Università degli Studi di Padova

FACOLTÀ DI SCIENZE STATISTICHE
Corso di laurea in Statistica e Informatica

Tesi di laurea

Confronto tra statistiche per l'ordinamento e la classificazione dei geni differenzialmente espressi

Candidato:
Daiana Gressani
Matricola 532690

Relatore:
Ch.ma Prof.ssa Monica Chiogna

Anno Accademico 2008–2009

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

La prima sequenza genomica ad essere stata pubblicata, nel 1995, fu quella di *Haemophilus influenzae*, batterio gram-negativo, con un genoma di circa 1,8 milioni di basi. Successivamente, nel 1996, è stato completato il sequenziamento del primo genoma eucariotico, quello del lievito *Saccharomyces cerevisiae*, che comprende circa 13 milioni di basi organizzate in sedici cromosomi. Infine, nel 2003, è stato completamente codificato il genoma umano grazie all'omonimo progetto. Per sequenza del genoma umano, si intende la sequenza completa del DNA, presente nei 22 autosomi più la coppia di cromosomi sessuali.

Nonostante i più sofisticati sistemi disponibili, si riescono ad interpretare, solo parzialmente, gli elementi funzionali contenuti in un genoma e, ancor meno, a comprendere il significato dell'informazione genomica nella sua globalità. Gli acidi nucleici offrono un metodo di indagine basato sulla specificità di ibridazione di due eliche complementari, che possono fungere da sonde per l'identificazione e la quantificazione di specifici mRNA. Il problema principale consiste nell'identificare le sequenze di DNA che sono trascritte in RNA messaggero (mRNA) per essere poi tradotte in proteine. L'analisi dell'insieme degli RNA trascritti o trascrittoma, consente di indagare direttamente a livello di RNA.

Il profilo trascrizionale riflette lo stato funzionale di una cellula; di conseguen-

za, capire in quali circostanze un gene si è espresso è essenziale per comprenderne la funzione. Un aiuto importante è dato dalla regolazione dei geni, vale a dire quando si accendono e spengono in risposta a particolari situazioni.

Lo studio dei profili di espressione genica mediante tecnologia microarray, consente di quantificare contemporaneamente centinaia o migliaia di mRNA presenti in un determinato campione mediante un unico esperimento. Ciò permette di ottenere un profilo di espressione o fenotipo molecolare, altamente dettagliato, che consente di: approfondire le conoscenze sui meccanismi fisiopatologici alla base delle malattie; la bioinformatica trova dunque nell'analisi di dati genomici un'area di indagine innovativa: individuare quali sono i geni attivi nel campione in esame; qual'è il loro livello di espressione e quali variazioni avvengono in condizioni patologiche.

Capitolo 1

Microarray

1.1 Tecnologia Microarray

I *microarray* a DNA sono formati da moltissime molecole di DNA (dette sonde), depositate in una posizione nota su un supporto in modo da formare una microgriglia (da cui il nome *microarray*), che le identifica in maniera univoca. Questi filamenti di DNA, collocati sulla superficie del supporto, sono utilizzati come sonde per misurare la quantità di altre molecole di DNA (anch'esse a singolo filamento) derivate dai trascritti di mRNA e contenute in una soluzione che viene depositata sulla superficie del *microarray*. Il supporto è di solito un vetrino per microscopio che ha circa le dimensioni di un pollice della mano.

Tre sono i principali approcci per la fabbricazione di microarray.

- *cDNA microarray*. Questo approccio consiste nel depositare, tramite un robot, una soluzione contenente le sonde di DNA sulla superficie del supporto solido. Le sonde possono essere costituite da cDNA (DNA complementare ottenuto da un trascritto di mRNA; che ha una lunghezza di 200-2400 basi) a singolo filamento oppure da oligonucleotidi (corte sequenze di nucleotidi con lunghezza di 50-100 basi) chimicamente pre-sintetizzati.

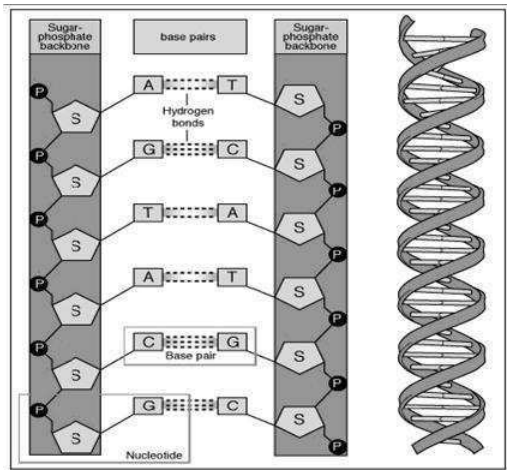

Figura 1.1: Ibridazione di DNA e RNA

I *microarray* fabbricati con questo procedimento, nel caso in cui le sonde sono formate da cDNA, sono chiamati *"cDNA microarray"* Ross [1995].

- *Affymetrix*: Si sintetizzano gli oligonucleotidi direttamente *(in situ)* sulla superficie del microarray, un'operazione che è eseguita principalmente con tecniche di tipo fotolitografico *(tipica di Affymetrix)* Lips [2005] e di stampa a getto *(metodo sviluppato da Rosetta Inpharmatics e concesso in licenza ad Agilent Technologies)* Hug [1999].

- *Microarray di Agilent con oligonucleotidi sintetizzati in situ*: Questi *microarray* sono realizzati sintetizzando oligonucleotidi direttamente sulla superficie di un vetrino con un processo di stampa a getto.

1.1.1 Caratteristiche dei cDNA microarray

I *cDNA microarray* sono realizzati depositando su un vetrino di pochi centimetriquadrati, in fissate locazioni dette spot prodotti di PCR (la *Polymerase Chai Reaction* è una tecnica comunemente utilizza in biologia per amplificare frammenti di DNA), derivati da cloni di cDNA che sono stati opportunamente selezionati da database di dominio pubblico. Ogni clone di cDNA è rappresentativo di uno specifico gene o EST *(Expressed Sequence Tag)* Nagaraj [2007]. Le EST sono corte sequenze di DNA (lunghezza 200-500 basi) che hanno una singola occorrenza nel DNA codificante del genoma umano e, nell'ambito dei microarray, sono utilizzate per caratterizzare ipotetici nuovi geni parzialmente sequenziati. Il prodotto di PCR prima di essere depositato è parzialmente depurato, la soluzione risultante è trasferita sulla superficie del vetrino tramite un gruppo di penne disposte a griglia, guidati da un braccio robotico. La superficie del vetrino, prima di depositare la soluzione contenente il cDNA, è ricoperta con composti chimici per permettere il fissaggio del cDNA.

A volte c'è sostanziale variabilità nelle dimensioni e nella forma degli spot. La dimensione dello spot influisce sulla quantità di cDNA disponibile per l'ibridazione e la forma influisce sull'analisi dell'immagine. Nell' applicazione di questa tecnologia, due campioni di mRNA sono separatamente retrotrascritti ed etichettati con differenti marcatori, per essere successivamente miscelati ed ibridati congiuntamente su un unico microarray. Di conseguenza, il cDNA *microarray* consente di misurare i livelli relativi di espressione di due campioni e quindi appartiene a quel genere di *microarray* che denominiamo "a due marcatori".

1.1.2 Genechip Affymetrix

Per sintetizzare gli oligonucleotidi direttamente sulla superficie di un vetrino quadrato, con lato di 1,28 cm, *Affymetrix* utilizza un processo fotolitografico simile a quello impiegato per fabbricare i chip per computer, in modo tale da riuscire a produrre in serie *microarray* standard. Il processo fotolitografico è controllato da maschere che permettono il passaggio della luce solo in zone prestabilite della superficie del microarray. La luce è impiegata, di volta in volta, per avviare il processo chimico che lega la singola base alla sequenza già sintetizzata sul microarray dell'oligonucleotide specifico. Visto che la produzione di maschere è abbastanza costosa, questo tipo di tecnologia è poco flessibile e non è particolarmente adatta per fabbricare microarray personalizzati. Ogni oligonucleotide ha una lunghezza di 25 basi ed è sintetizzato in circa 10 milioni di copie su un'area quadrata del vetrino con lato di 24 mm, denominata cella delle sonde. Per rappresentare ogni gene si utilizza un insieme di 11-20 oligonucleotidi diversi nella forma di PM (*Perfect Match*) e altrettanti nella forma di MM (*Mismatch*). In particolare, gli oligonucleotidi nella forma di PM sono perfettamente complementari all'mRNA trascritto da una particolare regione del gene in questione, mentre quelli nella forma di MM sono identici ai loro corrispettivi PM eccetto che per il nucleotide nella posizione centrale (il tredicesimo) della sequenza, il quale è sostituito dal suo complementare (la cella con gli oligonucleotidi PM e la cella con gli oligonucleotidi associati nella forma di MM formano una coppia di celle). Gli oligonucleotidi nella forma di MM sono utilizzati per scovare ibridazioni non specifiche del suo associato nella forma di PM; questo è un particolare importante soprattutto per quantificare l'mRNA espresso debolmente. Questo *microarray* è utilizzato per quantificare i livelli di espressione di un solo campione di mRNA, opportunamente etichettato con un marcatore fluorescente, e quindi appartiene a quel genere di *microarray* denominiato "a un marcatore".

1.1.3 Microarray sintetizzati in situ

Questi *microarray* sono realizzati sintetizzando oligonucleotidi direttamente sulla superficie di un vetrino con un processo di stampa a getto, simile a quello adottato dalle comuni stampanti a getto di inchiostro. Questa tecnologia è molto flessibile, in quanto il processo di stampa del *microarray* è controllato da un computer e quindi le sequenze di ogni oligonucleotide sono definite in un file. Tutto questo rende particolarmente efficiente il processo di fabbricazione di *microarray* personalizzati. Un'altra caratteristica positiva è la regolarità di forma e dimensione degli spot che rappresentano uno specifico gene. Questo tipo di *microarray* è "a due marcatori".

1.2 Misura del livello di espressione genica

Per misurare il livello di espressione genica in un campione di tessuto o di una linea cellulare sono necessarie quattro fasi:

- *estrazione dell' RNA;*
- *marcatura;*
- *ibridazione;*
- *acquisizione dell'immagine;*

Le fasi che sono descritte di seguito si riferiscono ad un esperimento condotto con un cDNA microarray e definite tramite un protocollo specifico della tecnologia microarray.

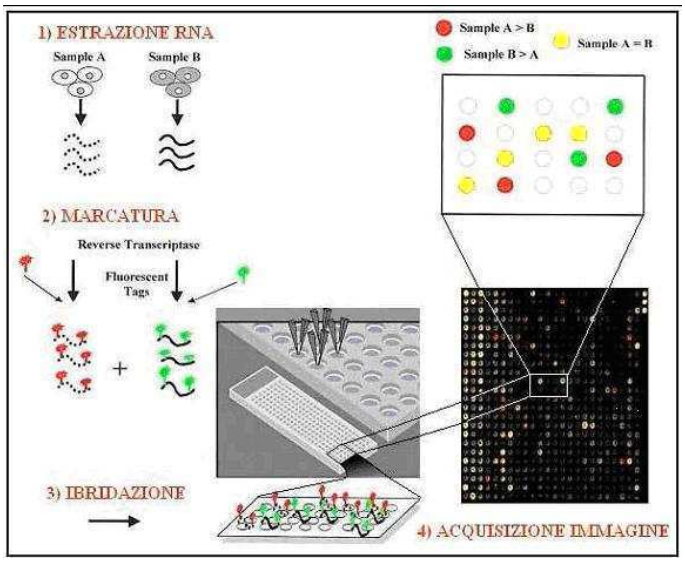

Figura 1.2: Fasi di un esperimento condotto con cDNA microarray

Estrazione e Marcatura

Una fase preliminare, eseguita nel caso di studio di tessuti non omogenei, è la dissezione di piccole popolazioni di cellule utilizzando strumenti a laser. L'estrazione delle macromolecole di RNA dal campione cellulare è effettuata tramite opportuni kit commerciali adatti a questo scopo. Ricavato l'RNA totale si procede con la fase di marcatura. L'operazione di marcatura può essere eseguita direttamente incorporando un nucleotide modificato legato con uno dei due marcatori fluorescenti Cy3 (verde) e Cy5 (rosso), e indirettamente, tramite un processo di retro trascrizione dell'mRNA, per ottenere il cDNA che è una macromolecola più stabile [Loftus [1999]].

La marcatura diretta avviene ricavando il cDNA con incorporati i nucleotidi modificati già legati al marcatore fluorescente. La marcatura indiretta è ottenuta ricavando il cDNA con incorporati i nucleotidi modificati legati ad un gruppo chimico reagente che solo successivamente è fatto reagire al marcatore fluorescente. La marcatura indiretta richiede un incremento in tempo di lavoro che però è larga-

mente compensata da un aumento in sensibilità, da una minor propensione all'errore sistematico causato dai marcatori (nella marcatura diretta Cy5 è più difficile da incorporare rispetto a Cy3) e costi più bassi.

1.2.1 Ibridazione

I cDNA marcati sono miscelati in una soluzione che viene depositata sulla superficie del microarray per iniziare la fase di ibridazione. L'ibridazione è un processo nel quale due filamenti singoli di DNA complementari si combinano creando l'usuale doppia elica di DNA. Negli esperimenti con cDNA *microarray*, l'ibridazione avviene tra i cDNA marcati in soluzione (*target*) e i cDNA fissati al supporto di vetro (sonde). [*Southern et al., 1975*] Sono molti i fattori che influiscono sul processo come: temperatura, umidità, concentrazione salina, volume della soluzione e l'operatore che prepara la procedura. Terminato il periodo d'incubazione la superficie del microarray viene lavata, per rimuovere la soluzione di ibridazione in eccesso e assicurare che solo i cDNA target complementari ai rispettivi cDNA sonda rimangano legati, riducendo casi di ibridazione non specifica.

1.2.2 Acquisizione dell'immagine

Per ottenere un'immagine della superficie del *microarray* ibridato viene utilizzato uno scanner. Lo scanner attraverso un laser eccita i marcatori fluorescenti presenti sulla superficie del *microarray*, i quali emettono una radiazione luminosa ad una particolare lunghezza d'onda che viene rilevata da un tubo fotomoltiplicatore. I marcatori fluorescenti utilizzati, Cy3 e Cy5, sono eccitati da radiazioni con frequenze diverse; di conseguenza lo sono anche le radiazioni emesse e quindi il segnale è distinguibile. Le lunghezze d'onda della radiazione di eccitazione e di emissione per il marcatore Cy3 sono rispettivamente 550 nm e 581 nm, mentre

per il marcatore Cy5 sono 649 nm e 670 nm Tian [2007]. La quantità di radiazione emessa da un particolare punto della superficie del microarray dipende dalla quantità del marcatore che si sta eccitando (e di conseguenza dalla quantità del cDNA ibridato di un particolare campione). Per ottenere l'intera scansione della superficie, il laser deve essere focalizzato su ogni punto del microarray.
La prima fase di analisi, dopo la procedura di laboratorio, è quella di elaborazione dell'immagine Smyth [2004] , in cui tramite opportuni software si ottengono valori numerici da associare ad ogni spot, tra i quali l'intensità del segnale in primo piano (*foreground*) ed in sottofondo (*background*).

1.3 Estrazione del segnale e correzioni

1.3.1 La quantizzazione o quantificazione

La *quantizzazione* o *quantificazione* è il processo che permette di ricavare dall'insieme delle informazioni relative ad uno spot un valore numerico, rappresentativo della concentrazione di mRNA di quel gene nel campione. Dalla quantizzazione, si ricava il rapporto delle intensità assolute sui due canali, detto *fold change*, che serve ad avere informazioni sul livello di espressione in un canale rispetto all'altro.

L'esito di questa operazione produce due immagini monocromatiche a 16 bit, corrispondenti a Cy3 e Cy5, memorizzate in un file TIFF, dove l'intensità di ogni pixel su ogni canale può essere quantificata da 65536 valori. È necessario che la risoluzione dell'immagine sia scelta in maniera tale che ogni spot sia descritto da un numero sufficiente di pixel da rendere robusto il calcolo dell'intensità associato allo spot, che è influenzato dal rumore tra *pixel*. Una buona scelta è quella di associare ad ogni spot almeno 50 *pixel*. Dalle immagini monocromatiche si ottiene l' immagine in falsi colori (rosso-verde) dei cDNA *microarray*.

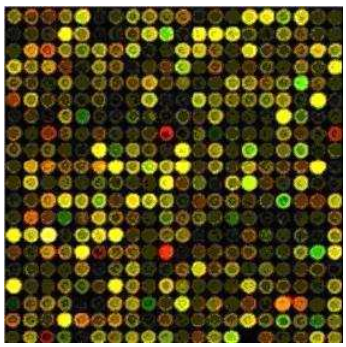

Figura 1.3: Immagine a colori di un esperimento con cDNA microarray

1.3.2 Correzione del Background

La presenza di un segnale di fondo sul microarray può essere dovuta a diversi fattori. Questi valori possono essere esclusi dall'insieme di dati attraverso la correzione o sottrazione del *background* [Ernst [2004], Yang and Speed [2002]]. Nella sottrazione locale del *background* viene identificato un intorno sufficientemente ampio centrato sullo *spot* e viene considerata la media o la mediana del *pixel* esterni allo *spot* ma interni alla zona di demarcazione come valore locale del rumore; questo valore viene poi sottratto alla media o mediana dello spot canale a canale. Con questa operazione è possibile gestire la variabilità locale del rumore, tuttavia non è un procedimento privo di rischi. Per ovviare agli inconvenienti della sottrazione del *background* è possibile calcolare il valore di *background* su sotto-griglie del *microarray*; in questo modo si conduce il calcolo su un ambito meno locale e si può riuscire a ricavare una stima del rumore anche su *array* particolarmente densi di *spot*.

Il parametro universalmente accettato per la misurazione degli effetti del rumore su un segnale è il rapporto segnale rumore (*Signal to Noise Ratio - SNR*), definito generalmente come:

$$SNR = \text{Mediana del segnale} \,/\, \text{STD del rumore} \tag{1.1}$$

dove al denominatore vi è la *deviazione standard (STD)* del rumore.
Molti software di analisi di microarray utilizzano il valore di SNR ricavato per ogni spot per escludere dal processo di normalizzazione quei dati che hanno un rumore troppo alto.

Correzione "PAIRED-SLIDE"

La correzione *"paired-slide"*si applica ad esperimenti nei quali due campioni diversi vengono ibridizzati su due *microarray* scambiando la marcatura, cioè il campione che sul primo microarray viene marcato in rosso, sul secondo sarà marcato in verde e viceversa per l'altro campione. Con questo metodo è possibile scalare i livelli di espressione relativa per i due microarray senza esplicitare la normalizzazione: questo procedimento prende il nome di *"self-normalization"*.
La validità di questa assunzione può essere verificata utilizzando un insieme di geni con livelli di espressione costante sui due canali.

