R-MarkDown-document: TNBC_TIL_analysis

Thomas Karn

May-18 2017

Table of Contents

SECTION-1 Selection of a gene expression based TNBC cohort from TCGA

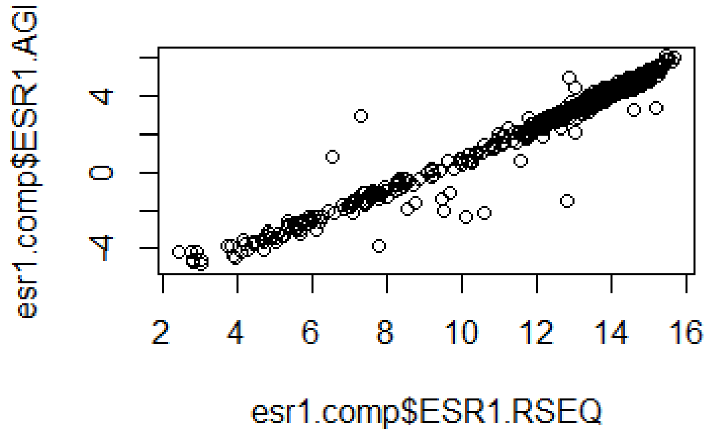
We use the cgdsr package to access data from the cBIO Portal.

```
library("cgdsr")
cbiop <- CGDS("http://www.cbioportal.org/public-portal/")
# getCancerStudies(cbiop)$cancer_study_id
clidat = getClinicalData(cbiop,"brca_tcga_all")
```

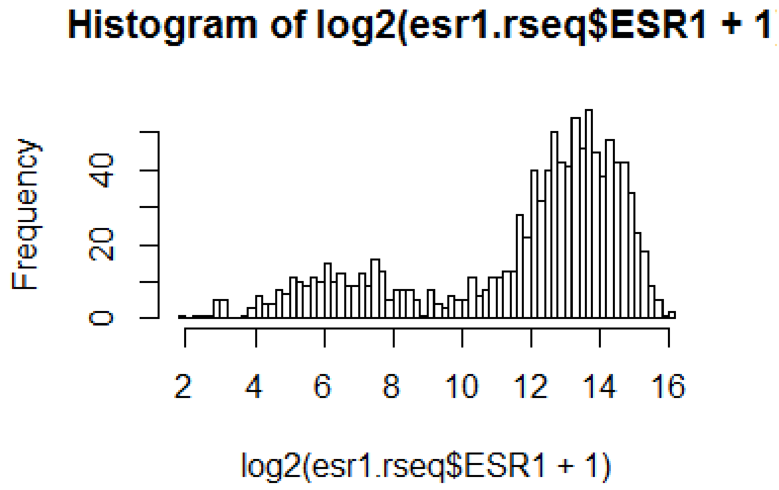
1.1 Analysis of correlation of ESR1 gene expression from RNA-Seq and Agilent microarray platform

```
esr1.rseq = getProfileData(cbiop,"ESR1","brca_tcga_rna_seq_v2_mrna",
"brca_tcga_all")
esr1.agi = getProfileData(cbiop,"ESR1","brca_tcga_mrna", "brca_tcga_all")
```

```r
# generate matrix of cases with both data for Agilent and RNA-Seq:
esr1.comp=as.data.frame(cbind(esr1.agi$ESR1, log2(esr1.rseq$ESR1+1))
        [(!is.nan(esr1.agi$ESR1)) & (!is.nan(esr1.rseq$ESR1)), ])
colnames(esr1.comp)=c("ESR1.AGI", "ESR1.RSEQ")

# correlation between Agilent and RNA-Seq:
plot(esr1.comp$ESR1.RSEQ, esr1.comp$ESR1.AGI)
```

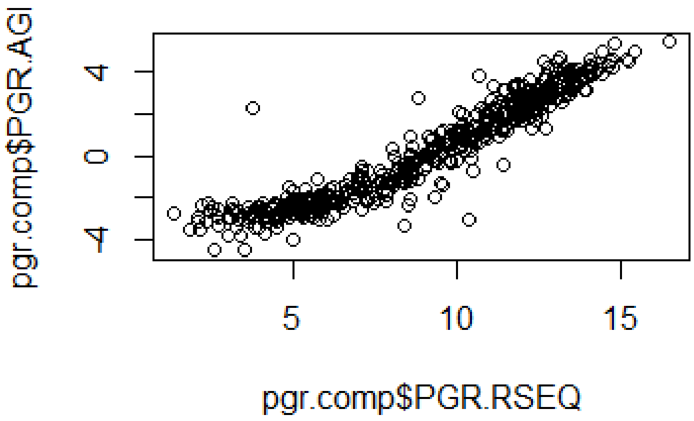


```r
cor(esr1.comp$ESR1.RSEQ, esr1.comp$ESR1.AGI)

## [1] 0.9821414

# bimodal distribution of RNA-Seq data
hist(log2(esr1.rseq$ESR1+1), breaks=80)
```

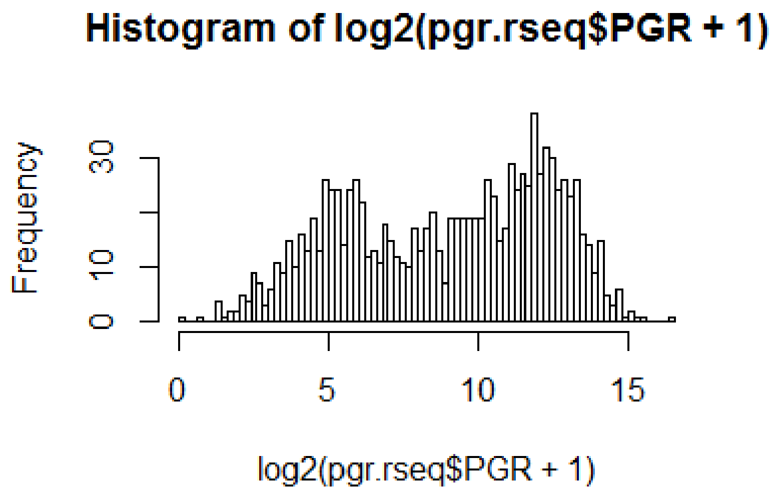


1.2 Analysis of correlation of PGR gene expression from RNA-Seq and Agilent microarray platform

```
pgr.rseq = getProfileData(cbiop,"PGR","brca_tcga_rna_seq_v2_mrna",
"brca_tcga_all")
pgr.agi = getProfileData(cbiop,"PGR","brca_tcga_mrna", "brca_tcga_all")

# generate matrix of cases with both data for Agilent and RNA-Seq:
pgr.comp=as.data.frame(cbind(pgr.agi$PGR, log2(pgr.rseq$PGR+1))
          [(!is.nan(pgr.agi$PGR)) & (!is.nan(pgr.rseq$PGR)), ])
colnames(pgr.comp)=c("PGR.AGI", "PGR.RSEQ")
# correlation between Agilent and RNA-Seq:
plot(pgr.comp$PGR.RSEQ, pgr.comp$PGR.AGI)
```

