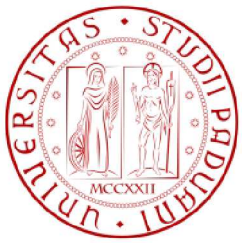
UNIVERSITÀ DEGLI STUDI DI PADOVA



Facoltà di Scienze Statistiche

Corso di Laurea in Scienze Statistiche, Demografiche e Sociali

Tesi di Laurea

Normalizzazione per dati di microarray a singolo canale

Uno studio comparativo

Relatore:

Dott. Chiara Romualdi

Laureando:

Federico Rotolo

Anno Accademico 2008/2009

Indice generale

0. Introduzione

0.1. Espressione genica e microarray

I geni, segmenti di DNA, portano alla sintesi di proteine attraverso due fasi distinte: la trascrizione e la traduzione. Durante la prima fase, il frammento di DNA viene copiato in un filamento chiamato RNA messaggero (mRNA), mentre durante la seconda fase il filamento di mRNA viene tradotto in proteine, sequenze di aminoacidi. Lo studio dell'espressione genica consiste nella misurazione della quantità di mRNA trascritto per ciascun gene presente nel DNA, al fine di trovare gruppi di geni più o meno espressi nell'organismo. A seconda della quantità di copie trascritte (mRNA), i caratteri o le funzioni biologiche regolate dal segmento di codice genetico in questione saranno più o meno espressi.

Dal momento che l'*acido ribonucleico* (RNA) è un composto piuttosto instabile, mentre l'*acido deossiribonucleico* (DNA) è molto stabile, di fatto si lavora con una copia inversa dell'mRNA, il *DNA complementare* (cDNA).

L'interesse principale degli studi di espressione genica è l'analisi comparativa del livello di espressione tra due situazioni differenti: tra due tessuti diversi, tra due condizioni biologiche, tra pazienti sani e pazienti malati e così via.

Gli studi di espressione genica sono largamente usati e affrontano problematiche diverse. Sono utilizzati, per esempio, per comprendere meglio alcuni meccanismi biologici latenti, per individuare sottogruppi di malattie, per esaminare la risposta ai farmaci, per classificare i pazienti in gruppi diagnostici e prognostici.

In questi casi l'obiettivo è l'individuazione dei geni che sono *differenzialmente espressi* (DE) nei due gruppi in esame, detti anche *geni attivi*, e la loro separazione da quelli *egualmente espressi* (EE) o *inattivi*.

Tradizionalmente, gli studi di biologia molecolare e di genetica classica permettevano lo studio intensivo di uno o pochi geni alla volta. Grazie al finanziamento e al successo dei progetti di ricerca genomica e al conseguente aumento del numero di geni identificati, si è reso necessario lo sviluppo di tecnologie che permettessero un'analisi di geni su vasta scala. A questo proposito è stata messa a punto una nuova tecnica, quella dei *microarray*.

I microarray sono vetrini da microscopio su cui vengono depositate numerose molecole di DNA (spot) a singola elica, disposte a griglia, che consentono di quantificare simultaneamente l'attività di decine di migliaia di geni.

I chip sfruttano una proprietà importante del DNA, ossia l'appaiamento tra basi complementari: la *Timina* con *l'Adenina* e la *Guanina* con la *Citosina*.

Quando i geni sono DE, nelle cellule dei due campioni in esame si avrà un numero di molecole di mRNA molto diverso. Per misurare questa quantità si estrae l'mRNA dai due tipi di tessuti (o gruppi di trattamento), lo si converte in *DNA complementare* (cDNA), perché più stabile, e vi si lega un marcatore fluorescente. Si ibrida poi questa soluzione (cDNA con marcatore) sul chip. Durante l'ibridazione i cDNA marcati riconoscono i loro rispettivi segmenti di DNA complementare in modo proporzionale alla loro concentrazione nel tessuto studiato. I chip vengono, quindi, sottoposti a lettura attraverso un laser (scansione). Il laser eccita i fluorofori del marcatore che di conseguenza emette luce in modo proporzionale alla quantità di fluoroforo attaccato, quindi in modo proporzionale alla quantità di cDNA attaccato. Per ogni spot si avrà un'emissione di luce. L'immagine risultante viene poi analizzata da appositi software che per ogni spot quantificano l'intensità del segnale.

0.1.1. Microarray a singolo e a doppio canale

I tipi di tecnologie disponibili per microarray sono molteplici, ma possono essere ricondotte a due grandi famiglie: a *doppio canale* o a *singolo canale*.

Nella tecnologia a doppio canale ogni chip contiene materiale biologico proveniente da due campioni di due gruppi diversi, marcati con fluorofori diversi (normalmente Cy3 per il colore verde e Cy5 per il rosso), in modo da interferire il meno possibile tra loro e in modo da poter operare un confronto al netto di tanti fattori specifici del singolo esperimento.

Ibridando il campione misto sullo stesso vetrino, la misura ottimale che si è soliti prendere in considerazione per ogni spot è il rapporto tra le intensità dei due canali. Questo sistema è piuttosto buono per il confronto semplice tra due gruppi, ma rende meno affidabile la misura di valori assoluti di intensità e molto difficile il confronto tra più di due gruppi.

Purtroppo però la presenza contemporanea di due fluorofori può spesso portare ad avere forti errori sistematici per singolo esperimento, motivo per cui negli ultimi anni i chip a singolo canale hanno visto un forte incremento e sviluppo.

Gli array a singolo canale, di cui *Affymetrix* rappresenta il sistema principale, sono impiegati, con diversi tipi di disegno di studio, per numerosi esperimenti: comparativi semplici, più complessi per confronti multipli, temporali, per la ricerca di sottogruppi e per test diagnostici.

Nella tecnologia a canale singolo ogni esperimento è effettuato su un singolo campione biologico; per questo è possibile che ci sia un effetto specifico del singolo esperimento, difficilmente controllabile nel confronto tra i gruppi. D'altra parte in questo modo si rende molto più semplice il confronto tra più di due gruppi, gli esperimenti sono maggiormente indipendenti tra loro e si ottengono valori assoluti di espressione, più affidabili per il confronto tra studi diversi.

Quindi se da un lato la tecnologia a singolo canale sembra essere maggiormente riproducibile, dall'altro lo stesso studio, se effettuato con tecnologia a doppio canale, richiede la metà degli esperimenti, il che rende il singolo canale più costoso.

I metodi utilizzati e sviluppati nel presente lavoro fanno riferimento a microarray a canale singolo, che permettono quindi la stima del livello assoluto di espressione genica.

In questo caso il confronto tra due diverse situazioni, come il quello tra *casi* e *controlli,* può essere effettuato soltanto con due esperimenti effettuati separatamente sui due campioni.

0.2. I dati trattati

Il grado di espressione dei geni, come illustrato, è misurato attraverso la luminescenza degli spot ottenuti da un campione biologico il cui cDNA è stato ibridato con marcatori fluorescenti. Pertanto il dato su cui si lavora è una misura di luminescenza, per ciascun gene (valore compreso tra 0 e 2^{16}).

La matrice dei dati di conseguenza è costituita da un numero di righe ng pari al numero di geni presi in considerazione e da un numero di colonne pari al numero di ripetizioni dell'esperimento.

