TCGA data download

Antonio Ahn

2018-02-21

Contents

Introduction

This document outlines the approach i took to download data from level 3 and 4 TCGA data repositry. The clinical information, RNA-seq data and methylation 450K data was downloaded using the `RTCGAToolbox`(Samur, M. K. 2014) package and then processed or cleaned-up for downstream analysis.

The `RTCGAToolbox` package retreives data from the broads institutive firehose database. Detailed information on how to use this package and the included functions are available in the Vignette, online PH525Xseries courses and Youtube videos.

Load the library

```
library(RTCGAToolbox)
```

A list of available cancer types are found using the `getFirehoseDatasets` command. The run dates and analyze datas are found using the `getFirehoseRunningDates` and the `getFirehoseAnalyzeDates` commands.

```
getFirehoseDatasets()
```

```
##  [1] "ACC"      "BLCA"     "BRCA"     "CESC"     "CHOL"     "COADREAD"
##  [7] "COAD"     "DLBC"     "ESCA"     "FPPP"     "GBMLGG"   "GBM"
```

```
## [13] "HNSC"     "KICH"     "KIPAN"    "KIRC"     "KIRP"     "LAML"
## [19] "LGG"      "LIHC"     "LUAD"     "LUSC"     "MESO"     "OV"
## [25] "PAAD"     "PCPG"     "PRAD"     "READ"     "SARC"     "SKCM"
## [31] "STAD"     "STES"     "TGCT"     "THCA"     "THYM"     "UCEC"
## [37] "UCS"      "UVM"
```

```
head(getFirehoseRunningDates())
```

```
## [1] "20160128" "20151101" "20150821" "20150601" "20150402" "20150204"
```

```
head(getFirehoseAnalyzeDates())
```

```
## [1] "20160128" "20150821" "20150402" "20141017" "20140715" "20140416"
```

Clinical and RNA-seq data - download and processing

The clinical information and normalised RNA-seq data is downloaded. The methylation data is downloaded separately from the clinical and RNA-seq data because its size was too large for my computer to handle. Thus the methylation data was downloaded in the DSM3735 server.

Skin cutaneous melanoma (SKCM) is selected and "20151101" is used as the rundate. The default file size is 500 mb and this limit is extended using `fileSizeLimit`.

```
readDataMel <- getFirehoseData (dataset="SKCM", runDate="20151101",forceDownload = TRUE,
    clinical=TRUE, RNASeq2GeneNorm=TRUE, fileSizeLimit= 3000)
```

```
##
Read 0.0% of 20533 rows
Read 48.7% of 20533 rows
Read 97.4% of 20533 rows
Read 20533 rows and 474 (of 474) columns from 0.077 GB file in 00:00:12
```

The RNA-seq data (RNAseq2_Gene_Norm) contains gene expression levels generated using **MapSplice** for alignment and **RNA-Seq by Expectation-Maximization (RSEM)** for quantification. RSEM values are calculated using an algorithm that estimate abundances at the gene level to generate TPM (Transcripts Per Million) values. TPM is similar to FPKM and RPKM in that it accounts for multiple variables including library size and gene length. For normalisation, TPM values are divided by the 75th percentile (3rd quartile) and multiplied by 1000.

Extracting the data

Extract the clinical and RNA-seq data.

```
clinMel <- getData(readDataMel, "clinical")
rnaseqMel <- getData(readDataMel, "RNASeq2GeneNorm")
```

Data cleaning of clinical and RNA-seq information

The identifiers are structured differently between the clinical and RNA-seq data. The identifiers in the RNA-seq data are transformed to be the same as the ones in the clinical data. Duplicate RNA-seq data are removed and any RNA-seq without clinical information or any clinical information without RNA-seq data are removed.

Here i need to add in what the TCGA name means. For example, TCGA-3N-A9WB-06A-11R-A38C-07, what does each section mean. Is there a difference in the raw identifier names between the duplicates?

