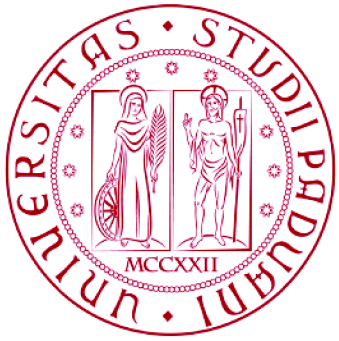
Università degli studi di Padova

Corso di Laurea in Scienze Statistiche Demografiche e Sociali



ANALISI DI DATI DI ESPRESSIONE GENICA:

UN CASO STUDIO SULLE LEUCEMIE LINFATICHE ACUTE DI TIPO B

Relatrice : Prof.ssa Chiara Romualdi

Dipartimento di Biologia

Laureanda: Nadia Panebianco

Anno accademico 2011/ 2012

Indice

CAPITOLO 1

INTRODUZIONE

1.1 Ambiente di lavoro

L'ambiente di lavoro che si andrà ad illustrare ed analizzare si riferisce all'ambiente molecolare. Le molecole organiche semplici si possono associare formando dei polimeri (ovvero delle macromolecole che costituiscono delle catene). Nelle cellule viventi lunghi polimeri formano proteine, acidi ribonucleici (RNA) o desossiribonucleici (DNA). Una sequenza di acidi nucleici (DNA o RNA), composta da regioni trascritte e regioni regolatorie, corrisponde al gene che è l'unità ereditaria fondamentale degli organismi viventi. Di conseguenza i geni sono fondamentali per lo sviluppo fisico e comportamentale dell'essere vivente.

Le sequenze geniche possono essere di tipo codificanti o non codificanti. La maggior parte dei geni codifica attraverso le proteine, mentre per quanto riguarda quelle non codificanti, producono RNA non codificante, il quale ha un ruolo fondamentale nella biosintesi delle proteine e nell'espressione genica. Ci si concentrerà in particolare sull'espressione genica, che è il processo attraverso cui l'informazione contenuta in un gene viene convertita in una macromolecola funzionale (tipicamente una proteina). Nello studio che si andrà ad affrontare quindi la matrice di dati grezzi verrà poi convertita in una matrice di espressione genica.

1.2 L'utilizzo di microarray (matrici ad alta densità)

Per quanto riguarda i dati grezzi corrispondono ad un campione di immagini, derivanti dalla quantità di DNA (probe) fissata sul vetrino e poi riprodotta tramite uno scanner a laser, formanti un array (come rappresentati in Figura 1). In un array quindi sono presenti migliaia di geni. Questo procedimento è prodotto utilizzando il metodo Affymetrix (Affymetrix Inc., 2001a,b).

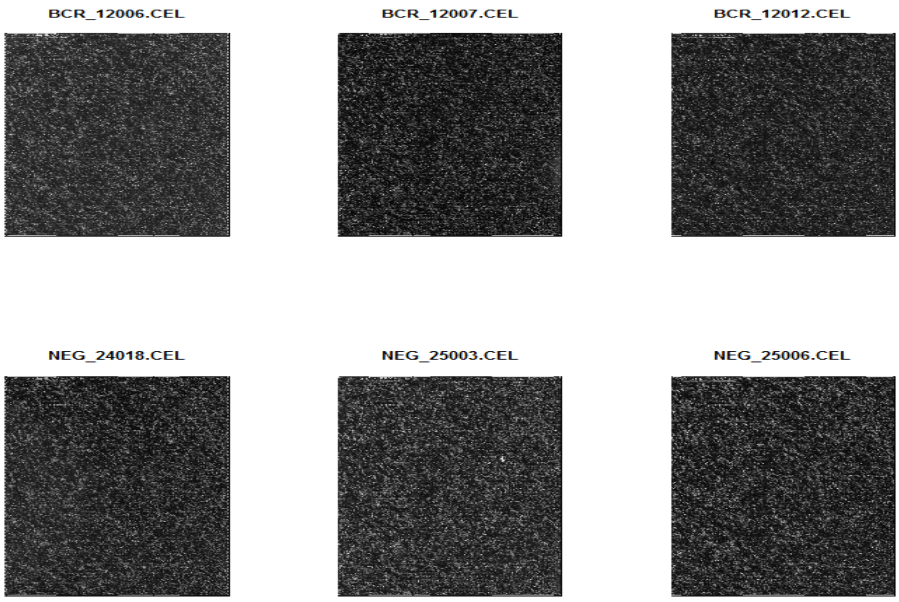


Figura 1: Campioni BCR/ABL positivi e campioni negativi (NEG) riguardo la leucemia linfocitica acuta di tipo B.

Per illustrare nel dettaglio tale procedimento bisogna dire innanzitutto che gli mRNA giocano un ruolo cruciale nella differenzazione e nello sviluppo della cellula, ma anche nelle patologie e in una serie di malattie e tumori. Per poterli studiare quindi essi vengono prima estratti dalle cellule, convertiti in cDNA (trascrittasi inversa) e marcati attraverso una sonda fluorescente. Si procede poi con la tecnica dell'ibridazione inversa, che

consiste appunto nel fissare i probe su un supporto e rilevare il target (sequenza di DNA o RNA che si vuole analizzare) al fine di ricavare un profilo di espressione. Il cDNA target rimarrà legato alla sonda e in tal modo si rileverà la sua posizione. Dopo l'ibridazione si rimuove il cDNA che non si è legato. Quindi la tonalità fluorescente varia in base alla prevalenza alla prevalenza di mRNA: se lo spot ha una gradazione sul verde vuol dire che la quantità di RNA espressa da un gene nelle cellule di interesse è bassa (down regolata), mentre se lo spot è sul rosso vuol dire che l'espressione genica è aumentata (up regolata). Successivamente la fluorescenza viene rilevata attraverso uno scanner a laser, che acquisisce un'immagine per ogni fluoroforo. Attraverso poi appositi software si convertono i segnali in una gamma di colori in base alla loro intensità. Perciò il segnale rilevato verrà poi convertito in valore numerico.

1.3 Normalizzazione

Un passo rilevante dell'analisi di microarray è la normalizzazione dei dati, importante per differenziare le variazioni reali nei livelli di espressione genica e le variazioni dovute al processo di misura. Come differenze di intensità dovute ai parametri di scansione, alla diversa potenza di laser.

Quindi per rendere comparabili i valori di espressione tra tutti gli esperimenti si procede con la normalizzazione della matrice di espressione.

Ci sono diverse tecniche che permettono di normalizzare i dati e ognuna si basa su metodologie diverse; in questo caso si è deciso di utilizzare:

1) la normalizzazione quantile, basata sulla trasformazione delle distribuzioni di intensità di ogni specifico array, assegnando ad ogni intensità lo stesso valore in base al quantile di appartenenza (Parmigiani et al., 2003) ;
2) la normalizzazione basata sulla trasformazione logaritmica generalizzata, in cui si rende costante la varianza fra i dati.

Si sono confrontate le due normalizzazioni (utilizzando MAplot, analisi dei clusters, χ^2 test) al fine di scegliere quella che spiega meglio i dati, per poterla poi adoperare nelle analisi successive.

1.4 Geni differenzialmente espressi e scopo della tesi

La differenza tra una sequenza genica e un'analisi di espressione è che la prima descrive ciò che la cellula è in grado di fare, mentre la seconda descrive ciò che realmente sta facendo. Di conseguenza i profili di espressione genica possono mostrare, per esempio, come le cellule reagiscono ad un particolare trattamento. Perciò lo scopo dell'analisi è quello di individuare, sotto determinate condizioni, i geni differenzialmente espressi (ovvero i geni che reagiscono in maniera diversa a particolari stimoli) ed inoltre studiare le alterazioni di espressione genica di geni appartenenti ad uno specifico pathway (Parmigiani et al., 2003).

A tal proposito vengono adoperati diversi test, prima per identificare i geni differenzialmente espressi tra cui quelli sovra e sotto espressi e successivamente per individuare i gruppi di geni, identificati con specifiche annotazioni e appartenenti a determminati pathway, che risultano significativi per l'analisi statistica.

In particolare, dopo aver normalizzato i dati, si identificano i geni differenzialmente espressi utilizzando i test *Empirical Bayes*, *Significance Analysis of Microarray* (test SAM), *Rank Prodouct*. I risultati sono stati confrontati tra loro e si sono rilevati quei geni che risultano comuni come differenzialmente espressi per tutti e tre i test.

Successivamente sono state rilevate le gene set tra i geni differenzialmente espressi. Una lista di geni infatti è facilmente interpretabile se i geni mostrano delle similarità nella loro annotazione funzionale (Goeman et al.,

2007). I termini delle annotazioni dei raggruppamenti sono ottenuti dalle librerie della Gene Ontology (che considerano la "Biological Process", la "Molecular Function" e la "Cellular Component") (Ashburner et al., 2000) o dei pathway Kegg (Ogata et al, 1999).

Ciò che è interessante osservare è rilevare i raggruppamenti che risultano significativamente diversi in media e quindi sarà compito del biologo studiare il fenotipo di appartenenza e le relazioni che esistono in esso.

Anche in questo sono stati eseguiti diversi test e si sono confrontati tra loro i risultati rispetto ai patterns ottenuti. I test per le gene set analysis si dividono in due grandi categorie che fanno capo a due diverse ipotesi nulle: ipotesi nulla competitiva (test dell'Ipergeometrica) e ipotesi nulla indipendente (il Global Test) (vedere capitolo 5 per ulteriori dettagli). Infine si esegue un'analisi sull'effetto di un determinato pathway (SPIA), è un tipo di analisi mista che combina il risultato ottenuto dall'analisi di arricchimento considerando però la perturbazione di un determinato pathway sotto una specifica condizione.

Per l'analisi statistica si è utilizzato il pacchetto statistico "R" (Hahne et al., 2008)

CAPITOLO 2

ANALISI DESCRITTIVA

2.1 Presentazione dei dati

I dati usati provengono da uno studio, condotto dall'Università "La Sapienza" di Roma, riguardante i profili dell'espressione genica sulla leucemia linfatica acuta di tipo B (Chiaretti et al., 2005). Vengono utilizzati microarray di oligonucleotidi ad alta densità (Affymetrix U95Av2) per definire i profili di espressione genica della Leucemia Linfatica Acuta (ALL) su un campione composto da128 pazienti (di cui 43 donne e 85 maschi) a cui è stata diagnosticata tale malattia (aventi più del 90% di cellule leucemiche) e che devono ancora sottoporsi ad un trattamento convenzionale (la chemioterapia). Tutti i pazienti, inoltre, provengono dalla clinica italiana GINEMA di Roma, negli anni tra il 1996 e il 2000, con un'età mediana di 29anni (appartenenti ad un range tra i 15 e i 58 anni).

La leucemia è un tumore delle cellule del sangue: le cellule normali che si ritrovano nel sangue (globuli rossi, globuli bianchi e piastrine) prendono origine da cellule immature (dette anche cellule staminali o blasti) che si trovano nel midollo osseo, cioè in quella parte di tessuto spugnoso contenuto all'interno delle ossa.

Nelle persone affette da leucemia vi è una proliferazione incontrollata di queste cellule, che interferisce quindi con la crescita e lo sviluppo delle normali cellule del sangue.

La tipologia di leucemia prende poi il nome in base alle cellule da cui ha origine il tumore. La cellula staminale, durante le varie fasi di maturazione, dà origine a cellule di tipo mieloide e cellule di tipo linfoide (cioè linfocitarie nulle, di tipo B, pre-B e T). Questa seconda tipologia di cellule sono quelle considerate per lo studio. Da queste si differenzieranno successivamente i globuli rossi o eritrociti, le piastrine e i globuli bianchi (leucociti e linfociti). Infine una malattia è considerata acuta quando questa si sviluppa molto velocemente.

Sebbene l'origine cellulare delle leucemie possono essere facilmente determinate da studi fenotipici (ovvero studi che rilevano dei fenotipi di appartenenza, che includono le informazioni relative a tutte le possibili caratteristiche delle cellule), il meccanismo di trasformazione delle diverse leucemie restano sconosciute. In questo contesto il profilo dell'espressione genica definisce un nuovo approccio per esplorare i meccanismi e le trasformazioni delle cellule maligne.

Si è osservato che il 25-30% dei pazienti presentavano un'ALL di tipo T, mentre il restante 70-75% erano di tipo B. Prendendo in considerazione quest'ultimo tipo di leucemia la numerosità campionaria si riduce a 95 pazienti.

E' noto che anche la più semplice disfunzione è dovuta ad un'alterazione genica, ovvero a dei riarrangiamenti molecolari che corrispondono a delle traslocazioni cromosomiche dovute ad un errato scambio di cromosomi che portano quindi a delle mutazioni geniche. Di conseguenza in base al tipo di traslocazione di cui un individuo è affetto aumenta la probabilità di incorrere in una determinata patologia.

Per quanto riguarda la leucemia linfatica acuta di tipo B si sono osservati tre riarrangiamenti molecolari (BCR/ABL, ALL1/AF4, E2A/PBX1) che hanno avuto una traslocazione su 10 molecole diverse. In più si sono presi in considerazione pazienti senza riarrangiamenti, cioè quei campioni che non mostrano un riarrangiamento molecolare evidente (che saranno chiamati NEG).

Nell'analisi seguente si terrà conto in particolare dei riarrangiamenti BCR/ABL e NEG. Infatti le rispettive traslocazioni molecolari incidono sulla presenza dell'ALL per il 39% dei casi per il primo e il 43% per quelli NEG, inoltre si è già osservato che i profili di espressione genica di questi due gruppi sono tra loro più simili rispetto agli altri gruppi molecolare, infatti non sono raggruppabili in clusters distinti e ciò porta a sostenere che includono una serie di sottogruppi comuni.

Dall'analisi citogenetica si è osservato che i campioni positivi BCR/ABL hanno avuto una traslocazione per ben 37 pazienti su 95. Mentre si riscontano 41 campioni di tipo NEG, quindi per l'analisi verranno utilizzati un totale di 78 campioni (41 NEG e 37 positivi che saranno chiamati BCR).

2.2 Campioni BCR/ ABL positivi e campioni negativi

Osservando i dati grezzi Affymetrix si contano 12625 geni. Ad ogni spot viene associato un valore di intensità assoluto (Figura 2.

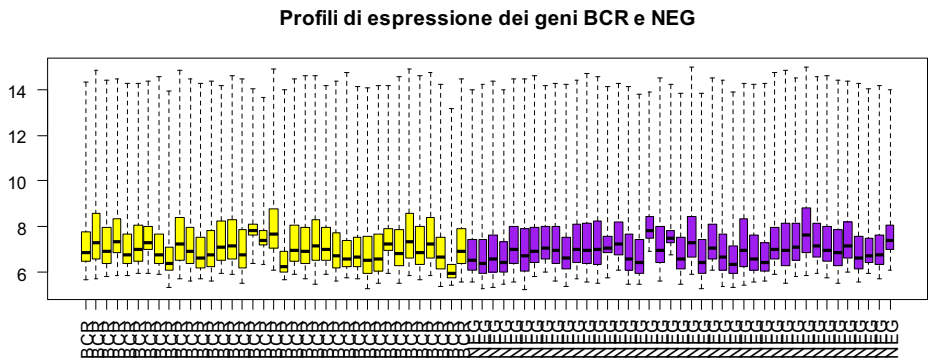


Figura 2: Box-plot degli array BCR/ABL positivi (giallo) e negativi (viola).

La scatola dei box-plot si distribuisce tra il 25esimo percentile e il 75esimo percentile. Si può osservare che le unità campionarie dei vari gruppi si distribuiscono maggiormente in un range tra 6 e 9 del logaritmo dell'espressione genica, anche se i valori più alti si distribuiscono fino al valore 15 del logaritmo dell'espressione genica. Sia per i campioni BCR e sia per quelli NEG la mediana sembra essere posizionata intorno al 50esimo percentile globale anche se per alcuni campioni si rivela essere un po' più bassa. I dati quindi si distribuiscono in maniera abbastanza omogenea all'interno delle singole scatole (anche se tendenti verso i valori più bassi, con valori anomali che estendono maggiormente verso valori alti); ma si osservano differenze di distribuzioni fra i vari box-plot, indipendentemente dall'appartenere all'espressione genica dei campioni positivi BCR/ABL o negativi, quindi con o senza riarrangiamento molecolare.

2.3 Matrice di espressione

Dopo una prima analisi dei dati grezzi viene estratta la matrice di espressione genica per tutti i 78 campioni, che rappresenta la matrice di interesse per lo studio.

Per capire meglio come si presenta una matrice di espressione se ne riportata di seguito una sua parte.

	BCR_62002.CEL	BCR_62003.CEL	BCR_65005.CEL	BCR_68003.CEL
1	5.075174	5.147494	4.787492	4.369448
2	9.069295	9.302555	9.415955	8.846713
3	4.955827	5.192957	4.700480	4.605170
4	9.049232	9.281637	9.416378	8.826441
5	4.463607	4.595120	4.406719	4.091006

	BCR_84004.CEL	NEG_01010.CEL	NEG_04007.CEL	NEG_04008.CEL
1	5.105945	5.337538	4.828314	4.753590
2	9.676104	8.858155	9.126306	9.134247
3	5.484797	5.579730	5.517453	4.882802
4	9.671808	8.855949	9.144521	9.168476
5	4.929425	4.877485	4.366913	4.115780

Si osservano per riga i primi quattro geni dei 12625 geni totali, con il corrispondente valore di espressione per ogni campione.

Per evidenziare l'esistenza di differenze sistematiche di espressione lungo i livelli medi di intensità si usa il grafico MAplot (Figura 3). Consiste nel confrontare i campioni di dati con un vettore di riferimento costituito dalla mediana per riga della matrice di espressione.

L'asse delle ascisse è identificata con la lettera A e rappresenta la media della log-intensità: `1/2log(x*y)`; mentre l'asse delle ordinate è identificata con la lettera M e riproduce la differenza delle log-intensità:

`log(x/y)`, dove x è uguale al logaritmo dell'intensità del campione e y è uguale al logaritmo dell'intensità di riferimento.

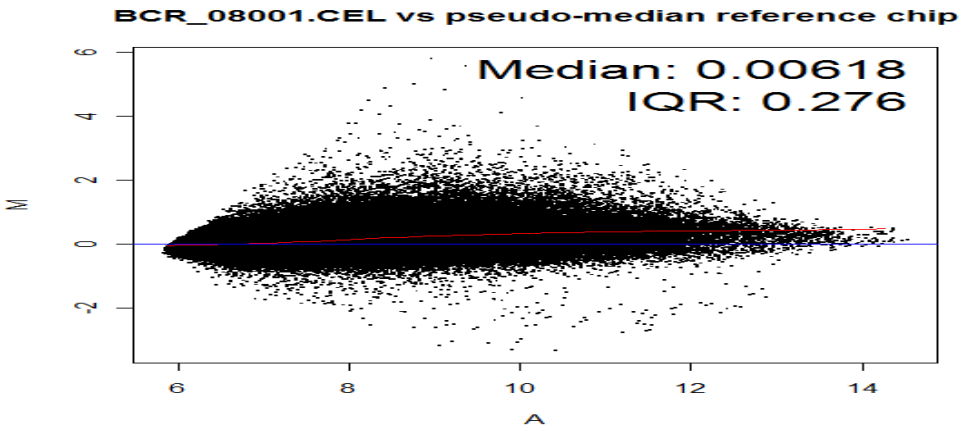


Figura 3: Grafico MAplot.

Nonostante sembra esserci una propensione dei dati che tende verso valori più alti rispetto ai valori dell'esperimento di riferimento, si osserva che la mediana (linea blu) è pari a zero così come l' IQR, vale a dire il range inter-quartile (definito come la differenza tra il 75esimo e il 25esimo percentile), di conseguenza possiamo considerare che i dati hanno una buona distribuzione, cioè abbastanza omogenei tra loro.

CAPITOLO 3

NORMALIZZAZIONE DEI DATI DI ESPRESSIONE

3.1 Introduzione

Le assunzioni che riguardano la maggior parte delle procedure di normalizzazione includono:

- Pochi geni differenzialmente espressi. Ci si aspetta infatti meno del 10% di geni differenzialmente espressi.
- Simmetria tra sovra e sotto espressi: I geni differenzialmente espressi siano egualmente ripartiti tra geni sovraespressi e geni sottoespressi;
- La differenziale espressione non dipenda dalla media del segnale.

Di seguito si analizzeranno due metodi di normalizzazioni: la normalizzazione quantile e la normalizzazione con la trasformazione logaritmica generalizzata, al fine di stabilire quale delle due normalizzazioni si adatta meglio ai dati per il proseguimento dell'analisi.

