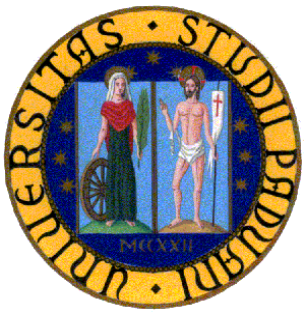
# UNIVERSITA' DEGLI STUDI DI PADOVA

## FACOLTA' DI SCIENZE STATISTICHE

### CORSO DI LAUREA
### STATISTICA E GESTIONE DELLE IMPRESE




### TESI DI LAUREA

# Algoritmo DEDS:
# un'analisi mediante studi di simulazione.

**Relatore: Ch.ma Prof.ssa Monica Chiogna**

**Laureanda: Elisa Rosso**
**470943-GEI**

**ANNO ACCADEMICO 2006-2007**

# INDICE

Capitolo 5
GLI STUDI DI SIMULAZIONE



APPENDICE



Bibliografia

# Introduzione

I microarray a Dna stanno diventando sempre più comuni nella ricerca biologica e medica, in quanto essi consentono lo studio simultaneo di centinaia di geni e permettono di ottenere informazioni sull'espressione genica di un intero livello genomico, come non accadeva in precedenza.
Un obiettivo comune degli esperimenti su microarray è la scoperta di differenti espressioni geniche, tra campioni ottenuti sotto diverse condizioni; attualmente questo è un argomento di grande interesse per i ricercatori che da un lato, in ambito farmacologico, sono interessati all'individuazione di nuovi farmaci e dall'altro studiano le basi molecolari ad esempio di malattie oncologiche o complesse.
I microarray operano monitorando simultaneamente i livelli di espressione di migliaia di geni: questa è un'innovazione straordinaria se si pensa che i ricercatori, in precedenza, analizzavano solo un gene alla volta. In questo modo vengono generati molti data set multivariati che presentano numerose "sfide analitiche": bisogna infatti prevedere, in primo luogo, l'aggiustamento delle varie fonti di variabilità, che sorgono dalle varie fasi dell'esperimento ed, in secondo luogo, si devono usare metodi di analisi che sono adattati alle nuove strutture di dati, costituiti da centinaia di variabili (geni) e da campioni di piccole dimensioni (array), aventi poche o nulle ripetizioni.
Lo scopo di queste analisi è quello di identificare in mezzo a molti geni, per i quali sono state ottenute misure di espressione, proprio quelli a cui è associata una risposta di interesse. Nel momento in cui avviene il confronto delle misure di espressione tra gruppi o condizioni differenti vengono identificati i geni "importanti", che verranno detti "differenzialmente espressi" (DE), che si distinguono da quelli chiamati "equivalentemente espressi" (EE).
L'operazione di identificazione dei geni differenzialmente espressi può essere divisa in due fasi fondamentali: la classificazione dei dati e la scelta

dei dati significativi. Nella prima fase, si richiede per ciascun gene la specificazione di una statistica o l'applicazione di un procedimento in grado di "catturare" l'evidenza dell'espressione differente (DE) focalizzandosi su una risposta dicotomica, attraverso il confronto fra due gruppi. Nella seconda fase, detta di "scelta", si richiede la specificazione di un livello per fissare arbitrariamente la significatività dell'espressione ad esempio stabilendo un valore critico. Da un punto di vista operativo, portare a termine la prima fase è meno difficoltoso del compimento della seconda, anche se esistono delle criticità in quanto per l'analisi dei dati su microarray non esiste una statistica ottima e raramente c'è una motivazione che porta in modo univoco alla scelta di una particolare statistica.

Lo schema in Figura -0.1- illustra le fasi che si incontrano quando si decide di rispondere ad una domanda di natura medica o biologica, attraverso l'uso della tecnologia dei dna microarray. Le domande più frequenti riguardano la ricerca oncologica, l'immunologia, l'analisi della mutazione genica per identificare i geni correlati alle diverse malattie. Una volta formulato un interrogativo, si crea un disegno sperimentale che sia adatto alla tipologia di domanda posta in esame e si stabiliscono, quindi, alcune linee operative che verranno successivamente applicate nella fase dell'esperimento. Si esegue dunque l'esperimento e si ottengono le immagini relative alle espressioni geniche che devono essere analizzate. Successivamente, vengono applicate diverse tecniche di normalizzazione al fine di togliere dai dati tutta quella variabilità che non ha origine biologica. I dati così ottenuti vengono analizzati attraverso diverse tecniche statistiche ed, infine, i risultati devono essere verificati, valutati ed interpretati.

In questo lavoro ci si propone di sviluppare ed illustrare un nuovo approccio all'analisi dei dati di microarray chiamato DEDS (Differential Expression via Distance Synthesis), che integra le differenti misure di espressione differenziale e unisce statistiche differenti, attraverso uno schema di sintesi. In particolare, si utilizza un data set simulato, che

rappresenta un insieme di dati provenienti da un esperimento su microarray con due diverse condizioni, in cui vi sono geni differenzialmente espressi. Infine, si dimostra che il metodo regge favorevolmente il confronto con le migliori statistiche individuali, mentre raggiunge in modo robusto proprietà che mancano alle singole statistiche.

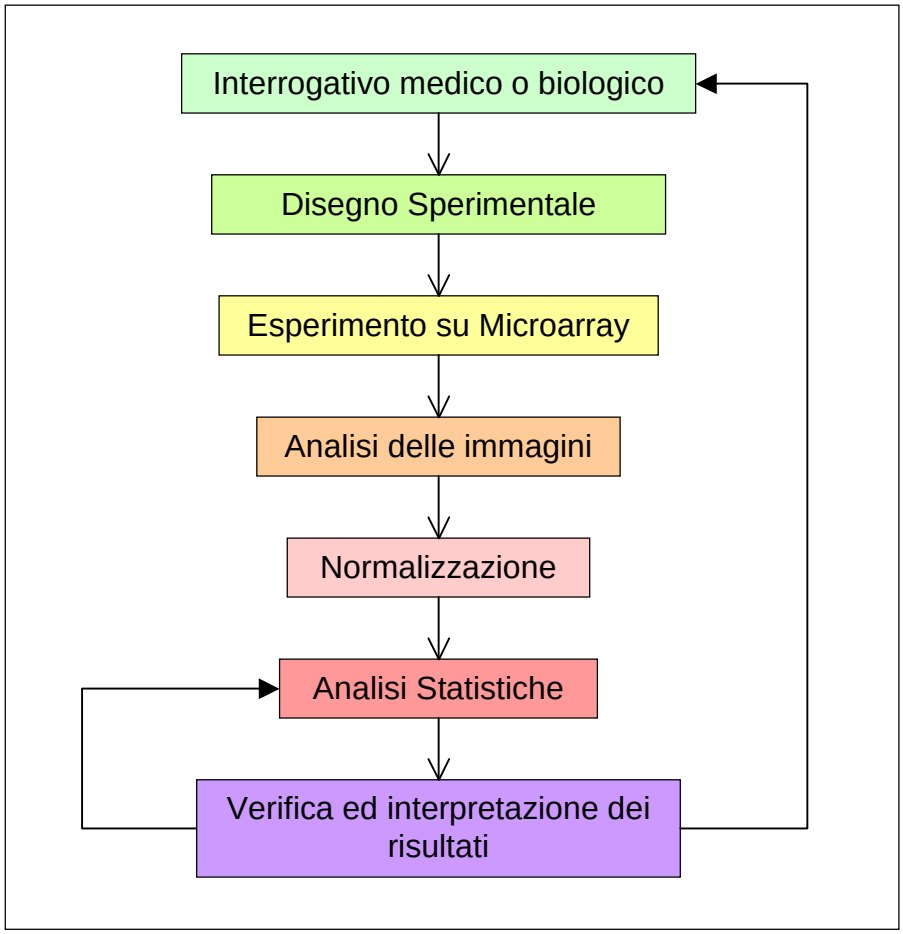



**Figura -0.1- Fasi dell'esperimento.**

# Capitolo 1
# GLI ESPERIMENTI CON DNA MICROARRAY

## 1.1 Alcuni richiami alla struttura della cellula

Le cellule sono le unità funzionali e strutturali biologiche di base di ogni essere vivente ed in natura se ne trova una grandissima varietà.
La cellula è un'unità indivisibile ed è costituita da un insieme di strutture, per lo più formate da molecole proteiche, quali:

- la membrana plasmatica, che ha la funzione di regolare gli scambi della cellula con l'ambiente esterno;
- il citoplasma, che costituisce la massa cellulare all'interno della membrana in cui sono contenuti:
    - i mitocondri, all'interno dei quali avvengono le reazioni chimiche della respirazione cellulare,
    - i ribosomi, che svolgono la funzione di sintetizzare le proteine,
    - i vacuoli, in cui si trovano le riserve nutritive della cellula,
    - i lisosomi, che controllano la distruzione di eventuali corpi esterni,
    - il reticolo endoplasmatico, attraverso cui le sostanze sono trasportate,
    - il complesso del Golgi, che regola gli scambi di sostanze verso l'esterno della cellula,
    - i centrioli, che svolgono una funzione importante nella riproduzione cellulare;

- il nucleo (parte centrale e più importante della cellula) che svolge la funzione di dirigere tutte le attività e provvedere alla riproduzione della cellula.

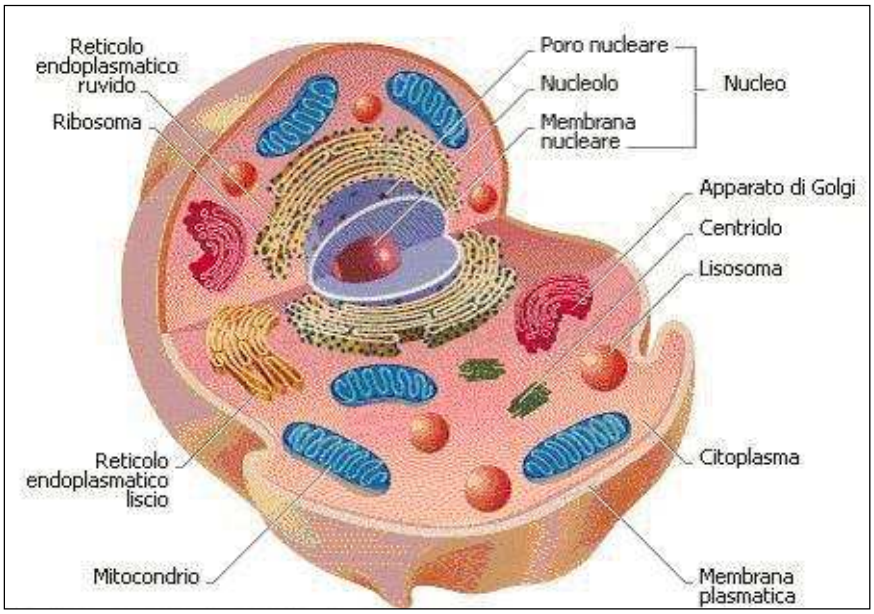



**Figura -1.1- La struttura della cellula**

All'interno del nucleo, in particolare nel Dna, risiede tutta l'informazione genetica. Il Dna o acido desossiribonucleico è formato da due lunghi filamenti uniti tra loro e avvolti a spirale l'uno sull'altro, in modo da formare una doppia elica. Ciascun filamento è costituito dalle unità fondamentali del Dna che sono chiamate nucleotidi, formate da una molecola di acido fosforico, da una di zucchero e da quattro basi azotate: citosina, guanina, timina e adenina. La possibilità dei vari nucleotidi di disporsi in successione e quantità diverse fa sì che, in natura, ogni specie sia caratterizzata da una diversa molecola di Dna e quindi da specifici cromosomi. Ogni gruppo di tre basi azotate, nell'ordine in cui si susseguono, rappresenta un aminoacido, l'unità di base delle proteine, e la parte di Dna formata dalle triplette che codificano una specifica proteina prende il nome di gene. L'insieme di tutti i geni costituisce il patrimonio

genetico di un individuo, detto anche genoma, che è unico ed è costituito da molecole di Dna.

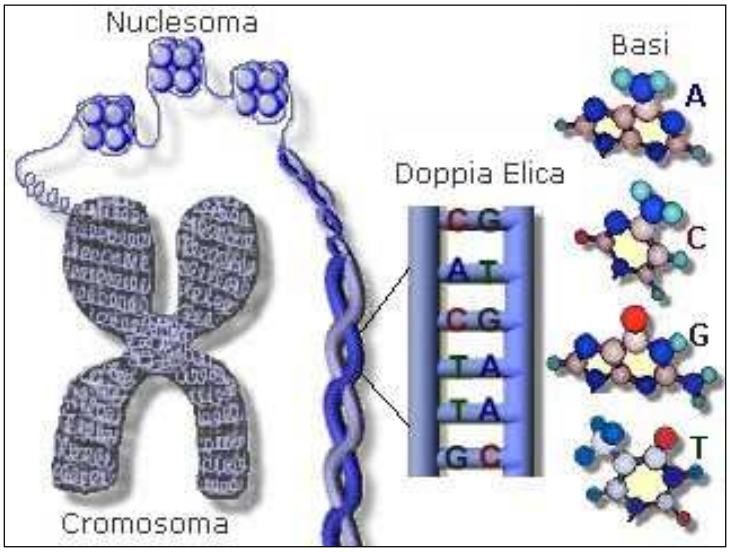



**Figura -1.2- I cromosomi**

Attraverso la sintesi proteica, che avviene all'interno dei ribosomi, le informazioni contenute nei geni possono essere tradotte in proteine. L'Rna o acido ribonucleico è una catena singola di nucleotidi che ha il compito di trasferire le informazioni dal Dna ai ribosomi. L'Rna è l'intermediario mediante il quale l'informazione, presente nel Dna, viene utilizzata per la costruzione delle proteine e ha una struttura simile a quella del Dna, infatti, presenta le seguenti basi azotate: citosina, adenina, guanina e uracile. Durante la sintesi proteica il Dna apre la doppia elica e agisce come se fosse uno stampo per la sintesi della molecola di Rna, detto "messaggero" (RNAm), nella quale verrà trascritta l'informazione dei vari geni. Ultimata la trascrizione l'Rna si stacca, esce dal nucleo e si fissa nei ribosomi: la molecola di Rna messaggero deve essere tradotta. Il processo di traduzione coinvolge l'Rna ribosomiale (RNAr) e l'Rna di trasporto (RNAt). Attraverso l'Rna "di trasporto" vengono agganciati gli amminoacidi in modo complementare all'Rna "messaggero" finché la costruzione della proteina non avviene in modo completo. La sintesi di una proteina avviene quindi

per tappe: la catena polipeptidica della proteina si forma per l'aggiunta di un amminoacido per volta.

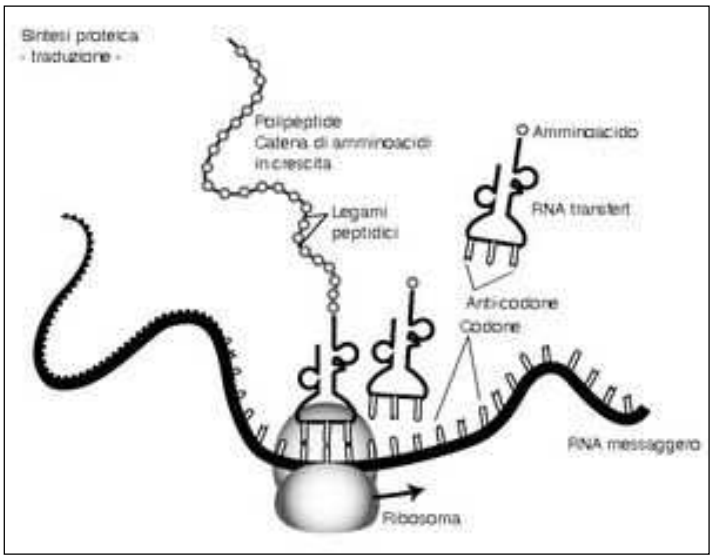



**Figura -1.3- La sintesi proteica**

## 1.2 I Dna Microarray

La tecnologia dei Dna microarray permette di studiare il livello di attivazione di migliaia di geni in contemporanea: i geni espressi in quantità diverse determinano, infatti, una enorme varietà di proteine ma, a questo punto ci si chiede come avvenga l'identificazione di questi geni.

I microarray a Dna, chiamati anche Dna chip, sono formati da un insieme di microscopiche sonde di Dna depositate su un supporto solido di vetro o plastica, in una posizione nota, come per formare una microgriglia che consente un'individuazione univoca.

Ogni sonda è costituita da un segmento di Dna a singola elica, che rappresenta un gene; nel loro insieme, le sonde di un Dna chip rappresentano tutti, o la maggior parte, dei geni di un organismo. La tecnologia dei Dna microarray sfrutta la proprietà strutturale che ha il Dna

di appaiarsi tra basi azotate complementari, infatti le basi si appaiano in questo modo: timina con adenina e guanina con citosina.
Nel momento in cui si devono confrontare i geni relativi alle cellule di due tessuti, che provengono da due differenti condizioni, ad esempio un tessuto sano ed uno malato, si può notare che alcuni geni sono "differenzialmente espressi" ed altri "equivalentemente espressi", a seconda della condizione da cui derivano. Quando i geni sono differenzialmente espressi nelle cellule del tessuto è presente un numero elevato di molecole di Rna messaggero. Si estrae pertanto l'Rna dai due tipi di tessuti e si procede, attraverso una tecnica di "ibridazione inversa", alla conversione dell'Rna messaggero nella copia più stabile chiamata "cDna". Grazie ad una sonda fluorescente i "cDna" sono marcati: il verde, ad esempio, sarà il colore per le cellule malate e il rosso per le cellule sane.

