Clustering of genomic data

La Serena Data Science

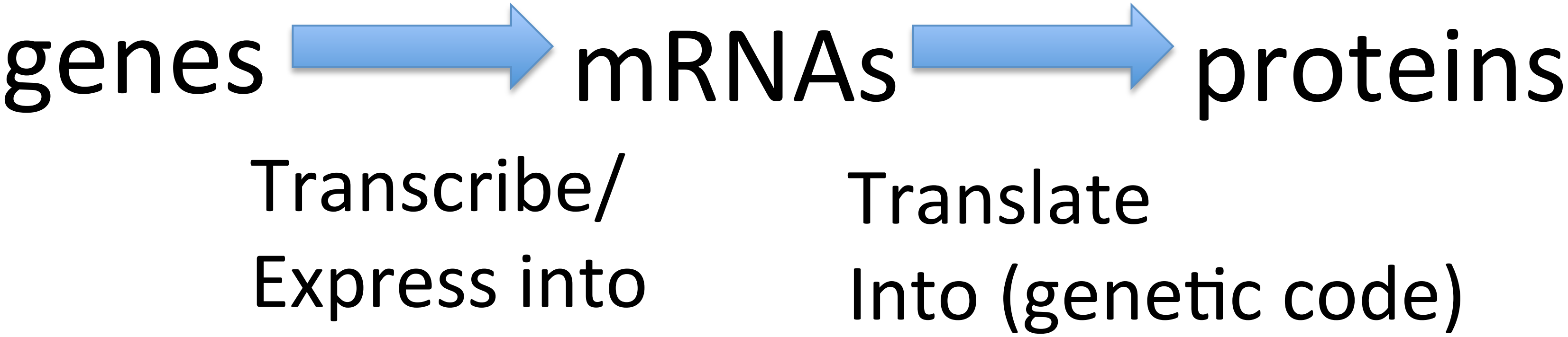
Rodrigo Assar, PhD

Assistant Professor Human Gene?cs Program School of Medicine, U. Chile

Genomic data

- Genes: DNA sequences that store the information about our phenotype conditions
  - Appearance
  - Predisposition to diseases
- Genes can express to make active these conditions
  - Cells especification by functions
  - Biological functioning
  - Manifestation/response of/to diseases

From genes to proteins



genes → mRNAs → proteins

Transcribe/
Express into

Translate
Into (genetic code)

Microarrays

- Measuring expression of many genes for different conditions
  - With stimulus (drug) versus with out (reference)

    -> resposing genes
  - With disease versus with out (reference)

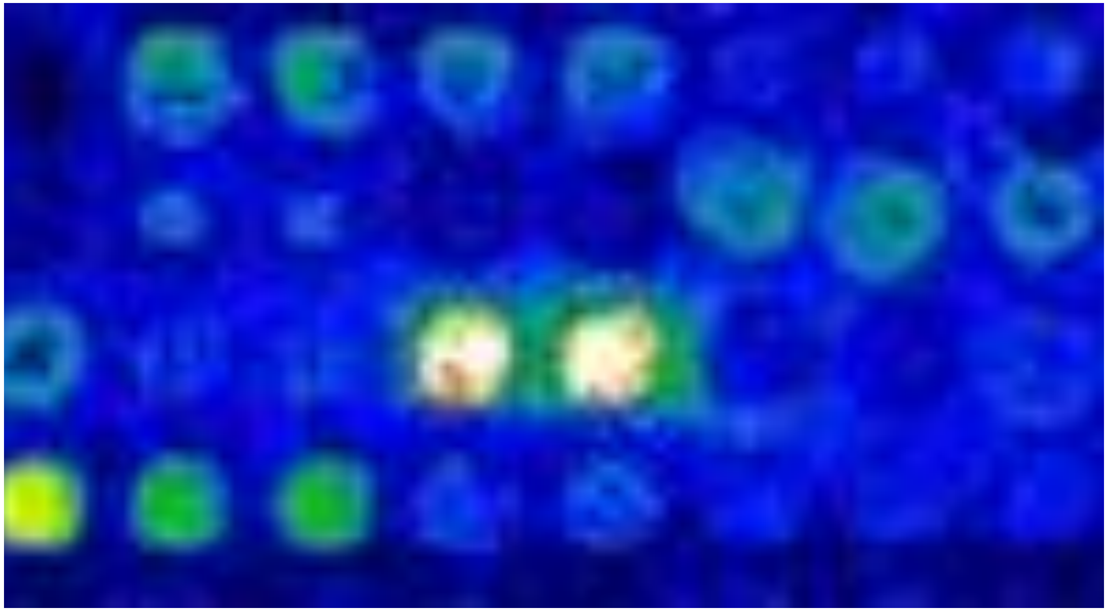
    -> genes that manifest the disease

Obtention of gene expression profiles and networks

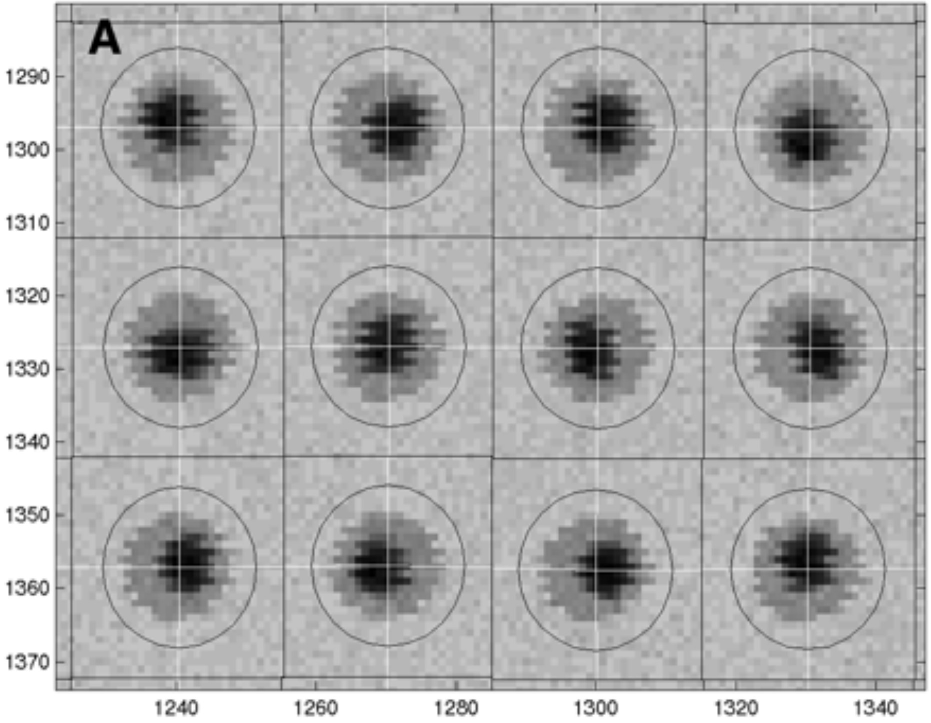
Imaging analysis of microarrays

Quantification of intensities that estimate expressions

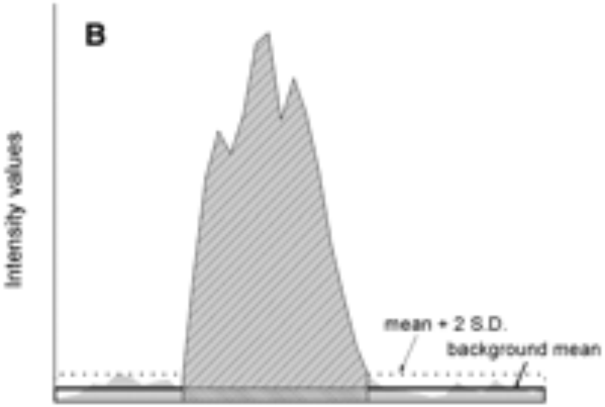
More than one spot representing one gene



Quantification ideas



Wang et al., 2001



(A) Images divided in pixels,

(B) Circular zones separating signal from background (noise),

(C) Quality estimations consider local background, stauration,

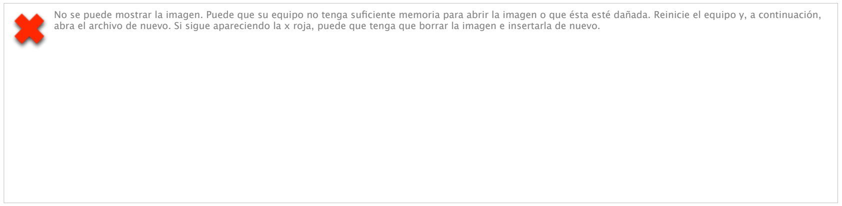
(D) Consideration of spots significantly different from background.

Differential expression: log-ratio

- Reference channel(Cy3 green fluorophore)
- Treatment channel(Cy5, red)

- Luminiscence quantification I

- Log-ratio index to compare conditions:



No se puede mostrar la imagen. Puede que su equipo no tenga suficiente memoria para abrir la imagen o que ésta esté dañada. Reinicie el equipo y, a continuación, abra el archivo de nuevo. Si sigue apareciendo la x roja, puede que tenga que borrar la imagen e insertarla de nuevo.

Interpretation of log-ratio values

Log-ratio	Interpretation
next to 0	treatment does not provoke change
positive	treatment increases expression (over expression)
1	Treatment duplicates the expression
negative	Treatment reduces the expression
-1	Treatment reduces the expression into half part

Example: PAM50

- Breast cancer subtypes give information about prognosis and orientate treatments-> improve assignations using genomic data.
- Microarray expression provide a comparative measure of patient vs healthy subject differences per gene.
- Typical human genome-wide microarray provides information about 20.000 variables (gene expressions).
- The problem is to identify variables with significant changes.
- **Paper by Parker et al., 2009**

Breast cancer sub-types

Name	Trend (with Ki67 also)
Luminal A	ER+, PR+, HER2-
Luminal B	ER+, PR+, HER2+
HER2-enriched	ER-, PR-, HER2+
Basal-like(Triple Negative)	ER-, PR-, HER2-
Normal-like	Different

PAM50 idea

- Characterize each patient sample by gene expressions
- Clusterize patients
- Reduce the number of dimensions (genes): 50
- Interpret clusters as subtypes
- Cluster techniques:
  - **Hierarchical clustering**
  - K-means

PAM50 results

columns: 189 samples

rows:1906 genes

Clusters (from left to right):