Nel caso più comune gli studi sono orientati ad un confronto tra due condizioni biologiche di interesse che denoteremo come *casi* e *controlli*. Il numero di colonne è quindi dato dalla somma di repliche biologiche su campioni di *casi* e su campioni di *controllo.*

A differenza della tecnologia a doppio canale, le prove su *casi* e *controlli* sono totalmente indipendenti; è possibile quindi che il disegno sperimentale sia sbilanciato, con un numero di esperimenti sui *casi* diverso da quello sui *controlli*. Per ragioni di semplicità considereremo, almeno in una prima fase, situazioni bilanciate, con pari numero di ripetizioni per i due gruppi; in questo modo il numero di colonne è naturalmente $2 \cdot ne$, con *ne* che indica il numero di esperimenti per ciascun gruppo.

Tabella 0.I: Esempio di tabella dei dati di espressione

	Casi			*Controlli*		
gene 1	$x'_{1,1}$	$\cdots$	$x'_{1,ne}$	$x''_{1,1}$	$\cdots$	$x''_{1,ne}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$
gene k	$x'_{k,1}$	$\cdots$	$x'_{k,ne}$	$x''_{k,1}$	$\cdots$	$x''_{k,ne}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$
gene ng	$x'_{ng,1}$	$\cdots$	$x'_{ng,ne}$	$x''_{ng,1}$	$\cdots$	$x''_{ng,ne}$

0.3. Fonti di errore e normalizzazione

Le analisi di dati di microarray poggiano sull'ipotesi che le intensità di fluorescenza misurate siano rappresentative dell'effettivo livello di espressione.

La complessità dei protocolli sperimentali dei microarray rende questa tecnologia molto variabile e soggetta a distorsioni sistematiche a volte significative. Per questo, prima di poter confrontare in maniera appropriata i livelli di espressione, sono necessarie alcune manipolazioni e trasformazioni al fine di attenuare valori affetti da aberrazioni casuali o variazioni sistematiche così da riportare tutti i dati su livelli comparabili.

Le normalizzazioni possibili sono diverse e muovono da presupposti differenti. Nel presente lavoro si propone una nuova normalizzazione per dati di array a singolo canale, valutandone la bontà attraverso uno studio comparativo con tre metodi di normalizzazione tra quelli più comunemente utilizzati: *Variance Stabilizing Normalization (Vsn)*, *Quantile Normalization (Quantile)* e Ciclic *Loess Normalization (CLoess)*.

Queste normalizzazioni sono molto usate perché funzionano piuttosto bene nella maggior parte degli studi, ma poggiano su assunzioni circa le caratteristiche dei dati che non sempre possono essere soddisfatte: *Vsn* e *CLoess* richiedono che la maggior parte dei geni sia EE e *Cloess* assume anche che, ad ogni livello di intensità, i geni sovraespressi e quelli sottoespressi siano grossomodo bilanciati. Invece *Quantile* ipotizza che tra le ripetizioni non ci siano livelli medi di intensità molto diversi.

Proprio a causa di questi vincoli *Xiong et al. (2008)* hanno proposto un metodo libero da assunzioni, basato sull'algoritmo *Generalizazed Procrustes Normalization (GPA)*.

In questo studio si propone, in tre versioni leggermente diverse, una implementazione ciclica del metodo *GPA*.

La valutazione della bontà di questi diversi metodi avverrà grazie al test *Significance Analysis of Microarray* (SAM) sulla differenza di espressione media tra i due gruppi. Elaborato appositamente per lo studio di espressione di genica, come suggerito dal nome, esso è costituito da una versione permutazionale di un test *t di Student* per la differenza tra medie, opportunamente corretto con una moderazione che attenua l'effetto dell'eteroschedasticità dei dati. Anche questo test verrà illustrato nel dettaglio nel *Paragrafo 1.3.1*.

Una prima valutazione dei diversi metodi di normalizzazione farà uso di dati simulati attraverso il modello proposto da *Balagurunathan* (2002), opportunamente adattato alla situazione di interesse. Grazie all'utilizzo di dati simulati è possibile conoscere quali dati sono stati generati come geni DE e quali no; in questo modo si può calcolare l'effettiva sensibilità e specificità del *test SAM* dopo le diverse normalizzazioni e confrontarle per mezzo di curve *Receiver Operating Characteristic* (curve ROC).

Una successiva fase di valutazione della bontà dei modelli avverrà operando su dati reali. I dati utilizzati sono estrapolati da un progetto molto più ampio, finanziato recentemente dalla FDA (Food Drug Administraction) con lo scopo di valutare la riproducibilità degli esperimenti di microarray. Il disegno sperimentale prevede l'utilizzo di 4 campioni: A, mRNA derivante da pool di diversi tessuti sani umani, B, mRNA di cervello umano, C mistura di 75% A e 25% di B e D mistura di 25% di A e 75% di B. In questo modo la bontà delle normalizzazioni viene valutata nella misura di geni differenzialmente espressi che mostrano in media andamenti crescenti (A>C>D>B) o decrescenti (B>D>C>A).

0.4. Scopo e struttura del lavoro

Nel corso del presente lavoro sarà presentato n confronto tra alcuni comuni metodi di normalizzazione di dati di microarray, proponendo tre versioni cicliche del metodo *GPA*. I vantaggi teorici di questo metodo sono l'assenza delle assunzioni sulle quali si basano gli altri metodi, come la *GPA* proposta da *Xiong et al.* (2008) e la robustezza offerta dalla implementazione ciclica, presa a prestito dal metodo *Loess Ciclica*, suggerito da *Bolstad et al.* (2003).

Nel *Capitolo 1* saranno presentati nel dettaglio gli strumenti utilizzati per la simulazione dei dati, per la normalizzazione e per la valutazione della bontà dei diversi metodi.

Nel *Capitolo 2* saranno mostrati i risultati delle diverse fasi di studio: la simulazione dei dati, le normalizzazioni dei dati simulati e il loro confronto, le normalizzazioni dei dati reali e il loro confronto.

Nel *Capitolo 3* verranno discussi i risultati e illustrate le conclusioni a cui si è giunti.

1. Materiali e metodi

1.1. I dati simulati

Il presente studio prende in considerazione due diversi scenari. Il primo scenario prevede la normalizzazione di una matrice di espressione genica i cui esperimenti derivano tutti da uno stesso laboratorio (quindi tendenzialmente caratterizzati da distorsioni sperimentali simili). Il secondo scenario, invece, prevede la normalizzazione di un'unica matrice costituita dall'unione di due matrici di espressione generate rispettivamente in due laboratori differenti (quindi con un evidente fattore di laboratorio che li differenzia). Questo secondo scenario è molto comune negli studi di biologia dei sistemi con approccio computazionale.

1.1.1. Perché usare dati simulati

La prima parte del lavoro si concentra sul confronto dei metodi di normalizzazione utilizzando dati simulati. Le matrici di espressione sono state generate utilizzando un algoritmo in grado di modificare il dato di intensità dei microarray in modo da renderlo più o meno distorto nelle sue diverse fasi sperimentali.

Il valore grezzo di intensità è modificato da diversi fattori che aggiungono, ciascuno, variabilità e distorsione al segnale effettivo. I fattori che possono intervenire sono moltissimi e sono legati a tutte le fasi sperimentali dell'esperimento: dalla fabbricazione del microarray alla preparazione del materiale biologico, dalla disposizione di questo sul vetrino alla lettura della luminosità da parte dello scanner ottico.

L'uso di dati simulati al posto di dati reali offre la possibilità di conoscere le diverse componenti aleatorie che hanno contribuito a generare ogni singola osservazione. Ciò permette di valutare le conclusioni a cui porta ogni modello di normalizzazione, alla luce delle caratteristiche effettive del processo generatore dei dati; nelle applicazioni reali questa informazione ovviamente non è nota, dato che è proprio l'obiettivo conoscitivo di tutto il processo.

