Lab #6
Multiple testing

In this lab, we will be working with an Affymetrix data set that was run on the human HGU95A array. This experiment was designed to assess the gene expression events in the frontal cortex due to aging. A total of 18 male and 12 female postmortem brain samples were obtained to assess this.

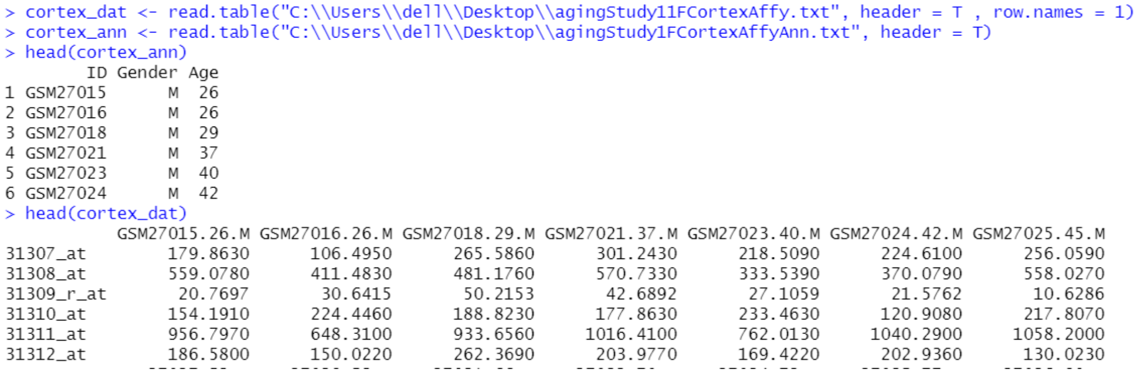
The analysis that we are interested in conducting is a direct follow up to the previous lab of differential expression. We first want to identify those genes/probes that are differentially expressed in the frontal cortex between old and young subjects, then between males and females. Next, we would like to evaluate the differences between a couple of multiple testing adjustment methods. As explained in the lecture and the course website, multiple testing is a necessary step to reduce false positives when conducting more than a single statistical test. You will generate some p-value plots to get an idea of the how conservative some methods are compared to others.

I have identified 2 gene vectors for you to use below, so do not calculate the t-test or adjustments on the entire array of genes/probes.

For the second part of this lab, you will be working with RNA-sequencing data from The Cancer Genome Atlas (TCGA), specifically a breast invasive carcinoma dataset of 119 patient tumors. The data matrix and annotation files are on the course website. We will be trying to confirm an observation from a meta-analysis performed by Mehra et al, 2005 in Cancer Research. The authors identified the gene (using arrays) and protein (using immunohistochemistry) GATA3 as a prognostic factor in breast cancer, where patients with low expression of GATA3 experienced overall worse survival. The PubMed abstract is here: http://www.ncbi.nlm.nih.gov/pubmed/16357129.

1.) Download the GEO Brain Aging study from the class website. Also obtain the annotation file for this data frame.

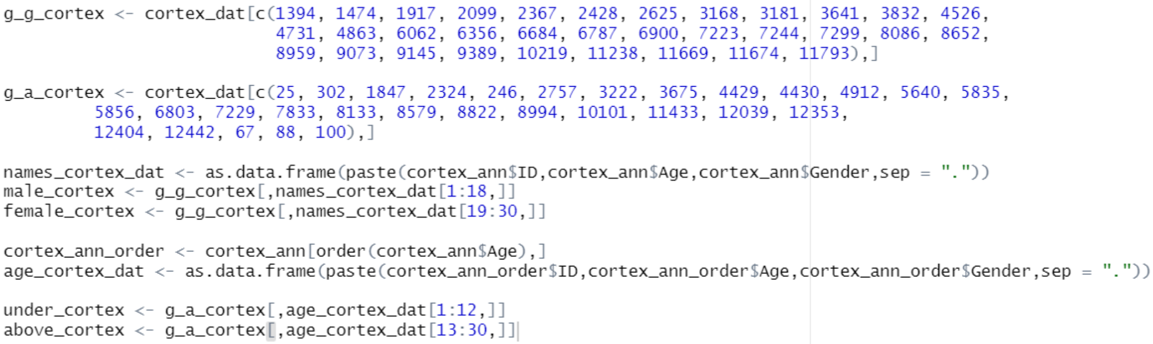
2.) Load into R, using read.table() function and the header=T/row.names=1 arguments for each data file.



```
> cortex_dat <- read.table("C:\\Users\\dell\\Desktop\\agingStudy11FCortexAffy.txt", header = T , row.names = 1)
> cortex_ann <- read.table("C:\\Users\\dell\\Desktop\\agingStudy1FCortexAffyAnn.txt", header = T)
> head(cortex_ann)
        ID Gender Age
1 GSM27015      M  26
2 GSM27016      M  26
3 GSM27018      M  29
4 GSM27021      M  37
5 GSM27023      M  40
6 GSM27024      M  42
> head(cortex_dat)
           GSM27015.26.M GSM27016.26.M GSM27018.29.M GSM27021.37.M GSM27023.40.M GSM27024.42.M GSM27025.45.M
31307_at        179.8630      106.4950      265.5860      301.2430      218.5090      224.6100      256.0590
31308_at        559.0780      411.4830      481.1760      570.7330      333.5390      370.0790      558.0270
31309_r_at       20.7697       30.6415       50.2153       42.6892       27.1059       21.5762       10.6286
31310_at        154.1910      224.4460      188.8230      177.8630      233.4630      120.9080      217.8070
31311_at        956.7970      648.3100      933.6560     1016.4100      762.0130     1040.2900     1058.2000
31312_at        186.5800      150.0220      262.3690      203.9770      169.4220      202.9360      130.0230
```