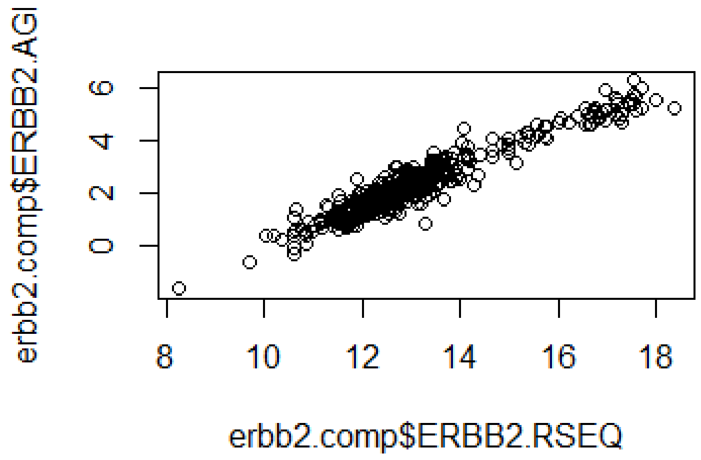


```
cor(pgr.comp$PGR.RSEQ, pgr.comp$PGR.AGI)

## [1] 0.9499931

# bimodal distribution of RNA-Seq data
hist(log2(pgr.rseq$PGR+1), breaks=80)
```



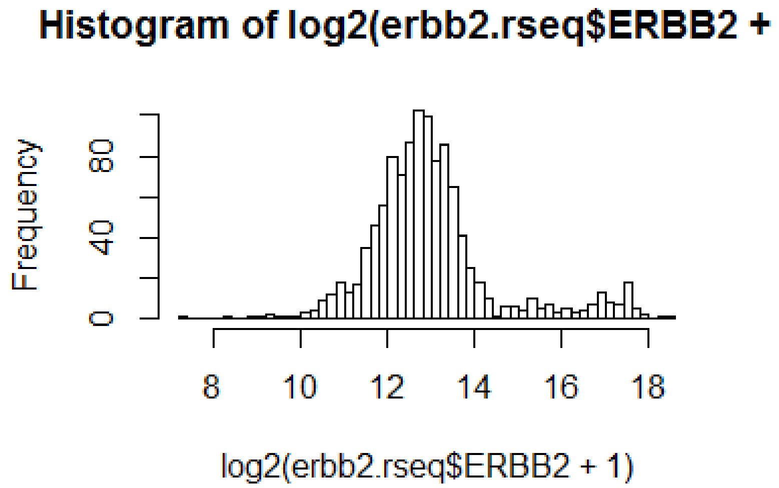
1.3 Analysis of correlation of HER2 gene expression from RNA-Seq and Agilent microarray platform

```
erbb2.rseq = getProfileData(cbiop,"ERBB2","brca_tcga_rna_seq_v2_mrna",
"brca_tcga_all")
erbb2.agi = getProfileData(cbiop,"ERBB2","brca_tcga_mrna", "brca_tcga_all")

# generate matrix of cases with both data for Agilent and RNA-Seq:
erbb2.comp=as.data.frame(cbind(erbb2.agi$ERBB2, log2(erbb2.rseq$ERBB2+1))
            [(!is.nan(erbb2.agi$ERBB2)) & (!is.nan(erbb2.rseq$ERBB2)), ])
colnames(erbb2.comp)=c("ERBB2.AGI", "ERBB2.RSEQ")
# correlation between Agilent and RNA-Seq:
plot(erbb2.comp$ERBB2.RSEQ, erbb2.comp$ERBB2.AGI)
```



```
cor(erbb2.comp$ERBB2.RSEQ, erbb2.comp$ERBB2.AGI)
```

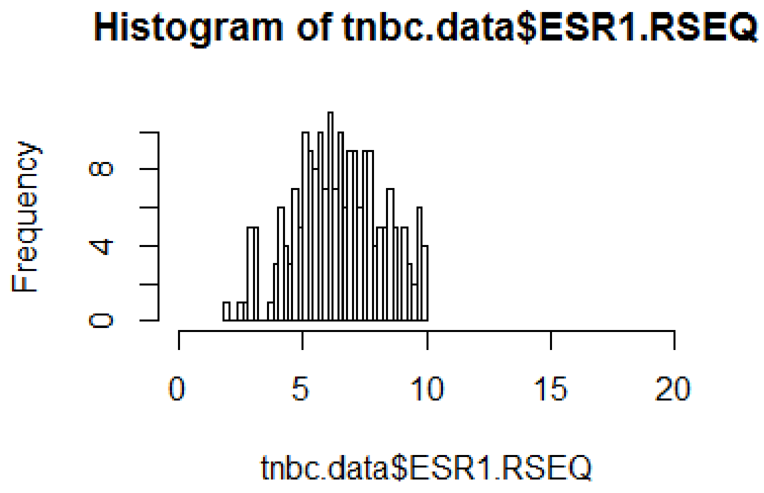
```
## [1] 0.9547622

# bimodal distribution of RNA-Seq data
hist(log2(erbb2.rseq$ERBB2+1), breaks=80)
```



