

# Jorge\_MLBrainstem

Jorge

2023-06-23

```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

library(readxl)
xl_NewDenC <- read_excel("~/xl_NewDenC.xlsx")

## New names:
## • `` -> `...1`

xl_NewDenC <- subset(xl_NewDenC, select = -c(...1, voxRowNum, X, Y, Z))

#filtering through data
TEST_307 <- filter(xl_NewDenC, `Structure-ID` == 307)

#Removing Structure-ID label
TEST_307 <- TEST_307 %>%
  select(-c(`Structure-ID`))

#Size of data
dim(TEST_307)

## [1] 90 4344

#New data frame created
head(TEST_307)

## # A tibble: 6 × 4,344
##   `Snap47-3` `Nvl-5` `Gtf3c2-7` `Pcdh9-10` `Ndufv3-12` `Mageh1-17` `Cdca7-26`
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 0.0661 0.00282 0.00161 0.00070 0.0140 0.00108 0.0354
## 2 -1 0.0000737 -1 -1 -1 -1 0.0214
## 3 0.0242 0.00116 -1 0.00574 0.00568 0.000208 0.0125
## 4 0.0300 0.00155 0.000195 0.0106 0.00617 -1 0.0963
## 5 0.0926 0.00340 0.00442 0.00011 0.0201 0.00182 0.0627
## 6 -1 0.000445 0.0000415 0.00208 0.00198 -1 0.0601
## # I 4,337 more variables: `Prkag2-29` <dbl>, `Dopey2-31` <dbl>,
## # `Anln-33` <dbl>, `Hepacam-35` <dbl>, `Setd5-37` <dbl>, `Cnih3-39` <dbl>,
## # `Thap1-44` <dbl>, `9130024F11Rik-51` <dbl>, `Mansc1-53` <dbl>,
## # `Pmch-55` <dbl>, `Prune2-57` <dbl>, `Acta1-60` <dbl>, `Actc1-62` <dbl>,
## # `Actr10-64` <dbl>, `Actr2-67` <dbl>, `Adamts1-69` <dbl>,
## # `9130024F11Rik-71` <dbl>, `Adcy2-73` <dbl>, `Adcy3-75` <dbl>,
## # `Adcy4-77` <dbl>, `Adcy5-79` <dbl>, `Adcy6-81` <dbl>, `Adcy9-83` <dbl>, ...
```

## Generating Correlation Matrix

Below is the code responsible for generating the correlation matrix of Structure-ID: 307.

```
library(corrplot)

## corrplot 0.92 loaded

library(doMC)

## Loading required package: foreach

## Loading required package: iterators

## Loading required package: parallel

#Creating correlation matrix
correlation_307 <- cor(TEST_307)

## Warning in cor(TEST_307): the standard deviation is zero

#Adding logical vector that indicates which genes are N/A
is_na <- is.na(correlation_307)

#Pulling the names of the columns in the correlation matrix
column_names <- rownames(correlation_307)

#Finding the indices of the logical vector that are TRUE
indices_Pruned <- which(is_na)

#Pulling the names of the indices that were found
na_columns <- column_names[indices_Pruned]

#Printing the names of the columns that have N/A
head(na_columns, 31)

## [1] "Adcy9-83" "Col6a1-290" "Erdr1-378"
## [4] "Enc1-390" "Extl3-411" "Jam2-602"
## [7] "Lcam-620" "Lig1-637" "Nrtn-741"
## [10] "Odz3-752" "Pcskin-777" "Prkc2-841"
## [13] "Sst-1001" "Vamp2-1092" "Ywhag-1124"
## [16] "Pex14-1397" "Tmem2-1409" "Plxnb1-1561"
## [19] "Zfp423-1607" "Trip4-1924" "Ankrd33-2006"
## [22] "Adrm1-2019" "Crat-2119" "Fbln3-2205"
## [25] "Ndufs1-2421" "Traf7-2527" "Mapk3-32576"
## [28] "Pim2-32583" "Limk2-61002" "Bcap29-104826"
## [31] "C230081A13Rik-73931624"

#Removing Inactive Genes
nu_TEST_307 <- TEST_307 %>% select(-c(`Adcy9-83`, `Col6a1-290`, `Erdr1-378`, `Enc1-390`, `Extl3-411`, `Jam2-602`,
`Lcam-620`, `Lig1-637`, `Nrtn-741`, `Odz3-752`, `Pcskin-777`, `Prkc2-841`, `Sst-1001`, `Vamp2-1092`, `Ywhag-1124`,
`Pex14-1397`, `Tmem2-1409`, `Plxnb1-1561`, `Zfp423-1607`, `Trip4-1924`, `Ankrd33-2006`, `Adrm1-2019`, `Crat-2119`,
`Fbln3-2205`, `Ndufs1-2421`, `Traf7-2527`, `Mapk3-32576`, `Pim2-32583`, `Limk2-61002`, `Bcap29-104826`,
`C230081A13Rik-73931624` ))

#Creating correlation matrix again
nu_correlation_307 <- cor(nu_TEST_307)
```

