Processing TCGA mRNA expression data

Humberto Ortiz-Zuazaga humberto.ortiz@upr.edu

03/26/2015

Introduction

We will be using the RTCGAToolbox, an experimental package for analysis of data from The Cancer Genome Atlas.

The analisis in this workshop borrows heavily from the RTCGAToolbox tutorial, available online. See the RTCGAToolbox tutorial.

The Toolbox uses `limma` and other bioconductor tools to do the analysis of the data. You can get more information and guidance for further analysis from the limma User's Guide, and the bioconductor website.

The analysis of RNASeq data is similar to microarray analysis. I have given other workshops on the analysis of these kinds of data.

1. Ortiz-Zuazaga, Humberto (2014): Microarray Analysis With Bioconductor Workshop. figshare. doi: 10.6084/m9.figshare.1251183 Retrieved 13:33, Nov 25, 2014 (GMT)
2. Using limma for microarray and RNA-Seq analysis. A workshop for the Research Design, Biostatistics and Clinical Research Ethics (DBE) key function of the PRCTRC, San Juan, PR, March 7, 2013. Handout

Instalation

To install the R TCGA Toolbox you need to run these commands once:

```
source("http://bioconductor.org/biocLite.R")
biocLite("limma")
biocLite("devtools")
library(devtools)
install_github("mksamur/RTCGAToolbox")
```

After that, you can load the toolbox with just one command.

```r
library(RTCGAToolbox)
```

Obtaining TCGA Data

The toolbox provides functions to examine the TCGA data available at the Broad Institute:

```r
getFirehoseDatasets()
```

```
##  [1] "ACC"      "BLCA"     "BRCA"     "CESC"     "COAD"     "COADREAD"
##  [7] "DLBC"     "ESCA"     "GBM"      "HNSC"     "KICH"     "KIRC"
## [13] "KIRP"     "LAML"     "LGG"      "LIHC"     "LUAD"     "LUSC"
## [19] "MESO"     "OV"       "PAAD"     "PCPG"     "PRAD"     "READ"
## [25] "SARC"     "SKCM"     "STAD"     "TGCT"     "THCA"     "THYM"
## [31] "UCEC"     "UCS"      "UVM"
```

These data are subject to the conditions of use available at the site. Please make sure you comply with all the conditions of use.

Each study has been updated multiple times:

```
stddata = getFirehoseRunningDates()
stddata
```

```
##  [1] "20150204" "20141206" "20141017" "20140902" "20140715" "20140614"
##  [7] "20140518" "20140416" "20140316" "20140215" "20140115" "20131210"
## [13] "20131114" "20131010" "20130923" "20130809" "20130715" "20130623"
## [19] "20130606" "20130523" "20130508" "20130421" "20130406" "20130326"
## [25] "20130309" "20130222" "20130203" "20130116" "20121221" "20121206"
## [31] "20121114" "20121102" "20121024" "20121020" "20121018" "20121004"
## [37] "20120913" "20120825" "20120804" "20120725" "20120707" "20120623"
## [43] "20120606" "20120525" "20120515" "20120425" "20120412" "20120321"
## [49] "20120306" "20120217" "20120124" "20120110" "20111230" "20111206"
## [55] "20111128" "20111115" "20111026"
```

Different experiments are updated on different dates.

```
gisticDate = getFirehoseAnalyzeDates(last=3)
gisticDate
```

```
## [1] "20141017" "20140715" "20140416"
```

We could obtain the data for breast cancer, including copy number variants, the clinical data, RNASeq and SNP with a single call. This would download the data and cache a local copy.

```
brcaData = getFirehoseData (dataset="BRCA", runDate="20150204",
                            gistic2_Date="20141017",
               Clinic=TRUE, RNAseq_Gene=TRUE, mRNA_Array=FALSE, Mutation=TRUE)
```

For this workshop, we will work instead with a sample dataset included with the toolbox. It contains RNASeq data for 100 genes in 800 participants.

```
data(RTCGASample)
brcaData = a2
```