1.4 Normalizzazione dei dati

Molti fattori possono influire e distorcere i risultati di un esperimento di microarray:

- disomogeneità del processo di deposizione delle sonde;
- quantità iniziali diverse di RNA;
- diversa efficienza di incorporazione dei due fluorocromi durante il procedimento di marcatura dei campioni;
- disomogeneità di ibridizzazione sul vetrino;

- diversa efficienza di emissione dei due fluorocromi;
- diversa efficienza dello scanner nel leggere i due canali di fluorescenza.

Tutti questi fattori possono influenzare i dati causando spostamenti nelle distribuzioni dei rapporti delle intensità dei due fluorofori. È necessaria quindi una normalizzazione dei dati atta ad eliminare distorsioni sistematiche *(bias)*.
Il processo di normalizzazione è necessario anche per confrontare dati provenienti da repliche dello stesso materiale. Solitamente in un esperimento di microarray le repliche fra vetrini possono essere di due tipi:

- repliche sperimentali: quando l'mRNA sui due vetrini proviene dalla stessa estrazione servono per ottenere una stima migliore dell'espressione di un gene;
- repliche biologiche: quando l'mRNA proviene da campioni biologici dello stesso tipo ma distinti (ad esempio individui diversi) e consentono di stimare l'errore *random*. È necessario che la normalizzazione tenga conto del disegno dell'esperimento: la sua scorretta applicazione, infatti, invalida completamente il dato e, di conseguenza, i risultati delle analisi sullo stesso. Si parla di normalizzazione *within-array* quando la tecnica scelta viene applicata ad ogni vetrino singolarmente, nell'intento di correggere gli errori sistematici su ogni array preso come unità a sé e indipendentemente dal disegno sperimentale, mentre si fa una normalizzazione *between-arrays* quando si cerca di ottenere un dato uniforme considerando sia il disegno sperimentale applicato che il tipo di campione biologico. In ciascuna di queste situazioni è necessario scegliere un gruppo di geni da utilizzare per la normalizzazione.

Questi possono essere:

- Tutti i geni sull'array. Quasi tutti i geni sull'array possono essere utilizzati per la normalizzazione quando è possibile prevedere che solo una porzione relativamente piccola di geni varierà significativamente in espressione fra i due campioni di mRNA.

- Geni espressi in maniera costante. Invece di utilizzare tutti i geni per la normalizzazione, si può scegliere di usare un piccolo sottoinsieme rappresentato dai geni *housekeeping*, cioè quei geni che mantengono lo stesso livello di espressione in condizioni sperimentali differenti. Non è facile identificare questo sottoinsieme, ma spesso è possibile trovare un gruppo di geni che si comportano da *housekeeping* nelle condizioni sperimentali considerate. Una limitazione nell'utilizzo dei geni *housekeeping* è che essi tendono ad essere espressi molto e quindi potrebbero non essere rappresentativi di altri geni di interesse.

- Controlli. Un'alternativa alla normalizzazione con geni housekeeping è l'utilizzo di controlli spiked o di una serie di sequenze di controllo a concentrazione scalare (*titration*). Nel metodo dei controlli *spiked*, sequenze sintetiche di DNA o sequenze selezionate da organismi differenti da quello studiato sono depositate sull'array e incluse nei due differenti campioni di mRNA in esame con identica concentrazione. Queste sequenze di controllo possono essere utilizzate per la normalizzazione perché daranno origine a segnali di uguale intensità nei due canali. Nell'approccio della serie *titration*, si utilizzano spot dello stesso gene a concentrazione scalare, con uguale intensità sui due canali nel range considerato, in modo da monitorare l'amplificazione lineare

della risposta in intensità rispetto alla concentrazione.

La normalizzazione può anche essere effettuata con un esperimento di calibrazione, in cui viene realizzato un *microarray* utilizzando marcatori differenti per due porzioni uguali di mRNA prelevato dalla stessa cellula. Le intensità dei due marcatori dovrebbero essere uguali e la differenza tra i due marcatori viene utilizzata come fattore di normalizzazione. La differenza correlata all'intensità delle due tinte viene rappresentata attraverso un MA-plot che in ascissa ha il logaritmo della media geometrica delle due intensità:

$$(A = (logCy3 + logCy5)/2) \tag{1.2}$$

mentre in ordinata vi è il logaritmo del rapporto dei due canali

$$M = log(Cy5/Cy3) \tag{1.3}$$

Per questo

$$\widehat{M} = \widehat{f}(A)$$

viene stimata attraverso un metodo di regressione non parametrica che consente la normalizzazione del logaritmo del rapporto delle tinte dei dati di espressione Lönnsted and Speed [2002], tramite $\tilde{M} = M - \hat{M}$. Un' ulteriore fonte di distorsione, conosciuta come *print-tip effect*, è causata dalla differenza di conformazione delle puntine dei *robot* che depositano il materiale genetico su ogni *spot* in quantità diverse. Una soluzione possibile è utilizzare più puntine per depositare il probe cDNA su spot diversi: la variabilità delle misure di espressione è infatti dovuta a questa fonte di distorsione, che può in questo modo essere considerata in fase di normalizzazione.

1.5 Obiettivi Tesi

Questo elaborato ha lo scopo di confrontare due statistiche, SAM proposta da Tusher and Chu [2001] e la statistica B Smyth [2004], elaborate per l'individuazione di geni differenzialmente espressi. Verranno confrontati gli ordinamenti introdotti da entrambe sui valori osservati dei test. Nel secondo capitolo verranno introdotte le due statistiche per l'analisi dei geni e nel terzo capitolo saranno introdotte le varie analisi finalizzate ad individuare geni differenzialmente espressi di *microarray* nelle due statistiche, con cellule sotto due condizioni sperimentali. Nelle terzo e quarto capitolo verranno confrontati gli ordinamenti e le distribuzioni delle due statistiche mentre nel quinto verranno applicate ad un insieme di dati reali riguardanti pazienti affetti da leucemia mieloide e cronica. Le simulazioni e le analisi proposte sono state condotte utilizzando il software statistico R disponibile al sito `http//www.r-project.org`; in particolare è stata impiegata la libreria *Marray, Limma, eBayes* e le funzioni contenute nel pacchetto R *Ihaka and Gentleman [1996]* chiamato *siggenes* (per SAM) che implementa i metodi bayesiani empirici, utili all'identificazione dei geni differenzialmente espressi. In appendice ad ogni capitolo è riportato il codice R di cui si è fatto uso, in particolare le funzioni create per eseguire le simulazioni opportune.

Capitolo 2

Statistica t e Sam

I *microarray* danno la possibilità di misurare i livelli di espressione di decine di migliaia di geni. La questione statistica importante, in questi esperimenti è: quali delle diverse migliaia di geni sono differenzialmente espressi? Per rispondere a queste domande si usano test per il confronto di medie, come, per esempio, il test t di Student nel caso di due campioni. In questi test succede frequentemente, che, per ragioni sperimentali, ci siano dei geni che presentano stime delle varianze molto piccole. Seppur pochi, questi modificano i risultati del test t facendoli diventare artificialmente grandi, creando, di conseguenza degli errori di identificazione. Per ovviare a questi problemi, sono stati introdotti dei test che, attraverso l'utilizzo di statistiche t-moderate, riducono la distorsione delle stime.

I due metodi per identificare i geni differenzialmente espressi considerati nel presente lavoro sono: la procedura *Significance analysis of microarray* (Sam), introdotta da Tusher and Chu [2001] e la statistica t utilizzata all'interno dell'approccio bayesiano empirico proposto da Smyth [2004]. È utile anticipare che in questo contesto si inserisce anche il problema dei test multipli. Pertanto le procedure sopra riportate spesso utilizzano il controllo del *family-wise error rate* (Westfall & Young), oppure il *False Discovery Rate* (FDR)(Benjamini [1995] e

Benjamini and Yekutieli [2001]) per la gestione del problema della molteplicità.

2.1 La procedura Significance analysis of microarray

La cosiddetta *Significance analysis of microarray* (Sam), è una tecnica che seleziona, tra un vastissimo insieme di geni, quelli che sono differenzialmente espressi in un esperimento di *microarray*. Si supponga data una matrice di dati X, contenente i livelli di espressione $x_{jg}, g = 1, \ldots, G$ $j = 1 \ldots, n$ di G geni su n campioni biologici. Dal momento che siamo interessati a 2 classi di dati, ad esempio malati/sani, chiameremo 1 la risposta per ognuno dei campioni $n1$ nel gruppo, 2 per i campioni $n2$ nel gruppo 2. Quindi per il gene g-esimo sia avrà:

$$\text{gruppo1}: \qquad x_{1g}, x_{2g}, \ldots, x_{n1g},$$

$$\text{gruppo2}: \qquad x_{1g}, x_{2g}, \ldots, x_{n2g},$$

Sam si basa su un t-test modificato, che mette a confronto due distribuzioni attraverso la verifica dell'ipotesi stabilita:

$$\begin{array}{ll} \text{H0:} & \mu_{1g} = \mu_{2g} \\ \text{H1:} & \mu_{1g} \neq \mu_{2g} \end{array}$$

dove μ_{1g} rappresenta la media di espressione della prima popolazione, da cui proviene il gruppo1 (per esempio, i pazienti malati), mentre μ_2 rappresenta la media di espressione della seconda popolazione (esempio: i pazienti sani). Considerando che la variabile presa in considerazione riguarda l'espressione di un gene, l'ipotesi può essere letta come:
Supponiamo di avere G geni su n array sotto due differenti condizioni. Definiamo

H0: il gene g non è differenzialmente espresso
H1: il gene g è differenzialmente espresso

$\overline{x}_{1g}$ e $\overline{x}_{2g}$ l'espressione media del gene g sotto la condizione 1 e 2 rispettivamente, e definiamo s_g la deviazione standard per il g-esimo gene

$$s_g = \sqrt{(\frac{1}{n_1} + \frac{1}{n_2}) \cdot \frac{\sum_1 (x_{ig} - \overline{x}_{1g})^2 + \sum_2 (x_{ig} - \overline{x}_{2g})^2}{N - 2}} \tag{2.1}$$

Quindi, un ragionevole test statistico per l'assegnazione dell'espressione differenziale del gene è basato sulla statistica t standard :

$$t_g = \frac{\overline{x}_{2g} - \overline{x}_{1g}}{s_g}, \tag{2.2}$$

dove a numeratore c'è la differenza dei livelli medi di espressione tra il gruppo 1 e il gruppo 2 per il gene g-esimo e a denominatore la deviazione standard. Per geni con bassi livelli di espressione e poche replicazioni sperimentali, si possono ottenere varianze molto piccole, e valori t molto grandi. La procedura Sam utilizza una statistica t-modificata che prevede l'uso, a denominatore, di una piccola costante s_0, il *fudge factor*, per evitare che geni con bassi livelli di espressione dominino il risultato dell'analisi. Tusher and Chu [2001] introducono un approccio non parametrico che porta al rimpicciolimento di s_g verso un scelta adatta di s_0. La statistica Sam modificata è:

$$d_g = \frac{\overline{x}_{2g} - \overline{x}_{1g}}{s_g + s_0} \tag{2.3}$$

dove, analogamente alla statistica t_g, al numeratore vi è la differenza fra le medie delle misure relative alle due condizioni 1 e 2 per il gene g-esimo, e al denominatore vi è la somma fra la stima della deviazione standard s_g del numeratore e il valore additivo s_0, *il fudge factor*.

Specificatamente, s_0 è scelto come il percentile del valore di s_g che produce un coefficiente di variazione di d_g approssimativamente costante ad una funzione di s_g. Questo ha l'effetto aggiunto di diminuire valori grandi di d_g per geni che assumono valore pressoché costante sui due stati biologici. Questo è calcolato dal seguente algoritmo fornito da Tusher and Chu [2001]:

1. Calcolare i 100 percentili $q_k, k = 1, \ldots, 100$, dei valori di s_i.
2. dato $\alpha \in R\{0{,}0.005\}$
   - calcolare $d_i^\alpha = \frac{r_i}{(s_i + s^\alpha)}$ dove s^α rappresenta il quantile α dei valori s_i e $s^0 = q_0 = min_{i=1,\ldots,m} s_i$
   - calcolate $v_k^\alpha = 1.4826.MADd_i^\alpha \| s_i \in [q_{k-1}, q_k], k = 1, \ldots, 100$
   - calcolare il coefficiente di variazione $CV(\alpha)$ dei valori v_k^α
3. Predisporre $\hat{\alpha} = argmin_{\alpha \in R}\{CV(\alpha)\}$ e $s_0 = s^{\hat{\alpha}}$

2.1.1 Procedura Sam

Il metodo *Significance analysis of Microarray* (SamTusher and Chu [2001]), decide la soglia sopra e sotto la quale vengono definiti i geni differenzialmente espressi, sulla base dei dati. Piuttosto che usare una regola standard definita come $|d_g| > t$ per definire i geni significativi (es. avere una soglia simmetrica $\pm\delta$), Sam ricava i punti di soglia t_1 e t_2 dai dati e usa le soglie di rifiuto $d_g < t_1$ o $d_g > t_2$. Questo può portare a un test molto più potente in situazioni dove molti geni sono espressi quindi differenzialmente espressi, o vice-versa. Dati i punti di soglia, il metodo Sam stima il *false discovery rate* come segue:

1. Calcola le statistiche ordinate

$$d_{(1)} \leq d_{(2)} \ldots \leq d_{(G)}.$$

2. Fissa un numero B di permutazioni di etichette. Per ogni permutazione b calcola la statistica $d^{*b}_{(g)}$ e le corrispondenti statistiche ordinate

$$d^{*b}_{(1)} \leq d^{*b}_{(2)} \ldots \leq d^{*b}_{(G)}.$$

In pratica permuta casualmente i valori del gene tra i due gruppi e ricalcola il valore di d_g chiamandolo $d^{*b}_{(g)}$. Dalle B permutazioni, stima il valore atteso della statistica dato da:

$$\overline{d}_g = (\frac{1}{B}) \sum_{b=1}^{B} d^{*b}_g$$

per $g = 1, 2, 3, \ldots, G$.

3. Rappresenta i valori $d_{(g)}$ verso i valori $\overline{d}_{(g)}$. Per fissare la soglia Δ, si parte dall'origine, e muovendosi verso destra, si trova il primo gene g indicato con g_2 tale che: $d_{(g)} - \overline{d}_{(g)} > \Delta$. Tutti i geni oltre g_2 sono chiamati "significativamente positivi". Similmente, partendo dall'origine, muovendosi verso sinistra si trova il primo gene g indicato con g_1 tale che $\overline{d}_{(g)} - d_{(g)} \leq \Delta$. Tutti i geni prima g_1 sono dichiarati "significativamente negativi". Per ogni Δ, si definisce il punto superiore $t_2(\Delta)$ come il più piccolo d_g tra i geni "significativamente positivi", analogamente viene definito per la soglia più bassa $t_1(\Delta)$.

4. Se $t_1(\Delta) > t_2(\Delta)$ avremo quindi $t_2(\Delta) = t_1(\Delta) = 0$.

Come menzionato precedentemente, a denominatore la statistica Sam somma una costante di aggiustamento chiamata s_0 per ogni statistica d_g. In genere viene fissato s_0 uguale al valore medio di s_g. Sotto questa procedura se $\Delta' > \Delta$ quindi $t_2(\Delta') \geq t_2(\Delta)$ e $t_1(\Delta') \geq t_1(\Delta)$ si viene a creare una regione di rifiuto molto usata nella verifica d'ipotesi.

2.1.2 Stima del False Discovery Rate

Per fissare la regione di rifiuto ossia (fissare Δ) FDR e pFDR utilizzando le quantità definite come:

$$FDR(\Delta) = E\left[\frac{V(\Delta)}{R(\Delta)} \mid R(\Delta) > 0\right] Pr(R(\Delta) > 0),$$
$$pFDR(\Delta) = E\left[\frac{V(\Delta)}{R(\Delta)} \mid R(\Delta) > 0\right],$$

$$\begin{aligned} \text{con} V(\Delta) &= \#d_g\text{: gene } g\text{-esimo immutati e } d_g \leq t_1(\Delta) \text{ o } d_g \geq t_2(\Delta) \\ \text{e} R(\Delta) &= \#d_g : d_g \leq t_1(\Delta) \text{o } d_g \geq t_2(\Delta) \end{aligned} \tag{2.4}$$

Perciò $V(\Delta)$ è il numero di geni falsi positivi e $R(\Delta)$ è il numero di geni significativi per Sam, dato dall'indice Δ della soglia. Storey [2002] sviluppa la seguente stima di FDR e pFDR data da Δ:

$$\begin{aligned} \widehat{FDR}_{\Delta'}(\Delta) &= \hat{\pi}_0(\Delta') \cdot \frac{R^0(\Delta)}{R(\Delta) \vee 1}, \\ \widehat{pFDR}_{\Delta'}(\Delta) &= \hat{\pi}_0(\Delta') \cdot \frac{R^0(\Delta)}{Pr(R^0(\Delta) > 0) \cdot [R(\Delta) \vee 1]} \end{aligned} \tag{2.5}$$

$R(\Delta)$ è definito come prima ma:

$$R^0(\Delta) = \frac{\sum_{b=1}^{B} \#\{d_g^{b*} : d_g^{b*} \mid t_1(\Delta) \text{o} d_g^{b*} \geq t_2(\Delta)\}}{B},$$

$$Pr(R^0(\Delta) > 0) = \frac{\#\{b : \#\{d_g^{b*} : d_g^{b*} \mid t_1(\Delta) \text{o} d_g^{b*} \geq t_2(\Delta)\} > 0\}}{B}$$

L'ultimo $\hat{\pi}_0$ è una stima complessiva che l'ipotesi nulla sia vera (non variando i geni). Questa stima dipende dalla nostra scelta di un altro Δ'. Il metodo Sam sceglie un Δ' tale che $R^0(\Delta') = G/2$ (es: Metà delle ipotesi nulle cade nella regione definita da Δ') ma Δ' potrebbe essere anche scelto in maniera tale da

minimizzare la distorsione *bias* e la varianza tra le stime (Storey [2001b]). La stima scelta è data da:

$$\hat{\pi}_0(\Delta') = \frac{G - R(\Delta')}{G - R^0(\Delta')} \tag{2.6}$$

Il denominatore rappresenta il numero di geni non significativi per Δ' quando l'ipotesi nulla è vera; il numeratore rappresenta il numero di geni osservati che non sono significativi per Δ'. Una soddisfacente scelta di Δ' è basata nel numero di geni che non cambiano alla creazione,, così $\hat{\pi}_0(\Delta') \approx \frac{G_0}{G}$, questo è facile da dimostrare dato che $E[\hat{\pi}_0(\Delta')] \geq \frac{G_0}{G}$.
Una domanda sorge naturale su questo argomento: come scegliere la regione di rifiuto (esempio: il valore Δ). Possiamo usare una stima dei punti seguendo queste 3 opzioni.