Changing patient identifier names

The identifiers are structure differently between the clinical and RNA-seq data.

```
dim(clinMel)
```

```
## [1] 470  18
```

```
head(clinMel)
```

```
##              Composite Element REF years_to_birth vital_status
## tcga.3n.a9wb                 value             71            1
## tcga.3n.a9wc                 value             82            0
## tcga.3n.a9wd                 value             82            1
## tcga.bf.a1pu                 value             46            0
## tcga.bf.a1pv                 value             74            0
## tcga.bf.a1px                 value             56            1
##              days_to_death days_to_last_followup
## tcga.3n.a9wb           518                  <NA>
## tcga.3n.a9wc          <NA>                  2022
## tcga.3n.a9wd           395                  <NA>
## tcga.bf.a1pu          <NA>                   387
## tcga.bf.a1pv          <NA>                    14
## tcga.bf.a1px           282                  <NA>
##              days_to_submitted_specimen_dx pathologic_stage
## tcga.3n.a9wb                           426         stage ia
## tcga.3n.a9wc                          1644        stage iia
## tcga.3n.a9wd                           183       stage iiia
## tcga.bf.a1pu                             0        stage iic
## tcga.bf.a1pv                             0        stage iic
## tcga.bf.a1px                             0       stage iiib
##              pathology_T_stage pathology_N_stage pathology_M_stage
## tcga.3n.a9wb               t1a                nx                m0
## tcga.3n.a9wc               t2b                nx                m0
## tcga.3n.a9wd               t2a               n1a                m0
## tcga.bf.a1pu               t4b                n0                m0
## tcga.bf.a1pv               t4b                n0                m0
## tcga.bf.a1px               t4b               n2a                m0
##              melanoma_ulceration melanoma_primary_known Breslow_thickness
## tcga.3n.a9wb                  no                   yes               0.7
## tcga.3n.a9wc                 yes                   yes               1.8
## tcga.3n.a9wd                  no                   yes              1.25
## tcga.bf.a1pu                 yes                   yes                13
## tcga.bf.a1pv                 yes                   yes                 9
## tcga.bf.a1px                 yes                   yes                12
##              gender date_of_initial_pathologic_diagnosis radiation_therapy
## tcga.3n.a9wb   male                                2012                no
## tcga.3n.a9wc   male                                2009                no
## tcga.3n.a9wd   male                                2013                no
## tcga.bf.a1pu female                                2010                no
## tcga.bf.a1pv female                                2010                no
## tcga.bf.a1px   male                                2010                no
```

```
##               race              ethnicity
## tcga.3n.a9wb white not hispanic or latino
## tcga.3n.a9wc white not hispanic or latino
## tcga.3n.a9wd white not hispanic or latino
## tcga.bf.a1pu white                  <NA>
## tcga.bf.a1pv white                  <NA>
## tcga.bf.a1px white not hispanic or latino
```

```
dim(rnaseqMel)
```

```
## [1] 20501   473
```

```
rnaseqMel[1:5,1:5]
```

```
##       TCGA-3N-A9WB-06A-11R-A38C-07 TCGA-3N-A9WC-06A-11R-A38C-07
## A1BG                      381.0662                      195.1822
## A1CF                        0.0000                        0.0000
## A2BP1                       0.0000                        0.0000
## A2LD1                     250.1979                      160.7548
## A2ML1                       7.2698                        0.0000
##       TCGA-3N-A9WD-06A-11R-A38C-07 TCGA-BF-A1PU-01A-11R-A18S-07
## A1BG                      360.8794                      176.3994
## A1CF                        0.7092                        0.0000
## A2BP1                       6.3830                        1.2987
## A2LD1                      97.1986                      163.2338
## A2ML1                       0.0000                        7.7922
##       TCGA-BF-A1PV-01A-11R-A18U-07
## A1BG                      216.8470
## A1CF                        0.0000
## A2BP1                       0.0000
## A2LD1                      60.8727
## A2ML1                       0.5977
```

The identifiers in the RNA-seq data are transformed to be the same as the ones in the clinical data.

```
rid = tolower(substr(colnames(rnaseqMel),1,12))
rid = gsub("-", ".", rid)

table(rid %in% rownames(clinMel)) #all 473 RNA-seqMel samples have corresponding clinical details
```

```
##
## TRUE
##  473
```

```
length(intersect(rid,rownames(clinMel)))
```

```
## [1] 469
```

```
# 469 patients out of 470 have RNA-seq data

colnames(rnaseqMel) = rid
head(colnames(rnaseqMel))
```

```
## [1] "tcga.3n.a9wb" "tcga.3n.a9wc" "tcga.3n.a9wd" "tcga.bf.a1pu"
## [5] "tcga.bf.a1pv" "tcga.bf.a1px"
```