## Screening Through Correlation Matrix

### Correlation Threshold: 90%

Below is my attempt at screening through the produced correlation matrix (nu\_correlation\_307) and returning specific gene pairs whose correlation coefficient values are equal to or greater than 90%.

```
library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

# METHOD 1 #
#Screening through matrix and returning columns (genes) that have 90% correlation or greater
highlyCorrelated90 <- findCorrelation(nu_correlation_307, cutoff = .90, names = TRUE)

#Printing a sample of the returned columns (genes)
head(highlyCorrelated90)

## [1] "Ml1t11-89" "Calm1-179" "Chd2-255" "D16Ert472e-318"
## [5] "Dlx3-350" "Ripply3-367"

tail(highlyCorrelated90)

## [1] "Rbm11-1411" "Taok1-74819235" "Prkcq-77869867" "Ctla4-75079745"
## [5] "Cdc27-293466" "Slc6a15-73929609"

# METHOD 2 #
#Finding all column pairs that have a correlation coefficient greater than 0.9
correlated_pairs90 <- which(abs(nu_correlation_307) > 0.9, arr.ind = TRUE)

#Printing the number of pairs found in total
totalAT90 <- length(correlated_pairs90) - 4313
totalAT90

## [1] 86857

#Printing a sample of the specified column pairs.
head(correlated_pairs90)

## row col
## Snap47-3 1 1
## Ml1t11-89 32 1
## Lynx1-655 179 1
## Rwd02b-1507 493 1
## Nr4a1-1792 577 1
## Gabrb1-2247 751 1

tail(correlated_pairs90)

## row col
## Tbr1-79591351 4170 4313
## Mecp2-79908808 4219 4313
## Stub1-79912552 4228 4313
## Lhx9-112646029 4289 4313
## Oxt-112648396 4300 4313
## Thbs4-112651416 4313 4313
```

## Screening Through Correlation Matrix

### Correlation Threshold: 80%

Below is my attempt at screening through the produced correlation matrix (nu\_correlation\_307) and returning specific gene pairs whose correlation coefficient values are equal to or greater than 80%.

```
# METHOD 1
#Screening through matrix and returning columns (genes) that have 80% correlation or greater
highlyCorrelated80 <- findCorrelation(nu_correlation_307, cutoff = .80, names = TRUE)

#Printing a sample of the returned columns (genes)
head(highlyCorrelated80)

## [1] "9130024F11Rik-71" "Ml1t11-89" "Akap8-91" "Bace2-143"
## [5] "Pofut2-147" "Btg3-168"

tail(highlyCorrelated80)

## [1] "Tro-74819332" "Hbb-b1-67936011" "Fryl-70431430" "Krt173-74750042"
## [5] "Cdc27-293466" "Slc6a15-73929609"

# METHOD 2
#Finding all column pairs that have a correlation coefficient greater than 0.8
correlated_pairs80 <- which(abs(nu_correlation_307) > 0.8, arr.ind = TRUE)

#Printing the number of pairs found in total
totalAT80 <- length(correlated_pairs80) - 4313
totalAT80

## [1] 167913

#Printing a sample of the specified column pairs.
head(correlated_pairs80)

## row col
## Snap47-3 1 1
## Aqt-87 31 1
## Ml1t11-89 32 1
## Lynx1-655 179 1
## Atp50-1158 331 1
## Crabbp1-1179 338 1

tail(correlated_pairs80)

## row col
## Stub1-79912552 4228 4313
## Ptger4-112203056 4272 4313
## Lhx9-112646029 4289 4313
## Nr2f2-112646890 4293 4313
## Oxt-112648396 4300 4313
## Thbs4-112651416 4313 4313
```

## Screening Through Correlation Matrix

### Correlation Threshold: 70%

Below is my attempt at screening through the produced correlation matrix (nu\_correlation\_307) and returning specific gene pairs whose correlation coefficient values are equal to or greater than 70%.

```
# METHOD 1
#Screening through matrix and returning columns (genes) that have 70% correlation or greater
highlyCorrelated70 <- findCorrelation(nu_correlation_307, cutoff = 0.70, names = TRUE)

#Printing a sample of the returned columns (genes)
head(highlyCorrelated70)

## [1] "Actc1-62" "Actr2-67" "9130024F11Rik-71" "Adcy5-79"
## [5] "Ml1t11-89" "Akap8-91"

tail(highlyCorrelated70)

## [1] "Synj2-71964267" "E2f1-2171" "Ttc390-70634139" "Pax9-2466"
## [5] "Slc6a15-73929609" "Trp73-73929579"

# METHOD 2
#Finding all column pairs that have a correlation coefficient greater than 0.7
correlated_pairs70 <- which(abs(nu_correlation_307) > 0.7, arr.ind = TRUE)

#Printing the number of pairs found in total
totalAT70 <- length(correlated_pairs70) - 4313
totalAT70

## [1] 429229

#Printing a sample of the specified column pairs.
head(correlated_pairs70)

## row col
## Snap47-3 1 1
## Pcdh9-10 4 1
## Hepacam-35 11 1
## Aqt-87 31 1
## Ml1t11-89 32 1
## Kit1-616 169 1

tail(correlated_pairs70)

## row col
## Ptger4-112203056 4272 4313
## Kdelr3-112203270 4273 4313
## Lhx9-112646029 4289 4313
## Nr2f2-112646890 4293 4313
## Oxt-112648396 4300 4313
## Thbs4-112651416 4313 4313
```