1. Per un Δ fissato e calcolare $\widehat{FDR}_{\Delta'}(\Delta)$ o $\widehat{pFDR}_{\Delta'}(\Delta)$. Quindi per un grande G e una debole dipendenza , o se i geni immutati sono indipendenti abbiamo

   $$E[\widehat{FDR}_{\Delta'}(\Delta)] \geq FDR(\Delta)$$

   e

   $$E[\widehat{pFDR}_{\Delta'}(\Delta)] \geq pFDR(\Delta)$$

2. Scegliere un livello α anticipatamente al FDR di controllo da noi desiderato. Prendere il più piccolo Δ in modo tale che $\hat{FDR}_{\Delta'}(\Delta) \leq \alpha$. Per un grande G e una debole dipendenza tra i geni, questo garantisce che il $\widehat{FDR}(\Delta) \leq \alpha$. Se i geni immutati sono indipendenti e il limite Δ che consideriamo è a $\delta \geq \Delta'$, quindi incuranti della dimensione di G, avremo $FDR(\Delta) \leq \alpha$.

3. Calcolare $\widehat{FDR}_{\Delta'}(\Delta)$ e $\widehat{pFDR}_{\Delta'}(\Delta)$ e mantenendo per entrambi $\Delta > 0$.

Quindi per un grande G e una debole dipendenza, avremo:

$$\overset{min}{\Delta} [\widehat{FDR}_{\Delta'}(\Delta) - FDR(\Delta)] \geq 0$$

e

$$\overset{min}{\Delta} [\widehat{pFDR}_{\Delta'}(\Delta) - pFDR(\Delta)] \geq 0.$$

La stima dei geni significativamente espressi dipende dalla soglia scelta. Generalmente si determina la soglia Δ in base al FDR o pFDR che si vuole. Un basso FDR corrisponde a Δ più alti, di conseguenza il numero di geni classificato come differenzialmente espressi diminuisce. Sam cerca di utilizzare il valore di soglia, irrobustito da una valutazione statistica del risultato ottenuto. Il metodo iterativo per individuare le soglie, garantisce un maggiore controllo sui dati, dato che la loro fluttuazione viene controllata. La permutazione, consente di considerare indipendenti i dati differenzialmente espressi e sottoespressi, così facendo, si ottengono due soglie che possono portare ad un intervallo simmetrico. L'uso di permutazioni, negli esperimenti che utilizzano un numero piccolo di campioni, migliora la qualità dell'informazione, generando insiemi di dati, coerenti con l'esperimento realizzato. Purtroppo, nel caso opposto, ossia con campioni molto numerosi o con molti geni per microarray, il carico computazionale diviene estremamente oneroso e può risultare ingestibile se non si dispone di un adeguato supporto *hardware*. L'utilizzo di un parametro statistico come il *False Discovery Rate*, permette un'immediata stima del livello di affidabilità dell'insieme di geni selezionati come differenzialmente espressi, evidenziando la percentuale di errori di I tipo che si commette scegliendo il valore di Δ.

2.2 Statistica t-moderata

La statistica B o fattore di *Bayes* appare per la prima volta nel modello di Smyth [2004] che sulla base del modello gerarchico parametrico di *Lönnsted and Speed [2002]*, elabora l'espressione del logaritmo a posteriori per la discriminazione dei geni differenzialmente espressi.
La Statistica B si prefigge di identificare i geni differenzialmente espressi, attraverso le probabilità a posteriori. Sulla base del modello sviluppato da *Lönnsted and Speed [2002]*, Smyth [2004], cerca di generalizzare la Statistica B *(odds a posteriori)* per poterla applicare a tutti i tipi di disegno sperimentali.
Gli esperimenti presi in considerazione da Smyth [2004] sono rappresentabili, per ogni gene da un modello lineare:

$$Y_g = X\alpha_g + \varepsilon_g \tag{2.7}$$

dove Y_g è un vettore risposta dei *log-ratio* normalizzati, X è una matrice nota, α_g è il vettore di coefficienti gene-specifico ed $\varepsilon_g \sim N(0, \sigma_g^2)$. Supponendo di disporre di n osservazioni, la variabile risposta per il g-esimo gene viene indicata come segue:

$$Y_g^T = (y_{1g}, y_{2g}, y_{3g}, \ldots, y_{ng}). \tag{2.8}$$

Dati i parametri, tutti i geni e le replicazioni si assumono indipendenti e il valore medio dell'espressione del gene g è dato da:

$$E(Y_g) = X\alpha_g = \mu_g, \tag{2.9}$$

dove μ_g è il valore medio dell'espressione del gene g e la varianza è data da:

$$Var(Y_g) = W_g\sigma_g^2, \tag{2.10}$$

dove W_g è una matrice nota di pesi definita non negativa. Si noti che il vettore y_g potrebbe avere dei valori mancanti e la matrice W_g potrebbe avere sulla diagonale pesi che sono uguali a zero. Spesso, si incorre in contrasti d'interesse biologico definiti da: $\beta_g = C^T\alpha_g$; assumiamo che il valore del contrasto β_{jg} sia uguale a zero cioè che la probabilità associata a tale evento sia la seguente: dove p_j indica

Tabella 2.1: Ipotesi

$\beta_{jg} = 0$	con Probabilità $1 - p_j$
$\beta_{jg} \neq 0$	con Probabilità p_j

la proporzione di geni differenzialmente espressi.
Per ottenere gli stimatori dei coefficienti, Smyth [2004] assume che il modello lineare sia adeguato alla variabile risposta per ogni gene, così facendo si ricavano gli stimatori dei coefficienti: $\hat{\alpha_g}$, s_g^2, σ_g^2, e la stima della matrice di varianze e covarianze degli stimatori $\hat{\alpha}_g$, data da:

$$Var(\hat{\alpha_g}) = V_g s_g^2$$

con V_g è una matrice definita positiva e indipendente da s_g^2 .
Gli stimatori dei contrasti definiti da $\hat{\beta}_g = C^T\hat{\alpha}_g$ con matrice di covarianza

$$Var(\hat{\beta}_g) = C^T V_g C s_g^2,$$

sono distribuiti come segue:

$$\hat{\beta}_g \approx N(\beta_g, C^T V_g C \sigma_g^2)$$

e la varianza residua $s_g^2 \sim \chi^2$. Essendo v_{jg} il g-esimo elemento diagonale della matrice V_g, le assunzioni delle distribuzioni degli stimatori si possono riassumere come:

$$\hat{\beta}_g | \beta_g, \sigma_g^2 \sim N(\beta_{jg}, V_{jg}\sigma_g^2) \tag{2.11}$$

e

$$s_g^2 | \sigma_g^2 \sim \frac{\sigma_g^2}{k_g} \chi_{k_g}^2,$$

dove k_g rappresenta i gradi di libertà residui del modello lineare per il gene g-esimo. Sotto questi presupposti la statistica t_{jg} segue una distribuzione t-student con v_g gradi di libertà

$$t_{jg} = \frac{\hat{\beta}_{jg}}{s_g \sqrt{V_{jg}}} \quad \sim t_{k_g}. \tag{2.12}$$

È importante trarre vantaggio dalla struttura parallela dei dati. Lo scopo è capire come i coefficienti ignoti β_{jg} e le varianze ignote σ_g^2 varino tra i geni, introducendo delle distribuzioni a priori sui parametri. Si suppone che la distribuzione a priori per σ_g^2 sia un χ^2 inverso con k_0 gradi di libertà:

$$\frac{1}{\sigma_g^2} \sim \frac{1}{k_0 s_0^2} \sim \chi_{k_0}^2,$$

dove s_0^2 è l'iperparametro per la distribuzione di σ_g^2. Sotto le stesse ipotesi formulate nella Tabella 2.1, dove p_j rappresenta la proporzione attesa di geni differenzialmente espressi, la distribuzione attesa per i geni differenzialmente espressi, $P(\beta_{jg} \neq 0) = p_j$, viene assunta uguale a quella di un'osservazione a priori uguale a zero con varianza pari a $V_{j0}\sigma_g^2$, di conseguenza costituisce una a priori coniugata

per il modello normale:

$$\beta_{jg}|\sigma_g^2, \beta_{jg} \neq 0 \sim N(0, V_{j0}\sigma_g^2). \tag{2.13}$$

La formula 2.13 costituisce un'a priori coniugata per il modello normale in 2.11. Con questo modello, la media a posteriori di $\sigma_g^2|s_g^2$ diventa:

$$\tilde{s}_g^2 = E(\sigma_g^2|s_g^2) = \frac{k_0 s_0^2 + k_g s_g^2}{k_0 + k_g}.$$

I valori della varianza calcolata a posteriori sembrano essere molto simili ai valori a priori. Viene così definita la statistica t-moderata:

$$\tilde{t}_{jg} = \frac{\hat{\beta}_{jg}}{\tilde{s}_g \sqrt{V_g}}. \tag{2.14}$$

Questa statistica rappresenta un approccio bayesiano, nel quale le usuali varianze campionarie della statistica t vengono sostituite dalle varianze a posteriori. La statistica t-moderata ($\tilde{t}_{jg}$) ha distribuzione indipendente da s_g^2 (Smyth [2004]), di conseguenza tale statistica sotto $H0 : \beta_{\text{jg}} = 0$ si distribuisce come una t-student con gradi di libertà $k_g + k_0$. La somma dei gradi di libertà per $\tilde{t}_{jg}$ al posto di t_{jg} esprime l'informazione aggiuntiva data dal modello gerarchico.
La statistica t-moderata viene utilizzata per fare il primo confronto con la statistica Sam.

2.2.1 Statistica B

Date le distribuzioni marginali di $\tilde{t}_{jg}$ e s_g^2, si ottiene la quota a posteriori indicata con O. Quest'ultima è utilizzata per l'ordinamento dei geni differenzialmente espressi:

$$\begin{aligned} O_{jg} &= \frac{\Pr(\beta_{jg} \neq 0 | \tilde{t}_{jg}, s_g^2)}{\Pr(\beta_{jg} = 0 | \tilde{t}_{jg}, s_g^2)} \\ &= \frac{\Pr(\beta_{jg} \neq 0, \widetilde{t}_{jg}, s_g^2)}{\Pr(\beta_{jg} = 0, \widetilde{t}_{jg}, s_g^2)} \\ &= \left(\frac{p_j}{1 - p_j} \frac{\Pr(\tilde{t}_{jg} | \beta_{jg} \neq 0)}{\Pr(\tilde{t}_{jg} | \beta_{jg} = 0)} \right). \end{aligned} \tag{2.15}$$

Dato che $\tilde{t}_{\mathrm{g}}$ e s_g^2 sono indipendenti e s_g^2 non dipende da β_{jg}, sostituendo la quantità $\tilde{t}_g$ si ottiene :

$$O_{jg} = \frac{p_j}{1 - p_j} \left(\frac{v_{jg}}{v_{jg} + v_{j0}} \right)^2 \left(\frac{\tilde{t}_{jg}^2 + k_0 + k_g}{\tilde{t}_{jg}^2 \frac{v_{jg}}{v_{jg} + v_{0j}} + k_0 + k_g} \right)^{\frac{(1 + k_0 + k_g)}{2}}. \tag{2.16}$$

In accordo con Lönnstedt [2005], si ha

$$O_{jg} = \frac{p_j}{1 - p_g} \left(\frac{v_{jg}}{v_{jg} + v_{j0}} \right)^{\frac{1}{2}} \exp\left(\frac{\widetilde{t}_{jg}^2}{2} \frac{v_{jg}}{v_{j0} + v_{j0}} \right)$$

Lönnstedt [2005] propongono la seguente scala preferibile a O:

$$B_{jg} = \log\left(O_{jg}\right), \tag{2.17}$$

I dati si assumono normalizzati a rimozione del *dye bias* (Huber and Heydebreck [2002]). Gli iperparametri s_0, k_0, V_{j0} del modello gerarchico vengono stimati dai dati (Smyth [2004]); si calcolano delle stime partendo dalle varianze campionarie s_g^2 e dalle statistiche t-moderate $\tilde{t}_{jg}$. Gli iperparametri s_g^2, d_0 vengono stima-

ti mediante il metodo dei momenti, eguagliando i valori empirici dei primi due momenti della quantità $\log s_g^2$ a quelli attesi. È preferibile usare $\log s_g^2$ perchè i momenti della funzione $\log s_g^2$ sono definiti per qualsiasi grado di libertà e la sua distribuzione è molto più prossima ad una distribuzione Normale. Le stime di V_{0j}, invece, si ottengono eguagliando le statistiche ordinate $\left|\tilde{t}_{jg}\right|$ ai loro valori nominali. Concludendo, l'odds dell'espressione differenziale ottenuto dalle replicazioni $Y_g^T = (y_{1g}, y_{2g}, y_{3g}, \ldots, y_{ng})$ è il seguente:

$$O_g = \frac{p f_0(x_{n1g}) f_0(x_{n2g})}{(1-p) f_0(x_{n1g} x_{n2g})} \tag{2.18}$$

Capitolo 3

Confronto tra ordinamenti Statistica t e Sam

In questo capitolo verranno esplorate le differenze tra la statistica t e la statistica Sam nell'ordinamento dei geni.
Per procedere a questo confronto è necessario simulare le espressioni di un fissato numero di geni. Si è scelto di simulare $m = 5000$ geni per $n = 30$ soggetti, di cui 15 appartenenti al gruppo 1 e 15 al gruppo 2, utilizzando il modello LogNormale-Normale. In tale modello si ipotizza che la distribuzione relativa alla trasformata logaritmica della singola espressione genica sia normale; inoltre, si assume che il coefficiente di variazione sia costante per tutti i geni analizzati. La simulazione prevede la creazione di una matrice, relativa all'espressione di 5000 geni, accompagnata da un vettore (TRUE,FALSE) contenente l'indicatore per l'uguale o diversa espressione degli stessi. La proporzione dei geni differenzialmente espressi costituisce un parametro a priori del modello di simulazione. Ai geni contenuti nella matrice dei dati simulata, di cui si riportano a titolo esemplificativo le prime due righe in Tabella 3.1, vengono applicate entrambe le statistiche (Statistica t e Sam) al fine di confrontare l'ordinamento dei geni indotto da queste. Dal confronto ci

si aspetta un elenco ordinato delle statistiche con risultati molto rassomigliante, ovvero, ci si attende che il posizionamento dei geni sia pressoché simile.

ESEMPIO L'analisi Sam è condotta in *R* mediante una funzione denominata *Sam* che produce i valori della statistica, indicati con tt, secondo l'espressione definita da 2.3.

Tabella 3.1: Esempio di due vettori di valori simulati attraverso il modello LogNormale-Normale

	Vettori simulati				
Vet	Array1	Array2	Array3	Array4	Array5
1	1375.20130	928.34789	618.4948	836.6578	2449.743
2	17.83708	75.89608	96.4696	841.6083	516.908
Vet	Array6	Array7	Array8	Array9	Array10
1	5688.6095	6108.3971	926.14526	3919.577	566.0107
2	530.4466	76.8493	79.97694	495.510	242.9652
Vet	Array11	Array12	Array13	Array14	Array15
1	849.59205	3171.0374	5611.7356	831.93504	5406.0518
2	93.53275	240.2117	133.0847	87.88198	188.3337
Vet	Array16	Array17	Array18	Array19	Array20
1	689.8100	564.7635	5463.36297	2516.4695	3518.488
2	550.7303	70.5422	42.89879	134.3112	277.958
Vet	Array21	Array22	Array23	Array24	Array25
1	1221.867	1710.0878	3531.7236	1444.72934	3610.28817
2	770.005	824.5962	153.8877	79.53863	54.95756
Vet	Array26	Array27	Array28	Array29	Array30
1	5254.6711	9946.8500	4770.0874	861.5947	313.1627
2	166.3142	186.9594	116.6845	275.8799	232.7087

A titolo esemplificativo, si riportano in Tabella 3.2, e in Tabella 3.3, i valori osservati della statistica per i 2 geni precedentemente mostrati in Tabella 3.1.

Tabella 3.2: Valori osservati e livello di significatività stimato per i 2 geni in Tabella 3.1

	Valori Samr	
	vett1	vett2
t-moderata	-0.4680639	-0.1605618
p-value	0.674536	0.887174

La statistica t è reperibile tramite l'utilizzo della funzione *eBayes* contenuta nella libreria *Limma*. Per completezza, si riportano in Tabella 3.3 i valori osservati della statistica t per i geni in Tabella 3.1

Tabella 3.3: I valori della statistica t calcolati per i 2 geni contenuti nella Tabella 3.1

	Valori Statistica t	
	vett1	vett2
t	4.981604	3.978449
p-value	2.719119e-05	4.284312e-04

3.1 Ordinamento dei geni nelle due statistiche

Dopo aver simulato i geni, si procede con l'analisi degli ordinamenti effettuati dalle due statistiche Sam e t. La proporzione p di geni differenzialmente espressi è stata posta uguale a 0.1. Nel nostro caso, avendo una matrice di 5000 geni, ci si aspetta che la proporzione di geni differenzialmente espressi simulati sia, in media, uguale a 500. Entrambe le statistiche utilizzano un test a due code, quindi per identificare i geni differenzialmente espressi vengono presi in considerazione i primi 250 e gli ultimi 250 geni nell'ordinamento. Per confrontare i due ordinamenti, e per una più chiara rappresentazione grafica, è utile considerare la posizione di

un gene nell'ordinamento. Nella 3.1, per 10 geni vengono mostrate, a titolo esemplificativo, le posizioni nel vettore ordinato per la statistica t, e le corrispondenti posizioni nel vettore ordinato della statistica Sam. Le posizioni nei due ordinamenti vengono collegate da un segmento per una più chiara lettura.