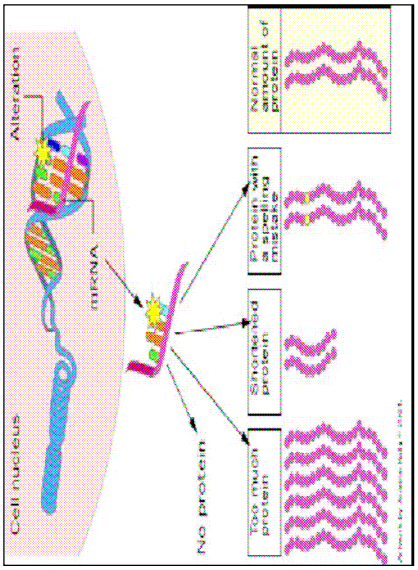



**Figura -1.4- L'identificazione dei geni**

I "cDna" sono applicati sul chip e nel momento in cui il "cDna" trova la sequenza di basi complementare vi si appaia: in quel punto del microarray si ha un'emissione di fluorescenza che indica l'espressione di quel determinato gene. I chip vengono successivamente analizzati con uno

scanner a laser, che permette l'acquisizione di un immagine per ogni fluoroforo e successivamente, si usano dei software appositi per convertire i segnali in una gamma di colori che dipende dalla loro intensità.

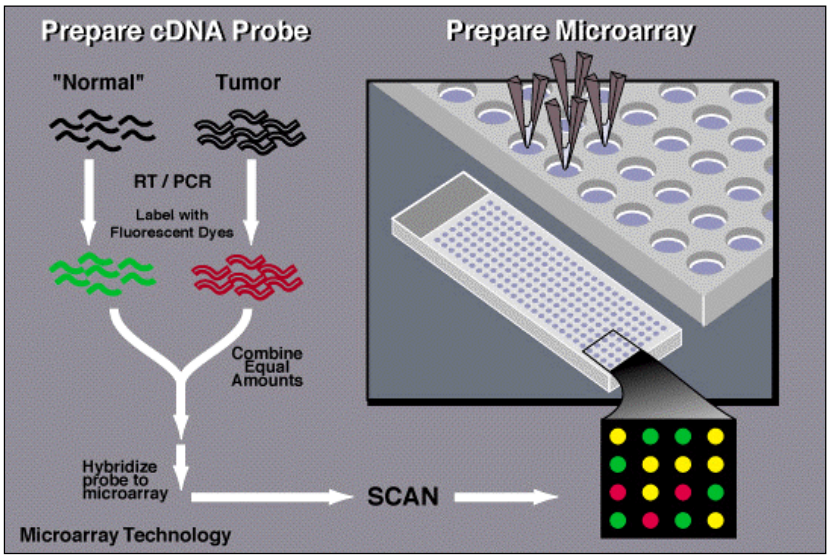



**Figura -1.5- l'esperimento su Dna microarray.**

A questo punto dell'esperimento si ottiene una mappa, chiamata profilo di espressione, in cui vi sono diversi colori, ad esempio il rosso indica un gene espresso solo nel tessuto sano, il verde indica che l'espressione riguarda solo il tessuto malato e si hanno anche altre gradazioni che indicano i livelli diversi di espressione relativi ad entrambi i tessuti. Il segnale rilevato dallo scanner viene poi sottoposto ad alcuni algoritmi di filtrazione e di pulizia per poi essere convertito in dati numerici.

I dati devono essere aggiustati al fine di rimuovere eventuali errori sistematici nelle misurazioni dovuti alle immagini che non sono sempre correttamente visualizzate: possono esserci, infatti, depositi irregolari, sovraesposizione degli spot oppure un background eccessivo.

Senza questi aggiustamenti i risultati che si ottengono non potrebbero essere confrontati. Si rende necessario, infine, un processo di "normalizzazione" dei dati, per rimuovere le distorsioni sistematiche e non

biologiche e in modo tale da poter confrontare il livello di espressione all'interno dello stesso array e tra array diversi.

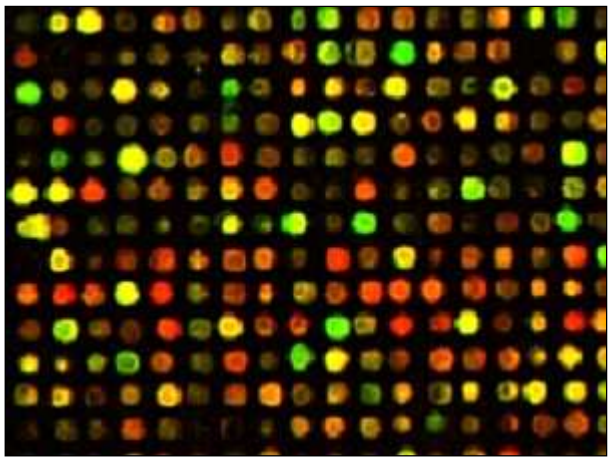


**Figura -1.6- La mappa di espressione genica**

# CAPITOLO 2
# LA SIMULAZIONE DEI DATI

## 2.1 Caratteristiche di base delle simulazioni

La tecnica del dna microarray consente lo studio di migliaia di geni contemporaneamente ma, dal punto di vista statistico, l'analisi dei dati provenienti da questi esperimenti presenta alcune difficoltà: si hanno infatti a disposizione misure di migliaia di geni con poche repliche per ciascun gene ("large p and small n"). I metodi bayesiani empirici si adattano bene a questo tipo di analisi di dati, in quanto sfruttano l'informazione condivisa da tutti i geni, utile per capire la variabilità del sistema.

Per valutare la capacità del metodo proposto in questo lavoro, di identificare geni differenzialmente espressi, è stata sviluppata una procedura per la generazione di dati "artificiali" (una simulazione) aventi caratteristiche analoghe ai dati prodotti dagli esperimenti di dna microarray.

Per la simulazione dei geni, provenienti da esperimenti di dna microarray, vengono utilizzati due modelli mistura di tipo gerarchico. L'assunto di base è riconducibile alle seguenti relazioni: un gene equivalentemente espresso rappresenta la realizzazione di una variabile avente la stessa distribuzione di probabilità nei gruppi; se invece un gene è differenzialmente espresso, proviene da distribuzioni di probabilità diverse.

Lo schema di simulazione è basato su un modello gerarchico a due strati, in cui uno strato tiene conto della variabilità di espressione relativa a tutti i geni. I dati simulati rappresentano esperimenti in cui le cellule sono

sottoposte a due differenti condizioni e, per ogni condizione, vi sono diverse repliche.

Il modello mistura è specificato dalla distribuzione sulla singola osservazione, che caratterizza la variabilità relativa alle misure ripetute di un gene e, da una seconda componente, che descrive la variabilità di queste medie tra geni. La distribuzione di probabilità delle misure di espressione relative a un gene sono di tipo parametrico e le stime dei parametri dipendono dalla variabilità tipica di ogni singolo esperimento di microarray.

I modelli alla base delle simulazioni sono il modello Gamma-Gamma ed il modello LogNormale-Normale di seguito descritti.

## 2.2 Il modello Gamma-Gamma

Nel modello Gamma-Gamma (GG) la distribuzione sulla singola osservazione è di tipo Gamma con parametro di forma $\alpha > 0$ e media $\mu_g$, il parametro di scala sarà dato da $\lambda_g = \alpha / \mu_g$. Per la quantità $\lambda_g = \alpha / \mu_g$, si assume una distribuzione Gamma con parametro di forma $\alpha_0$ e parametro di scala $\nu$ ($\lambda_g \sim Ga(\alpha_0, \nu)$). I parametri del modello Gamma-Gamma sono quindi: $\theta = (\alpha, \alpha_0, \nu)$.

Oltre a questi parametri si deve stimare la proporzione di miscuglio del modello mistura, ovvero la probabilità a priori $p$ che un gene appartenga ad una delle condizioni.

La simulazione del modello Gamma-Gamma avviene per $N$ geni sotto due diverse condizioni, con $m$ repliche per ogni condizione. I parametri di questa simulazione per il modello Gamma-Gamma sono stati fissati pari a $(\alpha = 10, \alpha_0 = 0.9, \nu = 0.5)$, con una probabilità a priori che un gene sia differenzialmente espresso fissata a $p = 0.2$.

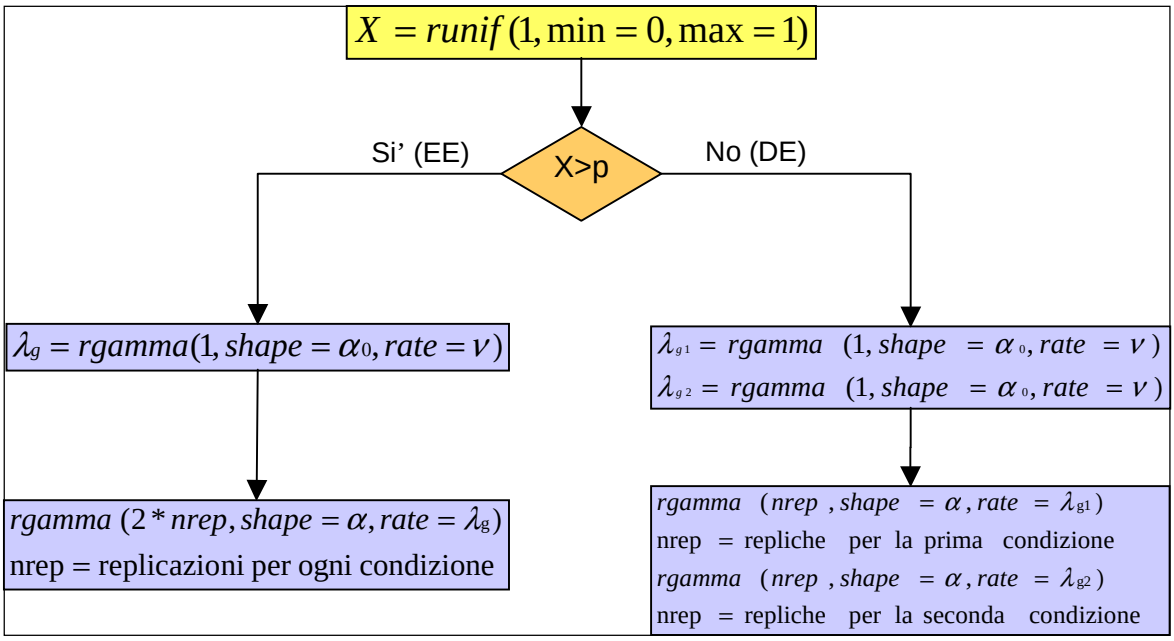



**Figura -2.1- Il diagramma illustra lo schema seguito per la simulazione del modello Gamma-Gamma, questa procedura è ripetuta *N* volte, per simulare *N* geni, sotto due condizioni diverse e con *m* repliche per ogni condizione.**

## 2.3 Il modello LogNormale-Normale

Nel modello LogNormale-Normale si ipotizza che la distribuzione relativa alla trasformata logaritmica della singola osservazione sia normale con media $\mu_g$, che dipende dal singolo gene e varianza $\sigma^2$ comune per tutti i geni ($N(\mu_g, \sigma^2)$). Accanto alla distribuzione normale si ha la distribuzione di $\mu_g \sim N(\mu_0, \tau_0^2)$. I parametri coinvolti in questo modello sono quindi: $\theta = (\mu, \mu_0, \tau_0^2)$.

La simulazione del modello LogNormale-Normale avviene per $N$ geni posti sotto due condizioni, con $m$ repliche per ogni condizione. I parametri di questa simulazione sono $(\mu_0 = 2.3, \sigma = 0.3, \tau = 1.39)$, anche in questo caso la probabilità a priori che un gene sia differenzialmente espresso è fissata a $p = 0.2$.

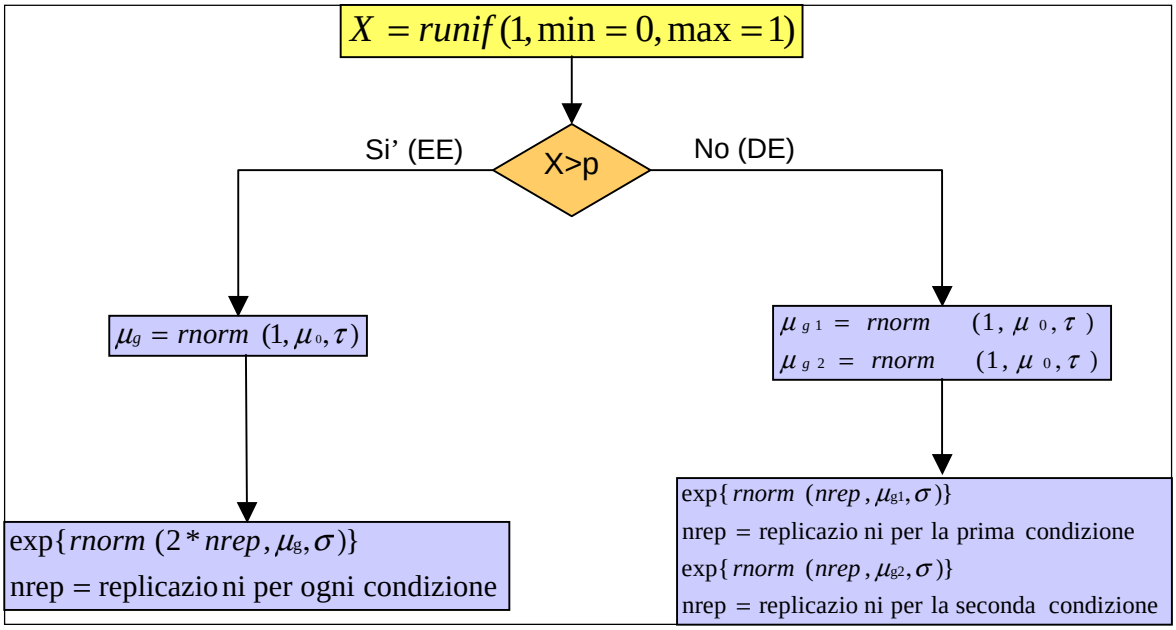



**Figura -2.2- Il diagramma illustra lo schema seguito per la simulazione del modello LogNormale-Normale, questa procedura è ripetuta *N* volte, per simulare *N* geni, sotto due condizioni diverse e con *m* repliche per ogni condizione.**

Le simulazioni sono state effettuate attraverso il software statistico R. I codici inerenti a queste simulazioni sono inseriti all'interno delle funzioni che calcolano l'algoritmo Deds.

Il primo caso illustra la simulazione con il modello Gamma-Gamma, mentre il secondo illustra la simulazione con il modello LogNormale-Normale.

# CAPITOLO 3
# DEDS: UN NUOVO APPROCCIO ALL'ANALISI DI DATI PROVENIENTI DA DNA MICROARRAY

## 3.1 Un nuovo approccio attraverso sintesi di statistiche test.

L'espressione differente dei geni può essere studiata attraverso diverse statistiche che presentano però alcune problematiche, inerenti alla complessa variabilità presente in tutti i dati provenienti da esperimenti su microarray. Tutte le fasi dell'esperimento sono affette da rumore, di conseguenza, se un esperimento è eseguito ad esempio due volte, nelle stesse condizioni, molti geni riporteranno diversi livelli di espressione. I campioni a disposizione sono sempre in numero limitato, in quanto tali esperimenti sono molto costosi e le risorse a disposizione sono spesso limitate.

I ricercatori devono attuare una scelta accurata della statistica da utilizzare e, successivamente, devono valutare i risultati che ottengono anche alla luce dei limiti presenti nel metodo che hanno scelto.

La tecnica più semplice per stabilire l'insieme dei geni che presenta un'espressione differenziale è chiamata "Fold Change", un metodo che fa uso del rapporto fra le intensità di segnale, misurate in entrambi i canali. I valori che si ottengono sono sempre compresi tra $[0,+\infty]$: se il rapporto $FC = 1$ si ha un'espressione simile in entrambi i canali, invece con $FC \neq 1$ si ha una diversa espressione tra i due canali. Utilizzando questo metodo

non si evidenzia però una valutazione statistica degli errori che possono essere commessi nell'affermare che un gene è differenzialmente espresso.
Una statistica largamente impiegata in questa tipologia di analisi è la statistica t di Student, usata per il confronto tra le medie di due gruppi. Anche questa statistica presenta alcuni limiti: infatti, si incontrano difficoltà nella stima della varianza, nel caso in cui si disponga di un campione di piccole dimensioni. Per stimare la varianza in modo più preciso, sono stati proposti diversi approcci mirati a stabilizzare la varianza o a contenere gli errori.
Al fine di superare le limitazioni appena esposte, si descrive in questo lavoro un nuovo approccio all'analisi dei geni provenienti da microarray, che integra statistiche diverse in un unico algoritmo chiamato DEDS ("Differential Expression via Distance Synthesis"). Questo approccio si fonda sul principio che i geni che presentano valori elevati, per tutte le statistiche prese in considerazione, possono essere definiti "differenzialmente espressi", con più sicurezza rispetto ai geni che presentano misure elevate solo per una singola statistica.
In questo approccio si utilizza il data set ottenuto grazie agli esperimenti su dna microarray ed altri data set, ottenuti attraverso varie permutazioni del data set di partenza.
Una volta scelte alcune statistiche significative nell'analisi di espressioni geniche, si procede calcolando i valori dei test su tutti i data set a disposizione; si ottengono in questo modo le misure inerenti ai dati osservati e permutati. Si calcolano, successivamente, i valori massimi delle statistiche test al fine di ottenere, per ciascuna statistica, un valore che sia massimo globale sia per i test su dati osservati sia per i test su dati permutati. Il valore massimo così ottenuto può non corrispondere ad un valore osservato.
Nella fase successiva del metodo vengono calcolate le distanze tra i valori delle statistiche ottenuti dal data set osservato ed il massimo globale. Intuitivamente, i valori che più si avvicinano al massimo globale sono quelli che corrispondono ai geni differenzialmente espressi.

Nel paragrafo seguente si descrivono le varie fasi del procedimento in modo analitico, mentre, in Appendice, è riportato il codice dell'algoritmo, implementato con il software statistico R.