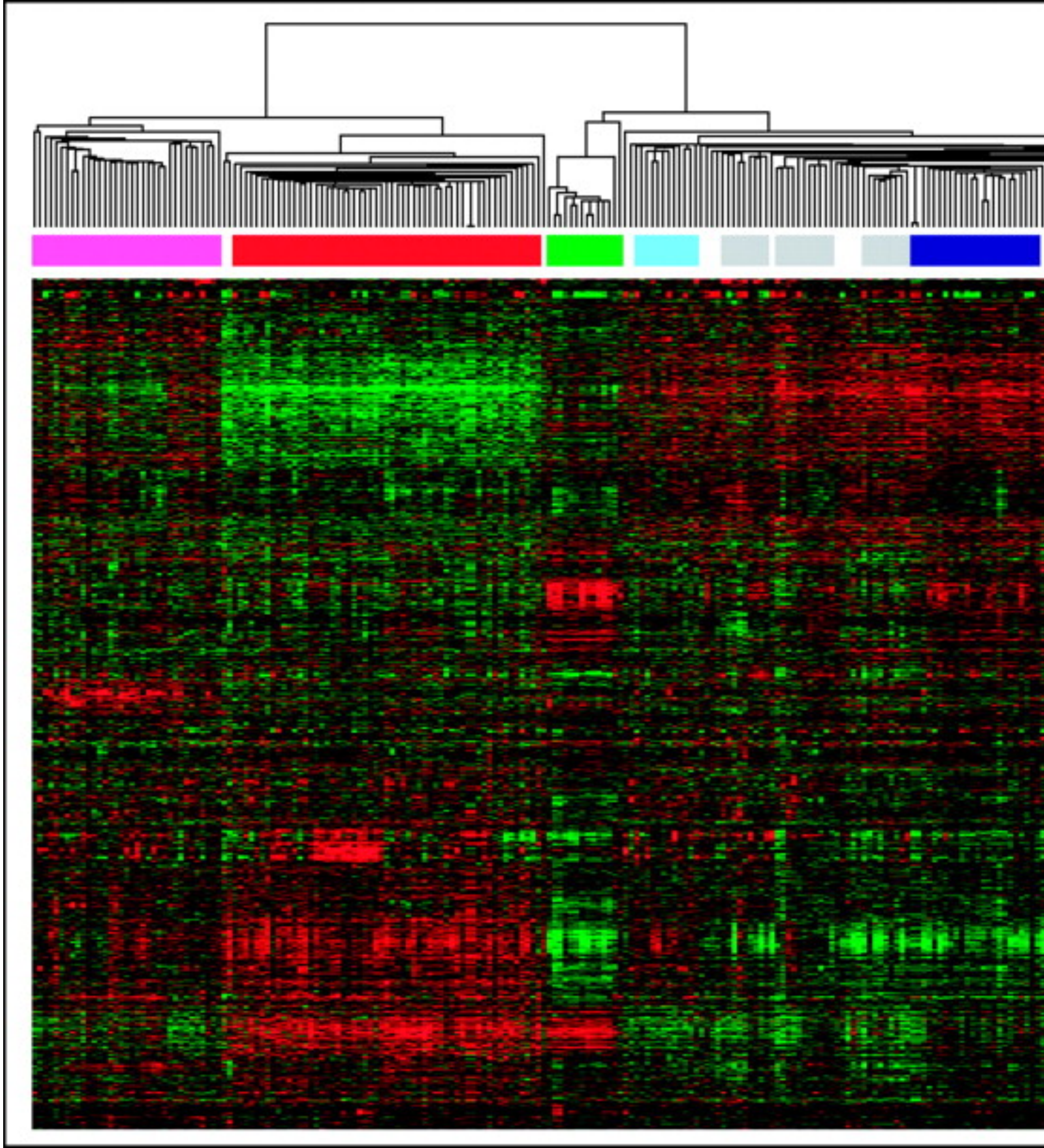
Her2-enriched (pink)

Basal-like

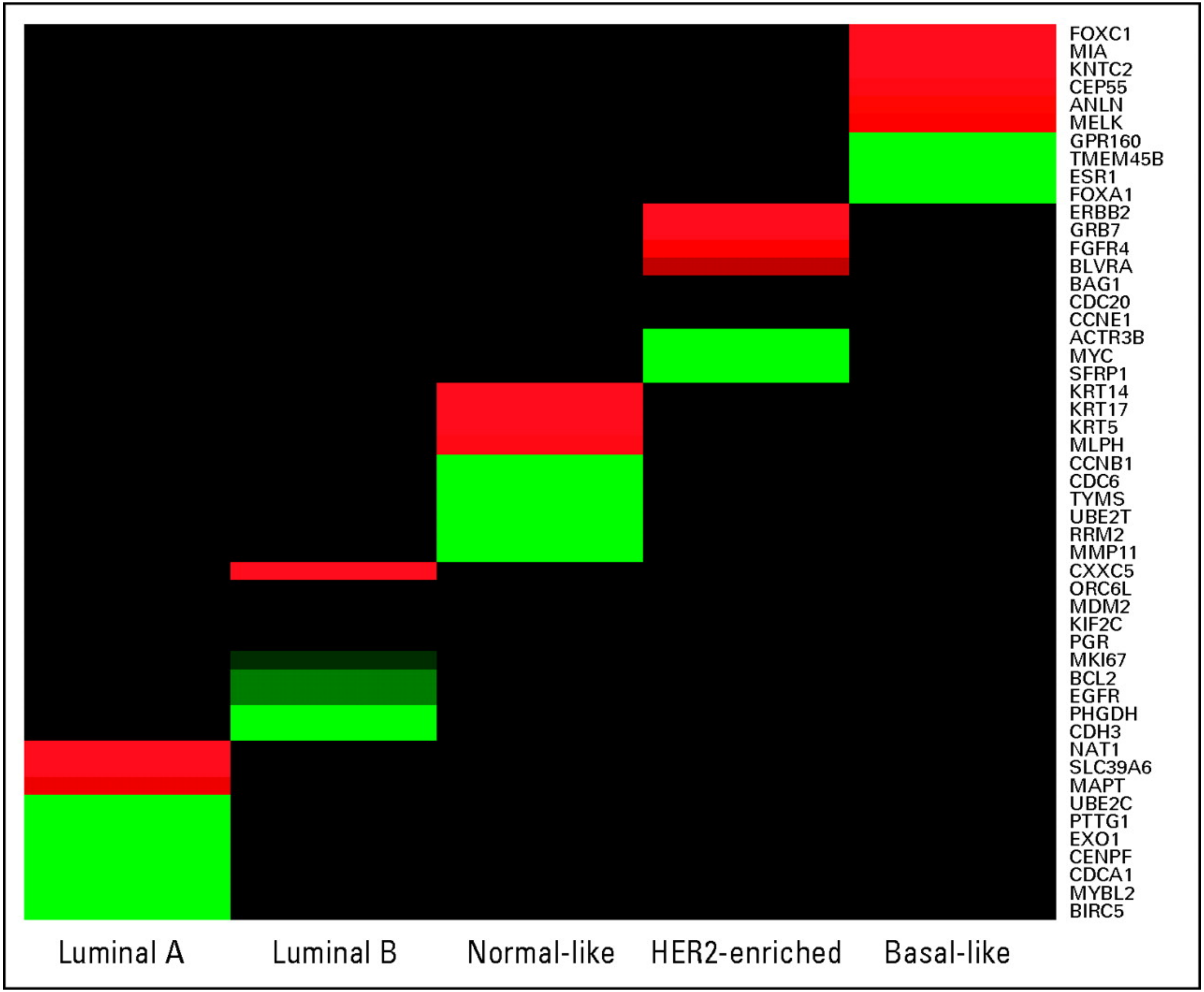
Normal-like

Luminal B (light blue)

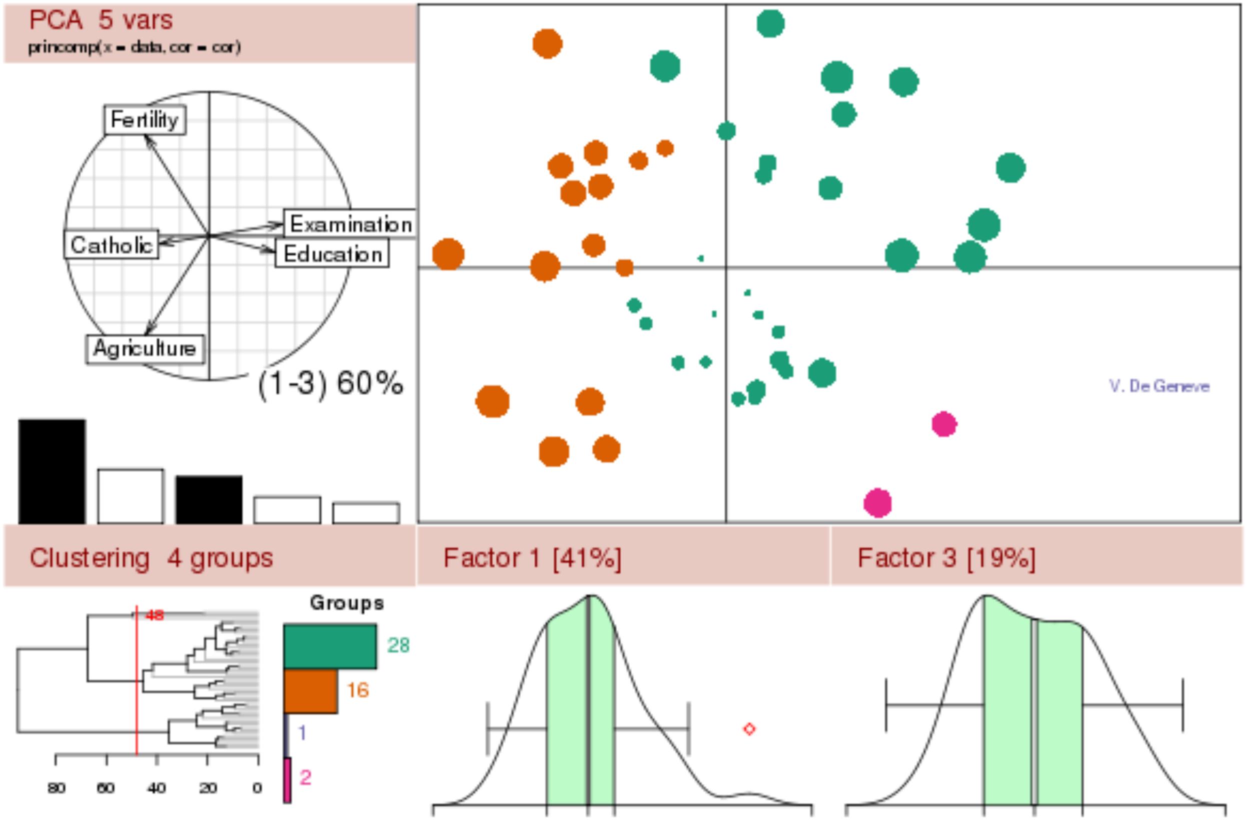
Luminal A (dark blue)



PAM50 results after cleaning: profiles



Playing with data: R (Rstudio)

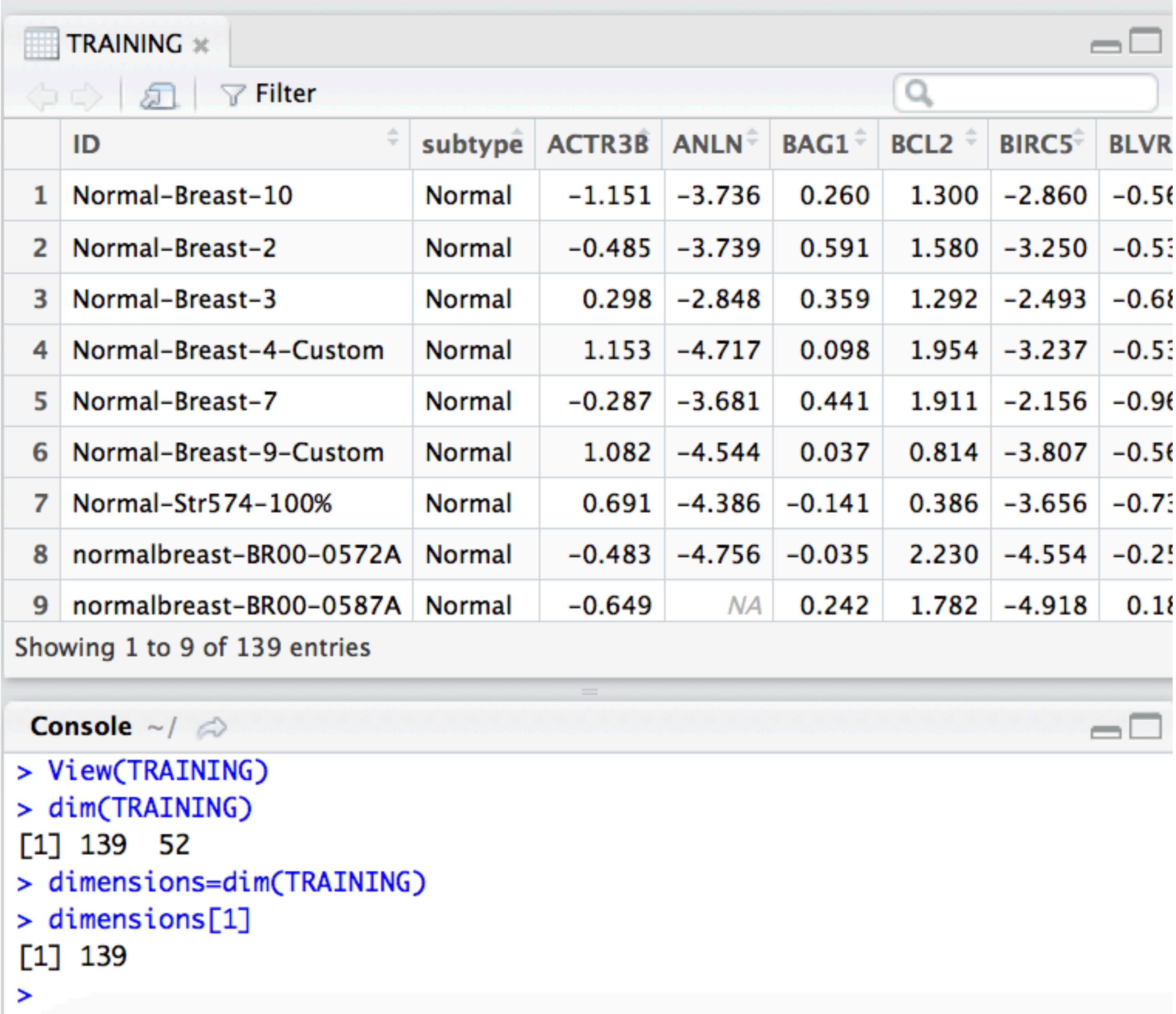


In R with training data

- Hierarchical clustering
- hclust function (from package stats)

```
>load(“GenomicClusteringLS.RData”)
>hccomplete=hclust(dist(TRAINING[,3:52]),
method = "complete")
>hcsingle=hclust(dist(TRAINING[,3:52]), method
= ”single")
>hccentroid=hclust(dist(TRAINING[,3:52]),
method = ”centroid")
```