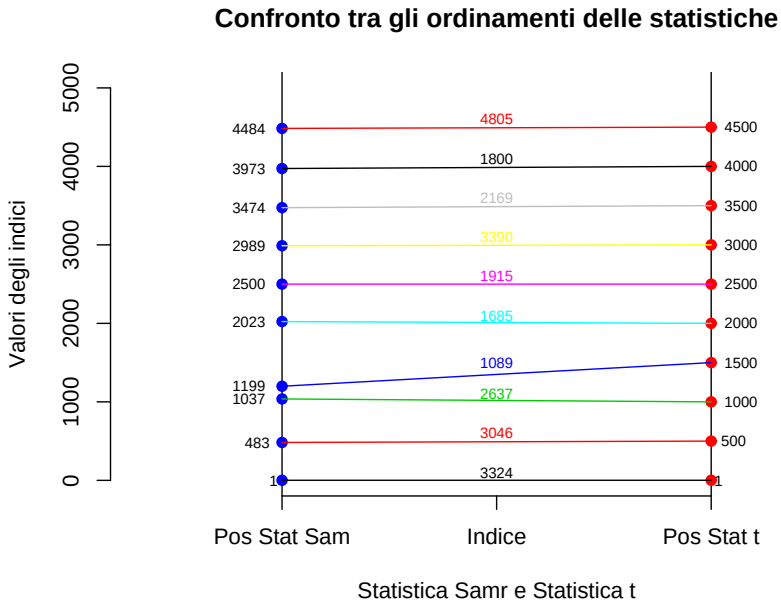

Figura 3.1: Ordinamento geni

I 10 geni considerati in Figura 3.1, hanno ranghi simili nell'ordinamento. È interessante focalizzare il confronto sull'ordinamento dei geni differenzialmente espressi, che rappresentano la parte di geni che, in genere, si è più interessati a studiare.

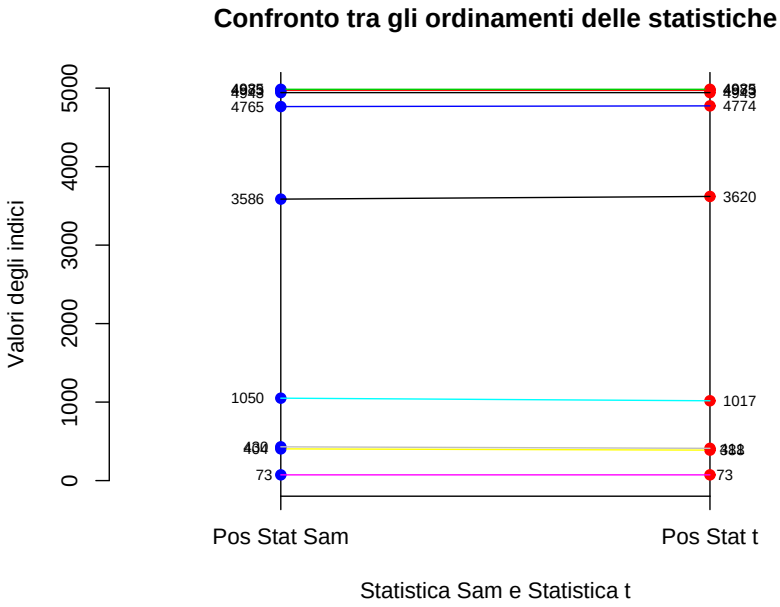

Figura 3.2: Ordinamento dei geni differenzialmente espressi.

Il grafico dei confronti tra gli ordinamenti delle statistiche (Figura 3.2) riporta 10 geni differenzialmente espressi e la loro posizione, dopo l'ordinamento, sembra essere molto simile. La Figura 3.2 segnala che l'ordinamento dei geni per le due statistiche sembra essere rassomigliante. Infatti, osservando le corrispondenze tra le posizioni della statistica Sam e la statistica t, emergono che a stessi geni fanno riferimento posizioni non molto differenti. È utile illustrare graficamente le densità delle due statistiche per i geni differenzialmente espressi e non. Idealmente, i valori stimati della statistica per i geni differenzialmente espressi dovrebbero avere una distribuzione bimodale, mentre per i restanti geni dovrebbero avere un andamento a campana simile a quello di una distribuzione normale.

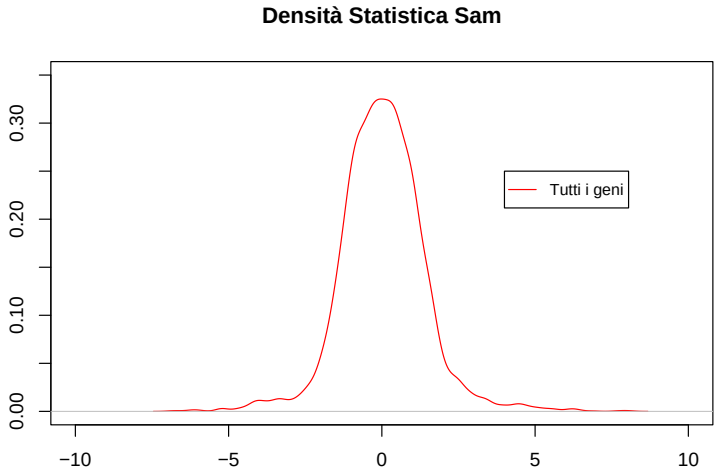

Figura 3.3: Stima non parametrica della densità della statistica Sam su tutti i geni

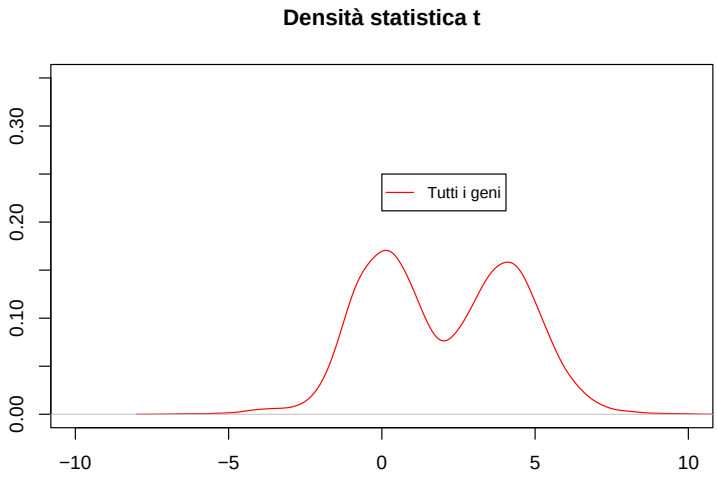

Figura 3.4: Stima non parametrica della densità della statistica t su tutti i geni

Notiamo, dalla Figura 3.3, che per la statistica Sam la densità sembra essere ideale. Infatti, per i geni differenzialmente espressi mostra due mode: come ci si aspettava, i geni differenzialmente espressi producono valori per la statistica che si trovano sulla coda della distribuzione rispetto ai geni non espressi. Questi producono valori che trovano il loro massimo in una moda centrale posizionata nel minimo locale tra le due mode dei geni differenzialmente espressi. Nella Figura 3.4 la statistica t sembrerebbe avere un andamento meno chiaro. Infatti, per i geni non espressi, la densità della statistica ha due mode, mentre per i geni differenzialmente espressi (Figura 3.6) mostra un andamento che ha una distribuzione che si spalma lungo tutto l'asse e non evidenzia una moda. Nelle Figure 3.5 e 3.6 vengono rappresentate le densità dei geni differenzialmente espressi e non espressi per entrambe le statistiche, in maniera tale da rendere più evidenti le differenze tra di esse.

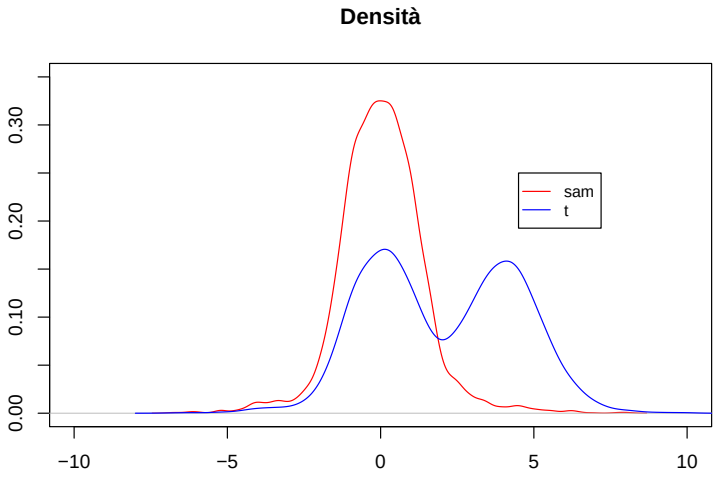

Figura 3.5: Stima non parametrica della densità della statistica Sam e t su tutti i geni

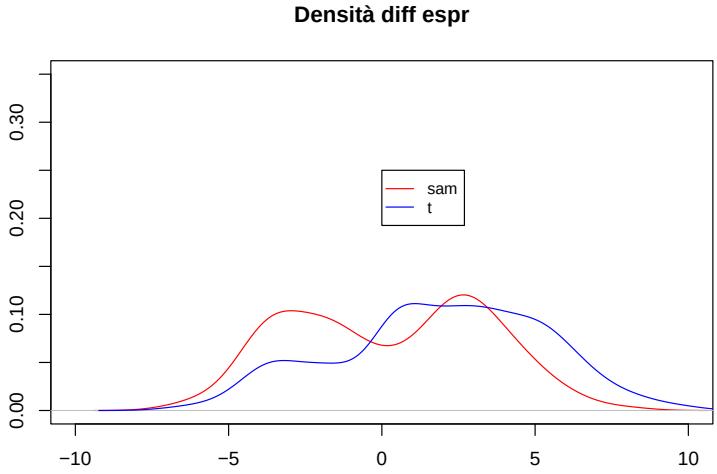

Figura 3.6: Stima non parametrica della densità della statistica Sam e t su i geni differenzialmente espressi

Per studiare le capacità discriminanti delle due statistiche, vengono riportate le matrici di confusione calcolate su 100 replicazioni della simulazione. Dall'analisi delle Tabelle 3.4 e 3.5 contenenti i valori medi delle 100 replicazioni, si può notare che i valori tendono ad assestarsi a quote simili. I geni non espressi vengono identificati abbastanza bene da entrambe le statistiche (Stat t: 4079, Stat Sam: 4074), mentre per i geni differenzialmente espressi sia la statistica t che la statistica Sam sembrerebbero non produrre un risultato ottimale. Per un'analisi più approfondita analizziamo le matrici di correlazione per i livelli di significatività delle due statistiche.

Tabella 3.4: Matrice di confusione per i geni differenzialmente espressi e non espressi identificati dalla statistica t (medie di 100 replicazioni)

GENI	False: Non espressi	True: Diff espressi
False: Non espressi	4079	453
True: Diff espressi	421	47

Tabella 3.5: Matrice di confusione per i geni differenzialmente espressi e non espressi identificati dalla statistica Sam (medie di 100 replicazioni)

GENI	False: Non espressi	True: Diff espressi
False: Non espressi	4074	458
True: Diff espressi	426	42

Tabella 3.6: Matrice di confusione per i geni differenzialmente espressi e non espressi identificati dalla statistica t con i livelli di espressione (medie di 100 replicazioni)

GENI	False: Non espressi	True: Diff espressi
False: Non espressi	4051	450
True: Diff espressi	449	50

Tabella 3.7: Matrice di confusione per i geni differenzialmente espressi e non espressi identificati dalla statistica Sam attraverso i livelli di espressione

GENI	False: Non espressi	True: Diff espressi
False: Non espressi	4216	284
True: Diff espressi	468	32

Dalle Tabelle 3.6 e 3.7, t sembra classificare correttamente un maggior numero di geni (Sam: 32, t: 50), mentre nelle Tabelle 3.5 e 3.4 contenenti le matrici di confusione, le due statistiche identificavano un numero simile di geni differenzialmente espressi. Dalle matrici di correlazione si può concludere che entrambe le statistiche non sembrano riconoscere un numero considerevole di geni differenzialmente espressi, e che sia t che Sam si comportano in modo simile. Il sostanziale accordo tra gli esiti dell'analisi Sam e t è visibile anche dall'analisi degli indici di correlazione dei livelli di significatività osservati.
Gli indici di correlazione, ottenuti da 3 simulazioni di 5000 geni, tra la statistica Sam e t sono riportati in Tabella 3.8 mentre in Figura 3.7 vengono rappresentati i diagrammi di dispersione per ogni simulazione. I valori della tabella ci portano alla conclusione che esiste una fortissima correlazione tra i risultati delle due analisi: i valori dell'indice di correlazione tra i valori osservati delle due statistiche indica una correlazione positiva molto alta, dimostrando che le due statistiche lavorano in maniera simile.

Tabella 3.8: Indici di correlazione

Correlazione tra:	Indici di correlazione per i valori della statistica t e Sam 1simul.	2simul.	3simul.
p-value t p-value Sam	0,99215	0,99466	0,99291
t-mod(stat t) tt(Sam)	0,99844	0,99893	0,99861

È interessante notare che i livelli di significatività osservati dalla statistica t sembrano sempre essere maggiori o uguali a quelli della statistica Sam.

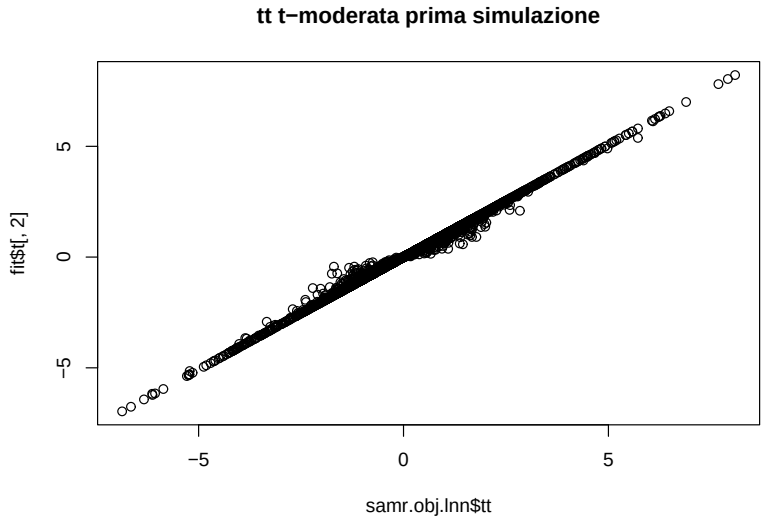

Figura 3.7: Diagramma di dispersione dei valori simulati(I simulazione)

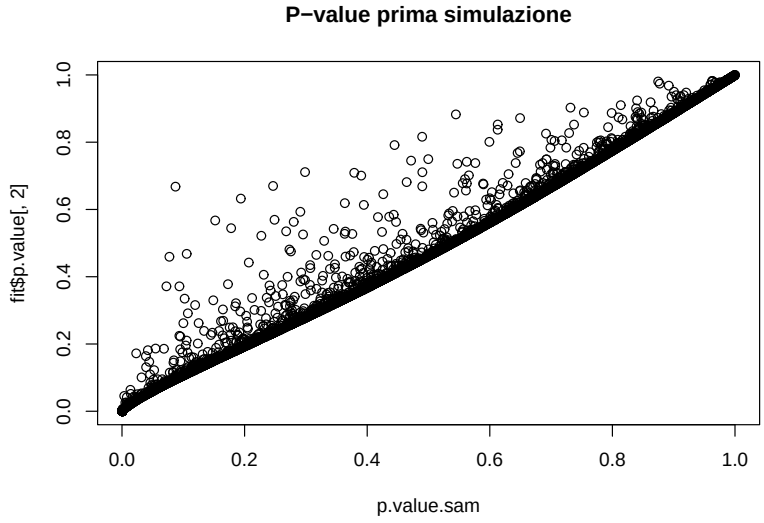

Figura 3.8: Diagramma di dispersione dei livelli di espressione(I simulazione)

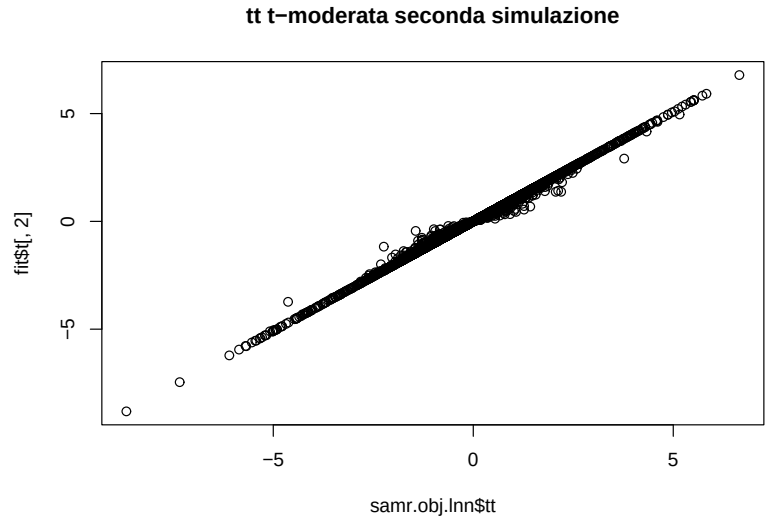

Figura 3.9: Diagramma di dispersione dei valori simulati(II simulazione)

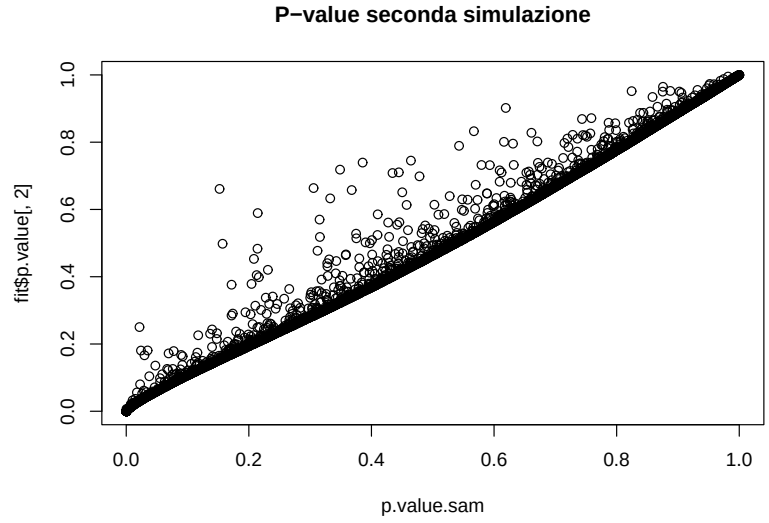

Figura 3.10: Diagramma di dispersione dei livelli di espressione(II simulazione)

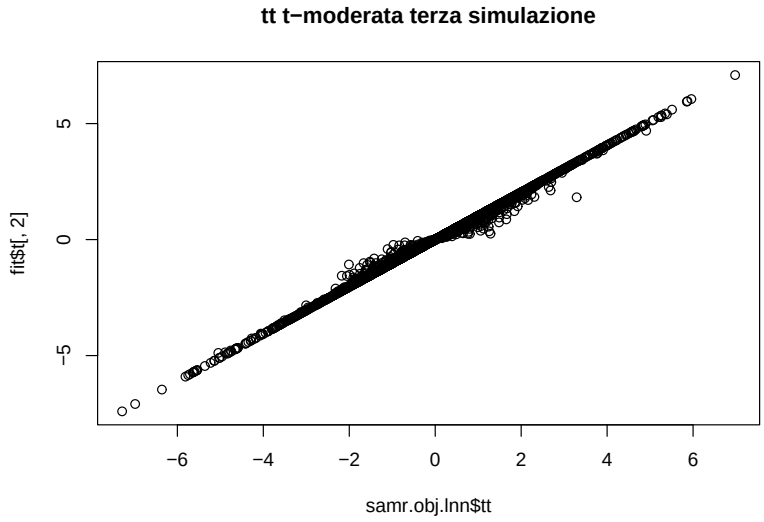

Figura 3.11: Diagramma di dispersione dei valori simulati(III simulazione)

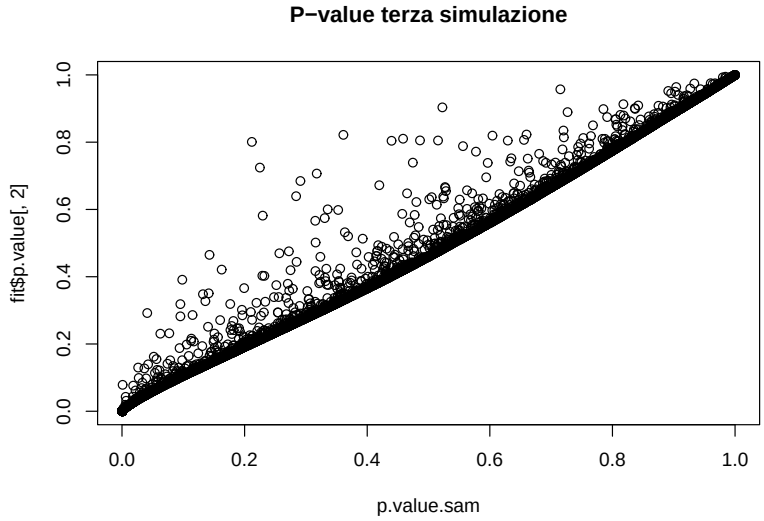

Figura 3.12: Diagramma di dispersione dei livelli di espressione(III simulazione)

Capitolo 4

Ordinamenti Statistica B e Sam

In questo capitolo verranno trattate le differenze tra la statistica Sam e la statistica B, proposta da Lönnstedt [2005]:

$$B_{jg} = \log\left(O_{jg}\right) \tag{4.1}$$

Il risultato usato è il logaritmo del rapporto per i due colori oppure la log intensità dei singoli canali di dati.

