Università degli Studi di Catania
Dipartimento di Matematica e Informatica
Corso di Laurea in Informatica Magistrale

Aparo Antonino

Estrazione di Biomarcatori e clustering

Relazione progetto Bioinformartica

Prof. Ferro

Anno Accademico 2015/2016

Indice

1. Introduzione

L'obiettivo del progetto è l'estrazione di biomarcatori e il clustering dei dati di uno specifico dataset fornito.
Il dataset contiene informazioni riguardanti le visite medica di pazienti con età compresa tra 1 anno e 17 anni al fine di diagnosticare l'autismo.

Il progetto è stato implementato utilizzando il **linguaggio R**, in particolar modo sono stati utilizzati i seguenti package:

- ***limma***, ***pamr***, ***Biobase*** e ***gplot*** per la ricerca e visualizzazione dei biomarcatori partendo da valori di concentrazione di composti molecolari;
- ***cluster*** e ***fpc*** per la ricerca e l'analisi dei cluster;
- ***ggplot2*** per la creazione di plot utili alla comprensione dei dati e dei risultati.

E' possibile scaricare l'intero elaborato cliccando nel seguente link.

2. Analisi del dataset

E' stato fornito un file *xslx* dal quale sono stati esportati due *csv*:

- "**dati**" contenente le informazioni sulla visita; tra queste le più rilevanti troviamo la data del controllo, il tipo di paziente (sano o patologico), e la sua età; inoltre per i pazienti patologici sono riportati anche i risultati dei test ADIR e ADOS e la comunità di alloggio Socio Sanitaria (CSS) alla quale appartengono.
- "**concentrazioni**" dove sono presenti le concentrazioni riscontrate durante la visita.

Il dataset è stato manipolato per modificare alcuni *tipi* assegnati di default da R, escludere le osservazioni contenenti dati non pronti (in "dati") ed effettuare una selezione delle concentrazioni per eliminare quelle di pazienti non presenti nell'altra tabella. Successivamente per effettuare il clustering è stata creata un unica tabella chiamata "**db**" effettuata tramite operazione di **JOIN** tra le tabelle dati e concentrazioni.

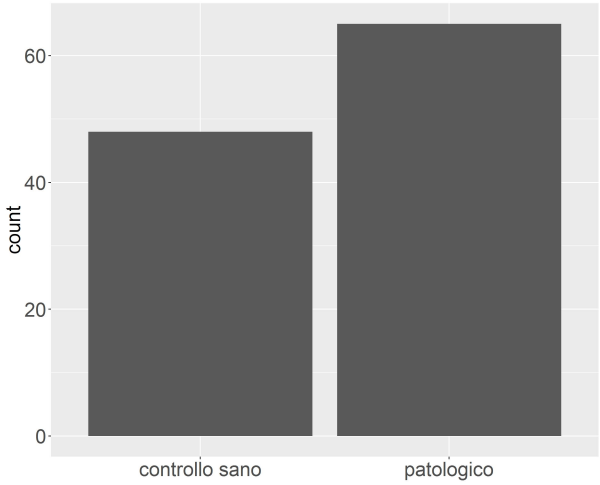
Sommari statistici e plot

Di seguito è riportato un breve sommario statistico della tabella *dati* (clinici).

```
  sample.id             dob                  eta                    tipo    regressione    rm
Length:113         Min.   :1997-08-17   Min.   : 1.000   controllo sano:48   No  :41     No  :16
Class :character   1st Qu.:2006-06-17   1st Qu.: 3.000   patologico    :65   Si  :23     Si  :49
Mode  :character   Median :2009-07-14   Median : 6.000                       Si  : 1     NA's:48
                   Mean   :2008-12-01   Mean   : 6.589                       NA's:48
                   3rd Qu.:2012-02-18   3rd Qu.: 9.000
                   Max.   :2015-06-20   Max.   :17.000
                   NA's   :2            NA's   :1
    ados.a          ados.b          ados.c          ados.d          adir.a          adir.b
Min.   : 0.000   Min.   : 0.000   Min.   :0.000   Min.   :0.000   Min.   : 0.00   Min.   : 0.000
1st Qu.: 2.000   1st Qu.: 5.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 9.00   1st Qu.: 5.000
Median : 5.000   Median : 8.000   Median :2.000   Median :2.000   Median :13.00   Median : 8.000
Mean   : 4.508   Mean   : 8.415   Mean   :2.185   Mean   :1.908   Mean   :13.28   Mean   : 8.369
3rd Qu.: 6.000   3rd Qu.:12.000   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:17.00   3rd Qu.:12.000
Max.   :12.000   Max.   :15.000   Max.   :6.000   Max.   :5.000   Max.   :28.00   Max.   :16.000
NA's   :48       NA's   :48       NA's   :48      NA's   :48      NA's   :48      NA's   :48
    adir.c           qit             css
Min.   : 0.000   Min.   : 40.00   0      :48
1st Qu.: 3.000   1st Qu.: 58.00   7      :15
Median : 5.000   Median : 77.50   6      :14
Mean   : 5.477   Mean   : 76.04   8      :11
3rd Qu.: 8.000   3rd Qu.: 90.75   4      : 7
Max.   :16.000   Max.   :120.00   1      : 4
NA's   :48       NA's   :87       (Other):14
```