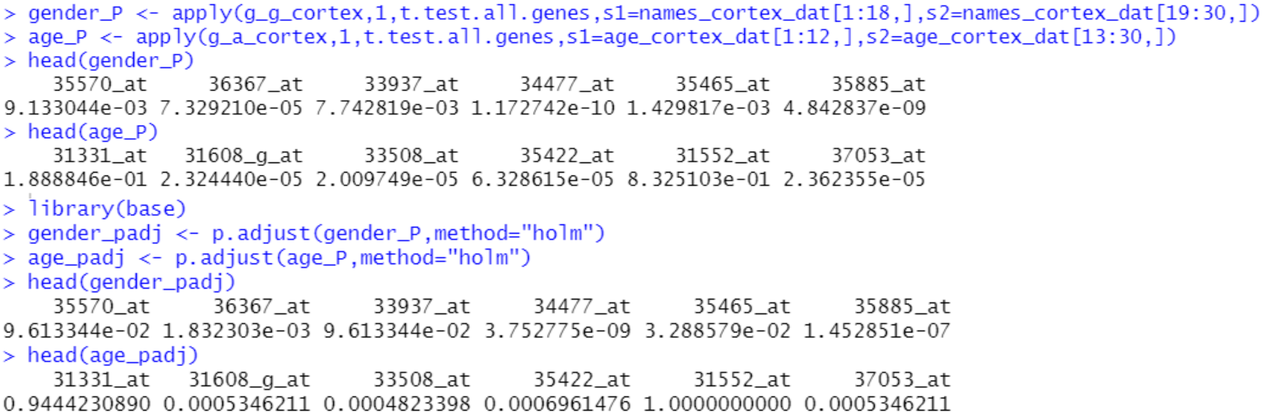
3.) Prepare 2 separate vectors for comparison. The first is a comparison between male and female patients. The current data frame can be left alone for this, since the males and females are all grouped together. The second vector is comparison between patients >= 50 years of age and those < 50 years of age.

To do this, you must use the annotation file and logical operators to isolate the correct arrays/samples.



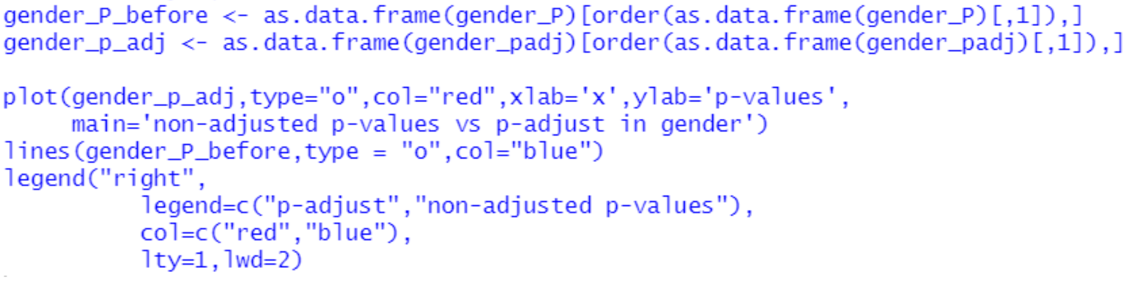
```
g_g_cortex <- cortex_dat[c(1394, 1474, 1917, 2099, 2367, 2428, 2625, 3168, 3181, 3641, 3832, 4526,
                           4731, 4863, 6062, 6356, 6684, 6787, 6900, 7223, 7244, 7299, 8086, 8652,
                           8959, 9073, 9145, 9389, 10219, 11238, 11669, 11674, 11793),]

g_a_cortex <- cortex_dat[c(25, 302, 1847, 2324, 246, 2757, 3222, 3675, 4429, 4430, 4912, 5640, 5835,
         5856, 6803, 7229, 7833, 8133, 8579, 8822, 8994, 10101, 11433, 12039, 12353,
         12404, 12442, 67, 88, 100),]

names_cortex_dat <- as.data.frame(paste(cortex_ann$ID,cortex_ann$Age,cortex_ann$Gender,sep = "."))
male_cortex <- g_g_cortex[,names_cortex_dat[1:18,]]
female_cortex <- g_g_cortex[,names_cortex_dat[19:30,]]

cortex_ann_order <- cortex_ann[order(cortex_ann$Age),]
age_cortex_dat <- as.data.frame(paste(cortex_ann_order$ID,cortex_ann_order$Age,cortex_ann_order$Gender,sep = "."))

under_cortex <- g_a_cortex[,age_cortex_dat[1:12,]]
above_cortex <- g_a_cortex[,age_cortex_dat[13:30,]]
```

4.) Run the t.test function from the notes using the first gene vector below for the gender comparison. Then use the second gene vector below for the age comparison. Using these p-values, use either p.adjust in the base library or mt.rawp2adjp in the multtest library to adjust the values for multiple corrections with the Holm's method.