4.1 Confronto tra la statistica B e Sam per i dati simulati

In Figura 4.1 vengono riportati gli ordinamenti delle due statistiche. Il grafico ci mostra che l'ordinamento della statistica B cambia rispetto alla statistica t. Le differenze nel posizionamento dei geni sembrano aumentare: per alcuni geni l'ordinamento sembra essere opposto, infatti a indici molto alti nella statistica Sam corrispondono indici piccoli nella statistica B. Questo comportamento è confermato anche dalla Figura 4.2 che rappresenta l'ordinamento dei geni differen-

zialmente espressi. In alcuni casi, cambia la posizione assegnata: Sam assegna ai geni indici alti e bassi, al contrario B associa, quasi esclusivamente, indici bassi. Questo è supportato dalla Figura 4.1 e dal confronto dei geni differenzialmente espressi (Figura4.2). In conclusione come per la statistica t, precedentemente discussa, anche la statistica B riconosce i geni differenzialmente espressi essendo quest'ultima funzione della statistica t.

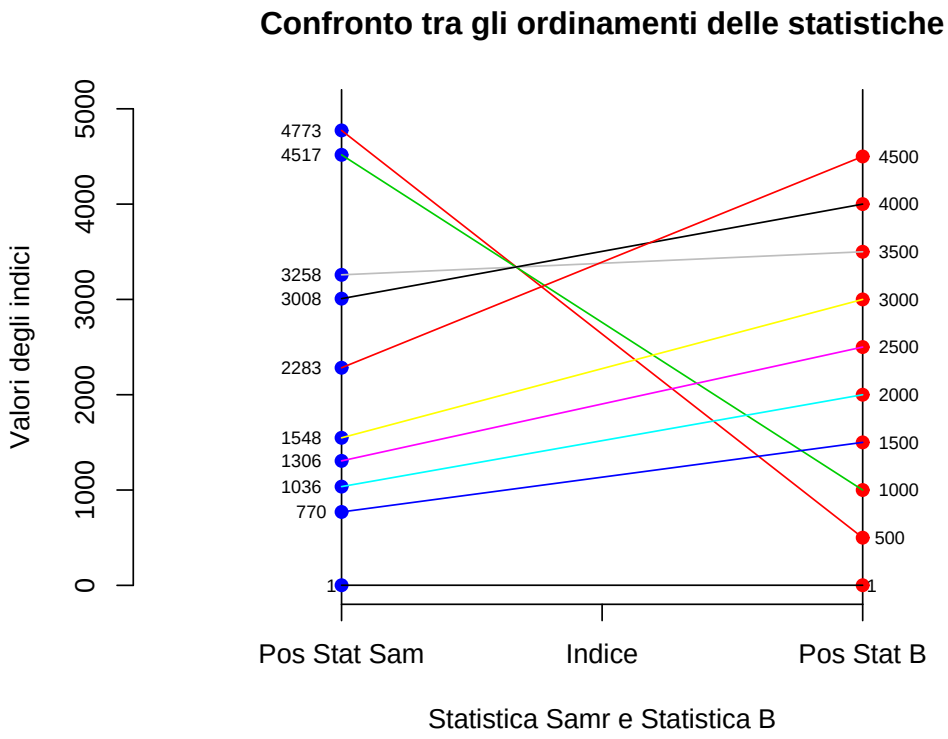

Figura 4.1: Confronto tra gli ordinamenti per Sam e B.

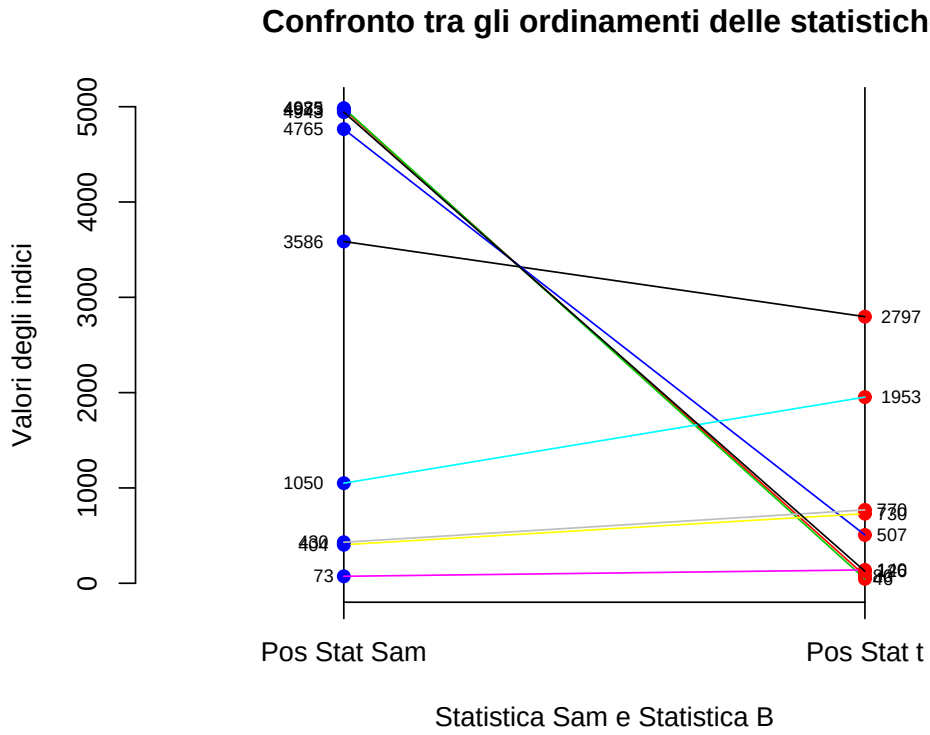

Figura 4.2: Confronto dei geni differenzialmente espressi per Sam e B.

In Figura 4.3 viene riportata la stima non parametrica della densità di tutti i geni per la statistica B che a differenza della statistica t, presentata in Figura 3.4 nella pagina 36, ha un solo punto di massimo. La stima della densità dei geni differenzialmente espressi per la statistica B sembra concentrarsi nell'intervallo $[-2.198; -2{,}194]$ quindi nella parte iniziale del grafico, che coincide con la densità dei non espressi.

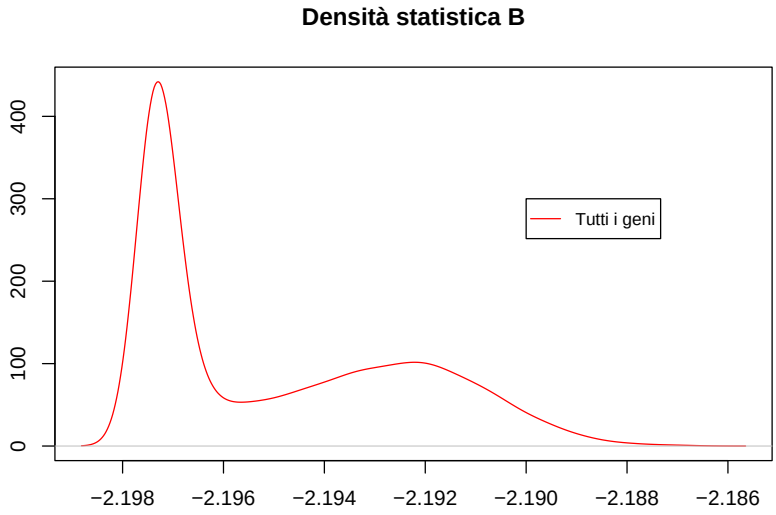

Figura 4.3: Stima non parametrica della densità di tutti i geni per la statistica B.

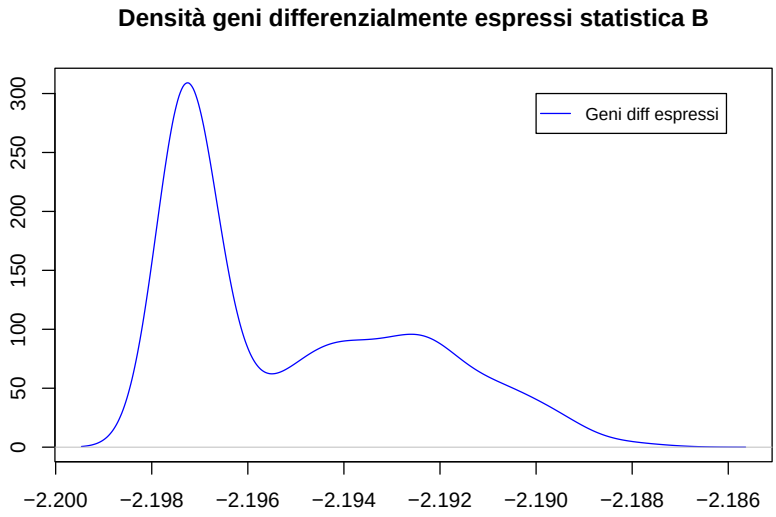

Figura 4.4: Stima non parametrica della densità dei geni espressi per la statistica B.

Tabella 4.1: Matrice di confusione per i geni differenzialmente espressi e non espressi identificati dalla statistica B

GENI	False: Non espressi	True: Diff espressi
False: Non espressi	4051	450
True: Diff espressi	449	50

Per verificare se la statistica B identifica un numero diverso di geni differenzialmente espressi, riportiamo in Tabella 4.1 la matrice di confusione per i dati simulati, contenente i valori medi per 100 replicazioni. Sia la statistica B che la statistica t identificano un numero di geni differenzialmente espressi molto basso, come ci si aspettava dato che B è una funzione di t, infatti la Tabella 4.1 riporta dei risultati molto simili a quelli della Tabella 3.6 per la statistica t.

Tabella 4.2: Indice di correlazione tra i valori associati ai geni dalla statistiche Sam e B

	Correlazione
statistica tt e B	0.02414133

Sebbene la statistica B riconosce un numero di geni differenzialmente espressi pari a quello della statistica Sam, la correlazione tra le due statistiche risulta essere bassa. La Figura 4.5 ci illustra una distribuzione di tipo simmetrico rispetto allo zero, dove ha la sua massima concentrazione.

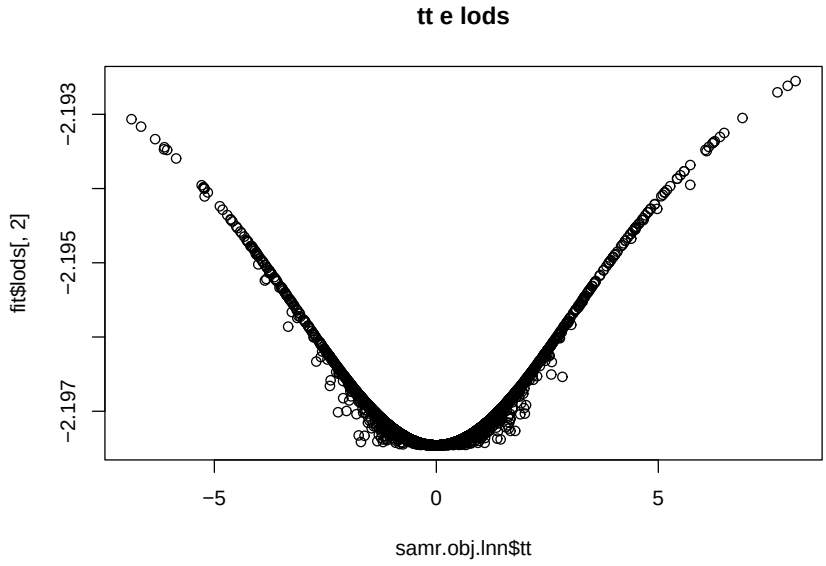

Figura 4.5: Diagramma di dispersione dei dati simulati per la statistica B e Sam

Capitolo 5

Leucemie: le due statistiche applicate alla realtà

Le leucemie sono neoplasie delle cellule progenitrici dell'ematopoiesi, le quali perdono le loro capacità di maturare e differenziarsi, ciò che conduce ad una proliferazione incontrollata[Cavalli [2006]]. La patologia ha sede iniziale nel midollo osseo e si divide in acuta o cronica: acuta se la proliferazione interessa cellule non in grado di differenziarsi e crescere completamente, croniche nel caso in cui la proliferazione interessi cellule capaci di differenziarsi e maturare. L'incidenza globale di tutte le leucemie è di circa 10 nuovi casi all'anno per 100.000 abitanti. Salvo la leucemia linfatica acuta (LLA), che è prevalentemente una malattia infantile e adolescenziale, le altre leucemie, sia acute che croniche, sono una malattia dell'età avanzata. Fino a una quindicina di anni fa la valutazione prognostica delle leucemie acute si basava principalmente sulla situazione clinica, oggi i fattori prognostici principali sono di natura biologica: citogenetica, favorevoli o sfavorevoli (LLA,LLC), presenza di proteine di membrana con funzione biologica negativa (ad es. FLT3 nella LMA), presenza o meno del riarrangiamento del gene delle catene pesanti (LLC), ecc, [Cavalli [2006]]. La nuova valutazione prognostica ha

portato ad essere molto utili gli esperimenti di microarray svolti su soggetti affetti da leucemie, individuando i geni differenzialmente espressi tra soggetti sani e malati o tra persone colpite dalla diverse patologie.

5.1 Leucemie croniche

Le leucemie croniche si dividono in mieloidi e linfoidi. I dati trattati contengono un solo tipo di leucemia mieloide acuta AML*(Acute myeloid leukemia)*, mentre le leucemia linfoidi si distinguono in:

- LLA-B *(acute linfoblastic leukemia B)* quando sono interessati i linfociti di tipo B
- LLA-T *(acute linfoblastic leukemia T)* quando sono intaccati i linfociti di tipo T
- LLA-Tr *(acute linfoblastic leukemia traslocate)* che comporta una particolare disomogeneità cromosomica dovuta alla traslocazione di alcuni frammenti di DNA all'interno dei cromosomi.

5.1.1 Leucemia linfoblastica acuta LLA- T

Le leucemie rappresentano la più comune neoplasia maligna nei bambini, ammontano a circa $1/3$ di tutti i tumori pediatrici. Approssimativamente $3/4$ di tutti i casi di leucemia in età pediatrica sono di tipo LLA. Il picco d'incidenza della LLA è compreso tra i 2 e i 5 anni. L'incidenza è leggermente più alta nel sesso maschile che in quello femminile, e questa discriminanza si fa più evidente durante l'età adolescenziale. La predominanza maschile è più marcata nei casi di leucemia di tipo LLA a cellule T.

L'analisi molecolare delle più comuni alterazioni genetiche, oltre a dimostrare l'origine della malattia da un unico clone (clonalità), ha contribuito a comprendre meglio la patogenesi LLA. La patogenesi LLA, produce delle alterazioni che si manifestano con delle trasformazioni in senso neoplastico delle cellule staminali emopoietiche o dei loro pro-genitori *"committed"*mediante il mutamento di alcune funzioni cellulari, come il mantenimento o la promozione di una illimitata capacità di auto-rinnovamento, il sovvertimento del controllo della normale proliferazione, il blocco della differenziazione e la promozione della resistenza ai segnali di morte cellulari programmata (apoptosi). La LLA può manifestarsi con emorragie o infezioni, astenia, letargiadolori ossei, artalgie o rifiuto della deambulazione. Si tratta di una malattia primaria a carico del midollo osseo, qualunque organo o tessuto può essere infiltrato dalle cellule tumorali. Il sottogruppo delle LLA a cellule T *(acute linfoblastic leukemia T)* (10-15% dei casi), rispetto alle LLA a precursori B *(acute linfoblastic leukemia)*, è caratterizzato da predominanza nel sesso maschile, massa mediastinica in cira metà dei pazienti, età mediana più alta, conta leucocitaria più elevata e normali livelli di emoglobina alle diagnosi, ed infine le leucemie di tipo LLA-Tr *(acute linfoblastic leukemia traslocate)* che comporta una particolare disomogeneità cromosomica dovuta alla traslocazione di alcuni frammenti di DNA all'interno dei cromosomi. Di seguito vengono riportate gli stadi di differenziazione linfoide e immunofenotipizzazione nella LLA in età pediatrica a precursori T e B.