Examining the TCGA data

What's in the data? There is clinical data on each participant.

```r
dim(brcaData@Clinical)
```

```
## [1]  22 989
```

```r
colnames(brcaData@Clinical[,1:20])
```

```
##  [1] "Hybridization.REF" "tcga.a1.a0sb"      "tcga.a1.a0sd"
##  [4] "tcga.a1.a0se"      "tcga.a1.a0sf"      "tcga.a1.a0sg"
##  [7] "tcga.a1.a0sh"      "tcga.a1.a0si"      "tcga.a1.a0sj"
## [10] "tcga.a1.a0sk"      "tcga.a1.a0sm"      "tcga.a1.a0sn"
## [13] "tcga.a1.a0so"      "tcga.a1.a0sp"      "tcga.a1.a0sq"
## [16] "tcga.a2.a04n"      "tcga.a2.a04p"      "tcga.a2.a04q"
## [19] "tcga.a2.a04r"      "tcga.a2.a04t"
```

```r
brcaData@Clinical[,1]
```

```
##  [1] "Composite Element REF"
##  [2] "yearstobirth"
##  [3] "vitalstatus"
##  [4] "daystodeath"
##  [5] "daystolastfollowup"
##  [6] "primarysiteofdesease"
##  [7] "neoplasm.diseasestage"
##  [8] "pathology.T.stage"
##  [9] "pathology.N.stage"
## [10] "pathology.M.stage"
## [11] "dccuploaddate"
## [12] "gender"
## [13] "dateofinitialpathologicdiagnosis"
## [14] "daystolastknownalive"
## [15] "radiationtherapy"
## [16] "histologicaltype"
## [17] "radiations.radiation.regimenindication"
## [18] "number.of.lymph.nodes"
## [19] "gleason_score"
## [20] "psa_value"
## [21] "days_to_psa"
## [22] "batchnumber"
```

```r
brcaData@Clinical[2:7,c(1:3,10)]
```

```
##       Hybridization.REF tcga.a1.a0sb tcga.a1.a0sd tcga.a1.a0sk
## 2          yearstobirth           70           59           54
## 3           vitalstatus            0            0            1
## 4           daystodeath         <NA>         <NA>          967
## 5    daystolastfollowup          259          437         <NA>
## 6  primarysiteofdesease       breast       breast       breast
## 7 neoplasm.diseasestage      stage i    stage iia    stage iia
```

We also have raw RNASeq read data for each participant.

```
dim(brcaData@RNASeqGene)
```

```
## [1] 100 878
```

```
brcaData@RNASeqGene[1:5,1:3]
```

```
##              TCGA-A1-A0SB-01A-11R-A144-07 TCGA-A1-A0SD-01A-11R-A115-07
## CST1|1469                               6                         8415
## COL10A1|1300                         2133                        20889
## MMP13|4322                              8                         4080
## IBSP|3381                               1                          627
## MMP11|4320                            692                        27243
##              TCGA-A1-A0SE-01A-11R-A084-07
## CST1|1469                            1217
## COL10A1|1300                        19640
## MMP13|4322                           1278
## IBSP|3381                              59
## MMP11|4320                          39701
```

The column names encode the disease status of the samples. We can separate the tumor samples from the control samples by using some R code. As Dr. Gonzalez mentioned in the workshop, a large part of any analysis in R will be manipulating the data to extract information on the samples or the genes.