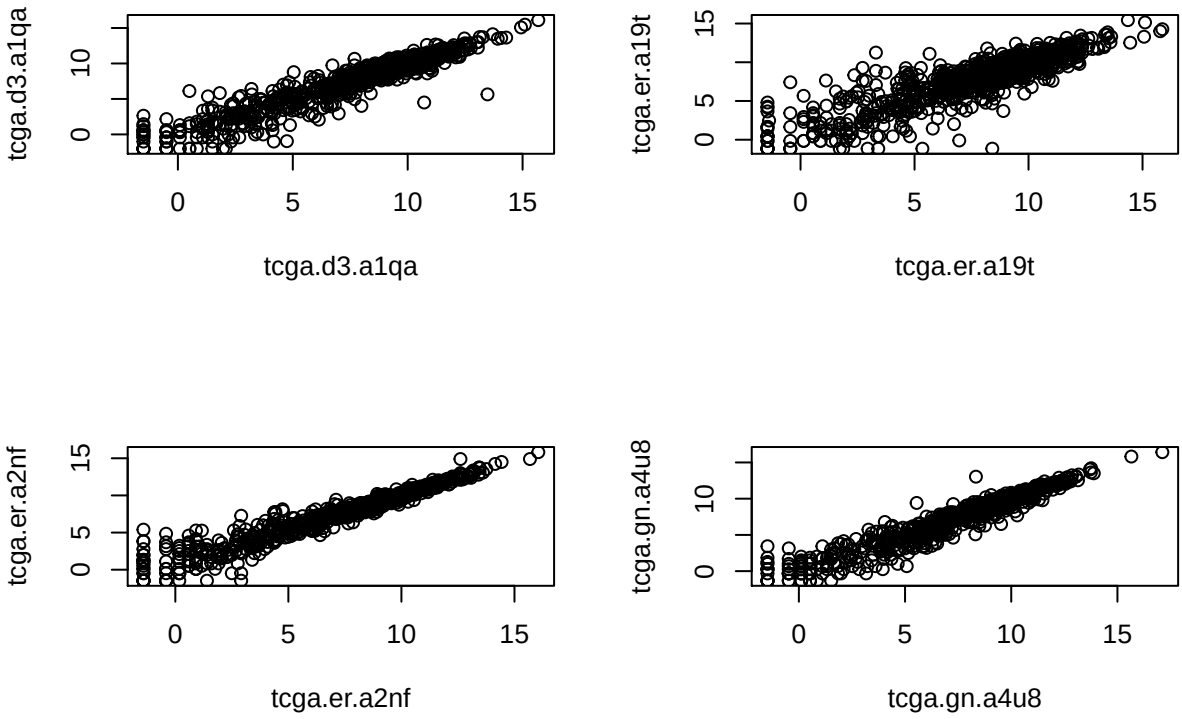
Remove duplicated samples

Samples with duplicated names are removed. These are samples that have 2 RNA-seq data for some reason. The data between the replicates are very similar and thus we remove the second duplicate.

```
duplicatedSamples <- which(duplicated(colnames(rnaseqMel))) # 4  duplicate samples

duplicatedSampleNames<-colnames(rnaseqMel)[duplicated(colnames(rnaseqMel))]

rnaseqMel_duplicated <-rnaseqMel[,colnames(rnaseqMel) %in% duplicatedSampleNames]

colnames(rnaseqMel_duplicated)
```

```
## [1] "tcga.d3.a1qa" "tcga.d3.a1qa" "tcga.er.a19t" "tcga.er.a19t"
## [5] "tcga.er.a2nf" "tcga.er.a2nf" "tcga.gn.a4u8" "tcga.gn.a4u8"
```

```
par(mfrow=c(2,2))
plot(log2(rnaseqMel_duplicated[1001:2000,1:2]))
plot(log2(rnaseqMel_duplicated[1001:2000,3:4]))
plot(log2(rnaseqMel_duplicated[1001:2000,5:6]))
plot(log2(rnaseqMel_duplicated[1001:2000,7:8]))
```




```
# it is not obvious which of the duplicates to keep, so we drop the second
```

```
rnaseqMel = rnaseqMel[,-which(duplicated(colnames(rnaseqMel)))] # getting rid of the duplicate

dim(rnaseqMel) # from 473 samples to 469
```

```
## [1] 20501   469
```

```
length(intersect(colnames(rnaseqMel),rownames(clinMel)))
```

```
## [1] 469
```

```
length(rownames(clinMel)) # there is 1 sample in clinMel which there is absent in rnaseqMel
```

```
## [1] 470
```

```
clinMel <-clinMel[intersect(colnames(rnaseqMel),rownames(clinMel)),]
dim(clinMel)
```

```
## [1] 469  18
```

```
table(colnames(rnaseqMel)==rownames(clinMel)) # patient names are in the same order
```

```
##
## TRUE
##  469
```

Create an expression set

An expression set is created to store the log2 transformed RNA-seq data and the clinical information.

```
library(Biobase)
```

```
readES = ExpressionSet(as.matrix(log2(rnaseqMel+1)))
readES
```

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 20501 features, 469 samples
##   element names: exprs
## protocolData: none
## phenoData: none
## featureData: none
## experimentData: use 'experimentData(object)'
## Annotation:
```

```
exprs(readES)[1:3,1:3]
```

```
##       tcga.3n.a9wb tcga.3n.a9wc tcga.3n.a9wd
## A1BG      8.577679      7.61605    8.4993652
## A1CF      0.000000      0.00000    0.7733212
## A2BP1     0.000000      0.00000    2.8842072
```

```
pData(readES) = clinMel
```

Survival data clean-up and analysis

Survival analysis: background

Important pData information for survival analysis is "vital_status", "days_to_death" and "days_to_last_followup"

Information from here

To analyse overall survival, 3 variables in the `clinMel` data set is required, which are "vital_status", "days_to_death" and "days_to_last_followup".

```
dim(clinMel)
```

```
## [1] 469  18
```

```
str(clinMel[,c("vital_status","days_to_death","days_to_last_followup")])
```

```
## 'data.frame':    469 obs. of  3 variables:
##  $ vital_status         : chr  "1" "0" "1" "0" ...
##  $ days_to_death        : chr  "518" NA "395" NA ...
##  $ days_to_last_followup: chr  NA "2022" NA "387" ...
```

```
clinMel[1:5,c("vital_status","days_to_death","days_to_last_followup")]
```

```
##              vital_status days_to_death days_to_last_followup
## tcga.3n.a9wb            1           518                  <NA>
## tcga.3n.a9wc            0          <NA>                  2022
## tcga.3n.a9wd            1           395                  <NA>
## tcga.bf.a1pu            0          <NA>                   387
## tcga.bf.a1pv            0          <NA>                    14
```

- **vital_status:** "1" means deceased and "0" means still alive.
- **days_to_death:** With patients who are deacesed, the days_to_death variable gives the number of days before death.
- **days_to_last_followup:** With patients who are still alive, the days_to_last_followup variable gives the number of days before the last follow-up.

Survival data: Exploratory analysis

For most patients (391 patients), the days_to_death and days_to_last_followup are mutually exlcusive; if theres an NA in days_to_death then there is a number to DaystoLastfollowup and vice versa.

```
table(!is.na(clinMel[,"days_to_death"]) & is.na(clinMel[,"days_to_last_followup"]))
```

```
##
## FALSE  TRUE
##   318   151
```

```
table(is.na(clinMel[,"days_to_death"]) & !is.na(clinMel[,"days_to_last_followup"]))
```

```
##
## FALSE  TRUE
##   229   240
```

However there are some patients with both days_to_last_followup and days_to_death (69 patients). Also there are patients with both of these variables as NA (9 patients).

