survival_analysis

TCGA survival analysis using Gene Expression data

Load library

```
library('survival')
library('survminer')
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ggpubr
```

```
##
## Attaching package: 'survminer'
```

```
## The following object is masked from 'package:survival':
##
##     myeloma
```

```
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(knitr) # For rendering purpose
```

Load Gene Expression cohort data

Load the Gene expression data. Each row corresponds to gene while each column corresponds to patients.

```
data= read.table("GeneExpression_log2Counts_TNBC_HER2pos_ERpos_Normal_normalized.txt", sep="\t", header=
kable(data[1:3,1:4])
```

GeneName	TCGA-E2-A15E-06A-11R-A12D-07	TCGA-D8-A1JM-01A-11R-A13Q-07	TCGA-AN-A0AT-01A-11R-A034-07
ABCC8	7.940489	3.150607	-0.4084395
CHRDL2	3.150412	11.307954	6.1765230
SERPINB3	0.000000	6.747542	5.0178252

Load sample information data

This file contains information of each patient,

```
sampleinfo = read.table("sample_group_info.txt", sep="\t", header=T,check.names=F, stringsAsFactor=F)
rownames(sampleinfo) = sampleinfo$sampleName;
sampleinfo$patientid = substr(sampleinfo$sampleName,1, 12);
# Print 1st sample information
print(sampleinfo[1,])
```

```
##                                                sampleName stageCode
## TCGA-E2-A15E-06A-11R-A12D-07 TCGA-E2-A15E-06A-11R-A12D-07  StageIIA
##                                             EPHsubtype menopauseStatus
## TCGA-E2-A15E-06A-11R-A12D-07 ERPositivePRPositiveHER2             Pre
##                                        ethnicity  race groupName    patientid
## TCGA-E2-A15E-06A-11R-A12D-07 NOTHISPANICORLATINO WHITE     ERpos TCGA-E2-A15E
```

Stage summary

```
table(sampleinfo$stageCode)
```

```
##
## NotApplicable        StageI       StageIA       StageIB       StageII
##           112            84            72             6             5
##      StageIIA      StageIIB     StageIII     StageIIIA     StageIIIB
##           300           235             1           138            22
##     StageIIIC       StageIV        StageX       Unknown
##            59            17             9            10
```

EPHsubtype

```
table(sampleinfo$groupName)
```

```
##
##   ERpos HER2pos  Normal    TNBC
##     805      37     112     116
```

Load Clinical Information of each patient

```
clinical = read.table("TCGA-BRCA_clinical.csv", sep=",", header=T,check.names=F, stringsAsFactor=F)
rownames(clinical) = clinical$submitter_id;
print(clinical[1,])
```

```
##              submitter_id classification_of_tumor last_known_disease_status
## TCGA-3C-AAAU TCGA-3C-AAAU            not reported              not reported
##                             updated_datetime primary_diagnosis tumor_stage
## TCGA-3C-AAAU 2016-09-02T19:08:49.101859-05:00             c50.9     stage x
##              age_at_diagnosis vital_status morphology days_to_death
## TCGA-3C-AAAU            20211        alive     8520/3            NA
##              days_to_last_known_disease_status days_to_last_follow_up state
## TCGA-3C-AAAU                                NA                    4047    NA
##              days_to_recurrence                          diagnosis_id
## TCGA-3C-AAAU                 NA 8cfb8afb-b915-5255-865b-a5923f47b351
##               tumor_grade tissue_or_organ_of_origin days_to_birth
## TCGA-3C-AAAU not reported                     c50.9        -20211
##              progression_or_recurrence prior_malignancy
## TCGA-3C-AAAU              not reported     not reported
##              site_of_resection_or_biopsy created_datetime cigarettes_per_day
## TCGA-3C-AAAU                      c50.9               NA                  NA
##              weight alcohol_history alcohol_intensity bmi years_smoked height
## TCGA-3C-AAAU     NA              NA                NA  NA           NA     NA
##                                    exposure_id gender year_of_birth  race
## TCGA-3C-AAAU 72f0be98-dffa-5d35-88fe-f9ca774d6db0 female          1949 white
##                                   demographic_id              ethnicity
## TCGA-3C-AAAU cee0a94c-1d9e-5650-a500-a6b021fe138d not hispanic or latino
##              year_of_death bcr_patient_barcode disease
## TCGA-3C-AAAU            NA        TCGA-3C-AAAU    BRCA
```