TRAINING PAM50 data

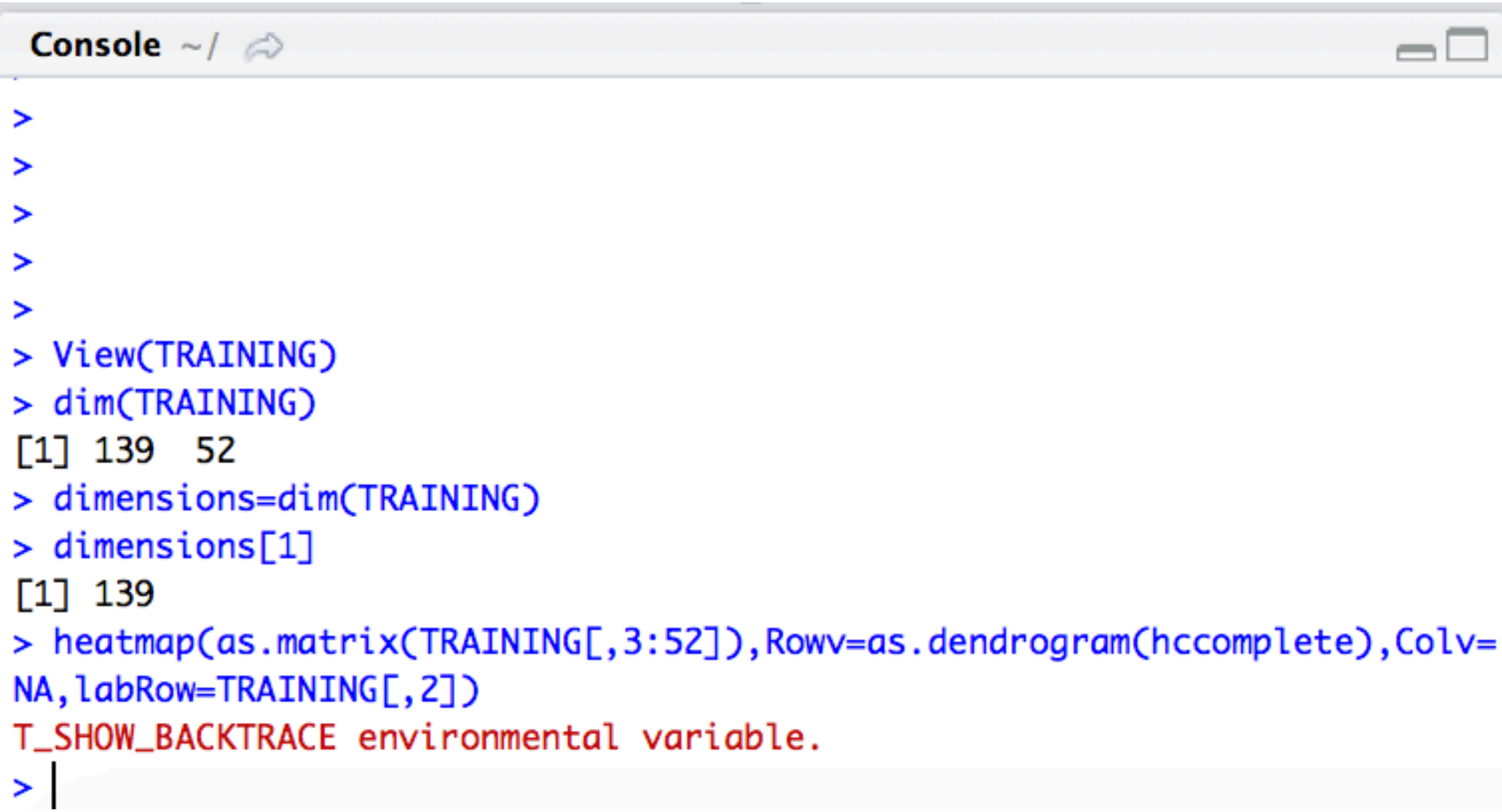


	ID	subtype	ACTR3B	ANLN	BAG1	BCL2	BIRC5	BLVR
1	Normal-Breast-10	Normal	-1.151	-3.736	0.260	1.300	-2.860	-0.56
2	Normal-Breast-2	Normal	-0.485	-3.739	0.591	1.580	-3.250	-0.53
3	Normal-Breast-3	Normal	0.298	-2.848	0.359	1.292	-2.493	-0.68
4	Normal-Breast-4-Custom	Normal	1.153	-4.717	0.098	1.954	-3.237	-0.53
5	Normal-Breast-7	Normal	-0.287	-3.681	0.441	1.911	-2.156	-0.96
6	Normal-Breast-9-Custom	Normal	1.082	-4.544	0.037	0.814	-3.807	-0.56
7	Normal-Str574-100%	Normal	0.691	-4.386	-0.141	0.386	-3.656	-0.73
8	normalbreast-BR00-0572A	Normal	-0.483	-4.756	-0.035	2.230	-4.554	-0.25
9	normalbreast-BR00-0587A	Normal	-0.649	*NA*	0.242	1.782	-4.918	0.18

Showing 1 to 9 of 139 entries

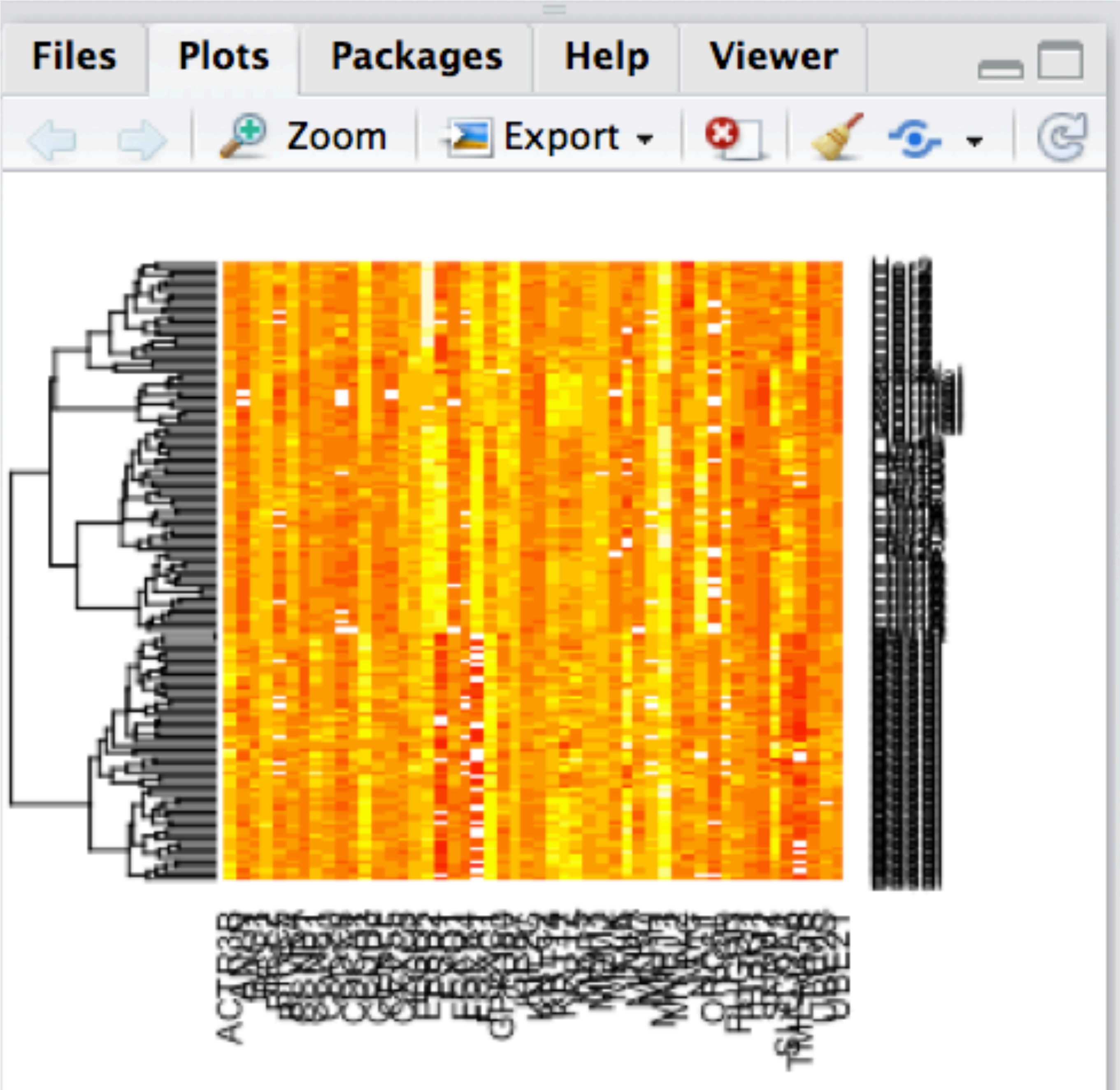
```
> View(TRAINING)
> dim(TRAINING)
[1] 139  52
> dimensions=dim(TRAINING)
> dimensions[1]
[1] 139
>
```

Visualization of clusters



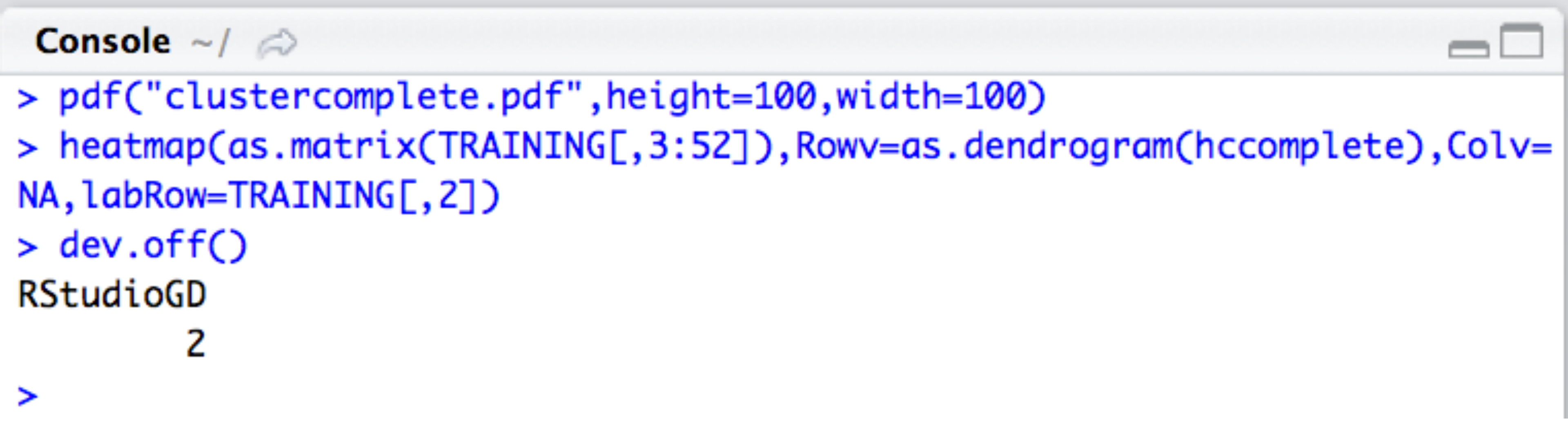
```
>
>
>
>
>
> View(TRAINING)
> dim(TRAINING)
[1] 139  52
> dimensions=dim(TRAINING)
> dimensions[1]
[1] 139
> heatmap(as.matrix(TRAINING[,3:52]),Rowv=as.dendrogram(hccomplete),Colv=
NA,labRow=TRAINING[,2])
T_SHOW_BACKTRACE environmental variable.
> |
```