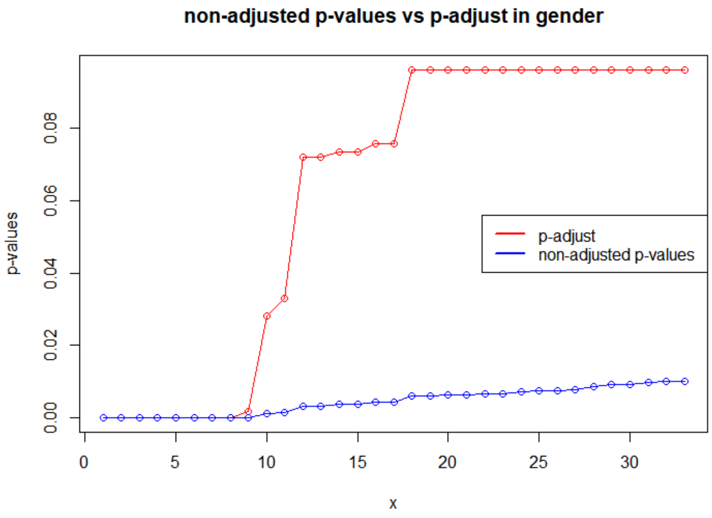


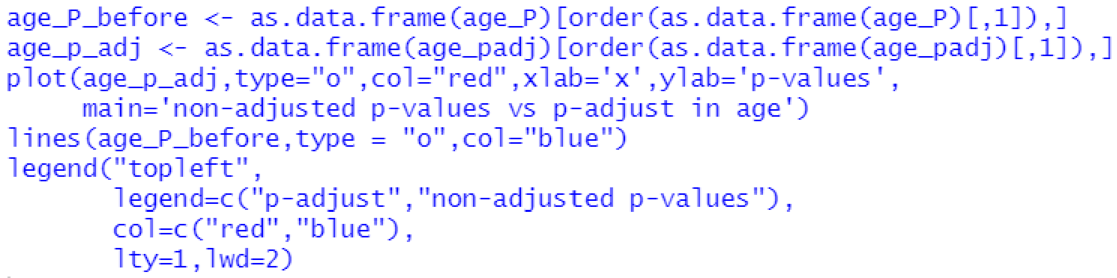
```
> gender_P <- apply(g_g_cortex,1,t.test.all.genes,s1=names_cortex_dat[1:18,],s2=names_cortex_dat[19:30,])
> age_P <- apply(g_a_cortex,1,t.test.all.genes,s1=age_cortex_dat[1:12,],s2=age_cortex_dat[13:30,])
> head(gender_P)
    35570_at     36367_at     33937_at     34477_at     35465_at     35885_at 
9.133044e-03 7.329210e-05 7.742819e-03 1.172742e-10 1.429817e-03 4.842837e-09 
> head(age_P)
    31331_at   31608_g_at     33508_at     35422_at     31552_at     37053_at 
1.888846e-01 2.324440e-05 2.009749e-05 6.328615e-05 8.325103e-01 2.362355e-05 
> library(base)
> gender_padj <- p.adjust(gender_P,method="holm")
> age_padj <- p.adjust(age_P,method="holm")
> head(gender_padj)
    35570_at     36367_at     33937_at     34477_at     35465_at     35885_at 
9.613344e-02 1.832303e-03 9.613344e-02 3.752775e-09 3.288579e-02 1.452851e-07 
> head(age_padj)
    31331_at   31608_g_at     33508_at     35422_at     31552_at     37053_at 
0.9444230890 0.0005346211 0.0004823398 0.0006961476 1.0000000000 0.0005346211 
```

5.) Sort the adjusted p-values and non-adjusted p-values and plot them vs. the x-axis of numbers for each comparison data set. Make sure that the two lines are different colors. Also make sure that the p-values are sorted before plotting.

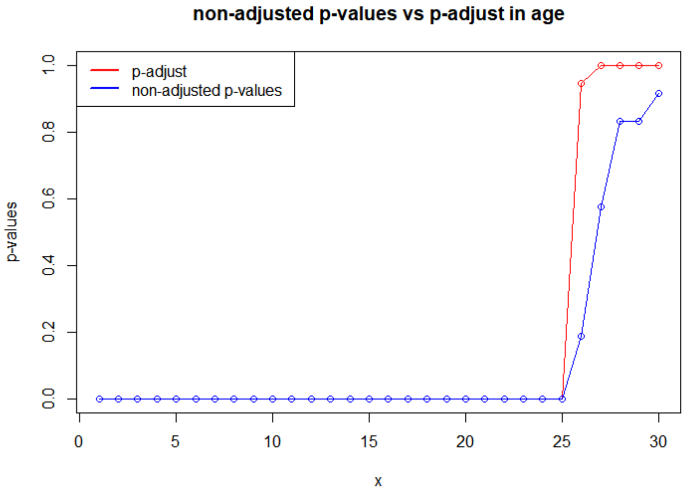


```
gender_P_before <- as.data.frame(gender_P)[order(as.data.frame(gender_P)[,1]),]
gender_p_adj <- as.data.frame(gender_padj)[order(as.data.frame(gender_padj)[,1]),]

plot(gender_p_adj,type="o",col="red",xlab='x',ylab='p-values',
     main='non-adjusted p-values vs p-adjust in gender')
lines(gender_P_before,type = "o",col="blue")
legend("right",
          legend=c("p-adjust","non-adjusted p-values"),
          col=c("red","blue"),
          lty=1,lwd=2)
```