Map TCGA patiend id to EPH status (ERpos/TNBC/HER2pos/Normal)

```
ephstatus = sampleinfo[colnames(data)[2:ncol(data)],"groupName"];
names(ephstatus) = colnames(data)[2:ncol(data)];
print(head(ephstatus))
```

```
## TCGA-E2-A15E-06A-11R-A12D-07 TCGA-D8-A1JM-01A-11R-A13Q-07
##                       "ERpos"                       "ERpos"
## TCGA-AN-A0AT-01A-11R-A034-07 TCGA-EW-A3E8-01B-11R-A24H-07
##                        "TNBC"                       "ERpos"
## TCGA-BH-A0E0-01A-11R-A056-07 TCGA-D8-A1JS-01A-11R-A13Q-07
##                        "TNBC"                       "ERpos"
```

Map clinical data with gene expression data

```
# Consider gene1
rowno=1;
```

```
# All sample names
samplename = colnames(data)[2:ncol(data)]


# All patient names
patientname = sampleinfo[samplename,"patientid"];

# Clinical data
tempclindata = clinical[,c("submitter_id","days_to_death","days_to_last_follow_up","vital_status")];
tempclindata = tempclindata[patientname,];
tempclindata$ephstatus = ephstatus[samplename];
tempclindata$genexp = unlist(data[rowno,samplename]);

notDead <- is.na(tempclindata$days_to_death)

if (any(notDead == TRUE)) {
       tempclindata$days_to_death[notDead] <- tempclindata[notDead, "days_to_last_follow_up"]
}

tempclindata$s <- grepl("dead", tempclindata$vital_status, ignore.case = TRUE)
tempclindata$ephstatus <- as.factor(tempclindata$ephstatus)
head(tempclindata)
```

```
##              submitter_id days_to_death days_to_last_follow_up vital_status
## TCGA-E2-A15E TCGA-E2-A15E           630                    630        alive
## TCGA-D8-A1JM TCGA-D8-A1JM           590                    590        alive
## TCGA-AN-A0AT TCGA-AN-A0AT            10                     10        alive
## TCGA-EW-A3E8 TCGA-EW-A3E8          1035                   1035        alive
## TCGA-BH-A0E0 TCGA-BH-A0E0           134                    134        alive
## TCGA-D8-A1JS TCGA-D8-A1JS           371                    371        alive
##              ephstatus     genexp     s
## TCGA-E2-A15E     ERpos  7.9404889 FALSE
## TCGA-D8-A1JM     ERpos  3.1506068 FALSE
## TCGA-AN-A0AT      TNBC -0.4084395 FALSE
## TCGA-EW-A3E8     ERpos 11.0814223 FALSE
## TCGA-BH-A0E0      TNBC  6.0500304 FALSE
## TCGA-D8-A1JS     ERpos 12.9273611 FALSE
```

Fit Proportional Hazards Regression Model

```
res.cox <- coxph(Surv(days_to_death,s) ~ ephstatus + genexp, data =  tempclindata)
summary(res.cox)
```

```
## Call:
## coxph(formula = Surv(days_to_death, s) ~ ephstatus + genexp,
##     data = tempclindata)
##
##   n= 1069, number of events= 172
##    (1 observation deleted due to missingness)
##
```

```
##                     coef exp(coef) se(coef)     z Pr(>|z|)
## ephstatusHER2pos 0.71723   2.04875  0.43440 1.651   0.0987 .
## ephstatusNormal  1.04443   2.84178  0.18411 5.673 1.41e-08 ***
## ephstatusTNBC    0.61329   1.84650  0.31137 1.970   0.0489 *
## genexp           0.01827   1.01844  0.03352 0.545   0.5857
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                  exp(coef) exp(-coef) lower .95 upper .95
## ephstatusHER2pos     2.049     0.4881    0.8744     4.800
## ephstatusNormal      2.842     0.3519    1.9809     4.077
## ephstatusTNBC        1.847     0.5416    1.0030     3.399
## genexp               1.018     0.9819    0.9537     1.088
##
## Concordance= 0.587  (se = 0.028 )
## Likelihood ratio test= 29.69  on 4 df,   p=6e-06
## Wald test            = 33.22  on 4 df,   p=1e-06
## Score (logrank) test = 35.78  on 4 df,   p=3e-07
```