- AUL
  - Pro B ALL
    - Pre B ALL; Trans pre B ALL; Mature B ALL
  - Immature T ALL; Common T ALL; Mature T ALL

5.1.2 Leucemia mieloide acuta

Circa $1/5$ dei casi di leucemia acuta dell'età infantile è rappresentato da LMA. L'incidenza annuale riportata nei Paesi Industrializzati è di circa 1 caso ogni 100.000 soggetti di età inferiore a 15 anni. In contrasto con la leucemia di tipo LLA, non si rileva nessun picco di incidenza fino all'età di 10 anni, mentre si rileva un aumento progressivo durante l'adolescenza. La diagnosi di LMA richiede la presenza di almeno un 30% di cellule blastiche nell'aspirato midollare, quindi i sinotmi più frequenti sono: febbre, pallore, diatesi emorragica. I sintomi più frequenti sono quelli legati all'infiltrazione blastica midollare. Febbre, pallore, diatesi emorragica sono frequentemente riscontrati. Anche per quanto riguarda la LMA esistono sottogruppi ben individuabili sulla base di criteri morfologici, genetici e di espressione genica. Il gruppo cooperativo FAB ha classificato la LMA in 8 sottogruppi:

- M1 $(11 - 19\%)$
- M2 $(25 - 30\%)$
- M3 $(3 - 12\%)$
- M4 $(15 - 23\%)$
- M4Eo$(2 - 6\%)$
- M5 $(13 - 29\%)$
- M6 $(1 - 5\%)$
- M7 $(4 - 14\%)$
- Sarcoma granulocitico $(0 - 1\%)$
- M0 $(1 - 6\%)$

5.2 I dati

Il dataset preso in esame contiene i profili genetici di 22 soggetti affetti da leucemia cronica. Nel dataset sono presenti le misure di espressione genica per 4992 geni, misurate con la tecnologia *cDNA microarrays* e successivamente normalizzati. Le misure relative a ciascun gene sono fornite come logaritmo in base 2 delle misure di fluorescenza, e i soggetti sono così suddivisi:

- 10 soggetti affetti da leucemia linfoblastica acuta(LLA B);
- 5 soggetti affetti da leucemia linfoblastica acuta T(LLA T);
- 3 soggetti affetti da leucemia linfoblastica acuta traslocata (ALL-Tr);
- 4 soggetti affetti da leucemia mieloide acuta (LMA).

5.2.1 Statistiche a confronto su dati reali

Le varie tipologie di leucemia possono essere divise in 2 gruppi, uno relativo ai soggetti con leucemia LMA ed uno relativo ai soggetti aventi leucemia LLA, che comprende tutte le 3 leucemie (LLA-B, LLA-T, LLA-Tr). Il sistema di ipotesi relativo ad ogni gene è il seguente:

$$\begin{aligned} &\text{H0:} \quad \mu_{LLA} = \mu_{LMA} \\ &\text{H1:} \quad \mu_{LLA} \neq \mu_{LMA} \end{aligned}$$

dove μ_{LLa} indica la media di espressione del primo campione (leucemia di tipo LLA), mentre μ_{LMA} indica la media di espressione del secondo campione (leucemia di tipo LMA). I primi due vettori dei valori sono riportati in Tabella 5.1 e nelle Tabelle 5.2 e 5.3 vengono indicati i valori delle due statistiche per i vettori considerati.

Tabella 5.1: Espressione dei primi due geni sui pazienti

	Vettori dati reali			
Vet	Array1	Array2	Array3	Array4
1	1.614100	-0.1307659	1.2417820	1.1184640
2	-0.178517	-0.6490136	0.2571654	0.4190591
Vet	Array5	Array6	Array7	Array8
1	0.4471580	0.9487120	0.1625535	1.774503
2	0.3062375	-0.2217738	-0.1733950	0.426461
Vet	Array9	Array10	Array11	Array12
1	1.4619440	1.401684	1.2850120	1.236589
2	-0.2466636	0.000000	0.5086526	0.488475
Vet	Array13	Array14	Array15	Array 16
1	1.4010780	1.3564520	1.8727680	2.32395200
2	0.3570121	-0.6327633	-0.2801606	0.07814755
Vet	Array17	Array18	Array19	Array 20
1	0.9380847	1.4463510	1.0649850	1.80499100
2	0.0000000	-0.1603578	0.1319201	-0.06880237
Vet	Array21	Array22		
1	0.59697340	0.9493493		
2	-0.06507724	-0.8516950		

Tabella 5.2: Valori osservati e livello di significatività stimato per i dati in Tabella 5.1 per la statistica Sam

	Valori Sam 3 riguardanti pazienti leucemici	
	vett1	vett2
t-moderata	0.3157865	0.9507895
p-value	0.7029708	0.2599780

Tabella 5.3: Valori osservati e livello di significatività stimato per i dati in Tabella 5.1 per la statistica t

	Valori t per i dati riguardanti pazienti leucemici	
	vett1	vett2
t-moderata	0.3636673	1.1232297
p-value	0.71950167	0.27318381

È utile effettuare lo stesso procedimento di confronto fatto per i dati simulati. A tal fine, la Figura 5.1 riporta la rappresentazione dell'ordinamento, e delle corrispondenze tra le due statistiche, per 10 geni.

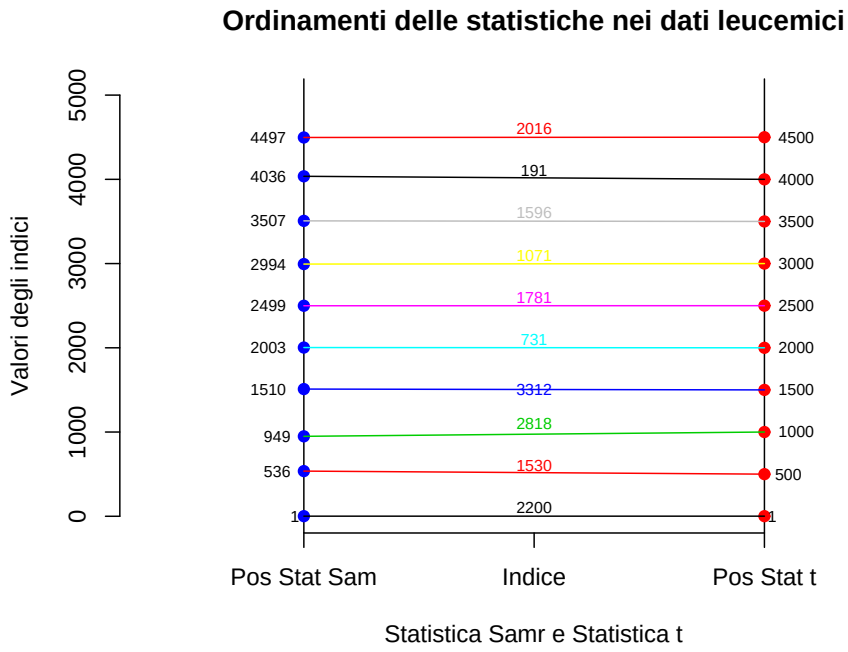

Figura 5.1: Confronto tra ordinamenti delle due statistiche per i pazienti leucemici.

La Figura 5.1 conferma il comportamento simile delle due statistiche 5.1; dobbiamo tener presente, comunque, che i dati presi in considerazione sono stati normalizzati prima di essere passati alle statistiche, operazione che consente di ricondursi ad una situazione simile a quella prodotta con i dati simulati. Il grafico 5.2 riporta il confronto per 10 geni dichiarati differenzialmente espressi. Per poter estrarre i geni differenzialmente espressi dall'ordinamento delle due statistiche, si sono presi i primi 250 geni e gli ultimi 249, poichè sia la statistica Sam sia la statistica t utilizzano un test a due code. Successivamente sono stati scelti 10 geni definiti differenzialmente espressi per Sam, per i quali sono state calcolate le posizioni per la statistica t ed infine rappresentati.

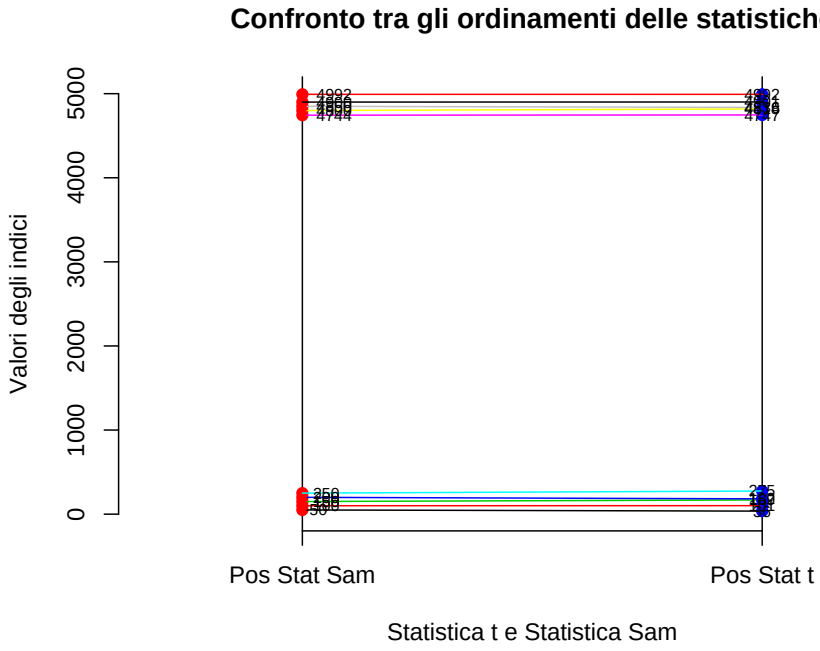

Figura 5.2: Ordinamento dei geni differenzialmente espressi per i dati reali

La Figura 5.2 indica che l'ordinamento dei geni differenzialmente espressi è molto simile. Entrambe le statistiche collocano i geni differenzialmente espressi nelle

code dell'ordinamento, ciò è confermato anche dal grafico della densità dei geni per la statistica Sam (Figura 5.3): infatti i geni differenzialmente espressi presentano due punti di massimo che si trovano sulle code della distribuzione dei geni non espressi. Anche in questo caso, quindi, i geni presentano ranghi simili, come per i dati simulati.

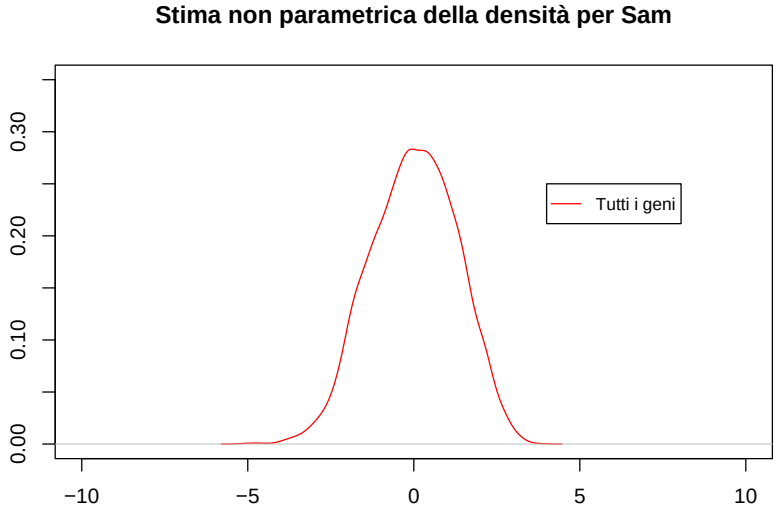

Figura 5.3: Stima non parametrica della densità della statistica Sam per i dati relativi a pazienti leucemici

La Figura 5.3 presenta le stesse caratteristiche della Figura 3.3 nella pagina 36; la medesima cosa avviene per la stima non parametrica della densità della statistica per i geni differenzialmente espressi (Figura 5.6), dove si nota un andamento bimodale ben definito nel quale i due punti di massimo si trovano nelle code della densità di tutti i geni. La stima della densità della statistica t (Figura 5.4) sembra avere un andamento pressoché simile alla stima non parametrica della densità effettuata su tutti i geni, però per i geni differenzilamente espressi (Figura 5.6), non presenta le stesse caratteristiche della statistica Sam. Infatti i valori per i ge-

ni differenzialmente espressi sembrano distribuirsi lungo tutto l'asse e presentano un punto di massimo che viene quasi a coincidere con la moda dei valori della statistica t per i geni non espressi. In conclusione, per quanto riguarda i geni differenzialmente espressi, Sam sembrerebbe posizionare tali geni nelle code della densità stimata su tutti i geni, mentre t sembra avere un andamento meno chiaro.

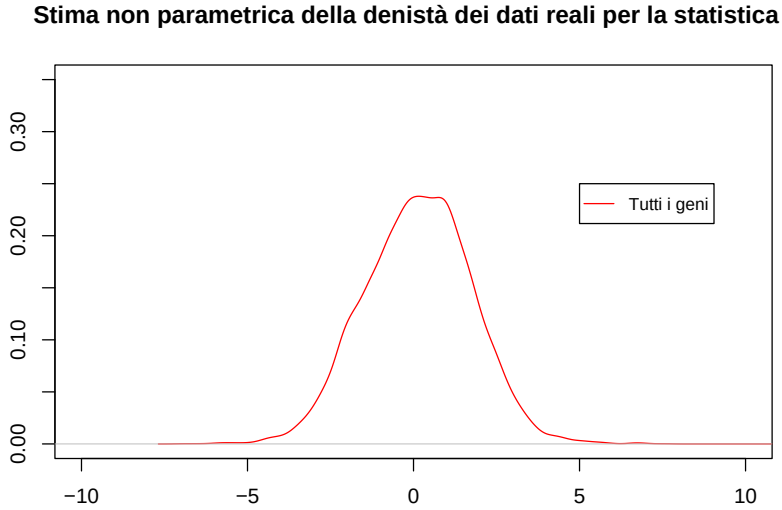

Figura 5.4: Stima non parametrica della densità della statistica t per i dati relativi a pazienti affetti da leucemia

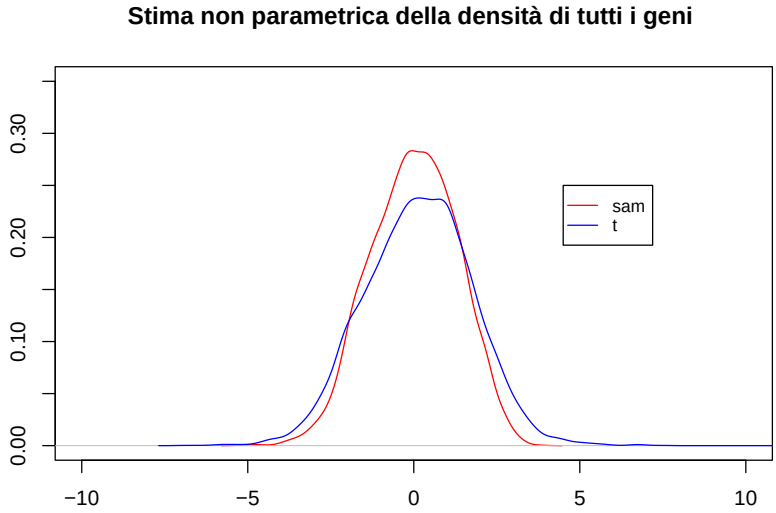

Figura 5.5: Stima non parametrica della densità per la statistica Sam e t per tutti i geni dei dati relativi ai pazienti affetti da leucemia

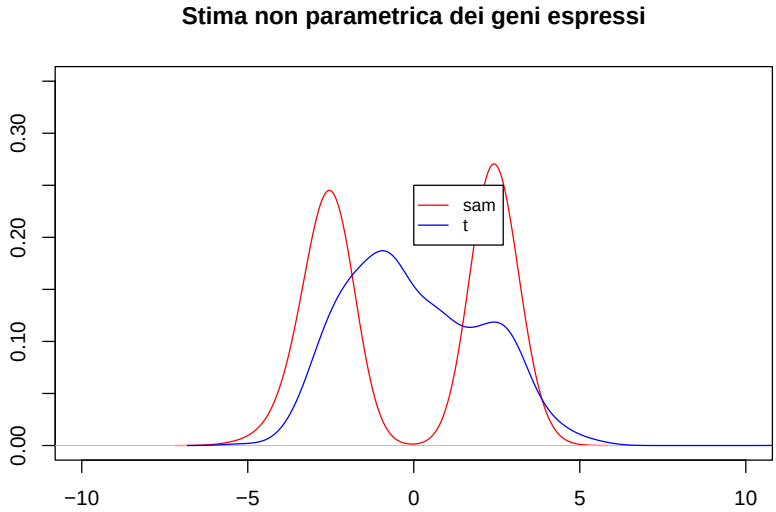

Figura 5.6: Stima non parametrica della densità per la statistica Sam e t per i geni differenzialmente espressi nei dati riguardanti pazienti affetti da leucemia

Gli indici di correlazione tra i valori generati dalla due statistiche, per i dati reali, confermano le osservazioni fatte per i dati simulati: esiste una forte correlazione tra la statistica Sam e t.

Tabella 5.4: Indici di correlazione tra i valori della statistica Sam e t per i dati reali

	Correlazione
p-value tt e t-moderata	0.9994106
statistica tt e t-moderata	0.9996757

Di seguito riportiamo i diagrammi di dispersione dei valori associati ai geni dei dati relativi ai soggetti affetti da leucemici (Tabella 5.4), dove i livelli di significatività della statistica t sembrano essere molto simili a quelli della statistica Sam.

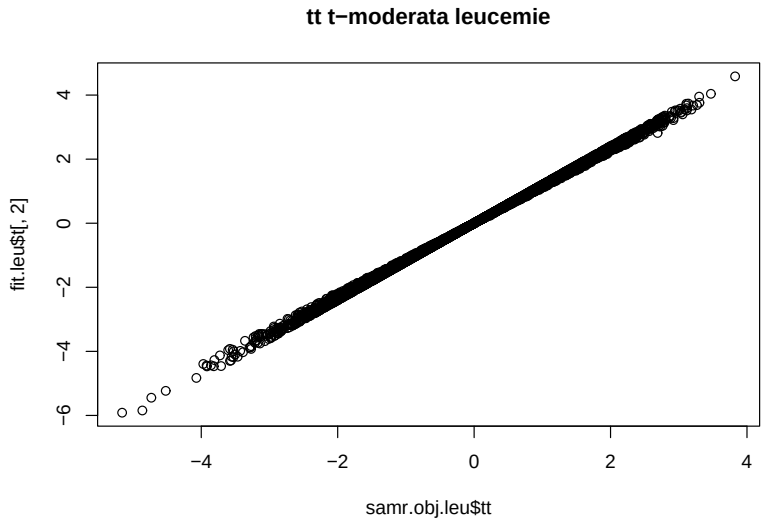

Figura 5.7: Diagramma di dispersione tra i valori delle due statistiche per i dati reali

Per studiare più in dettaglio le differenze tra gli ordinamenti delle due statistiche, proseguiamo verificando di quante posizioni differiscono le posizioni dei geni nell'ordinamento delle due statistiche. Riportiamo in Tabella 5.5 la distribuzione delle differenze, in valore assoluto, delle posizioni dei geni differenzialmente espressi.