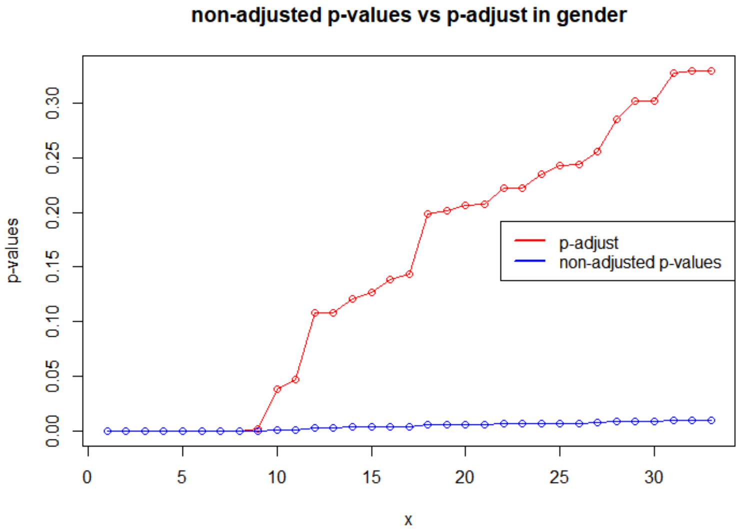


```
age_P_before <- as.data.frame(age_P)[order(as.data.frame(age_P)[,1]),]
age_p_adj <- as.data.frame(age_padj)[order(as.data.frame(age_padj)[,1]),]
plot(age_p_adj,type="o",col="red",xlab='x',ylab='p-values',
     main='non-adjusted p-values vs p-adjust in age')
lines(age_P_before,type = "o",col="blue")
legend("topleft",
       legend=c("p-adjust","non-adjusted p-values"),
       col=c("red","blue"),
       lty=1,lwd=2)
```

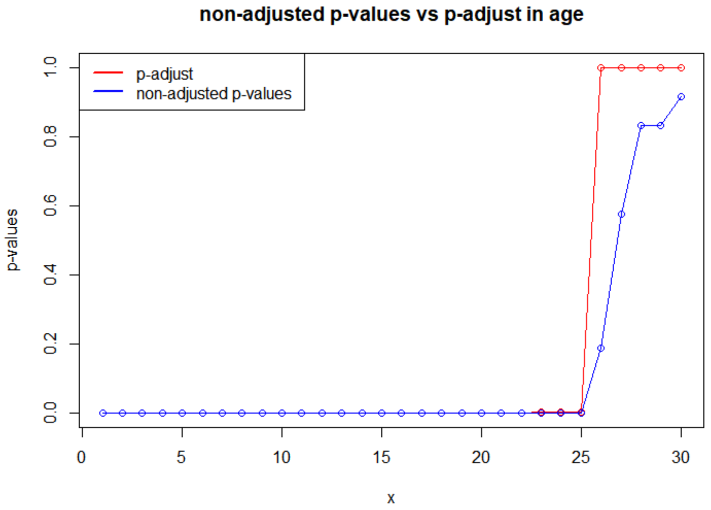


6.) Repeat #4 and #5 with the Bonferroni method.

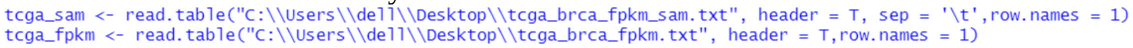
```
gender_padj_bonf <- p.adjust(gender_P,method="bonferroni")
age_padj_bonf <- p.adjust(age_P,method="bonferroni")

gender_P_before <- as.data.frame(gender_P)[order(as.data.frame(gender_P)[,1]),]
gender_p_bonf_adj <- as.data.frame(gender_padj_bonf)[order(as.data.frame(gender_padj_bonf)[,1]),]

plot(gender_p_bonf_adj,type="o",col="red",xlab='x',ylab='p-values',
     main='non-adjusted p-values vs p-adjust in gender')
lines(gender_P_before,type = "o",col="blue")
legend("right",
       legend=c("p-adjust","non-adjusted p-values"),
       col=c("red","blue"),
       lty=1,lwd=2)
```



```
age_P_before <- as.data.frame(age_P)[order(as.data.frame(age_P)[,1]),]
age_p_bonf_adj <- as.data.frame(age_padj_bonf)[order(as.data.frame(age_padj_bonf)[,1]),]
plot(age_p_bonf_adj,type="o",col="red",xlab='x',ylab='p-values',
     main='non-adjusted p-values vs p-adjust in age')
lines(age_P_before,type = "o",col="blue")
legend("topleft",
       legend=c("p-adjust","non-adjusted p-values"),
       col=c("red","blue"),
       lty=1,lwd=2)
```

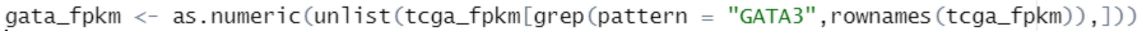


7.) Read in the $\log_2$ normalized fragments per kb per million mapped reads (FPKM) data matrix and annotation files. This is RNA-sequencing data that has normalized read counts on a similar scale to microarray intensities.