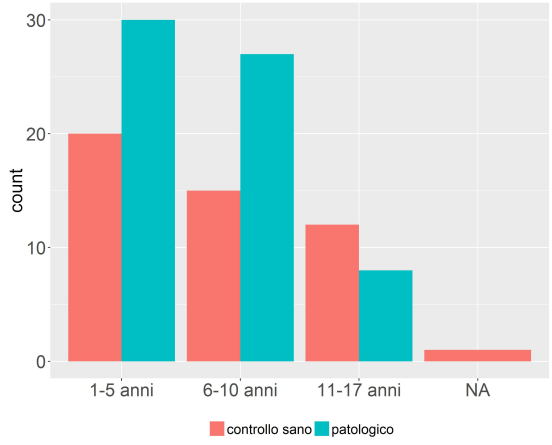
Si evince che le visite sono stati effettuati tra il 17/08/1997 e il 20/06/2015, a 48 pazienti sani e 65 patologici di età compresa tra 1 e 17 anni con un Quoziente Intellettivo (*qit*) medio di 76,04 assegnati a differenti centri CSS.

E' stato utilizzato il package *ggplot2 per* i seguenti grafici (istogrammi e barplot).

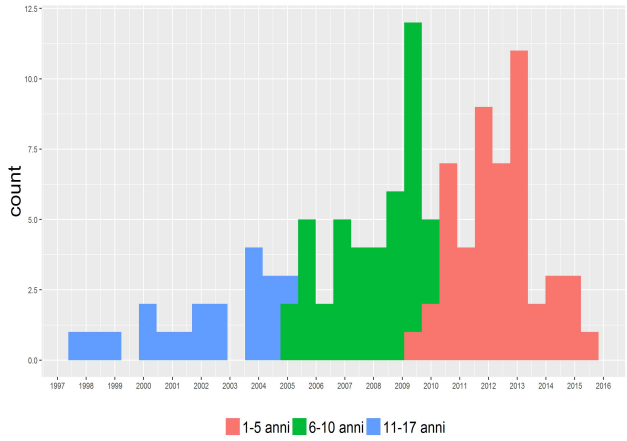


Per una migliore visualizzazione in alcuni dei grafici successivi sono state raggruppate le età dei pazienti in 3 gruppi: 1-5 anni, 6-10 anni,11-7 anni.

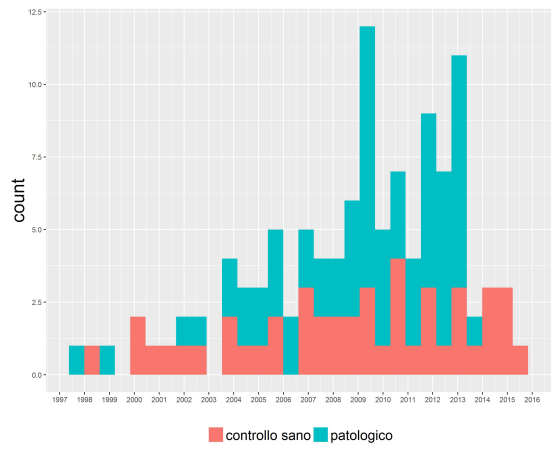
Barplot età in relazione al tipo



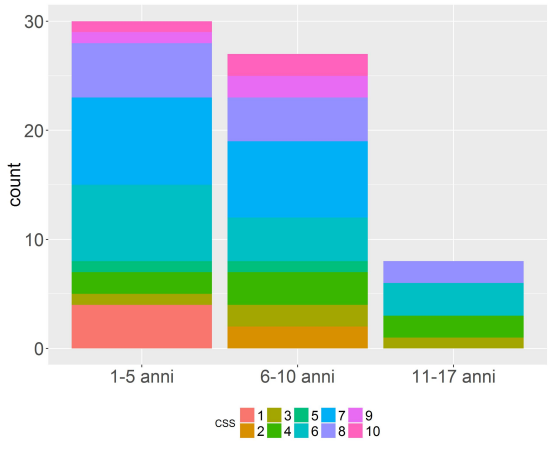
Istogramma date visite in relazione all'età



Istogramma date visite in relazione al tipo



Barplot età dei patologici in relazione al css



Dall'osservazione dei grafici possiamo facilmente dedurre che:

- La visita è stata fatta principalmente a bambini compresi tra 1 e 5 anni, mentre in minoranza sono i pazienti dagli 11 ai 17 anni; inoltre mentre per la fascia di età dagli 11 ai 17 anni sono maggiori i controlli sani (anche se di poco) per le altre 2 fasce sono maggiori i patologici.

- L'età dei pazienti che effettuano la visita diminuisce con l'aumentare degli anni; si osserva infatti che nell'ultimo periodo in cui si sono registrate i dati in possesso, i controlli sono stati principalmente fatti a bambini tra 1 anno e 5 anni, mentre nei primi anni principalmente ad adolescenti.