3.2 Normalizzazione quantile

Attraverso la normalizzazione quantile si pone come obiettivo rendere uguali le distribuzioni empiriche di tutti gli arrays, prendendone uno come riferimento. Attraverso la Figura 4 sarà possibile osservare immediatamente l'adattamento dei dati con tale metodo.

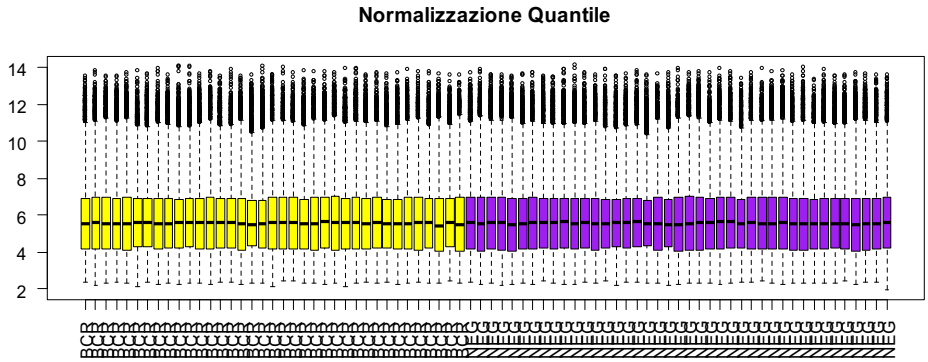


Figura 4: Box-plot dei dati di espressione normalizzati utilizzando il metodo 'quantile'.

Innanzitutto dai valori sull'ordinata si nota immediatamente come la scala di valori, normalizzando, si sia drasticamente abbassata rispetto ai valori molto alti che si potevano notare dalla precedente matrice di espressione non normalizzata.

Dai Box-plot si osserva un buon adattamento della normalizzazione quantile sui dati, poiché i box- plot sono omogeneamente distribuiti per entrambi i gruppi BCR e NEG. Infatti il primo e il terzo interquartile corrispondono al valore 4 e 7 del logaritmo del gene di espressione con mediana pari a 5.5 circa. Però i baffi del box-plot si estendono fino a 2 per i valori più bassi e fino a 11 per quelli maggiori con valori anomali che tendono verso geni di espressioni sempre più elevati.

3.3 Normalizzazione con la trasformazione logaritmica generalizzata

Si assume una struttura di generazione del segnale del tipo:

$$Y_g = \alpha_g + \mu_g e^{\eta_g} + \varepsilon_g \qquad \eta_g \sim \mathrm{N}(0, \sigma_\eta^2)$$

$$\varepsilon_g \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

In cui: Y_g è il livello di espressione del gene g per singolo canale;

α_g riporta il rumore medio di background;

μ_g è il vero valore di espressione del gene.

Per $\mu_g \rightarrow 0$ (livelli di espressione bassi), la misura di espressione la si può approssimare come:

$$Y_g \approx \alpha_g + \varepsilon_g \qquad Y_g \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

Mentre per $\mu_g \rightarrow +\infty$ (livelli di espressione alti), si approssima come :

$$Y_g \approx \mu_g e^{\eta_g} \qquad Y_g \sim \log \mathrm{N}(\log \mu_g, \sigma_\eta^2)$$

Quando il livello di espressione è intermedio tra i due casi estremi, la misura di espressione si distribuisce come una combinazione lineare di una Normale e di una Log-Normale. Di conseguenza tramite un'opportuna trasformazione dei dati si trova una funzione f(Y) sufficientemente liscia, tale che la varianza asintotica della trasformata sia costante.

Dopo opportuni passaggi si giunge quindi alla funzione:

$$f(Y_g) = \log\left(Y_g - \alpha_g + \sqrt{(Y_g - \alpha_g)^2 + c}\right)$$

Trasformazione logaritmica generalizzata (GLOG)

È una funzione monotona crescente per tutti i valori di Y_g positivi o negativi. Inoltre se μ_g tende a zero la funzione è approssimativamente lineare, mentre per valori elevati di μ_g assume una distribuzione logaritmica. Infine la varianza asintotica dei dati trasformati è costante e uguale a S^2 (con $S^2 = e^{\sigma_\eta^2}\left(e^{\sigma_\eta^2} - 1\right)$).

È interessante quindi osservare anche su grafico (Figura 5) come si comporta la deviazione standard sui dati con quest'altra normalizzazione confrontandola con la deviazione standard dei dati non normalizzati.

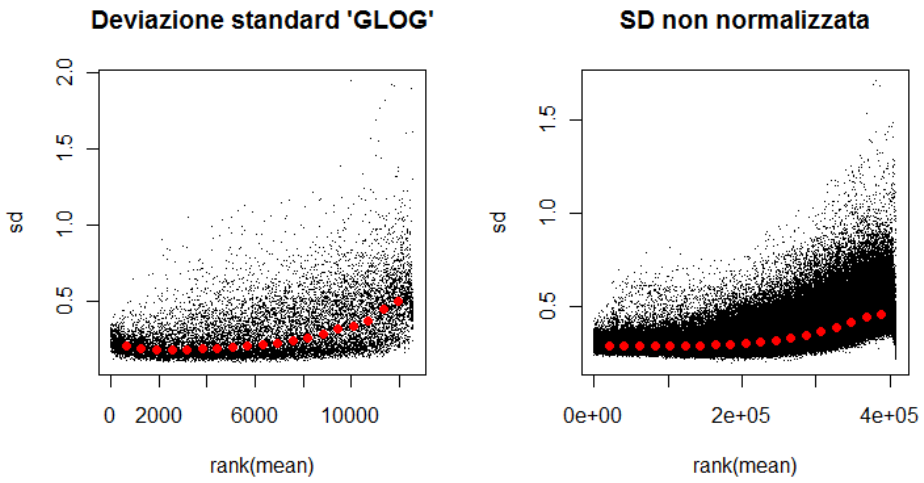


Figura 5: Deviazione Standard della matrice di espressione normalizzata col metodo della trasformazione logaritmica generalizzata e Deviazione Standard della matrice di espressione non normalizzata.

Si osserva che sia per i dati normalizzati GLOG con varianza costante che per i dati non normalizzati, per valori alti il trend tende ad assumere un andamento monotono crescente, rimanendo però costante intorno allo zero.

Così come per la normalizzazione quantile, si osserva anche in questo caso l'adattamento dei dati normalizzati col metodo GLOG tramite il grafico dei box-plot osservabili nella Figura 6.

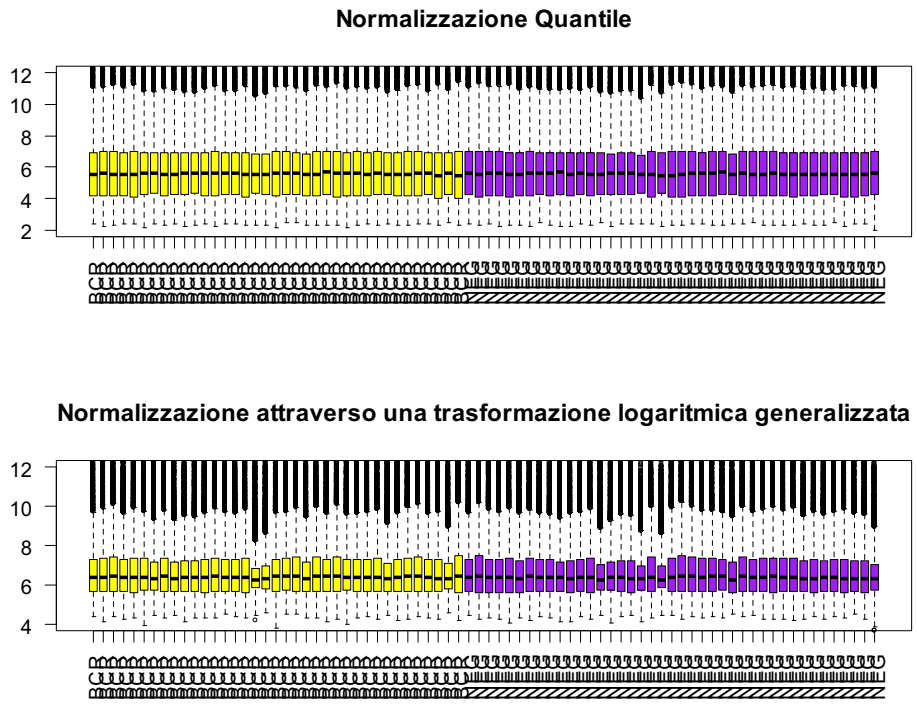


Figura 6: Confronto dei box- plot della matrice di espressione normalizzata col metodo quantile e normalizzata usando il metodo della trasformazione logaritmica generalizzata (glog).

Viene riproposto anche il grafico della normalizzazione quantile per poter fare un confronto più immediato fra le due normalizzazioni. Notiamo che con la normalizzazione quantile i box-plot sono meglio allineati, però in entrambi i casi la mediana nei vari campioni resta costante ed inoltre, nella normalizzazione con la trasformazione logaritmica generalizzata lo scarto interquartile fra il 75esimo e il 25esimo percentile è molto più basso rispetto ai box-plot della normalizzazione quantile, ciò vuol dire che i dati sono concentrati in un range più piccolo e di conseguenza più omogenei fra loro. Per questo motivo si tende a preferire una trasformazione dei dati effettuata col metodo della trasformazione logaritmica generalizzata.

3.4 Qual è la normalizzazione che si adatta meglio ai dati?

Al fine di decidere quale delle due normalizzazioni adottare per il proseguimento dell'analisi dei dati, si inizia con un'analisi descrittiva osservando la correlazione tra queste.

In particolare si confrontano tutti i 78 campioni della normalizzazione quantile con i corrispettivi della normalizzazione GLOG, quindi si otterranno 78 valori di correlazioni (vedere Appendice) tra le due normalizzazioni. Il grafico riassuntivo che ne deriva è il box-plot di Figura 7.

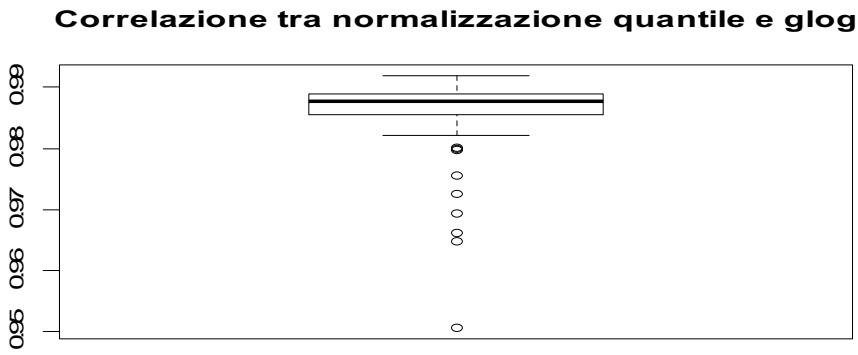


Figura 7: Correlazione tra i valori di espressione ottenuti con i due metodi di normalizzazione.

La correlazione tra i valori di espressione ottenuti attraverso la normalizzazione quantile e glog è da definire soddisfacente, in quanto l'intero box-plot, compresi i sette campioni anomali, si trova nell'intervallo [0.9; 1]. Per tanto tra i valori di espressione delle due normalizzazioni vi è una forte dipendenza lineare.

Dopo aver osservato il plot MA per la matrice di espressione non normalizzata, vediamo ora come si riproduce il rapporto delle log-intensità con le due normalizzazioni.

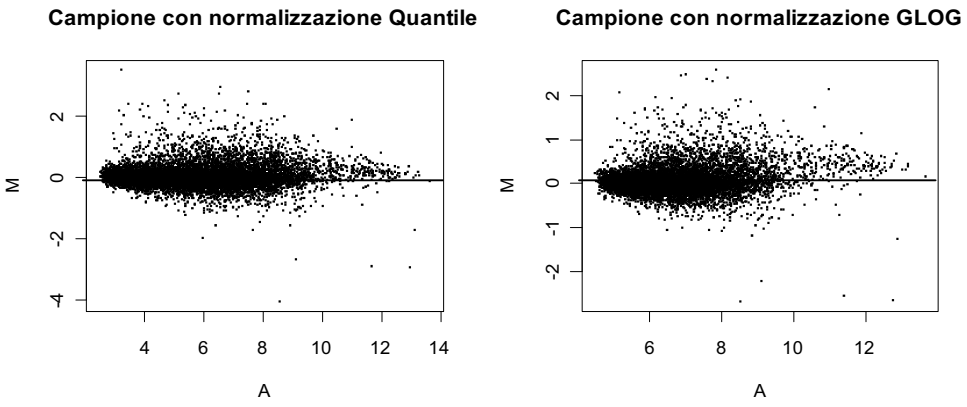


Figura 8: MAplot della matrice di espressione normalizzata col metodo quantile e col metodo glog.

Notiamo dalla Figura 8 che in entrambi i casi si tende ad avere una dispersione dei dati tendenti un po' di più verso valori positivi, quindi con differenze di log intensità rispetto al chip di riferimento, ma nonostante ciò entrambe le normalizzazioni si distribuiscono intorno allo zero. Anche attraverso il plot MA si osserva un range dei valori dei dati più ristretto per la normalizzazione GLOG rispetto alla quantile e quindi con una concentrazione di dati più appaiati, rispetto ai dati della normalizzazione quantile.

3.4.1 Analisi dei cluster

Si prosegue la comparazione confrontando gli alberi e i raggruppamenti ottenuti tra le diverse normalizzazioni.

L'analisi dei clusters si basa sul principio di organizzare gli oggetti, che in questo caso sono rappresentati dai campioni, in una gerarchia o in una

struttura raggruppata ad albero in base alle dissimilarità tra i profili. Queste dissimilarità vengono calcolate in base ad opportune matrici di distanza tra i campioni. Inoltre viene anche scelto il metodo che, attraverso un appropriato criterio, accorperà le unità.

Per lo studio viene applicata la distanza euclidea (opportuna per dati continui), in cui il quadrato della distanza euclidea degli oggetti standardizzati è proporzionale alla correlazione degli oggetti originali. La formula aritmetica che viene operativamente applicata per le distanze tra i profili di espressione è:

$d_{xy} = \|x - y\| = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$

In cui x e y rappresentano i p vettori delle misure sugli oggetti che devono essere clusterizzati.

Infine per quanto riguarda il metodo di accorpamento dei profili di espressione è stato scelto il metodo completo. Esso consiste nell'ottenere gruppi caratterizzati da una notevole somiglianza interna.

In cui

$$d(A, B) = \max(d(x, y) |\ (x \epsilon A), (\ y \epsilon B))$$

ISu un totale di 78 campioni si sono osservati tali raggruppamenti (Figura 9):

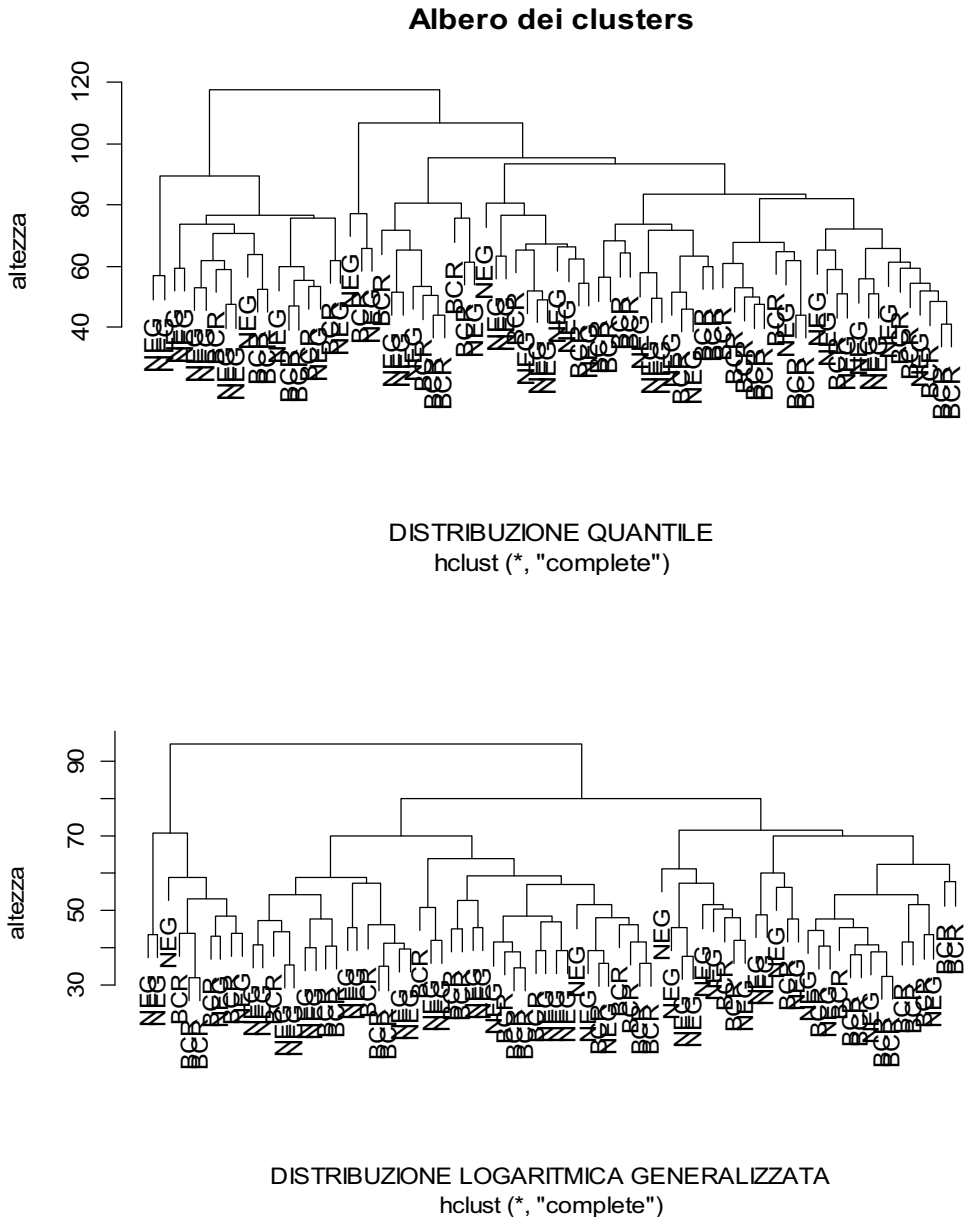


Figura 9: Analisi dei cluster rispetto alle distribuzioni delle due normalizzazioni.

Nella distribuzione quantile sembra che il cluster principale sia quello più alto, a sinistra contiene un certo raggruppamento, mentre a destra si sviluppano a scala il resto dei campioni. Per quanto riguarda, invece, la distribuzione logaritmica generalizzata si possono individuare due grandi gruppi (corrispondenti alle altezze maggiori), uno di "sostegno" e l'altro sottostante a questo che contiene, divisi sui due lati i restanti campioni.

Poiché i campioni sono tanti e graficamente, per entrambi i casi, la struttura ad albero tende ad essere più complessa da leggere, si decide di scomporla in gruppi più piccoli e di selezionare i due gruppi (visto che nei dati vi sono due gruppi) tagliando l'albero ad un'altezza predefinita.

Figura 10: Raggruppamento Cluster in 2 classi.

Per non perdere le distribuzioni rispetto ai due gruppi di geni BCR e NEG, questi vengono sintetizzati nella tabella sottostante.

Tabella 1: Distribuzione dei cluster per classe in entrambe le normalizzazioni.

Classe QUANTILE	BCR	NEG
1	30	29
2	7	12

Classe GLOG	BCR	NEG
1	32	36
2	5	5

Complessivamente per entrambe le normalizzazioni si hanno proporzioni molto simili, vale a dire nei cluster in cui si osserva un numero maggiore di geni BCR si contano anche in maniera numerosa i geni NEG. Inoltre anche la struttura ad albero è molto simile.

Per la distribuzione quantile il cluster 1 conta 59 campioni (30 BCR e 29 NEG), mentre il cluster 2 contiene 7 campioni del gruppo BCR e 12 per il gruppo NEG. Per quanto riguarda, invece, la distribuzione logaritmica generalizzata è sempre il cluster 1 ad essere il raggruppamento maggiore:

con 68 campioni complessivi (32 BCR e 36 NEG) di conseguenza rappresenta quasi l'intero campionamento, invece il cluster 2 ha solo 10 campioni, equamente distinti tra campioni BCR e NEG.

Come esito dal grafico dei clusters, non vengono rilevate differenze rilevanti tra le due normalizzazioni, entrambe possiedono lo stesso tipo di cluster (in cui i campioni hanno delle relazioni più affini) che ingloba quasi la maggior parte dei campioni.