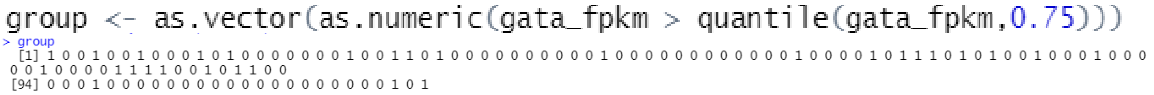
```
tcga_sam <- read.table("C:\\Users\\dell\\Desktop\\tcga_brca_fpkm_sam.txt", header = T, sep = '\t',row.names = 1)
tcga_fpkm <- read.table("C:\\Users\\dell\\Desktop\\tcga_brca_fpkm.txt", header = T,row.names = 1)
```

8.) Use grep to subset the data matrix only by gene ‘GATA3’ and make sure to cast this vector to numeric.



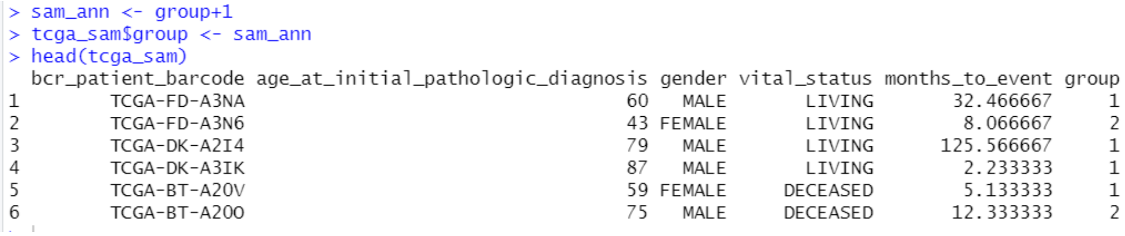
```
gata_fpkm <- as.numeric(unlist(tcga_fpkm[grep(pattern = "GATA3",rownames(tcga_fpkm)),]))
```

9.) Create a binary (1/0) vector for the patients where the **upper** 25% expression of GATA3 is coded as 1 and all other patients are coded as 0. Call this new variable ‘group’.



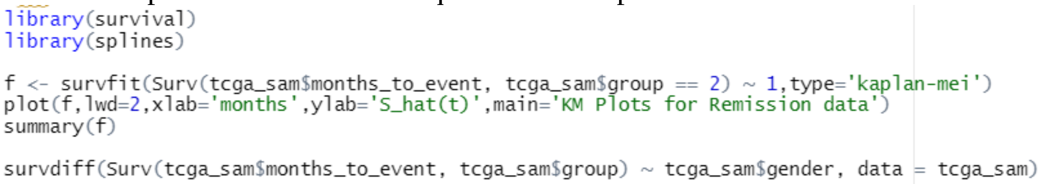
```
group <- as.vector(as.numeric(gata_fpkm > quantile(gata_fpkm,0.75)))
> group
  [1] 1 0 0 1 0 0 1 0 0 0 1 0 1 0 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 1 1 1 0 1 0 1 0 0 1 0 0 0 1 0 0 0
 0 0 1 0 0 0 0 1 1 1 1 0 0 1 0 1 1 0 0
 [94] 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1
```

10.) Create a data matrix with the ‘group’ variable you created in #9 and the remaining variables in the annotation file.

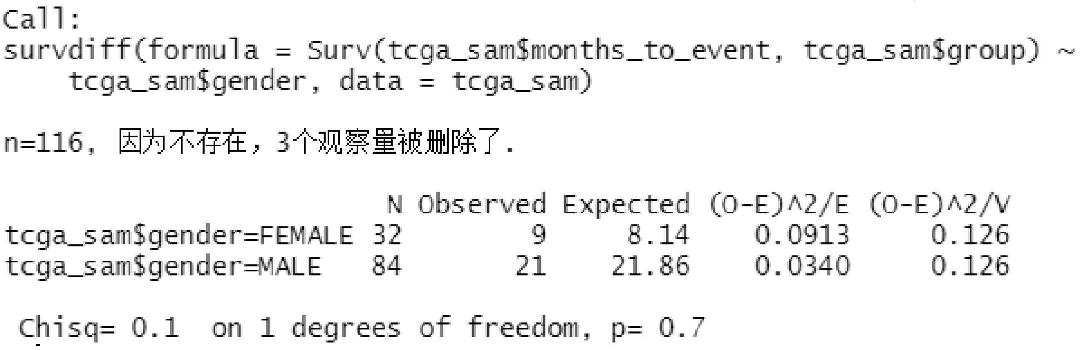


```
> sam_ann <- group+1
> tcga_sam$group <- sam_ann
> head(tcga_sam)
  bcr_patient_barcode age_at_initial_pathologic_diagnosis gender vital_status months_to_event group
1        TCGA-FD-A3NA                                  60   MALE       LIVING       32.466667     1
2        TCGA-FD-A3N6                                  43 FEMALE       LIVING        8.066667     2
3        TCGA-DK-A2I4                                  79   MALE       LIVING      125.566667     1
4        TCGA-DK-A3IK                                  87   MALE       LIVING        2.233333     1
5        TCGA-BT-A20V                                  59 FEMALE     DECEASED        5.133333     1
6        TCGA-BT-A20O                                  75   MALE     DECEASED       12.333333     2
```