- Il rapporto tra il numero di pazienti sani e patologici che si sottopongono alla visita di fatto non è cambiato nel corso del tempo, anche se negli ultimi anni presenti si registrano solo pazienti sani.

- L'età dei pazienti è ben distribuita nei vari centri css.

A seguire viene riportato il sommario statistico della tabella *concentrazioni*.

```
 sample.id             ALA               ARG               CIT               GLY        LEU\\ILE\\PRO-OH
Length:129         Min.   : 84.56   Min.   : 0.00   Min.   : 1.03   Min.   :  0.0   Min.   : 72.36
Class :character   1st Qu.:161.11   1st Qu.: 9.32   1st Qu.:19.36   1st Qu.:162.8   1st Qu.:110.46
Mode  :character   Median :186.70   Median :17.01   Median :24.10   Median :206.2   Median :124.61
                   Mean   :197.90   Mean   :18.48   Mean   :24.00   Mean   :206.3   Mean   :134.61
                   3rd Qu.:230.47   3rd Qu.:24.14   3rd Qu.:28.02   3rd Qu.:245.7   3rd Qu.:154.12
                   Max.   :497.63   Max.   :55.42   Max.   :45.25   Max.   :484.3   Max.   :260.40
   Leu/Pro            MET               ORN               PHE            PHE/TYR            PRO
Min.   :0.240   Min.   : 0.00   Min.   : 28.96   Min.   :25.37   Min.   :0.0000   Min.   : 31.61
1st Qu.:0.840   1st Qu.: 9.92   1st Qu.: 64.84   1st Qu.:40.75   1st Qu.:0.7300   1st Qu.:120.21
Median :1.060   Median :12.48   Median : 74.75   Median :45.07   Median :0.8300   Median :143.89
Mean   :1.128   Mean   :12.62   Mean   : 78.63   Mean   :47.02   Mean   :0.8844   Mean   :159.86
3rd Qu.:1.290   3rd Qu.:15.05   3rd Qu.: 89.84   3rd Qu.:51.41   3rd Qu.:0.9900   3rd Qu.:188.15
Max.   :4.290   Max.   :30.46   Max.   :152.39   Max.   :89.64   Max.   :2.3300   Max.   :419.39
      SA              TYR               VAL               C0              C2              C3
Min.   :0.0000   Min.   :  0.00   Min.   : 91.41   Min.   : 8.32   Min.   : 3.31   Min.   :0.390
1st Qu.:0.6575   1st Qu.: 45.99   1st Qu.:130.34   1st Qu.:16.74   1st Qu.: 7.81   1st Qu.:0.880
Median :0.9400   Median : 55.32   Median :151.61   Median :19.55   Median :10.96   Median :1.130
Mean   :0.9520   Mean   : 57.54   Mean   :156.24   Mean   :20.07   Mean   :10.82   Mean   :1.247
3rd Qu.:1.2450   3rd Qu.: 68.69   3rd Qu.:177.08   3rd Qu.:23.32   3rd Qu.:13.01   3rd Qu.:1.490
Max.   :1.8500   Max.   :115.67   Max.   :279.75   Max.   :35.38   Max.   :26.51   Max.   :6.910
NA's   :1
      C4           C4OH\\C3DC            C5              C5:1           C5DC\\C6OH        C5OH\\C4DC
Min.   :0.0000   Min.   :0.00000   Min.   :0.01000   Min.   :0.000000   Min.   :0.00000   Min.   :0.1000
1st Qu.:0.1000   1st Qu.:0.05000   1st Qu.:0.06000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.3000
Median :0.1500   Median :0.06000   Median :0.08000   Median :0.010000   Median :0.02000   Median :0.3800
Mean   :0.1668   Mean   :0.06969   Mean   :0.08798   Mean   :0.009223   Mean   :0.03101   Mean   :0.3879
3rd Qu.:0.2100   3rd Qu.:0.09000   3rd Qu.:0.10000   3rd Qu.:0.010000   3rd Qu.:0.06000   3rd Qu.:0.4700
Max.   :0.6600   Max.   :0.33000   Max.   :0.27000   Max.   :0.040000   Max.   :0.17000   Max.   :0.7200
                                                     NA's   :26
```