Tabella 5.5: Frequenze assolute delle differenze tra posizioni dei geni differenzialmente espressi

posizioni	frequenze	posizioni	frequenze
0	22	26	5
1	26	27	11
2	31	28	6
3	21	29	3
4	20	30	6
5	25	31	6
6	22	32	4
7	23	33	2
8	19	34	5
9	17	35	1
10	16	36	1
11	16	37	3
12	15	38	4
13	13	39	3
14	13	40	2
15	16	41	1
16	14	43	1
17	18	44	1
18	10	45	2
19	12	46	1
20	10	47	3
21	14	48	3
22	9	50	1
23	8	51	1
24	4	57	1
25	7	62	1

Dalla Tabella 5.5 e dalla Figura 5.8, emerge che le frequenze assolute più alte cadono nelle prime posizioni: questo significa che le differenze tra le posizioni, prese in valore assoluto, sono molto basse. Le frequenze assolute indicano che l'intervallo maggiore è compreso tra $[0 - 20]$, la differenza massima 62 appare molto piccola, dato che il range varia da $[0; 4992]$. Visto che le differenze tra gli ordinamenti dei geni differenzialmente espressi sono molto piccole, diventa di particolare importanza la scelta della soglia per la dichiarazione dei geni differenzialmente espressi. Nel caso trattato si era fissata una proporzione di geni differenzialmente espressi pari a $p = 0.1$, quindi pari a 499 geni. Di conseguenza, le due statistiche, Sam e t, forniscono un ordinamento similare e quindi un riconoscimento dei geni differenzialmente espressi rassomigliante. È interessante

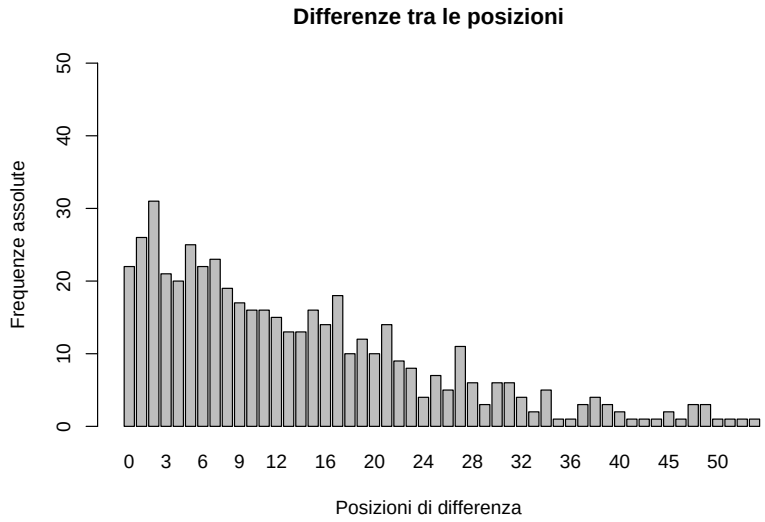

Figura 5.8: Differenze tra le posizioni dei geni differenzialmente espressi

confrontare gli ordinamenti ottenuti dalla statistica t, sui dati reali, con la statistica B che introduce la quota a posteriori. Per prima cosa riportiamo in Figura 5.9 il confronto tra gli ordinamenti della statistica Sam e la statistica B per i pazienti

affetti da leucemia.

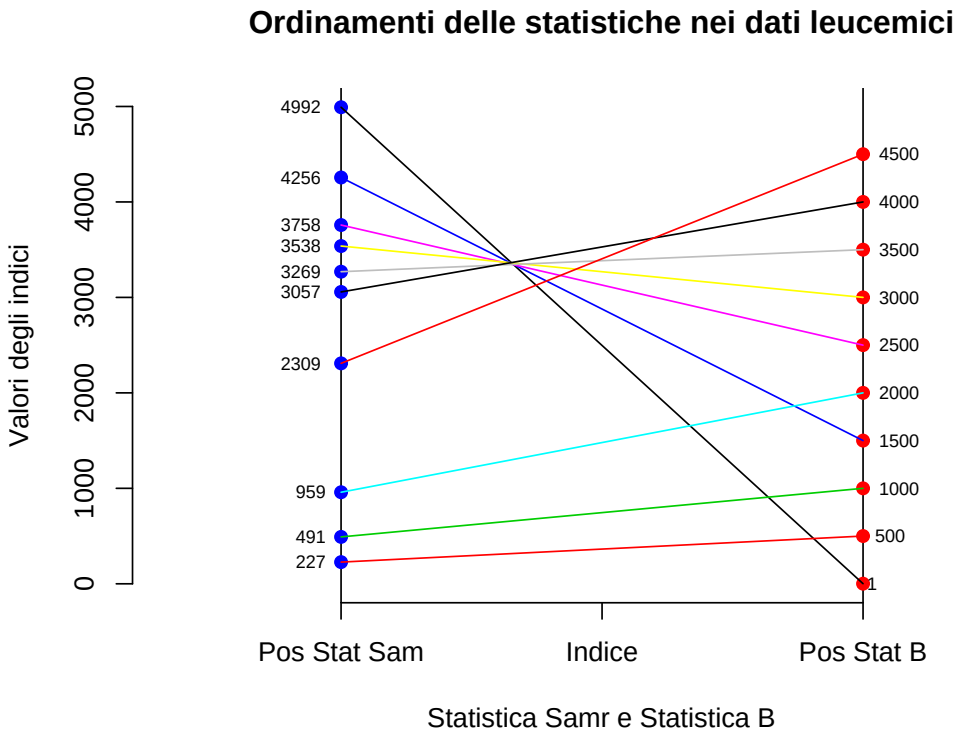

Figura 5.9: Confronto ordinamenti Sam e B per i dati reali

A differenza della Figura 5.1, emerge che l'ordinamento tra le due statistiche, Sam e B, non è speculare come per la statistica Sam e t, infatti in alcuni casi la posizione dell'indice è opposta, come era emerso per i dati simulati (Figura 4.4). La differenza tra l'ordinamento dei geni diventa ancora più chiara nel confronto degli ordinamenti tra i geni differenzialmente espressi, Figura 5.10, dove a indici sia alti che bassi della statistica Sam la statistica B attribuisce, ai geni differenzialmente espressi, indici piccoli.

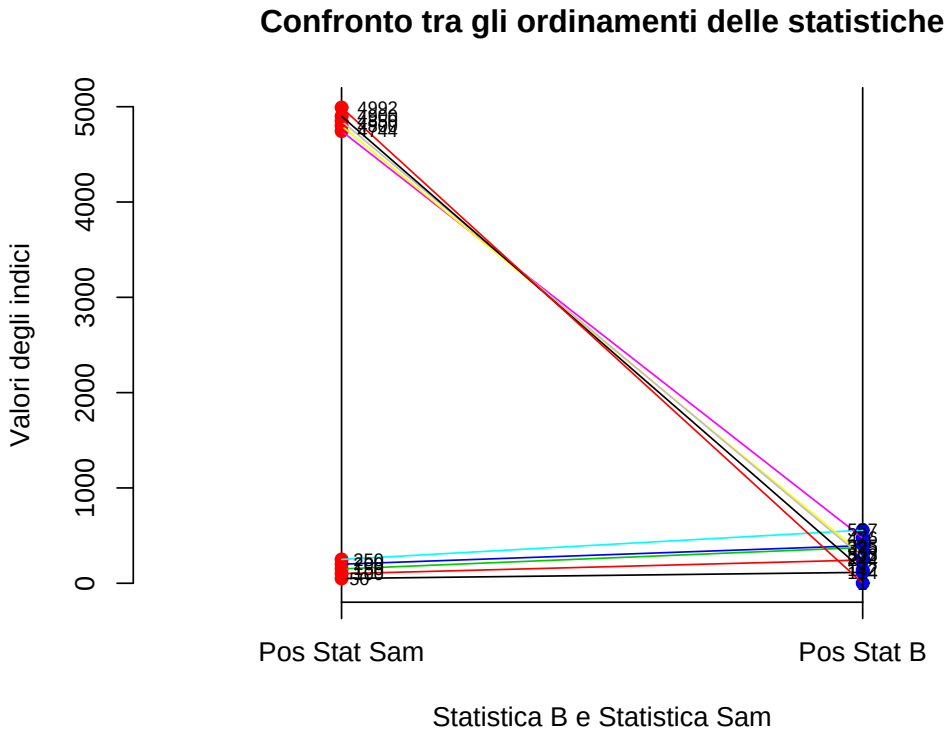

Figura 5.10: Confronto ordinamento geni differenzialmente tra Sam e B

Per un'ulteriore verifica riportiamo la stima non parametrica della densità della statistica B per i dati reali. In Figura 5.11 e Figura 5.12 vengono riportate le distribuzioni, rispettivamente di tutti i geni e dei geni differenzialmente espressi per la statistica B, dove, a conferma delle considerazioni fatte per la Figura 5.10, si evince che la distribuzione dei geni differenzialmente espressi per i dati riguardanti pazienti affetti da leucemia si localizza all'inizio del nostro grafico, con una sola moda. I valori associati ai geni non espressi dalla statistica B, rispecchiano l'andamento della Figura 4.3, Figura 4.4 per i dati simulati.

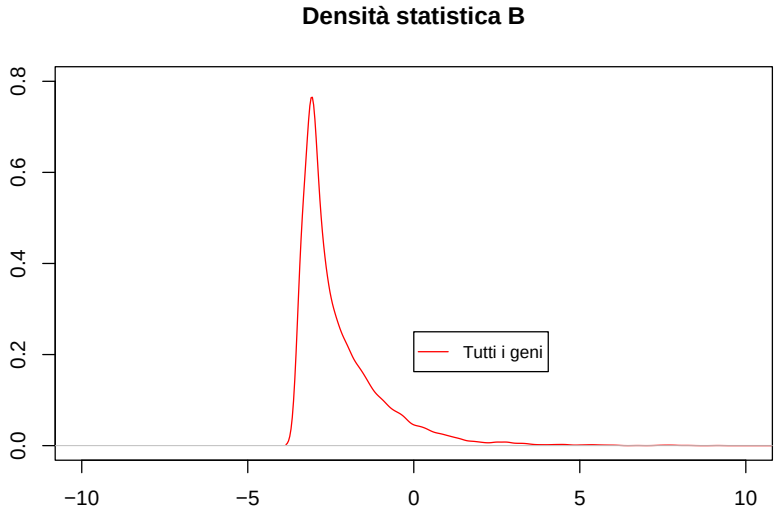

Figura 5.11: Stima non parametrica della densità per la statistica B sui dati riguardanti pazienti leucemici

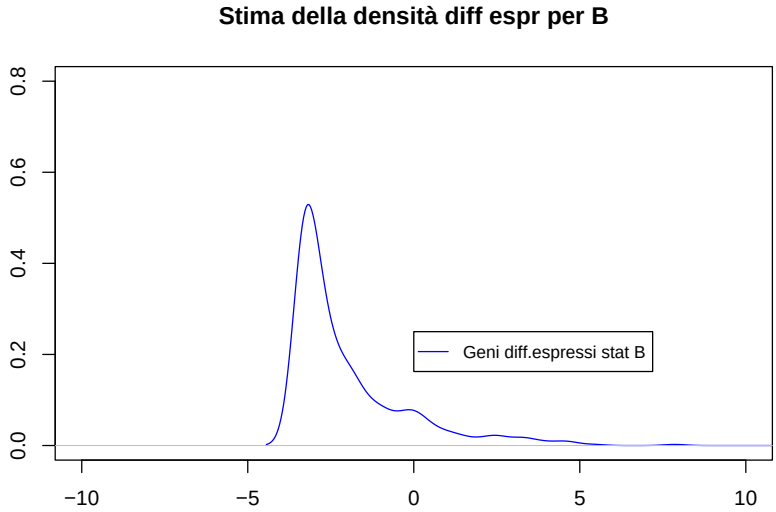

Figura 5.12: Stima non parametrica della densità per i geni differenzialmente espressi per la satistica B

L'indice di correlazione tra i valori della statistica B e Sam conferma il debole legame tra i valori delle statistiche, Figura 5.13.

Tabella 5.6: Valori di correlazione tra la statistica Sam e B per i dati riguardanti pazienti affetti da leucemia

	Correlazione
statistica tt e B	-0.09438399

La Figura 5.13, evidenzia lo stesso comportamento osservato nei dati simulati (Figura 4.5 nella pagina 49), quindi una situazione speculare attorno allo zero.

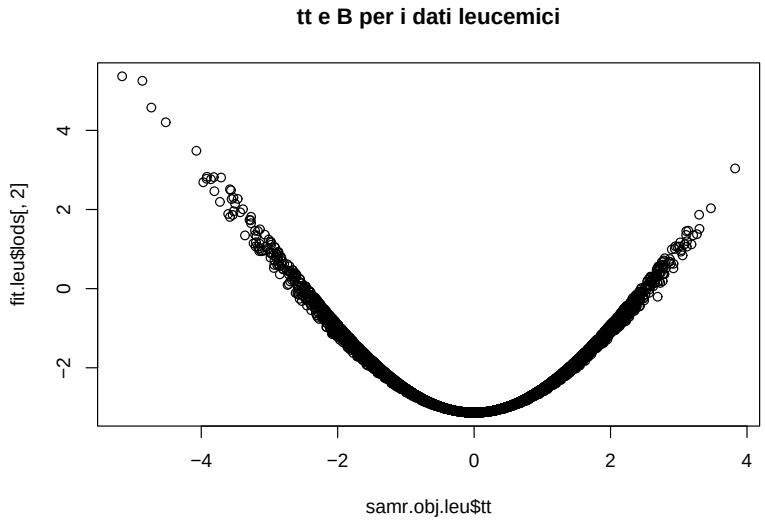

Figura 5.13: Correlazione tra la statistica Sam e B per i dati riguardanti pazienti affetti da leucemia

Capitolo 6

Conclusioni

Le analisi effettuate, hanno lo scopo di confrontare il lavoro tra due statistiche che adottano due test leggermente differenti. Sam utilizza una statistica t-modificata che prevede l'uso, a denominatore, di una piccola costante s_0, il *fudge factor* per evitare che geni con bassi livelli di espressione dominino il risultato dell'analisi. La statistica t non utilizza fattori di correzione a denominatore: lavora con le probabilità a posteriori, e le usuali varianze della statistica t vengono sostituite dalle varianze a posteriori.

Le due statistiche prese in esame presentano dei risultati nell'ordinamento e nella classificazione molto simili. Sia nel data set dei 5000 geni simulati, attraverso il modello log Normale Normale, sia nei dati riguardanti i pazienti affetti da leucemia, i risultati dell'ordinamento tra le due statistiche sembrano variare di poco.

Il confronto tra l'ordinamento dei geni differenzialmente espressi dalla statistica Sam e la statistica t (Figura 3.2 nella pagina 35) sembrano somigliarsi. Nelle simulazioni grazie alla conoscenza della diversa o uguale espressione dei geni alla tabella dei livelli di espressione, è stato possibile calcolare quanti di questi fossero classificati correttamente dalle due statistiche. Nelle medesime condizioni

di applicazione, per 100 replicazioni, si è potuto notare che entrambe le statistiche riconoscono un numero esiguo di geni differenzialmente espressi. Sono stati calcolati i valori medi della matrice di confusione per 100 replicazioni. Sam, mediamente, classifica con correttezza 42 geni differenzialmente espressi mentre t 47, su un totale di 500 geni (Tabelle 3.4,3.5). Sono state calcolate anche le matrici di confusione dei livelli medi di espressione, che confermano quanto emerso dalle matrici di confusione precedenti. Sam riconosce 32 geni differenzialmente espressi, mentre t 50; per quanto concerne i geni non espressi, sia Sam che la statistica t presentano valori abbastanza alti nella corretta classificazione(Tabelle 3.6, 3.7). Dalle matrici di confusione, il numero di geni classificati correttamente da Sam e t, risulta essere basso rispetto alle nostre attese, calcolate nell'ordine dei 500 geni differenzialmente espressi. Ovviamente, più grande è la proporzione di geni dichiarati differenzialmente espressi, più grande sarà il numero di geni classificati correttamente da entrambe le statistiche. Le densità stimate per i dati simulati sembrano comportarsi in maniera diversa (Figure 3.53.6). Per la statistica Sam, hanno un andamento ideale dove la stima non parametrica della densità dei geni differenzialmente espressi presenta due mode che si posizionano alla fine e all'inizio del massimo della densità stimata per tutti i geni. La distribuzione non parametrica per i geni data dalla statistica t ha andamento meno chiaro, dato che non sembra esservi una divisione netta tra geni differenzialmente espressi e non. Questo potrebbe indicare che la statistica Sam riconosce con maggiore sensibilità i geni differenzialmente espressi da quelli che non lo sono. Gli indici di correlazione tra i valori delle due statistiche, calcolati per tre simulazioni differenti, fanno emergere una correlazione positiva che coincide con l'indice di correlazione dei livelli di espressione, anch'essi positivi. I diagrammi di dispersione delle tre simulazioni effettuate (Figure 3.7, 3.9, 3.11) e quelli relativi ai livelli di espressione (Figure 3.8,3.10, 3.12), mostrano come i livelli di significatività os-

servati dalla statistica t sembrano sempre essere maggiori o uguali a quelli della statistica Sam. I dati relativi a pazienti leucemici hanno confermato le similarità emerse precedentemente. In questo data set si è scelto di suddividere i pazienti in due tipologie di leucemie: LLA, leucemia linfoblastica, e LMA, leucemia mieloide. Dato che entrambe le statistiche lavorano con test a due code è stato possibile confrontare i geni differenzialmente espressi. Infatti, sia t che Sam posizionano quest'ultimi all'inizio e alla fine dell'ordinamento. Sia la corrispondenza tra le posizioni assegnate al medesimo gene, che i valori di correlazione indicano un comportamento somigliante, sebbene la statistica Sam utilizzi una statistica t-modificata, che prevede l'uso a denominatore di una piccola costante s_0, il *fudge factor*, l'ordinamento dei geni differenzialmente espressi è simile a quello ottenuto con la statistica t, che lavora con probabilità a posteriori. Per poter confrontare l'ordinamento delle due statistiche sui geni differenzialmente espressi sono stati presi i primi 250 geni e gli ultimi 249, e successivamente confrontati. Le due statistiche sembrano presentare un ordinamento avente ranghi simili, notiamo infatti una corrispondenza molto alta sia tra i geni espressi (Grafico 5.2 nella pagina 57) che non espressi (Grafico 5.1 nella pagina 56). L'analisi della stima non parametrica delle distribuzioni coincide con quanto detto per i dati simulati. L'indice di correlazione tra i valori della statistica Sam e i valori della statistica t indica una correlazione positiva, confermata anche dall' indice di correlazione dei livelli medi di espressione anch'esso molto alto. Il diagramma di dispersione, Figura 5.7 tra i valori delle due statistiche applicate ai dati reali, conferma i valori degli indici di correlazione positivi. L'ipotesi di similarità è plausibile, anche a causa delle differenze calcolate nel posizionamento del medesimo gene, da parte di entrambe le statistiche. Le differenze calcolate sul posizionamento, riportate in Tabella 5.5 indicano come massima differenza 62, e la massima concentrazione di valori si posiziona nell'intervallo $0 - 20$, che conferma le ipotesi di similarità fatte prece-

dentemente. Calcolando che il range delle differenze può variare da $0 - 4992$ e la nostra massima differenza risulta essere 62, riscontriamo una somiglianza tra l'ordinamento dei geni da parte delle due statistiche. È importante tener presente che i dati reali presi in considerazione sono stati normalizzati prima di essere passati alle statistiche, operazione che consente di ricondursi ad una situazione simile a quella prodotta con i dati simulati. Le due analisi, una condotta sui dati simulati, e l'altra riguardante i dati su pazienti leucemici, portano allo stesso risultato: la classificazione e l'ordinamento delle due statistiche è pressochè simile. Data la similarità nell'ordinamento delle due statistiche, diventa di particolare importanza la scelta della soglia, al di sopra e al di sotto della quale si definisce un gene differenzialmente espresso poichè dalle analisi effettuate è sufficiente la differenza di qualche posizione per includere o meno un gene nel gruppo dei differenzialmente espressi.

