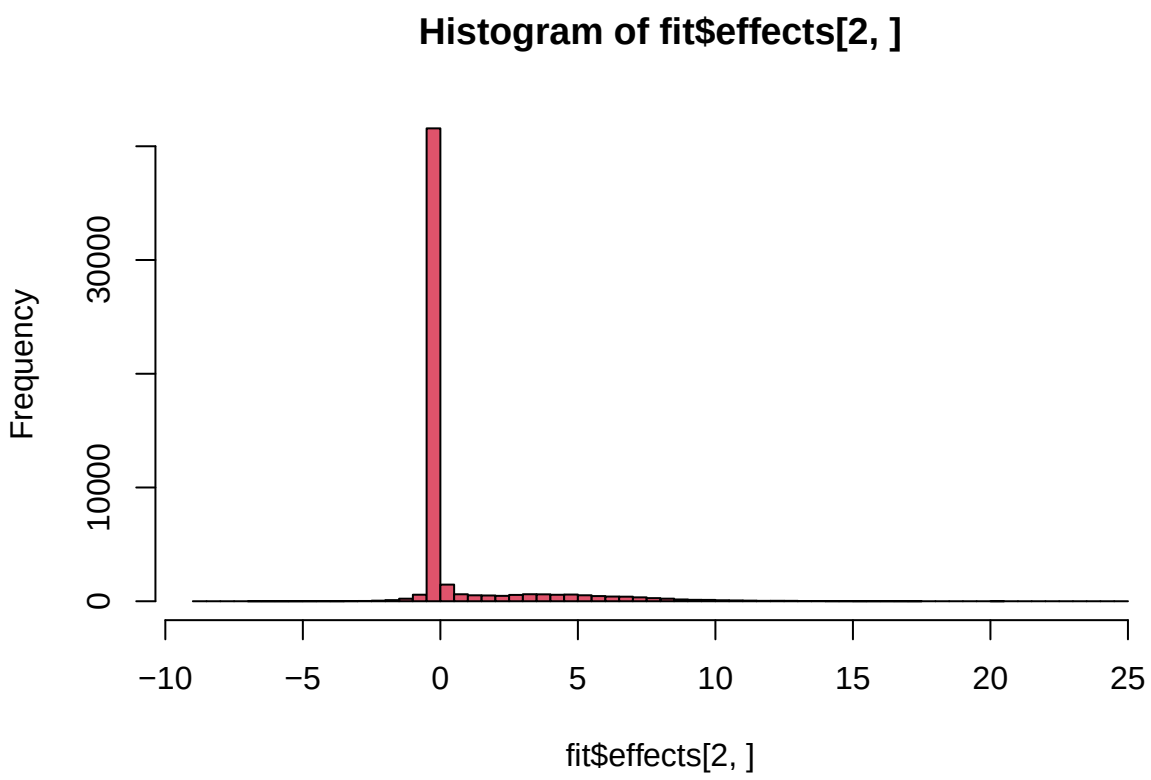
```
## [1] 129
```

```
nrow(fit$effects)
```

```
## [1] 129
```

6.Perform the log2(data + 1) transform. Then fit a regression model to each sample using population as the outcome. Do this using the `lm.fit` function (hint: don't forget the intercept). What is the effects matrix?

```
hist(fit$effects[2,],col=2,breaks=100)
```




```
nrow(fit$effects)
```

```
## [1] 129
```

9.Why is it difficult to distinguish the study effect from the population effect in the Montgomery Pickrell dataset from ReCount?

Load the Bodymap data with the following command

```
con =url("http://bowtie-bio.sourceforge.net/recount/ExpressionSets/bodymap_eset.RData")
load(file=con)
close(con)
bm = bodymap.eset
edata = exprs(bm)
pdata_bm=pData(bm)
ls()
```