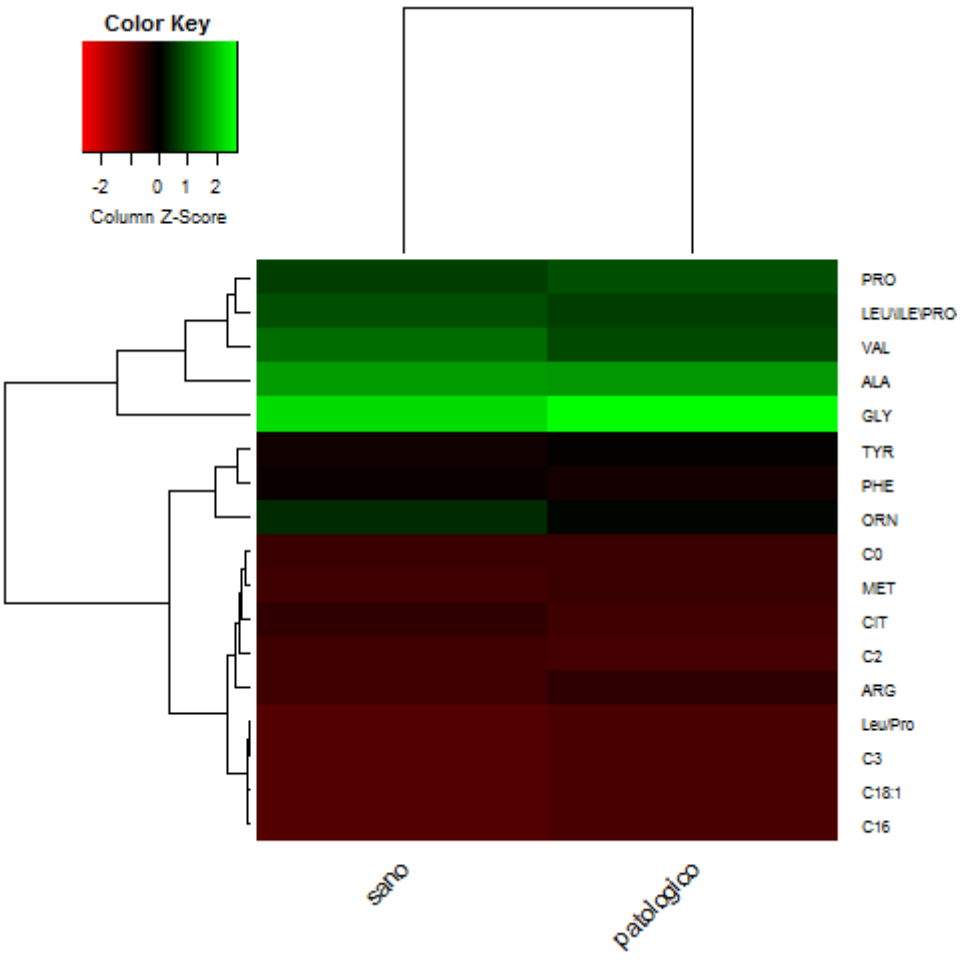
```
       C6                C6DC              C8                C8:1              C10               C10:1
 Min.   :0.00000   Min.   :0.0100   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.02000
 1st Qu.:0.02000   1st Qu.:0.1100   1st Qu.:0.0400   1st Qu.:0.04000   1st Qu.:0.0700   1st Qu.:0.06000
 Median :0.03000   Median :0.1400   Median :0.0600   Median :0.05000   Median :0.0900   Median :0.07000
 Mean   :0.03566   Mean   :0.1429   Mean   :0.0738   Mean   :0.05628   Mean   :0.1122   Mean   :0.07977
 3rd Qu.:0.05000   3rd Qu.:0.1800   3rd Qu.:0.0900   3rd Qu.:0.06000   3rd Qu.:0.1300   3rd Qu.:0.10000
 Max.   :0.10000   Max.   :0.3200   Max.   :0.3400   Max.   :0.37000   Max.   :0.6000   Max.   :0.18000
     C10:2             C12              C12:1             C14              C14-OH            C14:1
 Min.   :0.00000   Min.   :0.00000   Min.   :0.01000   Min.   :0.03000   Min.   :0.00000   Min.   :0.01000
 1st Qu.:0.00000   1st Qu.:0.03000   1st Qu.:0.04000   1st Qu.:0.06000   1st Qu.:0.01000   1st Qu.:0.05000
 Median :0.01000   Median :0.04000   Median :0.05000   Median :0.08000   Median :0.01000   Median :0.06000
 Mean   :0.00592   Mean   :0.04473   Mean   :0.05364   Mean   :0.08326   Mean   :0.00814   Mean   :0.06713
 3rd Qu.:0.01000   3rd Qu.:0.06000   3rd Qu.:0.06000   3rd Qu.:0.10000   3rd Qu.:0.01000   3rd Qu.:0.08000
 Max.   :0.04000   Max.   :0.15000   Max.   :0.20000   Max.   :0.17000   Max.   :0.02000   Max.   :0.22000
 NA's   :80
     C14:2             C16              C16-OH            C16:1             C16:1-OH          C18
 Min.   :0.00000   Min.   :0.450   Min.   :0.00000   Min.   :0.02000   Min.   :0.01000   Min.   :0.2300
 1st Qu.:0.01000   1st Qu.:0.860   1st Qu.:0.01000   1st Qu.:0.04000   1st Qu.:0.03000   1st Qu.:0.4800
 Median :0.02000   Median :1.060   Median :0.01000   Median :0.05000   Median :0.04000   Median :0.5900
 Mean   :0.02093   Mean   :1.101   Mean   :0.00907   Mean   :0.05326   Mean   :0.04426   Mean   :0.6129
 3rd Qu.:0.03000   3rd Qu.:1.320   3rd Qu.:0.01000   3rd Qu.:0.06000   3rd Qu.:0.05000   3rd Qu.:0.7200
 Max.   :0.05000   Max.   :1.870   Max.   :0.03000   Max.   :0.13000   Max.   :0.12000   Max.   :1.1700
      C18              C18-OH            C18:1             C18:1-OH
 Min.   :0.2300   Min.   :0.00000   Min.   :0.510   Min.   :0.00000
 1st Qu.:0.4800   1st Qu.:0.01000   1st Qu.:0.910   1st Qu.:0.01000
 Median :0.5900   Median :0.01000   Median :1.100   Median :0.01000
 Mean   :0.6129   Mean   :0.00857   Mean   :1.138   Mean   :0.01512
 3rd Qu.:0.7200   3rd Qu.:0.01000   3rd Qu.:1.330   3rd Qu.:0.02000
 Max.   :1.1700   Max.   :0.02000   Max.   :1.980   Max.   :0.05000
                  NA's   :52
```