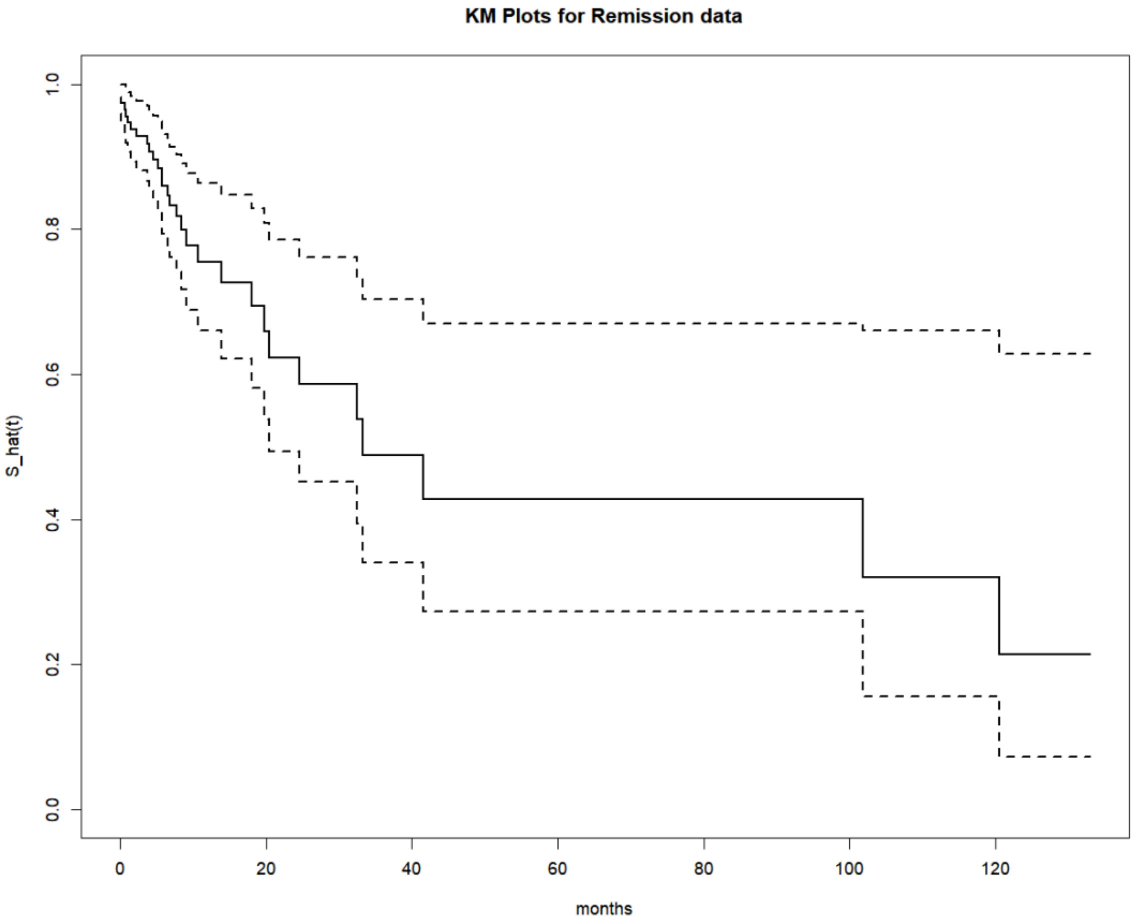
11.) Run a Kaplan-Meier (KM) analysis to determine if a difference in survival experience exists between the two GATA3 expression groups using the survdiff function. Extract the p-value from the chi squared test output.



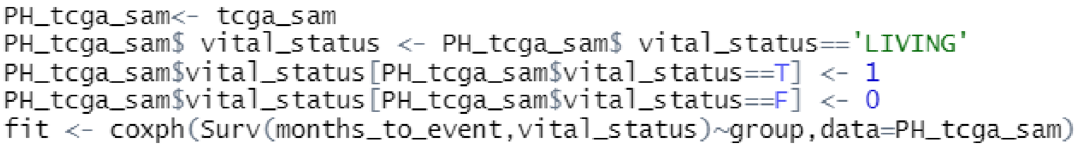
```
library(survival)
library(splines)

f <- survfit(Surv(tcga_sam$months_to_event, tcga_sam$group == 2) ~ 1,type='kaplan-mei')
plot(f,lwd=2,xlab='months',ylab='S_hat(t)',main='KM Plots for Remission data')
summary(f)

survdiff(Surv(tcga_sam$months_to_event, tcga_sam$group) ~ tcga_sam$gender, data = tcga_sam)
```