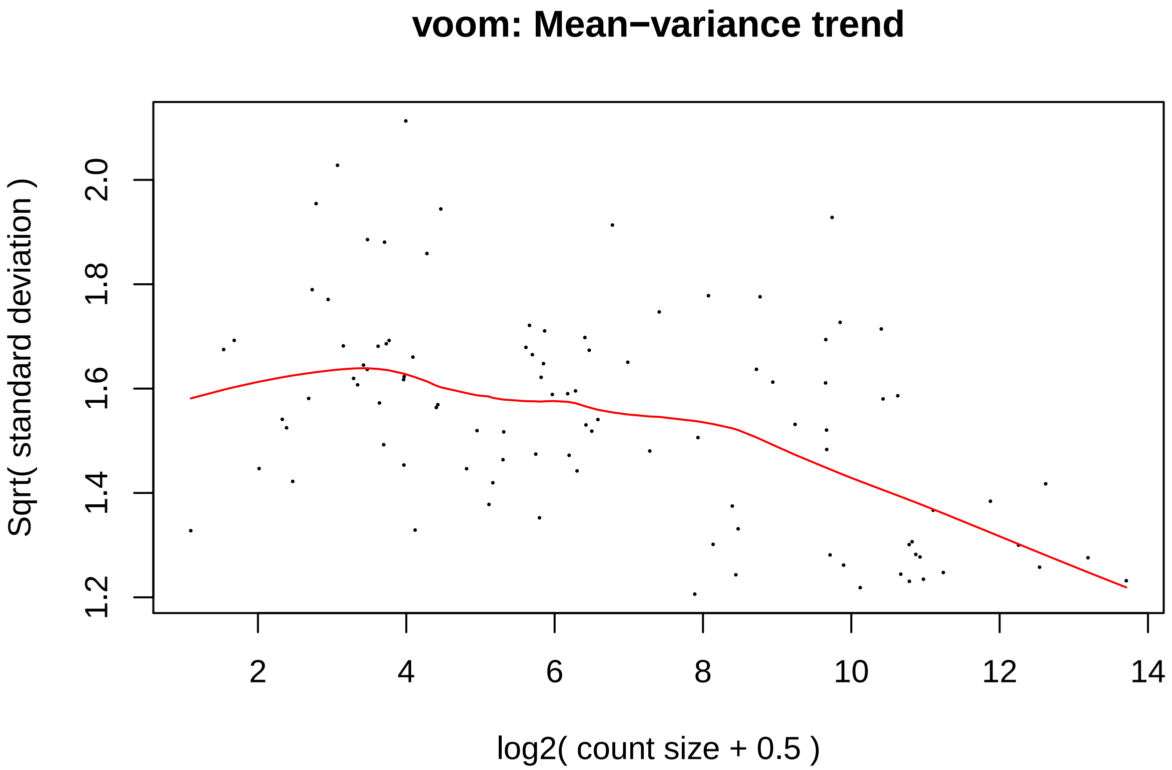
The limma users guide has example code to extract information from many different kinds of studies, and the U54 BEBiC or the HPCf helpdesk may be able to assist you in creating appropriate data analysis scripts.