3. Ricerca dei Biomarcatori

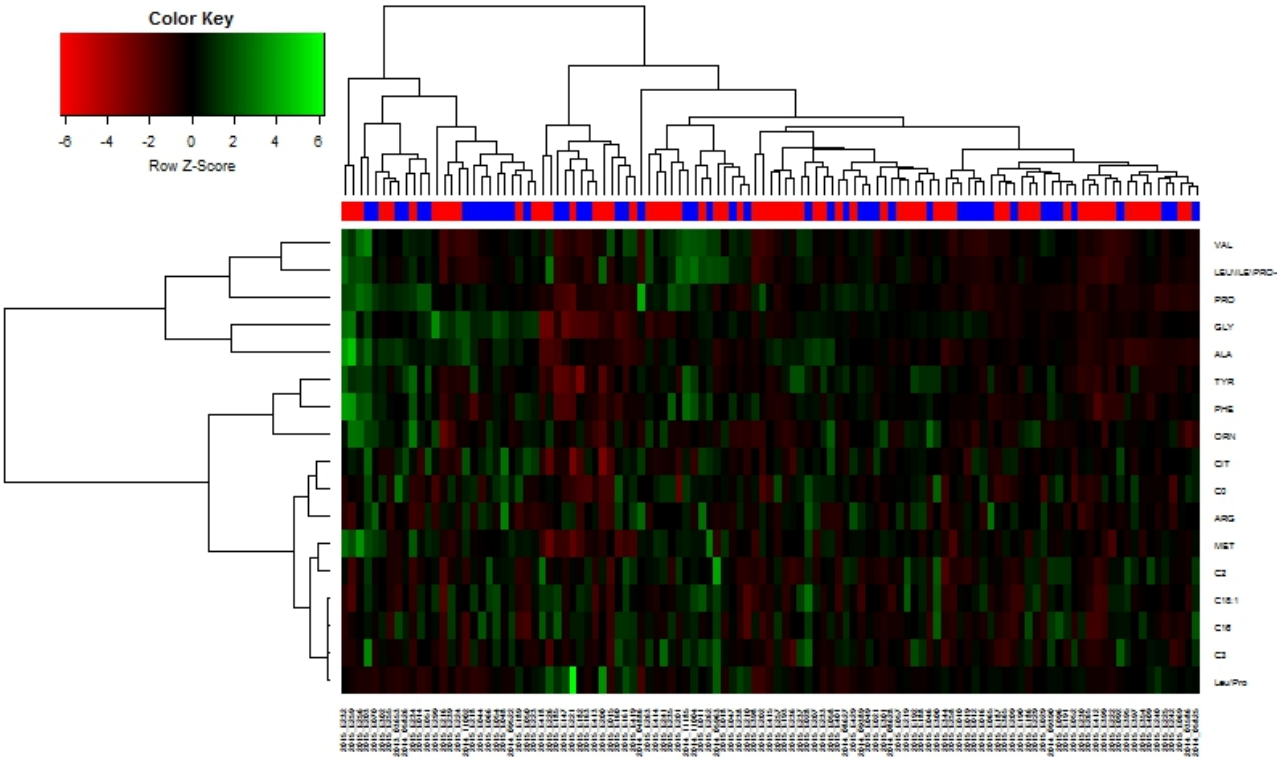
In questo capitolo vengono mostrati i risultati ottenuti dai packege di R ***limma*** e ***pamr*** per effettuare l'estrazione di biomarcatori.
Sono state necessarie delle manipolazioni alla struttura del dataset originario per poter applicare le funzioni presenti nelle librerie di R.

Analisi base con limma

Dalla analisi effettuata tramite il package *limma*, specificando come **p-value** massimo di 0.01 e come **logFoldChange** minimo di 1, si sono ricavati i risultati rappresentati dalle **heatmap** seguenti.