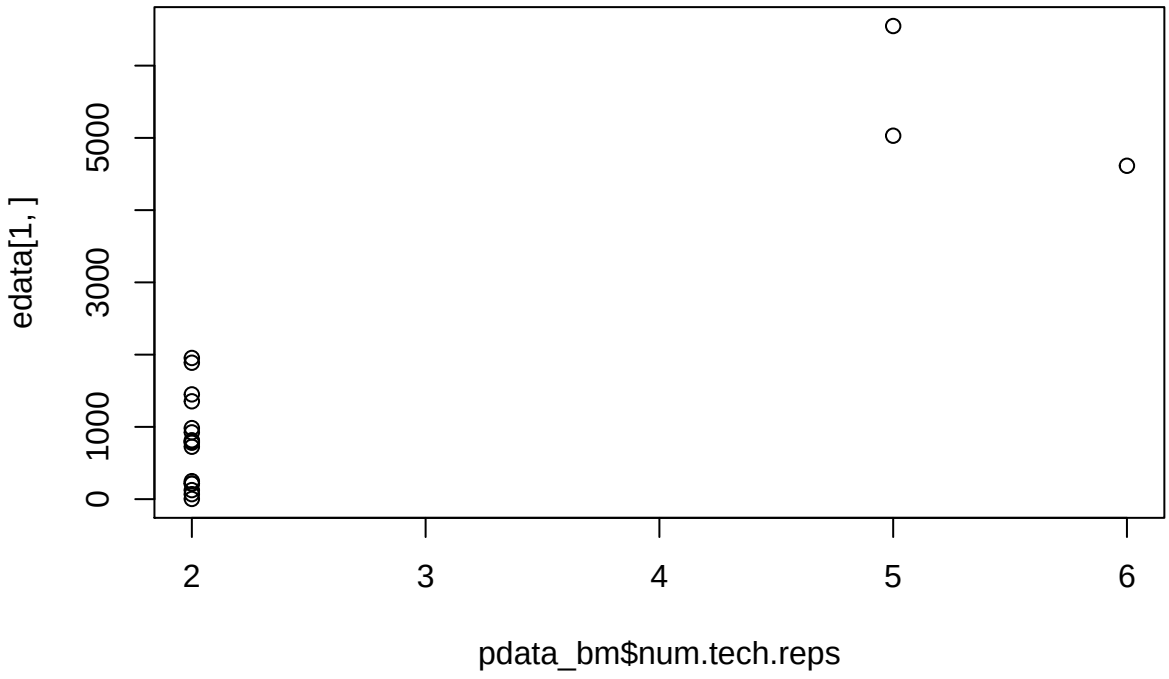
```
##  [1] "bm"             "bodymap.eset"   "con"            "edata"
##  [5] "edata_centered" "edata1"         "edata2"         "edata3"
##  [9] "fdata"          "fit"            "kmeans1"        "mod"
## [13] "montpick.eset"  "mp"             "pc1"            "pc2"
## [17] "pc3"            "pdata"          "pdata_bm"       "svd3"
```

3.Fit a linear model relating the first gene's counts to the number of technical replicates, treating the number of replicates as a factor. Plot the data for this gene versus the covariate. Can you think of why this model might not fit well?

```
edata = as.matrix(edata)
lm1 = lm(edata[1,] ~ as.factor(pdata_bm$num.tech.reps))
tidy(lm1)
```

```
## # A tibble: 3 x 5
##   term                                estimate std.error statistic      p.value
##   <chr>                                  <dbl>     <dbl>     <dbl>        <dbl>
## 1 (Intercept)                             784.      166.      4.72 0.000230
## 2 as.factor(pdata_bm$num.tech.reps)5     5005.      498.     10.0  0.0000000258
## 3 as.factor(pdata_bm$num.tech.reps)6     3830.      685.      5.59 0.0000404
```

```
plot(pdata_bm$num.tech.reps,edata[1,], col=1)
abline(lm1$coeff[1],lm1$coeff[2], col=2,lwd=3)
```



4.Fit a linear model relating he first gene's counts to the age of the person and the sex of the samples. What is the value and interpretation of the coefficient for age?

```
edata = as.matrix(edata)
lm2 = lm(edata[1,] ~ pdata_bm$age + pdata_bm$gender)
tidy(lm2)
```

```
## # A tibble: 3 x 5
##   term             estimate std.error statistic  p.value
##   <chr>               <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)        2332.     438.       5.32  0.000139
## 2 pdata_bm$age        -23.9      6.49    -3.69  0.00274
## 3 pdata_bm$genderM   -207.     236.      -0.877 0.397
```

7.Fit many regression models to the expression data where `age` is the outcome variable using the `lmFit` function from the `limma` package (hint: you may have to subset the expression data to the samples without missing values of age to get the model to fit). What is the coefficient for age for the 1,000th gene? Make a plot of the data and fitted values for this gene. Does the model fit well?