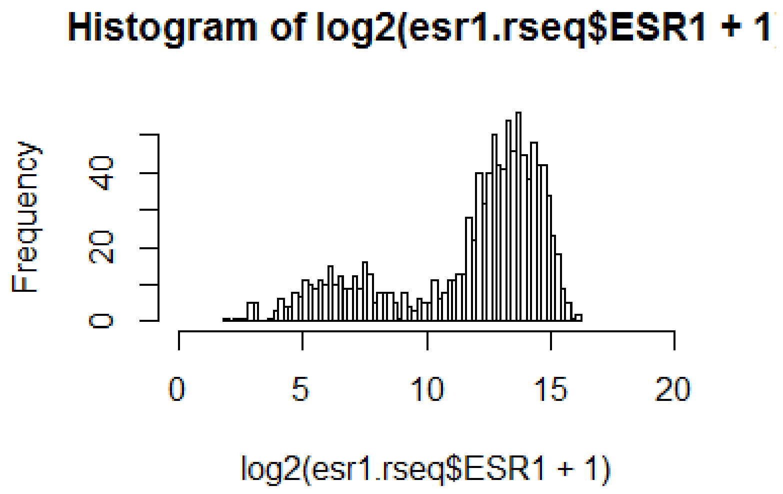
1.4 Generate TNBC dataset

```
# Select tnbc/dnbc based on cutoffs from distribution of RNA-Seq
# define a logical selection vector
tnbc.group= !is.na(esr1.rseq) & !is.na(erbb2.rseq) &
    (log2(esr1.rseq$ESR1+1)<10) &  (log2(erbb2.rseq$ERBB2+1)<14)
colnames(tnbc.group)="tnbc"
sum(na.omit(tnbc.group))

## [1] 208

# Generate tnbc dataset
tnbc.data= cbind(log2(esr1.rseq$ESR1+1)[tnbc.group],
                  log2(pgr.rseq$PGR+1)[tnbc.group],
                  log2(erbb2.rseq$ERBB2+1)[tnbc.group])
row.names(tnbc.data)= row.names(tnbc.group)[tnbc.group]
colnames(tnbc.data)=c("ESR1.RSEQ", "PGR.RSEQ", "ERBB2.RSEQ")

# Merge of Clinical data and tnbc dataset
# find subset in clidat corresponding to tnbc
clidat.sel=clidat[row.names(clidat)%in% row.names(tnbc.data),]
# merge tnbc.data and clinical data, left outer join:
tnbc.data= merge(tnbc.data, clidat.sel, by="row.names", all.x =TRUE)
#  "merge" creates resorted dataframe with the row.names
#      as a new first column "Row.names"
# rebuild structure (row.names):
row.names(tnbc.data)=tnbc.data$Row.names
tnbc.data=tnbc.data[,colnames(tnbc.data)!= "Row.names"]
```

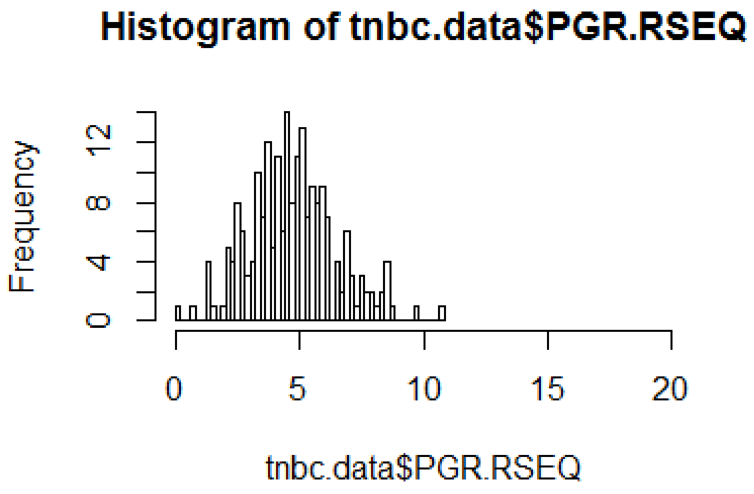
```
# check residual receptor expression in tnbc dataset:
hist(tnbc.data$ESR1.RSEQ, xlim=c(0,20), breaks=40) # tnbc group
```



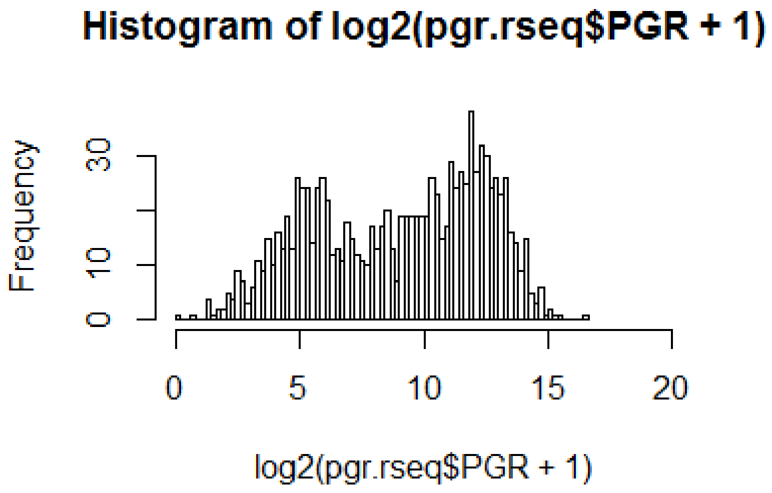
```
hist(log2(esr1.rseq$ESR1+1),xlim=c(0,20), breaks=80) # all samples
```



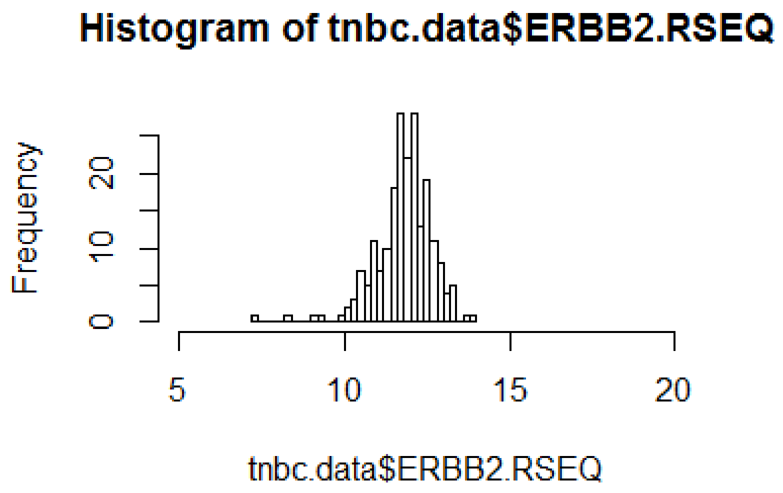
```
hist(tnbc.data$PGR.RSEQ, xlim=c(0,20), breaks=40) # tnbc group
```