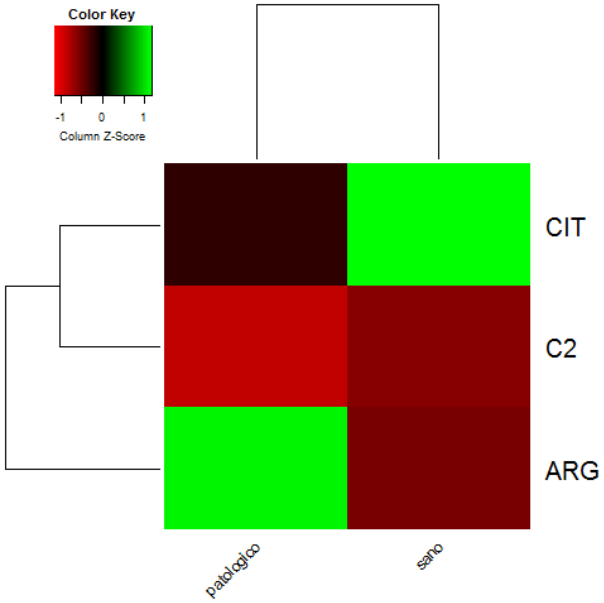
Le righe indicano i biomarcatori estratti ai quali viene assegnato uno Z-score nel caso di pazienti sani e patologici.



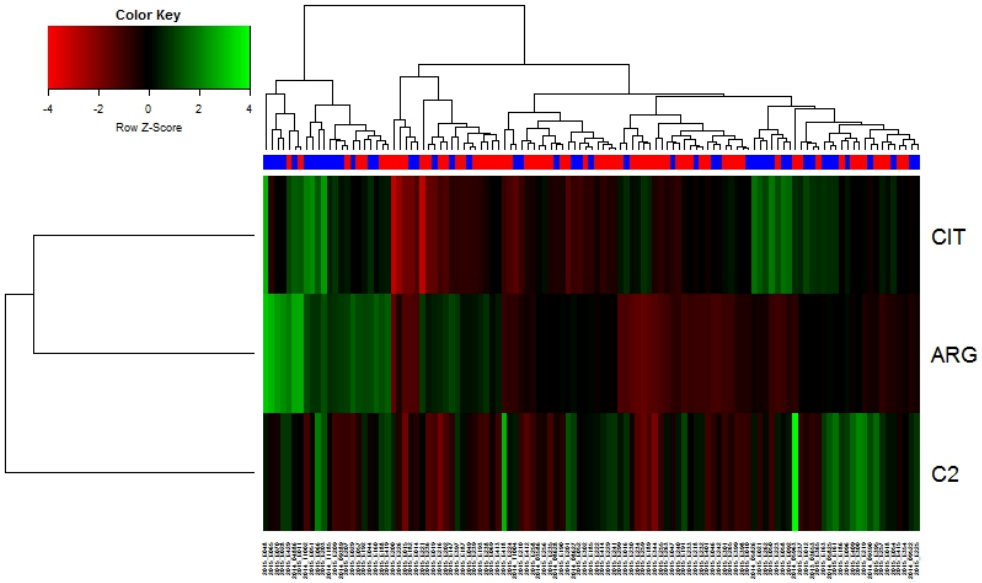
Dalle heatmap è possibile notare come questa analisi abbia prodotto risultati non soddisfacenti: molte delle concentrazioni estratte come biomarcatori hanno simile Z-score sia nel caso di pazienti sani che di pazienti patologici.

Analisi avanzata con limma

Viene aggiunto un **test di Caso/Controllo** per affinare l'analisi precedente, specificando in questo caso un **p-value** massimo di 0.1 e un **logFoldChange** minimo di 1.5.



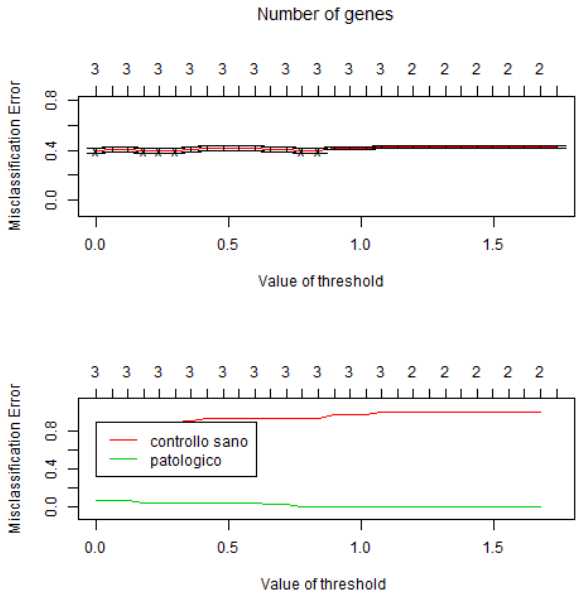
L'heatmap mostra come risultati di questa analisi siamo migliori rispetto al caso “base”; la componente ARG potrebbe essere un buon biomarcatore, sovraespressa per pazienti patologici, sottoespressa per quelli sani. La componente C2 da risultati non troppo soddisfacenti, avendo Z-score simili per entrambi i tipi di paziente.