## 3.2 Deds: un algoritmo delle permutazioni di base

1. Predisposizione del dataset di partenza.

   A ciascuno degli $i$ , $i = 1,...,N$ , geni presenti nel data set $T$ simulato sono applicate statistiche $t_j$ , $j = 1,...,J$ appropriate, in modo tale da ottenere i valori delle statistiche $t_{ij}$. Nel nostro caso vengono applicati tre test statistici: la statistica t-student, il test di Wilcoxon ed un test Semiparametrico. Successivamente, per $i = 1,...,N$ , si calcolano i valori massimi per ciascuna statistica, ottenendo il vettore $E_0 = (\max_i(t_{i1}),...,\max_i(t_{iJ}))$ .

2. Individuazione del massimo assoluto $E$.

   a. Per $b$ , $b = 1,...,B$ , si ottengono le permutazioni per riga del dataset di partenza $T$ , in modo tale da ridisporre i valori simulati in modo casuale. Per ogni dataset permutato si ricalcolano, per ciascuno degli $N$ geni, le $J$ statistiche come indicato al punto precedente, ottenendo i valori $t_{ij}^b$. I risultati così ottenuti sono salvati in file esterni. Vengono calcolati, successivamente, i valori massimi $E_b = (\max_i(t_{i1}^b),...,\max_i(t_{iJ}^b))$ per le statistiche $j = 1,...,J$ calcolate.

   b. Si prendono in considerazione i vettori $E_b$ e si dispongono in una matrice chiamata $E_B$ per $b = 1,...,B$. Si procede calcolando i massimi per colonna della matrice $E_B$ , al fine di ottenere $E_p = (\max_b(t_{b1}),...,\max_b(t_{bJ}))$ .

c. Dal confronto dei vettori $E_p$ ed $E_0$, si ottiene il vettore $E$, che corrisponde al massimo globale: $E = \max(E_p, E_0)$.

3. Calcolo della distanza da ciascun gene al massimo globale.

   Per calcolare la distanza $d$ da ciascun gene al massimo assoluto $E$ si utilizza la seguente formula:

$$d_i = \frac{(t_{i1} - E_1)^2}{MAD(t_1)^2} + \frac{(t_{i2} - E_2)^2}{MAD(t_2)^2} + ... + \frac{(t_{iJ} - E_J)^2}{MAD(t_J)^2}$$

   dove MAD è la deviazione assoluta dalla mediana.

   Infine le distanze ottenute sono ordinate in modo crescente ottenendo così:

$$d_{(1)} \le d_{(2)} \le ... \le d_{(N)}.$$

## 3.3 “False Discovery Rate”: come valutare i risultati ottenuti.

Il metodo proposto in questo lavoro, per analizzare dati provenienti da dna microarray, integra le singole statistiche in un unico algoritmo in cui si ottiene un vettore ordinato ( $d_{(1)} \le d_{(2)} \le ... \le d_{(N)}$ ) di distanze dei geni dal massimo assoluto $E$. I geni per i quali le distanze calcolate sono “significativamente piccole” sono molto vicini al massimo assoluto $E$ e quindi si possono considerare differenzialmente espressi. Per identificare questi geni si utilizza il “False Discovery Rate” (FDR) che è definito come il rapporto tra il numero di geni identificati in modo errato e il numero dei geni differenzialmente espressi nei dati originali. Il FDR è il livello che ci si aspetta di avere di identificazioni errate tra i geni che sono identificati come differenzialmente espressi.

Il FDR viene calcolato attraverso le seguenti fasi:

1. si prendono in considerazione le matrici che contengono i valori dei test $t_{i,j}^{b}$ relativi ai *B*-data set ottenuti dalla permutazione del data set di partenza: per ciascuna viene calcolato il vettore ordinato delle distanze $d_{(1)}^{b} \le d_{(2)}^{b} \le ... \le d_{(N)}^{b}$ dal massimo assoluto $E$;

2. per ogni gene contenuto nel vettore ordinato $d_{(1)} \le d_{(2)} \le ... \le d_{(N)}$ e per ogni vettore $d_{(1)}^{b} \le d_{(2)}^{b} \le ... \le d_{(N)}^{b}$ con $b = 1,...,B$, si calcola il numero di geni che soddisfano alle relazioni:

$$\begin{matrix} FP_{(1)}^{b} := \#(d_{(j)}^{b} \le d_{(1)}) \\ FP_{(2)}^{b} := \#(d_{(j)}^{b} \le d_{(2)}) \\ .... \\ FP_{(N)}^{b} := \#(d_{(j)}^{b} \le d_{(N)}) \end{matrix};$$

3. per $i$, $i = 1,...,N$, si calcolano le mediane dei valori contenuti nella matrice $FP$:

$$\begin{matrix} medFP_{(1)} = mediana(FP_{(1)}^{b}) \\ .... \\ medFP_{(N)} = mediana(FP_{(N)}^{b}) \end{matrix};$$

4. il valore che indica il FDR per l'$i-esimo$ gene ordinato è calcolato come il rapporto tra le mediane del numero dei geni chiamati in modo errato "Differenzialmente Espressi" e l'indice *i* corrispondente al gene. Al denominatore del rapporto si ha la media del numero di valori significativi per le $J$ statistiche considerate.

$$FDR = \begin{matrix} medFP_{(1)} / 1 \\ medFP_{(2)} / 2 \\ .... \\ medFP_{(N)} / N \end{matrix}.$$

La criticità che si incontra nel momento in cui si deve identificare i geni differenzialmente espressi riguarda la specificazione del livello del FDR: si possono considerare, ad esempio, differenzialmente espressi geni che hanno un valore di FDR sotto lo 0.01 o 0.05. La scelta di questi parametri si rivela molto delicata in quanto il numero di geni che vengono identificati come differenzialmente espressi cambia significativamente, a seconda del livello prescelto.

# CAPITOLO 4
# LE STATISTICHE TEST USATE NELL'ALGORITMO

## 4.1 Tipologie diverse di test statistici

Tra i metodi usati, per fare inferenza su una popolazione, si possono distinguere i metodi classici o parametrici ed i metodi non parametrici.
Nel caso dei test parametrici, prima dell'applicazione, è fondamentale che siano soddisfatti alcuni assunti che riguardano la popolazione d'origine. La prima di queste assunzioni è l'indipendenza dei gruppi campionari: i campioni dovrebbero essere estratti casualmente dalla popolazione, in modo tale che i fattori aleatori siano distribuiti in modo casuale. Il secondo assunto è la normalità delle distribuzioni, da cui deriva la relazione tra popolazione dei dati e medie dei campioni: il teorema del limite centrale dice, infatti, che se si estraggono casualmente campioni di dimensione n, da una popolazione con media $\mu$ e varianza $\sigma^2$, i cui dati hanno una forma di distribuzione non normale, le loro medie saranno distribuite normalmente, con media generale $\mu$ ed errore standard $\sigma^2/\sqrt{n}$. Il terzo assunto riguarda l'omoschedasticità o omogeneità delle varianze, indispensabile in quanto nelle statistiche parametriche è possibile verificare se esistono differenze significative tra medie campionarie solo quando i gruppi a confronto hanno la stessa varianza.
Nel caso in cui ci si accorge che i suddetti assunti non sono rispettati, si può ricorrere ai metodi statistici non parametrici, i quali non dipendono dalla forma di distribuzione della popolazione, non si basano sui parametri

ed è possibile applicarli anche a dati qualitativi. I metodi non parametrici si fondano sulle statistiche di rango od ordine, piuttosto che sulle osservazioni in sé; essi non utilizzano la media, ma la mediana come misura della tendenza centrale e vengono indifferentemente applicati sia alle variabili casuali discrete sia alle continue. I metodi non parametrici richiedono poche assunzioni sulle caratteristiche della popolazione dalla quale il campione è stato estratto, in particolare non richiedono la normalità e sono molto meno rigorosi. Tali metodi sono, inoltre, meno sensibili ai valori anomali e portano a conclusioni più generali, tuttavia essi sfruttano in modo meno completo l'informazione contenuta nei dati ed hanno quindi una potenza inferiore. I metodi non parametrici sono comunque adatti a problemi relativamente semplici, come il confronto tra medie e tra varianze sempre relativamente ad un solo fattore. Con strutture di dati complesse, in cui si vogliono considerare contemporaneamente più fattori e covariate, non esistono alternative al modello parametrico.

"In generale, è forse meglio considerare i metodi non parametrici come un insieme di tecniche cui far riferimento quando gli assunti teorici standard hanno una validità relativamente dubbia. Infine, torna spesso utile poter confermare i risultati di un test di significatività, basato sulla teoria normale, mediante l'applicazione di un appropriato test non parametrico" (Armitage P., Statistica Medica. Metodi statistici per la ricerca in Medicina, McGraw Hill, Libri Italia).

In questo lavoro si utilizzano tre statistiche, la t di Student per due gruppi, che appartiene ai metodi parametrici, il test Semiparametrico, in cui si assume nota solo la distribuzione di uno dei due gruppi e il test di Wilcoxon, analogo non parametrico del test t di Student per campioni indipendenti.

## 4.2 Il test t di Student

Nelle ricerche sperimentali, le situazioni più ricorrenti sono quelle del confronto tra due medie campionarie: in questi casi la distribuzione t di Student può essere derivata dal rapporto tra la differenza delle due medie campionarie ed il suo errore standard. Il test di significatività tra due medie campionarie comporta un'ipotesi nulla, secondo la quale le due medie a confronto sono estratte dalla stessa popolazione. Sotto $H_0$, le differenze effettivamente riscontrate nelle medie campionarie sarebbero imputabili solo a variazioni casuali, come effetti dovuti al campionamento. Mediante l'inferenza sulle medie calcolate sui dati di due campioni, si determina la probabilità di ottenere tra loro differenze significative da quelle sperimentalmente osservate, nel caso in cui l'ipotesi nulla sia vera. Se questa probabilità risulta maggiore del convenzionale livello di significatività $\alpha = 0.05$ , si accetta l'ipotesi nulla altrimenti, si deve ragionevolmente rifiutarla e si afferma quindi l'esistenza di una differenza tra le due medie. Il test t di Student può essere usato sia per campioni dipendenti sia per campioni indipendenti; in questo lavoro si prenderanno in esame solamente campioni indipendenti.

Sia $x = (x_1,...,x_n)$ un campione casuale semplice di numerosità $n$ , proveniente da una distribuzione Normale con media $\mu_x$ e varianza $\sigma_x^2$ ( $N(\mu_x,\sigma_x^2)$ ) e sia $y = (y_1,...,y_m)$ un altro campione casuale semplice di numerosità $m$ , proveniente da una distribuzione Normale di media $\mu_y$ e varianza $\sigma_y^2$ ( $N(\mu_y,\sigma_y^2)$ ), il sistema di ipotesi che si deve verificare è:

$$\begin{cases} H_0 : \mu_x = \mu_y \\ H_1 : \mu_x \neq \mu_y \end{cases} \text{ oppure } \begin{cases} H_0 : \mu_x - \mu_y = 0 \\ H_1 : \mu_x - \mu_y \neq 0 \end{cases}$$

in questo caso, i gradi di libertà del *t* sono uguali a $n+m-2$.

Il valore del test *t* è ottenuto mediante la formula:

$$t_{(n+m-2)} = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{S \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

in cui: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\bar{y} = \frac{\sum_{i=1}^{m} y_i}{m}$ e $S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{m} (y_i - \bar{y})^2}{n+m-2}}$.

Si rifiuta l'ipotesi nulla $H_0$ se il valore della statistica test *t* risulta essere molto piccolo o molto grande: si calcolano allora i valori $t_{\frac{\alpha}{2}}$ e $t'_{\frac{\alpha}{2}}$ tali che

$\Pr(t \le t_{\frac{\alpha}{2}} \mid H_0) = \frac{\alpha}{2}$ e $\Pr(t \ge t'_{\frac{\alpha}{2}} \mid H_0) = \frac{\alpha}{2}$.

In altre parole, l'ipotesi nulla viene rifiutata se il valore osservato di *t* soddisfa ad una delle seguenti disuguaglianze:

$$t \le t_{\frac{\alpha}{2}} \qquad t \ge t'_{\frac{\alpha}{2}}$$

Al contrario, viene accettata quando il valore osservato di t soddisfa la seguente disequazione:

$$t_{\frac{\alpha}{2}} \le t \le t'_{\frac{\alpha}{2}}.$$

## 4.3 Il test Semiparametrico

La seconda statistica utilizzata nell'algoritmo è il Test Semiparametrico. Date due variabili casuali X ed Y, si dispone di due campioni $x = (x_1, ..., x_n)$ e $y = (y_1, ..., y_m)$ con $n > m$. Per quanto riguarda la variabile X non si assume nessuna distribuzione nota, mentre si suppone che la variabile Y abbia

distribuzione nota $F_y(Y,\theta)$ ed in particolare, questa distribuzione è una Normale i cui parametri sono $\theta=(\mu_y,\sigma_y^2)$.

Il sistema di ipotesi che si vuole verificare vede l'uguaglianza delle medie nell'ipotesi nulla ed un'alternativa bilaterale:

$$\begin{cases} H_0 : \mu_x = \mu_y \\ H_1 : \mu_x \neq \mu_y \end{cases}$$

Questo sistema di verifica di ipotesi equivale al seguente:

$$\begin{cases} H_0 : \Pr[X > Y;\vartheta] = \Pr[X < Y;\vartheta] \\ H_1 : \Pr[X > Y;\vartheta] \neq \Pr[X < Y;\vartheta] \end{cases} = \begin{cases} H_0 : \rho = \rho_0 = 0.5 \\ H_1 : \rho \neq \rho_0 \end{cases}$$

indicando la $\Pr[X > Y;\vartheta] = \rho$.

Per la stima dei parametri $\vartheta$ e $\rho$ si utilizza la stima di massima verosimiglianza.

Facendo inferenza sul campione $y=(y_1,...,y_m)$ di cui si dispone, si calcolano le stime:

$$\hat{\vartheta} = (\hat{\mu}_y, \hat{\sigma}_y^2) = \left( \frac{1}{m}\sum_{i=1}^{m} y_i, \frac{1}{m}\sum_{i=1}^{m} (y_i - \hat{\mu}_y)^2 \right)$$

$$\hat{\rho} = \frac{1}{n}\sum_{i=1}^{n} S(x_i;\vartheta) = \frac{1}{n}\sum_{i=1}^{n} \{1 - F_Y(x_i;\vartheta)\} = \frac{1}{n}\sum_{i=1}^{n} \left\{ 1 - \phi\left( \frac{x_i - y_i}{\sigma_y} \right) \right\}$$

in cui $\phi(\cdot)$ è la funzione di ripartizione della normale standard.

Il valore della statistica test è $t = \dfrac{\hat{\rho} - \rho_0}{\varpi / \sqrt{n}}$, che si distribuisce come una normale standard in quanto, asintoticamente, $\sqrt{n}(\hat{\rho} - \rho_0) \sim N(0, \varpi^2)$.

Per quanto riguarda il parametro $\varpi^2$, si può ottenere una stima attraverso l'espressione:

$$\varpi^2 = \hat{\varpi}_S^2 + \frac{n}{m}\hat{\beta}^T \Omega \hat{\beta}$$

in cui $\hat{\varpi}_S^2 = \sum_{i=1}^{n} [S(x_i; \hat{\rho})]^2$ e $\Omega$ è la matrice di varianze e covarianze di $\sqrt{n(\vartheta - \vartheta_0)}$, che nel caso in esame è $\Omega = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$ e $\hat{\beta} = \begin{bmatrix} \frac{1}{n}\hat{\sigma}_y \sum_{i=1}^{n} \phi\left(\frac{x_i - \hat{\mu}_y}{\hat{\sigma}_y}\right) \\ \frac{1}{2n\hat{\sigma}_y^2} \sum_{i=1}^{n} \phi\left(\frac{x_i - \hat{\mu}_y}{\hat{\sigma}_y}\right) \end{bmatrix}$

in cui $\phi(\cdot)$ è la funzione di ripartizione della normale standard.

Dato una livello di significatività $\alpha$ si possono calcolare i limiti dell'intervallo di confidenza per $\rho$. Essendo $\rho$ definito come la $\Pr[X > Y]$, esso deve necessariamente appartenere all'intervallo $[0,1]$.

Con il metodo appena descritto, può tuttavia accadere che i limiti dell'intervallo di confidenza cadano fuori di questo dominio e si rende necessaria un trasformazione di tipo "logit".

Si definisce, dunque, la quantità $\tau = \log\left(\frac{\rho}{1-\rho}\right)$ ed il sistema di ipotesi da verificare diventa:

$$\begin{cases} H_0 : \tau = \tau_0 \\ H_1 : \tau \neq \tau_0 \end{cases}$$

Sotto l'ipotesi $H_0$, $\hat{\tau} = \log\left(\frac{\hat{\rho}}{1-\hat{\rho}}\right)$ si distribuisce asintoticamente come una Normale $N\left(0, \frac{\varpi^2}{n\rho^2(1-\rho)^2}\right)$ e, attraverso la standardizzazione, si può ottenere una statistica distribuita come una normale $N(0,1)$. Fissando ora

un livello di significatività $\alpha$ si può ottenere un intervallo di confidenza per $\tau$. Applicando la funzione inversa della trasformata "logit" agli estremi dell'intervallo per $\tau$, si può trovare un intervallo di confidenza per $\rho$ che, per come è stato costruito, appartiene all'intervallo [0,1].

## 4.4 Il test di Wilcoxon-Mann-Whitney

Il test di Wilcoxon-Mann-Whitney è uno dei test non parametrici più potenti usato in alternativa al test parametrico t di Student, quando non sono verificate le assunzioni di base.

Questo test, chiamato anche test della somma dei ranghi, utilizza completamente l'informazione del rango. Inizialmente proposto da F. Wilcoxon nel 1945, fu generalizzato ed esteso al caso di due campioni indipendenti, con differenti numerosità, da H. B. Mann e D. R. Whitney nel 1947.

Il metodo di Wilcoxon-Mann-Whitney (WMW test) richiede che:

- le due popolazioni a confronto siano distribuite in modo continuo ed abbiano la stessa forma rispetto alla simmetria (entrambe simmetriche o entrambe asimmetriche);
- i dati siano misurati con una scala almeno ordinale.