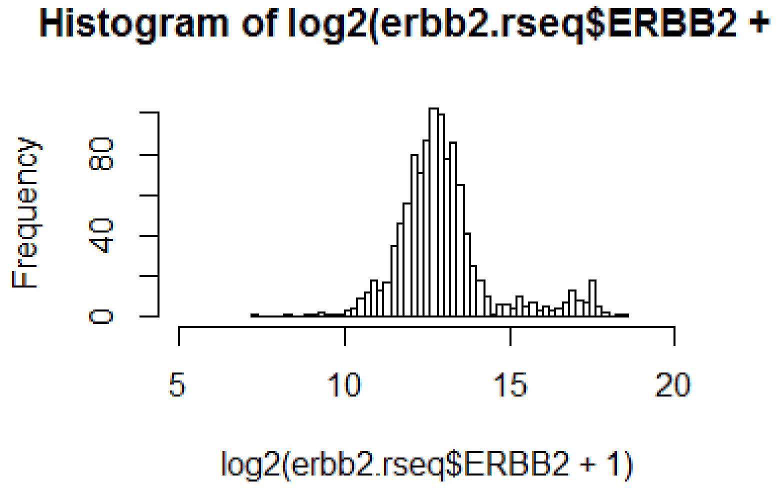
```r
hist(log2(pgr.rseq$PGR+1),xlim=c(0,20), breaks=80) # all samples
```



```r
hist(tnbc.data$ERBB2.RSEQ, xlim=c(5,20),  breaks=40) # tnbc group
```

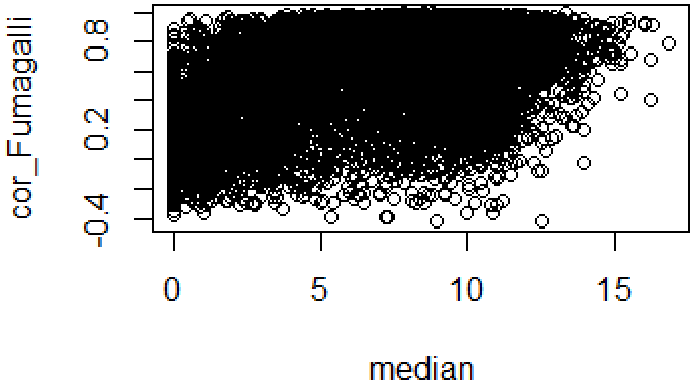


```
hist(log2(erbb2.rseq$ERBB2+1),xlim=c(5,20), breaks=80) # all samples
```

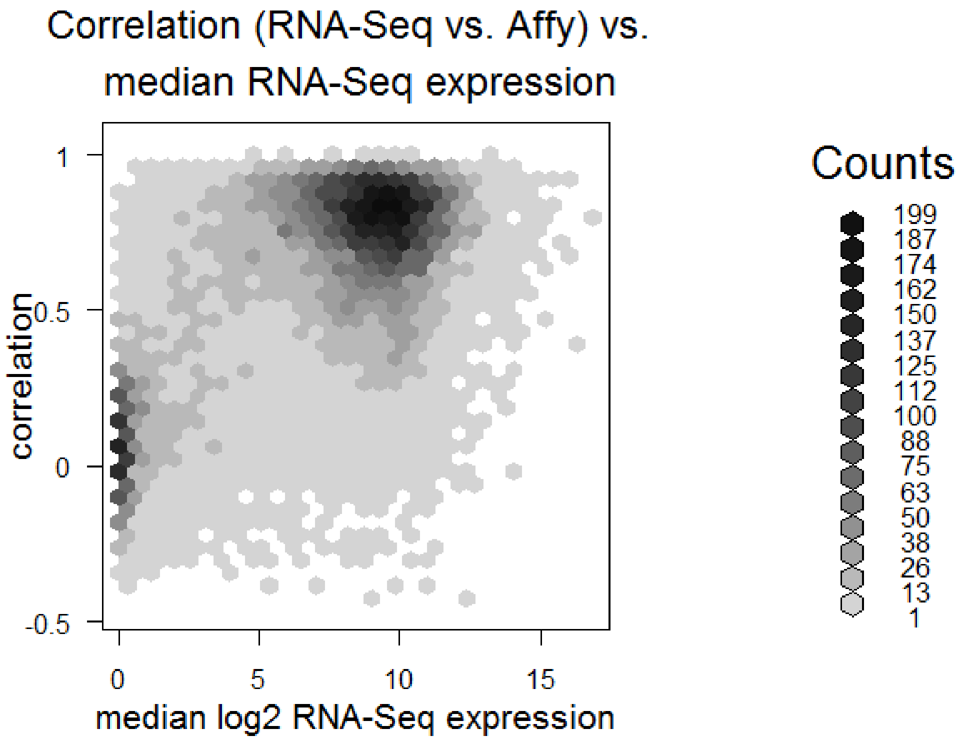


SECTION-2 Gene filtering in RNA-Seq data

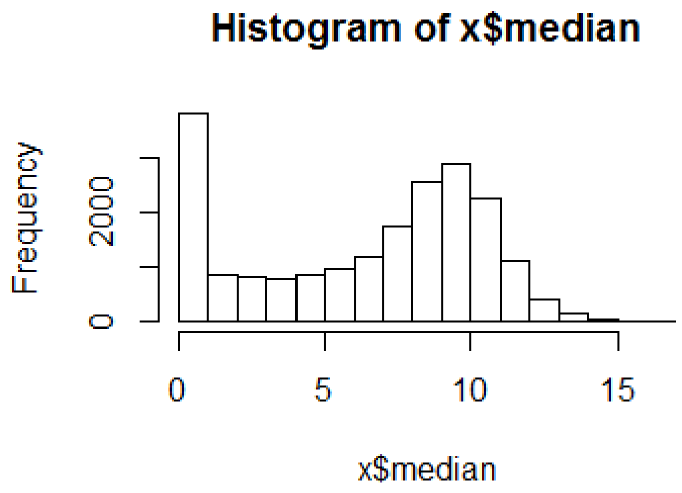
```
# Spearman correlation values between RNA-Seq and Affymetrix microarray
#  for 16,097 Jetset probes for 57 paired frozen breast cancer samples
#  can be obtained from:
#  Suppl.Tab.S2 of Fumagalli et al. 2014, PubmedID 25412710

n208.FumagCorrel <-
read.delim("2016_05_31_median_mean_n208RNASeq_vs_FumagalliCorrel.txt")

# Plot median expression vs Spearman correlation coefficient
x=n208.FumagCorrel[,c(1,3)]
plot(x)
```