Analisi con pamr

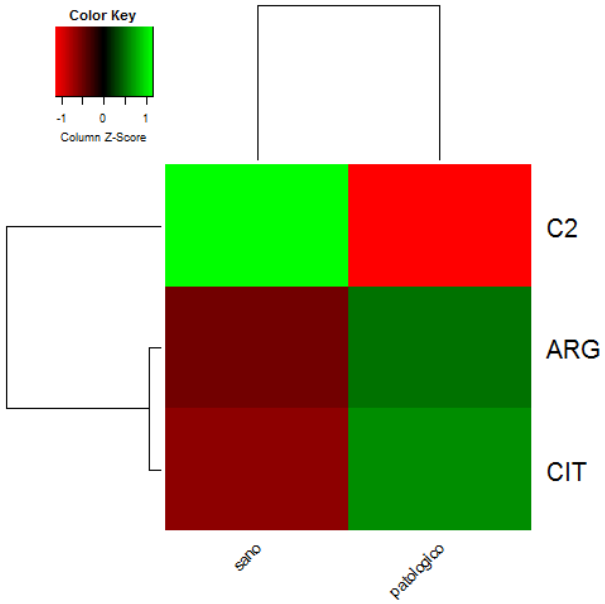
Tramite il package *pamr* è stato possibile migliorare ancora la ricerca dei biomarcatori analizzando anche le combinazioni dei possibili sottogruppi di concentrazioni attraverso una serie di **test statistici**.

Pamr riceve in input l'elenco dei biomarcatori individuato da *limma* (analisi avanzata); viene prima effettuato un test di cross-validazione per capire come funziona il modello addestrato. Il grafico seguente ci mostra la percentuale globale di errori e la percentuale di errore per ogni classe, che verranno poi usate per selezionare la soglia migliore.

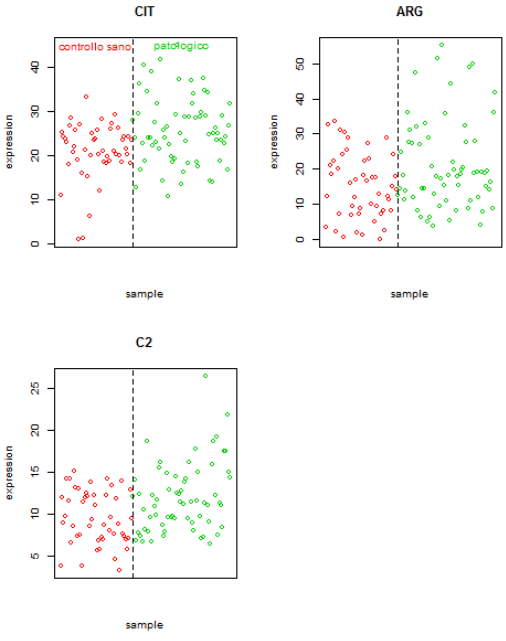


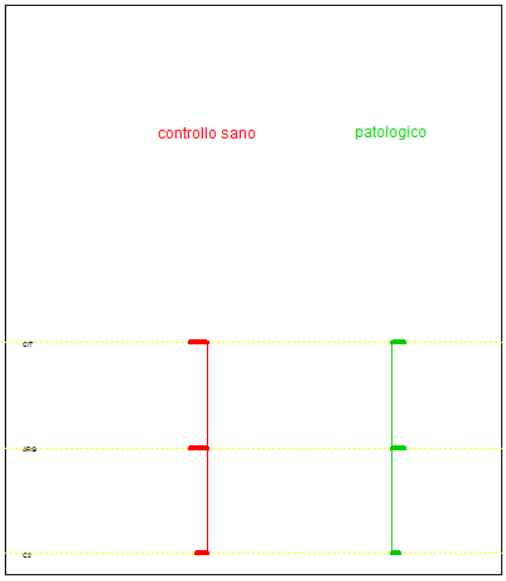
Si può notare un errore molto alto per quanto riguarda i controlli sani, una probabilità media di errore di 0.4 e che la soglia scelta è 0 (nella maggior parte delle esecuzioni).

I biomarcatori estratti sono visualizzabili nella seguente heatmap.



Il package *pamr* mette a disposizione anche una serie di funzione per generare immagini che ci aiutano a vedere graficamente i risultati.