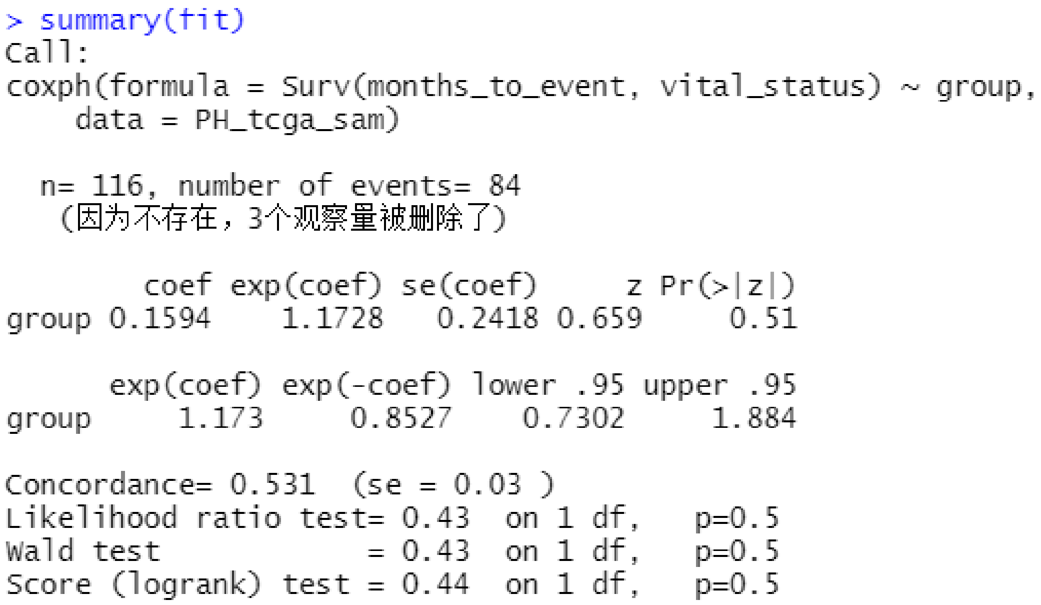
```
Call:
survdiff(formula = Surv(tcga_sam$months_to_event, tcga_sam$group) ~ 
    tcga_sam$gender, data = tcga_sam)

n=116, 因为不存在，3个观察量被删除了.

                        N Observed Expected (O-E)^2/E (O-E)^2/V
tcga_sam$gender=FEMALE 32        9     8.14    0.0913     0.126
tcga_sam$gender=MALE   84       21    21.86    0.0340     0.126

 Chisq= 0.1  on 1 degrees of freedom, p= 0.7
```



12.) Now run a Cox proportion hazard (PH) regression model on just the grouping variable (i.e. no other covariates) and extract both the p-value and hazard ratio from the output.



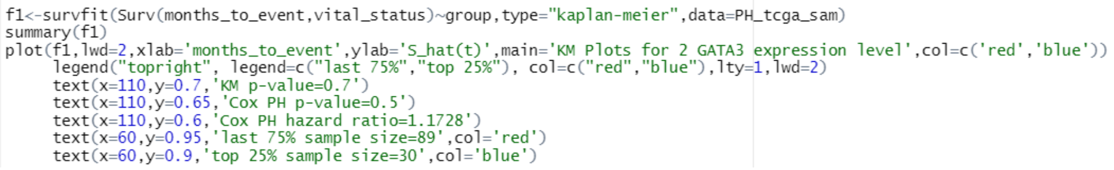
```
PH_tcga_sam<- tcga_sam
PH_tcga_sam$ vital_status <- PH_tcga_sam$ vital_status=='LIVING'
PH_tcga_sam$vital_status[PH_tcga_sam$vital_status==T] <- 1
PH_tcga_sam$vital_status[PH_tcga_sam$vital_status==F] <- 0
fit <- coxph(Surv(months_to_event,vital_status)~group,data=PH_tcga_sam)
```