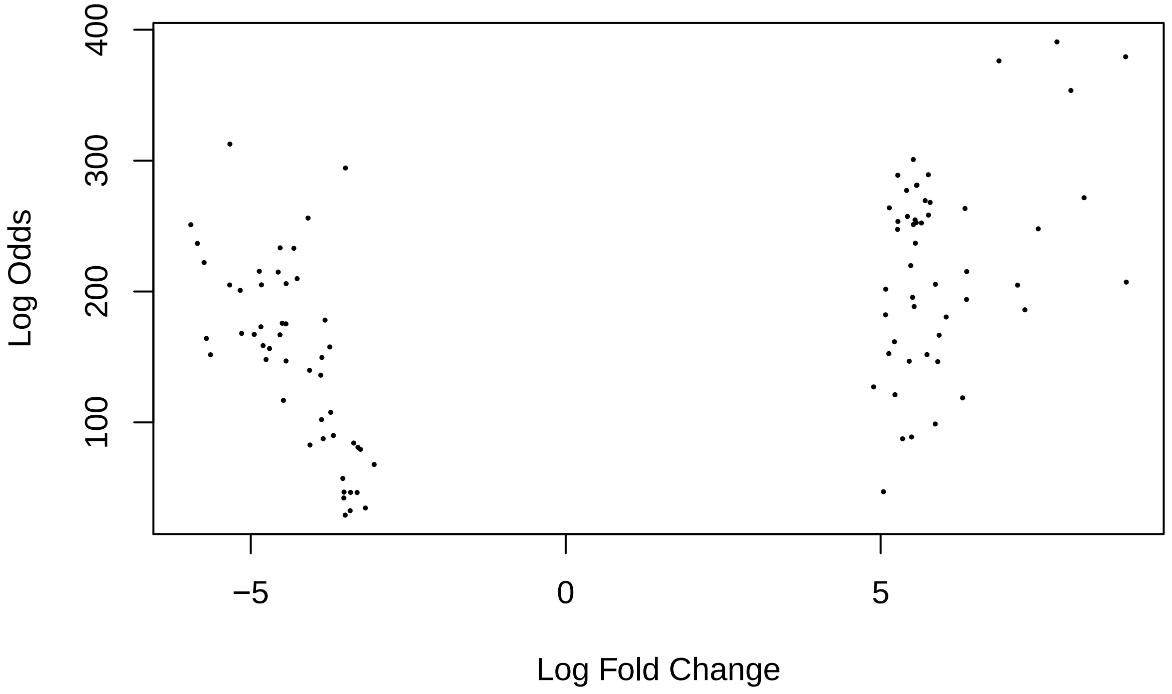
```
sampleIDs <- colnames(brcaData@RNASeqGene)
barcode <- rep(x = "", length(sampleIDs))
for (j in 1:length(sampleIDs)) {
  barcode[j] <- unlist(strsplit(sampleIDs[j], split = "-"))[4]
}
sampleIDs1 <- substr(barcode, 1, nchar(barcode) - 1)
sampleIDs1 <- as.numeric(sampleIDs1)
normalSamples <- sampleIDs[sampleIDs1 < 20 & sampleIDs1 > 9]
tumorSamples <- sampleIDs[sampleIDs1 < 10]
```