Dati i campioni indipendenti ed identicamente distribuiti $x = (x_1,...,x_n)$ e $y = (y_1,...,y_m)$, provenienti rispettivamente dalle variabili casuali $X$ e $Y$, le ipotesi da verificare sono:

$$\begin{cases} H_0 : Me_x = Me_y \\ H_1 : Me_x \neq Me_y \end{cases}$$

dove con $Me_x$ e $Me_y$ si intendono rispettivamente le mediane del gruppo *x* e del gruppo *y*.

Inizialmente i dati appartenenti ai due gruppi vengono combinati in un'unica serie, disponendo i valori in ordine crescente ed assegnando ad ognuno un rango. Viene successivamente calcolata la somma dei ranghi, chiamata $W_n$, del gruppo con il numero di dati minore (supponiamo che $n < m$), o nel caso di uguale numerosità, si calcola indifferentemente la somma dei ranghi di uno dei due gruppi. Se l'ipotesi $H_0 : Me_x = Me_y$ è vera, i valori del gruppo prescelti sono casualmente mescolati con quelli dell'altro gruppo; il valore di $W_n$ tende allora ad una media attesa $\mu_W$, che dipende dal numero di osservazioni ( $n$ ed $m$) dei due gruppi secondo la relazione:

$$\mu_W = \frac{n \cdot (n + m + 1)}{2}.$$

Se l'ipotesi $H_0 : Me_x = Me_y$ è falsa e quindi è vera l'ipotesi alternativa, il valore di $W_n$ tende ad essere maggiore o minore del valore atteso $\mu_W$, in rapporto alla coda della distribuzione nella quale è collocata la tendenza centrale del gruppo con meno dati. Il valore di $W_n$ può tendere verso uno dei due estremi:

- un valore minimo, dato dalla somma degli $n$ ranghi minori,
- un valore massimo, determinato dalla somma degli $n$ ranghi maggiori.

La significatività della differenza tra le mediane dei due gruppi può essere valutata confrontando il valore di $W_n$ calcolato con il valore atteso $\mu_W$.

Nel caso di grandi campioni ( $n$ oppure $m > 10$) la statistica $W_n$ segue una distribuzione approssimativamente Normale e la significatività del test può essere valutata attraverso la distribuzione $N(0,1)$. Per valutare la significatività di $W_n$, si utilizza la statistica test $Z$, corretta per continuità sommando $\pm 0.5$ al valore di $W_n$, in modo tale che lo scarto tra osservato ed atteso sia minore:

$$Z = \frac{(W_n \pm 0.5) - \mu_W}{\sigma_W^2}$$

in cui:

$$\mu_W = \frac{n \cdot (n+m+1)}{2}$$

$$\sigma_W^2 = \frac{n \cdot m \cdot (n+m+1)}{12}.$$

Nel caso in cui i punteggi non sono valutati con una scala continua, come richiede il metodo, si possono avere diversi valori uguali, od osservazioni ex-aequo (ties). Nella trasformazione in ranghi, ad ognuna di queste osservazioni viene assegnata la media dei ranghi dei valori uguali. La media rimane invariata, ma la varianza $\sigma_W^2$ è minore e quindi necessita di una correzione, diventando:

$$\sigma_W^2 = \frac{n \cdot m}{N(N-1)} \cdot \left( \frac{N^3 - N}{12} - \sum_{j=1}^{g} \frac{t_j^3 - t_j}{12} \right)$$

in cui $N = n + m$ e la stima dei *t* è condotta come nel test di Wilcoxon per due campioni dipendenti.
La correzione per i ties diminuisce il valore della deviazione standard e quindi aumenta il valore della statistica *Z*.

# Capitolo 5
# GLI STUDI DI SIMULAZIONE

## 5.1 Le tabelle Test/Realtà

Mediante il metodo illustrato precedentemente si ottiene la misura che rappresenta il “False Discovery Rate”, successivamente si assegna un livello di significatività che nel caso in esame è pari a $0.05$ infine, se il FDR presenta una misura superiore al livello fissato, i geni vengono identificati come equivalentemente espressi, al contrario, se il FDR è inferiore al livello di significatività, risultano differenzialmente espressi.
A questo punto occorre valutare quali siano gli errori che si commettono nel momento in cui vengono effettuate le elaborazioni dei dati. Il rischio è che alcuni geni che sono differenzialmente espressi vengano identificati come equivalentemente espressi e che, al contrario, alcuni geni equivalentemente espressi siano riconosciuti, in modo errato, come differenzialmente espressi. Nel primo caso i geni erroneamente identificati sono detti “falsi negativi”, mentre nel secondo caso sono detti “falsi positivi”.
Gli studi di simulazione sono di grande utilità nella valutazione degli errori commessi nell’identificazione dei geni in quanto, trattandosi di dati simulati, si conosce a priori la vera natura del gene e si può dunque confrontare la realtà con i risultati proposti dall’algoritmo. I valori che si ottengono vengono confrontati con i dati reali e i risultati vengono inseriti in tabelle chiamate “Test/Realtà”, che illustrano il numero di geni correttamente ed erroneamente identificati.

| | | Realtà | |
|---|---|---|---|
| | | EE | DE |
| Test | EE | VERI NEGATIVI ($d$) | FALSI NEGATIVI ($b$) |
| | DE | FALSI POSITIVI ($c$) | VERI POSITIVI ($a$) |

**Figura -5.1- Tabella "Test/Realtà".**

La sensitività del test è la capacità di individuare geni differenzialmente espressi quando questi sono realmente differenzialmente espressi:

$$\Pr(DE_T \mid DE_R) = \frac{a}{a+c}.$$

Il rischio che si incontra nell'elaborazione dei dati è quello di identificare geni differenzialmente espressi come equivalentemente espressi, incorrendo in un errore di II° tipo $\beta$:

$$\Pr(EE_T \mid DE_R) = \frac{c}{a+c}.$$

La specificità del test è la capacità di identificare come equivalentemente espressi geni che sono realmente tali:

$$\Pr(EE_T \mid EE_R) = \frac{d}{b+d}.$$

Se il test non è specifico identifica come differenzialmente espressi geni che in realtà sono equivalentemente espressi, incorrendo in un errore di I° tipo $\alpha$:

$$\Pr(DE_T \mid EE_R) = \frac{b}{b+d}.$$

## 5.2 Analisi dei risultati ottenuti con l'algoritmo Deds mediante simulazione da modello Gamma-Gamma.

L'analisi che viene proposta ha lo scopo di illustrare la capacità dell'algoritmo Deds di identificare geni differenzialmente espressi. Per il modello Gamma-Gamma vengono simulati data set contenenti $N = 2000$ geni aventi numerosità pari a $m = 10$ , $15$ , $20$ posti sotto due differenti condizioni. I parametri utilizzati dal modello sono $(\alpha = 10, \alpha_0 = 0.9, \nu = 0.5)$ e $p = 0.2$. Il numero delle permutazioni della matrice di partenza eseguite dall'algoritmo è $B = 500$. Per ciascuna simulazione vengono calcolati i valori relativi alla specificità del test, alla sensitività e agli errori di I° e II° tipo. Ciascuna simulazione viene ripetuta dieci volte in modo tale da valutare il comportamento dell'algoritmo in media: i valori ottenuti sono riportati nelle tabelle sotto elencate.

Simulazioni dal modello Gamma-Gamma con m=10

| Simulazione | Test | Realtà | | Specificità | Errore Alfa | Sensitività | Errore Beta |
|---|---|---|---|---|---|---|---|
| | | EE | DE | | | | |
| 1 | EE | 1547 | 76 | 0,9532 | 0,0468 | 0,8125 | 0,1875 |
| | DE | 57 | 247 | | | | |
| 2 | EE | 1562 | 82 | 0,9501 | 0,0499 | 0,7973 | 0,2027 |
| | DE | 59 | 232 | | | | |
| 3 | EE | 1530 | 98 | 0,9398 | 0,0602 | 0,7720 | 0,2280 |
| | DE | 70 | 237 | | | | |
| 4 | EE | 1529 | 85 | 0,9473 | 0,0527 | 0,7477 | 0,2523 |
| | DE | 82 | 243 | | | | |
| 5 | EE | 1507 | 103 | 0,9360 | 0,0640 | 0,7432 | 0,2568 |
| | DE | 85 | 246 | | | | |
| 6 | EE | 1528 | 70 | 0,9562 | 0,0438 | 0,7768 | 0,2232 |
| | DE | 77 | 268 | | | | |
| 7 | EE | 1524 | 82 | 0,9489 | 0,0511 | 0,7679 | 0,2321 |
| | DE | 78 | 258 | | | | |
| 8 | EE | 1520 | 81 | 0,9494 | 0,0506 | 0,7582 | 0,2418 |
| | DE | 81 | 254 | | | | |
| 9 | EE | 1518 | 89 | 0,9446 | 0,0554 | 0,7726 | 0,2274 |
| | DE | 73 | 248 | | | | |
| 10 | EE | 1514 | 91 | 0,9433 | 0,0567 | 0,7665 | 0,2335 |
| | DE | 78 | 256 | | | | |
| | | Valori in media | | 0,9469 | 0,0531 | 0,7715 | 0,2285 |

**Tabella -5.1- Valori ottenuti da dieci simulazioni dal modello Gamma-Gamma con m=10.**

| Simulazioni dal modello Gamma-Gamma con m=15 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Simulazione | Test | Realtà | | Specificità | Errore Alfa | Sensitività | Errore Beta |
| | | EE | DE | | | | |
| 1 | EE | 1534 | 73 | | | | |
| | DE | 75 | 259 | 0,9546 | 0,0454 | 0,7754 | 0,2246 |
| 2 | EE | 1535 | 81 | | | | |
| | DE | 80 | 249 | 0,9499 | 0,0501 | 0,7568 | 0,2432 |
| 3 | EE | 1534 | 73 | | | | |
| | DE | 75 | 259 | 0,9546 | 0,0454 | 0,7754 | 0,2246 |
| 4 | EE | 1503 | 75 | | | | |
| | DE | 88 | 275 | 0,9525 | 0,0475 | 0,7576 | 0,2424 |
| 5 | EE | 1523 | 79 | | | | |
| | DE | 74 | 273 | 0,9507 | 0,0493 | 0,7867 | 0,2133 |
| 6 | EE | 1528 | 71 | | | | |
| | DE | 70 | 273 | 0,9556 | 0,0444 | 0,7959 | 0,2041 |
| 7 | EE | 1529 | 89 | | | | |
| | DE | 72 | 258 | 0,9450 | 0,0550 | 0,7818 | 0,2182 |
| 8 | EE | 1534 | 73 | | | | |
| | DE | 75 | 259 | 0,9546 | 0,0454 | 0,7754 | 0,2246 |
| 9 | EE | 1503 | 75 | | | | |
| | DE | 88 | 275 | 0,9525 | 0,0475 | 0,7576 | 0,2424 |
| 10 | EE | 1533 | 74 | | | | |
| | DE | 72 | 257 | 0,9540 | 0,0460 | 0,7812 | 0,2188 |
| | | Valori in media | | 0,9524 | 0,0476 | 0,7744 | 0,2256 |

**Tabella -5.2- Valori ottenuti da dieci simulazioni dal modello Gamma-Gamma con m=15.**

| Simulazioni dal modello Gamma-Gamma con m=20 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Simulazione | Test | Realtà | | Specificità | Errore Alfa | Sensitività | Errore Beta |
| | | EE | DE | | | | |
| 1 | EE | 1524 | 70 | | | | |
| | DE | 86 | 270 | 0,9561 | 0,0439 | 0,7584 | 0,2416 |
| 2 | EE | 1585 | 56 | | | | |
| | DE | 55 | 248 | 0,9659 | 0,0341 | 0,8185 | 0,1815 |
| 3 | EE | 1548 | 74 | | | | |
| | DE | 47 | 265 | 0,9544 | 0,0456 | 0,8494 | 0,1506 |
| 4 | EE | 1545 | 63 | | | | |
| | DE | 63 | 266 | 0,9608 | 0,0392 | 0,8085 | 0,1915 |
| 5 | EE | 1519 | 79 | | | | |
| | DE | 68 | 269 | 0,9506 | 0,0494 | 0,7982 | 0,2018 |
| 6 | EE | 1502 | 77 | | | | |
| | DE | 91 | 281 | 0,9512 | 0,0488 | 0,7554 | 0,2446 |
| 7 | EE | 1534 | 84 | | | | |
| | DE | 63 | 260 | 0,9481 | 0,0519 | 0,8050 | 0,1950 |
| 8 | EE | 1526 | 78 | | | | |
| | DE | 69 | 264 | 0,9514 | 0,0486 | 0,7928 | 0,2072 |
| 9 | EE | 1516 | 61 | | | | |
| | DE | 77 | 283 | 0,9613 | 0,0387 | 0,7861 | 0,2139 |
| 10 | EE | 1531 | 75 | | | | |
| | DE | 73 | 273 | 0,9533 | 0,0467 | 0,7890 | 0,2110 |
| | | Valori in media | | 0,9553 | 0,0447 | 0,7961 | 0,2039 |

**Tabella -5.3- Valori ottenuti da dieci simulazioni dal modello Gamma-Gamma con m=20.**

Osservando le tabelle sopra riportate si nota che i valori in media relativi all'indice di specificità sono sensibilmente in crescita all'aumentare della numerosità campionaria e variano nell'intervallo $[0.946,0.955]$. Per quanto riguarda la sensitività i valori variano nell'intervallo $[0.771,0.796]$ mostrando anche in questo caso una lieve crescita all'aumentare della numerosità campionaria. Nel caso in esame si potrebbe concludere che, all'aumentare della numerosità campionaria, l'algoritmo risulta più preciso.

## 5.3 Analisi dei risultati ottenuti con l'algoritmo Deds mediante simulazione da modello Log-Normale Normale.

Le simulazioni effettuate con il modello Log-Normale Normale forniscono data set di $N = 2000$ geni e numerosità pari a $m = 10, 15, 20$. I parametri utilizzati dal modello sono pari a $(\mu_0 = 2.3, \sigma = 0.3, \tau = 1.39)$ e $p = 0.2$. Il numero delle permutazioni del data set di partenza è pari a $B = 500$. Anche per quanto riguarda il modello Log-Normale Normale le simulazioni sono state ripetute dieci volte e i risultati ottenuti sono stati riportati nelle tabelle "Test/Realtà" con i relativi calcoli degli indici di specificità, sensitività ed errori di I° e II° tipo.

Come emerge dalle tabelle di seguito riportate, i valori relativi alla specificità in media variano nell'intervallo $[0.947,0.957]$, si nota quindi una sensibile crescita all'aumentare della numerosità campionaria. Per quanto riguarda la sensitività, i valori in media variano nell'intervallo $[0.746,0.772]$, anche in questo caso si osserva una lieve crescita all'aumentare della numerosità campionaria.

Come nel caso precedente si può intuire che all'aumentare della numerosità campionaria l'algoritmo tende ad essere più preciso.

| Simulazioni dal modello LogNormale Normale con m=10 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Simulazione | Test | Realtà | | Specificità | Errore Alfa | Sensitività | Errore Beta |
| | | EE | DE | | | | |
| 1 | EE | 1499 | 108 | | | | |
| | DE | 94 | 223 | 0,9328 | 0,0672 | 0,7035 | 0,2965 |
| 2 | EE | 1521 | 88 | | | | |
| | DE | 70 | 248 | 0,9453 | 0,0547 | 0,7799 | 0,2201 |
| 3 | EE | 1477 | 76 | | | | |
| | DE | 81 | 276 | 0,9511 | 0,0489 | 0,7731 | 0,2269 |
| 4 | EE | 1469 | 86 | | | | |
| | DE | 85 | 283 | 0,9447 | 0,0553 | 0,7690 | 0,2310 |
| 5 | EE | 1543 | 89 | | | | |
| | DE | 70 | 234 | 0,9455 | 0,0545 | 0,7697 | 0,2303 |
| 6 | EE | 1500 | 89 | | | | |
| | DE | 81 | 250 | 0,9440 | 0,0560 | 0,7553 | 0,2447 |
| 7 | EE | 1503 | 86 | | | | |
| | DE | 92 | 227 | 0,9459 | 0,0541 | 0,7116 | 0,2884 |
| 8 | EE | 1525 | 76 | | | | |
| | DE | 101 | 233 | 0,9525 | 0,0475 | 0,6976 | 0,3024 |
| 9 | EE | 1508 | 75 | | | | |
| | DE | 85 | 266 | 0,9526 | 0,0474 | 0,7578 | 0,2422 |
| 10 | EE | 1516 | 68 | | | | |
| | DE | 87 | 250 | 0,9571 | 0,0429 | 0,7418 | 0,2582 |
| | | Valori in media | | 0,9471 | 0,0529 | 0,7459 | 0,2541 |

**Tabella -5.4- Valori ottenuti da dieci simulazioni dal modello Log-Normale Normale con m=10.**

| Simulazioni dal modello LogNormale Normale con m=15 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Simulazione | Test | Realtà | | Specificità | Errore Alfa | Sensitività | Errore Beta |
| | | EE | DE | | | | |
| 1 | EE | 1537 | 68 | | | | |
| | DE | 71 | 238 | 0,9576 | 0,0424 | 0,7702 | 0,2298 |
| 2 | EE | 1519 | 75 | | | | |
| | DE | 77 | 253 | 0,9529 | 0,0471 | 0,7667 | 0,2333 |
| 3 | EE | 1501 | 78 | | | | |
| | DE | 88 | 258 | 0,9506 | 0,0494 | 0,7457 | 0,2543 |
| 4 | EE | 1566 | 67 | | | | |
| | DE | 76 | 242 | 0,9590 | 0,0410 | 0,7610 | 0,2390 |
| 5 | EE | 1508 | 75 | | | | |
| | DE | 80 | 261 | 0,9526 | 0,0474 | 0,7654 | 0,2346 |
| 6 | EE | 1544 | 75 | | | | |
| | DE | 77 | 231 | 0,9537 | 0,0463 | 0,7500 | 0,2500 |
| 7 | EE | 1516 | 76 | | | | |
| | DE | 79 | 255 | 0,9523 | 0,0477 | 0,7635 | 0,2365 |
| 8 | EE | 1532 | 76 | | | | |
| | DE | 85 | 253 | 0,9527 | 0,0473 | 0,7485 | 0,2515 |
| 9 | EE | 1544 | 64 | | | | |
| | DE | 75 | 236 | 0,9602 | 0,0398 | 0,7588 | 0,2412 |
| 10 | EE | 1536 | 70 | | | | |
| | DE | 58 | 257 | 0,9564 | 0,0436 | 0,8159 | 0,1841 |
| | | Valori in media | | 0,9548 | 0,0452 | 0,7646 | 0,2354 |