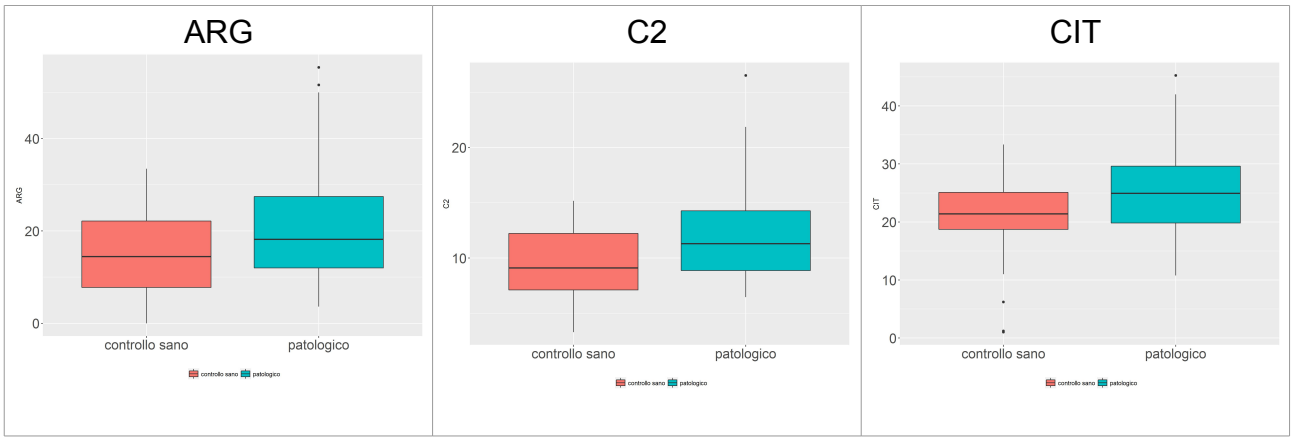




Il grafico a sinistra ci mostra per ogni biomarcatore individuato come si organizzano i valori di espressione dei campioni.
Il grafico a destra ci indica in base alla classe come si comporta mediamente ogni possibile biomarcatore selezionato da *pamr*.

Valutazione

Il semplice ma efficace metodo utilizzato per la valutazione dei biomarcatori ottenuti è l'utilizzo dei **boxplot**. Tramite la libreria *ggplot* si sono genarati le seguenti immagini.



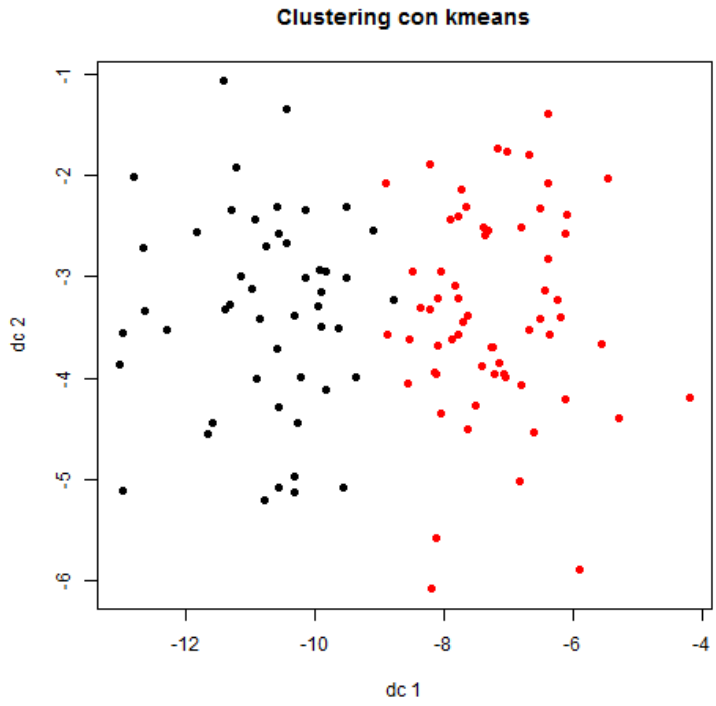
Nonostante si siano scelti dei p-value bassi, i boxplot ci mostrano come queste concentrazioni non sono ottimi biomarcatori, poiché i box si sovrappongono e non è quindi possibile distinguere tra i due casi .

4. Clustering

Tramite il package *cluster* di R sono stati applicati i metodi di **clustering gerarchico** e **k-means**.
E' stata creata un unica tabella chiamata "*db*" (113 osservazioni di 48 attributi) selezionando i sample comuni tra le tabelle dati e concentrazioni (includendo di dati soltanto le informazioni riguardanti il tipo di paziente e l'età).

K-means

Sono stati plotatti i cluster ottenuti rispetto alla prima e alla seconda funzione discriminante mediante il metodo *plotcluster* del package *fpc.*



Inoltre si è valutato quanto i cluster ottenuti distinguano i tipi sano e patologico, i risultati sono scarsi:

```
Clustering con kmeans:

   1  2
1 15 32
2 34 31
```

Le colonne della matrice rappresentano i due cluster mentre le righe rispettivamente sani e patologici.

Clustering gerarchico

Sono state applicate alcuni metodi di clustering gerarchico (con diverse misure di distanza) al fine di identificare **2 cluste**r differenti. Grazie al metodo *cluster.stats* del package *fpc* sono stati poi ricavati il diametro medio dei cluster e la separazione per ogni metodo applicato, al fine di valutarne il migliore.
Si è cercato inoltre anche qui di verificare se i cluster ottenuti identificassero i tipi sano e patologico ma ancora una volta i risultati sono scarsi:

```
Clustering gerarchico distanza di manhattan metodo single
   groups
     1  2
  1 46  2
  2 65  0
Media diametri cluster:  783.3842
Separazione:  456.1167

Clustering gerarchico distanza minkowski metodo complete:
   groups
     1  2
  1 42  6
  2 62  3
Media diametri cluster:  420.7447
Separazione:  74.75716

Clustering gerarchico distanza euclidea metodo ward.D:
   groups
     1  2
  1 26 22
  2 42 23
Media diametri cluster:  398.5227
Separazione:  42.2895
```

Le colonne della matrice rappresentano i due cluster mentre le righe rispettivamente sani e patologici.
Infine sono stati generati i **dendogrammi** dei cluster ottenuti.