Osserviamo inoltre cosa ci suggerisce il test del X^2 di Pearson, per verificare l'osservazione tra la divisione dei campioni reali e quelle ottenute dalla cluster analisi, calcolando utilizzando sempre i valori che sono stati adoperati per la costruzione dei clusters attraverso il metodo completo.

```
        Pearson's Chi-squared test

data: table(cluster.QUANTILE$cluster, colori)
X-squared = 0.6387, df = 1, p-value = 0.4242
```

```
        Pearson's Chi-squared test

data:  table(cluster.GLOG$cluster, colori)
X-squared = 0.0273, df = 1, p-value =  0.8688
```

Per entrambe le normalizzazioni i p-values risultano non significativi e anche molto alti, quindi non utili per trarre delle conclusioni per decidere quale delle due normalizzazioni sia migliore per i dati.
Riportiamo infine la funzione "silhouette" che raccoglie i valori delle altezze tra un campione e il suo vicino: ovvero

$$s(i) = \left(b(i) - a(i) \Big/ \max\left(a(i), b(i)\right) \right)$$

Di seguito (Figura 11 e 12) si riportano i grafici della funzione silhouette, che descrive i due gruppi di claster per ogni normalizzazione

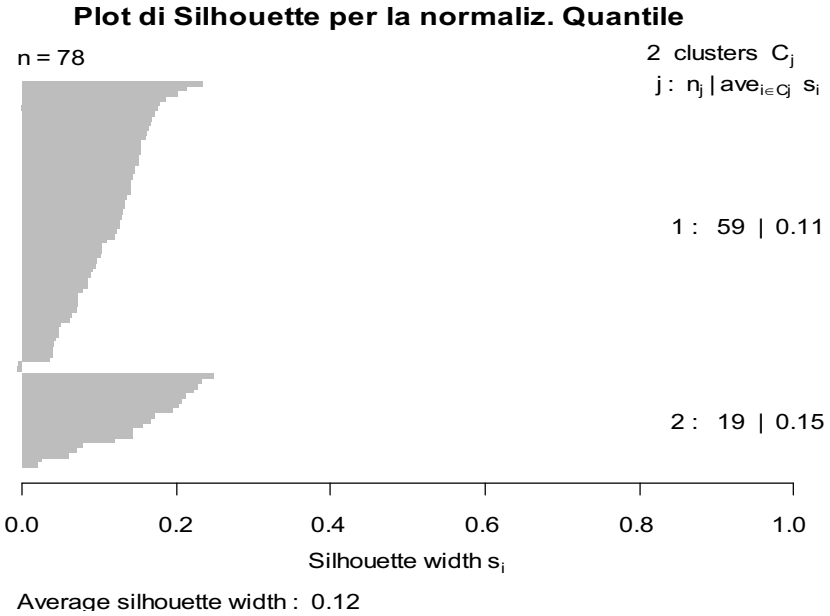


Figura 11: Grafico silhouette per la normalizzazione quantile.

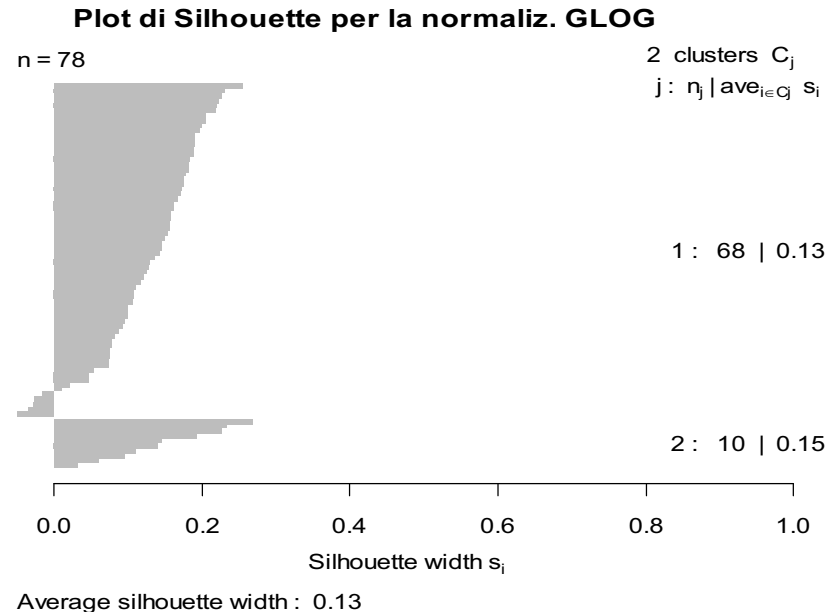


Figura 12: Grafico Silhouette per la normalizzazione GLOG.

Osserviamo valori bassi della funzione silhouette, più vicini allo zero che ad uno per entrambe le normalizzazioni, ciò porta a far pensare che la clusterizzazione utilizzata non è ottimale.

Analizzando i vari aspetti studiati si è notato che non ci sono differenze sostanziali tra la normalizzazione quantile e la normalizzazione basata sulla trasformazione logaritmica generalizzata: entrambe mantengono una mediana costante fra i diversi campioni, attraverso i plot MA si è osservato una leggera tendenza dei dati verso valori positivi (così come lo era per i dati grezzi), ma entrambe le normalizzazioni erano abbastanza simmetriche intorno allo zero ed infine anche dall'analisi dei clusters si è osservato un raggruppamento dei campioni BCR e NEG molto simile fra le due normalizzazioni.

Dovendo necessariamente sceglierne una per proseguire l'analisi, si ritiene opportuno considerare la normalizzazione GLOG, poiché per costruzione ha la capacità di mantenere la varianza costante, quindi avere dei dati più omogenei al suo interno.

CAPITOLO 4
INFERENZA STATISTCA PER L'IDENTIFICAZIONE DEI GENI DIFFERENZIALMENTE ESPRESSI

4.1 Introduzione

L'identificazione di geni differenzialmente espressi si raggiunge tramite la selezione dei geni che risultano avere un valore di espressione significativamente diverso tra due o più condizioni sperimentali.
Bisogna quindi eseguire un test per ogni gene, che abbia come ipotesi nulla l'uguaglianza dell'espressione del gene nelle due condizioni e come ipotesi alternativa l'espressione differenziale.

$$\begin{cases} H_{0,g}: \mu_{BCR}(g) = \mu_{NEG}(g) \\ H_{1,g}: \mu_{BCR}(g) \neq \mu_{NEG}(g) \end{cases}$$

In cui g=1…n e $\mu_{BCR}(g)$ rappresenta il valore atteso dell'espressione del gene g nella prima condizione (per il gruppo BCR) e $\mu_{NEG}(g)$ lo è invece nella seconda condizione (per il gruppo NEG).

A tal proposito verranno utilizzati dei Test come l'Empirical Bayes e il Test sulla Significance Analysis of Microarray (SAM) che appartengono alla classe dei Test t- Moderati. Questo tipo di test trasformano i risultati dei t-test per ogni gene, in modo da ottenere una statistica in cui la varianza al denominatore è modificata in modo da tenere conto delle deviazioni standard di tutti i geni dell'array.

Successivamente invece, per utilizzare e osservare un approccio diverso, si ricercano i geni differenzialmente espressi attraverso la statistica non parametrica del Rank Product (vedere paragrafo 4.4).

Si osserveranno quindi i risultati delle tre statistiche, estraendo infine i geni comuni differenzialmente espressi che sono risultati tali per le diverse statistiche.

4.2 Test Empirical Bayes

Il Test Empirical Bayes (EB) identifica i geni differenzialmente espressi utilizzando appunto un approccio Bayesiano empirico calcolando la probabilità a posteriori (di essere sregolato) per ogni gene.
Si assume per ogni gene un modello del tipo:

$$y_g = X\alpha_g + \varepsilon_g$$

In particolare Y_gè il vettore dei valori di espressione, X è la matrice disegno, α_g è il vettore dei parametri ed ε_g un fattore d'errore non necessariamente normale.

Inoltre si avrà che:

$E\,(y_g) = X\alpha_g$

$Var\,(y_g) = W_g\sigma_g^2$ dove W_g è una matrice di pesi non negativa nota

La statistica che di solito si usa è: $t_{gi} = \frac{\widehat{\beta}_{gi}}{s_g\sqrt{v_{gi}}} \sim t_{d_g}$

In cui $\hat{\beta}_{gi}|\,\beta_{gi}, \sigma_g^2 \sim N(\beta_{gi}, v_{gi}\sigma_g^2)$

con $\hat{\beta}_{gi} = C^T\hat{\alpha}_g$ e C invece sono i contrasti

Ed $s_g^2|\ \sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi_{d_g}^2$ dove d_g sono i gradi di libertà residui nel modello lineare per il gene g, s_g^2 è la stima di σ_g^2

Utilizzando questo modello la media a posteriori di σ_g^2 dato i s_g^2 è s_g^{*2} dove

$$s_g^{*2} = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

E la statistica test sarà: $t^*_{gi} = \frac{\hat{\beta}_{gi}}{s^*_g \sqrt{v_{gi}}} \sim t_{d_g + d_0}$

Osserviamo le prime dieci righe dell'output riguardante il test dell'Empirical Bayes sui dati normalizzati col metodo della trasformazione logaritmica generalizzata.

	ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
715	1636_g_at	0.9902055	9.263929	8.598257	6.086662e-13	7.684410e-09	18.303864
9823	39730_at	1.0018965	9.077702	8.336314	1.973023e-12	1.245471e-08	17.249154
714	1635_at	0.9927335	8.133635	7.180006	3.431551e-10	1.444111e-06	12.598185
757	1674_at	0.9508351	6.159706	6.683980	3.024510e-09	9.546111e-06	10.626715
10605	40504_at	0.7910088	5.804627	6.055673	4.502942e-08	1.071835e-04	8.175999
10300	40202_at	1.5355233	8.840548	6.026540	5.093871e-08	1.071835e-04	8.064042
2457	32434_at	1.0830481	5.926475	5.841771	1.108426e-07	1.999125e-04	7.358101
9931	39837_s_at	0.3781118	7.545638	5.769463	1.499229e-07	2.365971e-04	7.083892
11381	41274_at	0.3147187	6.189806	5.615255	2.842011e-07	3.986711e-04	6.503272
3810	33774_at	0.9338825	8.218654	5.444343	5.729878e-07	7.233971e-04	5.866897

La colonna ID è l'identificativo del gene, la colonna del log Fold Change rappresenta i cambiamenti percentuali dei valori di y, la terza colonna riporta i valori medi di espressione, seguita dalla colonna riguardante i valori della statistica t, e per finire le ultime due colonne riguardano i valori del p-value, da notare che sono estremamente piccoli e quindi altamente significativi contro l'ipotesi nulla, vale a dire che vi è differenza in media tra i geni con e senza riarrangiamento molecolare. Mentre l'Adjusted p-value (FDR) misura la significatività per il test con ipotesi multipla (Westfall and Young, 1993), in cui si definisce il livello per entrare o meno nel processo di una singola ipotesi H_j, considerando però i valori delle altre statistiche test coinvolte. Uno degli aspetti negativi è che la combinazione di molte migliaia di geni aumenta la probabilità di errore e quindi può spesso portare a rilevare dei falsi positivi (FDR: *false discovery rate*), come per esempio particelle di cDna che si sono legate alle sonde e invece non dovevano, oppure una errata sovra o sotto stima di espressione. L'ultima colonna infine riporta i valori della statistica Bayesiana.

Considerando che si possiedono 12625 geni, si contano 73 geni differenzialmente espresssi che si trovano in uno stato di incertezza, cioè appartenenti all'intervallo con un FDR (False Discovery Rate) tra (0.05;0.1), mentre sono 145 i geni differenzialmente espressi con FDR minore di 0.05 e quindi certamente significativi contro H_0.
Di seguito si riportano i primi geni differenzialmente espressi con FDR minore di 0.05:

ID	NOME DEL GENE
1036_at	"interleukin 15"
106_at	"runt-related transcription factor 3"
1107_s_at	"ISG15 ubiquitin-like modifier"
1134_at	"tyrosine kinase, non-receptor, 2"
1135_at	"G protein-coupled receptor kinase 5"
1140_at	"integrin, alpha E (antigen CD103, human mucosal lymphocyte antigen 1; alpha polypeptide)"

4.3 Test sulla Significance Analysis of Microarray **(SAM)**

La statistica SAM soddisfa sempre la verifica d'ipotesi:

$$\begin{cases} H_{0,g}: \mu_x(g) = \mu_y(g) \\ H_{1,g}: \mu_x(g) \neq \mu_y(g) \end{cases}$$

Tale che: $T_g = \frac{\bar{x}_g - \bar{y}_g}{s_{p,g}\left(\sqrt{\frac{1}{n_x} - \frac{1}{n_y}}\right) + s_0}$

Dove s_0 è una costante, denominata fudge factor,che ha il compito di evitare che i geni poco espressi dominino l'analisi. Inoltre è scelto come il percentile di s_p (calcolato attraverso un algoritmo) che rende il coefficiente di variazione di T_g costante, indipendentemente dai valori di $s_{p,g}$.

Bisogna considerare che i livelli di espressione vanno da valori prossimi allo zero a valori estremamente elevati, inoltre solitamente il numero di repliche per condizione sperimentale è molto basso. Tale caratteristica può quindi portare ad una maggiore probabilità di errore di primo tipo e per evitare ciò è necessario che la distribuzione di T_g sia indipendente dal livello di espressione dei geni. Si utilizza un approccio permutazionale che identifica i geni differenzialmente espressi confrontando le statistiche T_g originali con la quantità media ottenute su 100 permutazioni $T_g(E)$.

Si definisce per tanto una soglia oltre la quale si rifiuta l'ipotesi di uguaglianza di espressione, questa è data dalla disuguaglianza tra: $|T_g - T_g(E)| > \Delta$, dove Δ rappresenta la differenza tra il valore più grande e quello più piccolo per cui vale tale espressione. I geni estremi che non appartengono a tale differenza vengono definiti sovra espressi e sottoespressi.

Si osserva ciò attraverso il seguente la Figura 13:

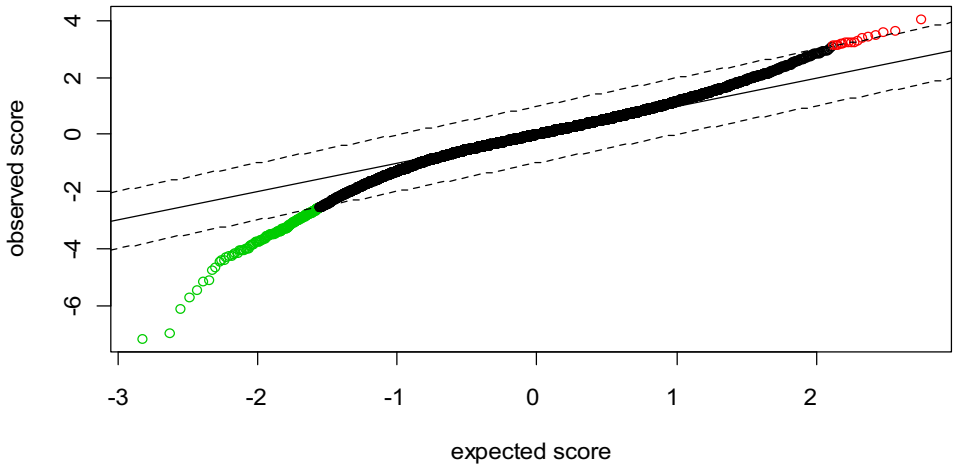


Figura 13:Grafico Sam

Si osserva che c'è un buon legame tra gli score attesi e quelli realmente osservati realizzando un rapporto monotono crescente. Il Δ è all'incirca pari a 2, i geni sovraespressi sono all'incirca maggiori di (2.5, 3), mentre

quelli sottoespressi sono minori all'incirca di (-1.5, -3). Infine si contano un numero maggiore di geni sottoespressi, infatti sono ben 191 geni, rispetto a quelli sovra espressi che sono solo 16, che corrispondono rispettivamente allo 0.001% e allo 0.01% rispetto ai 12625 geni considerati, essendo delle percentuali basse possiamo considerare buoni i risultati.

- CONFRONTO TRA IL TEST EBAYES E IL TEST SAM

Avendo fatto inferenza con due statistiche test diverse ma per valutare la stessa verifica d'ipotesi è interessante osservare i risultati delle due statistiche congiuntamente.

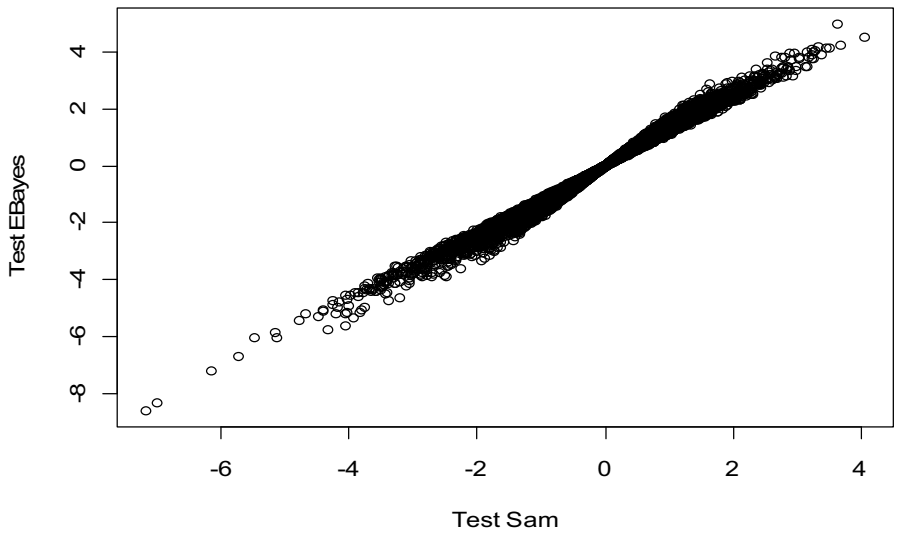


Figura 14: Confronto tra eBayes e Sam

Si noti come dalla Figura 14 come i due test siano molto concordi. Infatti si osserva che i risultati delle due statistiche test sono proporzionati, all'aumentare della statistica test Sam aumentano i valori della statistica Empirical Bayes.

Osservano però i geni differenzialmente espressi che risultano significativi sia per la statistica Empirical Bayes che per quella Sam ne risultano che sono 132 con un livello del p-value minore di 0.05. In particolare i primi geni differenzialmente espressi corrispondono a tale lista:

ID	NOME DEL GENE
1036_at	"interleukin 15"
106_at	"runt-related transcription factor 3"
1107_s_at	"ISG15 ubiquitin-like modifier"
1134_at	"tyrosine kinase, non-receptor, 2"
1135_at	"G protein-coupled receptor kinase 5"
1140_at	"integrin, alpha E (antigen CD103, human mucosal lymphocyte antigen 1; alpha polypeptide)"

Si osserva che la lista dei primi geni più significativi è uguale alla lista trovata con la Statistica dell'Empirical Bayes.

4.4 Test Rank Prodouct

Anche il test Rank Prod identifica i geni differenzialmente espressi. Viene denominato così poiché utilizza il prodotto dei ranghi attraverso una statistica non parametrica.

Dati n geni e k repliche si assume che $r_{g,i}$ è il rango del gene g nella i-esima replica e il prodotto dei ranghi sarà uguale ad:

$$RP_g = \left(\prod_{i=1}^{k} r_{g,i} \right)^{\frac{1}{k}}$$

Se il valore della statistica RP è basso, allora vuol dire che sarà bassa la probabilità di osservare, in maniera casuale, il passo più alto della lista.

Per quanto riguarda la matrice di espressione sulla leucemia linfatica di tipo B, si osserva che i gruppi (BCR e NEG) di geni differenzialmente espressi del tipo up regolati con soglia minore di 0.05 sono 914 (0.07% dei geni), mentre del tipo down regolati sono 1025 (0.08% dei geni).

Osserviamo le tabelle di geni sovra (Table 1) e sotto espressi (Table 2):

$Table1

	gene.index	RP/Rsum	FC:(class1/class2)	pfp	P.value
38514_at	8596	149.6021	0.4151	0.0000	0.0000
37014_at	7082	381.5067	0.4956	0.0000	0.0000
39318_at	9408	398.2625	0.6348	0.0000	0.0000
31525_s_at	1539	452.6770	0.7274	0.0000	0.0000
39878_at	9972	470.6205	0.5961	0.0000	0.0000...