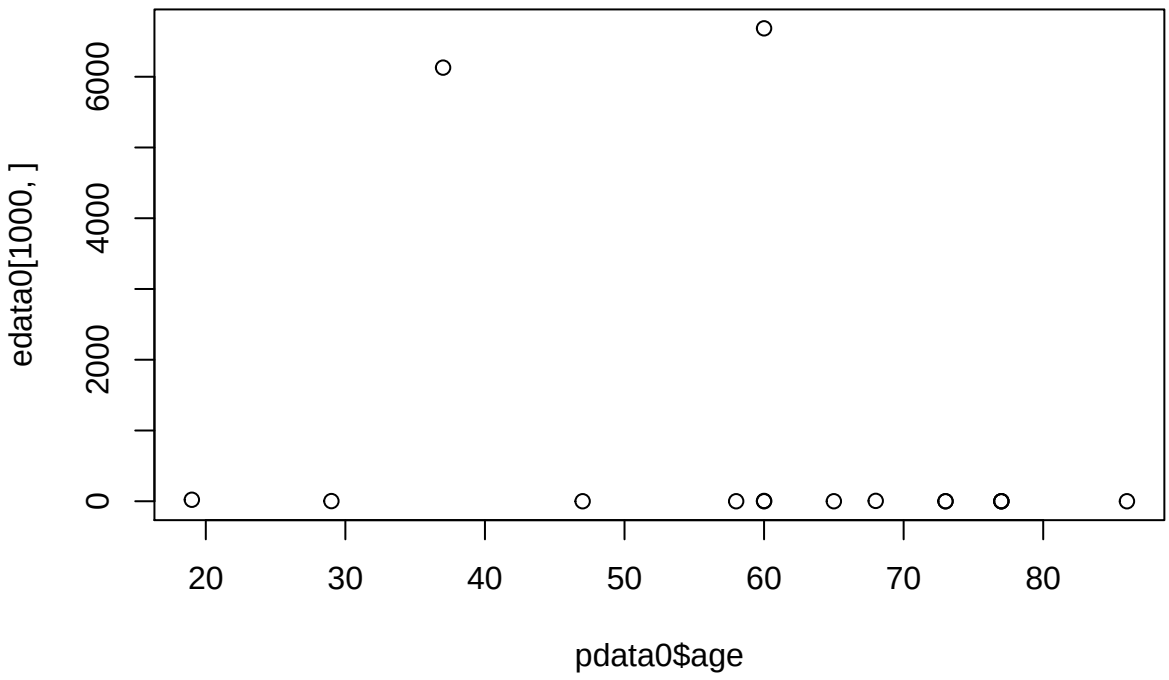
```
pdata0 = as.data.frame(na.omit(pdata_bm))
edata0 = edata[,-c(11,12,13)]

mod_adj = model.matrix(~ pdata0$age)
fit_limma = lmFit(edata0,mod_adj)
names(fit_limma)
```

```
##  [1] "coefficients"     "rank"             "assign"           "qr"
##  [5] "df.residual"      "sigma"            "cov.coefficients" "stdev.unscaled"
##  [9] "pivot"            "Amean"            "method"           "design"
```

```
fit_limma$coefficients[1000,]
```

```
## (Intercept)  pdata0$age
##  2469.87375   -27.61178
```

```
plot(pdata0$age,edata0[1000,], col=1)
abline(fit_limma$coeff[1],fit_limma$coeff[2], col=2,lwd=3)
```



8.Fit many regression models to the expression data where `age` is the outcome variable and `tissue.type` is an adjustment variable using the `lmFit` function from the `limma` package (hint: you may have to subset the expression data to the samples without missing values of age to get the model to fit). What is wrong with this model?

```
mod_adj = model.matrix(~ pdata0$age + pdata0[,3])
fit_limma = lmFit(edata0,mod_adj)
```

```
## Coefficients not estimable: pdata0[, 3]white_blood_cell pdata0[, 3]mixture
```

```
## Warning: Partial NA coefficients for 52580 probe(s)
```