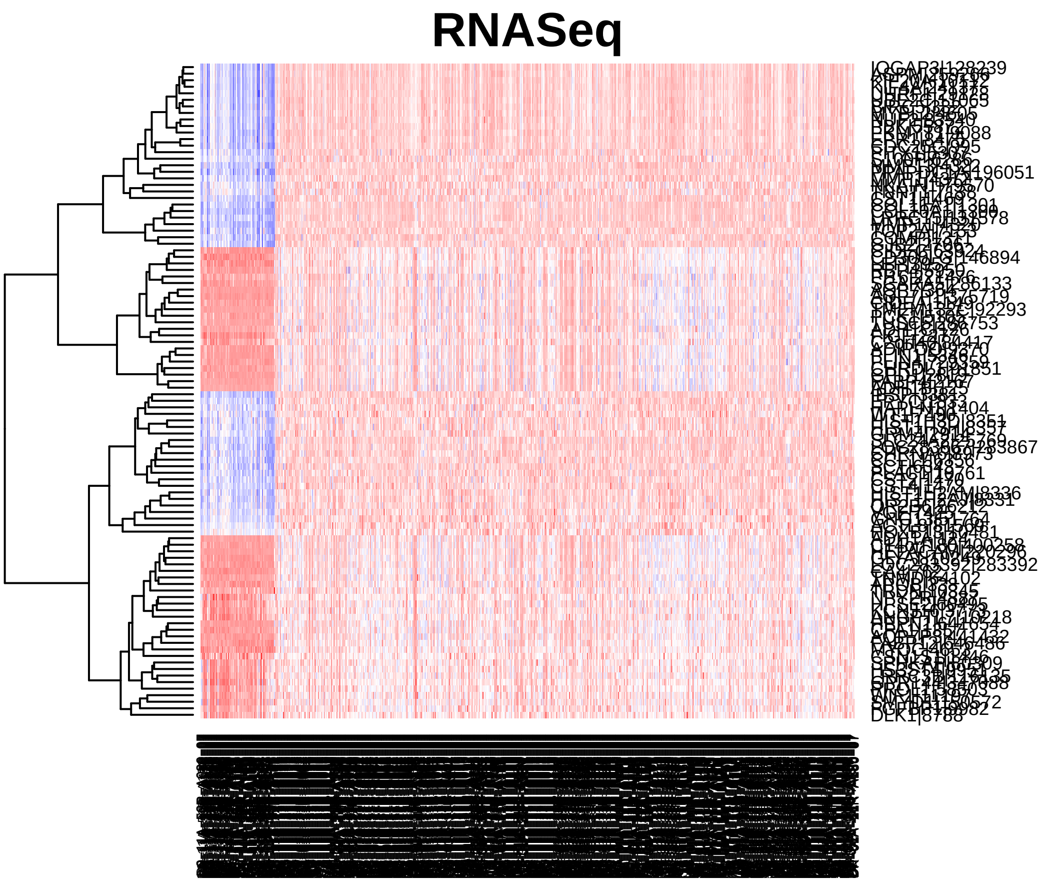
Differential gene expression analysis

Run differential expression analysis on tumor vs normal with sample data.

```
diffGeneExprs = getDiffExpressedGenes(dataObject=a2)
```


voom: Mean−variance trend
Sqrt(standard deviation)
2.0
1.8
1.6
1.4
1.2
2
4
6
8
10
12
14
log2(count size + 0.5)


Log Odds
400
300
200
100
−5
0
5
Log Fold Change



The real analysis could be run as follows

```
diffGeneExprs = getDiffExpressedGenes(dataObject=brcaData,DrawPlots=TRUE,
                                      adj.method="BH",adj.pval=0.05,raw.pval=0.05,
                                      logFC=2,hmTopUpN=10,hmTopDownN=10)
```

Reporting the results

The `diffGeneExprs` object contains the results for each type of data present in our data object. The sample data only has RNASeq data (not microarray).

```
for(i in 1:length(diffGeneExprs)) {
   writeLines(diffGeneExprs[[i]]@Dataset);
   print(head(diffGeneExprs[[i]]@Toptable))
}
```

```
## RNASeq
##                      logFC   AveExpr         t       P.Value     adj.P.Val
## MMP11|4320        7.798190 15.851413  36.01997 8.493105e-176 8.493105e-174
## COL10A1|1300      8.888113 14.398666  35.26459 5.509317e-171 2.754659e-169
## PPAPDC1A|196051   6.877850 10.280566  35.06334 1.061985e-169 3.539950e-168
## IBSP|3381         8.019494  8.339241  33.51301 9.188515e-160 2.297129e-158
## LEP|3952         -5.330958  9.426745 -30.68711 1.596039e-141 3.192078e-140
## PKMYT1|9088       5.518106 12.263864  29.92449 1.381221e-136 2.302035e-135
##                          B
## MMP11|4320      390.7043
## COL10A1|1300    379.3535
## PPAPDC1A|196051 376.2074
## IBSP|3381       353.5686
## LEP|3952        312.6208
## PKMYT1|9088     300.8328
```

Summarizing multiple sources of information

The Toolbox also allows you to combine information from multiple data sources, including SNP (mutations), copy number variantion, RNASeq and microarrays.

```
library(RCircos)
data(hg19.ucsc.gene.locations)
getReport(dataObject=a2,
          DGEResult1=diffGeneExprs[[1]],
          geneLocations=hg19.ucsc.gene.locations)
```

```
##
## RCircos.Core.Components initialized.
## Type ?RCircos.Reset.Plot.Parameters to see how to modify the core components.

## Track No: 1  (in) differential gene expression data 1
## Track No: 2  (in) copy number data

## Outside track mutations!

## pdf
##   2
```

You can see the resulting graph by opening `BRCA-reportImage.pdf`.

Acknowledgements

This workshop was sponsored by the UPR/MDACC Partnership for Excellence in Cancer Research (NIH-NCI 2U54CA096297-11) and the Puerto Rico Clinical and Translational Research Consortium (NIH-NIMHD 8U54MD007587-03). Humberto Ortiz-Zuazaga is supported in part by the Puerto Rico IDeA Network Biomedical Research Excellence award (NIH NIGMS P20GM103475-13).