```
survivalVariables <- c("days_to_last_followup","vital_status","days_to_death")

index <- !is.na(clinMel$"days_to_death") & !is.na(clinMel$"days_to_last_followup")
clinMel[index,survivalVariables]
```

```
##              days_to_last_followup vital_status days_to_death
## tcga.d3.a2jn                  1709            1          2022
## tcga.d3.a8gm                  2897            1          3259
## tcga.d9.a1jx                   195            1           216
## tcga.d9.a3z1                   345            1           468
## tcga.d9.a3z4                   119            1           519
## tcga.d9.a4z2                    93            1           190
## tcga.d9.a4z6                   338            1           561
## tcga.da.a1hw                   820            0          1096
## tcga.da.a1i0                   594            1           620
```

```
## tcga.da.a1i2 5088 1 5370
## tcga.da.a1i4 823 1 1093
## tcga.da.a1i8 1368 1 1640
## tcga.da.a1ia 1887 1 2005
## tcga.da.a1ib 825 0 1235
## tcga.da.a1ic 1926 1 2071
## tcga.da.a3f2 1025 1 1032
## tcga.da.a3f3 151 1 319
## tcga.da.a3f5 6826 1 6873
## tcga.da.a95y 302 1 430
## tcga.eb.a3y7 0 1 326
## tcga.eb.a42y 440 1 721
## tcga.eb.a44n 45 1 205
## tcga.eb.a44r 309 1 315
## tcga.eb.a4iq 414 1 636
## tcga.eb.a4p0 -2 1 326
## tcga.eb.a550 6 1 264
## tcga.eb.a57m 399 1 472
## tcga.eb.a5fp 6 1 454
## tcga.eb.a5kh 543 1 619
## tcga.eb.a5se 0 1 401
## tcga.eb.a5sf 0 1 369
## tcga.eb.a5vu 12 1 321
## tcga.eb.a6qz -3 1 352
## tcga.eb.a6r0 467 1 608
## tcga.ee.a29b 2452 1 2588
## tcga.ee.a29c 1455 1 2402
## tcga.ee.a29q 1136 1 2030
## tcga.ee.a29s 1701 1 1864
## tcga.ee.a2gd 9568 1 10346
## tcga.ee.a2gj 2717 1 3266
## tcga.ee.a2gn 2767 1 3106
## tcga.ee.a2gr 435 1 1301
## tcga.ee.a2gs 1691 1 2470
## tcga.ee.a2ml 6176 1 6590
## tcga.ee.a3ad 112 1 875
## tcga.ee.a3ag 714 1 1265
## tcga.ee.a3ji 4504 1 4648
## tcga.er.a19j 196 1 196
## tcga.er.a2nb 486 1 857
## tcga.er.a2nf 498 1 877
## tcga.er.a2ng 951 1 1490
## tcga.er.a3et 2443 1 2829
## tcga.er.a3ev 1429 1 1429
## tcga.er.a42k 206 1 394
## tcga.fr.a726 263 1 305
## tcga.fr.a8yd 896 1 1103
## tcga.fs.a1ze 1225 1 1413
## tcga.fs.a4fc 1504 1 1655
## tcga.fs.a4fd 2369 1 2454
## tcga.fw.a3tu 1446 1 1691
## tcga.gn.a26d 1204 1 1460
## tcga.gn.a4u7 266 1 317
## tcga.gn.a4u9 384 1 673
```

```
## tcga.od.a75x                  8966            1          9061
## tcga.we.a8k5                  1654            1          1860
## tcga.we.a8zr                   133            1           274
## tcga.we.a8zy                  1330            1          1506
## tcga.xv.aazw                    18            1           393
## tcga.yg.aa3o                  1096            1          1154
```

```
dim(clinMel[index,survivalVariables])
```

```
## [1] 69  3
```

```
index <- is.na(clinMel[,"days_to_death"]) & is.na(clinMel[,"days_to_last_followup"])
clinMel[index,survivalVariables]
```

```
##              days_to_last_followup vital_status days_to_death
## tcga.d3.a3c1                  <NA>            0          <NA>
## tcga.d3.a3c3                  <NA>            0          <NA>
## tcga.d3.a51g                  <NA>            0          <NA>
## tcga.d3.a8go                  <NA>            1          <NA>
## tcga.er.a19o                  <NA>            1          <NA>
## tcga.fr.a3yo                  <NA>            0          <NA>
## tcga.rp.a695                  <NA>            0          <NA>
## tcga.rp.a6k9                  <NA>            0          <NA>
## tcga.yd.a9tb                  <NA>            0          <NA>
```

```
dim(clinMel[index,survivalVariables])
```

```
## [1] 9 3
```

There are also some patients with a negative days_to_last_followup. What does this mean?

```
survivalVariables <- c("days_to_last_followup","vital_status","days_to_death")

index <- which(clinMel[,"days_to_death"] < 0 | clinMel[,"days_to_last_followup"] < 0)

clinMel[index,survivalVariables]
```

```
##              days_to_last_followup vital_status days_to_death
## tcga.eb.a430                    -2            0          <NA>
## tcga.eb.a4p0                    -2            1           326
## tcga.eb.a6qz                    -3            1           352
```

Survival analysis: merge days_to_death and days_to_last_followup

Here i merge days_to_death and days_to_last_followup to create a new variable called new_death. Most are simple to handle because they are mutually exlcusive; if there's an NA in days_to_death then there is a number to days_to_last_followup and vice versa. However, as shown above, some patients have values to both variables with different number of days which i am unsure what that means. Also some patients have an NA to both variables.

Here i create a new variable called new_death.

- If patient has deceased (1 in vital status), the days_to_death is selected
- If patient is alive (0 in vital status), days_to_last_followup is selected

```
mergeOS <- ifelse(clinMel[,"vital_status"]==1, clinMel[,"days_to_death"], clinMel[,"days_to_last_follow

summary(mergeOS)
```

```
##    Length     Class      Mode
##       469 character character
```

```
clinMel$mergeOS <- as.numeric(mergeOS)
```

clinMel with the `mergeOS` parameter is re-loaded into `readES`.