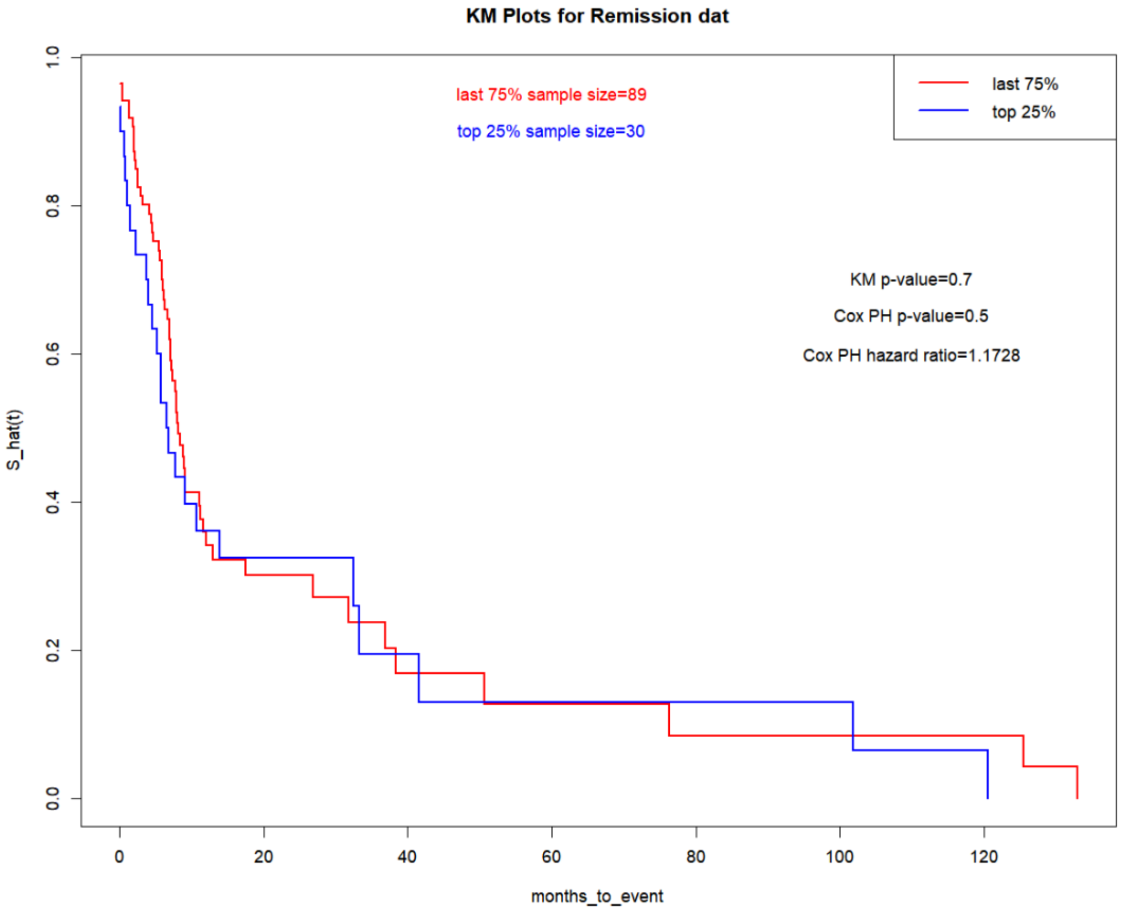
```
> summary(fit)
Call:
coxph(formula = Surv(months_to_event, vital_status) ~ group,
    data = PH_tcga_sam)

  n= 116, number of events= 84
   (因为不存在，3个观察量被删除了)

        coef exp(coef) se(coef)     z Pr(>|z|)
group 0.1594    1.1728   0.2418 0.659     0.51

      exp(coef) exp(-coef) lower .95 upper .95
group     1.173     0.8527    0.7302     1.884

Concordance= 0.531  (se = 0.03 )
Likelihood ratio test= 0.43  on 1 df,   p=0.5
Wald test            = 0.43  on 1 df,   p=0.5
Score (logrank) test = 0.44  on 1 df,   p=0.5
```

hazard ratio is 1.1728, p-value is 0.5

13.) Run the survfit() function only on the grouping variable (i.e. no other covariates) and plot the KM curves, being sure to label the two groups with a legend, two different colors for each line, and provide the KM p-value, Cox PH p-value, Cox PH hazard ratio, and sample sizes all in each of the two groups all on the plot.



```
f1<-survfit(Surv(months_to_event,vital_status)~group,type="kaplan-meier",data=PH_tcga_sam)
summary(f1)
plot(f1,lwd=2,xlab='months_to_event',ylab='S_hat(t)',main='KM Plots for 2 GATA3 expression level',col=c('red','blue'))
     legend("topright", legend=c("last 75%","top 25%"), col=c("red","blue"),lty=1,lwd=2)
     text(x=110,y=0.7,'KM p-value=0.7')
     text(x=110,y=0.65,'Cox PH p-value=0.5')
     text(x=110,y=0.6,'Cox PH hazard ratio=1.1728')
     text(x=60,y=0.95,'last 75% sample size=89',col='red')
     text(x=60,y=0.9,'top 25% sample size=30',col='blue')
```



14.) Does this result agree with the Mehra et al, study result?

I think the result did not correspond to Mehra et al result, because according to the #11-#13's result, the p_value did not show the significance among the 2 group of expression of GATA3.

Gene Vectors (indices for specific rows/genes)

# gender comparison gene vector
g.g <- c(1394, 1474, 1917, 2099, 2367, 2428, 2625, 3168, 3181, 3641, 3832, 4526, 4731, 4863, 6062, 6356, 6684, 6787, 6900, 7223, 7244, 7299, 8086, 8652, 8959, 9073, 9145, 9389, 10219, 11238, 11669, 11674, 11793)

# age comparison gene vector
g.a <- c(25, 302, 1847, 2324, 246, 2757, 3222, 3675, 4429, 4430, 4912, 5640, 5835, 5856, 6803, 7229, 7833, 8133, 8579, 8822, 8994, 10101, 11433, 12039, 12353, 12404, 12442, 67, 88, 100)