Appendice A

Modello Log Normale Normale

In questa appendice guarderemo più da vicino il modello da noi scelto per la simulazione dei nostri geni. Nel modello LogNormale-Normale si ipotizza che la distribuzione, relativa alla trasformata logaritmica della singola misurazione, sia normale. Si indica con $\widetilde{z}_{gi} = \log z_{gi}$ il logaritmo naturale della misura di espressione z_{gi}. La variabile $\widetilde{Z}_{gij} \| \mu_g$ si distribuisce come una $N(\mu_g; \sigma^2)$, con una varianza σ^2 comune per tutti i geni e con una media μ_g dipendente dal singolo gene. La distribuzione a priori di μ_g è una $N(\mu_0; \tau_0^2)$. In definitiva sono coinvolti 3 parametri $\Theta = (\mu; \sigma^2; \tau_0)$ e sfruttando le ipotesi precedentemente fatte si ricava la densità marginale delle osservazioni. Dato $\widetilde{Z}_{gi} = \mu_g + \epsilon_i$ con $\mu_g \sim N(\mu_0; \tau_0^2)$ e $\epsilon_i \sim N(0; \sigma_2)$ si ricavano le seguenti quantità:

$$E(\widetilde{Z}_{gi}) = E(\mu_g) + E(\epsilon_i) = \mu_0$$

$$Var(\widetilde{Z}_{gi}) = Var(\mu_g) + Var(\epsilon_i) = \tau_0^2 + \sigma^2$$

$$\text{Cov}(\widetilde{Z}_{gi}; \widetilde{Z}_{gj}) = E(\widetilde{Z}_{gi} \cdot \widetilde{Z}_{gj}) - E(\widetilde{Z}_{gi}) \cdot E(\widetilde{Z}_{gj})$$

$$= \left[E(\mu_g^2) + E(\mu_g \cdot \epsilon_j) + E(\mu_g \cdot \epsilon_i) + E(\epsilon_i \cdot \epsilon_j) \right] - \mu_0^2$$

$$= E(\mu_g^2) - \mu_0^2 = Var(\mu_g) = \tau_0^2$$

con $i \neq j$. In conclusione la variabile n-dimensionale $\widetilde{Z}_g$ è una normale multipla con la seguente struttura:

$$(\widetilde{Z}_g)_{n*1} = \begin{bmatrix} \widetilde{z}_{g1} \\ \widetilde{z}_{g2} \\ \dots \\ \widetilde{z}_{gn} \end{bmatrix} \sim N \left[\begin{bmatrix} \mu_0 \\ \mu_0 \\ \ddots \\ \mu_0 \end{bmatrix}, \begin{bmatrix} \sigma^2 + \tau_0^2 & \tau_0^2 & \dots & \tau_0^2 \\ \tau_0^2 & \sigma^2 + \tau_0^2 & \dots & \tau_0^2 \\ \vdots & \vdots & \ddots & \\ \tau_0^2 & \dots & \sigma^2 + \tau_0^2 & \end{bmatrix} \right]$$

In forma compatta $(\widetilde{Z}_g)_{n*1} \sim N(\_mu_0, \sum_n)$ con $\_mu_0 = (\mu_0, \mu_0, \dots, \mu_0)^T$ e $\sum_n = \sigma^2 I_n + \tau_0^2 M_n$ con I_n matrice d'identità $n * n$ e M_n matrice $n\ astn$ di tutti gli 1. Con questi risultati si possono calcolare le probabilità a posteriori di qualsiasi pattern di espressione secondo la formula e le misure di odds relative.

Appendice B

Codice R

```
set.seed(123456)
sim1=function(ngeni,nrep,mu0,sigma,tau,p,j)
{ncond=2
matrice=matrix(rep(NA,ngeni*ncond*nrep),ncol=ncond*nrep,byrow=T)
DE=rep(FALSE,ngeni)
for(i in 1:ngeni)
{if(runif(1)>p){ mu.g=rnorm(1,mu0,tau)
 matrice[i,]=exp(rnorm(nrep*ncond,mu.g,sigma))}
 else{
  mu.g1=rnorm(1,mu0,tau)
  mu.g2=rnorm(1,mu0,tau)
  cond1=exp(rnorm(nrep,mu.g1,sigma))
  cond2=exp(rnorm(nrep,mu.g2,sigma))
    matrice[i,]=c(cond1,cond2)
 DE[i]=TRUE}}
write.table(matrice,paste(1,''matricelnn.txt'',sep='' ''),
row.names=FALSE,col.names=FALSE, sep ''\\t'')
write.table(DE,paste(1,''DElnn.txt'',sep='' ''),
row.names=FALSE,col.names=FALSE)}
sim1(ngeni=5000,nrep=15,mu0=6.7,sigma=sqrt(0.80),tau=1.3,p=0.1,j)
```

```
matric<-read.table(paste(1,''matricelnn.txt'',sep=''''),h=F)
de<-read.table(paste(1,''DElnn.txt'',sep=''''),h=F)
R.lnn<-matrix(data=0,nrow=5000,ncol=15)
G.lnn<-matrix(data=0,nrow=5000,ncol=15)
matricelnn.globale<-cbind(dati.lnn@maRf,dati.lnn@maGf)
Statistica SAM
y<-c(rep(2,15),rep(1,15))
data.lnn<-list(x=matricelnn.globale,y=y,geneid=as.character
(1:nrow(matricelnn.globale)),
genenames=paste(''g'',as.character(1:nrow(matricelnn.globale))
,sep=''''),logged2=TRUE)
samr.obj.lnn<-samr(data.lnn,resp.type=
''Two class unpaired'',nperms=100)
nrighe<-0
l<-0.1
delta.table.lnn<-samr.compute.delta.table(samr.obj.lnn,dels=l)
siggenes.table.lnn<-samr.compute.siggenes.table
(samr.obj.lnn,del=l,data.lnn,delta.table.lnn)
ord.samr<-order(samr.obj.lnn\$tt,decreasing=T)
Statistica B
design <- cbind(Grp1=1,Grp2vs1=c(rep(1,15),rep(0,15)))
fitma<-lmFit(matricelnn.globale,design)
p2<-0.1
fit<-eBayes(fitma,proportion=p2)
ord<-order(fit\$t[,2],decreasing=T)
qualism<-ord[1:(p2*5000)]
totsm<-c(qualism[order(qualism)])
top.all.B<-topTable(fit,n=nrow(fitma),
coef=2,adjust=''fdr'',sort.by=''B'')
gene.eB<-top.all\$ID[1:25]
top.all.T<-topTable(fit,n=nrow(fitma),coef=''Grp2vs1'',
adjust=''fdr'')
top.all.M<-topTable(fit,n=nrow(fitma),
```

```
coef=NULL, adjust=''fdr'', sort.by=''t'', resort.by=''M'')
qqt(fit\$t,de=fit\$df.prior+fit\$df.residual,
main=''Moderated t'',xlim=range(-10,10))
abline(0,1)
Corrispondenza tra gli ordinamenti delle due statistiche
pos<-NULL
for(i in 1:5000)
{ valore<-ord[i]
j<-1
    while ((valore!=ord.samr[j])\&(j<=5000))
     j<-j+1
    if (valore==ord.samr[j])
        pos<-c(pos,j)}
vet<-rep(1000,5000)
vet1<-rep(4000,5000)
par(mfrow=c(1,1))
veta<-c(1,500,1000,1500,2000,2500,3000,3500,4000,4500)
ca<-pos[veta]
vatb<-ord.samr[ca]
vetc<-ord[veta]
plot(1:5000, 1:5000,axes=F, type = ''n'',
xlab=''Statistica Samr e Statistica t'', ylab=
''Valori degli indici'',
main=''Confronto tra gli ordinamenti delle statistiche'')
abline(v=4000)
abline(v=1000)
for (i in 1:10)
{(points(vet1[i],veta[i],col=''red'',pch=19))
(points(vet[i],ca[i], col = ''blue'',pch=19))
segments(vet[i],ca[i],vet1[i],veta[i],col=i)}
text(vet1[1:10],veta,label=(veta),cex=0.7,adj=-0.4)
text(vet[1:10],ca,label=(ca),cex=0.7,adj=1.5)
axis(1,c(1000,2500,4000),c(''Pos Stat Sam'',
```

```
''Indice'',''Pos Stat t''))
axis(2)
text(2500,100,labels=ord[1],cex=.7,col=1)
text(2500,600,labels=ord[2],cex=.7,col=2)
text(2500,1100,labels=ord[3],cex=.7,col=3)
text(2500,1500,labels=ord[4],cex=.7,col=4)
text(2500,2100,labels=ord[5],cex=.7,col=5)
text(2500,2600,labels=ord[6],cex=.7,col=6)
text(2500,3100,labels=ord[7],cex=.7,col=7)
text(2500,3600,labels=ord[8],cex=.7,col=8)
text(2500,4100,labels=ord[9],cex=.7,col=9)
text(2500,4600,labels=ord[10],cex=.7,col=2)
differenze<-NULL
lung<-5000
for (i in 1:5000)
{differenze[i]<-pos[i]-i}
diffesp<-NULL
ve<-NULL
for(i in 1:5000)
{ ve<-c(ve,de[i,1])
if (ve[i]==TRUE)
 diffesp<-c(diffesp,i)}
Calcolo dei geni differenzialmente espressi
difford<-NULL
diffsamr<-NULL
posizione<-NULL
for(i in 1:length(diffesp))
{for (j in 1:5000){
 if (diffesp[i]==ord[j])
   difford<-c(difford,j)
 if (diffesp[i]==ord.samr[j])
     diffsamr<-c(diffsamr,j)}}
```

Appendice C

Codice R per le leucemie

```
4992 geni 22 colonne 18 ALL e 4 AML
data.leu<-list(x=data,y=y,geneid=as.character(1:nrow(data)),
genenames=paste(''g'',as.character(1:nrow(data)),
sep='''''),logged2=TRUE)
samr.obj.leu<-samr(data.leu,resp.type=''Two class unpaired''
,nperms=100)
data.leu
samr.obj.leu
nrighe<-0
l<-0.1
delta.table.leu<-samr.compute.delta.table(samr.obj.leu,dels=l)
siggenes.table.leu<-samr.compute.siggenes.table
(samr.obj.leu,del=l,data.leu,delta.table.leu)
ord.sam.leu<-order(samr.obj.leu\$tt,decreasing=T)
design <- cbind(Grp1=1,Grp2vs1=c(rep(1,18),rep(0,4)))
fitma.leu<-lmFit(data,design)
p2<-0.1
fit.leu<-eBayes(fitma.leu,proportion=p2)
ord.leu<-order(fit.leu\$t[,2],decreasing=T)
qualism.leu<-ord.leu[1:(p2*4992)]
```

```
totsm.leu<-c(qualism.leu[order(qualism.leu)])
pos.leu<-NULL
posizioni1<-NULL
for(i in 1:4992)
{ valore<-ord.leu[i]
j<-1
   while ((valore!=ord.sam.leu[j])&(j<=4992))
    j<-j+1
   if (valore==ord.sam.leu[j])
       pos.leu<-c(pos.leu,j)}
posizioni1<-c(pos.leu[1:250],pos.leu[4743:4992])
a<-pos.leu[ord.leu]
diff.sam.leu<-NULL
diff.t.leu<-NULL
for(i in 1:4992)
{       if(i<=250|i>=4744){
                diff.sam.leu<-c(diff.sam.leu,ord.sam.leu[i])
                diff.t.leu<-c(diff.t.leu,ord.leu[i])}}
posizioni<-NULL
for(i in 1:499){
        for(j in 1:4992){
                if (diff.t.leu[i]==ord.sam.leu[j])
                posizioni<-c(posizioni,j)
                        }
                }
differenza1<-NULL
differenza2<-NULL
k<-4744
for(i in 1:length(posizioni))
{       if(i<=250)
        { differenza1<-c(differenza1,(i-posizioni[i]))}
        else {differenza2<-c(differenza2,(k-posizioni[i]))
        k<-(k+1)}       }
```

```
dif1<-abs(differenza1)
dif2<-abs(differenza2)
dif3<-c(dif1,dif2)
valori<-sort(dif3)
cont<-NULL
val<-NULL
for (i in 1:length(valori)){
        if(i==1) {
                val[i]<-valori[i]
                cont[i]<-length(which(valori==valori[i]))
                                }else{
                        for (j in 1:length(val)){
                                        if (valori[i] != val[j])
                                        p<-0
        else{
                                p<-1
                                break    }        }
                        if (p==0){
                                x<-length(val)
                                val[x+1]<-valori[i]
                                cont[x+1]<-
                                length(which(valori==valori[i]))}}}
d<-list(z=cbind(val,cont))
postscript(file=''posdif.ps'',
paper=''special'',horizontal=F,width=7,height=5)
barplot(d\$z[,2],names.arg=d\$z[,1],xlim=c(0,60)
,ylim=c(0,40),width=0.4,space=2.0,
xlab=''Posizioni di differenza'',
ylab=''Frequenze assolute'',
main=''Differenze tra le posizioni'')
dev.off()
```

Bibliografia

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple hypothesis testing under dependency. *The Annals of Statistics*, Vol. 29(No. 4), 2001.

Y. Benjamini, Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, serie B(57), 1995.

F. Costa A. Orecchia R. Cavalli, F. Cognetti. *Fondamenti di Oncologia Clinica*, volume Primo. Elsevier, Milano, 2006.

R. Storey J.D. Thusher V.G. Efron, B. Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, (23), 2001.

W. McClure Ernst. *Statitics for Microarray*, volume Primo. Jhon Wiley, States, 2004.

W. Huber and A.V. Heydebreck. Variance stabilization applied to microarray data calibration and to the quantification of diferential expression. *Bioinformatics*, 1(1), 2002.

J. et al. Hug. Multiple locus variable number tandem repeat analysis reveals genetic relationship within bacillus anthracis. *Journal of Clinial Microbiology*, 37(8), 1999.

R. Ihaka and Gentleman. R: A language for data analysis and garphics. *Journal of Computation and Graphical Statistics*, (5), 1996.

J. et alt Lips, E. Dierssen. Reliable hig throughput genotyping and loss of-heterozygosity detection in formalin-fixed, paraffin embedded tumors using single nucleotide polymorphism arrays. *Cancer res.*, 65(22), 2005.

Mauro Lise. Metodi bayesiani empirici parametrici. Tesi di laurea in statistica e informatica, Università degli Studi di Bologna, 2005.

I. Lönnsted and Speed. Replicated microarray data. *Statistica sinica*, 12(1), 2002.

T. Lönnstedt, I. Britton. Two hierarchical bayes models for cdna microarray gene expression. *Bioinformatics*, 1(1), 2005.

Y. Goden G. et al. Loftus, S.K. Chen. Informatic selection of a neural crest melanocyte cdna set for microarry analysis. Articol, The National Academy of Scienze of the United States of America, 1999.

R. Ranganathan S. Nagaraj, S. Gasser. A hitchhiker's guide to expressed sequence tag (est) analysis. *Bioinformatics*, 8(1), 2007.

M.A. Newton and C.S. Blattner F.R. Tsui K.W. others Kendziorski, C.M. Richmond. On differential variability of expression ratios:improving statistical inference about gene expression changes from microarray dataa. *Journal of Computational Biology*, 8, 2001.

S Parmigiani, R.A. Garrett, E.S. Irizarry, and S.L. Zeger. *The Analysis of Gene Expression Data: Methods and Software*. Springer, New York, 2003.

R. et al. Ross, D. Tibshirani. Clustering methods for the analysis of dna microarray data. Technical report, Department of Health Research and Policy Statistics, Genetics and Biochemistry, Standford University, 1995.

G.K. Smyth. Linear models and empirical bayes methods for assessing diferential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

JD. Storey. The positive false discovery rate: A bayesian interpretation and the q-value. *Journal of the Royal Statistical Society*, (25), 2001a.

JD. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society*, (B64), 2002.

R. Storey, JD. Tibshirani. Estimating false discobvery rates under dependence with application to dna microarry. *Journal of the Royal Statistical Society*, (56), 2001b.

J. et al. Tian, Y. Matthew. Structure based design of robust glucose biosensior using a thermotoga maritma perplasmic glucose binding protein. *A publication of the Protein Society*, 16(10), 2007.

R. Tusher, V.G. Tibshirani and G. Chu. *Significance Analysis of microarrays applied to the ionizing radiation response*. Proceedings of the National Academy of Sciences USA, USA, 2001.

M.J. Dudoit S. Yang, Y.H. Buckley and T.P. Speed. Comparison of methods for image analysis on cdna microarray data. *Journal of Computational Biology*, (11):108 – 136, 2002.

Y.H. Yang, S. Dudoit, P. Luu, V. Lin, D.M. Peng, and J. Ngai. *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation*, volume 40 Volume of LMS Lecture Notes -Monograph Series. Science and Statistics: A Festschrift for Terry Speed, USA, 2002-2003.