GSE32866, Genome-scale DNA methylation profiling of lung adenocarcinoma:

Paulyna Magana

Introduction

Load the following packages:

```
library(rmarkdown) # paged table
library(GEOquery) #geo query access
library(dplyr)
library(devtools)
library(ggplot2)
library(limma)
library(Glimma)
library(edgeR)
library(kableExtra)
library(ggrepel)
library(tinytex)
library(purrr)
Sys.setenv(VROOM_CONNECTION_SIZE = 25600000)
```

Importing the data

```
my_id <- "GSE32866"
gse <- getGEO(my_id)
```

Some datasets on GEO may be derived from different microarray platforms. Therefore the object gse is a list of different datasets. You can find out how many were used by checking the length of the gse object. Usually there will only be one platform and the dataset we want to analyse will be the first object in the list (gse[[1]]).

```
length(gse)
```

```
[1] 1
```

Extract the data

```
gse <- gse[[1]]
gse
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 27578 features, 55 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: GSM813533 GSM813534 ... GSM813587 (55 total)
  varLabels: title geo_accession ... tissue:ch1 (58 total)
  varMetadata: labelDescription
featureData
  featureNames: cg00000292 cg00002426 ... cg27665659 (27578 total)
  fvarLabels: ID Name ... ORF (38 total)
  fvarMetadata: Column Description labelDescription
experimentData: use 'experimentData(object)'
  pubMedIds: 22613842
Annotation: GPL8490
```

Exploratory analysis

The `exprs` function can retrieve the expression values as a data frame; with one column per-sample and one row per-gene.

```
pdata= pData(gse) #sample information
edata= exprs(gse) #expression data
fdata = fData(gse) #gene annotation
```

Make sure dimensions match up

The number of rows of the feature data should match the number of rows of the genomic data (both are the number of genes). The number of rows of the phenotype data should match the number of columns of the genomic data (both are the number of samples).

```
dim(fdata)
```

```
[1] 27578    38
```

```
#Dimension of edata
dim(pdata)
```

```
[1] 55 58
```

```
#Dimension of edata
dim(edata)
```

```
[1] 27578    55
```

Look at overall distributions

For visualisation and statistical analysis, we will inspect the data to discover what scale the data are presented in. The methods we will use assume the data are on a log2 scale; typically in the range of 0 to 16.

The `summary` function can then be used to print the distributions.

```
## exprs get the expression levels as a data frame and get the distribution
dat <- summary(exprs(gse))
m <- matrix(1:ncol(dat), 5)

for (i in 1:ncol(m)) {
 cat(kbl(dat[, m[, i]], 'latex', booktabs=TRUE), "\\newline")
}
```

GSM813533	GSM813534	GSM813535	GSM813536	GSM813537
Min. :0.00158	Min. :0.0088	Min. :0.003852	Min. :0.00412	Min. :0.001593
1st Qu.:0.02499	1st Qu.:0.0675	1st Qu.:0.034014	1st Qu.:0.04078	1st Qu.:0.021340
Median :0.08185	Median :0.1621	Median :0.088613	Median :0.10260	Median :0.066667
Mean :0.26751	Mean :0.2951	Mean :0.270265	Mean :0.28832	Mean :0.274648
3rd Qu.:0.52701	3rd Qu.:0.5223	3rd Qu.:0.526548	3rd Qu.:0.56454	3rd Qu.:0.557164
Max. :0.99686	Max. :0.9857	Max. :0.993288	Max. :0.99236	Max. :0.997155
NA's :118	NA's :351	NA's :23	NA's :66	NA's :1

GSM813538	GSM813539	GSM813540	GSM813541	GSM813542
Min. :0.001563	Min. :0.002816	Min. :0.01359	Min. :0.01291	Min. :0.00453
1st Qu.:0.021631	1st Qu.:0.023452	1st Qu.:0.08990	1st Qu.:0.07017	1st Qu.:0.03913
Median :0.071831	Median :0.059120	Median :0.15982	Median :0.14873	Median :0.09375
Mean :0.272946	Mean :0.272262	Mean :0.30666	Mean :0.29943	Mean :0.27430
3rd Qu.:0.548465	3rd Qu.:0.569431	3rd Qu.:0.54827	3rd Qu.:0.54477	3rd Qu.:0.54061
Max. :0.996068	Max. :0.996306	Max. :0.97459	Max. :0.97656	Max. :0.99097
NA's :4	NA's :3	NA's :25	NA's :16	NA's :36