```
pData(readES) = clinMel
```

Survival analysis: sanity check with t-stage

- t0 - patients without a known primary tumor.
- t1
- t2
- t3
- t4
- ti
- tx

```
library(survival)
```

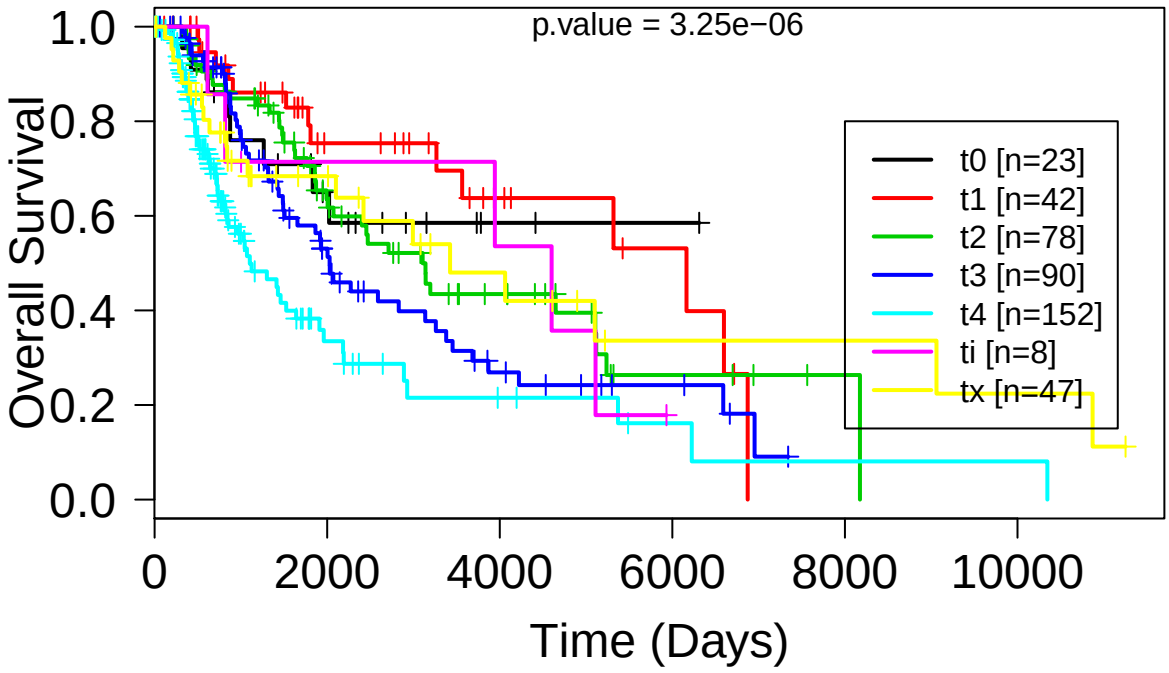
```
ev <- as.numeric(pData(readES)$vital_status)
fut <-as.numeric(pData(readES)$mergeOS)
su = Surv(fut, ev)

table(pData(readES)$pathology_T_stage)
```

```
##
##  t0  t1 t1a t1b  t2 t2a t2b  t3 t3a t3b  t4 t4a t4b tis  tx
##  23  10  22  10  32  31  15  14  39  37  15  25 112   8  47
```

```
table(substr(clinMel$pathology_T_stage,1,2))
```

```
##
##  t0  t1  t2  t3  t4  ti  tx
##  23  42  78  90 152   8  47
```

```
t_stage = factor(substr(clinMel$pathology_T_stage,1,2))

plot(survfit(su~t_stage),mark.time=TRUE, lwd=2, col=1:7, las=1, cex.axis=1.5)
mtext("Overall Survival", side=2, line=2.7, cex=1.5)
mtext("Time (Days)", side=1, line=2.8, cex=1.5)

ntab = table(t_stage)
ns = paste("[n=", ntab, "]", sep="")
legend(8000, .8, col=1:7, lwd=2, legend=paste(levels(t_stage), ns))
text(6000,1, paste("p.value = 3.25e-06 "))
```




```
summary(coxph(su~t_stage))
```

```
## Call:
## coxph(formula = su ~ t_stage)
##
##   n= 433, number of events= 203
##    (36 observations deleted due to missingness)
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## t_staget1 -0.2336    0.7917   0.4438 -0.526  0.59867
## t_staget2  0.2029    1.2250   0.3886  0.522  0.60157
## t_staget3  0.5013    1.6509   0.3817  1.313  0.18905
## t_staget4  1.0533    2.8671   0.3758  2.803  0.00507 **
## t_stageti  0.3598    1.4331   0.5717  0.629  0.52913
## t_stagetx  0.1951    1.2154   0.4253  0.459  0.64643
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## t_staget1    0.7917     1.2631    0.3317     1.889
## t_staget2    1.2250     0.8164    0.5719     2.624
## t_staget3    1.6509     0.6057    0.7813     3.488
## t_staget4    2.8671     0.3488    1.3726     5.989
## t_stageti    1.4331     0.6978    0.4673     4.395
## t_stagetx    1.2154     0.8228    0.5281     2.797
##
## Concordance= 0.626  (se = 0.023 )
## Rsquare= 0.074   (max possible= 0.992 )
## Likelihood ratio test= 33.1  on 6 df,   p=1.002e-05
## Wald test            = 33.51  on 6 df,   p=8.372e-06
## Score (logrank) test = 35.63  on 6 df,   p=3.251e-06
```