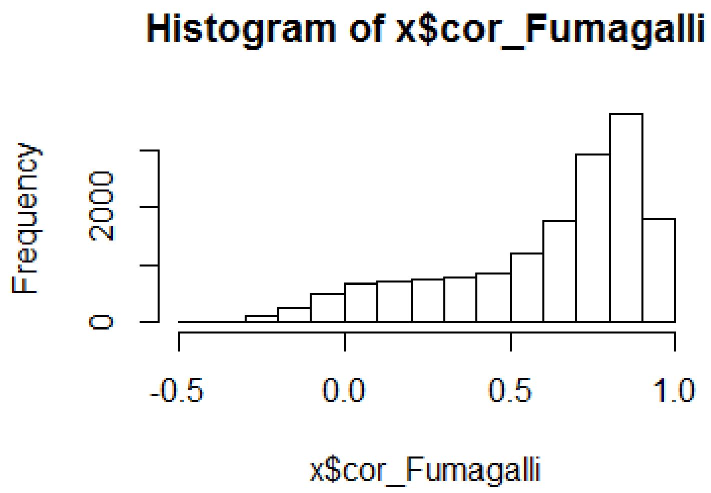
```
# Use hexbin plot to display the density of the scatter
library(hexbin)
plot(hexbin(x$median, x$cor_Fumagalli, xbins=30),
     xlab="median log2 RNA-Seq expression", ylab="correlation",
     main="Correlation (RNA-Seq vs. Affy) vs. \n median RNA-Seq expression")
```



```
# Distribution of median expression values
hist(x$median)
```



```r
# Distribution of Spearman correlation coefficients
hist(x$cor_Fumagalli)
```

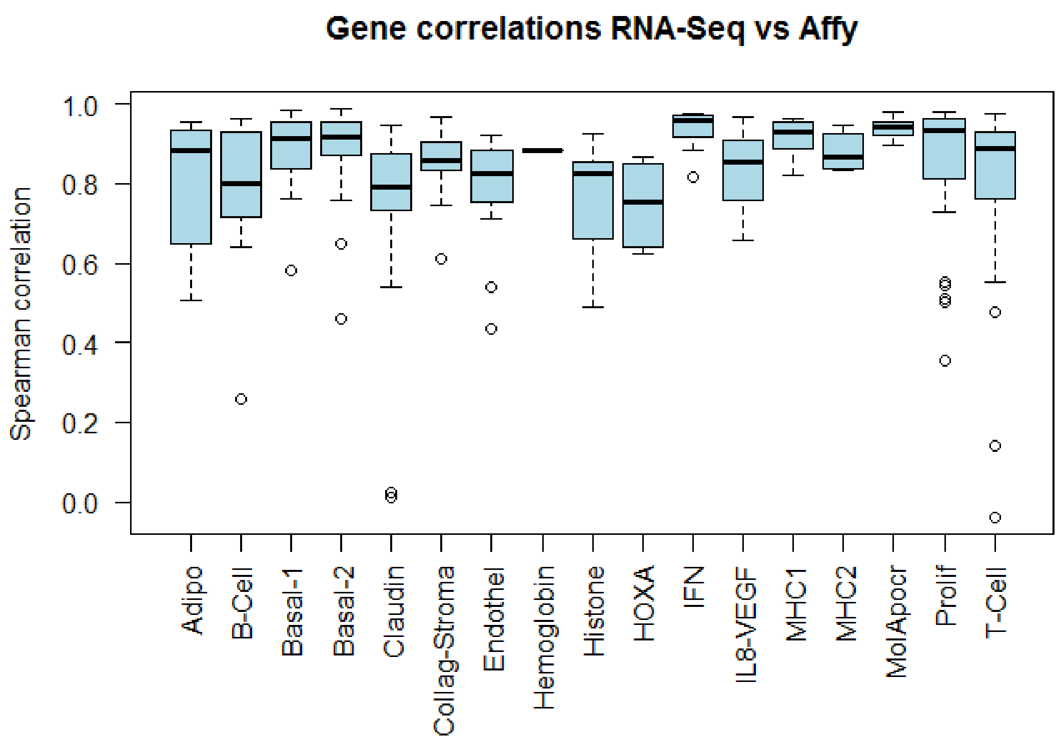


```r
rm(x)
```

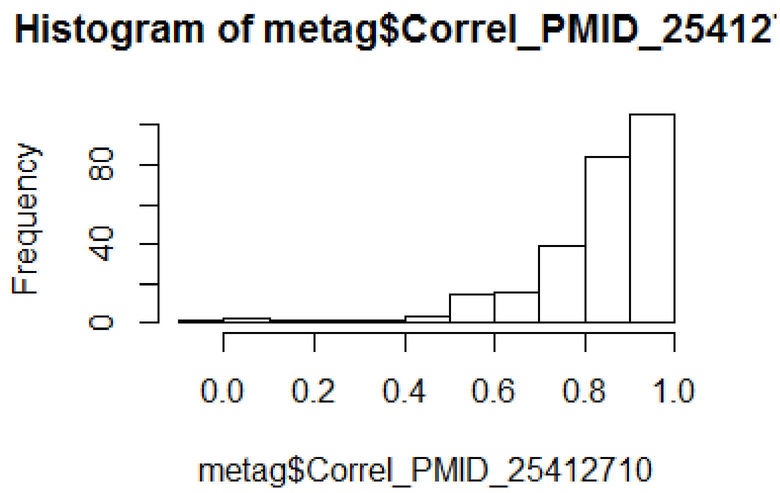
SECTION-3 Metagene construction

3.1 Metagene genes: RNA-Seq vs. Affy correlation

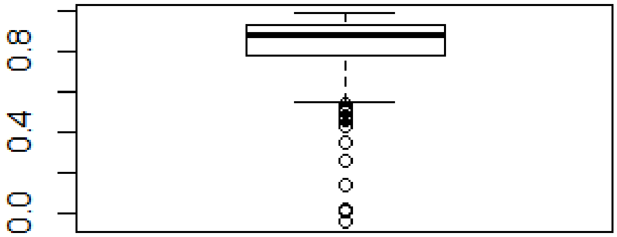
```r
metag <- read.delim("2016_06_01_TNBC-metagenes_gene_list.txt")

par(las = 2) # labels always perpendicular to the axis
par(mar=c(7,4,4,2)+0.1) # increase bottom margin
boxplot(Correl_PMID_25412710~TNBCmetagene_RNA.Seq,
         data=metag, notch=F, col="lightblue",
         ylab="Spearman correlation",
         main="Gene correlations RNA-Seq vs Affy" )
```



```
par(mar=c(5.1, 4.1, 4.1, 2.1))
hist(metag$Correl_PMID_25412710)
```



```
boxplot(metag$Correl_PMID_25412710)
```



```
median(metag$Correl_PMID_25412710, na.rm=T)

## [1] 0.8831346

summary(metag$Correl_PMID_25412710, na.rm=T)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
## -0.04109  0.77800  0.88310  0.82610  0.93210  0.98810       35
```