Nel nostro caso l'informazione di interesse riguarda la conoscenza a priori dei DE e degli EE. In condizioni ideali i geni DE, a differenza degli EE, presenterebbero livelli di espressione molto diversi tra *casi* e *controlli*, con un rapporto tra i due molto lontano da 1. Sarebbe dunque sufficiente identificare quei geni caratterizzati da $M = \log(x'/x'') \gg 0$ o $M \ll 0$, tuttavia, vista la grande quantità di fattori che possono influire sul livello di luminescenza misurata, il valore del log-rapporto non-normalizzato può, spesso, essere fuorviante.

1.1.2. Il modello di simulazione

Il modello che si è scelto di utilizzare per la simulazione dei dati di espressione è una variante semplificata ed adattata alla tecnologia a singolo canale del modello proposto da *Balagurunathan et al.* (2002). Quest'ultimo prevede, oltre al segnale grezzo, sedici fonti di variabilità che imitano l'effetto di diverse caratteristiche degli spot e dell'array. Tra queste, per esempio, si trovano la dimensione di ciascuno spot, la sua asimmetria, la vicinanza con gli altri spot, la luminescenza di fondo (*background*) del microarray, i graffi del vetrino e così via.

Nel presente lavoro le fonti di variabilità tenute in considerazione sono state solo alcune di quelle proposte da *Balagurunathan* e sono indicate nella *Tabella 1.I*.

Il modello proposto dagli Autori è specifico per dati di microarray secondo la tecnologia a doppio canale. Nel caso qui esaminato, riguardante invece microarray a singolo canale, è stato necessario modificare il modello in maniera da renderlo coerente con la situazione considerata.

Tabella 1.I: Struttura e parametri del modello di simulazione per dati di microarray

Fonte di segnale	***Distribuzione e parametri***
Livello generale di espressione del gene k-esimo	$E_k \sim Esp(1/\beta)$ β *: vedi tabelle 1.II e 1.III*
Livello di espressione del gene k-esimo nello spot i-esimo	$E_{k,i} \sim N(E_k; \alpha \cdot E_k)$ α : *vedi tabelle 1.II e 1.III*
Eventuale sovra/sottoespressione del gene k-esimo	$E_{k,i} \sim [N(E_k, \alpha \cdot E_k)] \cdot [t_k]$ $t_k = 10^{\pm 2 \cdot b_k}$ $b_k \sim Beta(1.7, 4.8)$
Funzione di risposta espressione-intensità nello spot i-esimo	$I_{k,i} = a_{3,i} \cdot (a_{0,i} + E_{k,i} \cdot (1 - e^{-E_{k,i}/a_{1,i}})^{a_{2,i}})$ $a_{0,i} \sim N(0; 0.5)$ $a_{1,i} \sim N(m_1; s_1)$ $a_{2,i} \sim N(m_2; s_2)$ $a_{3,i} \sim U(1, 1.5)$ m_1, s_1, m_2, s_2 *: vedi tabelle 1.II e 1.III*
Ulteriore rumore specifico dello spot i-esimo del gene k-esimo	$N(\alpha_{1,k,i} \cdot I_{k,i};\ \alpha_{2,k,i} \cdot \lvert \alpha_{1,k,i} \cdot I_{k,i} \rvert)$ $\alpha_{1,k,i} \sim U(1, 1.4)$ $\alpha_{2,k,i} \sim U(1, 1.3)$

1.1.3. I parametri scelti per le simulazioni

Come detto precedentemente, il modello di simulazione sarà utilizzato in due diverse fasi dello studio: nella prima fase si farà una valutazione della bontà dei metodi di normalizzazione su un singolo insieme di dati, mentre nella seconda fase, si farà una analoga valutazione per due insiemi di dati provenienti da due laboratori diversi.

In *Tabella 1.II* sono mostrati i parametri scelti per la prima simulazione.

Tabella 1.II: Parametri per la simulazione di dataset da singolo laboratorio

Parametro	***Valore scelto***
α	0.1
β	6000
m_1	500
s_1	50
m_2	-1.5
s_2	0.5

Per simulare due insiemi di dati provenienti da due laboratori diversi, con caratteristiche non del tutto paragonabili, sono stati usati parametri come riportato in *Tabella 1.III*.

Tabella 1.III: Parametri per la simulazione di dataset da due laboratori

Parametro	***Valore lab1***	***Valore lab2***
α	0.05	0.15
β	3000	9000
m_1	500	50
s_1	50	10
m_2	-1.5	0.0
s_2	0.0	0.5

Al fine di rendere i dati il più possibile simili a quelli di un esperimento reale, le simulazioni sono state effettuate mantenendo una proporzione di geni DE, sul totale, intorno al 5%, con equiprobabilità tra sovra- e sottoespressi.

1.1.3.I parametri scelti per le simulazioni

Per avere risultati sufficientemente affidabili e robusti si è scelto di generare insiemi di dati di 10,000 geni con un numero di ripetizioni pari a *i) 15 casi e 15 controlli* per il laboratorio singolo*, ii) 15 casi* e 15 *controlli* per il laboratorio 1, *10 casi* e *5 controlli* per il laboratorio 2 nel caso dei due laboratori. Inoltre, tutte le valutazioni saranno fatte ripetendo le simulazioni per 10 volte.

Cinquecento geni DE distribuiti in modo equiprobabile tra sovra- e sottoespressi costituiscono uno scenario generalmente accettato negli studi di espressione ma che a volte può essere disatteso. Per questo motivo si è valutata la bontà delle normalizzazioni in presenza di: 1) molti geni DE (intorno al 70%) equamente distribuiti tra sovra e sottoespressi e 2) molti geni DE (70%) asimmetricamente distribuiti tra sovra e sottoespressi (21% sottoespressi e 49% sovraespressi). In questo modo vengono meno due delle assunzioni di molti dei più comuni metodi di normalizzazione.

Tabella 1.IV: Riepilogo, per le tre fasi di simulazione, delle probabilità per ogni gene di essere differenzialmente espresso [P(DE)] e, tra questi, di essere sovraespresso [P(Up|DE)]

Fase	***P(DE)***	***P(Up|DE)***
1	0.05	0.50
2	0.70	0.50
3	0.70	0.70

1.1.4. L'imputazione dei dati mancanti: il metodo KNN

Per motivi dovuti a problemi sperimentali legati al processo di preparazione del *microarray* e di lettura del valore di luminescenza del chip, è possibile che alcuni valori di intensità non siano rilevati o che presentino un valore inferiore rispetto a quello di *background* locale, dando origine quindi a valori negativi o mancanti.

Anche nel caso di dati simulati è possibile avere intensità nulle o negative, da considerarsi, quindi, come dati mancanti. Alcune tecniche semplici di imputazione dei dati mancanti prevedono l'assegnazione del valore 0 o della media di riga, cioè del livello medio di espressione del gene tra tutti i gruppi e tutti gli esperimenti, senza alcuna considerazione per la struttura di correlazione dei dati.