GSM813543	GSM813544	GSM813545	GSM813546	GSM813547
Min. :0.003748	Min. :0.002923	Min. :0.01073	Min. :0.00363	Min. :0.00374
1st Qu.:0.025443	1st Qu.:0.022825	1st Qu.:0.05799	1st Qu.:0.04541	1st Qu.:0.02760
Median :0.066108	Median :0.060558	Median :0.11796	Median :0.14063	Median :0.06540
Mean :0.269724	Mean :0.263932	Mean :0.28764	Mean :0.29780	Mean :0.25911
3rd Qu.:0.542025	3rd Qu.:0.538074	3rd Qu.:0.54404	3rd Qu.:0.54639	3rd Qu.:0.51551
Max. :0.996605	Max. :0.994622	Max. :0.98549	Max. :0.99514	Max. :0.99406
NA's :1	NA's :5	NA's :6	NA's :42	NA's :2

GSM813548	GSM813549	GSM813550	GSM813551	GSM813552
Min. :0.002475	Min. :0.001802	Min. :0.00275	Min. :0.007388	Min. :0.00387
1st Qu.:0.022601	1st Qu.:0.021012	1st Qu.:0.03178	1st Qu.:0.053120	1st Qu.:0.04471
Median :0.074436	Median :0.059294	Median :0.09664	Median :0.107693	Median :0.12566
Mean :0.271670	Mean :0.264499	Mean :0.29360	Mean :0.286343	Mean :0.28056
3rd Qu.:0.533615	3rd Qu.:0.532461	3rd Qu.:0.58296	3rd Qu.:0.555106	3rd Qu.:0.52114
Max. :0.995359	Max. :0.997065	Max. :0.99442	Max. :0.985281	Max. :0.99471
NA	NA's :6	NA's :96	NA's :3	NA's :208

GSM813553	GSM813554	GSM813555	GSM813556	GSM813557
Min. :0.003055	Min. :0.01007	Min. :0.0099	Min. :0.00969	Min. :0.001198
1st Qu.:0.032660	1st Qu.:0.05484	1st Qu.:0.0568	1st Qu.:0.05408	1st Qu.:0.018511
Median :0.086232	Median :0.12322	Median :0.1227	Median :0.13335	Median :0.051788
Mean :0.280567	Mean :0.28903	Mean :0.2733	Mean :0.28789	Mean :0.239002
3rd Qu.:0.559251	3rd Qu.:0.55005	3rd Qu.:0.4960	3rd Qu.:0.53638	3rd Qu.:0.455191
Max. :0.994143	Max. :0.98923	Max. :0.9836	Max. :0.98753	Max. :0.996707
NA	NA's :51	NA's :453	NA's :71	NA's :16

GSM813558	GSM813559	GSM813560	GSM813561	GSM813562
Min. :0.000755	Min. :0.000893	Min. :0.00000	Min. :0.00329	Min. :0.00288
1st Qu.:0.017597	1st Qu.:0.015929	1st Qu.:0.02894	1st Qu.:0.03900	1st Qu.:0.02200
Median :0.072014	Median :0.044893	Median :0.09338	Median :0.11016	Median :0.05827
Mean :0.277990	Mean :0.230163	Mean :0.26941	Mean :0.26246	Mean :0.25957
3rd Qu.:0.559703	3rd Qu.:0.429716	3rd Qu.:0.51380	3rd Qu.:0.46627	3rd Qu.:0.51087
Max. :0.997774	Max. :0.996079	Max. :1.00000	Max. :0.99187	Max. :0.99605
NA's :8	NA's :19	NA's :11	NA's :102	NA's :49