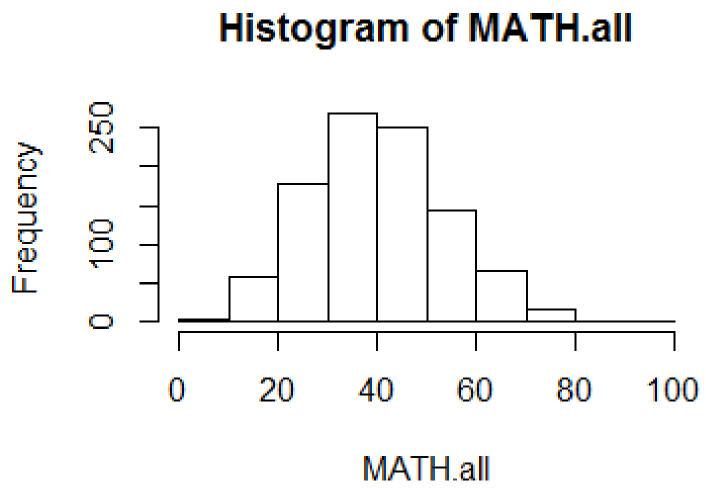
3.2 Metagene calculation from RNA-Seq expression

```
# load RNAseq data of 304 genes for 208 tnbc samples
# RNAseq data of 1218 TCGA BRCA can be downloaded from UCSC Xena browser
(https://tcga.xenahubs.net/download/TCGA.BRCA.sampleMap/HiSeqV2)

n304genes <- read.table("n208tnbc_n304genes_RNAseq.csv", header=TRUE,
sep=";")

# scale transposed expression data and re-transpose
n304.expr.sca= t(scale(t(n304genes[,5:212])))
colnames(n304.expr.sca)=colnames(n304genes[,5:212])

# calculate mean expression of each metag-cluster from scaled expression for
17 metagenes
metag17=array(NA,dim=c(0,17))
for (i in 1: ncol(n304.expr.sca)) {
  mdf= as.data.frame(as.list(by(n304.expr.sca[,i],
                             n304genes$MetagCluster17, mean)))
  rownames(mdf)=colnames(n304.expr.sca)[i]
  metag17=rbind(metag17, mdf)
}
rm(mdf)

# merge 17 metagene expression data with tnbc.data dataframe, left outer
join:
```

```
tnbc.data.meta17= merge(tnbc.data, metag17, by="row.names", all.x =TRUE)
# "merge" command results in resorting of dataframe and loss of row.names
#    but an additional new first column "Row.names"
# Assign new row.names from this additional column and then delete it
row.names(tnbc.data.meta17)=tnbc.data.meta17$Row.names
tnbc.data.meta17=tnbc.data.meta17[,colnames(tnbc.data.meta17)!= "Row.names"]
```

SECTION-4 MATH analysis of dispersion in mutant allele frequencies

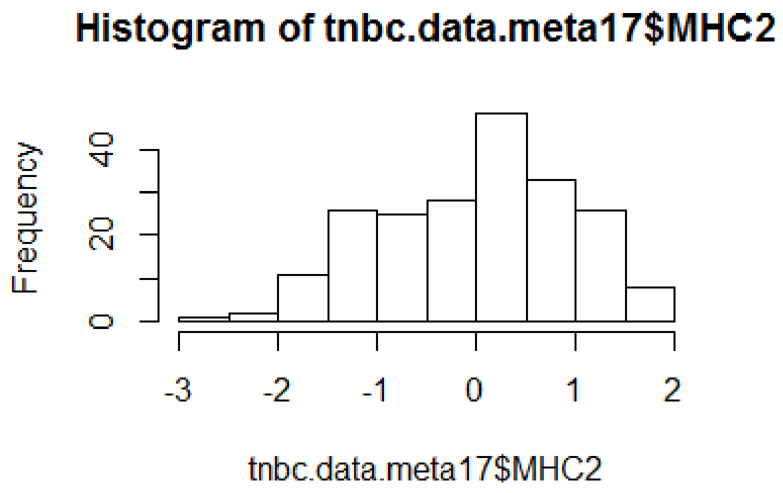
```
# Copy of maf file from TCGA
genome.wustl.edu_BRCA.IlluminaGA_DNASeq.Level_2.1.1.0.curated.somatic.maf.txt
(52MB) is available at https://portal.gdc.cancer.gov/legacy-
archive/files/50d6fb1d-5bb1-4a30-9e91-6d45bd9b1c3f

# The required variant allele frequencies have been extracted in the smaller
file used here: "VAF-
table_genome.wustl.edu_BRCA.IlluminaGA_DNASeq.Level_2.1.1.0.curated.somatic.m
af.txt""

maf.download <- read.delim(
    "VAF-
table_genome.wustl.edu_BRCA.IlluminaGA_DNASeq.Level_2.1.1.0.curated.somatic.m
af.txt")

all.maf = maf.download[,c("Hugo_Symbol", "Tumor_Sample_Barcode",
"tumor_vaf")]

TCGA_Sample=substr(all.maf$Tumor_Sample_Barcode, 1, 15)

all.maf = cbind(TCGA_Sample, all.maf)

# calculate for each sample the median of tumor_vaf values
med=by(all.maf$tumor_vaf, all.maf$TCGA_Sample, median)

# convert list to dataframe and transpose
med.df = t(as.data.frame(as.list(med)))
colnames(med.df)= "med.mut.AF"

# calculate MAD (Median Absolute Deviation) for each sample
MAD=by(all.maf$tumor_vaf, all.maf$TCGA_Sample, mad)

# convert list to dataframe and transpose
MAD.df= t(as.data.frame(as.list(MAD)))
colnames(MAD.df)= "MAD.mut.AF"

# calculate MATH (Mutant Allele Tumor Heterogeneity) as MATH=100*MAD/median
MATH.all =100 * MAD.df / med.df
colnames(MATH.all)= "MATH"

hist(MATH.all)
```