Heatmap of hierarchical clustering with method= complete

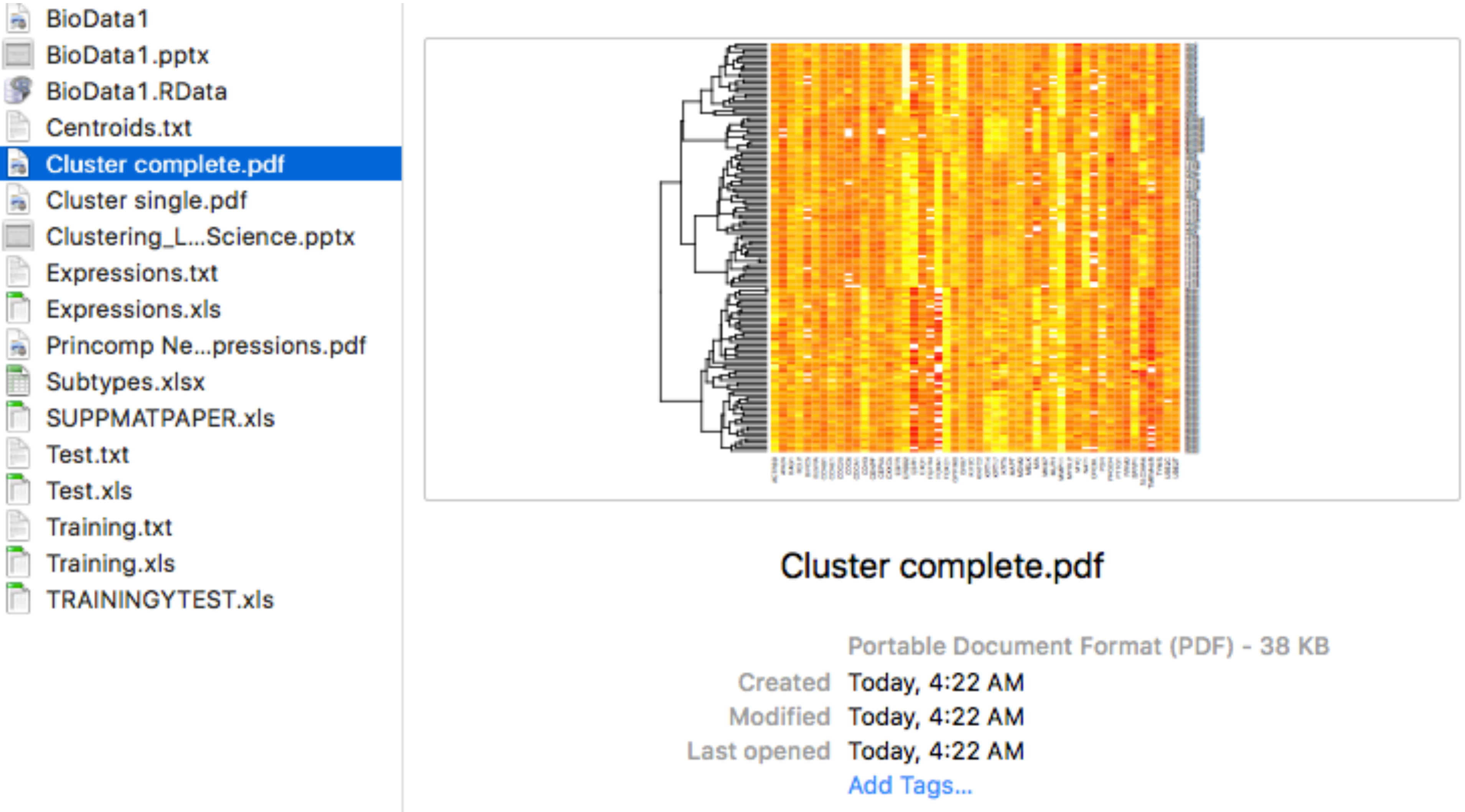


… Bigger

- Generation of files from command line (or export)



```
> pdf("clustercomplete.pdf",height=100,width=100)
> heatmap(as.matrix(TRAINING[,3:52]),Rowv=as.dendrogram(hccomplete),Colv=
NA,labRow=TRAINING[,2])
> dev.off()
RStudioGD
        2
>
```



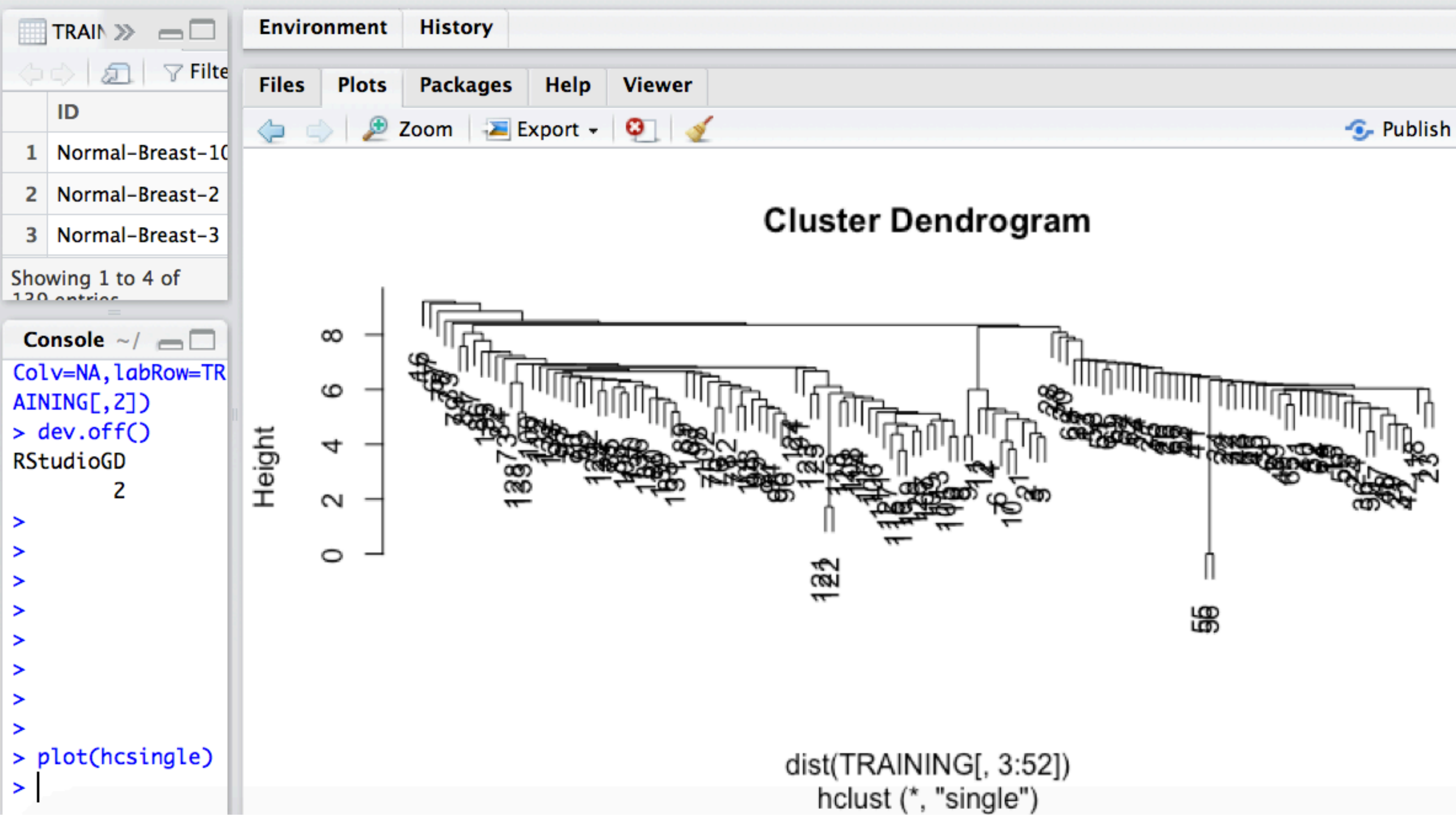
BioData1
BioData1.pptx
BioData1.RData
Centroids.txt
Cluster complete.pdf
Cluster single.pdf
Clustering_L...Science.pptx
Expressions.txt
Expressions.xls
Princomp Ne...pressions.pdf
Subtypes.xlsx
SUPPMATPAPER.xls
Test.txt
Test.xls
Training.txt
Training.xls
TRAININGYTEST.xls

Cluster complete.pdf

Portable Document Format (PDF) - 38 KB

Created	Today, 4:22 AM
Modified	Today, 4:22 AM
Last opened	Today, 4:22 AM

Dendrograms of clustering



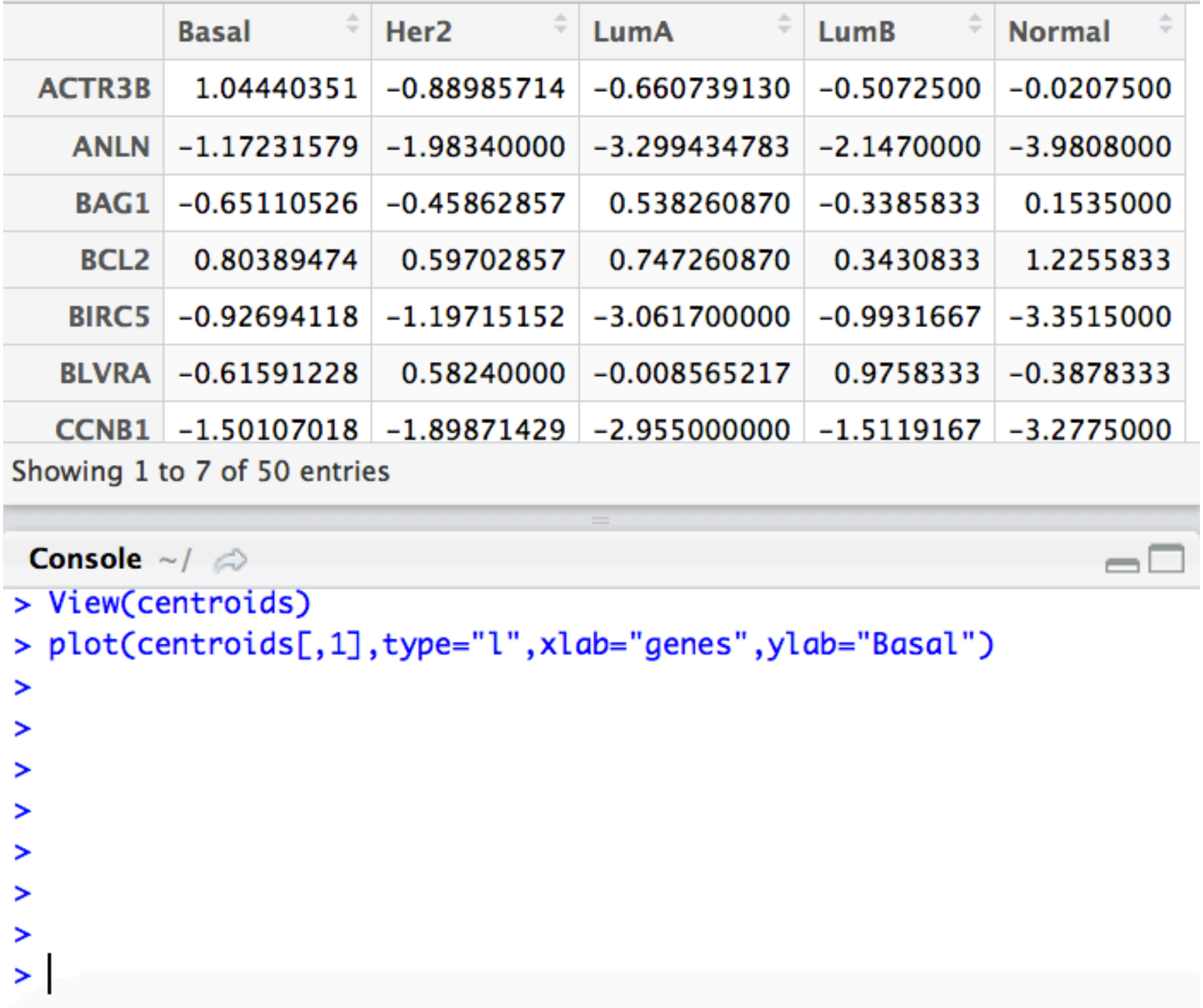
	ID
1	Normal-Breast-1
2	Normal-Breast-2
3	Normal-Breast-3