```
survdiff(su~t_stage)
```

```
## Call:
## survdiff(formula = su ~ t_stage)
##
## n=433, 36 observations deleted due to missingness.
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## t_stage=t0  23        8    12.39    1.5566    1.6661
## t_stage=t1  42       14    27.01    6.2702    7.3108
## t_stage=t2  76       39    48.79    1.9653    2.6200
## t_stage=t3  89       49    46.04    0.1908    0.2486
## t_stage=t4 152       68    38.33   22.9719   29.6834
## t_stage=ti   7        5     5.28    0.0148    0.0153
## t_stage=tx  44       20    25.16    1.0573    1.3201
##
##  Chisq= 35.6  on 6 degrees of freedom, p= 3.25e-06
```

There is a significant statistical difference in overall survival between the different T stages

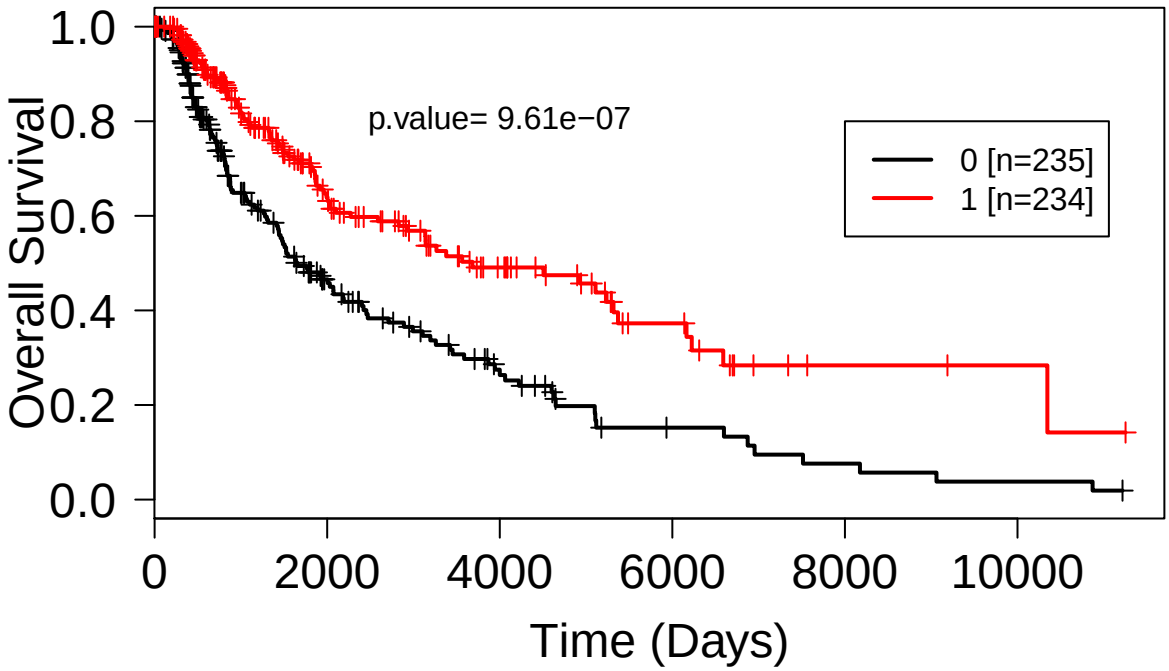
Survival analysis: sanity check with CD74

CD74 gene exprresion was found to be associated with good prognosis using TCGA data (Ekmekcioglu, S., et al 2016).

```
CD74 <- ifelse(exprs(readES)["CD74",] > median(exprs(readES)["CD74",]), 1, 0)
# higher than median is 1, lower than median is 0
CD74 <- as.factor(CD74)
table(CD74)
```

```
## CD74
##   0   1
## 235 234
```

```
ev <- as.numeric(pData(readES)$vital_status)
fut <-as.numeric(pData(readES)$mergeOS)
su = Surv(fut, ev)

plot(survfit(su~CD74),mark.time=TRUE, lwd=2, col=c("black","red"), las=1, cex.axis=1.5)
mtext("Overall Survival", side=2, line=2.7, cex=1.5)
mtext("Time (Days)", side=1, line=2.8, cex=1.5)

ntab = table(CD74)
ns = paste("[n=", ntab, "]", sep="")
legend(8000, .8, col= c("black","red"), lwd=2, legend=paste(levels(CD74), ns))
text(4000,0.8, paste("p.value= 9.61e-07"))
```




```
survdiff(su~CD74, data=clinMel)
```

```
## Call:
## survdiff(formula = su ~ CD74, data = clinMel)
##
## n=460, 9 observations deleted due to missingness.
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## CD74=0 230      133     97.2      13.2        24
## CD74=1 230       85    120.8      10.6        24
##
##  Chisq= 24  on 1 degrees of freedom, p= 9.61e-07
```

Higher CD74 gene expression is associated with a better prognosis.

Methylation 450K data - download and processing

The methylation 450K data-frame was too big (>6gb) to download or work with in my desktop (my dekstop freezes). It has 485,577 rows and 478 columns with each value having many digits. Therefore I had to use the DSM3735 server (based in the pathology department in Otago university) to download the data and then reduce the file size by lowering the number of decimal points for every beta-value. The size-reduced file was then moved to my desktop and loaded into R.

```
ssh -X aahn@dsm3735.otago.ac.nz # to login to the server

scp aahn@dsm3735.otago.ac.nz:/home/aahn/PDL1/TCGAMel1.RData /Users/antonioahn/Desktop # Move the RData
```

It is recommended to lower the worker/core usage to prevent crashing the server.