GSM813563	GSM813564	GSM813565	GSM813566	GSM813567
Min. :0.005434	Min. :0.003316	Min. :0.00608	Min. :0.003185	Min. :0.009481
1st Qu.:0.039251	1st Qu.:0.024736	1st Qu.:0.04581	1st Qu.:0.023066	1st Qu.:0.070661
Median :0.162052	Median :0.086374	Median :0.11263	Median :0.062453	Median :0.156346
Mean :0.322312	Mean :0.271873	Mean :0.28444	Mean :0.269006	Mean :0.297500
3rd Qu.:0.604959	3rd Qu.:0.518953	3rd Qu.:0.53472	3rd Qu.:0.528240	3rd Qu.:0.521266
Max. :0.989720	Max. :0.995040	Max. :0.98937	Max. :0.995625	Max. :0.979380
NA's :1	NA's :1	NA's :39	NA's :2	NA's :3

GSM813568	GSM813569	GSM813570	GSM813571	GSM813572
Min. :0.01552	Min. :0.00985	Min. :0.006009	Min. :0.004533	Min. :0.003998
1st Qu.:0.07648	1st Qu.:0.07542	1st Qu.:0.044466	1st Qu.:0.030680	1st Qu.:0.029502
Median :0.15532	Median :0.16813	Median :0.119853	Median :0.078872	Median :0.098945
Mean :0.28663	Mean :0.28440	Mean :0.290716	Mean :0.256968	Mean :0.302563
3rd Qu.:0.47719	3rd Qu.:0.47053	3rd Qu.:0.547162	3rd Qu.:0.484131	3rd Qu.:0.600460
Max. :0.97192	Max. :0.97799	Max. :0.987253	Max. :0.992672	Max. :0.992380
NA's :4	NA's :51	NA's :1	NA's :3	NA's :1

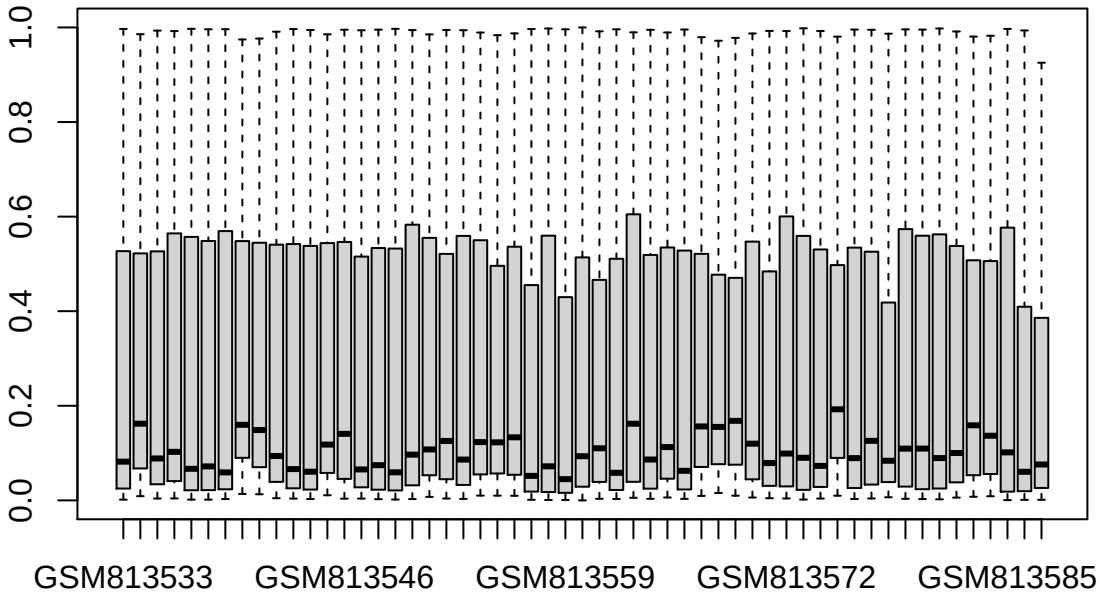
GSM813573	GSM813574	GSM813575	GSM813576	GSM813577
Min. :0.001719	Min. :0.003869	Min. :0.00984	Min. :0.002911	Min. :0.003975
1st Qu.:0.022253	1st Qu.:0.028332	1st Qu.:0.08979	1st Qu.:0.026097	1st Qu.:0.033384
Median :0.090038	Median :0.072875	Median :0.19254	Median :0.089301	Median :0.125818
Mean :0.283277	Mean :0.270009	Mean :0.30278	Mean :0.274118	Mean :0.286136
3rd Qu.:0.559240	3rd Qu.:0.530568	3rd Qu.:0.49747	3rd Qu.:0.534446	3rd Qu.:0.525828
Max. :0.998242	Max. :0.992547	Max. :0.98056	Max. :0.995558	Max. :0.995050
NA's :6	NA's :3	NA's :124	NA's :4	NA's :4