Showing 1 to 4 of 130 entries

Console ~/

```
Colv=NA,labRow=TR
AINING[,2])
> dev.off()
RStudioGD
        2
>
>
>
>
>
>
>
>
> plot(hcsingle)
>
```

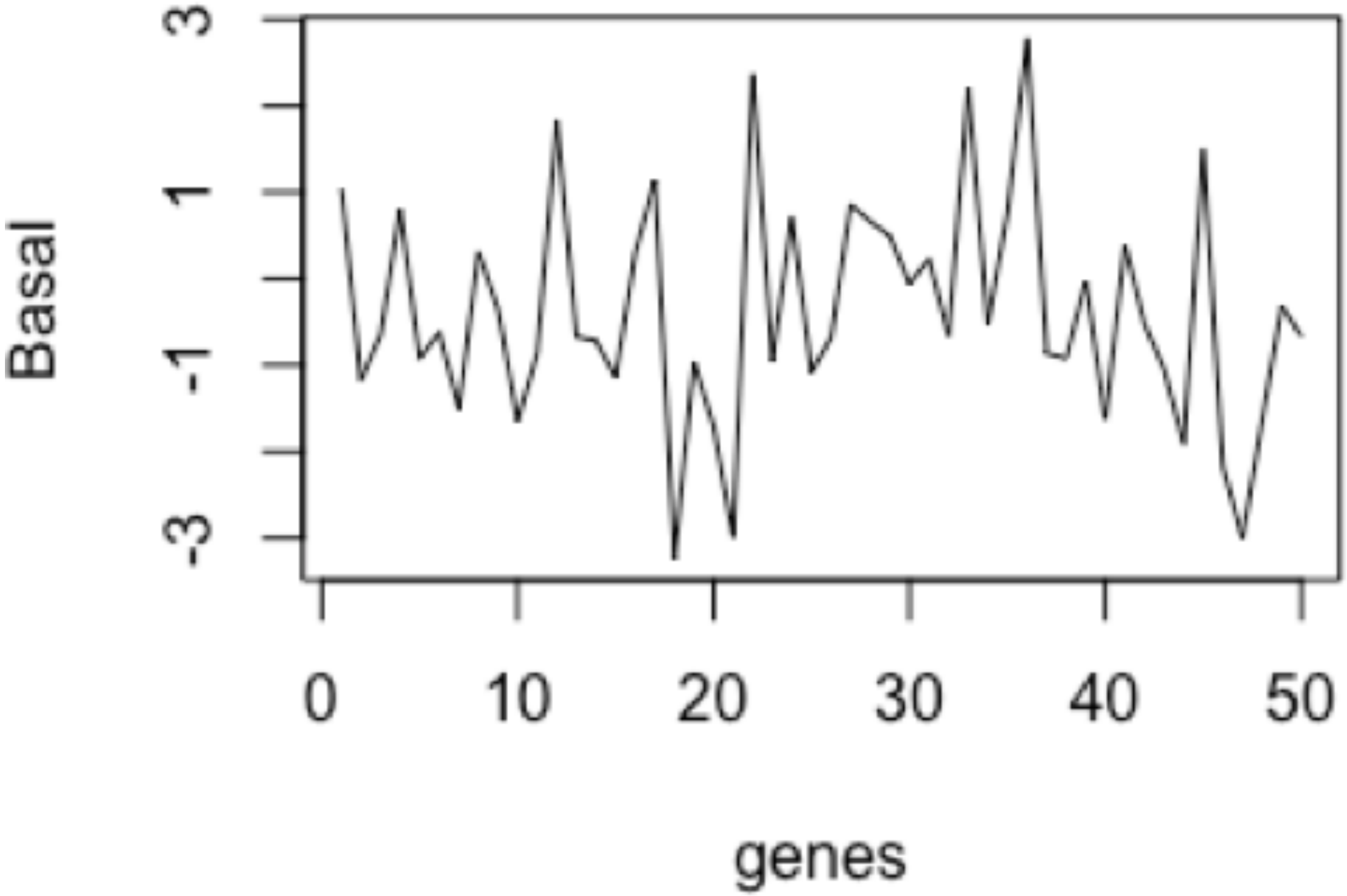
Centroids of clusters



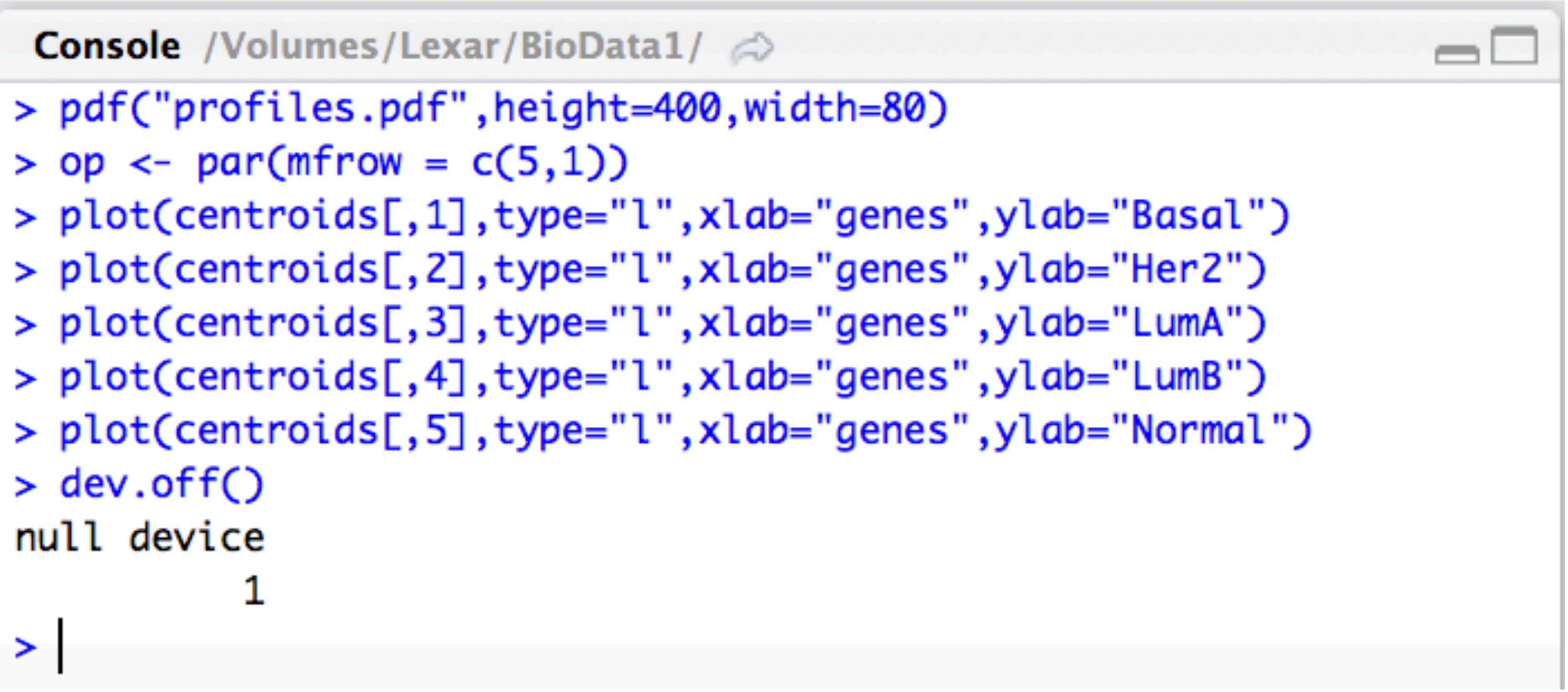
	Basal	Her2	LumA	LumB	Normal
ACTR3B	1.04440351	-0.88985714	-0.660739130	-0.5072500	-0.0207500
ANLN	-1.17231579	-1.98340000	-3.299434783	-2.1470000	-3.9808000
BAG1	-0.65110526	-0.45862857	0.538260870	-0.3385833	0.1535000
BCL2	0.80389474	0.59702857	0.747260870	0.3430833	1.2255833
BIRC5	-0.92694118	-1.19715152	-3.061700000	-0.9931667	-3.3515000
BLVRA	-0.61591228	0.58240000	-0.008565217	0.9758333	-0.3878333
CCNB1	-1.50107018	-1.89871429	-2.955000000	-1.5119167	-3.2775000

Showing 1 to 7 of 50 entries

```
> View(centroids)
> plot(centroids[,1],type="l",xlab="genes",ylab="Basal")
>
>
>
>
>
>
>
> |
```


Basal
3
1
-1
-3
0
10
20
30
40
50
genes

All the cluster profiles



```
> pdf("profiles.pdf",height=400,width=80)
> op <- par(mfrow = c(5,1))
> plot(centroids[,1],type="l",xlab="genes",ylab="Basal")
> plot(centroids[,2],type="l",xlab="genes",ylab="Her2")
> plot(centroids[,3],type="l",xlab="genes",ylab="LumA")
> plot(centroids[,4],type="l",xlab="genes",ylab="LumB")
> plot(centroids[,5],type="l",xlab="genes",ylab="Normal")
> dev.off()
null device
          1
> 
```

It was the same task again

- Exercise 1:

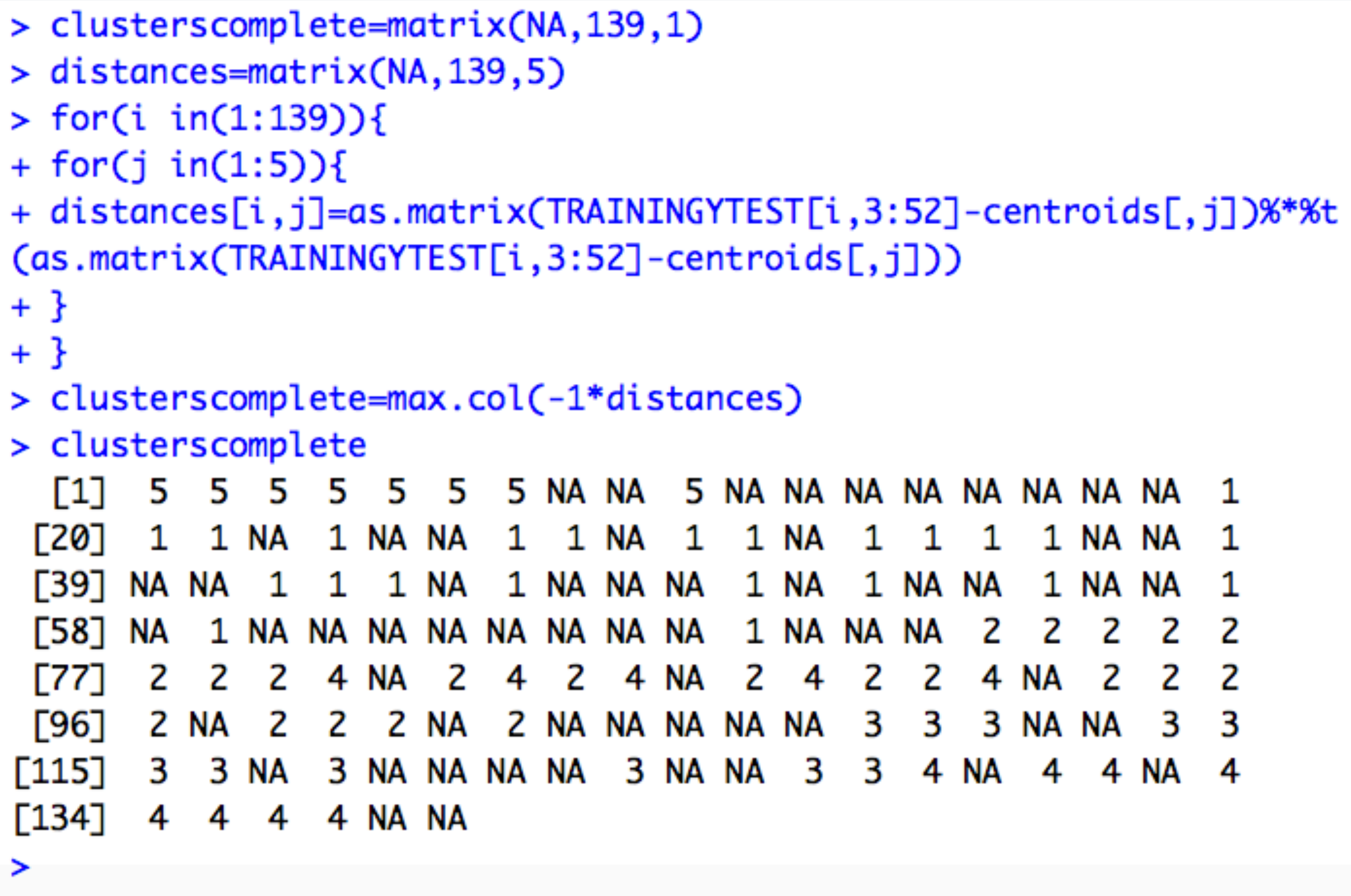
Do a R function (method) that receives the data.frames of centroids and plots the graphics.

Remember:

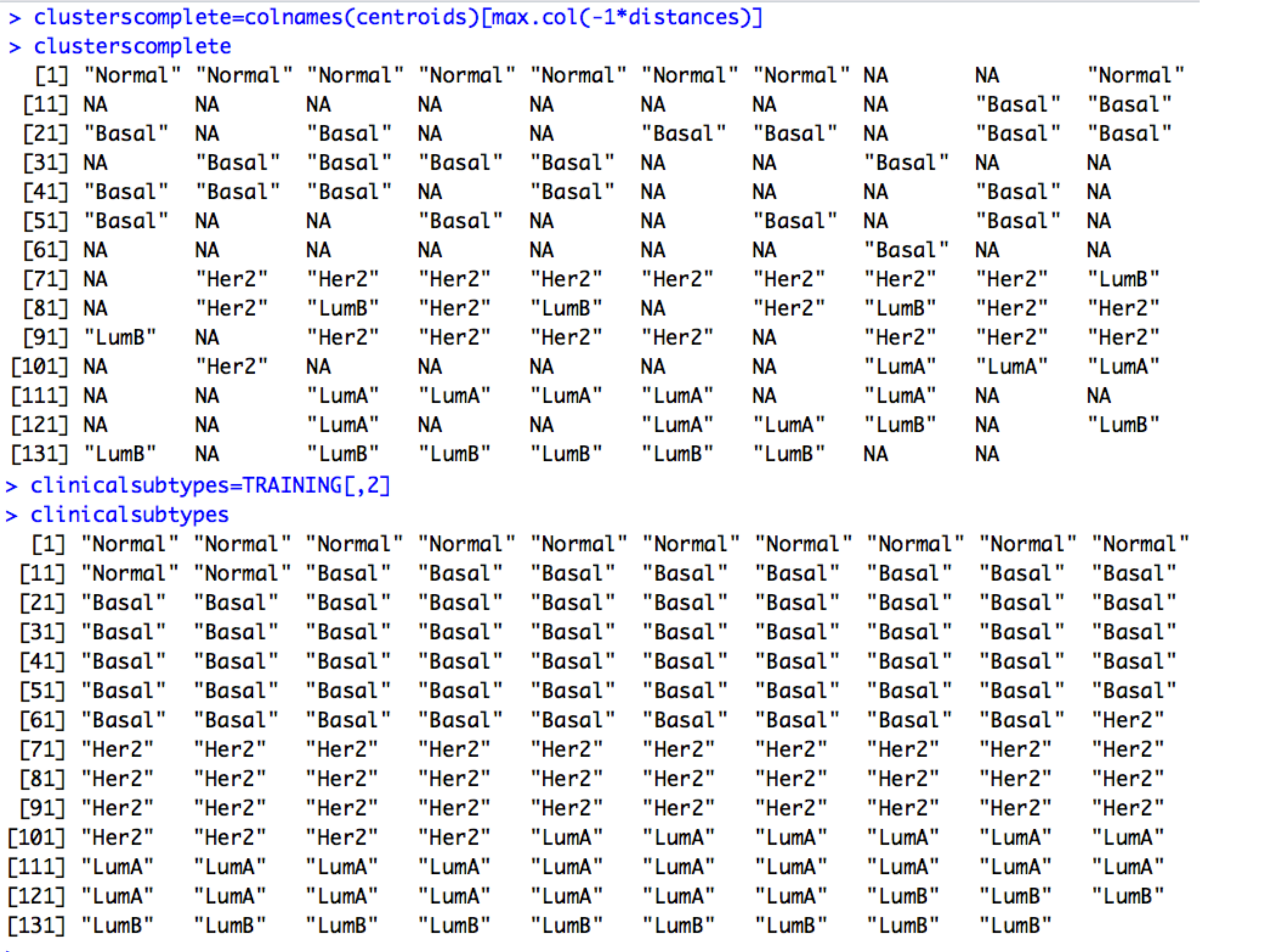
```
 functionname<-function(input1,
…,input2=defaultvalue2){

return(answer)
}
```

Given a set of patient profile assign the subtype



```
> clusterscomplete=matrix(NA,139,1)
> distances=matrix(NA,139,5)
> for(i in(1:139)){
+ for(j in(1:5)){
+ distances[i,j]=as.matrix(TRAININGYTEST[i,3:52]-centroids[,j])%*%t
(as.matrix(TRAININGYTEST[i,3:52]-centroids[,j]))
+ }
+ }
> clusterscomplete=max.col(-1*distances)
> clusterscomplete
  [1]  5  5  5  5  5  5  5 NA NA  5 NA NA NA NA NA NA NA NA  1
 [20]  1  1 NA  1 NA NA  1  1 NA  1  1 NA  1  1  1  1 NA NA  1
 [39] NA NA  1  1  1 NA  1 NA NA NA  1 NA  1 NA NA  1 NA NA  1
 [58] NA  1 NA NA NA NA NA NA NA NA  1 NA NA NA  2  2  2  2  2
 [77]  2  2  2  4 NA  2  4  2  4 NA  2  4  2  2  4 NA  2  2  2
 [96]  2 NA  2  2  2 NA  2 NA NA NA NA NA  3  3  3 NA NA  3  3
[115]  3  3 NA  3 NA NA NA NA  3 NA NA  3  3  4 NA  4  4 NA  4
[134]  4  4  4  4 NA NA
>
```

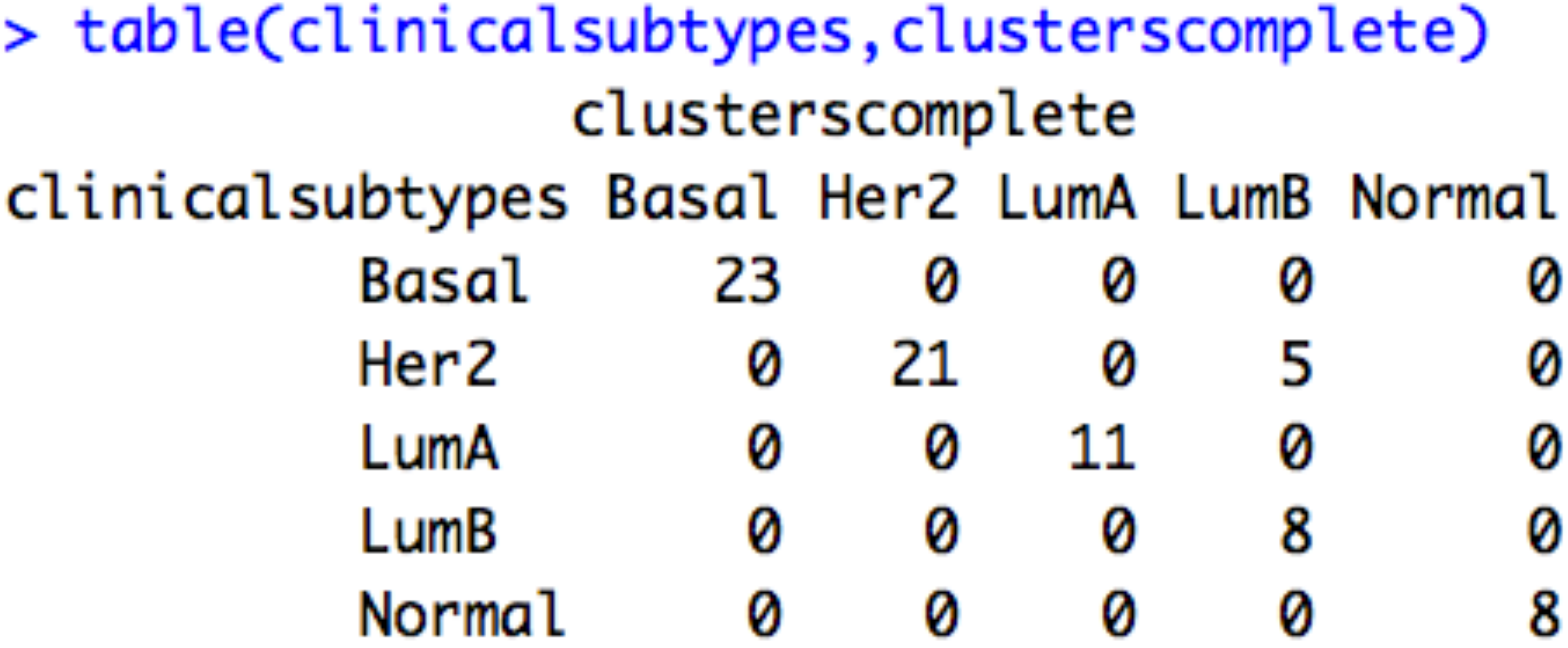


```
> clusterscomplete=colnames(centroids)[max.col(-1*distances)]
> clusterscomplete
  [1] "Normal" "Normal" "Normal" "Normal" "Normal" "Normal" "Normal" NA       NA       "Normal"
 [11] NA       NA       NA       NA       NA       NA       NA       NA       "Basal"  "Basal" 
 [21] "Basal"  NA       "Basal"  NA       NA       "Basal"  "Basal"  NA       "Basal"  "Basal" 
 [31] NA       "Basal"  "Basal"  "Basal"  "Basal"  NA       NA       "Basal"  NA       NA      
 [41] "Basal"  "Basal"  "Basal"  NA       "Basal"  NA       NA       NA       "Basal"  NA      
 [51] "Basal"  NA       NA       "Basal"  NA       NA       "Basal"  NA       "Basal"  NA      
 [61] NA       NA       NA       NA       NA       NA       NA       "Basal"  NA       NA      
 [71] NA       "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "LumB"  
 [81] NA       "Her2"   "LumB"   "Her2"   "LumB"   NA       "Her2"   "LumB"   "Her2"   "Her2"  
 [91] "LumB"   NA       "Her2"   "Her2"   "Her2"   "Her2"   NA       "Her2"   "Her2"   "Her2"  
[101] NA       "Her2"   NA       NA       NA       NA       NA       "LumA"   "LumA"   "LumA"  
[111] NA       NA       "LumA"   "LumA"   "LumA"   "LumA"   NA       "LumA"   NA       NA      
[121] NA       NA       "LumA"   NA       NA       "LumA"   "LumA"   "LumB"   NA       "LumB"  
[131] "LumB"   NA       "LumB"   "LumB"   "LumB"   "LumB"   "LumB"   NA       NA      
> clinicalsubtypes=TRAINING[,2]
> clinicalsubtypes
  [1] "Normal" "Normal" "Normal" "Normal" "Normal" "Normal" "Normal" "Normal" "Normal" "Normal"
 [11] "Normal" "Normal" "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal" 
 [21] "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal" 
 [31] "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal" 
 [41] "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal" 
 [51] "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal" 
 [61] "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Basal"  "Her2"  
 [71] "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "Her2"  
 [81] "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "Her2"  
 [91] "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "Her2"   "Her2"  
[101] "Her2"   "Her2"   "Her2"   "Her2"   "LumA"   "LumA"   "LumA"   "LumA"   "LumA"   "LumA"  
[111] "LumA"   "LumA"   "LumA"   "LumA"   "LumA"   "LumA"   "LumA"   "LumA"   "LumA"   "LumA"  
[121] "LumA"   "LumA"   "LumA"   "LumA"   "LumA"   "LumA"   "LumA"   "LumB"   "LumB"   "LumB"  
[131] "LumB"   "LumB"   "LumB"   "LumB"   "LumB"   "LumB"   "LumB"   "LumB"   "LumB"  
```