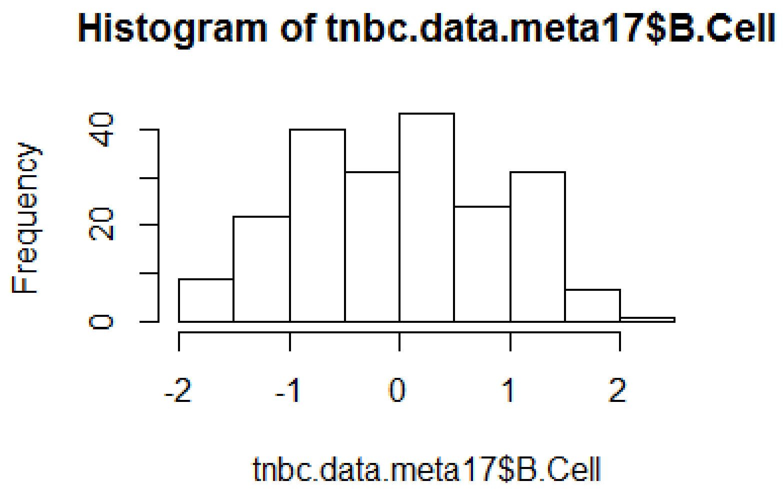
```
# Export MATH values:

# write.table(MATH.all, file="n982TCGA_MATH.txt",
#        row.names=TRUE, col.names = NA, quote=FALSE, sep="\t")
```

SECTION-5 Survival analysis

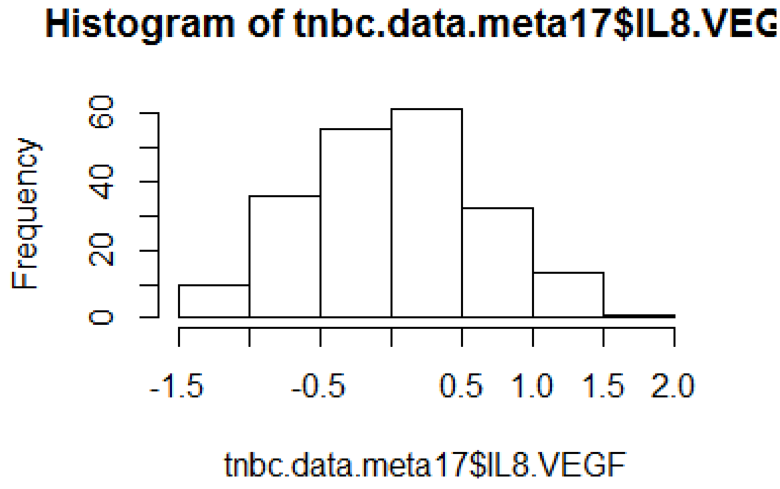
```
library("survival")

# Censor DFS at 120 months
dfs.120=tnbc.data.meta17$DFS_MONTHS
ev.120=tnbc.data.meta17$DFS_STATUS

for (i in 1:nrow(tnbc.data.meta17)) {
     if (is.na(tnbc.data.meta17$DFS_MONTHS[i]))
       {dfs.120[i]=NA ; ev.120[i]=NA}
       else
         { if (tnbc.data.meta17$DFS_MONTHS[i] > 120)
           {dfs.120[i]=120 ; ev.120[i]="DiseaseFree"}
           else {dfs.120[i]=tnbc.data.meta17$DFS_MONTHS[i] ;
ev.120=tnbc.data.meta17$DFS_STATUS}
         }
     }

# Add censored DFS to dataframe
tnbc.data.meta17=cbind(tnbc.data.meta17, dfs.120, ev.120)

# Distributions of MHC2 metagene, B-Cell metagen, and IL8VEGF metagene
hist(tnbc.data.meta17$MHC2)
```



```r
hist(tnbc.data.meta17$B.Cell)
```



```r
hist(tnbc.data.meta17$IL8.VEGF)
```



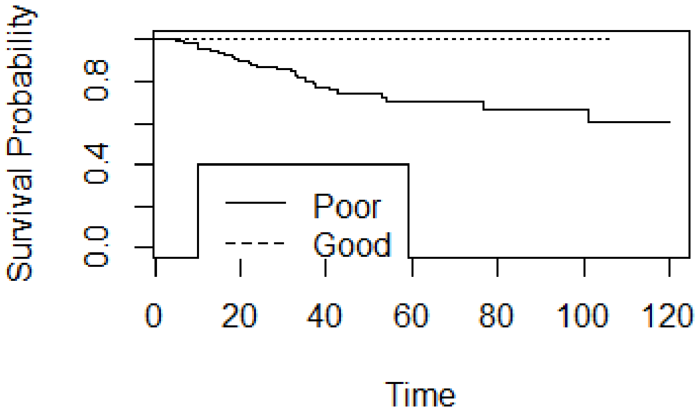
```
# Since no clear bimodality observed in distributions,
# we stay with previously established cutoffs for metagenes/signatures:
# MHC2 metagene: Upper quartile (Rody 2009, PMID 19272155)
# B-Cell metagene: Lower quartile (Rody 2011, PMID 21978456)
# IL8.VEGF metagene: Median split (Rody 2011, PMID 21978456)
```

5.1 MHC2/IL8VEGF signature

```
# Define upper quartile MHC2 metagene (based on Rody 2009, PMID 19272155)
MHC2.q4=tnbc.data.meta17$MHC2 > quantile(tnbc.data.meta17$MHC2, probs=0.75)
# Define below median IL8.VEGF metagene (cutoff from Rody 2011, PMID
21978456)
IL8.VEGF.q12=tnbc.data.meta17$IL8.VEGF < quantile(tnbc.data.meta17$IL8.VEGF,
probs=0.5)
# Define prognostic signature
MHC2.IL8.VEGF.sig = MHC2.q4 & IL8.VEGF.q12


## Check MHC2.IL8.VEGF.sig in Survival analysis
time=tnbc.data.meta17$dfs.120
censor= (tnbc.data.meta17$ev.120 =="Recurred/Progressed")
strata= MHC2.IL8.VEGF.sig
test=survfit(Surv(time, censor)~strata,conf.type="none")
summary(test)
```

```
## Call: survfit(formula = Surv(time, censor) ~ strata, conf.type = "none")
##
## 14 observations deleted due to missingness
##                 strata=FALSE
##    time n.risk n.event survival std.err
##    5.09    151       1    0.993 0.00660
##    6.80    149       1    0.987 0.00933
##    7.79    145       1    0.980 0.01149
##    9.89    138       1    0.973 0.01342
##    10.02    135       1    0.966 0.01513
##    10.28    134       1    0.958 0.01665
##    12.55    128       1    0.951 0.01812
##    12.71    126       1    0.943 0.01949
##    14.98    113       1    0.935 0.02103
##    16.10    109       1    0.926 0.02252
##    18.27    103       1    0.917 0.02403
##    18.50    102       1    0.908 0.02542
##    19.32     99       1    0.899 0.02677
##    21.91     89       1    0.889 0.02831
##    22.40     88       1    0.879 0.02974
##    23.95     82       1    0.868 0.03125
##    28.22     74       1    0.857 0.03295
##    31.90     69       1    0.844 0.03474
##    32.65     67       1    0.832 0.03643
##    33.31     63       1    0.818 0.03817
##    35.22     57       1    0.804 0.04011
##    36.79     53       1    0.789 0.04212
##    37.32     52       1    0.774 0.04396
##    40.70     47       1    0.757 0.04600
##    42.81     44       1    0.740 0.04807
##    53.02     37       1    0.720 0.05076
##    53.88     36       1    0.700 0.05315
##    76.54     21       1    0.667 0.06017
##   101.05     11       1    0.606 0.07957
##
##                  strata=TRUE
##       time n.risk n.event survival std.err
```

```
plot(test, lty=c(1,3), xlab="Time", ylab="Survival Probability")
legend(10, 0.4, c("Poor", "Good") , lty=c(1,2))
```