GSM813578	GSM813579	GSM813580	GSM813581	GSM813582
Min. :0.00654	Min. :0.003284	Min. :0.002743	Min. :0.002231	Min. :0.005838
1st Qu.:0.03869	1st Qu.:0.029137	1st Qu.:0.023585	1st Qu.:0.025043	1st Qu.:0.038127
Median :0.08380	Median :0.109404	Median :0.109505	Median :0.089401	Median :0.100111
Mean :0.24641	Mean :0.294090	Mean :0.289157	Mean :0.282871	Mean :0.280244
3rd Qu.:0.41828	3rd Qu.:0.573597	3rd Qu.:0.559674	3rd Qu.:0.562494	3rd Qu.:0.537907
Max. :0.98813	Max. :0.995953	Max. :0.995473	Max. :0.997839	Max. :0.991527
NA's :6	NA	NA's :4	NA's :16	NA's :4

GSM813583	GSM813584	GSM813585	GSM813586	GSM813587
Min. :0.00773	Min. :0.00872	Min. :0.000538	Min. :0.001063	Min. :0.0012
1st Qu.:0.05315	1st Qu.:0.05574	1st Qu.:0.018411	1st Qu.:0.019389	1st Qu.:0.0262
Median :0.15874	Median :0.13669	Median :0.101503	Median :0.060440	Median :0.0759
Mean :0.28648	Mean :0.28092	Mean :0.294703	Mean :0.228942	Mean :0.2264
3rd Qu.:0.50774	3rd Qu.:0.50595	3rd Qu.:0.576664	3rd Qu.:0.409377	3rd Qu.:0.3859
Max. :0.98072	Max. :0.98240	Max. :0.996908	Max. :0.995679	Max. :0.9941
NA's :131	NA's :280	NA's :4	NA's :22	NA's :871

A boxplot can also be generated to see if the data have been normalised. If so, the distributions of each sample should be highly similar.

```
boxplot(edata,outline=FALSE)
```




```
kable(table(pdata$characteristics_ch1.1),  booktabs = T)
```

Var1	Freq
tissue: Lung tumor	28
tissue: Normal lung	27

```
kable(table(pdata$characteristics_ch1.6,pdata$characteristics_ch1.1), booktabs = T)
```

	tissue: Lung tumor	tissue: Normal lung
gender: Female	14	13
gender: Male	11	11
gender: NA	3	3

Inspect the clinical variables

Data submitted to GEO contain sample labels assigned by the experimenters, and some information about the processing protocol. All these data can be extracted by the pData function.

```
sampleInfo <- pData(gse)
```

```
## source_name_ch1 and characteristics_ch1.1 seem to contain factors we might need for the analysis
sampleInfo <- select(sampleInfo, source_name_ch1,characteristics_ch1.1)
```

```
#rename to more convenient column names
sampleInfo <- rename(sampleInfo, patient=characteristics_ch1.1, group = source_name_ch1)
kable(sampleInfo, longtable = T, booktabs = T, caption = "SampleInfo") %>%
kable_styling(latex_options = c("repeat_header"))
```

Table 1: SampleInfo

	group	patient
GSM813533	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813534	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813535	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813536	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813537	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813538	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813539	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813540	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813541	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813542	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813543	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813544	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813545	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813546	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813547	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813548	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813549	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813550	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813551	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813552	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813553	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813554	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813555	Fresh frozen macrodissected tissue	tissue: Normal lung