**Tabella -5.5- Valori ottenuti da dieci simulazioni dal modello Log-Normale Normale con m=15.**

| Simulazioni dal modello LogNormale Normale con m=20 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Simulazione | Test | Realtà | | Specificità | Errore Alfa | Sensitività | Errore Beta |
| | | EE | DE | | | | |
| 1 | EE | 1541 | 57 | | | | |
| | DE | 78 | 246 | 0,9643 | 0,0357 | 0,7593 | 0,2407 |
| 2 | EE | 1533 | 72 | | | | |
| | DE | 78 | 249 | 0,9551 | 0,0449 | 0,7615 | 0,2385 |
| 3 | EE | 1508 | 72 | | | | |
| | DE | 79 | 276 | 0,9544 | 0,0456 | 0,7775 | 0,2225 |
| 4 | EE | 1516 | 73 | | | | |
| | DE | 77 | 246 | 0,9541 | 0,0459 | 0,7616 | 0,2384 |
| 5 | EE | 1516 | 61 | | | | |
| | DE | 77 | 274 | 0,9613 | 0,0387 | 0,7806 | 0,2194 |
| 6 | EE | 1498 | 73 | | | | |
| | DE | 73 | 260 | 0,9535 | 0,0465 | 0,7808 | 0,2192 |
| 7 | EE | 1512 | 70 | | | | |
| | DE | 76 | 263 | 0,9558 | 0,0442 | 0,7758 | 0,2242 |
| 8 | EE | 1553 | 60 | | | | |
| | DE | 80 | 233 | 0,9628 | 0,0372 | 0,7444 | 0,2556 |
| 9 | EE | 1524 | 72 | | | | |
| | DE | 62 | 277 | 0,9549 | 0,0451 | 0,8171 | 0,1829 |
| 10 | EE | 1505 | 71 | | | | |
| | DE | 85 | 267 | 0,9549 | 0,0451 | 0,7585 | 0,2415 |
| | | | Valori in media | 0,9571 | 0,0429 | 0,7717 | 0,2283 |

**Tabella -5.6- Valori ottenuti da dieci simulazioni dal modello Log-Normale Normale con m=20.**

Confrontando i due modelli utilizzati per le simulazioni (Gamma-Gamma e Log-Normale Normale) possiamo osservare che i valori relativi a specificità e sensitività sono tra loro molto simili e abbastanza stabili. Con questo tipo di analisi si può quindi affermare che l'algoritmo Deds è abbastanza robusto.

## 5.4 Confronto tra i risultati ottenuti mediante l'algoritmo Deds e le singole statistiche.

L'algoritmo Deds presentato in questo lavoro si basa sulla sintesi di singole statistiche, che stimano le stesse quantità di interesse e si propone di superare, attraverso un metodo meno rigoroso, le limitazioni che le singole statistiche possiedono. Si propone quindi un confronto tra le tabelle "Test/Realtà" ottenute mediante l'algoritmo Deds e le tabelle

ottenute considerando singolarmente i test t-Student, Semiparametrico e Wilcoxon.

Le tabelle sotto elencate si riferiscono allo studio di simulazione effettuato attraverso il modello Gamma-Gamma con numerosità pari a $m = 10,15,20$.

| Simulazione dal modello Gamma-Gamma con m=10 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Deds | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1547 | 76 | Sensitività= | 0,7584 | Specificità= | 0,9561 |
| | DE | 57 | 247 | beta= | 0,2416 | alfa= | 0,0439 |

| Simulazione dal modello Gamma-Gamma con m=10 | | | | | | | |
|---|---|---|---|---|---|---|---|
| t-test | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1534 | 44 | Sensitività= | 0,7994 | Specificità= | 0,9721 |
| | DE | 70 | 279 | beta= | 0,2006 | alfa= | 0,0279 |

| Simulazione dal modello Gamma-Gamma con m=10 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Semiparametrico | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1542 | 52 | Sensitività= | 0,8138 | Specificità= | 0,9674 |
| | DE | 62 | 271 | beta= | 0,1862 | alfa= | 0,0326 |

| Simulazione dal modello Gamma-Gamma con m=10 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Wilcoxon | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1546 | 51 | Sensitività= | 0,8242 | Specificità= | 0,9681 |
| | DE | 58 | 272 | beta= | 0,1758 | alfa= | 0,0319 |

**Tabella -5.7- Confronto tra l'algoritmo Deds e le singole statistiche mediante simulazione da modello Gamma-Gamma e numerosità m=10.**

Nella tabella -5.7- viene proposto il confronto dei risultati ottenuti dall'algoritmo Deds e le singole statistiche nella simulazione con modello Gamma-Gamma e numerosità $m = 10$. Come si può notare i valori dell'indice di sensitività variano nell'intervallo $[0.758, 0.824]$; il valore più elevato corrisponde al test di Wilcoxon mentre il più basso corrisponde all'algoritmo Deds. Per quanto riguarda invece l'indice di specificità, si può

notare che esso varia nell'intervallo $[0.956,0.972]$, il valore più elevato corrisponde al t-test, mentre il più basso è relativo all'algoritmo Deds.

| Simulazione dal modello Gamma-Gamma con m=15 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Deds | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1534 | 73 | Sensitività= | 0,7754 | Specificità= | 0,9546 |
| | DE | 75 | 259 | beta= | 0,2246 | alfa= | 0,0454 |
| Simulazione dal modello Gamma-Gamma con m=15 | | | | | | | |
| t-test | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1536 | 34 | Sensitività= | 0,8032 | Specificità= | 0,9783 |
| | DE | 73 | 298 | beta= | 0,1968 | alfa= | 0,0217 |
| Simulazione dal modello Gamma-Gamma con m=15 | | | | | | | |
| Semiparametrico | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1543 | 36 | Sensitività= | 0,8177 | Specificità= | 0,9772 |
| | DE | 66 | 296 | beta= | 0,1823 | alfa= | 0,0228 |
| Simulazione dal modello Gamma-Gamma con m=15 | | | | | | | |
| Wilcoxon | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1545 | 36 | Sensitività= | 0,8222 | Specificità= | 0,9772 |
| | DE | 64 | 296 | beta= | 0,1778 | alfa= | 0,0228 |

**Tabella -5.8- Confronto tra l'algoritmo Deds e le singole statistiche mediante simulazione da modello Gamma-Gamma e numerosità m=15.**

Nella tabella -5.8- è riportato il confronto tra i risultati ottenuti dall'algoritmo Deds e l'analisi effettuata dalle singole statistiche mediante una simulazione con il modello Gamma-Gamma con numerosità $m=15$. Dall'analisi emerge che l'indice di sensitività varia nell'intervallo $[0.775,0.822]$ in cui si nota che il valore maggiore corrisponde al test di Wilcoxon e il minore all'algoritmo Deds. L'indice di specificità varia invece

nell'intervallo $[0.955,0.978]$: in questo caso i valori di tutte e tre le singole statistiche sono molto vicini e maggiori rispetto all'algoritmo Deds.

| Simulazione dal modello Gamma-Gamma con m=20 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Deds | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1524 | 70 | Sensitività= | 0,8125 | Specificità= | 0,9532 |
| | DE | 86 | 270 | beta= | 0,1875 | alfa= | 0,0468 |

| Simulazione dal modello Gamma-Gamma con m=20 | | | | | | | |
|---|---|---|---|---|---|---|---|
| t-test | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1531 | 34 | Sensitività= | 0,7948 | Specificità= | 0,9783 |
| | DE | 79 | 306 | beta= | 0,2052 | alfa= | 0,0217 |

| Simulazione dal modello Gamma-Gamma con m=20 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Semiparametrico | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1535 | 37 | Sensitività= | 0,8016 | Specificità= | 0,9765 |
| | DE | 75 | 303 | beta= | 0,1984 | alfa= | 0,0235 |

| Simulazione dal modello Gamma-Gamma con m=20 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Wilcoxon | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1528 | 38 | Sensitività= | 0,7865 | Specificità= | 0,9757 |
| | DE | 82 | 302 | beta= | 0,2135 | alfa= | 0,0243 |

**Tabella -5.9- Confronto tra l'algoritmo Deds e le singole statistiche mediante simulazione da modello Gamma-Gamma e numerosità m=20.**

La tabella -5.9- indica i risultati ottenuti con la simulazione effettuata mediante modello Gamma-Gamma con numerosità $m = 20$. In questo caso si assiste ad una inversione di tendenza per quanto riguarda l'indice di sensitività che varia nell'intervallo $[0.786,0.812]$: questa volta il valore maggiore corrisponde all'algoritmo Deds, mentre il valore più basso corrisponde al test di Wilcoxon. Per quanto riguarda invece l'indice di specificità le tre singole statistiche hanno valori maggiori e molto simili tra

loro e superano ancora una volta il valore corrispondente all'algoritmo Deds: i valori relativi a questo indicatore variano nell'intervallo $[0.953,0.978]$.

Analizzando in un'ottica più generale i risultati ottenuti si può notare come in tutti i casi considerati, all'aumentare della numerosità campionaria si ottiene un miglioramento per quanto riguarda l'indice di specificità, invece non si trovano collegamenti tra le variazioni dell'indice di sensitività e la numerosità campionaria.

I confronti tra le performance dell'algoritmo Deds e le singole statistiche sono stati effettuati anche per le simulazioni che utilizzano il modello Log-Normale Normale con numerosità pari a $m = 10,15,20$.

| Simulazione da modello LogNormale Normale con m=10 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Deds | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1513 | 50 | Sensitività= | 0,7784 | Specificità= | 0,9680 |
| | DE | 80 | 281 | beta= | 0,2216 | alfa= | 0,0320 |
| Simulazione da modello LogNormale Normale con m=10 | | | | | | | |
| t-test | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1513 | 50 | Sensitività= | 0,7784 | Specificità= | 0,9680 |
| | DE | 80 | 281 | beta= | 0,2216 | alfa= | 0,0320 |
| Simulazione da modello LogNormale Normale con m=10 | | | | | | | |
| Semiparametrico | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1531 | 51 | Sensitività= | 0,8187 | Specificità= | 0,9678 |
| | DE | 62 | 280 | beta= | 0,1813 | alfa= | 0,0322 |
| Simulazione da modello LogNormale Normale con m=10 | | | | | | | |
| Wilcoxon | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1515 | 51 | Sensitività= | 0,7821 | Specificità= | 0,9674 |
| | DE | 78 | 280 | beta= | 0,2179 | alfa= | 0,0326 |

**Tabella -5.10- Confronto tra l'algoritmo Deds e le singole statistiche mediante simulazione da modello Log-Normale Normale e numerosità m=10.**

La tabella -5.10- illustra i valori che si sono ottenuti dalla simulazione con numerosità $m=10$. Si nota che in questa simulazione il numero di geni identificati dall'algoritmo Deds è uguale al numero di geni identificati dal test t-Student e di conseguenza anche i valori degli indici di sensitività e specificità sono uguali tra loro. Per quanto riguarda l'indice di sensitività si osserva che varia nell'intervallo $[0.778,0.818]$, il valore maggiore si riferisce al test Semiparametrico mentre i più bassi all'algoritmo Deds e al test t-Student. Per quanto riguarda la specificità si nota che in tutti i casi rimane pressoché stabile con un valore pari a circa $0.96$.

| Simulazione da modello LogNormale-Normale con m=15 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Deds | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1537 | 68 | Sensitività= | 0,7702 | Specificità= | 0,9576 |
| | DE | 71 | 238 | beta= | 0,2298 | alfa= | 0,0424 |

| Simulazione da modello LogNormale-Normale con m=15 | | | | | | | |
|---|---|---|---|---|---|---|---|
| t-test | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1513 | 31 | Sensitività= | 0,7432 | Specificità= | 0,9799 |
| | DE | 95 | 275 | beta= | 0,2568 | alfa= | 0,0201 |

| Simulazione da modello LogNormale-Normale con m=15 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Semiparametrico | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1527 | 32 | Sensitività= | 0,7718 | Specificità= | 0,9795 |
| | DE | 81 | 274 | beta= | 0,2282 | alfa= | 0,0205 |

| Simulazione da modello LogNormale-Normale con m=15 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Wilcoxon | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1534 | 33 | Sensitività= | 0,7867 | Specificità= | 0,9789 |
| | DE | 74 | 273 | beta= | 0,2133 | alfa= | 0,0211 |

**Tabella -5.11- Confronto tra l'algoritmo Deds e le singole statistiche mediante simulazione da modello Log-Normale Normale e numerosità m=15.**

Nella tabella -5.11- si illustra il confronto nella simulazione effettuata mediante il modello Log-Normale Normale con numerosità $m=15$. I risultati ottenuti non mostrano sostanziali differenze rispetto ai casi precedenti. Per quanto riguarda la sensitività essa varia nell'intervallo $[0.743,0.786]$ il valore più elevato corrisponde al test di Wilcoxon mentre il valore minore corrisponde al test t-Student. L'indice di specificità rimane pressoché stabile per le tre singole statistiche, invece è minore il valore corrispondente all'algoritmo Deds ed i valori variano nell'intervallo $[0.958,0.98]$.

| Simulazione da modello LogNormale-Normale con m=20 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Deds | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1541 | 57 | Sensitività= | 0,7593 | Specificità= | 0,9643 |
| | DE | 78 | 246 | beta= | 0,2407 | alfa= | 0,0357 |

| Simulazione da modello LogNormale-Normale con m=20 | | | | | | | |
|---|---|---|---|---|---|---|---|
| t-test | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1535 | 27 | Sensitività= | 0,7667 | Specificità= | 0,9827 |
| | DE | 84 | 276 | beta= | 0,2333 | alfa= | 0,0173 |

| Simulazione da modello LogNormale-Normale con m=20 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Semiparametrico | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1542 | 30 | Sensitività= | 0,7800 | Specificità= | 0,9809 |
| | DE | 77 | 273 | beta= | 0,2200 | alfa= | 0,0191 |

| Simulazione da modello LogNormale-Normale con m=20 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Wilcoxon | | Realtà | | | | | |
| | | EE | DE | | | | |
| Test | EE | 1543 | 29 | Sensitività= | 0,7829 | Specificità= | 0,9816 |
| | DE | 76 | 274 | beta= | 0,2171 | alfa= | 0,0184 |

**Tabella -5.12- Confronto tra l'algoritmo Deds e le singole statistiche mediante simulazione da modello Log-Normale Normale e numerosità m=20.**

La tabella -5.12- illustra i risultati che si sono ottenuti dalla simulazione effettuata mediante il modello Log-Normale Normale con numerosità pari a $m = 20$. Per quanto riguarda gli indici di sensitività, si nota che essi variano nell'intervallo $[0.759, 0.782]$, il valore più elevato corrisponde al test di Wilcoxon mentre il più basso corrisponde all'algoritmo Deds. La specificità varia invece in $[0.964, 0.983]$ e si osserva che il valore maggiore si riferisce al test t-Student, mentre invece il valore minore appartiene all'algoritmo Deds.
In un'ottica più generale, analizzando i valori ottenuti da questa simulazione, si può notare come per l'indice di sensitività non ci sia una relazione con la numerosità campionaria, cosa che invece accade per quanto riguarda la specificità che vede in tutti i casi un miglioramento all'aumentare del campione.

## 5.5 Confronto dei risultati ottenuti tramite simulazioni da distribuzioni con parametri diversi.

L'analisi che viene proposta ha lo scopo di osservare le performance dell'algoritmo Deds nelle simulazioni effettuate mediante i modelli Gamma-Gamma e Log-Normale Normale con parametri diversi.
In particolare si vuole confrontare il modello Gamma-Gamma nel primo caso con parametri $(\alpha = 10, \alpha_0 = 0.9, \nu = 0.5)$, mentre nel secondo caso il modello Gamma-Gamma con parametri $(\alpha = 1, \alpha_0 = 1.1, \nu = 45.4)$. Per entrambi si considera la numerosità pari a $m = 10,15,20$. I grafici riportati nelle figure sotto indicate mostrano i geni simulati come equivalentemente o differenzialmente espressi mediante il modello Gamma-Gamma.
Nel primo caso i parametri utilizzati sono $(\alpha = 10, \alpha_0 = 0.9, \nu = 0.5)$, nel secondo caso sono invece $(\alpha = 1, \alpha_0 = 1.1, \nu = 45.4)$ e la numerosità è pari a $m = 10$. L'ascissa di ogni punto corrisponde alla media delle $m$ replicazioni per il singolo gene posto sotto la prima condizione, mentre l'ordinata è la media delle $m$ replicazioni di geni posti sotto la seconda condizione.