Do preprocessing

```
newdata=data.frame(ephstatus=unique(as.character(tempclindata$ephstatus)), genexp = sapply(unique(as.cha
head(newdata)
```

```
##         ephstatus   genexp
## ERpos       ERpos 8.511694
## TNBC         TNBC 3.098211
## Normal     Normal 7.768913
## HER2pos   HER2pos 5.412567
```

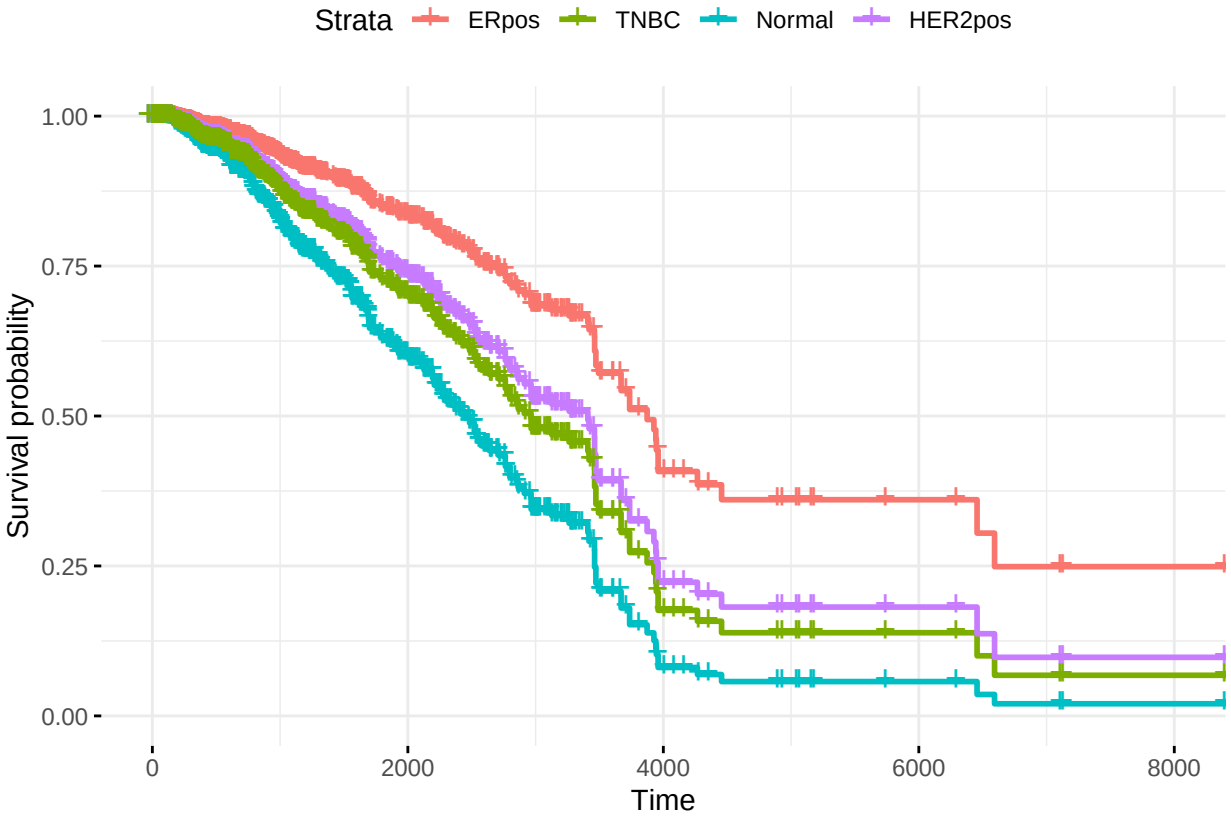
Create survival curves

```
fit <- survfit(res.cox, newdata = newdata)
print(fit)
```

```
## Call: survfit(formula = res.cox, newdata = newdata)
##
##            n events median 0.95LCL 0.95UCL
## ERpos   1069    172   3873    3492    4456
## TNBC    1069    172   3409    2573    3959
## Normal  1069    172   2469    2097    2965
## HER2pos 1069    172   2965    2192      NA
```

Drawing Survival Curves Using ggplot2

```
ggsurvplot(fit, data = newdata, conf.int = F, legend.labs=newdata$ephstatus,ggtheme = theme_minimal())
```


Strata
ERpos
TNBC
Normal
HER2pos
Survival probability
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000
8000
Time