```
library(BiocParallel)
registered()
register(MulticoreParam(workers=2))
```

The methylation data are the Beta-values from the 450K methylation arrays. In the DSM3735 server, the methylation data was downloaded, extracted and reduced in size.

```
readDataMel <- getFirehoseData (dataset="SKCM", runDate="20151101",forceDownload = TRUE,
    Clinic=FALSE, RNAseq2_Gene_Norm=FALSE,Methylation = TRUE, fileSizeLimit= 3000)

me450kMel = getData(readDataMel, "Methylation",1)
```

```
dim(me450kMel)
head(colnames(me450kMel))

probeinfo <- me450kMel[,1:3] # These columns have the "Gene_Symbol", "Chromosome" and "Genomic_Coordina

write.csv(probeinfo , "probeinfo.csv")
```

```
scp aahn@dsm3735.otago.ac.nz:/home/aahn/probeinfo.csv /Users/antonioahn/Desktop
```

```
me450kMel <- me450kMel[,4:478] # dropping the first 3 columns which contains the probe info
```

Change the identifier names in the methylation data

```
rid = tolower(substr(colnames(me450kMel),1,12))
rid = gsub("-", ".", rid)

colnames(me450kMel) <- rid

which(duplicated(colnames(me450kMel)))

me450kMel <- me450kMel[,!duplicated(colnames(me450kMel))] # dropping the second duplicate samples
dim(me450kMel)


table(duplicated(colnames(me450kMel)))
# me450kMel has 470 samples but rnaseqMel has 469. There is 1 extra sample in me450kMel.

table(colnames(me450kMel)%in%colnames(rnaseqMel))

me450kMel <- me450kMel[,colnames(me450kMel)%in%colnames(rnaseqMel)] # keeping only the matching samples

table(colnames(rnaseqMel)==colnames(me450kMel)) # Everything is in the same length and order.

table(rownames(clinMel)==colnames(me450kMel))
```

Reducing the size of the methylation 450K data

```
str(me450kMel) # this shows that all the values are characters.

me450kMel <- sapply(me450kMel, as.numeric)
me450kMel_rounded <- as.matrix(round(me450kMel, digits=3)) # Round to 3 digits

save.image("/home/aahn/Bioinformatics/RDatafiles/TCGAmelanoma_methylation.RData")
```

After i reduced the size of the methylation data to generate `me450kMel_rounded`, I saved into my computer for loading.

```
load("~/Dropbox/Education/Bioinformatics/5DataAnalysis/TCGAmelanoma/Methylation/TCGAmelanoma_methylatio

dim(me450kMel_rounded)
```

```
## [1] 485577    469
```

```
dim(probeinfo)
```

```
## [1] 485577      3
```

```
class(me450kMel_rounded)
```

```
## [1] "matrix"
```

```
me450kMel_rounded[1:3,1:3]
```

```
##            tcga.3n.a9wb tcga.3n.a9wc tcga.3n.a9wd
## cg00000029        0.517        0.419        0.215
## cg00000108           NA           NA           NA
## cg00000109           NA           NA           NA
```

Acquring methylation probe values for meTIL-score

It was demonstrated that methylation probe values can be used to determine the level of CD8 immune cells within bulk tumour (Jeschke, J., et al. 2017).

Beta-values of 5 CpG probes are needed to generate the meTIL-score. Here i did not use me450kMel_rounded but used the data prior to rounding to 3 decimal points.

```
meTIL_probes <- c("cg20792833","cg20425130","cg23642747","cg12069309","cg21554552") # the 5 CpG probes

me450kMel[1:3,1:6]
```

```
        X Gene_Symbol Chromosome Genomic_Coordinate
```

1 cg00000029 RBL2 16 53468112 2 cg00000108 C3orf35 3 37459206 3 cg00000109 FNDC3B 3 171916037 TCGA.3N.A9WB.06A.11D.A38H.05 TCGA.3N.A9WC.06A.11D.A38H.05 1 0.5167 0.4193 2 NA NA 3 NA NA

```
write.csv(me450kMel[me450kMel$X%in%probes_iwant,], file="meTIL_probes.csv")
```

The "meTIL_probes.csv" file is transfered from the server to my computer and then loaded.

```
meTIL_probes <- read.csv("~/Dropbox/Education/Bioinformatics/5DataAnalysis/TCGAmelanoma/Methylation/meT

dim(meTIL_probes)
```

```
## [1]   5 478
```

```
meTIL_probe_info <- meTIL_probes[,1:3] # separating out the probe info from the probe values
meTIL_probes <- meTIL_probes[,4:478]
```

Changing identifier names and removing duplicates as was done before.

```
rid = tolower(substr(colnames(meTIL_probes),1,12))
rid = gsub("-", ".", rid)

colnames(meTIL_probes) <- rid
```

```r
table(colnames(rnaseqMel)%in%colnames(meTIL_probes))
```

```
##
## TRUE
##  469
```

```r
# All of the RNA-seq patient identifiers are also in the methylation identifiers