Exercise,2

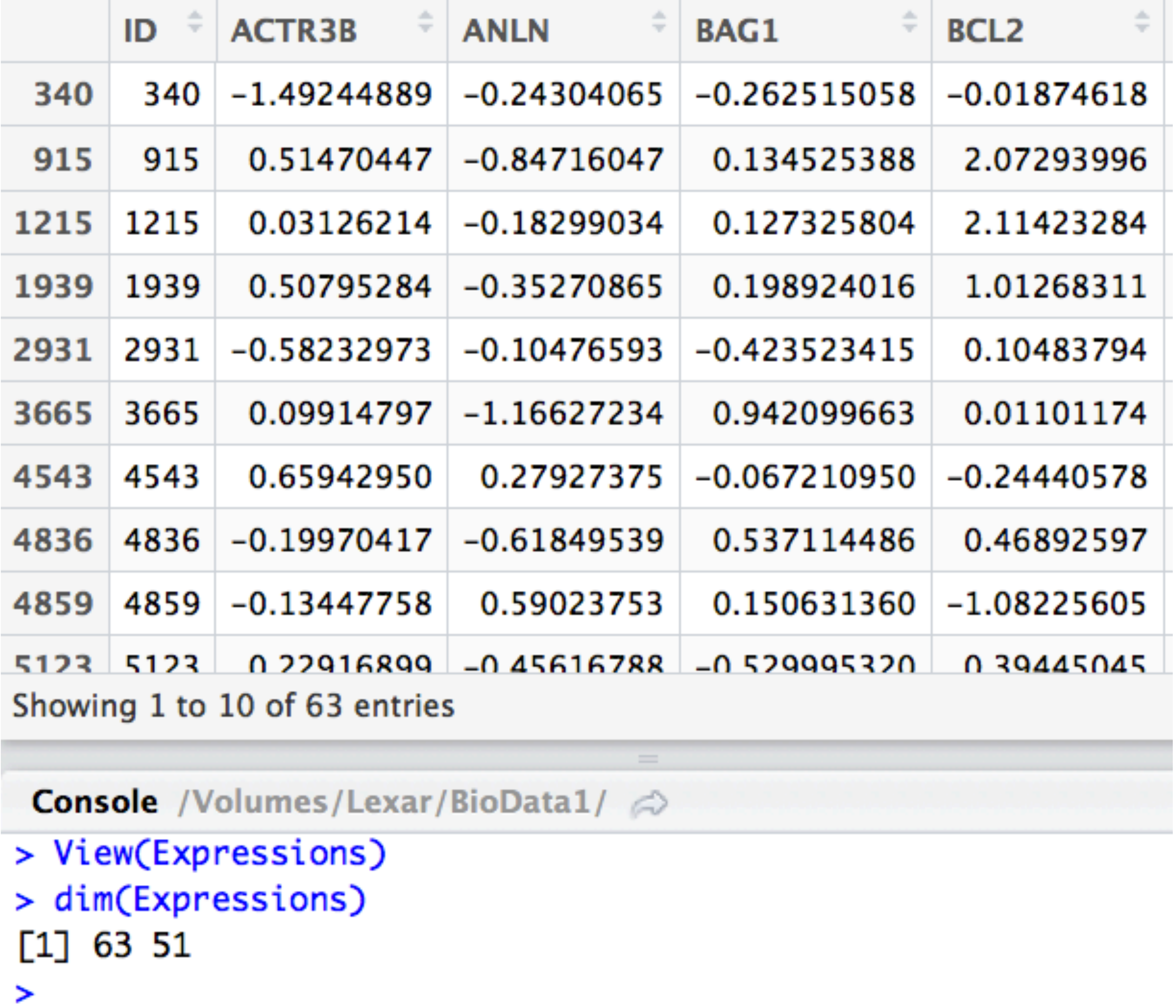
Do a R function that receives the data.frames of patient profiles and gives you assignations.

Comparing clinical annotations with predictions



```
> table(clinicalsubtypes,clusterscomplete)
                clusterscomplete
clinicalsubtypes Basal Her2 LumA LumB Normal
          Basal     23    0    0    0      0
          Her2       0   21    0    5      0
          LumA       0    0   11    0      0
          LumB       0    0    0    8      0
          Normal     0    0    0    0      8
```

New data: Chilean patients

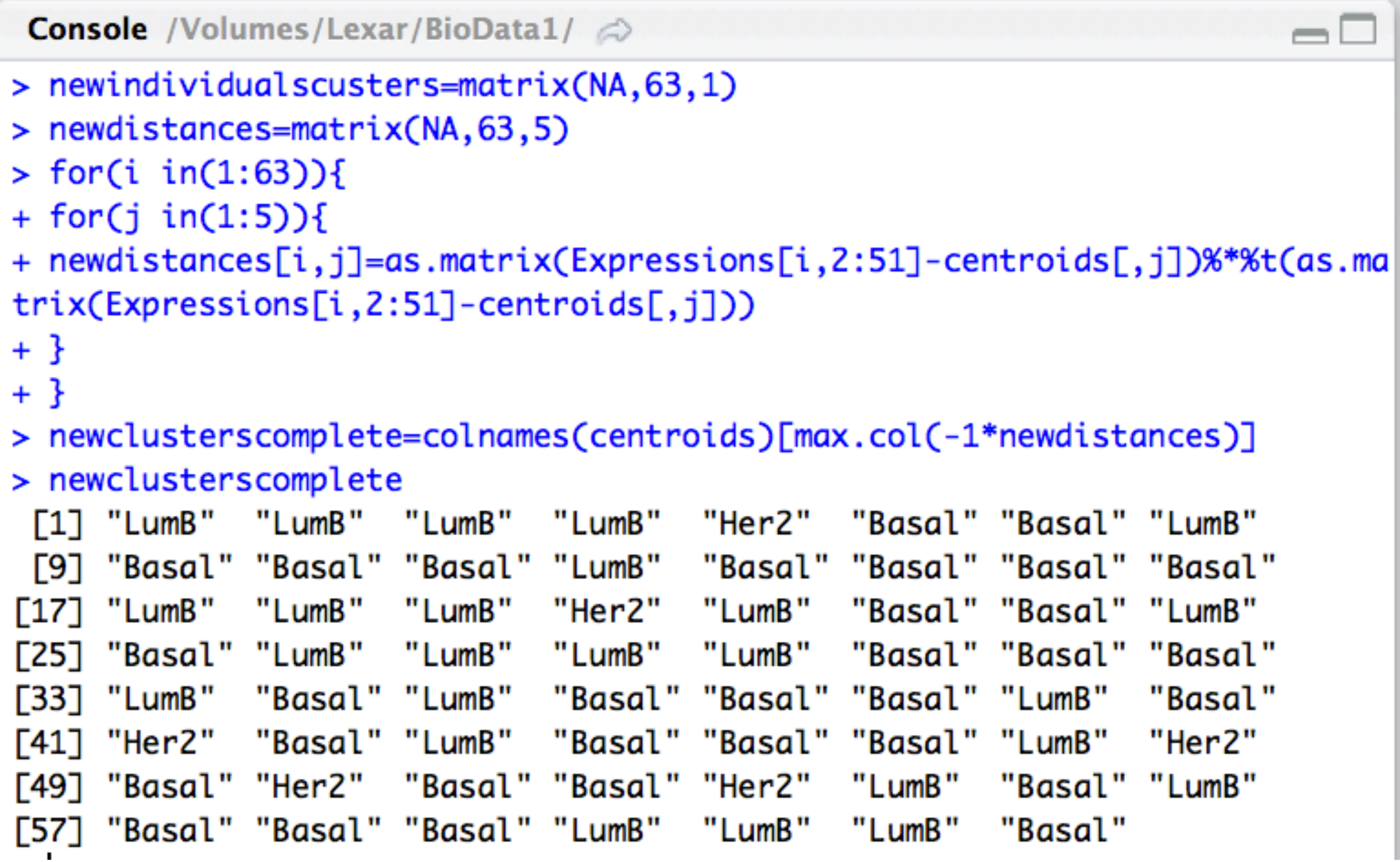


	ID	ACTR3B	ANLN	BAG1	BCL2
340	340	-1.49244889	-0.24304065	-0.262515058	-0.01874618
915	915	0.51470447	-0.84716047	0.134525388	2.07293996
1215	1215	0.03126214	-0.18299034	0.127325804	2.11423284
1939	1939	0.50795284	-0.35270865	0.198924016	1.01268311
2931	2931	-0.58232973	-0.10476593	-0.423523415	0.10483794
3665	3665	0.09914797	-1.16627234	0.942099663	0.01101174
4543	4543	0.65942950	0.27927375	-0.067210950	-0.24440578
4836	4836	-0.19970417	-0.61849539	0.537114486	0.46892597
4859	4859	-0.13447758	0.59023753	0.150631360	-1.08225605
5123	5123	0.22916899	-0.45616788	-0.529995320	0.39445045