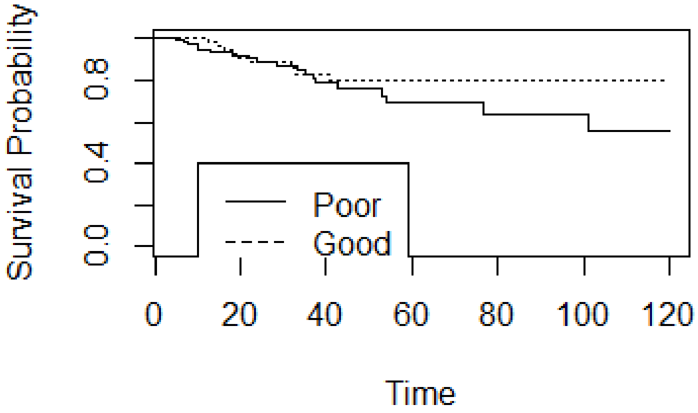


5.2 B-Cell/IL8VEGF signature

```
# Define B-Cell metagene above lowest quartile (cutoff from Rody 2011, PMID
21978456)
B.Cell.q234=tnbc.data.meta17$B.Cell > quantile(tnbc.data.meta17$B.Cell,
probs=0.25)
# Define below median IL8.VEGF metagene (cutoff from Rody 2011, PMID
21978456)
IL8.VEGF.q12=tnbc.data.meta17$IL8.VEGF < quantile(tnbc.data.meta17$IL8.VEGF,
probs=0.5)
# Define prognostic signature
B.Cell.IL8.VEGF.sig = B.Cell.q234 & IL8.VEGF.q12


## Check B.Cell.IL8.VEGF.sig in Survival analysis
time=tnbc.data.meta17$dfs.120
censor= (tnbc.data.meta17$ev.120 =="Recurred/Progressed")
strata= B.Cell.IL8.VEGF.sig
test=survfit(Surv(time, censor)~strata,conf.type="none")
summary(test)
```

```
## Call: survfit(formula = Surv(time, censor) ~ strata, conf.type = "none")
##
## 14 observations deleted due to missingness
##                 strata=FALSE
##    time n.risk n.event survival std.err
##    5.09    108       1    0.991 0.00922
##    6.80    106       1    0.981 0.01303
##    7.79    102       1    0.972 0.01607
##    9.89     97       1    0.962 0.01877
##   10.02     95       1    0.952 0.02113
##   10.28     94       1    0.942 0.02320
##   12.71     89       1    0.931 0.02524
##   18.27     71       1    0.918 0.02808
##   21.91     62       1    0.903 0.03129
##   23.95     58       1    0.887 0.03440
##   28.22     52       1    0.870 0.03774
##   33.31     45       1    0.851 0.04156
##   35.22     41       1    0.830 0.04544
##   36.79     37       1    0.808 0.04944
##   37.32     36       1    0.785 0.05292
##   42.81     29       1    0.758 0.05761
##   53.02     22       1    0.724 0.06448
##   53.88     21       1    0.689 0.07002
##   76.54     13       1    0.636 0.08230
##  101.05      8       1    0.557 0.10355
##
##                 strata=TRUE
##  time n.risk n.event survival std.err
##  12.6     61       1    0.984  0.0163
##  15.0     54       1    0.965  0.0241
##  16.1     52       1    0.947  0.0299
##   18.5     49       1    0.928  0.0350
##   19.3     45       1    0.907  0.0398
##   22.4     43       1    0.886  0.0441
##   31.9     34       1    0.860  0.0499
##   32.6     32       1    0.833  0.0551
##   40.7     27       1    0.802  0.0611
```

```
plot(test, lty=c(1,3), xlab="Time", ylab="Survival Probability")
legend(10, 0.4, c("Poor", "Good") , lty=c(1,2))
```



```
dir()

## [1] "2016_05_31_median_mean_n208RNASeq_vs_FumagalliCorrel.txt"
## [2] "2016_06_01_TNBC-metagenes_gene_list.txt"
## [3] "n208tnbc_n304genes_RNAseq.csv"
## [4] "TNBC_TIL_analysis_2017_05_18.Rmd"
## [5] "TNBC_TIL_analysis_2017_05_18_files"
## [6] "VAF-
table_genome.wustl.edu_BRCA.IlluminaGA_DNASeq.Level_2.1.1.0.curated.somatic.m
af.txt"

sessionInfo()

## R version 3.3.2 (2016-10-31)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 14393)
##
## locale:
## [1] LC_COLLATE=German_Germany.1252  LC_CTYPE=German_Germany.1252
## [3] LC_MONETARY=German_Germany.1252 LC_NUMERIC=C
## [5] LC_TIME=German_Germany.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
```

```
## other attached packages:
## [1] survival_2.40-1 hexbin_1.27.1   cgdsr_1.2.5
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.9       lattice_0.20-34   digest_0.6.12
##  [4] rprojroot_1.2     R.methodsS3_1.7.1 grid_3.3.2
##  [7] backports_1.0.5   magrittr_1.5      evaluate_0.10
## [10] stringi_1.1.2     R.oo_1.21.0       Matrix_1.2-8
## [13] rmarkdown_1.3     splines_3.3.2     tools_3.3.2
## [16] stringr_1.2.0     yaml_2.1.14       htmltools_0.3.5
## [19] knitr_1.15.1
```