$Table2

	gene.index	RP/Rsum	FC:(class1/class2)	pfp	P.value
40202_at	10300	110.7056	2.8989	0.0000	0.0000
37006_at	7074	169.9787	2.3635	0.0000	0.0000
36638_at	6703	210.0667	2.2178	0.0000	0.0000
37027_at	7095	274.4993	2.1596	0.0000	0.0000
36275_at	6336	311.9076	2.0972	0.0000	0.0000

La prima colonna rappresenta l'indice dei geni nel data set originale, la seconda colonna rappresenta invece il codice dei geni indicizzati, la terza il prodotto del rango completo per ogni gene. Infine la quinta e l'ultima colonna riportano i valori del pfp (proporzione dei falsi positivi, definito anche come FDR) e del p-value (minori di 0.05).

Osserviamo infine, tramite il diagramma di Venn (Figura15), il numero di geni che risultano differenzialmente espressi sia con la Statistica Test

Empirical Bayes, sia con la Statistica della Significance Analysis of Microarray e sia con la Statistica Rank Produoct.

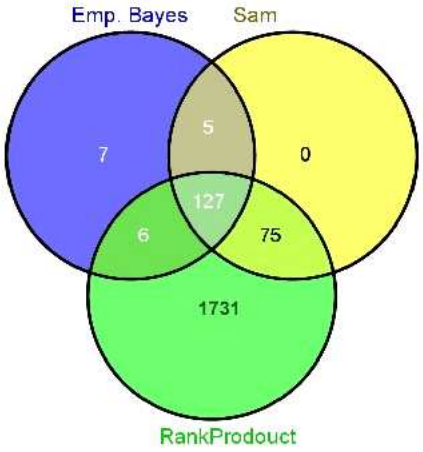


Figura 15: Diagramma di Venny sui geni comuni differenzialmente espressi.

Di conseguenza i geni comuni per i campione BCR che NEG risultano 127 fra geni sovra espressi o sotto espressi significativi per entrambi i gruppi, con un p-value inferiore a 0.05.

Nello specifico si riportano di seguito i primi 10 geni differenzialmente espressi dei 127 comuni.

ID	NOME DEL GENE
1036_at	"interleukin 15"
106_at	"runt-related transcription factor 3"
1134_at	"tyrosine kinase, non-receptor, 2"
1135_at	"G protein-coupled receptor kinase 5"
1211_s_at	"CASP2 and RIPK1 domain containing adaptor with death domain"
1249_at	"GRB2-associated binding protein 1"
1326_at	"caspase 10, apoptosis-related cysteine peptidase"
1467_at	"epidermal growth factor receptor pathway substrate 8"

```
160027_s_at "insulin-like growth factor 2 receptor"

1635_at     "c-abl oncogene 1, receptor tyrosine kinase"
```

Di conseguenza si può concludere che i geni differenzialmente sovra e sotto espressi sono pochi, poiché solo lo 0.01% risultano significativi per la matrice di espressione. Nonostante ciò è comunque interessante capire come mai determinati geni assumono comportamenti di sregolatezza estremi mantenendo un'importanza significativa per l'analisi.

CAPITOLO 5

GENE SET ANALYSIS

5.1 Ipotesi competitiva e indipendente

Dopo aver eseguito le varie analisi differenziali si è giunti ad una lista di 127 geni differenzialmente espressi (up e down regolati). Per poter dare un significato a questa lista di geni si ricorre all'utilizzo di annotazioni funzionali come la GeneOntology (GO) e i pathway (KEGG) che favoriscono il raggruppamento dei geni in categorie. Questo tipo di analisi viene chiamata col nome di Gene Sets Analysis, basandosi su test fatti separatamente su ogni gene set, assumendo l'indipendenza tra questi.

A seconda della verifica d'ipotesi che ci si pone si intraprendono metodologie diverse. Infatti l'ipotesi nulla può essere di tipo competitiva o indipendente (Dinu et al., 2008):

- H_0 competitiva: La significatività di una gene set dipende non solo dai geni all'interno dell'insieme (quindi associati ad uno stesso fenotipo F), ma anche sugli altri geni in tutto l'array.
- H_0 indipendente: La significatività di una gene set dipende solo dai geni all'interno dell'insieme, quindi associati al fenotipo F.

I test con l'ipotesi nulla indipendente risultano più potenti e quindi danno risultati significativi per più gene sets.

Per quanto riguarda i pathwhay, si utilizza il tipo KEGG (Kyoto Encyclopedia of Genes and Genomes), dove i geni prodotti sono classificati in pathway (percorsi) di segnale e metabolici. Un pathway è un diagramma grafico di una reazione biochimica che ha la proprietà di controllare diversi enzimi.

Di seguito verranno applicate diversi test:

1) Ipergeometrica (appartenente ai test con ipotesi nulla competitiva), utilizzando tre categorie diverse di Gene Ontolog (Biological Process, Molecular Function, Cellular Component) e il pathway KEGG;
2) Global Test (con ipotesi nulla indipendente) e infine
3) un test con ipotesi mista, l'effetto del pathway considerato (SPIA).

I dati utilizzati per tali analisi sono quelli derivati dall'intersezione dei tre test precedentemente applicati per ottenere i geni differenzialmente espressi. Quindi si hanno un totale di 127 geni.

5.2 Gene sets analysis con la Ipergeometrica (Competitivo)

Data un'urna di cardinalità N con geni divisi in appartenenti alla categoria G di cardinalità g e non appartenenti a G di cardinalità N-g, l'ipergeometrica descrive la probabilità di ottenere esattamente k geni nella categoria funzionale G di numerosità h estraendo senza reinserimento r geni.

Osserviamo le risposte al test utilizzando tre diverse categorie Gene Ontology (Biological Process, Molecular Function e Cellular Component) e successivamente anche il pathway KEGG.

Sotto si riportano per semplicità le prime dieci.

1) Biological Process

```
Gene to GO BP  test for over-representation

1479 GO BP ids tested (21 have p < 0.001)

Selected gene set size: 103

    Gene universe size: 7757

    Annotation package: hgu95av2
```

summary(BiologicalProcess.iper)

```
       GOBPID       Pvalue OddsRatio   ExpCount Count Size                                Term
1  GO:0023052 1.979842e-06  2.564548 32.9833699    56 2484                           signaling
2  GO:0007165 3.883118e-05  2.324002 24.5250741    43 1847                 signal transduction
...
16 GO:0035023 7.693176e-04  7.759184  0.7303081     5   55      regulation of Rho protein signal
                                                                                transduction
...
19 GO:0010562 8.871118e-04  5.931097  1.1286580     6   85     positive regulation of phosphorus
                                                                          metabolic process
20 GO:0045937 8.871118e-04  5.931097  1.1286580     6   85      positive regulation of phosphate
                                                                          metabolic process
```

Sono risultati significativi, con un p-value minore di 0.001, 21 gruppi di geni su 1479 classi della Biological Process. L'informazione più notevole è data dalla colonna dell' Odds Ratio, che misura il rischio di quanto un fattore può influire sulla malattia. Essendo tutti maggiore di 1 vuol dire che quelle classi di geni influiscono negativamente sull'insorgenza della leucemia ALL di tipo B. I gruppi che hanno l'odds ratio più alto nella gene set sono "regulation of Rho protein signal transduction" (OR=7.7), "positive regulation of phosphorus metabolic process" (OR=5.9), "positive regulation of phosphate metabolic process" (OR=5.9).

2) Molecular Function

```
Gene to GO MF  test for over-representation

318 GO MF ids tested (9 have p < 0.001)

Selected gene set size: 107

    Gene universe size: 7940

    Annotation package: hgu95av2
```

summary(MolecularFunction.iper)

	GOMFID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1	GO:0030234	5.163326e-05	3.180107	6.96712846	19	517	enzyme regulator activity
2	GO:0030674	6.195563e-05	7.993382	1.01070529	7	75	protein binding, bridging
3	GO:0043621	3.629284e-04	28.215144	0.14823678	3	11	protein self-association
4	GO:0005088	4.370759e-04	8.880529	0.64685139	5	48	Ras guanyl- nucleotide exchange factor activity
5	GO:0005534	5.350296e-04	149.180952	0.04042821	2	3	galactose binding
6	GO:0005085	7.446330e-04	6.144950	1.09156171	6	81	guanyl-nucleotide exchange factor activity
7	GO:0060090	7.572291e-04	7.787115	0.72770781	5	54	molecular adaptor activity
8	GO:0030695	8.108739e-04	3.688108	3.00516373	10	223	GTPase regulator activity
9	GO:0060589	9.615165e-04	3.601154	3.07254408	10	228	nucleoside-triphosphatase regulator activity

Per quanto riguarda la gene-sets analisys basata sulla Molecolar Function sono risultati significativi, con un p-value minore di 0.001, 9 gruppi di geni su 318. Osservando nuovamente il rischio relativo quelle molecole che influiscono, sull'insorgenza e l'aggravare della leucemia linfatica acuta, sembrano essere le molecole riguardante il " galactose binding" con un odds ratio estremamente alta pari a 149, tale indice suggerisce un comportamento di tale molecola sicuramente rilevante per l'insorgenza

della malattia, inoltre importante è anche la molecola “protein self-association” (OR=28).

3) Cellular Component

```
Gene to GO CC  test for over-representation

201 GO CC ids tested (5 have p < 0.001)

Selected gene set size: 103

    Gene universe size: 8008

    Annotation package: hgu95av2
```

`summary(CellularComponent.iper)`

	GOCCID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1	GO:0005924	0.0001903108	8.087629	0.8489011	6	66	cell-substrate adherens junction
2	GO:0005925	0.0001903108	8.087629	0.8489011	6	66	focal adhesion
3	GO:0030055	0.0002634814	7.578286	0.9003497	6	70	cell-substrate junction
4	GO:0031091	0.0004273903	8.911565	0.6431069	5	50	platelet alpha granule
5	GO:0005576	0.0009619642	2.190009	13.9168332	26	1082	extracellular region

Infine rispetto alla componente cellulare si rilevano significativi solo 5 gruppi di geni su 201.

Anche in questo caso non si evidenziano gruppi di geni attivi come fattori di protezione ma sono tutti maggiori di 1, tra le classi di categorie con un fattore di rischio più alto troviamo quelli appartenenti alla gene set del “platelet alpha granule”, della “cell-substrate adherens junction” e della “focal adhesion”.

In sintesi adoperando le tre categorie “Biological Process”, “Molecular Function”, “Cellular Component” sono state estratte rispettivamente 21, 9

e 5 classi significative con un p- value minore di 0.001, per essere più conservativi.

4) Pathway KEGG

```
Gene to KEGG  test for over-representation

88 KEGG ids tested (4 have p < 0.05)

Selected gene set size: 51

    Gene universe size: 2899

    Annotation package: hgu95av2
```

`summary(KEGG.iper)`

	KEGGID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1	05416	0.008766142	5.486953	0.8256177	4	57	Viral myocarditis
2	04150	0.022103823	5.498397	0.6083499	3	42	mTOR signaling pathway
3	05130	0.023523257	5.359375	0.6228344	3	43	Pathogenic Escherichia coli infection
4	04622	0.046371453	4.029481	0.8111332	3	56	RIG-I-like receptor signaling pathway

Per quanto riguarda l'analisi utilizzando il pathway KEGG si prende in considerazione una significatività minore di 0.05 (poiché per 0.01 non è stato rilevato nessuno gruppo) e si contano solo 4 gruppi di geni significativi su 88. In questo caso si osserva che i diversi pathways hanno tutti un rischio relativo intorno a 5 e 4 per l'ultimo pathway.

5.3 Gene sest Analysis con il Global Test (Test indipendente)

Il Global Test è un'inferenza basata su modelli GLM, in particolare è uno score test basato sulla derivata della funzione di verosimiglianza (Liu et Al., 2007).

In questo caso non sono più di interesse solo i geni differenzialmente espressi ma tutti i 12625 geni all'interno dei 78 campioni, che verranno distribuite in determinate classi.

Iniziando a considerare i pathway KEGG, osserviamo dall'output del summary che sono 199 le classi di geni sottoposte alla statistica test, di cui però con una successiva analisi si osserva che sono 119 quelle che possiedono un p-value minore di 0.05.

Di seguito si riportano le prime classi di geni maggiormente significative.

	BH	alias	p-value	Statistic	Expected	Std.dev	#Cov
04360	2.73e-05	Axon guidance	1.37e-07	7.9155	1.30	0.534	166
05416	9.45e-05	Viral myocarditis	1.49e-06	7.1694	1.30	0.609	97
04012	9.45e-05	ErbB signaling pathway	1.81e-06	5.6858	1.30	0.480	170

Facendo parte dell'ipotesi nulla indipendente, il Global Test analizza solo i geni all'interno della propria gene set osservando la significatività dei geni contro l'associazione al fenotipo F.

Si osserva una deviazione standard molto bassa, a rilevare, quindi, una certa omogeneità tra i geni all'interno di ogni gene set.

Riscontriamo delle gene sets altamente significative contro l'ipotesi nulla, quindi dobbiamo sostenere che ci siano delle dipendenze tra i geni che appartengono ad un certo fenotipo e che quindi vanno ad influenzare lo sviluppo della leucemia linfatica acuta (ALL).

Tra i pathways delle gene sets maggiormente significativi troviamo il "Axon guidance" (riguarda lo sviluppo neurale, attraverso cui i neuroni inviano segnali per raggiungere gli obiettivi corretti), il "Viral myocarditis" (infezione virale del muscolo cardiaco) e il "ErbB signaling" (è un pathway associato allo sviluppo delle malattie neurodegenerative).

Si può dire che l'ambiente neurologico ha un ruolo importante per lo sviluppo della malattia ALL anche se questi sono solo alcuni delle categorie rilevate come significative.

Osserviamo graficamente, dalla Figura 16, come si comporta, per esempio, la classe Viral Myocarditis rispetto alla verifica d'ipotesi, anche se si è già notato che è altamente significativa.

E'una classe molto numerosa, e la presenza predominante delle barre di colore rosso rispetto a quelle verdi mostra una negativa associazione di risposta al test e anche una maggiore numerosità dei geni sovra-espressi nei campioni BCR e NEG. Invece l'altezza delle barre misura la significatività dei geni all'interno del pathway.

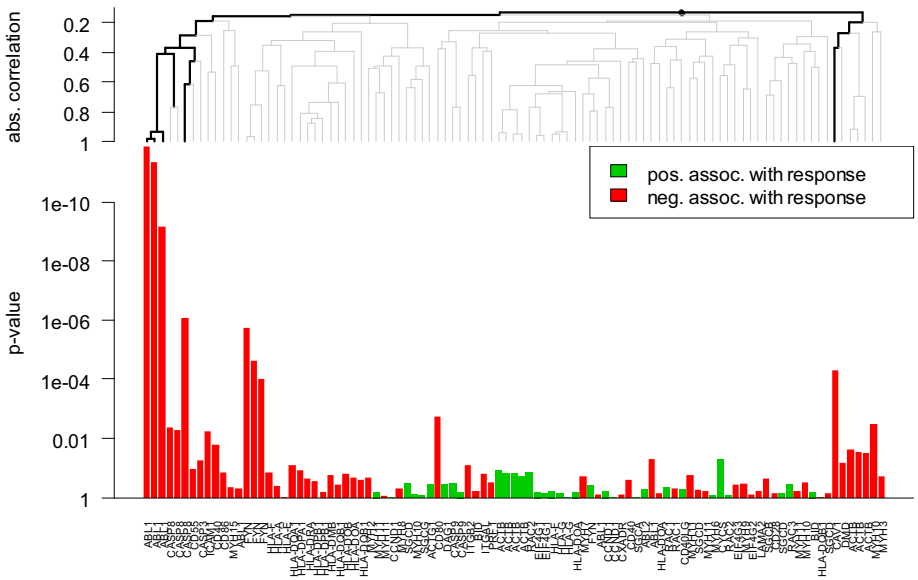


Figura 16: Global Test per la classe Viral Myocarditis.

Notiamo che i primi 3 geni ABL1, CASP8, FYN e SGCD sono tutti altamente significativi.

Non potendo definire quale dei due metodi sia migliore (Ipergeometrica o Global Test), poiché verificano ipotesi diverse, utilizzando sempre il pathway KEGG, estrapoliamo quelle gene- sets che sono risultate significative sia con il metodo dell'ipergeometrica e sia con il global test.

Osserviamo che l'intersezione tra i risultati del Global Test e quelli dell'Ipergeometrica, includono tutti i risultati di quest'ultimo. Infatti erano risultati quattro gruppi di geni dalla precedente analisi con i pathway KEGG:

```
"Viral myocarditis", "mTOR signaling pathway",
"Pathogenic Escherichia coli infection", "RIG-I-
like receptor signaling pathway"
```

Pertanto le gene set significative con il metodo dell'Ipergeometrica possono essere viste anche come un sotto gruppo della verifica d'ipotesi indipendente, tipica del Global Test.

E' possibile analizzare nel dettaglio i gruppi di geni appartenenti alle gene set risultate significative per entrambi i test (Figure 17, 18 e 19).

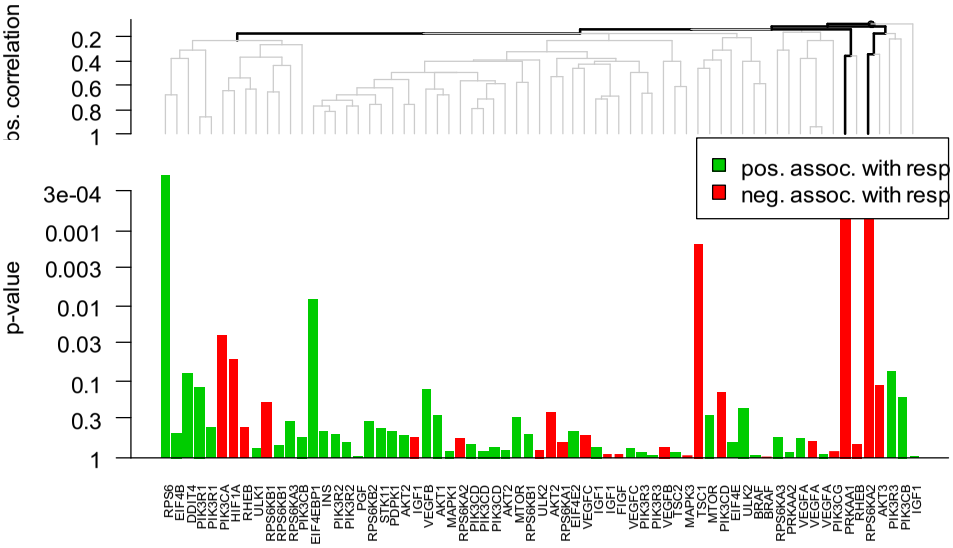


Figura 17: GENE SET "mTOR signaling pathway".

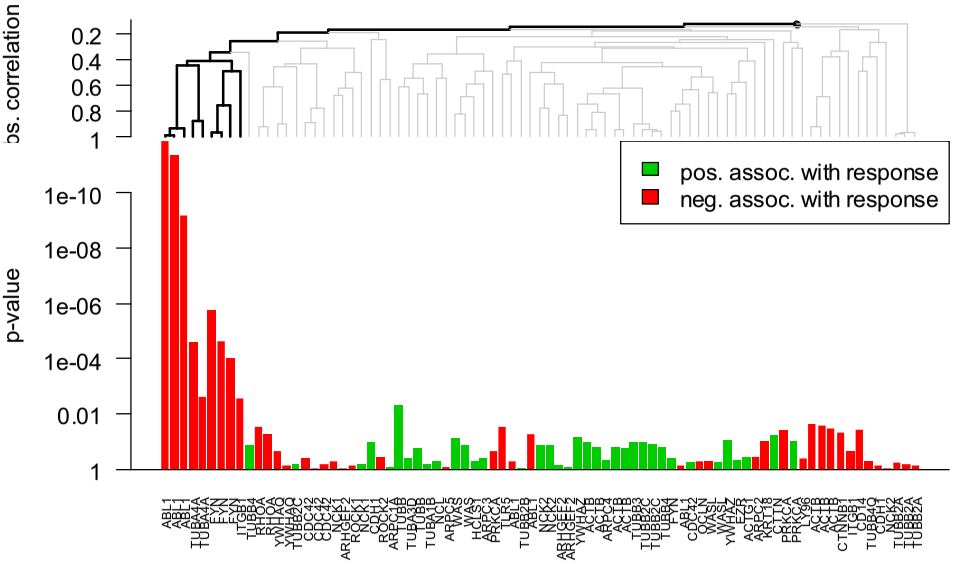
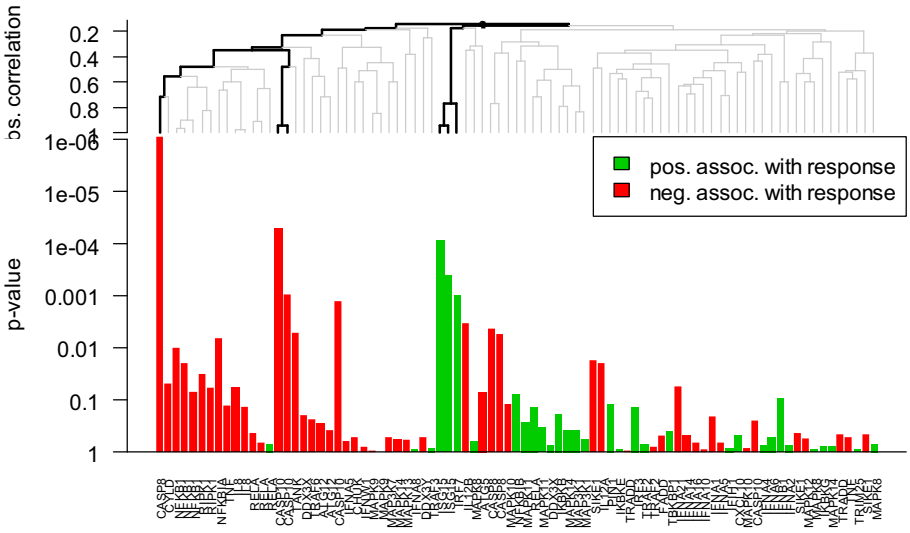


Figura 18: GENE SET "Pathogenic Escherichia coli infection".