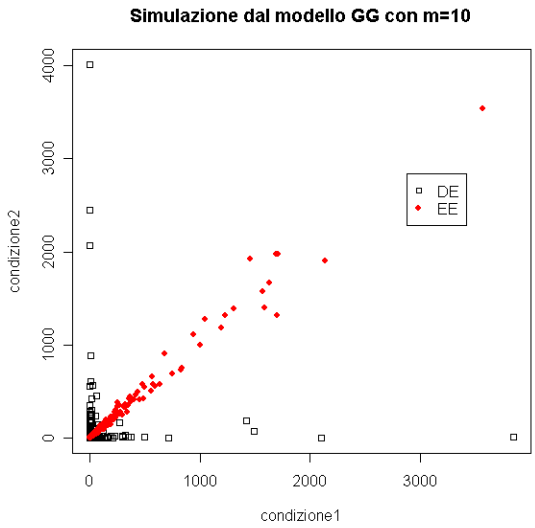



**Figura -5.2- Specificazione dell'espressione genica nel modello Gamma-Gamma simulato nel caso 1 con m=10.**

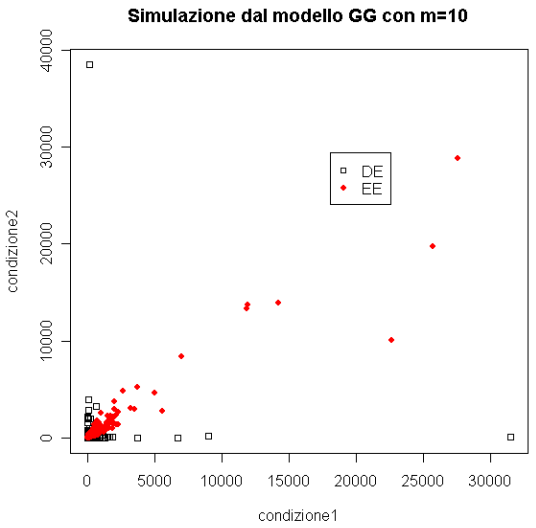



**Figura -5.3- Specificazione dell'espressione genica nel modello Gamma-Gamma simulato nel caso 2 con m=10.**

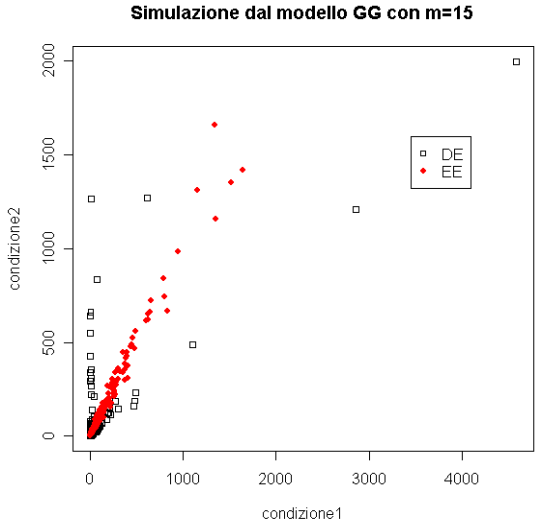



**Figura -5.4- Specificazione dell'espressione genica nel modello Gamma-Gamma simulato nel caso 1 con m=15.**

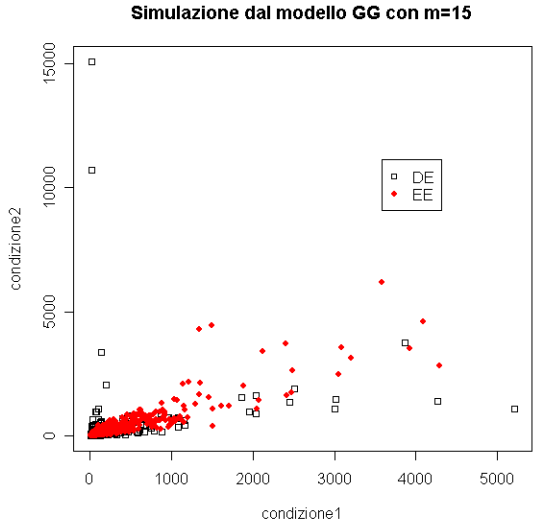



**Figura -5.5- Specificazione dell'espressione genica nel modello Gamma-Gamma simulato nel caso 2 con m=15.**

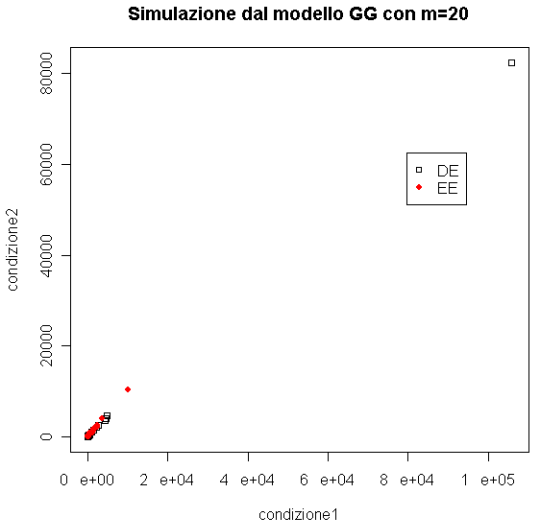



**Figura -5.6- Specificazione dell'espressione genica nel modello Gamma-Gamma simulato nel caso 1 con m=20.**

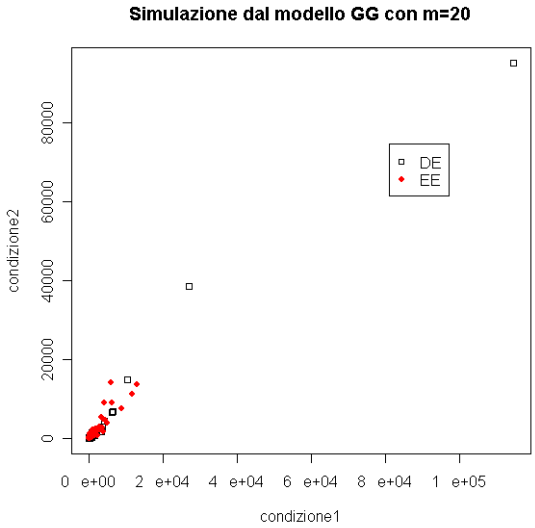



**Figura -5.7- Specificazione dell'espressione genica nel modello Gamma-Gamma simulato nel caso 2 con m=20.**

Sulla base di queste simulazioni si vogliono confrontare le performance dell'algoritmo Deds e si riportano nelle seguenti tabelle "Test/Realtà" i valori correttamente identificati ed erroneamente identificati.

CASO 1

Simulazione da modello Gamma Gamma con m=10

| | | Realtà | | | |
|---|---|---|---|---|---|
| | | EE | DE | | |
| Test | EE | 1547 | 76 | Specificità= | 0,9532 |
| | DE | 57 | 247 | alfa= | 0,0468 |
| | | | | Sensitività= | 0,8125 |
| | | | | beta= | 0,1875 |

Simulazione da modello Gamma Gamma con m=15

| | | Realtà | | | |
|---|---|---|---|---|---|
| | | EE | DE | | |
| Test | EE | 1534 | 73 | Specificità= | 0,9546 |
| | DE | 75 | 259 | alfa= | 0,0454 |
| | | | | Sensitività= | 0,7754 |
| | | | | beta= | 0,2246 |

Simulazione da modello Gamma Gamma con m=20

| | | Realtà | | | |
|---|---|---|---|---|---|
| | | EE | DE | | |
| Test | EE | 1524 | 70 | Specificità= | 0,9561 |
| | DE | 86 | 270 | alfa= | 0,0439 |
| | | | | Sensitività= | 0,7584 |
| | | | | beta= | 0,2416 |

CASO 2

Simulazione da modello Gamma Gamma con m=10

| | | Realtà | | | |
|---|---|---|---|---|---|
| | | EE | DE | | |
| Test | EE | 1535 | 315 | Specificità= | 0,8297 |
| | DE | 61 | 85 | alfa= | 0,1703 |
| | | | | Sensitività= | 0,5822 |
| | | | | beta= | 0,4178 |

Simulazione da modello Gamma Gamma con m=15

| | | Realtà | | | |
|---|---|---|---|---|---|
| | | EE | DE | | |
| Test | EE | 1524 | 282 | Specificità= | 0,8439 |
| | DE | 64 | 129 | alfa= | 0,1561 |
| | | | | Sensitività= | 0,6684 |
| | | | | beta= | 0,3316 |

Simulazione da modello Gamma Gamma con m=20

| | | Realtà | | | |
|---|---|---|---|---|---|
| | | EE | DE | | |
| Test | EE | 1508 | 304 | Specificità= | 0,8322 |
| | DE | 68 | 114 | alfa= | 0,1678 |
| | | | | Sensitività= | 0,6264 |
| | | | | beta= | 0,3736 |

**Tabella -5.13- Tabelle "Test/Realtà" e relativi indicatori sul modello Gamma-Gamma.**

Esaminando la tabella -5.13- emerge che nel primo caso l'indice di specificità aumenta sensibilmente all'aumentare della numerosità campionaria, invece l'indice di sensitività diminuisce. Per quanto riguarda il secondo caso i valori più elevati di specificità e sensitività si hanno con numerosità pari a $m = 15$. Nel primo caso l'errore $\alpha$ rimane abbastanza stabile pari a circa $0.04$ mentre nel secondo caso varia nell'intervallo $[0.156, 0.170]$, per quanto riguarda l'errore $\beta$ nel primo caso varia nell'intervallo $[0.187, 0.242]$ e si assiste ad un incremento nel secondo caso con una variazione tra $[0.331, 0.417]$.

Per quanto riguarda il modello Log-Normale Normale si vogliono confrontare le performance dell'algoritmo Deds nell'identificazione di geni provenienti da un modello Log-Normale Normale con parametri pari a $(\mu_0 = 2.3, \sigma = 0.3, \tau = 1.39)$ nel primo caso e con parametri $(\mu_0 = 6.58, \sigma = 0.9, \tau = 1.13)$ nel secondo caso, con numerosità per entrambi pari a $m = 10, 15, 20$.

Le figure sotto riportate rappresentano i geni che sono stati simulati come equivalentemente o differenzialmente espressi mediante il modello Log-Normale Normale.

L'ascissa di ogni punto corrisponde alla media delle $m$ replicazioni per il singolo gene sotto la prima condizione, invece l'ordinata corrisponde alla media delle $m$ replicazioni per il singolo gene sotto la seconda condizione.

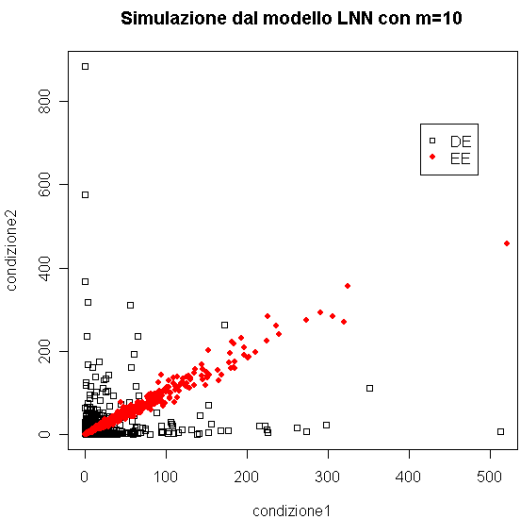



**Figura -5.8- Specificazione dell'espressione genica nel modello Log-Normale Normale simulato nel caso 1 con m=10.**

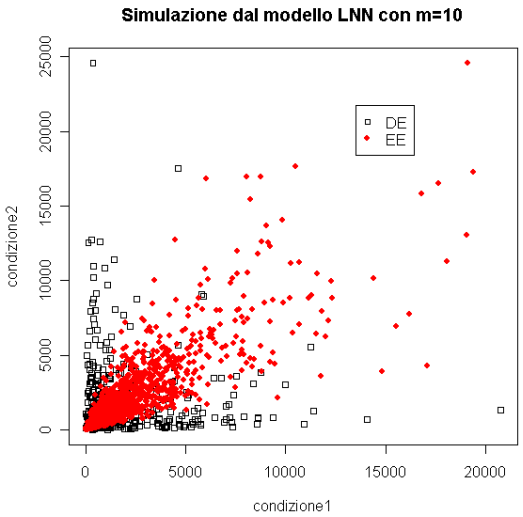



**Figura -5.9- Specificazione dell'espressione genica nel modello Log-Normale Normale simulato nel caso 2 con m=10.**

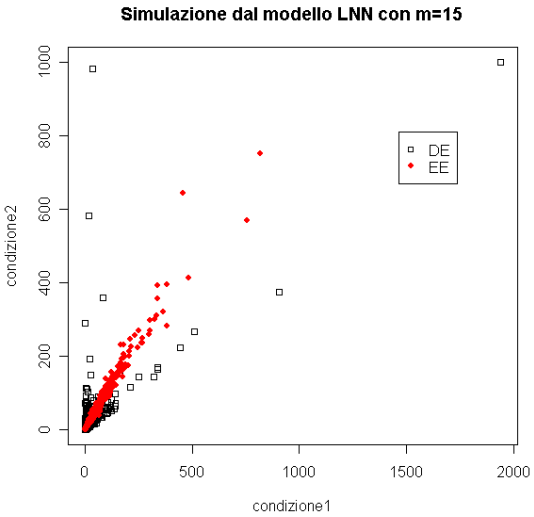



**Figura -5.10- Specificazione dell'espressione genica nel modello Log-Normale Normale simulato nel caso 1 con m=15.**

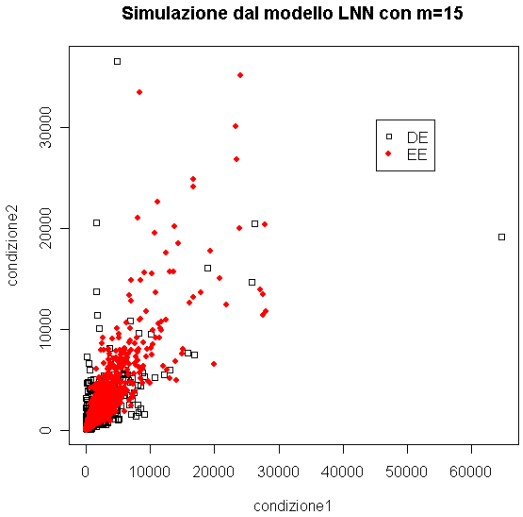



**Figura -5.11- Specificazione dell' espressione genica nel modello Log-Normale Normale simulato nel caso 2 con m=15.**

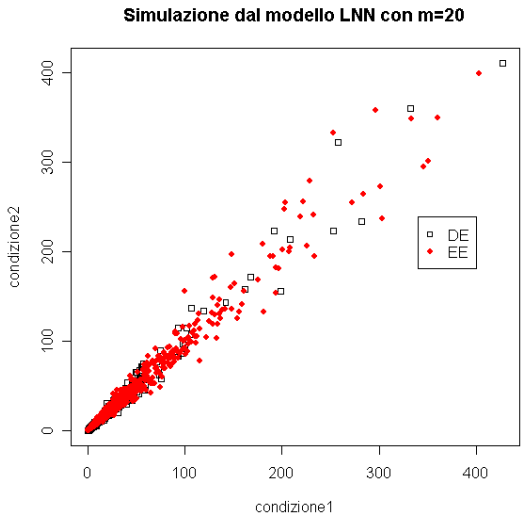



**Figura -5.12- Specificazione dell'espressione genica nel modello Log-Normale Normale simulato nel caso 1 con m=15.**

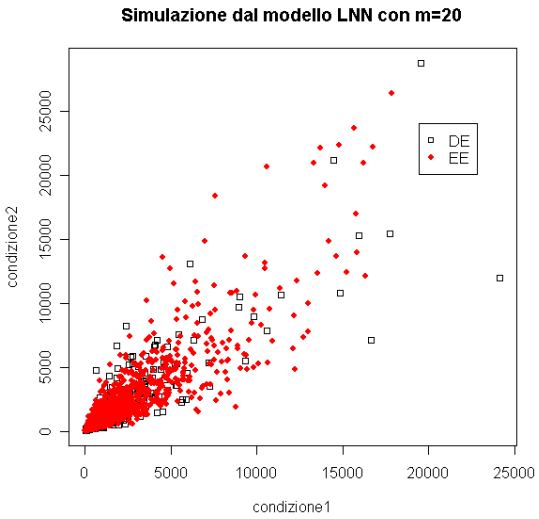



**Figura -5.13- Specificazione dell'espressione genica nel modello Log-Normale Normale simulato nel caso 1 con m=15.**

Sulla base di queste simulazioni vengono riportate le tabelle "Test/Realtà" relative alla corretta identificazione dei geni, con i relativi calcoli degli indicatori.