Nel 2001 *Troyanskaya et al.* hanno comparato diversi metodi di imputazione nell'ambito dei *microarray* ed hanno dimostrato come il metodo detto dei "k vicini più prossimi" (*K-nearest neighbours, KNN*) sia uniformemente migliore rispetto agli altri. L'algoritmo *KNN* imputa un valore tenendo conto del valore di espressione per il dato esperimento dei geni che sono mediamente più simili a quello in questione. Ovvero, per imputare la *i*-esima replica per il gene *j*-esimo, si considerano i *k* geni (con *k* a scelta) con profilo medio nei restanti $n-1$ esperimenti più simile a quello del gene *j*-esimo e con replica *i*-esima non mancante. La stima del valore da imputare al posto (j,i) è calcolata come la media del valore di espressione per la replica *i*-esima, pesata sulla similarità del profilo. Come indice di similarità tra i profili dei geni rispetto al gene *j*-esimo si usa la distanza euclidea.

Da notare che una imputazione di questo tipo ipotizza un meccanismo generatore dei dati mancanti non informativo; di fatto però il numero di dati mancanti imputati sarà di circa 1100-1300 valori per matrici di 10,000 geni e 30 esperimenti.

$$\frac{\#Valori_{MISSING}}{\#Valori_{TOT}} \approx \frac{1200}{(10000 \cdot 30)} = \frac{1200}{300000} = 0{,}004$$

Il peso dei dati mancanti sul totale sarà quindi nell'ordine del 4‰ ed eventuali arbitri nelle assunzioni del metodo di imputazione avranno un impatto piuttosto contenuto.

1.2. I metodi di normalizzazione

1.2.1. Un utile strumento : il grafico M-A

Dal momento che l'interesse principale degli studi di espressione genica, come detto in precedenza, è rivolto al confronto tra due condizioni biologiche (*casi* e *controlli*), non è importante studiare tanto le intensità assolute quanto valutare la differenza di intensità nei due campioni.

Quello che si è soliti fare a tale scopo è operare una trasformazione dei dati che mette in evidenza, al variare del livello medio dell'intensità di luminescenza, il differenziale di espressione tra *casi* e *controlli*, espresso come rapporto delle due misure. La trasformazione logaritmica è utilizzata per attenuare il peso di valori estremi e per smorzare la asimmetria intrinseca del fenomeno, che ha altissime concentrazioni su valori molto bassi e poche osservazioni su valori molto elevati.

Ogni coppia di valori $(x'_{k,a}; x''_{k,b})$ può essere trasformata in un'altra coppia $(A_{k,(a,b)}; M_{k,(a,b)})$, in cui A (*Amplitude*) è la media aritmetica delle log-intensità (e logaritmo della media geometrica delle intensità) e M (*Magnitude*) la differenza delle log-intensità (e logaritmo del rapporto delle intensità).

$$\begin{cases} A_{k,(a,b)} = \frac{1}{2}\cdot[\ln(x'_{k,a})]\cdot[\ln(x''_{k,b})] = \ln(\sqrt{x'_{k,b}\cdot x''_{k,b}}) \\ M_{k,(a,b)} = \ln(x'_{k,a}) - \ln(x''_{k,b}) = \ln\left(\frac{x'_{k,a}}{x''_{k,b}}\right) \end{cases} \quad (1.1)$$

Il grafico che vede sulle ascisse A e sulle ordinate M viene chiamato MA-plot ed è stato dimostrato essere la trasformazione migliore per identificare distorsioni sistematiche dipendenti dalla media delle intensità (*Yang et al. 2003*).

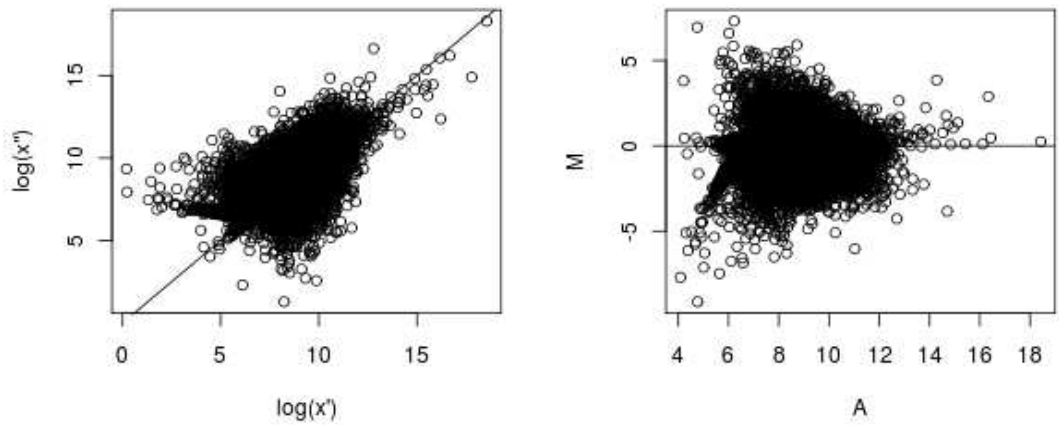


Grafico 1.I: Esempio di dati di intensità luminosa di microarray e trasformazione MA

In una situazione ideale i dati dovrebbero essere uniformemente distribuiti intorno all'asse orizzontale e dovrebbero presentare una dispersione grossomodo costante lungo i diversi livelli di intensità media A. I punti più distanti dall'asse delle ascisse rappresentano i geni con un rapporto tra le intensità nei due campioni molto diverso da 1, il che indica un gene attivo. Ancor prima, il grafico MA permette di rilevare eventuali andamenti sistematici nei dati, come asimmetrie rispetto all'asse $M=0$ o eteroschedasticità rispetto al livello medio A (come nella figura di esempio 1.I).

Nel caso di microarray a doppio canale i dati contengono già per loro struttura accoppiamenti tra valori di espressione di *casi* e *controlli*. Per questo motivo i grafici MA sono nati nell'ambito dello studio di distorsione dovuta ad un effetto di canale in presenza di microarray a doppio canale.

Nel caso del singolo canale la maggior parte dei disegni sperimentali rendono impossibile un'assegnazione uno-a-uno dei valori di intensità dei due tessuti. Quello che è possibile fare, in questa situazione, è generare i) tutte le possibili coppie tra casi e controlli, o ii) tutte le possibili coppie indipendentemente dal tipo di tessuto, a seconda delle necessità.

1.2.2. Le normalizzazioni di riferimento

I metodi proposti nel presente studio saranno esaminati in maniera comparativa con riferimento ai metodi di normalizzazione più diffusi e che al momento offrono le migliori prestazioni per dati di microarray a singolo canale.

I metodi presi in considerazione per la comparazione sono: *Loess Ciclica*, *Quantile* e *Vsn.*

1.2.2.i. Loess Ciclica

Questo approccio, applicato a tutte le coppie formate da osservazioni distinte, è stato proposto da *Bolstand et al.* nel 2003. Ogni coppia di intensità $(x'_{k,a}; x''_{k,b})$ viene trasformata in *MA*; sul grafico *MA* risultante, si calcola una curva di regressione locale *Loess* (Local Scatterplot Smoothing, *Cleveland and Devlin, 1988*). Le stime dei log-rapporti di intensità così ottenute sono $\hat{M}_{k,ab}$ e gli aggiustamenti $M'_{k,ab} = M_{k,ab} - \hat{M}_{k,ab}$, mentre le intensità stimate si ottengono attraverso la semplice inversa della trasformazione MA sulle coppie $(A_{k,ab};\ M'_{k,ab})$.

Dal momento che si lavora con più di due array, si ripete il procedimento su tutte le possibili coppie, effettuando tutte le normalizzazioni nel modo appena illustrato. Dopo aver normalizzato le coppie ed aver annotato l'aggiustamento per entrambi gli elementi, per ogni array *i*-esimo si avranno $(n_{array} - 1)$ aggiustamenti di cui si calcola una media semplice.