<u>Figura 19</u>: GENE SET "RIG-I- like receptor signaling pathway".

A parte nel pathway “mTOR signaling pathway”, in tutti i restanti si osserva una dominanza di geni sovra espressi.

A livello biologico saranno proprio quei geni le cui caratteristiche risulteranno interessanti da analizzare. Le gene set che hanno nel loro interno una maggiore associazione negativa sono il “Viral myocarditis” e il “RIG-I-like receptor signaling pathway”.

Confronteremo inoltre i risultati del Global Test sempre con quelli del metodo dell’Ipergeometrica, ma tenendo conto delle categorie della Gene Ontology. Quindi si ricalcolerà innanzitutto il Global Test rispetto a quest’ultima categoria e poi sarà possibile eseguire un confronto.

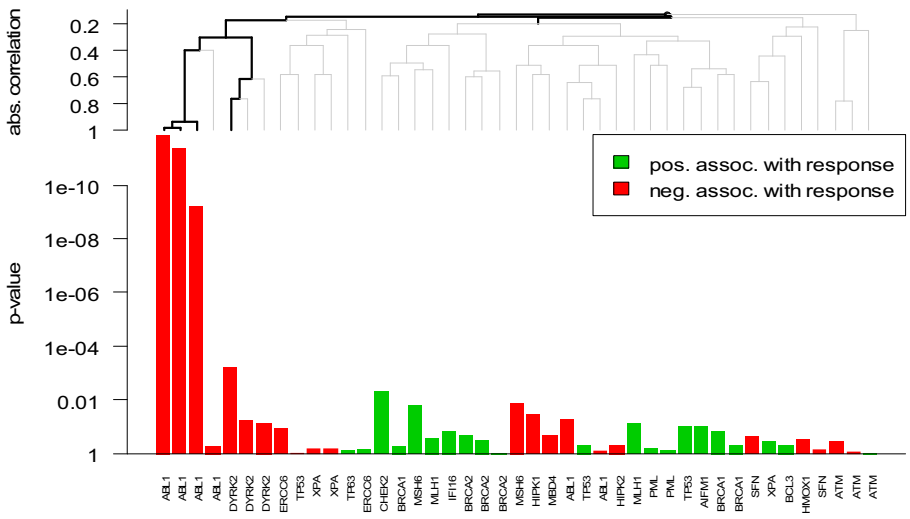
In questo caso i gruppi di geni selezionati sono altamente numerosi infatti si contano 11384 raggruppamenti totali e 698 quelli che hanno un p- value minore di 0.001.

Di cui riportiamo di seguito le prime righe:

	BY		alias	p-value	Statistic	Expected	Std.dev	#Cov
GO:0008630	8.896183e-07	DNA damage response, signal transduction r...		1.226904e-11	15.66650	1.298701	0.7121782	43
GO:0000115	8.896183e-07	regulation of transcription involved in S-...		1.863458e-11	35.99141	1.298701	1.4692178	6
GO:0000084	8.896183e-07	S phase of mitotic cell cycle		2.363959e-11	15.80640	1.298701	0.7468797	37
GO:0006298	1.630562e-06	mismatch repair		5.777132e-11	14.92165	1.298701	0.8172984	31
GO:0051353	2.578883e-06	positive regulation of oxidoreductase acti...		1.142133e-10	14.79898	1.298701	0.7398250	38
GO:0004715	8.109770e-06	non-membrane spanning protein tyrosine kin...		4.309975e-10	10.53133	1.298701	0.5764243	73

Le gene set derivanti con un p-value minore di 0.001 sono ben 698. Anche in questo caso si osserva una deviazione standard abbastanza bassa. Tra le gene set maggiormente significative troviamo le GO corrispondenti a "DNA damage response" e "regulation of transcription involved".

Osserviamo nuovamente il grafico (Figura 20) rispetto alla classe più significativa, quale la "DNA damage response".



<u>Figura 20</u>: Gene Ontology di "DNA damage response".

In questa classe di categoria la numerosità dei geni è più ridotta e quindi il grafico è più leggibile. Osserviamo che primi geni che hanno una significatività più alta, all'interno del gruppo, sono gli stessi geni ABL del gruppo "Viral myocarditis" rispetto al pathway KEGG.

Eseguendo quindi un'intersezioni fra le due analisi (Global Test e Ipergeometrica), rispetto alla Gene Ontology, ottniamo i seguenti risultati.

Gruppi di geni comuni rispetto alla Gene Onology Biological Process:

"signaling", "signal transduction", "regulation of apoptosi", "regulation of programmed cell death", "regulation of cell death","signal transmission", "signaling process", "regulation of signal transduction", "regulation of signaling process", "positive regulation of apoptosis", "positive regulation of programmed cell death", "positive regulation of cell death","induction of apoptosis", "induction of programmed cell death", "death", "anatomical structure morphogenesis", "apoptosis", "programmed cell death"

Si contano 18 categorie comuni.

Gruppi di geni comuni rispetto alla Gene Onology Molecular Function:

"enzyme regulator activity", "protein binding, bridging", "Ras guanyl- nucleotide exchange factor activity", "galactose binding", "molecular adaptor activity", "GTPase regulator activity", "nucleoside-triphosphatase regulator activity".

In questo caso invece sono 7 le categorie avute dall'intersezione.

Ed infine ecco i gruppi di geni comuni rispetto alla Gene Ontology Cellular Component:

"cell-substrate adherens junction", "focal adhesion", "cell-substrate junction", "platelet alpha granule", "extracellular region"

Le categorie comuni sono i 5 gruppi totali della Ipergeometrica rispetto alla Cellular Component.

5.4 Analisi sull'affetto del pathway considerato (SPIA) (Test misto)

Attraverso il metodo SPIA si combina il risultato ottenuto dall'analisi di arricchimento con un nuovo tipo di metodologia, che misura la perturbazione di pathway sotto un'opportuna condizione. Da ciò segue la possibilità di calcolare un p-value che misura la significatività globale del pathway, derivante dalla combinazione del p-value di arricchimento con quello di perturbazione.

Viene definito un fattore genico di perturbazione come: $PF_{(g_i)} = \Delta E(g_i) + \sum_{j=1}^{n} \underline{\beta}_{ij} \frac{PF_{(g_j)}}{N_{ds}(g_j)}$, in cui $\Delta E(g_i)$ rappresenta la misura normalizzata del cambio di espressione del gene g_i e $N_{ds}(g_j)$ è il numero di geni differenzialmente simili, tale equazione essenzialmente descrive il fattore di perturbazione PF per un gene g_i come una funzione lineare dei fattori di perturbazione di tutti i geni in un determinato pathway. Dalla differenza tra il fattore genico di perturbazione e il suo log fold change osservato è possibile ricavare l'accumulazione genica ($Acc(g_i) = PF_{(g_i)} - \Delta E(g_i)$). Di conseguenza la probabilità di osservare l'accumulo totale di perturbazione del pethway è data da: $P_{PERT} = P(T_A \geq t_A | H_0)$, in cui T_A è un valore al netto della perturbazione accumulata più estrema di t_A ($t_A = \sum_i Acc(g_i)$).

L'analisi d'effetto combina due tipi di analisi: la rappresentazione dei geni differenzialmente espressi in un dato pathway e la perturbazione irregolare del pathway. Questi due aspetti sono rilevati da due valori di probabilità indipendenti, P_{NDE} e P_{PERT}.

I risultati che si ottengono con il metodo SPIA hanno una sensitività e specificità più affidabile rispetto ai risultati ottenibili da altri metodi di analisi di pathways.

Il pacchetto SPIA di Bioconductor utilizza l'annotazione biologica KEGG, che include le descrizioni sulle interazioni dei geni ed inoltre ha la capacità di eseguire un'analisi automatica mirata ad identificare i geni che sono rilevanti in una determinata condizione.

In sintesi SPIA utilizza come dato il vettore dei geni differenzialmente espressi con il loro fold changes (descrive l'alterazione del gene in un gruppo rispetto all'altro), così come la topologia di pathways al fine di valutare la significatività dei pathways in base alle condizioni di studio.

Si inizia l'analisi SPIA utilizzando il vettore dei 127 geni differenzialmnente espressi comuni ai test eseguiti in precedenza.

Per lavorare con SPIA si necessita di un vettore con log2 fold changes per tutti i geni considerati ad essere differenzialmente espressi.. I geni devono però essere convertiti in Entrez- Gene. Poiché diversi probosets possono riferirsi ad uno stesso Entrez ID, ci sono due modi per ottenere un log fold change: la prima opzione è quella di usare il fold change del proboset più significativo per ogni gene, mentre la seconda opzione è quella di utilizzare la media del logaritmo dei fold-changes di tutti i probosets dello stesso gene. Utilizzeremo per i campioni leucemici questo ultimo metodo.

Di seguito si mostrano le prime cinque righe della matrice necessaria per l'utilizzo di SPIA.

```
             ID     logFC ENTREZ
1       1036_at 0.8512077   5595
2        106_at 0.6260991   7075
3       1134_at 0.7958113    643
```

```
5     1211_s_at 0.5940707   4319
```

Inoltre l'algoritmo SPIA attraverso il vettore ENTREZ e attraverso il logFC produce una tavola con di classifica dei pathways.

Di seguito vengono illustrate le prime dieci righe su 73, derivanti dall'analisi sull'effetto dei pathway.

	Name	ID	pSize	NDE	tA	pNDE	pPERT	pG	pGFdr	pGFWER	Status
1	Tight junc	04530	1	1	0.6119763	1	0.025	0.1172220	1	1	Activated
2	Long-term	04730	2	2	0.8285456	1	0.069	0.2534818	1	1	Activated
3	Cell cycle	04110	2	2	2.3694213	1	0.086	0.2969931	1	1	Activated
4	Epithelial	05120	5	5	-0.4279804	1	0.103	0.3371217	1	1	Inhibited
5	Notch sign	04330	1	1	-0.4193652	1	0.172	0.4747649	1	1	Inhibited
6	Regulation	04810	4	4	-1.3569760	1	0.220	0.5531081	1	1	Inhibited
7	Fc epsilon	04664	4	4	0.1395371	1	0.245	0.5895918	1	1	Activated
8	Prion dise	05020	3	3	0.4201464	1	0.251	0.5979579	1	1	Activated
9	Vascular s	04270	1	1	0.1395371	1	0.254	0.6020869	1	1	Activated
10	GnRH signa	04912	2	2	0.1395371	1	0.255	0.6034554	1	1	Activated

Si nota immediatamente che l'analisi ci restituisce i pathways (attraverso la colonna Name) dai più ai meno significativi rispetto all'indice di perturbazione e al p-value della probabilità combinata.

La colonna *Name* riporta quindi i nomi per ogni pathway, mentre l'ID è il corrispondente codice identificativo. Per quanto riguarda il *pSize* e la colonna *NDE* riportano il numero di geni totali e il numero di geni differenzialmente espressi all'interno del pathway corrispondente. Naturalmente i valori di *NDE* sono tutti maggiori di uno poiché l'analisi che si sta ponendo è gia su un vettore di geni differenzialmente espressi. Inoltre la colonna tA corrisponde al valore totale di accumulazione della perturbazione netta nel pathway (non sembra esserci una particolare

correlazione tra i valori positivi e negativi del t_A rispetto al p-value riguardante l'accumulazione di perturbazione netta. Per quanto riguarda il $pNDE$ è il p-value che rileva la significatività del dato pathway P_i come definito dall'analisi sulla rappresentazione del numero di geni differenzialmente espressi (N_{DE}) osservati nel pathway ($P_{NDE} = P(X \geq N_{DE} | H_0)$). L'ipotesi nulla H_0 suppone che i geni che appaiono come DE (diffrenzialmente espressi) in un determinato pathway siano completamente randomizzati. Da una prospettiva biologica, ciò significa che il pathway non è rilevante per la condizione sotto studio, Inoltre i valori del P_{NDE} sono ottenuti assumendo che N_{DE} (il numero di geni DE in un dato pathway) seguono una distribuzione ipergeometrica. Dai risultati ottenuti si può dedurre che il numero di geni DE, nei diversi pathway prominenti, non è rilevante per lo studio. Per quanto riguarda invece il $pPERT$ rappresenta la significatività sulla probabilità dell'ammontare delle perturbazioni misurate in ogni pathway. Osserviamo nello studio che solo il primo pathway ha una significatività minore di 0.05, nel senso che i fold changes osservati nel pathway restituiscono un impatto significativo in considerazione al numero di geni DE, localizzati in maniera casuale, sullo stesso pathway.

Infine P_g è il valore di una probabilità combinata. Per ultimo lo stato invece definito come inibited o activeted rappresenta il tipo di relazione esistente fra i geni/ proteine all'interno dei pathways, che quindi risulta essere per lo più di tipo attiva o inibitoria.

Si è osservato che solo il primo pathway ha una significatività del $pPERT$ minore di 0.05, ciò probabilmente è dovuto alla numerosità esigua dei geni (127) distribuiti nei pathway. Per tanto l'unico pathway più considerevole è il *Tight junc* (descrive un complesso di cellule le cui membrane si uniscono per formare una barriera impermeabile al fluido, fondamentale per le funzioni vitali).

Di seguito con la Figura 21 si mostra la funzionalità di tale pathway, si nota come attraverso di esso le cellule si restringono a formare come un nodo.

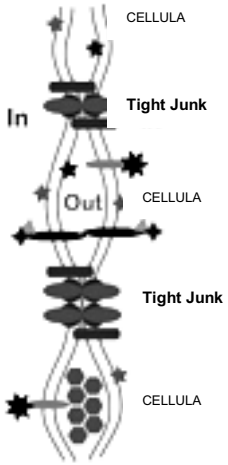


Figura 21: Immagine della funzionalità del Tight Junk.

Per capire meglio la costruzione di un pathway e in particolare quello di Tight Jung, osservare la sua struttura attraverso la Figura 22.

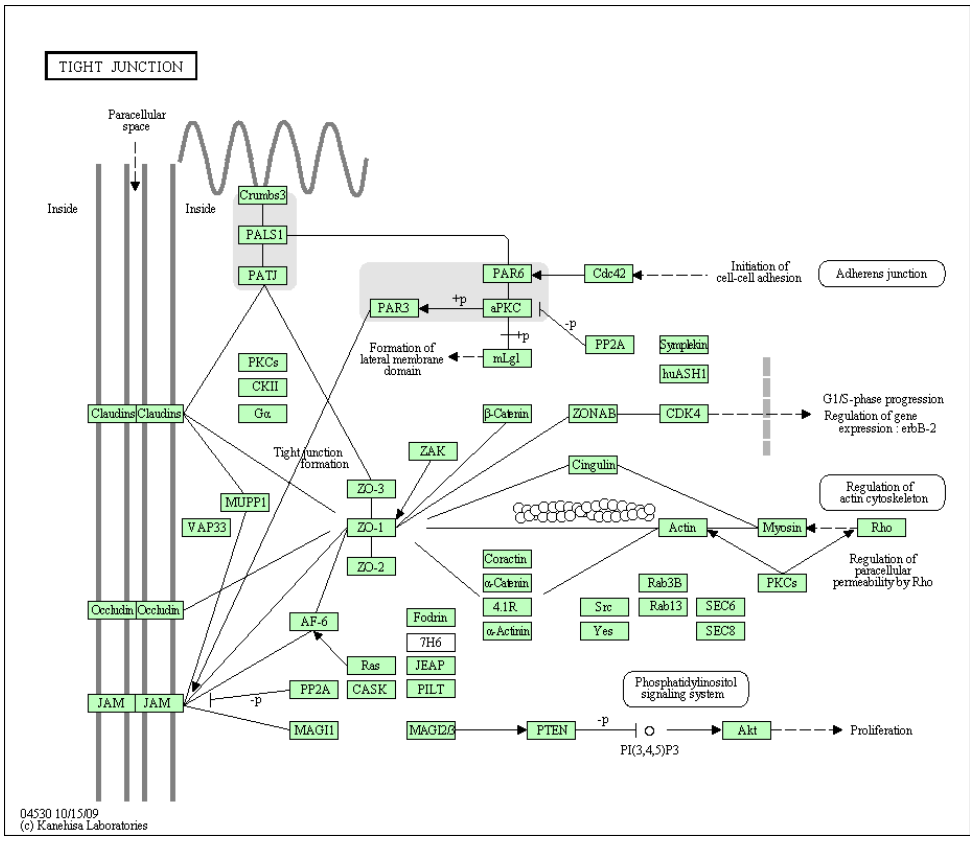


Figura 22: Pathway Tight Junk.

Attraverso il pathway è possibile osservare le relazioni tra i geni.

CAPITOLO 6

CONCLUSIONI

Nel presente studio si sono analizzati i dati di espressione su 78 campioni di leucemia linfatica acuta (ALL) di tipo B e a cui non è ancora stata somministrata una cura convenzionale (Chiaretti et al, 2005). In particolare ci si è focalizzati sullo studio delle alterazioni molecolari dovute ad un riarrangiamento genico chiamato BCR/ABL, in cui due geni si fondono insieme come risultato di una divisione cellulare anomala. I riarrangiamenti genici sono molto frequenti nei tumori e spesso portano ad un fenotipo tumorale più aggressivo. I 78 campioni sono suddivisi in 37 campioni BCR/ABL positivi e 41 campioni di controllo.

Lo studio ha identificato una serie di geni differenzialmente espressi (DE) tra pazienti affetti da leucemia linfatica acuta con e senza riarrangiamento cromosomico. Successivamente sono stati trovati quei pathway in cui questi geni sono coinvolti.

Per poterli studiare si è partiti innanzitutto dalla normalizzazione dei dati di espressione, applicando sia la normalizzazione quantile, sia la normalizzazione tramite la trasformazione logaritmica generalizzata. Si è osservato che entrambe normalizzano bene i dati poiché entrambe mantengono una mediana costante fra i diversi campioni BCR / ABL positivi e negativi. Attraverso il grafico MA si è osservata una leggera tendenza dei dati verso valori positivi (tendenza già osservata con i dati

grezzi), nonostante ciò si distribuiscono in maniera abbastanza simmetrica intorno allo zero, e che quindi non evidenziano importanti differenze rispetto al campione di riferimento. Infine anche attraverso l'analisi dei cluster si è osservato un comportamento molto simile tra le due normalizzazioni, in cui prevale un primo gruppo più numeroso che contiene la maggior parte dei campioni BCR/ABL e dei campioni negativi, in maniera proporzionale, e un secondo gruppo che riunisce i restanti campioni. Si è pertanto deciso di scegliere la normalizzazione che utilizza la trasformazione logaritmica generalizzata in quanto per costruzione ha l'obbiettivo di rendere la varianza tra geni costante, quindi con uno scarto interquartile più piccolo possibile, infatti si è notato attraverso i box-plot una distribuzione dei dati più compatta rispetto alla normalizzazione quantile.

Avendo scelto la normalizzazione da applicare ai dati si è passati all'identificazione di geni differenzialmente espressi, che quindi risultano avere un valore di espressione significativamente diverso tra le due condizioni sperimentali. Si sono usati tre approcci: i) un approccio Bayesiano, con il Test dell'Empirical Bayes, ii) un approccio classico attraverso il test Significance Analysis of Microarray (SAM), e iii) un approccio basato sui ranghi usando il test del Rank Prodouct. Confrontando le liste dei geni differenzialmente espressi si sono ottenuti 127 geni comuni differenzialmente espressi a tutti i test con una significatività minore dello 0.05.