CASO 1

Simulazione da modello LogNormale Normale con m=10

| | | Realtà | | | |
|---|---|---|---|---|---|
| | | EE | DE | Specificità= | 0,9328 |
| | | | | alfa= | 0,0672 |
| Test | EE | 1499 | 108 | Sensitività= | 0,7035 |
| | DE | 94 | 223 | beta= | 0,2965 |

Simulazione da modello LogNormale-Normale con m=15

| | | Realtà | | | |
|---|---|---|---|---|---|
| | | EE | DE | Specificità= | 0,9576 |
| | | | | alfa= | 0,0424 |
| Test | EE | 1537 | 68 | Sensitività= | 0,7702 |
| | DE | 71 | 238 | beta= | 0,2298 |

Simulazione da modello LogNormale-Normale con m=20

| | | Realtà | | | |
|---|---|---|---|---|---|
| | | EE | DE | Specificità= | 0,9643 |
| | | | | alfa= | 0,0357 |
| Test | EE | 1541 | 57 | Sensitività= | 0,7593 |
| | DE | 78 | 246 | beta= | 0,2407 |

CASO 2

Simulazione da modello LogNormale Normale con m=10

| | | Realtà | | | |
|---|---|---|---|---|---|
| | | EE | DE | Specificità= | 0,8318 |
| | | | | alfa= | 0,1682 |
| Test | EE | 1513 | 306 | Sensitività= | 0,5625 |
| | DE | 77 | 99 | beta= | 0,4375 |

Simulazione da modello LogNormale-Normale con m=15

| | | Realtà | | | |
|---|---|---|---|---|---|
| | | EE | DE | Specificità= | 0,8181 |
| | | | | alfa= | 0,1819 |
| Test | EE | 1498 | 333 | Sensitività= | 0,6108 |
| | DE | 65 | 102 | beta= | 0,3892 |

Simulazione da modello LogNormale-Normale con m=20

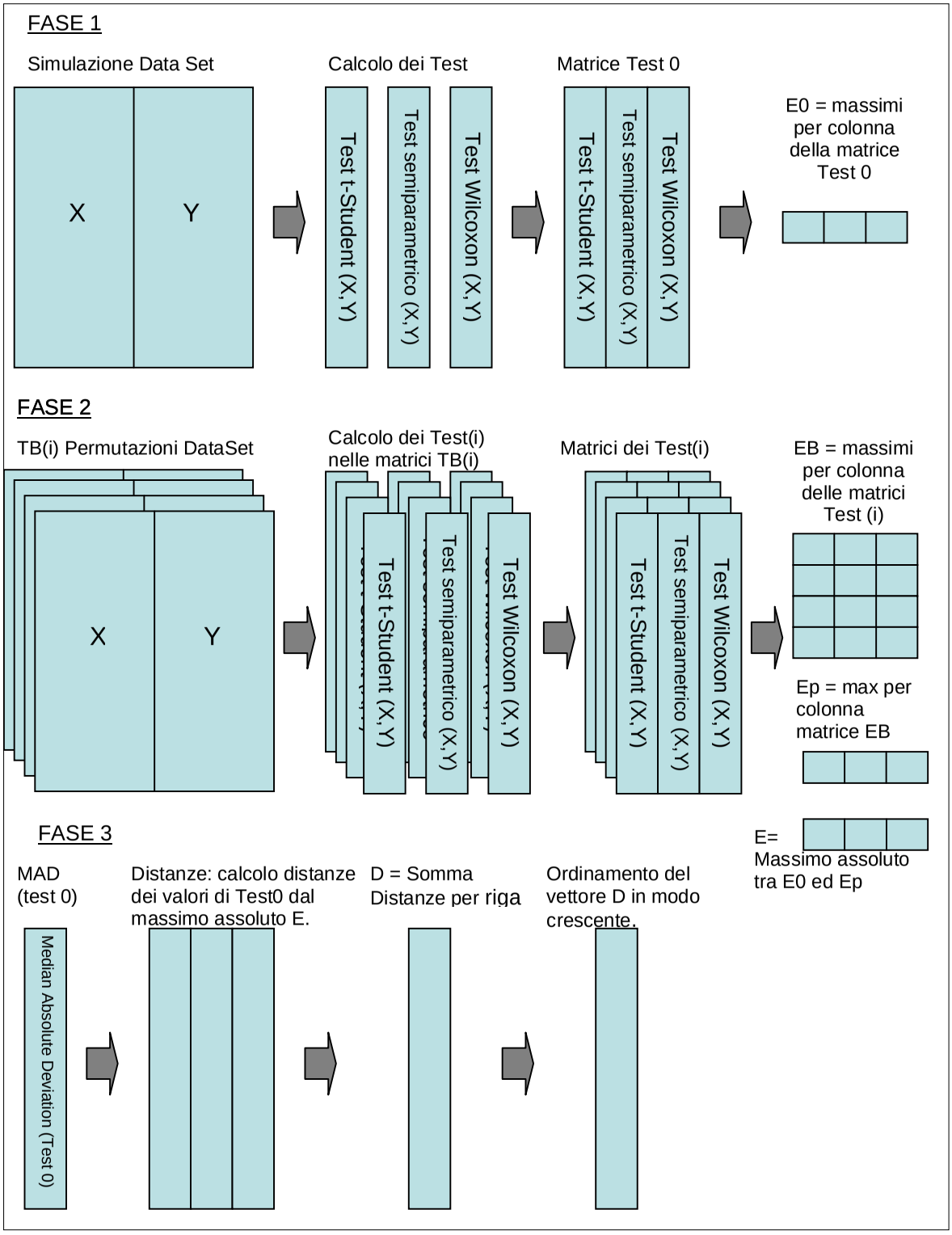
| | | Realtà | | | |
|---|---|---|---|---|---|
| | | EE | DE | Specificità= | 0,8122 |
| | | | | alfa= | 0,1878 |
| Test | EE | 1488 | 344 | Sensitività= | 0,5663 |
| | DE | 72 | 94 | beta= | 0,4337 |

**Tabella -5.14- Tabelle "Test/Realtà" e relativi indicatori sul modello Log-Normale Normale.**

Dall’analisi della tabella -5.14- si evince che per quanto riguarda il primo caso la specificità cresce all’aumentare della numerosità campionaria e varia nell’intervallo $[0.932,0.964]$, la sensitività ha invece un andamento irregolare che non dipende dal campione considerato, il valore maggiore risulta $0.770$ e corrisponde a $m=15$. Nel secondo caso la specificità diminuisce all’aumentare della numerosità campionaria, variando nell’intervallo $[0.812,0.831]$ e come nel caso precedente non emergono relazioni per quanto riguarda la sensitività, si evidenzia il valore maggiore pari a $0.611$ e corrispondente a $m=15$.

# Appendice

## Schema relativo all'algoritmo Deds.



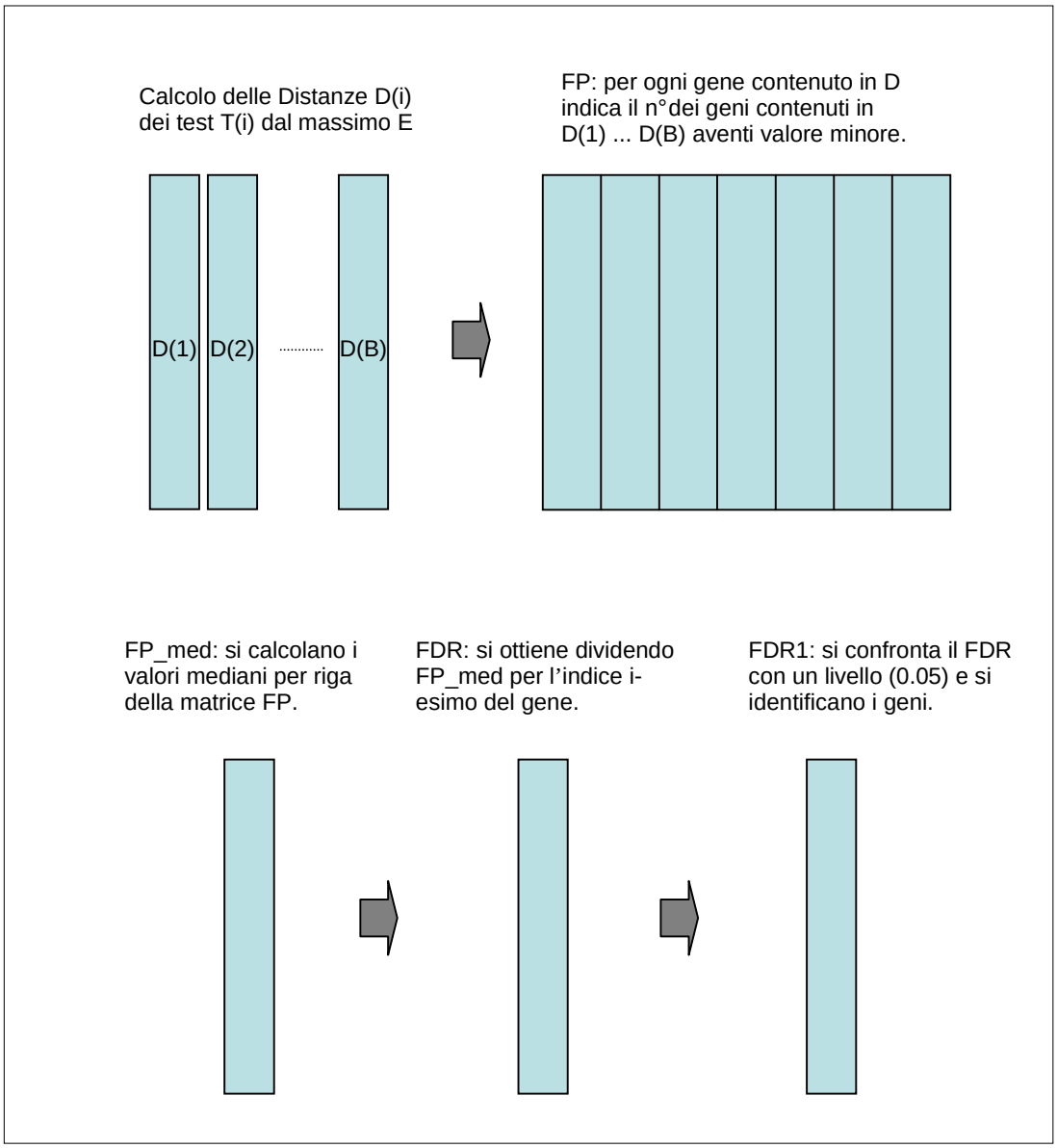



**Figura 1- Lo schema mostra le varie fasi dell'algoritmo DEDS.**

Codice R relativo all’algoritmo Deds con simulazione da modello Gamma-Gamma.

```
dedsGG<-function(m=15, N=2000, B=10000, alpha=10, alpha0=0.9,
nu=0.5, p=0.2){

library(permax)

#m = numerosità del campione
#N = geni simulati
#B = numero delle permutazioni del dataset simulato
#(alpha,alpha0,nu) = parametri del modello Gamma-Gamma

# SIMULAZIONE DEL DATASET #

Sim_GG<-function (N=N, m=m, alpha=alpha, alpha0=alpha0, nu=nu,
p=p){

ncond<-2
matrice<-matrix(rep(NA,N*ncond*m),ncol=ncond*m,byrow=T)

DE<-rep(FALSE,N)
for(i in 1:N){
      if(runif(1)>p){
            lambda<-rgamma(1,shape=alpha0,rate=nu)
            matrice[i,]<-rgamma(m*ncond, shape=alpha, rate=lambda)
      }
      else{
            lambda1<-rgamma(1, shape=alpha0, rate=nu)
            lambda2<-rgamma(1, shape=alpha0, rate=nu)
            cond1<-rgamma(m, shape=alpha, rate=lambda1)
            cond2<-rgamma(m, shape=alpha, rate=lambda2)
            matrice[i,]<-c(cond1,cond2)
            DE[i]<-TRUE
      }
      }
list(matrice=matrice, DE=DE)
      }


Simulazione<-Sim_GG(N=N, m=m, alpha=alpha, alpha0=alpha0, nu=nu,
p=p)
matrice<-Simulazione$matrice
DE<-Simulazione$DE
write.table(matrice, file="c:/tabelle tesi/matrice.txt",
row.names=F, col.names=F)
write.table(DE, file="c:/tabelle tesi/DE.txt", row.names=F,
col.names=F)

# FASE 1 #

# Calcolo delle J statistiche nel dataset simulato: all'interno del
# ciclo "for" viene calcolato il test t-student e il test di
# wilcoxon, successivamente si richiama la funzione
```

```
# semiparametric.test, che permette il calcolo del test
# semiparametrico. Si calcolano infine i valori massimi delle J
# statistiche.

x<-matrice[,1:m]
y<-matrice[,(m+1):(m*2)]

stat.t<-rep(0,N)
stat.wilc<-rep(0,N)
stat.sem<-rep(0,N)
p_t<-rep(0,N)
p_wilc<-rep(0,N)
p_sem<-rep(0,N)

      for(i in 1:N){
      stat.t[i]<-t.test(x[i,],y[i,],var.equal=TRUE)$statistic
      p_t[i]<-t.test(x[i,],y[i,],var.equal=TRUE)$p.value
      stat.wilc[i]<-wilcox.test(x[i,],y[i,])$statistic
      p_wilc[i]<-wilcox.test(x[i,],y[i,])$p.value
      }
stat.t<-matrix(abs(stat.t), ncol=1, byrow=TRUE)
stat.wilc<-matrix(abs(stat.wilc), ncol=1, byrow=TRUE)

sem<-semiparametric.test(x,y)
stat.sem<-sem$test
p_sem<-sem$pvalue
stat.sem<-matrix(abs(stat.sem), ncol=1, byrow=TRUE)

p_value<-cbind(p_t,p_sem,p_wilc)
p_value<-na.omit(p_value)

write.table(p_value, file="c:/tabelle tesi/p_value.txt",
row.names=F, col.names=F)

J<-ncol(p_value)

Test_na<-cbind(stat.t, stat.sem, stat.wilc)

id<-rep(FALSE,N)
      for(i in 1:N){
            if(any(is.na(Test_na[i,]))==TRUE)
            id[i]<-TRUE
            }
id<-matrix(id, ncol=1)      # vettore che indica con TRUE le righe
                            # della matrice che producono valori
                            # Na nei test.

DE1<-DE[id==FALSE]          # vettore di geni differenzialmente
                            # espressi senza i valori
                            # corrispondenti a Na.

write.table(id, file="c:/tabelle tesi/id.txt", row.names=F,
col.names=F)

Test0<-na.omit(Test_na)   # vengono omesse le righe contenenti Na.
percentuale0<-(N-nrow(Test0))*100/N   # percentuale di valori Na
                                      # sul totale
E0<-apply(Test0,2,max)
```

```
write.table(Test0, file="c:/tabelle tesi/Test0.txt", row.names=F,
col.names=F)
# write.table(E0, file="c:/tabelle tesi/E0.txt", row.names=F,
col.names=F)
# write.table(percentuale0, file="c:/tabelle tesi/perc0.txt",
row.names=F, col.names=F)

matrice1<-matrice[id==FALSE,]

# FASE 2 #

# Si ripete il calcolo dei test da nuovi dataset, ottenuti tramite
# B permutazioni del dataset "matrice 1". Questa matrice si ottiene
# simulando i dati e, dopo il calcolo dei test, omettendo le righe
# contenenti valori Na. Si calcolano successivamente i
# massimi delle J statistiche, e tra tutti i massimi si ricalcolano
# nuovamente i valori massimi, ottenendo in tal modo il massimo
# assoluto E.

filepath <- "c:/tabelle tesi/"

TB<-matrix(0, nrow(matrice1), m*2)
EB<-matrix(NA,B,J)

t_student<-rep(0, nrow(matrice1))
wilcoxon<-rep(0, nrow(matrice1))
semiparametrico<-rep(0, nrow(matrice1))

      for(i in 1:B){

      TB<-rowperm(matrice1)      #matrici permutate
      write.table(TB, file=paste(filepath, "TB" , i, ".txt"),
      row.names=F, col.names=F)
      x<-TB[,1:m]
      y<-TB[,(m+1):(m*2)]

            for(j in 1:nrow(matrice1)){
            t_student[j]<-t.test(x[j,],y[j,],
            var.equal=TRUE)$statistic
            wilcoxon[j]<-wilcox.test(x[j,],y[j,])$statistic
            }

      t_student<-matrix(t_student, ncol=1, byrow=TRUE)
      wilcoxon<-matrix(wilcoxon, ncol=1, byrow=TRUE)
      semi<-semiparametric.test(x,y)
      semiparametrico<-semi$test
      semiparametrico<-matrix(semiparametrico, ncol=1, byrow=TRUE)
      Test<-cbind(t_student, semiparametrico, wilcoxon)
            if(any(is.na(Test))==TRUE)
            next
      Test<-abs(Test)

      write.table(Test, file=paste(filepath,"Test", i, ".txt"),
      row.names=F, col.names=F)

      Eb<-apply(Test,2,max)
      # write.table(Eb, file=paste(filepath, "Eb", i, ".txt"),
      row.names=F, col.names=F)
```

```
      EB[i,]<-Eb
      write.table(EB, file="c:/tabelle tesi/EB.txt", row.names=F,
      col.names=F)

      }

EB<-na.omit(EB)
Ep<-rep(0,J)
Ep<-apply(EB,2,max)
# write.table(Ep, file="c:/tabelle tesi/Ep.txt", row.names=F,
col.names=F)

Em<-rbind(E0,Ep)
# write.table(Em, file="c:/tabelle tesi/Em.txt", row.names=F,
col.names=F)

E<-apply(Em,2,max)    #massimo assoluto

write.table(E, file="c:/tabelle tesi/E.txt", row.names=F,
col.names=F)

# FASE 3 #

# Si calcola la distanza di ciascun gene dal massimo globale E e si
# ordinano le distanze ottenute in modo crescente.

MAD<-rep(0,nrow(Test0))
      for(i in 1:nrow(Test0)){
      MAD[i]<-(mad(Test0[i,]))^2
      }
MAD<-matrix(MAD,nrow(Test0),1)
Distanze<-matrix(0,nrow(Test0),J)
D<-rep(0,nrow(Test0))

for(i in 1:nrow(Test0)){
      for(j in 1: J){
      Distanze[,j]<-((Test0[,j]-E[j])^2/MAD[i])
      }
D[i]<-sum(Distanze[i,])
      }
DD<-matrix(D, nrow(Test0), 1)
D<-sort(DD)
Ordine<-order(DD)
DE1<-DE1[Ordine]

write.table(D, file="c:/tabelle tesi/D.txt", row.names=F,
col.names=F)
write.table(Ordine, file="c:/tabelle tesi/Ordine.txt", row.names=F,
col.names=F)
write.table(DE1, file="c:/tabelle tesi/DE1.txt", row.names=F,
col.names=F)

# Valutazione dei dati ottenuti #

# Per ogni permutazione B si calcolano le distanze D dei Test(b)
# dal massimo assoluto E, infine, si ordinano i vettori delle
# distanze e si salvano i file ottenuti.
```

```
for (z in 1:B){
      Test<-read.table(file=paste(filepath,"Test", z, ".txt"))
      Test<-as.matrix(Test)
      MAD<-rep(0,nrow(Test))
            for(i in 1:nrow(Test)){
                  MAD[i]<-(mad(Test[i,]))^2
                  }
      MAD<-matrix(MAD,nrow(Test),1)
      Distanze<-matrix(0,nrow(Test),J)
      D<-rep(0,nrow(Test))

      for(i in 1:nrow(Test)){
            for(j in 1: J){
                  Distanze[,j]<-((Test[,j]-E[j])^2/MAD[i])
                  }
            D[i]<-sum(Distanze[i,])
            }
      D<-matrix(D,nrow(Test),1)
      D<-sort(D)
      write.table(D, file=paste(filepath,"D", z, ".txt"),
row.names=F, col.names=F)
      }

# Si calcola, per ogni gene, il numero di geni che hanno valore
# inferiore al dato gene, che sono contenuti in ogni vettore
# D(b),inerente alle distanze dei test ottenuti con le b
# permutazioni dal massimo globale E. I valori ottenuti
# sono contenuti nella matrice FP.

D<-read.table(file="c:/tabelle tesi/D.txt")
D<-as.matrix(D)
FP<-matrix(0, ncol=B, nrow=nrow(D))

for (j in 1:B){
k<-read.table(file=paste(filepath, "D", j, ".txt"))
k<-as.matrix(k)
      for(i in 1:nrow(D))
            FP[i,j]<-sum(k<=D[i])}
write.table(FP, file="c:/tabelle tesi/FP.txt", row.names=F,
col.names=F)

# Si calcola la mediana dei valori ottenuti nella matrice FP, per
# ogni gene.

FP_med<-rep(0,nrow(FP))
for (i in 1:nrow(FP)){
      FP_med[i]<-median(FP[i,])
      }
write.table(FP_med, file="c:/tabelle tesi/FP_med.txt", row.names=F,
col.names=F)

# Il valore del "False Discovery Rate" si ottiene mediante il
# rapporto tra le mediane appena calcolate e l'indice i, relativo
# all'i-esimo gene.

for (i in 1:length(FP_med))
FDR<-FP_med/i
```

```
write.table(FDR, file="c:/tabelle tesi/FDR.txt", row.names=F,
col.names=F)

FDR1<-FDR<0.05
write.table(FDR1, file="c:/tabelle tesi/FDR1.txt", row.names=F,
col.names=F)

      }
```