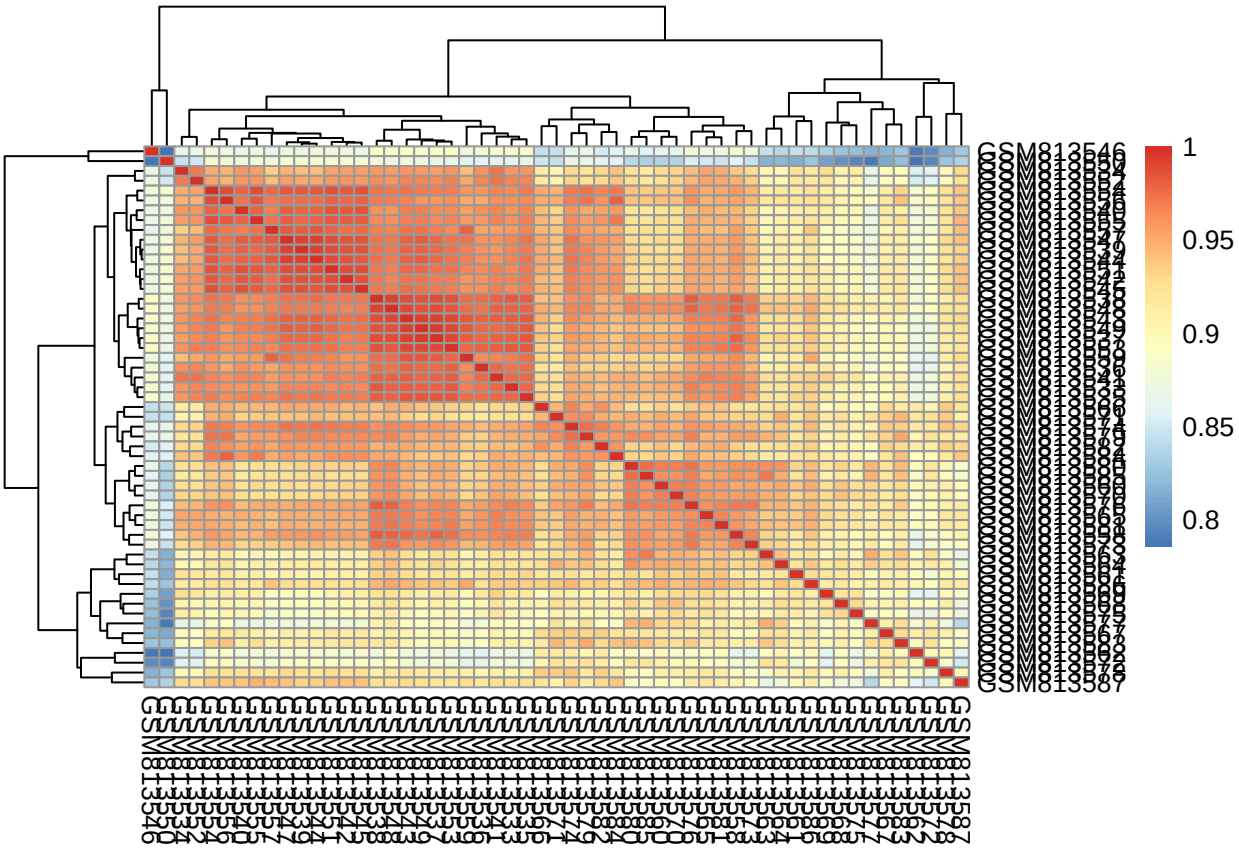
Table 1: SampleInfo *(continued)*

	group	patient
GSM813556	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813557	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813558	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813559	Fresh frozen macrodissected tissue	tissue: Normal lung
GSM813560	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813561	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813562	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813563	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813564	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813565	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813566	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813567	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813568	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813569	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813570	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813571	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813572	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813573	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813574	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813575	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813576	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813577	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813578	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813579	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813580	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813581	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813582	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813583	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813584	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813585	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813586	Fresh frozen macrodissected tissue	tissue: Lung tumor
GSM813587	Fresh frozen macrodissected tissue	tissue: Lung tumor

Sample clustering and Principal Components Analysis

Unsupervised analysis is a good way to get an understanding of the sources of variation in the data. It can also identify potential outlier samples.

```
library(pheatmap)
corMatrix <- cor(exprs(gse),use="c")
pheatmap(corMatrix)
```


1
0.95
0.9
0.85
0.8