Tra i geni differenzialmente espressi che sono risultati maggiormente significativi si è osservano l'*Interleukin 15*, il *Runt- related transcription factor 3*, la *Tyrosine kinase, non- receptor, 2*. Ciò che accomuna tali geni è che sono tutte relazionate a delle cellule NK, cioè a delle cellule *killer* responsabili del mutamento e del deterioramento del DNA e quindi importanti per lo sviluppo della leucemia linfatica acuta. I recettori delle cellule *killer* sono altamente significative poiché si sta analizzando l'espressione genica di pazienti con cellule malate per il 90% (dato che la

malattia è nello stato acuto) e quindi rappresentano la causa maggiore dell'estensione della malattia. Vita l'importanza diquesti geni si è ritenuto importante studiarne il loro coinvolgimento funzionale in specifiche vie di segnale.

A questo scopo si è utilizzata l'analisi di *Gene Set* utilizzando sia test con ipotesi competitiva (Ipergeometrica), indipendene (Global Test) e mista. I gruppi (*gene set*) considerati sono quelli della Gene Ontology (che individua i geni in relazione processi biologici, funzioni molecolari e localizzazione cellulare) e dei pathway KEGG.

Per ogni categoria di Gene Ontology (GO) e dei pathway KEGG si valutata la concordanza tra i risultati ottenuti dai test rappresentativi delle tre ipotesi citate sopra. La categoria GO dei "processi biologici" identifica 18 con una significatività minore dello 0.001, la classe GO "funzioni molecolari" ne identifica 7 e la categoria GO "localizzazione cellulare" 5. Infine per il pathway KEGG si sono individuati solo 4 pathway con significatività al 5% e sono il *Viral Myocarditis*, *mTOR signaling pathway*, il *Pathological Escherichia coli infection*, il *RIG- I- like receptor signaling pathway*.

In particolare per il pathway *Viral Myocarditis* (infezione virale al muscolo cardiaco) per il pathway *Pathogenic Escherichia coli infection* (infezione gastrointestinale) e per la GO *DNA damage response* si osserva la presenza significativa del gene ABL1 che è proprio uno dei due geni coinvolti nel riarrangiamento genico studiato. Altri geni che sono risultati altamente significativi all'interno dei diversi pathway sono CASP8, FYN, CAV1 E CD80 (pathway *Viral Myocarditis*), RPS6, PIK3CA, HIF1A, EIF4EBP1, TSC1, PRKAA1, RPS6KA2, PIK3R3 (pathway *mTOR signaling pathway*), FYN (pathway *Pathogenic Escherichia coli infection* - pathway *Viral Myocarditis*), TUBA4A per il pathway *Pathogenic Escherichia coli infection* e infine CASP8 (comune con pathway *Viral Myocarditis*), CASP10, ISG15, IRF7 (pathway *RIG- I- like receptor signaling pathway*).

Con l'analisi SPIA invece si è ottenuto un solo pathway significativo: Tight Junc (coinvolto nei processi di comunicazione cellulare), osservando il pathway però in questo caso non viene rilevato il gene ABL1.

Sarà compito del biologo, a questo punto, capire l'importanza di quei geni differenzialmente espressi e delle loro relazioni all'interno dei pathway identificati.

APPENDICE

(Par. 3.4) Correlazione tra i valori di espressione ottenuti con i due metodi di normalizzazione quantile vs trasformazione logaritmica generalizzata:

```
0.9850143 0.9903801 0.9892214 0.9875529 0.9891905 0.9900700 0.9823333 0.9877288 0.9798824 0.9887212
0.9869612 0.9862689 0.9883761 0.9879949 0.9891186 0.9881354 0.9506408 0.9662937 0.9878382 0.9854884
0.9899799 0.9883659 0.9919604 0.9891545 0.9889050 0.9864618 0.9858713 0.9887387 0.9879040 0.9797311
0.9880669 0.9909438 0.9907344 0.9900394 0.9859814 0.9726029 0.9897096 0.9864734 0.9862197 0.9859798
0.9845003 0.9876468 0.9875525 0.9884913 0.9860444 0.9875214 0.9847603 0.9885203 0.9885543 0.9900593
0.9692925 0.9859567 0.9860665 0.9843435 0.9802011 0.9905983 0.9647489 0.9887955 0.9907404 0.9864021
0.9897763 0.9889247 0.9850992 0.9859475 0.9877879 0.9875060 0.9858242 0.9884341 0.9897926 0.9907704
0.9840738 0.9841484 0.9880840 0.9895907 0.9876587 0.9820316 0.9839001 0.9755887
```

(Par. 4.2) Output dei valori del Test Empirical Bayes:

```
An object of class "MArrayLM"
$coefficients
               NEG     BCRvsNEG
100_g_at  8.244519  0.023907003
1000_at   7.752126  0.003669352
1001_at   6.178494  0.033676021
1002_f_at 5.557785 -0.028689267
1003_s_at 6.659039 -0.019000172
12620 more rows ...

$rank
[1] 2

$assign
NULL

$qr
$qr
            NEG   BCRvsNEG
[1,] -8.8317609 -4.1894250
[2,]  0.1132277 -4.4100701
[3,]  0.1132277  0.1070680
[4,]  0.1132277  0.1070680
[5,]  0.1132277  0.1070680
73 more rows ...

$qraux
[1] 1.113228 1.107068

$pivot
[1] 1 2

$tol
[1] 1e-07

$rank
[1] 2

$df.residual
```

```
$Amean
 100_g_at       1000_at       1001_at   1002_f_at
1003_s_at
 8.255860     7.753866     6.194468     5.544176
6.650027
12620 more elements ...

$method
[1] "ls"

$design
     NEG BCRvsNEG
[1,]   1        1
[2,]   1        1
[3,]   1        1
[4,]   1        1
[5,]   1        1
73 more rows ...

$df.prior
[1] 2.710529

$s2.prior
[1] 0.04511302

$var.prior
[1] 354.6648258   0.8586779

$proportion
[1] 0.01

$s2.post
  100_g_at      1000_at      1001_at   1002_f_at
1003_s_at
0.05351207   0.06527171   0.04802622   0.01819709
0.03922041
12620 more elements ...

$t
```

```
[1] 76 76 76 76 76
12620 more elements ...

$sigma
 100_g_at       1000_at       1001_at   1002_f_at
1003_s_at
0.2319733    0.2568865    0.2193858    0.1312903
0.1975101
12620 more elements ...
$cov.coefficients
                NEG    BCRvsNEG
NEG       0.02439024 -0.02439024
BCRvsNEG -0.02439024  0.05141727

$stdev.unscaled
               NEG  BCRvsNEG
100_g_at  0.1561738 0.2267538
1000_at   0.1561738 0.2267538
1001_at   0.1561738 0.2267538
1002_f_at 0.1561738 0.2267538
1003_s_at 0.1561738 0.2267538
12620 more rows ...

$pivot
[1] 1 2

$genes
[1]   "100_g_at"      "1000_at"         "1001_at"
"1002_f_at" "1003_s_at"
12620 more rows ...
```

```
                NEG    BCRvsNEG
100_g_at  228.2083  0.45576895
1000_at   194.2899  0.06333917
1001_at   180.5242  0.67768371
1002_f_at 263.8109 -0.93791606
1003_s_at 215.3017 -0.42310383
12620 more rows ...

$p.value
                     NEG  BCRvsNEG
100_g_at  8.325584e-113 0.6498111
1000_at   2.577082e-107 0.9496570
1001_at   8.271982e-105 0.4999595
1002_f_at 9.356406e-118 0.3511567
1003_s_at 8.079435e-111 0.6733733
12620 more rows ...

$lods
               NEG  BCRvsNEG
100_g_at  247.7489 -5.932806
1000_at   235.4259 -6.029991
1001_at   229.7557 -5.813174
1002_f_at 258.7106 -5.614104
1003_s_at 243.3068 -5.946486
12620 more rows ...

$F
[1] 49675.06 35923.30 31160.00 65877.87 43974.42
12620 more elements ...

$F.p.value
[1]  8.718776e-123  2.985425e-117  8.007008e-115
1.314560e-127 1.052205e-120
12620 more elements ...
```

(Par. 4.2) Geni differenzialmente espresso con FDR minore di 0.05 per la Statistica Empirical Bayes.

ID	NOME DEL GENE
1036_at	"interleukin 15"
106_at	"runt-related transcription factor 3"
1107_s_at	"ISG15 ubiquitin-like modifier"
1134_at	"tyrosine kinase, non-receptor, 2"
1135_at	"G protein-coupled receptor kinase 5"
1140_at	"integrin, alpha E (antigen CD103, human mucosal lymphocyte antigen 1; alpha polypeptide)"
1211_s_at	"CASP2 and RIPK1 domain containing adaptor with death domain"
1249_at	"GRB2-associated binding protein 1"
1326_at	"caspase 10, apoptosis-related cysteine peptidase"
1361_at	"telomeric repeat binding factor (NIMA-interacting) 1"
1467_at	"epidermal growth factor receptor pathway substrate 8"
160027_s_at	"insulin-like growth factor 2 receptor"
1635_at	"c-abl oncogene 1, receptor tyrosine kinase"
1636_g_at	"c-abl oncogene 1, receptor tyrosine kinase"
1674_at	"v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1"
174_s_at	"intersectin 2"
2002_s_at	"BCL2-related protein A1"
2039_s_at	"FYN oncogene related to SRC, FGR, YES"

31506_s_at NA

31786_at "KH domain containing, RNA binding, signal transduction associated 3"

32134_at "testis derived transcript (3 LIM domains)"

32148_at "FERM, RhoGEF (ARHGEF) and pleckstrin domain protein 1 (chondrocyte-derived)"

32310_f_at "postmeiotic segregation increased 2-like 1 pseudogene"

32434_at "myristoylated alanine-rich protein kinase C substrate"

32542_at "four and a half LIM domains 1"

32562_at "endoglin"

32724_at "phytanoyl-CoA 2-hydroxylase"

32747_at "aldehyde dehydrogenase 2 family (mitochondrial)"

32961_at "DENN/MADD domain containing 4A"

32979_at "GRB2-associated binding protein 1"

33232_at "cysteine-rich protein 1 (intestinal)"

33263_at "enolase superfamily member 1"

33325_at "ribosomal protein S6 kinase, 90kDa, polypeptide 2"

33362_at "CDC42 effector protein (Rho GTPase binding) 3"

33371_s_at "RAB31, member RAS oncogene family"

33440_at "zinc finger E-box binding homeobox 1"

336_at "thromboxane A2 receptor"

33774_at "caspase 8, apoptosis-related cysteine peptidase"

33924_at "DENN/MADD domain containing 5A"

33997_at "GRB2-associated binding protein 1"

34216_at "Kruppel-like factor 7 (ubiquitous)"

34237_at "HBS1-like (S. cerevisiae)"

34376_at "protein kinase (cAMP-dependent, catalytic) inhibitor gamma"

34472_at "frizzled homolog 6 (Drosophila)"

35051_at "carbonic anhydrase VI"

35125_at "ribosomal protein S6"

35162_s_at "activin A receptor, type IIA"

35549_at "ribosomal protein L39-like"

35625_at "CD97 molecule"

35664_at "multimerin 1"

35775_at "SET and MYND domain containing 2"

35831_at "ATPase, class II, type 9A"

35912_at "mucin 4, cell surface associated"

35933_f_at "postmeiotic segregation increased 2-like 3"

35951_at "neurexin 3"

35975_at "myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila);

	translocated to, 3"
36025_at	"Rho guanine nucleotide exchange factor (GEF) 17"
36119_at	"caveolin 1, caveolae protein, 22kDa"
36142_at	"ataxin 1"
36275_at	"sema domain, transmembrane domain (TM), and cytoplasmic domain, (semaphorin) 6A"
36334_at	"lymphocyte antigen 9"
36514_at	"cell growth regulator with ring finger domain 1"
36543_at	"coagulation factor III (thromboplastin, tissue factor)"
36577_at	"fermitin family homolog 2 (Drosophila)"
36591_at	"tubulin, alpha 4a"
36617_at	"inhibitor of DNA binding 1, dominant negative helix-loop-helix protein"
36795_at	"prosaposin"
36892_at	"integrin, alpha 7"
37014_at	"myxovirus (influenza virus) resistance 1, interferon-inducible protein p78 (mouse)"
37015_at	"aldehyde dehydrogenase 1 family, member A1"
37027_at	"AHNAK nucleoprotein"
37093_at	"endonuclease, polyU-specific"
37105_at	"cathepsin G"
37351_at	"uridine phosphorylase 1"
37363_at	"metastasis suppressor 1"
37398_at	"platelet/endothelial cell adhesion molecule"
37403_at	"annexin A1"
37762_at	"epithelial membrane protein 1"
37875_at	"glycoprotein A33 (transmembrane)"
37944_at	"GTP cyclohydrolase 1"
37951_at	"deleted in liver cancer 1"
38032_at	"synaptic vesicle glycoprotein 2A"
38052_at	"coagulation factor XIII, A1 polypeptide"
38062_at	"Rap guanine nucleotide exchange factor (GEF) 5"
38085_at	NA
38091_at	"lectin, galactoside-binding, soluble, 9"
38111_at	"versican"
38112_g_at	"versican"
38119_at	"glycophorin C (Gerbich blood group)"
38168_at	"inositol polyphosphate-4-phosphatase, type II, 105kDa"
38323_at	"carboxypeptidase, vitellogenic-like"
38352_at	"peptidylprolyl isomerase H (cyclophilin H)"
38432_at	"ISG15 ubiquitin-like modifier"
38546_at	"interleukin 1 receptor accessory protein"
38578_at	"CD27 molecule"

38631_at "tumor necrosis factor, alpha-induced protein 2"

38994_at "suppressor of cytokine signaling 2"

39070_at "fascin homolog 1, actin-bundling protein (Strongylocentrotus purpuratus)"

39143_at "nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 1"

39236_s_at "fatty acid amide hydrolase"

39263_at "2'-5'-oligoadenylate synthetase 2, 69/71kDa"

39317_at "cytidine monophosphate-N-acetylneuraminic acid hydroxylase (CMP-N-acetylneuraminate monooxygenase) pseudogene"

39319_at "lymphocyte cytosolic protein 2 (SH2 domain containing leukocyte protein of 76kDa)"

39329_at "actinin, alpha 1"

39330_s_at "actinin, alpha 1"

39532_at "Ras and Rab interactor 1"

39631_at "epithelial membrane protein 2"

39689_at "cystatin C"

39730_at "c-abl oncogene 1, receptor tyrosine kinase"

39753_at "integrin, alpha 5 (fibronectin receptor, alpha polypeptide)"

39837_s_at "zinc finger protein 467"

40051_at "translocation associated membrane protein 2"

40076_at "tumor protein D52-like 2"

40132_g_at NA

40167_s_at "WD repeat and SOCS box-containing 2"

40196_at "CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase-like"

40202_at "Kruppel-like factor 9"

40479_at "FYN oncogene related to SRC, FGR, YES"

40480_s_at "FYN oncogene related to SRC, FGR, YES"

40496_at "complement component 1, s subcomponent"

40504_at "paraoxonase 2"

40516_at "aryl hydrocarbon receptor"

40702_at "interferon, gamma"

40795_at "titin"

40818_at "recombination signal binding protein for immunoglobulin kappa J region"

40855_at "sterile alpha motif domain containing 4A"

40994_at "G protein-coupled receptor kinase 5"

41015_at "protein kinase, AMP-activated, alpha 1 catalytic subunit"

41071_at "serine peptidase inhibitor, Kazal type 2 (acrosin-trypsin inhibitor)"

41123_s_at "ectonucleotide pyrophosphatase/phosphodiesterase 2"

41174_at NA

41195_at "LIM domain containing preferred translocation partner in lipoma"

41257_at "calpastatin"

41274_at NA

41343_at	"CDP-diacylglycerol synthase (phosphatidate cytidylyltransferase) 2"
41439_at	"myosin IB"
41478_at	"tetratricopeptide repeat domain 28"
41612_at	"zinc finger protein 264"
41815_at	"spectrin repeat containing, nuclear envelope 2"
479_at	"disabled homolog 2, mitogen-responsive phosphoprotein (Drosophila)"
671_at	"secreted protein, acidic, cysteine-rich (osteonectin)"
754_s_at	"breakpoint cluster region"
760_at	"dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 2"
763_at	"glia maturation factor, beta"
766_at	"lectin, galactoside-binding, soluble, 9"

(Par. 4.3) Geni differenzialmente espressi significativi sia per la Statistica Empirical Bayes che per quella Sam.

ID	NOME DEL GENE
1036_at	"interleukin 15"
106_at	"runt-related transcription factor 3"
1107_s_at	"ISG15 ubiquitin-like modifier"
1134_at	"tyrosine kinase, non-receptor, 2"
1135_at	"G protein-coupled receptor kinase 5"
1140_at	"integrin, alpha E (antigen CD103, human mucosal lymphocyte antigen 1; alpha polypeptide)"
1211_s_at	"CASP2 and RIPK1 domain containing adaptor with death domain"
1249_at	"GRB2-associated binding protein 1"
1326_at	"caspase 10, apoptosis-related cysteine peptidase"
1467_at	"epidermal growth factor receptor pathway substrate 8"
160027_s_at	"insulin-like growth factor 2 receptor"
1635_at	"c-abl oncogene 1, receptor tyrosine kinase"
1636_g_at	"c-abl oncogene 1, receptor tyrosine kinase"
1674_at	"v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1"
174_s_at	"intersectin 2"
2002_s_at	"BCL2-related protein A1"
2039_s_at	"FYN oncogene related to SRC, FGR, YES"
31506_s_at	NA
31786_at	"KH domain containing, RNA binding, signal transduction associated 3"
32134_at	"testis derived transcript (3 LIM domains)"
32148_at	"FERM, RhoGEF (ARHGEF) and pleckstrin domain protein 1 (chondrocyte-derived)"
32310_f_at	"postmeiotic segregation increased 2-like 1 pseudogene"

32434_at "myristoylated alanine-rich protein kinase C substrate"

32542_at "four and a half LIM domains 1"

32562_at "endoglin"

32724_at "phytanoyl-CoA 2-hydroxylase"

32747_at "aldehyde dehydrogenase 2 family (mitochondrial)"

32961_at "DENN/MADD domain containing 4A"

32979_at "GRB2-associated binding protein 1"

33232_at "cysteine-rich protein 1 (intestinal)"

33263_at "enolase superfamily member 1"

33325_at "ribosomal protein S6 kinase, 90kDa, polypeptide 2"

33362_at "CDC42 effector protein (Rho GTPase binding) 3"

33371_s_at "RAB31, member RAS oncogene family"

33440_at "zinc finger E-box binding homeobox 1"

336_at "thromboxane A2 receptor"

33774_at "caspase 8, apoptosis-related cysteine peptidase"

33924_at "DENN/MADD domain containing 5A"

33997_at "GRB2-associated binding protein 1"

34216_at "Kruppel-like factor 7 (ubiquitous)"

34237_at "HBS1-like (S. cerevisiae)"

34376_at "protein kinase (cAMP-dependent, catalytic) inhibitor gamma"

34472_at "frizzled homolog 6 (Drosophila)"

35051_at "carbonic anhydrase VI"

35125_at "ribosomal protein S6"

35162_s_at "activin A receptor, type IIA"

35625_at "CD97 molecule"

35664_at "multimerin 1"

35831_at "ATPase, class II, type 9A"

35912_at "mucin 4, cell surface associated"

35951_at "neurexin 3"

35975_at "myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila);
translocated to, 3"

36025_at "Rho guanine nucleotide exchange factor (GEF) 17"

36119_at "caveolin 1, caveolae protein, 22kDa"

36142_at "ataxin 1"

36275_at "sema domain, transmembrane domain (TM), and cytoplasmic domain, (semaphorin) 6A"

36514_at "cell growth regulator with ring finger domain 1"

36543_at "coagulation factor III (thromboplastin, tissue factor)"

36577_at "fermitin family homolog 2 (Drosophila)"

36591_at "tubulin, alpha 4a"

36617_at "inhibitor of DNA binding 1, dominant negative helix-loop-helix protein"

36795_at "prosaposin"

37014_at "myxovirus (influenza virus) resistance 1, interferon-inducible protein p78 (mouse)"

37015_at "aldehyde dehydrogenase 1 family, member A1"

37027_at "AHNAK nucleoprotein"

37093_at "endonuclease, polyU-specific"

37105_at "cathepsin G"

37351_at "uridine phosphorylase 1"

37363_at "metastasis suppressor 1"

37398_at "platelet/endothelial cell adhesion molecule"

37403_at "annexin A1"

37762_at "epithelial membrane protein 1"