10.Set the seed using the command `set.seed(33353)` then estimate a single surrogate variable using the `sva` function after log2(data + 1) transforming the expression data, removing rows with rowMeans less than 1, and treating age as the outcome (hint: you may have to subset the expression data to the samples without missing values of age to get the model to fit). What is the correlation between the estimated surrogate for batch and age? Is the surrogate more highly correlated with `race` or `gender`?

```
edata2 = log2(edata0 + 1)
edata = edata2[rowMeans(edata2) > 1, ]

mod = model.matrix(~age,data=pdata0)
mod0 = model.matrix(~1, data=pdata0)
sva1 = sva(edata,mod,mod0,n.sv=2)
```

```
## Number of significant surrogate variables is:  2
## Iteration (out of 5 ):1  2  3  4  5
```

```
# why error
pdata0$batch
```

```
## NULL
```

```
# summary(lm(sva1$sv ~ pdata0$batch))
```

```
devtools::session_info()
```

```
## - Session info ---------------------------------------------------------------
##  setting  value
##  version  R version 4.0.5 (2021-03-31)
##  os       Windows 10 x64
##  system   x86_64, mingw32
##  ui       RTerm
##  language (EN)
##  collate  Chinese (Simplified)_China.936
##  ctype    Chinese (Simplified)_China.936
##  tz       Asia/Taipei
##  date     2021-10-11
##
## - Packages -------------------------------------------------------------------
##  package       * version  date       lib source
##  annotate        1.68.0   2020-10-27 [1] Bioconductor
##  AnnotationDbi   1.52.0   2020-10-27 [1] Bioconductor
##  backports       1.2.1    2020-12-09 [1] CRAN (R 4.0.3)
##  Biobase       * 2.50.0   2020-10-27 [1] Bioconductor
##  BiocGenerics  * 0.36.1   2021-04-16 [1] Bioconductor
##  BiocParallel  * 1.24.1   2020-11-06 [1] Bioconductor
##  bit             4.0.4    2020-08-04 [1] CRAN (R 4.0.5)
##  bit64           4.0.5    2020-08-30 [1] CRAN (R 4.0.5)
##  blob            1.2.2    2021-07-23 [1] CRAN (R 4.0.5)
##  broom         * 0.7.9    2021-07-27 [1] CRAN (R 4.0.5)
##  cachem          1.0.6    2021-08-19 [1] CRAN (R 4.0.5)
##  callr           3.7.0    2021-04-20 [1] CRAN (R 4.0.5)
##  cli             3.0.1    2021-07-17 [1] CRAN (R 4.0.5)
##  crayon          1.4.1    2021-02-08 [1] CRAN (R 4.0.5)
##  DBI             1.1.1    2021-01-15 [1] CRAN (R 4.0.5)
##  desc            1.4.0    2021-09-28 [1] CRAN (R 4.0.5)
##  devtools      * 2.4.2    2021-06-07 [1] CRAN (R 4.0.5)
##  digest          0.6.27   2020-10-24 [1] CRAN (R 4.0.5)
##  dplyr           1.0.7    2021-06-18 [1] CRAN (R 4.0.5)
##  edgeR           3.32.1   2021-01-14 [1] Bioconductor
##  ellipsis        0.3.2    2021-04-29 [1] CRAN (R 4.0.5)
##  evaluate        0.14     2019-05-28 [1] CRAN (R 4.0.5)
##  fansi           0.5.0    2021-05-25 [1] CRAN (R 4.0.5)
##  fastmap         1.1.0    2021-01-25 [1] CRAN (R 4.0.5)
##  fs              1.5.0    2020-07-31 [1] CRAN (R 4.0.5)
##  genefilter    * 1.72.1   2021-01-21 [1] Bioconductor
```

```