É dimostrato che dopo soltanto una o due iterazioni complete i cambiamenti ulteriormente apportati diventano molto piccoli. Per contro, dato che si lavora su tutte le possibili coppie di array, questo metodo è computazionalmente piuttosto oneroso.

L'idea di una normalizzazione effettuata ciclicamente su tutte le coppie di esperimenti sarà utilizzata nell'implementazione della *GPA ciclica* presentata nel *Paragrafo 1.2.3*.

1.2.2.ii. Normalizzazione Quantile

L'obiettivo della normalizzazione *Quantile* è di rendere uguali le distribuzioni empiriche di intensità di tutti gli array. Questo metodo è basato sulla considerazione che un grafico quantile-quantile è una linea perfettamente coincidente con la diagonale se e solo se la distribuzione dei due vettori di dati è la stessa. Questo concetto può essere esteso in n dimensioni in modo che n vettori di dati hanno la stessa distribuzione se e solo se il grafico quantile-quantile *n*-dimensionale è perfettamente allineato alla diagonale $(1/\sqrt{n},\ldots,1/\sqrt{n})$ dell'ipercubo di lato n. In questo modo è possibile far assumere la stessa distribuzione a un insieme di dati proiettando i punti del grafico quantile-quantile *n*-dimensionale sulla diagonale.

Sia $q_k=(q_{k1},\ldots,q_{kn})$, con $k=1..\mathrm{ng}$, il vettore del *k*-esimo quantile per tutti gli n array e sia $d=(1/\sqrt{n},\ldots,1/\sqrt{n})$ la diagonale unitaria. Per proiettare i quantili sulla diagonale si usa la proiezione di q su d

$$proj_d q_k=\left(\frac{1}{n}\sum_{j=1}^{n} q_{kj},\ldots,\frac{1}{n}\sum_{j=1}^{n} q_{kj}\right) \qquad (1.2)$$

Questo significa che si può dare a tutti gli array la stessa distribuzione sostituendo i valori del dataset originale con il quantile medio. Ciò giustifica il seguente algoritmo: data una matrice X di n array (colonne) e ng geni (righe)

1. si ordina ogni colonna di X, ottenendo X_{sort}
2. si assegna ad ogni elemento della *k*-esima riga di X_{sort} la media della riga stessa, ottenendo X'_{sort}
3. si calcola $X_{normalized}$, riordinando ogni colonna di X'_{sort} secondo l'ordine originale.

Un aspetto negativo di questo metodo è dato dal fatto che forza i valori dei quantili ad essere tutti uguali. Ciò potrebbe essere un problema nelle code, dove è possibile che un gene abbia lo stesso valore su tutti gli array, anche se di fatto si tratta di una situazione che si presenta molto raramente.

1.2.2.iii. Variance Stabilizing Normalization

Il metodo di trasformazione dei dati introdotto nel 2002 da *Durbin et al.* cerca di ovviare ad uno dei problemi principali accennati in precedenza: l'eteroschedasticità dei dati lungo il livello medio *A*. Questo metodo consente infatti di ottenere una distribuzione simmetrica, la cui varianza non dipende dal valore della media, ma è costante lungo tutto il range dei valori di espressione.

Secondo gli Autori i dati di espressione, una volta trasformati con il metodo *Variance Stabilizing Normalization (Vsn)*, sono conformi alle assunzioni dei più comuni metodi statistici (come l'ANOVA o la regressione) e possono perciò essere utilizzati senza ulteriori manipolazioni in un *test SAM* (v. *Paragrafo 1.3.1*).

La normalizzazione *Vsn* si basa sull'idea che un valore di intensità misurato per un singolo spot è il risultato di diverse componenti, legate dalla seguente relazione:

$$X_{k,i} = \alpha_{k,i} + \mu_{k,i} \cdot e^{\eta_{k,i}} + \epsilon_{k,i} \tag{1.3}$$

dove $X_{k,i}$ è il livello di luminescenza misurato per il gene *k*-esimo nell'*i*-esimo spot, $\alpha_{k,i}$ è il rumore medio di background, $\mu_{k,i}$ è il vero valore di espressione del gene e $\eta_{k,i}$ e $\epsilon_{k,i}$ sono termini di errore distribuiti secondo una distribuzione Normale: $\eta_{k,i} \sim N(0, \sigma_\eta^2)$ e $\epsilon_{k,i} \sim N(0, \sigma_\epsilon^2)$.

Per livelli bassi di espressione si ha che $\mu_{k,i} \to 0$ e che quindi il livello di luminescenza misurato è approssimabile con

$$X_{k,i} \approx \alpha_{k,i} + \epsilon_{k,i} \tag{1.4}$$

e ha una distribuzione di probabilità approssimata $X_{k,i} \sim N(\alpha_{k,i}, \sigma_\epsilon^2)$.

Analogamente per livelli molto alti di espressione si ha che $\mu_{k,i} \to +\infty$ e che quindi il primo e terzo termine della (1.3) sono trascurabili rispetto al secondo. Il livello di luminosità può quindi essere approssimato in questo caso con

$$X_{k,i} \approx \mu_{k,i} \cdot e^{\eta_{k,i}} \tag{1.5}$$

che ha distribuzione Log-Normale $X_{k,i} \sim \mathrm{LN}(\log \mu_{k,i}, \sigma_\eta^2)$.

In questo secondo caso, a differenza di prima, la varianza di $X_{k,i}$ è legata al livello medio di espressione in maniera crescente, poiché

$$Var(X_{k,i}) = \mu_{k,i}^2 \cdot S_\eta^2 \tag{1.6}$$

con $S_\eta = e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1)$.

Più in generale il livello di espressione $\mu_{k,i}$ ha un valore intermedio tra i due casi appena illustrati. La misura dell'espressione $X_{k,i}$ si distribuisce come una combinazione lineare di una Normale e di una Log-Normale. La varianza di $X_{k,i}$ si può quindi scrivere come

$$Var(X_{k,i}) = \mu_{k,i}^2 \cdot S_\eta^2 + \sigma_\epsilon^2 \tag{1.7}$$

In generale perciò la deviazione standard dipende in maniera lineare dal livello medio di espressione mentre, per le successive procedure statistiche, è necessario stabilizzare la varianza su tutto il range di luminosità misurata.

Si cerca quindi una funzione $f(\bullet)$ che renda costante la varianza asintotica $AV(\bullet)$ della trasformata.

$$AV(f(X_{k,i})) = f'(X_{k,i})^2 \cdot Var(X_{k,i}) = k^2 \tag{1.8}$$

dove $f'(\bullet)$ è la derivata della funzione ricercata e k^2 è invariante rispetto a tutte le componenti aleatorie e ai parametri del modello.

Per la (1.7) e la (1.8) si ha che $f'(X_{k,i})^2 = \frac{k^2}{Var(X_{k,i})} = \frac{k^2}{\mu_{k,i}^2 S_\eta^2 + \sigma_\epsilon^2}$, da cui

$$\int f'(X_{k,i})dy = \int \frac{k}{\sqrt{\mu_{k,i}^2 S_\eta^2 + \sigma_\epsilon^2}} dy \qquad (1.9)$$

Una soluzione dell'integrale (1.9) è

$$f(X_{k,i}) = \ln\left(X_{k,i} - \alpha_{k,i} + \sqrt{(X_{k,i} - \alpha_{k,i})^2 + c}\right) \qquad (1.10)$$

con $c = \sigma_\epsilon^2 / S_\eta^2$.