37875_at "glycoprotein A33 (transmembrane)"

37944_at "GTP cyclohydrolase 1"

37951_at "deleted in liver cancer 1"

38032_at "synaptic vesicle glycoprotein 2A"

38052_at "coagulation factor XIII, A1 polypeptide"

38062_at "Rap guanine nucleotide exchange factor (GEF) 5"

38085_at NA

38091_at "lectin, galactoside-binding, soluble, 9"

38111_at "versican"

38112_g_at "versican"

38119_at "glycophorin C (Gerbich blood group)"

38168_at "inositol polyphosphate-4-phosphatase, type II, 105kDa"

38323_at "carboxypeptidase, vitellogenic-like"

38546_at "interleukin 1 receptor accessory protein"

38578_at "CD27 molecule"

38631_at "tumor necrosis factor, alpha-induced protein 2"

38994_at "suppressor of cytokine signaling 2"

39070_at "fascin homolog 1, actin-bundling protein (Strongylocentrotus purpuratus)"

39143_at "nuclear factor of activated T-cells, cytoplasmic, calcineurin dependent 1"

39317_at "cytidine monophosphate-N-acetylneuraminic acid hydroxylase (CMP-N-acetylneuraminate monooxygenase) pseudogene"

39319_at "lymphocyte cytosolic protein 2 (SH2 domain containing leukocyte protein of 76kDa)"

39329_at "actinin, alpha 1"

39330_s_at "actinin, alpha 1"

39631_at "epithelial membrane protein 2"

39689_at "cystatin C"

39730_at "c-abl oncogene 1, receptor tyrosine kinase"

39753_at "integrin, alpha 5 (fibronectin receptor, alpha polypeptide)"

39837_s_at "zinc finger protein 467"

40051_at	"translocation associated membrane protein 2"
40076_at	"tumor protein D52-like 2"
40132_g_at	NA
40167_s_at	"WD repeat and SOCS box-containing 2"
40196_at	"CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase-like"
40202_at	"Kruppel-like factor 9"
40479_at	"FYN oncogene related to SRC, FGR, YES"
40480_s_at	"FYN oncogene related to SRC, FGR, YES"
40504_at	"paraoxonase 2"
40516_at	"aryl hydrocarbon receptor"
40702_at	"interferon, gamma"
40795_at	"titin"
40818_at	"recombination signal binding protein for immunoglobulin kappa J region"
40855_at	"sterile alpha motif domain containing 4A"
40994_at	"G protein-coupled receptor kinase 5"
41015_at	"protein kinase, AMP-activated, alpha 1 catalytic subunit"
41071_at	"serine peptidase inhibitor, Kazal type 2 (acrosin-trypsin inhibitor)"
41123_s_at	"ectonucleotide pyrophosphatase/phosphodiesterase 2"
41174_at	NA
41195_at	"LIM domain containing preferred translocation partner in lipoma"
41257_at	"calpastatin"
41274_at	NA
41343_at	"CDP-diacylglycerol synthase (phosphatidate cytidylyltransferase) 2"
41439_at	"myosin IB"
41478_at	"tetratricopeptide repeat domain 28"
41612_at	"zinc finger protein 264"
41815_at	"spectrin repeat containing, nuclear envelope 2"
479_at	"disabled homolog 2, mitogen-responsive phosphoprotein (Drosophila)"
671_at	"secreted protein, acidic, cysteine-rich (osteonectin)"
760_at	"dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 2"
763_at	"glia maturation factor, beta"
766_at	"lectin, galactoside-binding, soluble, 9"

(Par. 4.4) Geni differenzialmente espressi significativi sia per la Statistica Empirical Bayes, sia per la Statistica Sam e sia per quella Rank Prodouct.

ID	NOME DEL GENE
1036_at	"interleukin 15"
106_at	"runt-related transcription factor 3"
1134_at	"tyrosine kinase, non-receptor, 2"

1135_at	"G protein-coupled receptor kinase 5"
1211_s_at	"CASP2 and RIPK1 domain containing adaptor with death domain"
1249_at	"GRB2-associated binding protein 1"
1326_at	"caspase 10, apoptosis-related cysteine peptidase"
1467_at	"epidermal growth factor receptor pathway substrate 8"
160027_s_at	"insulin-like growth factor 2 receptor"
1635_at	"c-abl oncogene 1, receptor tyrosine kinase"
1636_g_at	"c-abl oncogene 1, receptor tyrosine kinase"
1674_at	"v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1"
174_s_at	"intersectin 2"
2002_s_at	"BCL2-related protein A1"
2039_s_at	"FYN oncogene related to SRC, FGR, YES"
31506_s_at	NA
31786_at	"KH domain containing, RNA binding, signal transduction associated 3"
32134_at	"testis derived transcript (3 LIM domains)"
32148_at	"FERM, RhoGEF (ARHGEF) and pleckstrin domain protein 1 (chondrocyte-derived)"
32434_at	"myristoylated alanine-rich protein kinase C substrate"
32542_at	"four and a half LIM domains 1"
32562_at	"endoglin"
32747_at	"aldehyde dehydrogenase 2 family (mitochondrial)"
32961_at	"DENN/MADD domain containing 4A"
32979_at	"GRB2-associated binding protein 1"
33232_at	"cysteine-rich protein 1 (intestinal)"
33263_at	"enolase superfamily member 1"
33325_at	"ribosomal protein S6 kinase, 90kDa, polypeptide 2"
33362_at	"CDC42 effector protein (Rho GTPase binding) 3"
33371_s_at	"RAB31, member RAS oncogene family"
33440_at	"zinc finger E-box binding homeobox 1"
336_at	"thromboxane A2 receptor"
33774_at	"caspase 8, apoptosis-related cysteine peptidase"
33924_at	"DENN/MADD domain containing 5A"
33997_at	"GRB2-associated binding protein 1"
34216_at	"Kruppel-like factor 7 (ubiquitous)"
34237_at	"HBS1-like (S. cerevisiae)"
34472_at	"frizzled homolog 6 (Drosophila)"
35051_at	"carbonic anhydrase VI"
35162_s_at	"activin A receptor, type IIA"
35625_at	"CD97 molecule"
35664_at	"multimerin 1"
35912_at	"mucin 4, cell surface associated"

35951_at "neurexin 3"

35975_at "myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 3"

36025_at "Rho guanine nucleotide exchange factor (GEF) 17"

36119_at "caveolin 1, caveolae protein, 22kDa"

36142_at "ataxin 1"

36275_at "sema domain, transmembrane domain (TM), and cytoplasmic domain, (semaphorin) 6A"

36514_at "cell growth regulator with ring finger domain 1"

36543_at "coagulation factor III (thromboplastin, tissue factor)"

36577_at "fermitin family homolog 2 (Drosophila)"

36591_at "tubulin, alpha 4a"

36617_at "inhibitor of DNA binding 1, dominant negative helix-loop-helix protein"

36795_at "prosaposin"

37015_at "aldehyde dehydrogenase 1 family, member A1"

37027_at "AHNAK nucleoprotein"

37105_at "cathepsin G"

37351_at "uridine phosphorylase 1"

37363_at "metastasis suppressor 1"

37398_at "platelet/endothelial cell adhesion molecule"

37403_at "annexin A1"

37762_at "epithelial membrane protein 1"

37875_at "glycoprotein A33 (transmembrane)"

37944_at "GTP cyclohydrolase 1"

37951_at "deleted in liver cancer 1"

38032_at "synaptic vesicle glycoprotein 2A"

38052_at "coagulation factor XIII, A1 polypeptide"

38062_at "Rap guanine nucleotide exchange factor (GEF) 5"

38111_at "versican"

38112_g_at "versican"

38119_at "glycophorin C (Gerbich blood group)"

38323_at "carboxypeptidase, vitellogenic-like"

38546_at "interleukin 1 receptor accessory protein"

38578_at "CD27 molecule"

38631_at "tumor necrosis factor, alpha-induced protein 2"

38994_at "suppressor of cytokine signaling 2"

39070_at "fascin homolog 1, actin-bundling protein (Strongylocentrotus purpuratus)"

39317_at "cytidine monophosphate-N-acetylneuraminic acid hydroxylase (CMP-N-acetylneuraminate monooxygenase) pseudogene"

39319_at "lymphocyte cytosolic protein 2 (SH2 domain containing leukocyte protein of 76kDa)"

39329_at "actinin, alpha 1"

39330_s_at "actinin, alpha 1"

39631_at "epithelial membrane protein 2"

39689_at "cystatin C"

39730_at "c-abl oncogene 1, receptor tyrosine kinase"

39753_at "integrin, alpha 5 (fibronectin receptor, alpha polypeptide)"

39837_s_at "zinc finger protein 467"

40051_at "translocation associated membrane protein 2"

40167_s_at "WD repeat and SOCS box-containing 2"

40196_at "CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase-like"

40202_at "Kruppel-like factor 9"

40479_at "FYN oncogene related to SRC, FGR, YES"

40480_s_at "FYN oncogene related to SRC, FGR, YES"

40504_at "paraoxonase 2"

40516_at "aryl hydrocarbon receptor"

40702_at "interferon, gamma"

40795_at "titin"

40818_at "recombination signal binding protein for immunoglobulin kappa J region"

40855_at "sterile alpha motif domain containing 4A"

40994_at "G protein-coupled receptor kinase 5"

41015_at "protein kinase, AMP-activated, alpha 1 catalytic subunit"

41123_s_at "ectonucleotide pyrophosphatase/phosphodiesterase 2"

41174_at NA

41257_at "calpastatin"

41274_at NA

41343_at "CDP-diacylglycerol synthase (phosphatidate cytidylyltransferase) 2"

41439_at "myosin IB"

41815_at "spectrin repeat containing, nuclear envelope 2"

479_at "disabled homolog 2, mitogen-responsive phosphoprotein (Drosophila)"

671_at "secreted protein, acidic, cysteine-rich (osteonectin)"

760_at "dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 2"

763_at "glia maturation factor, beta"

1107_s_at "ISG15 ubiquitin-like modifier"

1140_at "integrin, alpha E (antigen CD103, human mucosal lymphocyte antigen 1; alpha polypeptide)"

32310_f_at "postmeiotic segregation increased 2-like 1 pseudogene"

32724_at "phytanoyl-CoA 2-hydroxylase"

34376_at "protein kinase (cAMP-dependent, catalytic) inhibitor gamma"

35125_at "ribosomal protein S6"

35831_at "ATPase, class II, type 9A"

37014_at "myxovirus (influenza virus) resistance 1, interferon-inducible protein p78 (mouse)"

```
38085_at    NA
38091_at    "lectin, galactoside-binding, soluble, 9"
40076_at    "tumor protein D52-like 2"
40132_g_at  NA
41071_at    "serine peptidase inhibitor, Kazal type 2 (acrosin-trypsin inhibitor)"
41478_at    "tetratricopeptide repeat domain 28"
766_at      "lectin, galactoside-binding, soluble, 9"
```

(Par. 5.2) Gene set analysis con la Ipergeometrica:

2) `summary(Biological Process.iper)`

	GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1	GO:0023052	1.979842e-06	2.564548	32.9833699	56	2484	signaling
2	GO:0007165	3.883118e-05	2.324002	24.5250741	43	1847	signal transduction
3	GO:0042981	6.857773e-05	2.946796	8.4051824	21	633	regulation of apoptosis
4	GO:0043067	7.685534e-05	2.920840	8.4715741	21	638	regulation of programmed cell death
5	GO:0010941	8.224049e-05	2.905468	8.5114090	21	641	regulation of cell death
6	GO:0023060	1.281614e-04	2.152691	28.3226763	46	2133	signal transmission
7	GO:0023046	1.297218e-04	2.151274	28.3359546	46	2134	signaling process
8	GO:0009966	3.344443e-04	2.723148	8.0466675	19	606	regulation of signal transduction
9	GO:0023051	3.344443e-04	2.723148	8.0466675	19	606	regulation of signaling process
10	GO:0043065	4.708619e-04	3.278397	4.4615186	13	336	positive regulation of apoptosis
11	GO:0043068	4.843888e-04	3.267833	4.4747970	13	337	positive regulation of programmed cell death
12	GO:0010942	4.982451e-04	3.257333	4.4880753	13	338	positive regulation of cell death
13	GO:0006917	5.184087e-04	3.646493	3.3726956	11	254	induction of apoptosis
14	GO:0012502	5.358495e-04	3.631058	3.3859740	11	255	induction of programmed cell death
15	GO:0016265	5.495640e-04	2.396067	11.1936316	23	843	death
16	GO:0035023	7.693176e-04	7.759184	0.7303081	5	55	regulation of Rho protein signal transduction
17	GO:0009653	7.912655e-04	2.288472	12.2293412	24	921	anatomical structure morphogenesis
18	GO:0006915	8.522320e-04	2.396366	10.0915302	21	760	apoptosis
19	GO:0010562	8.871118e-04	5.931097	1.1286580	6	85	positive regulation of phosphorus metabolic process
20	GO:0045937	8.871118e-04	5.931097	1.1286580	6	85	positive regulation of phosphate metabolic process
21	GO:0012501	9.437242e-04	2.375004	10.1712002	21	766	programmed cell death

(Par. 5.3) Gene sets analysis con il Global Test, con un p-value minore di 0.05 utilizzando i pathway KEGG.

```
Model: linear regression.
Degrees of freedom: 78 total; 1 null; 1 + 12625 alternative.
Tested: 199 subsets.
```

Null distibution: asymptotic.

	BH	alias	p-value	Statistic	Expected	Std.dev	#Cov
04360	2.728629e-05	Axon guidance	1.371170e-07	7.915537	1.298701	0.5335180	166
05416	9.448630e-05	Viral myocarditis	1.485501e-06	7.169410	1.298701	0.6092224	97
04012	9.448630e-05	ErbB signaling pathway	1.807731e-06	5.685781	1.298701	0.4801277	170
05130	9.448630e-05	Pathogenic Escherichia coli infection	1.899222e-06	7.477967	1.298701	0.6220515	81
04520	1.375941e-03	Adherens junction	3.804007e-05	4.976540	1.298701	0.5186115	135
05217	1.375941e-03	Basal cell carcinoma	4.623152e-05	7.542982	1.298701	0.7277097	52
04610	1.375941e-03	Complement and coagulation cascades	4.839992e-05	4.947601	1.298701	0.5288759	73
05200	1.475903e-03	Pathways in cancer	5.933278e-05	4.303310	1.298701	0.4315098	567
05220	2.151233e-03	Chronic myeloid leukemia	9.729193e-05	5.324576	1.298701	0.5601334	166
04722	2.183191e-03	Neurotrophin signaling pathway	1.097081e-04	5.069843	1.298701	0.5301613	216
04210	4.065086e-03	Apoptosis	2.247032e-04	4.908909	1.298701	0.5373646	143
01040	4.360609e-03	Biosynthesis of unsaturated fatty acids	2.728141e-04	5.611244	1.298701	0.7088027	23
00903	4.360609e-03	Limonene and pinene degradation	2.848639e-04	6.514219	1.298701	0.8383263	10
00565	5.223786e-03	Ether lipid metabolism	3.675026e-04	6.006009	1.298701	0.7610400	24
04916	5.766991e-03	Melanogenesis	4.346978e-04	4.374472	1.298701	0.5209474	124
05020	6.010269e-03	Prion diseases	4.879956e-04	4.452273	1.298701	0.5739600	52
04060	6.010269e-03	Cytokine-cytokine receptor interaction	5.215691e-04	3.611691	1.298701	0.4173067	317
00340	6.010269e-03	Histidine metabolism	5.436424e-04	5.432849	1.298701	0.6973064	22
04630	6.163705e-03	Jak-STAT signaling pathway	5.884945e-04	3.714676	1.298701	0.4522980	206
04110	6.737675e-03	Cell cycle	6.771532e-04	4.419555	1.298701	0.5444360	196
04350	7.929605e-03	TGF-beta signaling pathway	8.526395e-04	3.972004	1.298701	0.5262389	137
00760	7.929605e-03	Nicotinate and nicotinamide metabolism	9.319470e-04	4.904290	1.298701	0.7135605	15
04622	7.929605e-03	RIG-I-like receptor signaling pathway	9.979135e-04	5.275734	1.298701	0.7199677	85
04530	7.929605e-03	Tight junction	1.026145e-03	4.212268	1.298701	0.5611644	161
04670	7.929605e-03	Leukocyte transendothelial migration	1.035563e-03	3.984451	1.298701	0.5354136	159
05216	7.929605e-03	Thyroid cancer	1.036029e-03	4.277702	1.298701	0.6022834	49
00053	8.128798e-03	Ascorbate and aldarate metabolism	1.102902e-03	5.496960	1.298701	0.7860646	14
04115	1.018714e-02	p53 signaling pathway	1.433366e-03	3.753849	1.298701	0.5177111	98
05210	1.057380e-02	Colorectal cancer	1.586831e-03	3.633535	1.298701	0.5137336	156
04620	1.057380e-02	Toll-like receptor signaling pathway	1.647016e-03	4.381235	1.298701	0.5956803	142
05412	1.057380e-02	Arrhythmogenic right ventricular cardiomyo...	1.647175e-03	4.108260	1.298701	0.6060178	107
04930	1.122926e-02	Type II diabetes mellitus	1.805710e-03	3.845556	1.298701	0.5560239	73
05215	1.275738e-02	Prostate cancer	2.174257e-03	3.611907	1.298701	0.5139256	169
04510	1.275738e-02	Focal adhesion	2.179654e-03	3.409271	1.298701	0.4797660	325
05221	1.331710e-02	Acute myeloid leukemia	2.342204e-03	3.617287	1.298701	0.5461509	107
04540	1.811229e-02	Gap junction	3.276596e-03	3.385698	1.298701	0.5231728	137
04621	1.979643e-02	NOD-like receptor signaling pathway	3.680743e-03	4.702299	1.298701	0.7254227	75
00790	1.996072e-02	Folate biosynthesis	3.846581e-03	4.981679	1.298701	0.8807962	12
04810	1.996072e-02	Regulation of actin cytoskeleton	3.911900e-03	3.214852	1.298701	0.4900351	284

04514	2.229575e-02	Cell adhesion molecules (CAMs)	4.564243e-03	3.216965	1.298701	0.5163962	151
04623	2.229575e-02	Cytosolic DNA-sensing pathway	4.703623e-03	4.422388	1.298701	0.7240072	51
04660	2.229575e-02	T cell receptor signaling pathway	4.705636e-03	3.404046	1.298701	0.5362321	181
04310	2.614105e-02	Wnt signaling pathway	5.648568e-03	3.120572	1.298701	0.4971752	192
00830	3.215352e-02	Retinol metabolism	7.358000e-03	3.476968	1.298701	0.6310000	56
00310	3.215352e-02	Lysine degradation	7.493610e-03	3.709952	1.298701	0.6684021	30
00640	3.215352e-02	Propanoate metabolism	7.500239e-03	4.424334	1.298701	0.7976945	29
04666	3.215352e-02	Fc gamma R-mediated phagocytosis	7.594047e-03	3.000191	1.298701	0.5081883	126
04150	3.304579e-02	mTOR signaling pathway	7.970844e-03	3.193506	1.298701	0.5565874	67
05213	3.485103e-02	Endometrial cancer	8.581409e-03	3.135625	1.298701	0.5599290	95
04020	3.718059e-02	Calcium signaling pathway	9.447877e-03	2.827299	1.298701	0.4730681	246
05222	3.718059e-02	Small cell lung cancer	9.692040e-03	3.011286	1.298701	0.5367610	153
05340	3.718059e-02	Primary immunodeficiency	9.715530e-03	4.026567	1.298701	0.7625549	47
04062	3.740345e-02	Chemokine signaling pathway	9.961723e-03	2.863315	1.298701	0.4782231	259
04910	3.936998e-02	Insulin signaling pathway	1.105455e-02	2.919432	1.298701	0.5092205	187
04740	3.936998e-02	Olfactory transduction	1.108895e-02	3.268584	1.298701	0.6147569	54
04080	3.936998e-02	Neuroactive ligand-receptor interaction	1.122260e-02	2.811456	1.298701	0.4874869	258
00270	3.936998e-02	Cysteine and methionine metabolism	1.127683e-02	3.818859	1.298701	0.7655558	32
04614	3.972885e-02	Renin-angiotensin system	1.170544e-02	5.140048	1.298701	1.0762062	17
00561	3.972885e-02	Glycerolipid metabolism	1.180065e-02	3.177194	1.298701	0.6105203	33
00564	3.972885e-02	Glycerophospholipid metabolism	1.210542e-02	3.241409	1.298701	0.6265016	48
04010	3.972885e-02	MAPK signaling pathway	1.217819e-02	2.666686	1.298701	0.4529479	428
00380	4.003115e-02	Tryptophan metabolism	1.247202e-02	3.429916	1.298701	0.6623219	42
04142	4.113175e-02	Lysosome	1.302161e-02	2.783841	1.298701	0.4862148	138
04664	4.220468e-02	Fc epsilon RI signaling pathway	1.357337e-02	2.949024	1.298701	0.5542979	120
05010	4.243783e-02	Alzheimer's disease	1.386160e-02	3.385059	1.298701	0.6508612	186
04650	4.563421e-02	Natural killer cell mediated cytotoxicity	1.513496e-02	2.766878	1.298701	0.5084399	171
04330	4.729865e-02	Notch signaling pathway	1.595853e-02	3.203134	1.298701	0.6390979	43
05110	4.729865e-02	Vibrio cholerae infection	1.616235e-02	3.243442	1.298701	0.6613751	72
05218	4.780318e-02	Melanoma	1.675762e-02	2.635666	1.298701	0.4908650	125
04140	4.780318e-02	Regulation of autophagy	1.708980e-02	3.193281	1.298701	0.6655277	31
00410	4.780318e-02	beta-Alanine metabolism	1.723654e-02	3.842282	1.298701	0.8475061	23
04070	4.780318e-02	Phosphatidylinositol signaling system	1.731211e-02	3.054383	1.298701	0.5939271	105
05014	4.780318e-02	Amyotrophic lateral sclerosis (ALS)	1.753584e-02	2.870028	1.298701	0.5642031	81
04742	5.003155e-02	Taste transduction	1.860470e-02	3.560030	1.298701	0.8015780	29
04914	5.028376e-02	Progesterone-mediated oocyte maturation	1.902508e-02	2.579535	1.298701	0.4903930	120
04920	5.028376e-02	Adipocytokine signaling pathway	1.920385e-02	2.911475	1.298701	0.5943105	93
04144	5.044189e-02	Endocytosis	1.951772e-02	2.487936	1.298701	0.4535281	223
04662	5.150457e-02	B cell receptor signaling pathway	2.039791e-02	2.715098	1.298701	0.5430298	121
05120	5.150457e-02	Epithelial cell signaling in Helicobacter ...	2.044654e-02	3.019625	1.298701	0.6244274	100
00650	5.256528e-02	Butanoate metabolism	2.113177e-02	3.298490	1.298701	0.7257239	34
00620	5.355408e-02	Pyruvate metabolism	2.188397e-02	3.397259	1.298701	0.7533553	39
04270	5.355408e-02	Vascular smooth muscle contraction	2.228750e-02	2.630227	1.298701	0.5167222	149