## Codice R relativo all'algoritmo Deds con simulazione da modello Log-Normale Normale.

```
dedsLNN<-function(m=15,N=2000,B=500,mu0=2.3,sigma=0.3,tau=1.39,p=0.2){

library(permax)

#m = numerosità del campione
#N = geni simulati
#B = numero delle permutazioni del dataset simulato
#(mu0,sigma,tau)=parametri del modello LogNormale-Normale

# SIMULAZIONE DEL DATASET #

Sim_LNN<-function(N=N,m=m,mu0=mu0,sigma=sigma,tau=tau,p=p){
      ncond=2
      matrice<-matrix(rep(NA,N*ncond*m),ncol=ncond*m,byrow=T)
      #vettore che indice un gene è differenzialmente espresso
      DE<-rep(FALSE,N)

      for (i in 1:N){
            if(runif(1)>p){
                  mu.g<-rnorm(1,mu0,tau)
                  matrice[i,]<-exp(rnorm(m*ncond,mu.g,sigma))
                  }
            else{
                  # different expression
                  mu.g1<-rnorm(1,mu0,tau)
                  mu.g2<-rnorm(1,mu0,tau)
                  cond1<-exp(rnorm(m,mu.g1,sigma))
                  cond2<-exp(rnorm(m,mu.g2,sigma))
                  matrice[i,]<-c(cond1,cond2)
                  DE[i]<-TRUE
                  }
      }
      list(matrice=matrice,DE=DE)
}

Simulazione<-Sim_LNN(N=N,m=m,mu0=mu0,sigma=sigma,tau=tau,p=p)
matrice<-Simulazione$matrice
DE<-Simulazione$DE

write.table(matrice,file="c:/tabelle
tesi/matrice.txt",row.names=F,col.names=F)
```

```
write.table(DE,file="c:/tabelle tesi/DE.txt",row.names=F,col.names=F)

# FASE 1 #

# Calcolo delle J statistiche nel dataset simulato: all'interno del
# ciclo "for" viene calcolato il test t-student e il test di wilcoxon,
# successivamente si richiama la funzione semiparametric.test, che
# permette il calcolo del test semiparametrico. Si calcolano infine i
# valori massimi delle J statistiche.

x<-matrice[,1:m]
y<-matrice[,(m+1):(m*2)]

stat.t<-rep(0,N)
stat.wilc<-rep(0,N)
stat.sem<-rep(0,N)
p_t<-rep(0,N)
p_wilc<-rep(0,N)
p_sem<-rep(0,N)

      for(i in 1:N){
      stat.t[i]<-t.test(x[i,],y[i,],var.equal=TRUE)$statistic
      p_t[i]<-t.test(x[i,],y[i,],var.equal=TRUE)$p.value
      stat.wilc[i]<-wilcox.test(x[i,],y[i,])$statistic
      p_wilc[i]<-wilcox.test(x[i,],y[i,])$p.value
      }
stat.t<-matrix(abs(stat.t), ncol=1, byrow=TRUE)
stat.wilc<-matrix(abs(stat.wilc), ncol=1, byrow=TRUE)

sem<-semiparametric.test(x,y)
stat.sem<-sem$test
p_sem<-sem$pvalue
stat.sem<-matrix(abs(stat.sem), ncol=1, byrow=TRUE)

p_value<-cbind(p_t,p_sem,p_wilc)
p_value<-na.omit(p_value)

write.table(p_value, file="c:/tabelle tesi/p_value.txt", row.names=F,
col.names=F)

J<-ncol(p_value)

Test_na<-cbind(stat.t, stat.sem, stat.wilc)

id<-rep(FALSE,N)
      for(i in 1:N){
            if(any(is.na(Test_na[i,]))==TRUE)
            id[i]<-TRUE
            }
id<-matrix(id, ncol=1)  # vettore che indica con TRUE le righe della
                        # matrice che producono valori Na nei test.
DE1<-DE[id==FALSE]      # vettore di geni differenzialmente espressi
                        # senza i valori corrispondenti a Na.
write.table(id, file="c:/tabelle tesi/id.txt", row.names=F, col.names=F)

Test0<-na.omit(Test_na)       # vengono omesse le righe contenenti Na.
percentuale0<-(N-nrow(Test0))*100/N # percentuale di valori Na sul totale
E0<-apply(Test0,2,max)
```

```
write.table(Test0, file="c:/tabelle tesi/Test0.txt", row.names=F,
col.names=F)
# write.table(E0, file="c:/tabelle tesi/E0.txt", row.names=F,
col.names=F)
# write.table(percentuale0, file="c:/tabelle tesi/perc0.txt",
row.names=F, col.names=F)

matrice1<-matrice[id==FALSE,]

# FASE 2 #
# Si ripete il calcolo dei test da nuovi dataset, ottenuti tramite B
# permutazioni del dataset "matrice 1". Questa matrice si ottiene
# simulando i dati e, dopo il calcolo dei test, omettendo le righe
# contenenti valori "Na". Si calcolano successivamente i massimi delle J
# statistiche, e tra tutti i massimi si ricalcolano nuovamente i valori
# massimi, ottenendo in tal modo il massimo assoluto E.

filepath <- "c:/tabelle tesi/"

TB<-matrix(0, nrow(matrice1), m*2)
EB<-matrix(NA,B,J)

t_student<-rep(0, nrow(matrice1))
wilcoxon<-rep(0, nrow(matrice1))
semiparametrico<-rep(0, nrow(matrice1))

      for(i in 1:B){

      TB<-rowperm(matrice1)        #matrici permutate
      write.table(TB, file=paste(filepath, "TB" , i, ".txt"),
      row.names=F, col.names=F)
      x<-TB[,1:m]
      y<-TB[,(m+1):(m*2)]

            for(j in 1:nrow(matrice1)){
            t_student[j]<-t.test(x[j,],y[j,],var.equal=TRUE)$statistic
            wilcoxon[j]<-wilcox.test(x[j,],y[j,])$statistic
            }

      t_student<-matrix(t_student, ncol=1, byrow=TRUE)
      wilcoxon<-matrix(wilcoxon, ncol=1, byrow=TRUE)
      semi<-semiparametric.test(x,y)
      semiparametrico<-semi$test
      semiparametrico<-matrix(semiparametrico, ncol=1, byrow=TRUE)
      Test<-cbind(t_student, semiparametrico, wilcoxon)
            if(any(is.na(Test))==TRUE)
            next
      Test<-abs(Test)

      write.table(Test, file=paste(filepath,"Test", i, ".txt"),
      row.names=F, col.names=F)

      Eb<-apply(Test,2,max)
      # write.table(Eb, file=paste(filepath, "Eb", i, ".txt"),
      row.names=F, col.names=F)

      EB[i,]<-Eb
```

```
      write.table(EB, file="c:/tabelle tesi/EB.txt", row.names=F,
      col.names=F)

      }

EB<-na.omit(EB)
Ep<-rep(0,J)
Ep<-apply(EB,2,max)
# write.table(Ep, file="c:/tabelle tesi/Ep.txt", row.names=F,
col.names=F)

Em<-rbind(E0,Ep)
# write.table(Em, file="c:/tabelle tesi/Em.txt", row.names=F,
col.names=F)

E<-apply(Em,2,max)      #massimo assoluto

write.table(E, file="c:/tabelle tesi/E.txt", row.names=F, col.names=F)

# FASE 3 #

# Si calcola la distanza di ciascun gene dal massimo globale E e si
# ordinano le distanze ottenute in modo crescente.

MAD<-rep(0,nrow(Test0))
      for(i in 1:nrow(Test0)){
      MAD[i]<-(mad(Test0[i,]))^2
      }
MAD<-matrix(MAD,nrow(Test0),1)
Distanze<-matrix(0,nrow(Test0),J)
D<-rep(0,nrow(Test0))

for(i in 1:nrow(Test0)){
      for(j in 1: J){
      Distanze[,j]<-((Test0[,j]-E[j])^2/MAD[i])
      }
D[i]<-sum(Distanze[i,])
      }
DD<-matrix(D, nrow(Test0), 1)
D<-sort(DD)
Ordine<-order(DD)
DE1<-DE1[Ordine]

write.table(D, file="c:/tabelle tesi/D.txt", row.names=F, col.names=F)
write.table(Ordine, file="c:/tabelle tesi/Ordine.txt", row.names=F,
col.names=F)
write.table(DE1, file="c:/tabelle tesi/DE1.txt", row.names=F,
col.names=F)

# Valutazione dei dati ottenuti #

# Per ogni permutazione B si calcolano le distanze D dei Test(b) dal
# massimo assoluto E, infine, si ordinano i vettori delle distanze e si
# salvano i file ottenuti.

for (z in 1:B){
      Test<-read.table(file=paste(filepath,"Test", z, ".txt"))
      Test<-as.matrix(Test)
```

```
      MAD<-rep(0,nrow(Test))
            for(i in 1:nrow(Test)){
                  MAD[i]<-(mad(Test[i,]))^2
                  }
      MAD<-matrix(MAD,nrow(Test),1)
      Distanze<-matrix(0,nrow(Test),J)
      D<-rep(0,nrow(Test))

      for(i in 1:nrow(Test)){
            for(j in 1: J){
                  Distanze[,j]<-((Test[,j]-E[j])^2/MAD[i])
                  }
            D[i]<-sum(Distanze[i,])
            }
      D<-matrix(D,nrow(Test),1)
      D<-sort(D)
      write.table(D, file=paste(filepath,"D", z, ".txt"), row.names=F,
      col.names=F)
      }

# Si calcola, per ogni gene, il numero di geni che hanno valore
# inferiore al dato gene, e che sono contenuti in ogni vettore
# D(b),inerente alle distanze dei test ottenuti con le b permutazioni
# dal massimo globale E.
# I valori ottenuti sono contenuti nella matrice FP.

D<-read.table(file="c:/tabelle tesi/D.txt")
D<-as.matrix(D)
FP<-matrix(0, ncol=B, nrow=nrow(D))

for (j in 1:B){
k<-read.table(file=paste(filepath, "D", j, ".txt"))
k<-as.matrix(k)
      for(i in 1:nrow(D))
            FP[i,j]<-sum(k<=D[i])}
write.table(FP, file="c:/tabelle tesi/FP.txt", row.names=F, col.names=F)

# Si calcola la mediana dei valori ottenuti nella matrice FP, per ogni
# gene.

FP_med<-rep(0,nrow(FP))
for (i in 1:nrow(FP)){
FP_med[i]<-median(FP[i,])
}
write.table(FP_med, file="c:/tabelle tesi/FP_med.txt", row.names=F,
col.names=F)

# Il valore del "False Discovery Rate" si ottiene con il rapporto tra le
# mediane appena calcolate e l'indice i, relativo all'i-esimo gene cui
# corrisponde il valore della mediana.

for (i in 1:length(FP_med))
FDR<-FP_med/i

write.table(FDR, file="c:/tabelle tesi/FDR.txt", row.names=F,
col.names=F)

FDR1<-FDR<0.05
```

```
write.table(FDR1, file="c:/tabelle tesi/FDR1.txt", row.names=F,
col.names=F)
}
```

## Codice R relativo al test semiparametrico.

```
semiparametric.test<-function(x,y,level=0.05)
{
pval<-level/2
z1<-qnorm(pval)
z2<-qnorm(1-pval)

# se le matrici vengono inserite con i pazienti sulle righe, le
# traspone ed inoltre salva in "m" ed "n" il numero di pazienti per
# ciascuna patologia.

n1<-nrow(x)
n2<-ncol(x)
m1<-nrow(y)
m2<-ncol(y)

if(n1<n2) {
      x<-t(x)
      n<-ncol(x)
      }
if(n1>=n2){
      n<-n2
      }
if(m1<m2){
      y<-t(y)
      m<-ncol(y)
      }
if(m1>=m2){
      m<-m2
      }
ngen<-nrow(y)

# fa corrispondere ad "x" la matrice con maggior numero di
# pazienti e ad "y" quella con numero minore.

if(m>n){
      z<-x
      x<-y
      y<-z
      rm(z)
      n<-ncol(x)
      m<-ncol(y)
      }

# calcolo della media, varianza, e deviazione standard
# (relativamente a y) e coefficienti "rho".

mean<-apply(y,1,mean)
var<-apply(y,1,var)
var<-((m-1)/m)*var
```

```
sd<-sqrt(var)         #mean,sd e var sono vettori p-dimensionali
S<-1-(apply(x,2,pnorm,mean=mean,sd=sd)) #S è una matrice pxn
rho<-1/n*(apply(S,1,sum))  #rho è un vettore p-dimensionale

#calcolo della deviazione standard dei coefficienti rho
#calcolo di w quadro esse

g<-(S-rho)^2
w2s<-(1/(n-1))*(apply(g,1,sum))

#matrice "omega"
omega1<-var
omega2<-2*(var^2)

#calcolo del primo elemento del vettore beta
zi<-(x-mean)/sd
densxi<-apply(zi,2,dnorm)
beta1<-(1/(n*sd))*apply(densxi,1,sum)

#calcolo del secondo elemento del vettore beta
si<-zi*densxi
beta2<-(1/(2*n*var))*apply(si,1,sum)

p1<-omega1*(beta1^2)
p2<-omega2*(beta2^2)
p<-(n/m)*(p1+p2)
var.rho<-(w2s+p)

#passaggio al logit

tau<-log(rho/(1-rho))
var.tau<-(var.rho)/(rho^2*(1-rho)^2*n)
toss.tau<-(tau)/sqrt(var.tau)
tau.inf<-z1*sqrt(var.tau)
tau.sup<-z2*sqrt(var.tau)

rho<-exp(tau)/(exp(tau)+1)
rho.inf<-exp(tau.inf)/(exp(tau.inf)+1)
rho.sup<-exp(tau.sup)/(exp(tau.sup)+1)
test<-toss.tau
var.test<-var.tau

#calcolo dei valori p e conta dei geni significativi
test<-toss.tau
non.agg<-2*(1-pnorm(abs(test)))
pvalue<-non.agg

list(rho=rho, rho.inf=rho.inf, rho.sup=rho.sup, test=test,
var.test=var.test, pvalue=pvalue)
}
```

# BIBLIOGRAFIA

Yee Hwa Yang, Yuanyuan Xiao, and Mark R. Segal, "Identifying differentially expressed genes from microarray experiments via statistic synthesis" (July 2, 2004). Center for Bioinformatics & Molecular Biostatistics. Paper deds.

Yuanyuan Xiao, at al. “Analysis of a Splice Array Experiment Elucidates Roles of Chromatin Elongation Factor Spt4-5 in Splicing”. (2005) Center for Bioinformatics & Molecular Biostatistics. Paper PLoS.

Armitage P., “Statistica Medica. Metodi statistici per la ricerca in Medicina”, McGraw Hill, Libri Italia.

Terry Speed, Yee Hwa Yang, Sandrine Dudoit, “Dynamic Modeling of complex Biomedical Systems. Statistics and microarray analysis: design, pre-processing and analysis.” Statistics, UC Berkeley, Biochemistry, Stanford. April 26-28, 2001 Washington, D.C.

Mocci, Simona “The use and analysis of microarray data”. Nature Publishing Group (2002)

Ruffino F., Valentini G., Fuselli M. “Valutazione di metodi di Gene Selection per l’analisi di esperimenti con Dna Microarray”. Dipartimento di Scienze dell’Informazione. Università di Milano.

Pace, L. e Salvan, S. “Introduzione alla statistica - II. Inferenza, verosimiglianza, modelli.” Cedam, Padova. (2001)

Cichitelli G., “Probabilità e Statistica”, Maggioli Ed., Rimini 1992.

Bortot, P., Ventura, L. e Salvan, A. (2000). “Inferenza statistica: applicazioni con S-Plus e R.”, Cedam, Padova.

Iacus, Masarotto, “Laboratorio di statistica con R”, McGraw Hill, 2003.

Hasko, E. “Individuazione di geni differenzialmente espressi: uno studio di simulazione.” (2005)

Cavallin, F. “Test alternativi per la selezione di geni differenzialmente espressi: uno studio di simulazione.” (2005)

Lise, M. “Metodi Bayesiani empirici per l’identificazione di geni differenzialmente espressi”. (2005)

www.microarraystation.com/it/dna-microarray-bioinformatics/

www.bepress.com/uwbiostat/index.html