La trasformazione (1.10) prende il nome d *Generalized Logarithm Transformation (GLOG)*, introdotta per la prima volta nel contesto dei microarray da *Munson (2001)*. La funzione $f(\bullet)$ è monotona crescente per tutti i valori di $X_{k,i}$, positivi o negativi. Per valori di $\mu_{k,i}$ tendenti a zero la funzione è approssimativamente lineare, mentre per valori elevati di $\mu_{k,i}$ assume una forma logaritmica. La varianza asintotica dei dati trasformati è costante ed uguale a S_η^2 .

1.2.3. La normalizzazione basata sull'Analisi Procrastica Generalizzata

Tutti gli approcci classici per la normalizzazione dei dati di microarray hanno assunzioni sulla distribuzione dei dati che possono risultare critiche, dal momento che non sempre sono valide nella pratica. *Xiong et al.* (2008) hanno proposto un nuovo metodo per microarray a doppio canale, basato sull'algoritmo *Generalized Procrustes Normalization (GPA).* Questo sfrutta una procedura puramente geometrica di adattamento di una nuvola di punti rispetto ad un'altra di riferimento, con il vantaggio di non richiedere ipotesi di tipo statistico e probabilistico.

La *GPA* è un metodo basato sui minimi quadrati per la trasformazione di matrici di coordinate, orientato alla massimizzazione della similarità. Esso opera mediante tre operazioni successive:

1. *traslazione*: il baricentro di ogni matrice è spostato in maniera rigida fino all'origine del sistema di riferimento, sottraendo ad ogni punto le coordinate del baricentro del suo gruppo;

2. *rotazione*: tutti i punti vengono mossi di un angolo fisso per ciascuna matrice, tenendo fissa la distanza dal baricentro;

3. *cambio di scala*: ogni punto viene allontanato o avvicinato al centro secondo una proporzione fissa.

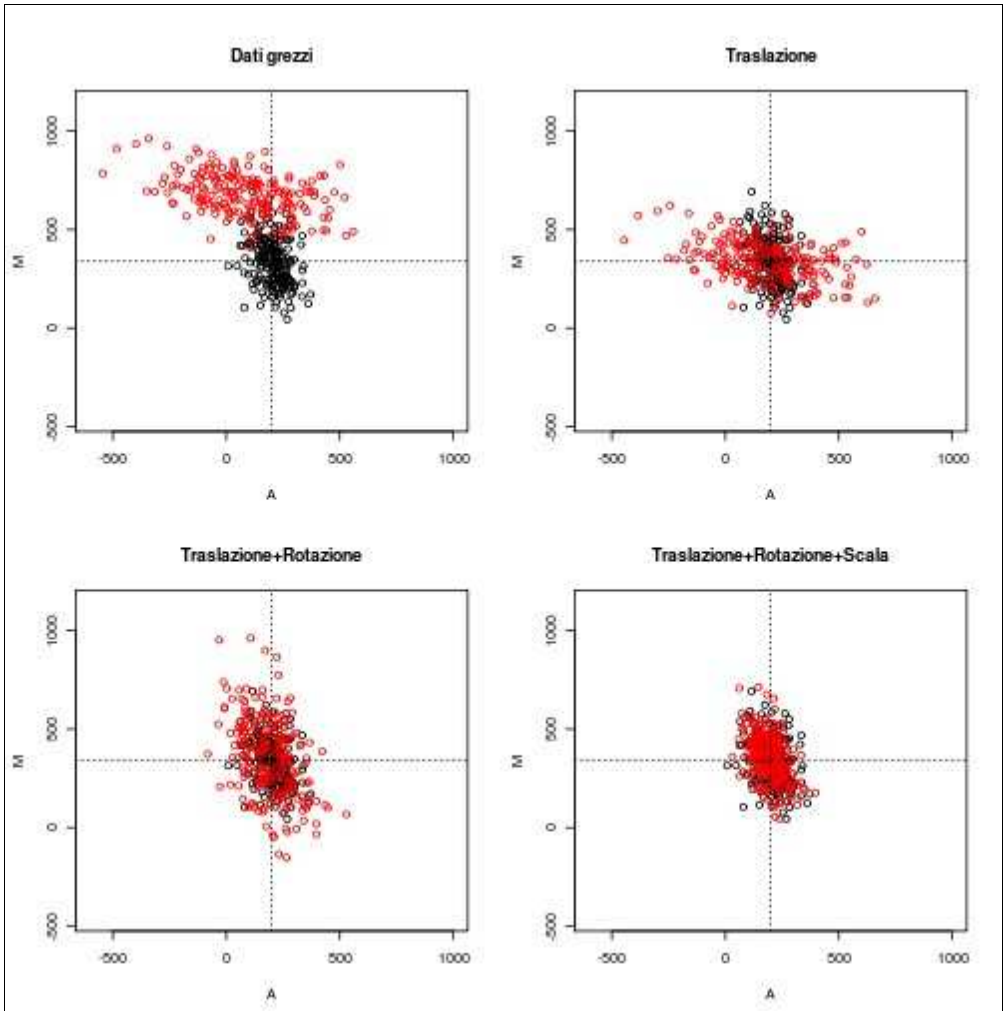


Grafico 1.II: Esempio di GPA per due nuvole di punti: i dati grezzi e i tre passi della normalizzazione

Nel lavoro gli Autori mostrano numerosi confronti tra la *GPA* e sei normalizzazioni molto comuni per array a doppio canale: *Global, Loess, Scale, Quantile, Vsn* e l'*Housekeeping method per array-boutique*, cioè microarray preparati ad hoc per determinate ricerche, nei quali la maggior parte dei geni potrebbe essere DE.

I confronti mostrano risultati promettenti i quali, unitamente all'assenza di assunzioni sulla distribuzione dei dati, rendono la *GPA* un metodo di normalizzazione molto versatile e robusto per diversi tipi di dataset. In particolare potrebbe risultare migliore per gli *array-boutique*, costruiti in maniera specifica per determinate situazioni e che per questo possiedono una struttura di espressione diversa, anche di molto, da quella ipotizzabile per un array *genome-wide*.

1.2.3.i. Da due canali ad uno: l'implementazione ciclica

La normalizzazione *GPA* presentata da *Xiong et al.* (2008) è pensata per la nuvola di punti del grafico *MA* proveniente da un array a doppio canale.

Nel caso di dati di microarray ad un solo canale, ogni osservazione di uno spot di *casi* è un valore singolo di intensità, senza nessuna naturale associazione con quello di uno specifico *controllo*. In questo modo tutte le possibili associazioni tra *casi* e *controlli* sono molte e nessuna è logicamente preferenziale rispetto alle altre.

Per ovviare a questo problema *Bolstand et al.* (2003) hanno adattato la normalizzazione *Loess,* specifica per tecnologia doppio canale, a quella a canale singolo basandosi su tutti gli *MA*-plot costruiti ciclicamente su tutti i possibili accoppiamenti tra i *casi* e i *controlli*.

Anche la *GPA*, essendo un metodo di trasformazione geometrica di nuvole di punti di un grafico *MA*, è per sua natura legata alla tecnologia a doppio canale, per il quale le coppie di valori osservati sono immediatamente rappresentabili come punti di un piano bidimensionale.

Anche in questo caso, come in quello della normalizzazione Loess implementata in maniera ciclica, è possibile ricondurre i dati di intensità singole a tutte le coppie *caso-controllo*.

Nel presente lavoro si mostreranno tre diversi modi di implementare ciclicamente la GPA, adattandola alla specificità dei microarray a singolo canale.