##  generics        0.1.0    2020-10-31 [1] CRAN (R 4.0.5)
##  glue            1.4.2    2020-08-27 [1] CRAN (R 4.0.5)
##  highr           0.9      2021-04-16 [1] CRAN (R 4.0.5)
##  htmltools       0.5.2    2021-08-25 [1] CRAN (R 4.0.5)
##  httr            1.4.2    2020-07-20 [1] CRAN (R 4.0.5)
##  IRanges         2.24.1   2020-12-12 [1] Bioconductor
##  knitr           1.36     2021-09-29 [1] CRAN (R 4.0.5)
##  lattice         0.20-45  2021-09-22 [1] CRAN (R 4.0.5)
##  lifecycle       1.0.1    2021-09-24 [1] CRAN (R 4.0.5)
##  limma         * 3.46.0   2020-10-27 [1] Bioconductor
##  locfit          1.5-9.4  2020-03-25 [1] CRAN (R 4.0.5)
##  magrittr        2.0.1    2020-11-17 [1] CRAN (R 4.0.5)
##  Matrix          1.3-4    2021-06-01 [1] CRAN (R 4.0.5)
##  matrixStats     0.61.0   2021-09-17 [1] CRAN (R 4.0.5)
##  memoise         2.0.0    2021-01-26 [1] CRAN (R 4.0.5)
##  mgcv          * 1.8-37   2021-09-23 [1] CRAN (R 4.0.5)
##  nlme          * 3.1-153  2021-09-07 [1] CRAN (R 4.0.5)
##  pillar          1.6.3    2021-09-26 [1] CRAN (R 4.0.5)
##  pkgbuild        1.2.0    2020-12-15 [1] CRAN (R 4.0.5)
##  pkgconfig       2.0.3    2019-09-22 [1] CRAN (R 4.0.5)
##  pkgload         1.2.2    2021-09-11 [1] CRAN (R 4.0.5)
##  prettyunits     1.1.1    2020-01-24 [1] CRAN (R 4.0.5)
##  processx        3.5.2    2021-04-30 [1] CRAN (R 4.0.5)
##  ps              1.6.0    2021-02-28 [1] CRAN (R 4.0.5)
##  purrr           0.3.4    2020-04-17 [1] CRAN (R 4.0.5)
##  R6              2.5.1    2021-08-19 [1] CRAN (R 4.0.5)
##  Rcpp            1.0.7    2021-07-07 [1] CRAN (R 4.0.5)
##  remotes         2.4.1    2021-09-29 [1] CRAN (R 4.0.5)
##  rlang           0.4.11   2021-04-30 [1] CRAN (R 4.0.5)
##  rmarkdown       2.11     2021-09-14 [1] CRAN (R 4.0.5)
##  rprojroot       2.0.2    2020-11-15 [1] CRAN (R 4.0.5)
##  RSQLite         2.2.8    2021-08-21 [1] CRAN (R 4.0.5)
##  rstudioapi      0.13     2020-11-12 [1] CRAN (R 4.0.5)
##  S4Vectors       0.28.1   2020-12-09 [1] Bioconductor
##  sessioninfo     1.1.1    2018-11-05 [1] CRAN (R 4.0.5)
##  stringi         1.7.5    2021-10-04 [1] CRAN (R 4.0.5)
##  stringr         1.4.0    2019-02-10 [1] CRAN (R 4.0.5)
##  survival        3.2-13   2021-08-24 [1] CRAN (R 4.0.5)
##  sva           * 3.38.0   2020-10-28 [1] Bioconductor
##  testthat        3.0.4    2021-07-01 [1] CRAN (R 4.0.5)
##  tibble          3.1.4    2021-08-25 [1] CRAN (R 4.0.5)
##  tidyr           1.1.4    2021-09-27 [1] CRAN (R 4.0.5)
##  tidyselect      1.1.1    2021-04-30 [1] CRAN (R 4.0.5)
##  usethis       * 2.0.1    2021-02-10 [1] CRAN (R 4.0.5)
##  utf8            1.2.2    2021-07-24 [1] CRAN (R 4.0.5)
##  vctrs           0.3.8    2021-04-29 [1] CRAN (R 4.0.5)
##  withr           2.4.2    2021-04-18 [1] CRAN (R 4.0.5)
##  xfun            0.26     2021-09-14 [1] CRAN (R 4.0.5)
##  XML             3.99-0.8 2021-09-17 [1] CRAN (R 4.0.5)
##  xtable          1.8-4    2019-04-21 [1] CRAN (R 4.0.5)
##  yaml            2.2.1    2020-02-01 [1] CRAN (R 4.0.5)
##
## [1] D:/R/R-4.0.5/library
```