00052	5.355408e-02	Galactose metabolism	2.233662e-02	3.121218	1.298701	0.6905298	24
05414	5.377842e-02	Dilated cardiomyopathy	2.270044e-02	2.839368	1.298701	0.5973014	133
00562	5.379371e-02	Inositol phosphate metabolism	2.297721e-02	2.783063	1.298701	0.5694473	67
00983	5.541174e-02	Drug metabolism - other enzymes	2.394678e-02	2.887326	1.298701	0.6280446	48
05410	5.601643e-02	Hypertrophic cardiomyopathy (HCM)	2.448960e-02	2.918363	1.298701	0.6330844	120
04512	6.259559e-02	ECM-receptor interaction	2.780218e-02	2.785198	1.298701	0.6150926	123
00280	6.259559e-02	Valine, leucine and isoleucine degradation	2.829777e-02	3.039793	1.298701	0.6981065	45
00531	6.259559e-02	Glycosaminoglycan degradation	2.830956e-02	2.689314	1.298701	0.5996906	27
05322	6.386370e-02	Systemic lupus erythematosus	2.920400e-02	2.822666	1.298701	0.6382542	98
04640	6.734331e-02	Hematopoietic cell lineage	3.113359e-02	2.371281	1.298701	0.4832042	129
01100	6.919115e-02	Metabolic pathways	3.233556e-02	2.524139	1.298701	0.5273562	957
04720	7.255232e-02	Long-term potentiation	3.456753e-02	2.552371	1.298701	0.5649266	116
00240	7.255232e-02	Pyrimidine metabolism	3.463553e-02	2.968067	1.298701	0.7268973	83
00980	7.559613e-02	Metabolism of xenobiotics by cytochrome P450	3.646848e-02	2.777795	1.298701	0.6705273	65
00051	7.565343e-02	Fructose and mannose metabolism	3.698976e-02	2.729295	1.298701	0.6646730	34
00230	7.565343e-02	Purine metabolism	3.725646e-02	2.636422	1.298701	0.6075698	147
00520	7.681173e-02	Amino sugar and nucleotide sugar metabolism	3.855930e-02	2.543863	1.298701	0.5913781	32
00910	7.681173e-02	Nitrogen metabolism	3.866776e-02	2.911127	1.298701	0.7567629	23
04146	7.681173e-02	Peroxisome	3.917279e-02	2.656526	1.298701	0.6387440	63
00750	7.681173e-02	Vitamin B6 metabolism	3.937084e-02	5.017841	1.298701	1.6353466	2
00532	7.829222e-02	Chondroitin sulfate biosynthesis	4.052311e-02	3.633126	1.298701	1.0751503	9
04130	7.994312e-02	SNARE interactions in vesicular transport	4.199371e-02	2.511659	1.298701	0.5941870	47
00330	7.994312e-02	Arginine and proline metabolism	4.218104e-02	2.579584	1.298701	0.6275089	63
05214	8.101432e-02	Glioma	4.315336e-02	2.355995	1.298701	0.5255099	125
05212	8.153142e-02	Pancreatic cancer	4.391412e-02	2.396788	1.298701	0.5486354	154
03430	8.153142e-02	Mismatch repair	4.424821e-02	3.446930	1.298701	1.0185226	32
05016	8.779151e-02	Huntington's disease	4.808680e-02	2.809005	1.298701	0.7606524	178
40	6783e-02	Pentose and glucuronate interconversions	4.995207e-02	2.963644	1.298701	0.8671005	15

(Par. 5.4.1) Matrice dei geni utilizzata per l'analisi SPIA:

	ID	logFC	ENTREZ		ID	logFC	ENTREZ
1	1036_at	0.8512077	5595	65	37944_at	0.7102577	3594
2	106_at	0.6260991	7075	66	37951_at	0.7885038	2323
3	1134_at	0.7958113	643	68	38052_at	0.5981327	5743
5	1211_s_at	0.5940707	4319	69	38062_at	0.7475573	864
6	1249_at	0.7071947	780	70	38111_at	0.6424862	2624
7	1326_at	0.7576054	5610	71	38112_g_at	0.7415832	6917
8	1467_at	0.7095067	3094	72	38119_at	0.6571295	5861
9	160027_s_at	0.8100172	5875	73	38323_at	0.7290780	3449
10	1635_at	0.5025247	5600	74	38546_at	0.6890169	3552
11	1636_g_at	0.5034060	7531	75	38578_at	0.5907391	5896
12	1674_at	0.5173329	8850	76	38631_at	0.7102066	5618
13	174_s_at	0.6377471	4090	78	39070_at	0.6609000	282808
14	2002_s_at	0.8215681	5428	79	39317_at	0.6065795	1544
15	2039_s_at	0.6756128	3984	80	39319_at	0.7416757	4953
16	31506_s_at	0.6710266	3598	81	39329_at	0.5923162	5335
17	31786_at	0.6718678	2956	82	39330_s_at	0.7270155	4582
18	32134_at	0.6300205	7480	83	39631_at	0.7919757	27
20	32434_at	0.4720305	10519	84	39689_at	0.8731327	5336
21	32542_at	0.5624156	3458	85	39730_at	0.4993431	1439
22	32562_at	0.7121400	3448	86	39753_at	0.7986608	2057
23	32747_at	0.7116706	1543	87	39837_s_at	0.7694440	627

25	32979_at	0.7375130	1302
27	33263_at	0.8024319	7156
28	33325_at	0.7907357	7294
29	33362_at	0.6229161	10114
30	33371_s_at	0.8460801	7150
31	33440_at	0.5753555	6732
32	336_at	0.8250821	3579
34	33924_at	0.7405835	7078
36	34216_at	0.6899221	3600
37	34237_at	0.7077830	574
38	34472_at	0.5639521	3091
39	35051_at	0.7337988	7060
40	35162_s_at	0.7056291	1946
41	35625_at	0.6293307	5918
42	35664_at	0.7997662	5893
43	35912_at	0.6561417	5970
44	35951_at	0.8022208	24138
46	36025_at	0.8363889	6258
48	36142_at	0.8356680	5463
49	36275_at	0.4768312	2315
51	36543_at	0.6475285	1052
52	36577_at	0.7111589	5982
53	36591_at	0.5275066	5984
55	36795_at	0.6982510	3603
56	37015_at	0.6555511	10810
57	37027_at	0.4630405	4916
58	37105_at	0.7726774	9970
60	37363_at	0.5925021	3587
62	37403_at	0.5793718	7301
63	37762_at	0.6655499	5756
64	37875_at	0.7918727	2322
89	40167_s_at	0.7168634	5575
90	40196_at	0.7471397	1907
91	40202_at	0.3449542	5519
93	40480_s_at	0.6178932	3082
94	40504_at	0.5779398	930
95	40516_at	0.5911423	1236
96	40702_at	0.8237764	7066
98	40818_at	0.6415812	10244
99	40855_at	0.8348620	3654
100	40994_at	0.8363302	322
101	41015_at	0.7885584	2908
103	41174_at	0.7456673	51561
105	41274_at	0.8040078	9636
106	41343_at	0.7560451	2041
107	41439_at	0.6806192	5154
109	479_at	0.7809390	4684
111	760_at	0.7694692	650
112	763_at	0.8000679	652
113	1107_s_at	1.6049689	1557
114	1140_at	1.3523041	1843
115	32310_f_at	1.2216389	8798
116	32724_at	1.5175235	2056
117	34376_at	1.4519868	3459
118	35125_at	1.3124514	10152
119	35831_at	1.5019471	1875
120	37014_at	2.0177215	1382
121	38085_at	1.3485153	2959
125	41071_at	1.3978860	283
126	41478_at	1.2560537	1464
127	766_at	1.3019332	5196

(Par. 5.4.1) Analisi SPIA:

	Name	ID	pSize	NDE	tA	pNDE	pPERT	pG	pGFdr	pGFWER	Status
1	Tight junc	04530	1	1	0.61197633	1	0.025	0.1172220	1	1	Activated
2	Long-term	04730	2	2	0.82854563	1	0.069	0.2534818	1	1	Activated
3	Cell cycle	04110	2	2	2.36942130	1	0.086	0.2969931	1	1	Activated
4	Epithelial	05120	5	5	-0.42798044	1	0.103	0.3371217	1	1	Inhibited
5	Notch sign	04330	1	1	-0.41936520	1	0.172	0.4747649	1	1	Inhibited
6	Regulation	04810	4	4	-1.35697600	1	0.220	0.5531081	1	1	Inhibited
7	Fc epsilon	04664	4	4	0.13953710	1	0.245	0.5895918	1	1	Activated
8	Prion dise	05020	3	3	0.42014640	1	0.251	0.5979579	1	1	Activated
9	Vascular s	04270	1	1	0.13953710	1	0.254	0.6020869	1	1	Activated
10	GnRH signa	04912	2	2	0.13953710	1	0.255	0.6034554	1	1	Activated
11	Progestero	04914	2	2	0.26868860	1	0.255	0.6034554	1	1	Activated
12	Dorso-vent	04320	1	1	-0.13906770	1	0.263	0.6142631	1	1	Inhibited
13	mTOR signa	04150	2	2	-0.25232321	1	0.265	0.6169267	1	1	Inhibited
14	ErbB signa	04012	4	4	0.13906770	1	0.271	0.6248275	1	1	Activated
15	Endometria	05213	1	1	0.13953710	1	0.272	0.6261313	1	1	Activated
16	MAPK signa	04010	7	7	-0.52376688	1	0.276	0.6313098	1	1	Inhibited
17	Chemokine	04062	6	6	1.88225263	1	0.279	0.6351556	1	1	Activated
18	Bladder ca	05219	1	1	0.13434430	1	0.281	0.6377016	1	1	Activated
19	Long-term	04720	1	1	0.26868860	1	0.284	0.6414938	1	1	Activated
20	PPAR signa	03320	1	1	0.10937340	1	0.323	0.6880233	1	1	Activated
21	Neurotroph	04722	9	9	-1.34514846	1	0.341	0.7078726	1	1	Inhibited
22	Renal cell	05211	3	3	-0.44108282	1	0.377	0.7447673	1	1	Inhibited

23	TGF-beta s	04350	8	8	-0.94101955	1	0.428	0.7912145	1	1 Inhibited
24	Oocyte mei	04114	3	3	0.60760967	1	0.449	0.8085288	1	1 Activated
25	Basal cell	05217	3	3	-0.29098500	1	0.484	0.8352245	1	1 Inhibited
26	Focal adhe	04510	5	5	-0.60924994	1	0.488	0.8381107	1	1 Inhibited
27	Insulin si	04910	2	2	0.20541232	1	0.507	0.8513768	1	1 Activated
28	VEGF signa	04370	5	5	-0.35466072	1	0.515	0.8567480	1	1 Inhibited
29	Wnt signal	04310	2	2	-0.08165010	1	0.557	0.8829509	1	1 Inhibited
30	Melanoma	05218	3	3	-0.50471785	1	0.609	0.9110256	1	1 Inhibited
31	Acute myel	05221	3	3	0.04349056	1	0.609	0.9110256	1	1 Activated
32	Small cell	05222	3	3	-0.06030586	1	0.610	0.9115208	1	1 Inhibited
33	T cell rec	04660	5	5	-0.09771695	1	0.615	0.9139718	1	1 Inhibited
34	Pancreatic	05212	2	2	-0.07087380	1	0.617	0.9149408	1	1 Inhibited
35	Viral myoc	05416	1	1	0.06496020	1	0.618	0.9154229	1	1 Activated
36	Gap juncti	04540	2	2	-0.32257205	1	0.637	0.9242778	1	1 Inhibited
37	Adipocytok	04920	2	2	0.02403664	1	0.671	0.9387197	1	1 Activated
38	Leukocyte	04670	4	4	0.17236439	1	0.742	0.9634173	1	1 Activated
39	Melanogene	04916	2	2	-0.24095759	1	0.754	0.9669016	1	1 Inhibited
40	B cell rec	04662	4	4	-0.06066878	1	0.775	0.9725415	1	1 Inhibited
41	Glioma	05214	4	4	-0.11955879	1	0.789	0.9759843	1	1 Inhibited
42	Natural ki	04650	7	7	-0.28488395	1	0.798	0.9780661	1	1 Inhibited
43	Non-small	05223	4	4	0.25304667	1	0.809	0.9804727	1	1 Activated
44	Toll-like	04620	6	6	-0.08600460	1	0.893	0.9940596	1	1 Inhibited
45	ECM-recept	04512	2	2	0.01734980	1	0.939	0.9981005	1	1 Activated
46	Fc gamma R	04666	5	5	0.04628940	1	0.942	0.9982845	1	1 Activated
47	Cytokine-c	04060	22	22	0.04134600	1	0.967	0.9994494	1	1 Activated
48	Apoptosis	04210	5	5	-0.01916431	1	0.968	0.9994824	1	1 Inhibited
49	Antigen pr	04612	2	2	-0.00533760	1	0.975	0.9996849	1	1 Inhibited
50	Axon guida	04360	4	4	0.01088479	1	0.977	0.9997334	1	1 Activated
51	Calcium si	04020	2	2	0.00000000	1	1.000	1.0000000	1	1 Inhibited
52	Neuroactiv	04080	2	2	0.00000000	1	1.000	1.0000000	1	1 Inhibited
53	Regulation	04140	3	3	0.00000000	1	1.000	1.0000000	1	1 Inhibited
54	Hedgehog s	04340	3	3	0.00000000	1	1.000	1.0000000	1	1 Inhibited
55	NOD-like r	04621	3	3	0.00000000	1	1.000	1.0000000	1	1 Inhibited
56	RIG-I-like	04622	5	5	0.00000000	1	1.000	1.0000000	1	1 Inhibited
57	Cytosolic	04623	3	3	0.00000000	1	1.000	1.0000000	1	1 Inhibited
58	Jak-STAT s	04630	13	13	0.00000000	1	1.000	1.0000000	1	1 Inhibited
59	Intestinal	04672	1	1	0.00000000	1	1.000	1.0000000	1	1 Inhibited
60	Type II di	04930	1	1	0.00000000	1	1.000	1.0000000	1	1 Inhibited
61	Type I dia	04940	2	2	0.00000000	1	1.000	1.0000000	1	1 Inhibited
62	Aldosteron	04960	1	1	0.00000000	1	1.000	1.0000000	1	1 Inhibited

63	Alzheimer'	05010	2	2	0.00000000	1	1.000	1.0000000	1	1 Inhibited
64	Amyotrophi	05014	1	1	0.00000000	1	1.000	1.0000000	1	1 Inhibited
65	Huntington	05016	1	1	0.00000000	1	1.000	1.0000000	1	1 Inhibited
66	Vibrio cho	05110	2	2	0.00000000	1	1.000	1.0000000	1	1 Inhibited
67	Colorectal	05210	2	2	0.00000000	1	1.000	1.0000000	1	1 Inhibited
68	Thyroid ca	05216	2	2	0.00000000	1	1.000	1.0000000	1	1 Inhibited
69	Chronic my	05220	2	2	0.00000000	1	1.000	1.0000000	1	1 Inhibited
70	Autoimmune	05320	2	2	0.00000000	1	1.000	1.0000000	1	1 Inhibited
71	Systemic l	05322	1	1	0.00000000	1	1.000	1.0000000	1	1 Inhibited
72	Allograft	05330	1	1	0.00000000	1	1.000	1.0000000	1	1 Inhibited
73	Graft-vers	05332	2	2	0.00000000	1	1.000	1.0000000	1	1 Inhibited

BIBLIOGRAFIA

Affymetrix Inc., Gene Chip Expressioni Analysis Technical Manual, Technical Report, 2001 a.

Affymetrix Inc., New Statistical Algorithms for Monitoring Gene

Expression On Gene Chip Probe Array, Technical note, http: //www.affymetrix.com/ pdf/ algorithms.pdf, 2001 b.

M. Ashburner et al. Gene Ontology: tool for the unification of biology.

Nat. Genet., 25, 25- 29. (anno 2000).

S. Chiaretti, Xiaochun Li, Robert Gentleman, Antonella Vitale, Kathy S.

Wang, Franco Mandelli, Robin Foà and Jerome Ritz. Gene Expression Profiles of B-lineageAdult Acute Lymphocytic Leukemia Reveal Genetic Patterns that Identify LineageDerivation and Distinct Mechanisms of Transformation. Clin Cancer Res 2005.

Dinu, J. D. Potter, T. Mueller, Q. Liu, A. J. Adewale, G. S. Jhangri, G. S.

Famulski, P. Halloran and Y.Yasui. Gene – Set Analysis and Reduction. PubMed Central 2008.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2638622/?tool=pubmed.

J. J. Goeman and P. Bühlmann. Analyzing Gene Expression Data in Terms

Of Gene Sets: Methodological Issues. Original Paper (Vol. 23 no. 8 2007, pages 980- 987; doi: 10.1093/bioinformatica/btm051) 2007.

F.Hahne, W.Huber, R.Gentleman, S. Falcon. Bioconductor Case Studies.

Springer 2008.

Q. Liu, I. Dinu, A. J Adewale, J. D Potter and Y. Yasui. Comparative

Evaluation of Gene- Set Analysis Methods. PubMed Central, 2007.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2238724/?tool=pubmed.

M. S. Massa, M. Chiogna and C. Romualdi. Gene Set Analysis Exploiting

the Topology of a Pathway. PubMed Central 2010. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2945950/?tool=pubmed.

H. Ogata et al. KEGG: Kyoto Encyclopedia of Genes and Genomes,

Nucleic Acids Res., 27, 29- 34. (anno 1999).

G. Parmigiani, E. S. Garret, R. A. Irizarry, S. L. Zeger. The Analysis of

Gene Expression Data. Springer 2003.

T.Speed. Statistical Analysis of Gene Expression Microarray Data.

Chapman & Hall/ Crc 2003.

A. L. Tarca, S. Draghici, P. Khatri, S. S. Hasssan, P. Mittal, J. Kim, C. J.

Kim, J. P. Kusanovic and R. Romero. A Novel Signaling Pathway Impact Analysis. PubMed Central 2009.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2732297/?tool=pubmed

P.H. Westfall and S.S. Young. Resampling- Based Multiple Testing:

Examples and Methods for p- Value Adjustement, New York: John Wiley & Sons, 1993.