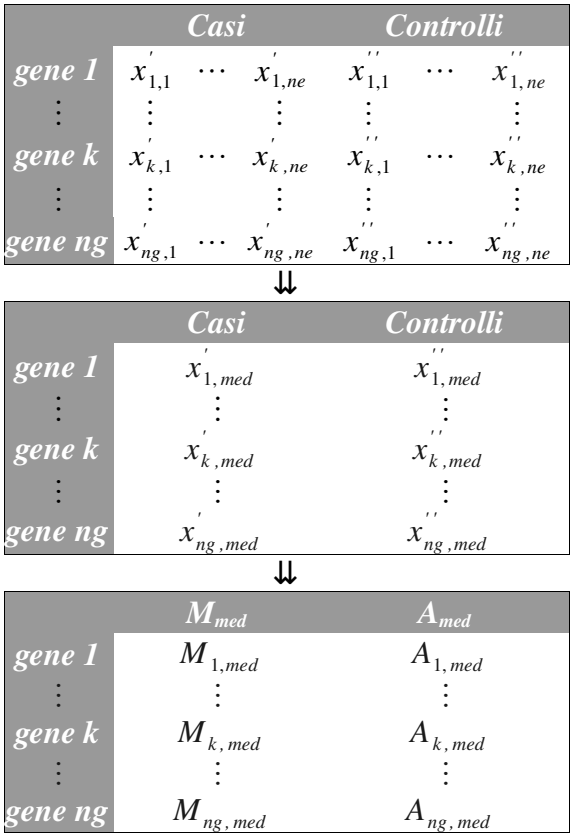
1.2.3.ii. La GPA mediana

La prima implementazione proposta per una *GPA* ciclica è quella che chiameremo *GPA mediana*, che non prende in considerazione tutte le possibili coppie di *casi-controlli,* ma è computazionalmente più semplice ed agile.

Come detto in precedenza, l'*Analisi Procrastica Generalizzata* consiste nell'adattare una nuvola di punti ad una nuvola preesistente; questa è detta *reference* o *target*, in quanto costituisce l'insieme di punti di riferimento o l'obiettivo a cui ci si vuole adattare.

La *GPA mediana* usa come *reference* la trasformazione *MA* dell'intensità mediana dei *casi* contro l'intensità mediana dei *controlli*.

Tabella 1.V: Costruzione del reference per la GPA mediana



Le coppie di intensità da normalizzare rispetto al *target* appena calcolato sono costituite, in questa versione semplificata, da tutte le possibili coppie del tipo $(x';x'')$. Se quindi, come accade nel caso più semplice, sia i *casi* che i *controlli* sono di numerosità ne, il numero totale di coppie da normalizzare sarà ne^2.

Una volta normalizzate tutte le coppie (M, A) con l'algoritmo GPA, queste vengono poi riconvertite in valori di intensità. L'intensità finale dell'esperimento *i* è quindi rappresentata dalla mediana delle intensità provenienti da quelle coppie di (M, A) a cui l'esperimento *i* apparteneva.

1.2.3.iii. Le GPA cicliche media e mediana

L'implementazione completa della GPA ciclica, ma computazionalmente più dispendiosa rispetto alla precedente, viene proposta in due versioni molto simili, diversificate per la statistica scelta come stima dell'intensità normalizzata: la *media* o la *mediana.*

La prima è una scelta più canonica e leggermente più rapida in presenza di molti dati, la seconda ha il pregio della robustezza rispetto a dati anomali. Si può ipotizzare che la prima sia preferibile in presenza di dati molto numerosi e di qualità piuttosto buona, con anomalie il cui peso risulta irrilevante, mentre la seconda può essere una scelta in caso di una quantità più contenuta di dati, tra i quali le anomalie possono esercitare un'influenza maggiore, o con un sospetto di molti dati aberranti.

Nel caso della *GPA ciclica media* o *mediana* il target è calcolato, per ogni gene, come la media (mediana) della trasformata *MA* di ogni possibile coppia di esperimenti, indipendentemente dal gruppo di appartenenza (*casi* o *controlli*). In questo caso il numero totale di coppie di osservazioni considerate è

$$N_{COPPIE} = \sum_{i=1}^{ne_{tot}} (i) = \left[\frac{ne_{tot}}{2} \cdot (ne_{tot} + 1) \right] \qquad (1.11)$$

con $ne_{tot} = ne_{casi} + ne_{controlli}$.

Come si può notare dal *Grafico 1.III*, all'aumentare della numerosità degli esperimenti, il numero di coppie considerate dalle *GPA cicliche* aumenta molto più velocemente rispetto a quella della *GPA mediana*.

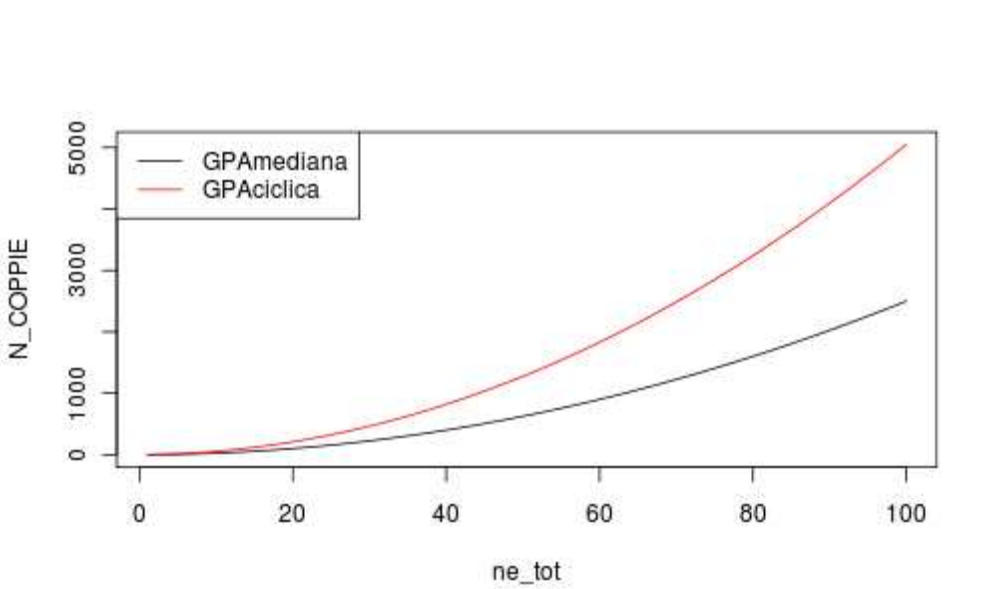


Grafico 1.III: Numero di coppie di osservazioni con cui lavorano la GPA mediana e le GPA cicliche a seconda del numero totale di esperimenti, nel caso semplice di un disegno di studio bilanciato

Una volta ottenuto il *target*, si creano tutte le possibili coppie di esperimenti, si convertono in *MA* e si utilizzano quindi per la normalizzazione. Su tutte queste coppie si eseguirà la procedura GPA e, per ogni esperimento, le intensità normalizzate saranno ottenute come media o mediana dei valori di intensità provenienti da tutte le $n_{tot}-1$ coppie che avevano come primo o come secondo elemento l'esperimento in questione.

Tuttavia, a differenza della *GPA mediana* che vedeva ogni esperimento comparire sempre al primo oppure sempre al secondo posto delle coppie da normalizzare, adesso ognuno compare a volte come primo e a volte come secondo.