Showing 1 to 10 of 63 entries

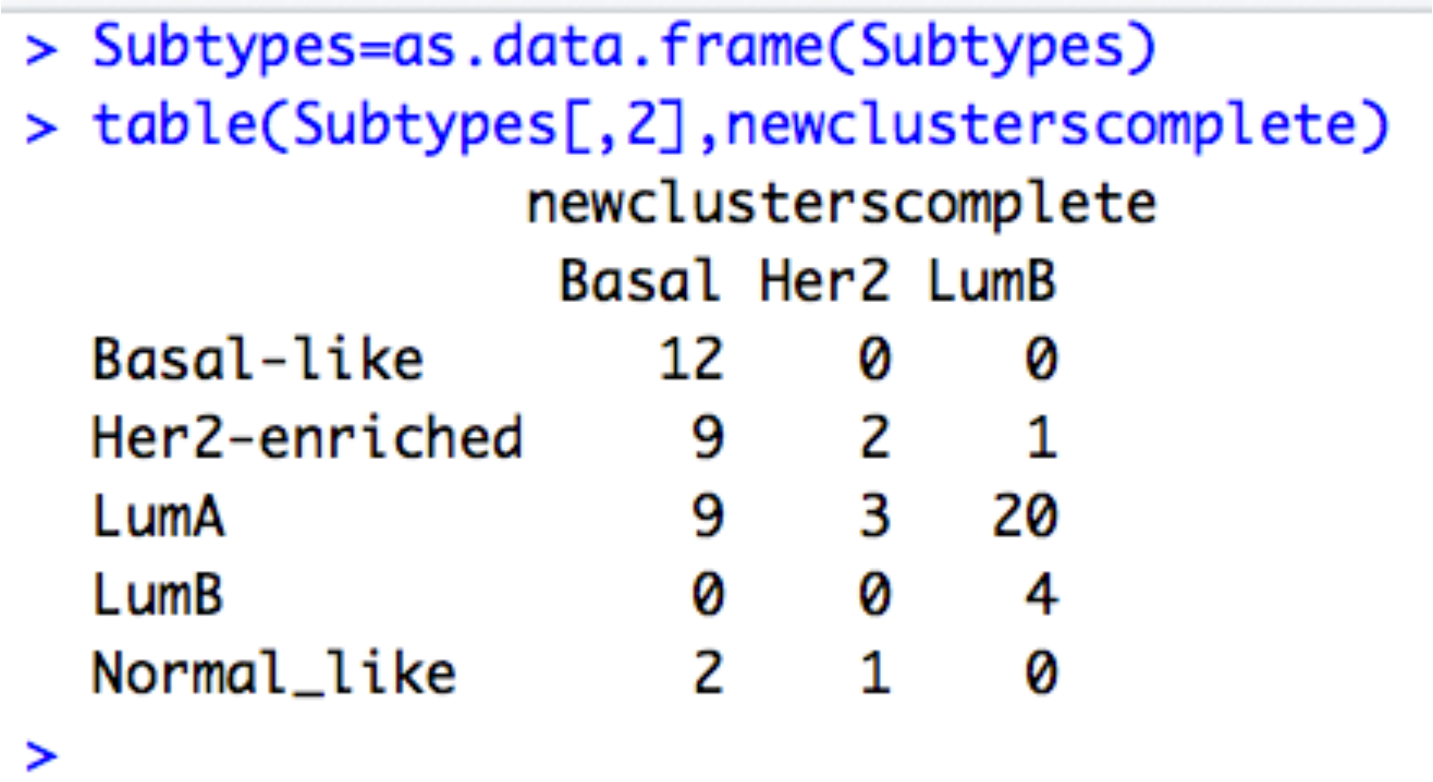
```
> View(Expressions)
> dim(Expressions)
[1] 63 51
>
```

Assignations



```
> newindividualscusters=matrix(NA,63,1)
> newdistances=matrix(NA,63,5)
> for(i in(1:63)){
+ for(j in(1:5)){
+ newdistances[i,j]=as.matrix(Expressions[i,2:51]-centroids[,j])%*%t(as.ma
trix(Expressions[i,2:51]-centroids[,j]))
+ }
+ }
> newclusterscomplete=colnames(centroids)[max.col(-1*newdistances)]
> newclusterscomplete
 [1] "LumB"  "LumB"  "LumB"  "LumB"  "Her2"  "Basal" "Basal" "LumB" 
 [9] "Basal" "Basal" "Basal" "LumB"  "Basal" "Basal" "Basal" "Basal"
[17] "LumB"  "LumB"  "LumB"  "Her2"  "LumB"  "Basal" "Basal" "LumB" 
[25] "Basal" "LumB"  "LumB"  "LumB"  "LumB"  "Basal" "Basal" "Basal"
[33] "LumB"  "Basal" "LumB"  "Basal" "Basal" "Basal" "LumB"  "Basal"
[41] "Her2"  "Basal" "LumB"  "Basal" "Basal" "Basal" "LumB"  "Her2" 
[49] "Basal" "Her2"  "Basal" "Basal" "Her2"  "LumB"  "Basal" "LumB" 
[57] "Basal" "Basal" "Basal" "LumB"  "LumB"  "LumB"  "Basal"
```

Comparing clinical annotations with predictions



```
> Subtypes=as.data.frame(Subtypes)
> table(Subtypes[,2],newclusterscomplete)
               newclusterscomplete
                Basal Her2 LumB
  Basal-like       12    0    0
  Her2-enriched     9    2    1
  LumA              9    3   20
  LumB              0    0    4
  Normal_like       2    1    0
>
```

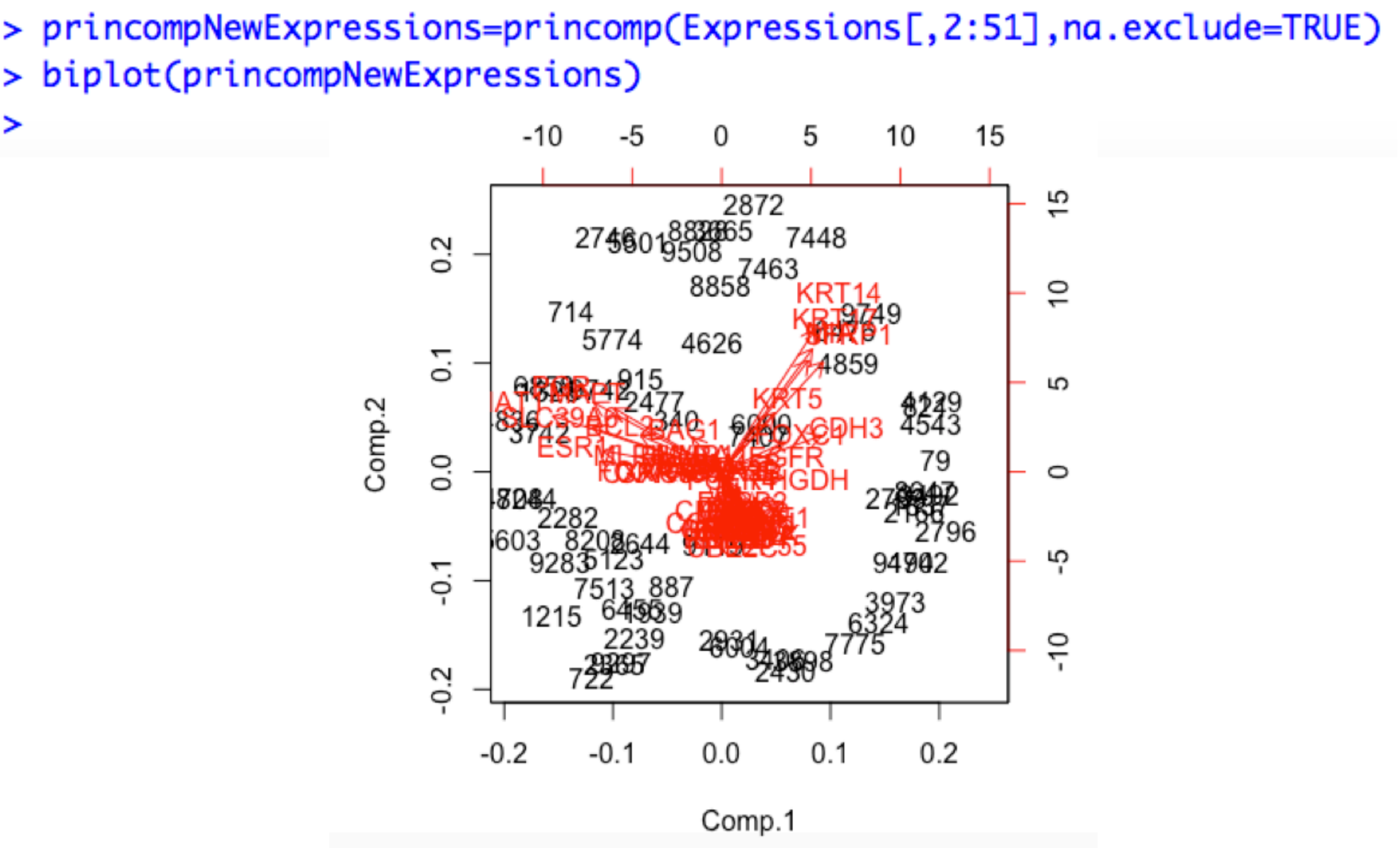
It was the same task again

- Exercise 3:

Do a R function that receives the two columns of labels and counts the percentaje of coincidences.

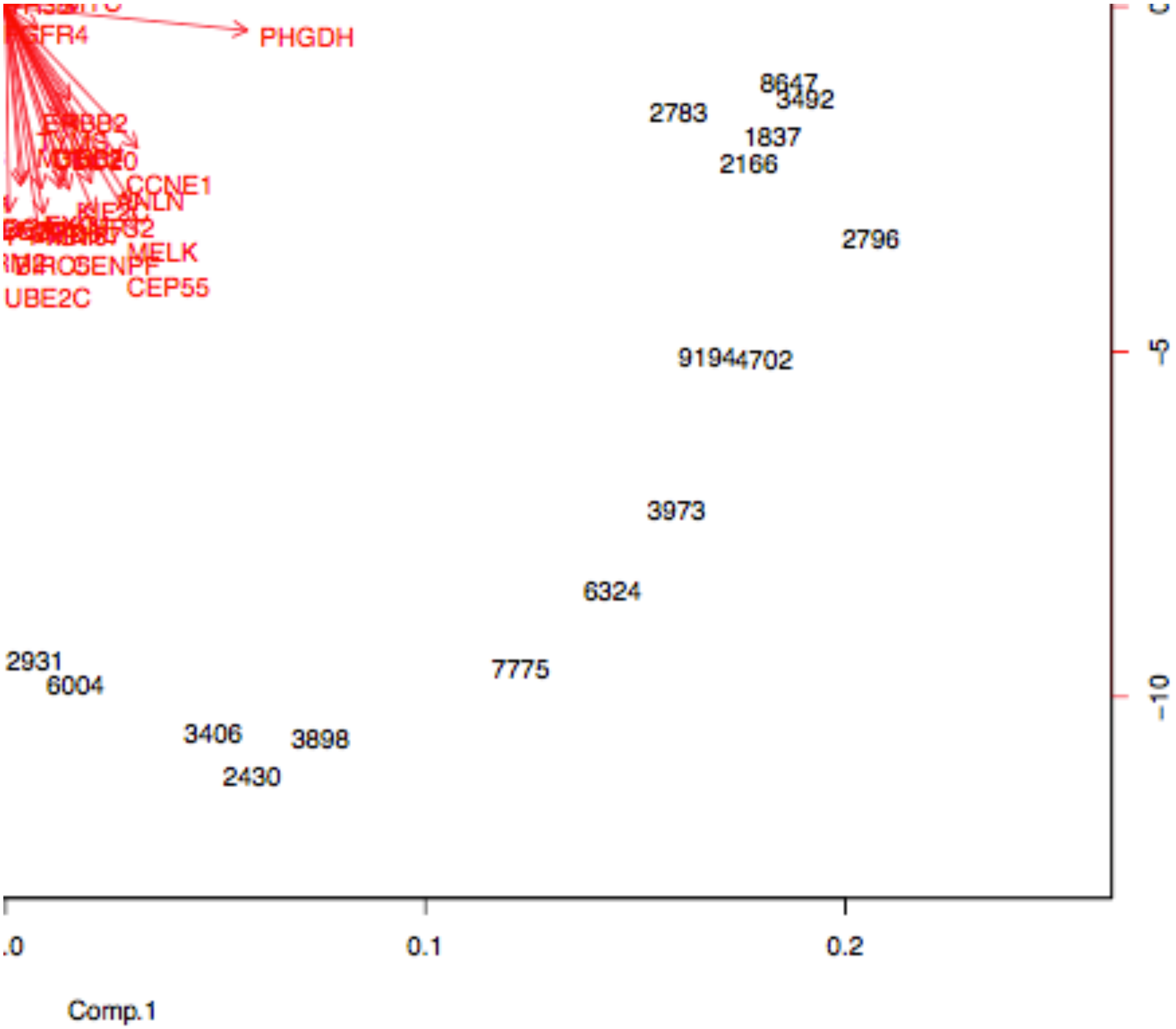
Which one works the best (try with all the difference alternatives) for our data (Chilean).

Principal Component Analysis



```
> princompNewExpressions=princomp(Expressions[,2:51],na.exclude=TRUE)
> biplot(princompNewExpressions)
>
```

… Bigger, zooming: Her2 quadrant



END

FIN

GRACIAS

THANKS

¿PREGUNTAS?

QUESTIONS?