which(duplicated(colnames(meTIL_probes))) # There are 5 duplicates
```

```
## [1]  37 316 324 388 415
```

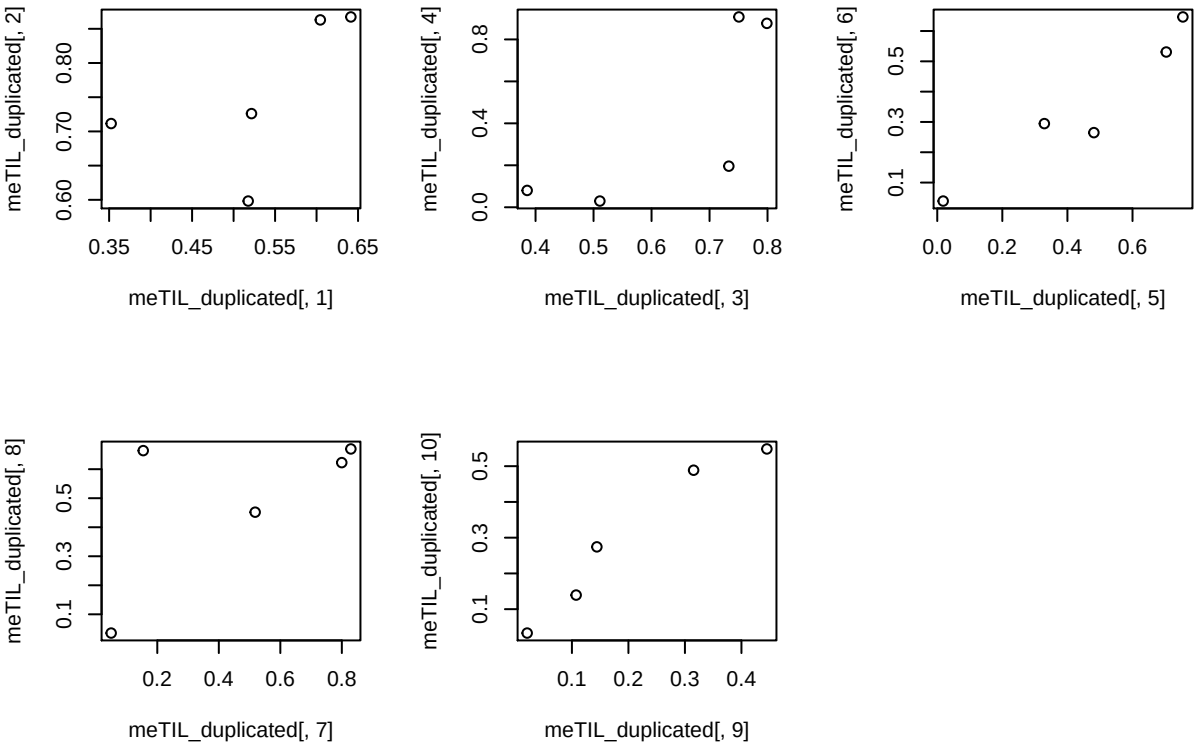
```r
colnames(meTIL_probes)[c(36,37,315,316,323,324,387,388,414,415)]
```

```
##  [1] "tcga.d3.a1qa" "tcga.d3.a1qa" "tcga.er.a19t" "tcga.er.a19t"
##  [5] "tcga.er.a2nf" "tcga.er.a2nf" "tcga.fw.a3r5" "tcga.fw.a3r5"
##  [9] "tcga.gn.a4u8" "tcga.gn.a4u8"
```

```r
duplicated_SampleNames <- colnames(meTIL_probes)[duplicated(colnames(meTIL_probes))]

meTIL_duplicated<- meTIL_probes[,colnames(meTIL_probes)%in%duplicated_SampleNames]
colnames(meTIL_duplicated)
```

```
##  [1] "tcga.d3.a1qa"   "tcga.d3.a1qa.1" "tcga.er.a19t"   "tcga.er.a19t.1"
##  [5] "tcga.er.a2nf"   "tcga.er.a2nf.1" "tcga.fw.a3r5"   "tcga.fw.a3r5.1"
##  [9] "tcga.gn.a4u8"   "tcga.gn.a4u8.1"
```

```r
par(mfrow=c(2,3))
plot(meTIL_duplicated[,1],meTIL_duplicated[,2])
plot(meTIL_duplicated[,3],meTIL_duplicated[,4])
plot(meTIL_duplicated[,5],meTIL_duplicated[,6])
plot(meTIL_duplicated[,7],meTIL_duplicated[,8])
plot(meTIL_duplicated[,9],meTIL_duplicated[,10])
```



There seems to be more variation in the methylation 450K data compared to the RNA-seq data within the duplicates. But I'm not sure which one to take so i will drop the second data.

```
meTIL_probes <- meTIL_probes[,!duplicated(colnames(meTIL_probes))] # dropping the duplicates

dim(meTIL_probes)
```

```
## [1]   5 470
```

```
dim(rnaseqMel)
```

```
## [1] 20501   469
```

```
table(colnames(meTIL_probes)%in%colnames(rnaseqMel))
```

```
##
## FALSE  TRUE
##     1   469
```

```
# Theres 1 extra sample in meTIL_probes which is not in rnaseqMel

meTIL_probes <- meTIL_probes[,colnames(meTIL_probes )%in%colnames(rnaseqMel)]

table(colnames(meTIL_probes) == colnames(rnaseqMel)) # Everything is in the same order and matches.
```

```
##
## TRUE
##  469
```

```
write.csv(meTIL_probe_info, file="meTIL_probe_info.csv")
write.csv(meTIL_probes, file="meTIL_probes.csv")
```

References

Ekmekcioglu, S., et al. 2016. “Inflammatory Marker Testing Identifies Cd74 Expression in Melanoma Tumor Cells, and Its Expression Associates with Favorable Survival for Stage III Melanoma.”

Jeschke, J., et al. 2017. “DNA Methylation-Based Immune Response Signature Improves Patient Diagnosis in Multiple Cancers.”

Samur, M. K. 2014. “RTCGAToolbox a New Tool for Exporting TCGA Firehose Data.”