La coppia *MA* media (mediana) normalizzata deve essere ottenuta tenendo conto di questo fatto. Un modo per fare ciò è il seguente:

1. si selezionano tutte le $n_{tot}-1$ coppie *MA* normalizzate che avevano l'esperimento *i*-esimo come primo o come secondo elemento
2. si calcola A_{medio} ($A_{mediano}$) come media (mediana) dell'elemento *A* delle coppie in (1)
3. si calcola M_{medio} ($M_{mediano}$) come media (mediana) di
   - *M* delle coppie in (1) che avevano l'esperimento *i*-esimo come primo elemento
   - *(-M)* delle coppie in (1) che avevano l'esperimento *i*-esimo come secondo elemento.

Le operazioni effettuate sono giustificate dalle relazioni esplicitate nella (1.12).

$$\begin{aligned} M &= \log(x/y) \rightarrow (-M) = \log(y/x) \\ A &= \log(x \cdot y) \rightarrow (-A) = \log(y \cdot x) = A \end{aligned} \tag{1.12}$$

Di fatto quello che si fa ai punti (2) e (3) è un riordinamento di tutte le coppie contenenti l'esperimento *i*-esimo, in modo che questo compaia sempre come primo elemento.

Così facendo è possibile riconvertire la coppia MA_{media} ($MA_{mediana}$) in due valori di intensità dei quali il primo è la normalizzazione dell'*i*-esimo array.

1.3. Metodi di confronto delle normalizzazioni

L'obiettivo del presente lavoro, come già detto in precedenza, consiste nel confronto della bontà dei diversi metodi di normalizzazione in termini di capacità di renderli adatti ad una analisi statistica inferenziale che sia in grado di individuare al meglio i geni effettivamente DE.

Questa analisi comparativa avverrà grazie a strumenti diversi a seconda dei dati utilizzati: simulati col metodo di *Balagurunathan*, riadattato come illustrato nel *Paragrafo 1.1*, e reali come descritti nel *Paragrafo 0.3*.

In entrambi i casi sarà disponibile una informazione fondamentale del disegno sperimentale: il gruppo di studio di appartenenza dei vari campioni. Nel caso di dati simulati, ogni caratteristica è nota e controllabile, dunque anche l'espressione differenziale, mentre nel caso di dati reali si ricorrerà alla tecnica della titolazione di due campioni biologici in quattro diverse proporzioni, i cui dettagli sono presentati nel seguito (*Paragrafo 1.3.3*).

Ciò che distingue le due situazioni è principalmente la conoscenza dei geni DE e di quelli EE. Nel primo caso il modello di simulazione è costruito in modo da tenere traccia di quali valori di intensità sono stati generati con una componente di sovra- o sottoespressione e della sua direzione, mentre nel secondo non si è in possesso di questa informazione, che costituisce infatti l'obiettivo conoscitivo delle analisi di dati reali di microarray.

1.3.1. Test Significance Analysis of Microarray

Uno strumento che verrà utilizzato in entrambe le fasi del confronto tra i metodi di normalizzazione è il test *Significance Analysis of Microarray (SAM)* proposto da *Tusher et al.* nel 2001. Il *test SAM* è una versione moderata del *test t* per la differenza tra medie, riadattato alle caratteristiche di dati di microarray. L'idea alla base di questo test è che se un gene è effettivamente DE, allora la differenza di luminosità tra i due gruppi sarà molto ampia, mentre se il gene è EE questa differenza si attesterà intorno al valore 0.

Il test, effettuato indipendentemente gene per gene, ha come ipotesi nulla l'espressione non differenziata tra i gruppi, due nel caso di dati simulati, quattro nel caso di dati reali. In generale per ciascun gene g, le ipotesi del *test SAM* con K gruppi sono:

$$\begin{aligned} H_{0,g}&: \mu_1(g) = ... = \mu_K(g) \\ H_{1,g}&: \bigcup_{k \neq j}^{k,j \leq K} \left(\mu_k(g) \neq \mu_j(g)\right) \end{aligned} \tag{1.13}$$

con $g = 1..\mathrm{ng}$ e $\mu_k(g)$ il valore atteso di intensità per il gene g nel gruppo k.

Per ciascun gene verrà calcolata la distanza d_g tra i valori medi di intensità nei gruppi, standardizzata per mezzo della deviazione standard s_g e della correzione s_0 che vincola il coefficiente di variazione ad un valore costante, annullando l'effetto di una eventuale eteroschedasticità.

$$d_g = \frac{r_g}{s_g - s_0} \tag{1.14}$$

Il primo caso riportato, più semplice, presenta dati appartenenti a due classi non appaiate; il numeratore r_g della (1.14) sarà allora la differenza tra i livelli medi dei due gruppi.

$$r_g = \overline{x}_{g,2} - \overline{x}_{g,1} \tag{1.15}$$

con $\{\overline{x}_{g,k} = \sum_{j \in C_k} x_{g,j} / n_k\}_{k=1,2}$ le medie campionarie dei due gruppi e $\{C_k\}_{k=1,2}$ i due gruppi di esperimenti di numerosità n_1 e n_2 .

La deviazione standard del gene *g* sarà definita come

$$s_g = \sqrt{\alpha \cdot \left(\sum_{j \in C_1} (x_{g,j} - \overline{x}_{g,1})^2 + \sum_{j \in C_2} (x_{g,j} - \overline{x}_{g,2})^2 \right)} \tag{1.16}$$

con $\alpha = \frac{(1/n_1 + 1/n_2)}{(n_1 + n_2 - 2)} = \frac{n_1 + n_2}{n_1 \cdot n_2 \cdot (n_1 + n_2 - 2)}$ che, nel caso bilanciato $n_1 = n_2 = n$, diventa più semplicemente $\frac{2n}{n^2 \cdot (2n - 2)} = \frac{1}{n \cdot (n-1)}$

Nel caso di dati reali, si incontrerà una situazione più complessa, con quattro gruppi $\{C_k\}_{k=1..4}$ di numerosità $\{n_k\}_{k=1..4}$ e media campionaria dell'intensità misurata $\{\overline{x}_{g,k} = \sum_{j \in C_k} x_{g,j} / n_k\}_{k=1..4}$. In questo caso il numeratore r_g (1.15) diventa

$$r_g = \sqrt{\left(\sum_{k=1}^{4} n_k / \prod_{k=1}^{4} n_k \right) \cdot \sum_{k=1}^{4} n_k \cdot \left(\overline{x}_{g,k} - \overline{x}_g \right)^2} \tag{1.17}$$

dove $\overline{x}_g = \left(\sum_{j \in C_k} \overline{x}_{g,j} \cdot n_k \right) / \sum_{k=1}^{4} n_k$ rappresenta la media campionaria totale del gene g. Analogamente anche la deviazione standard (1.16) diventa

$$s_g = \sqrt{\alpha \cdot \left(\sum_{k=1}^{4} \sum_{j \in C_k} (x_{g,j} - \overline{x}_{g,k})^2 \right)} \qquad (1.18)$$

con $\alpha = \frac{\sum^{4} 1/n_k}{\sum^{4} (n_k - 1)} = \frac{\sum^{4} n_k}{\prod^{4} n_k \cdot \left(\sum^{4} n_k - 4 \right)}$ che, nel caso bilanciato, diventa

$$\frac{4n}{n^4 \cdot (4n-4)} = \frac{1}{n^3 \cdot (n-1)}$$

Infine è dimostrato che s_0, costante di correzione del coefficiente di variazione, può essere calcolata come il novantesimo percentile campionario degli s_g.

In questo modo si ottiene un valore osservato della statistica test d_g per ogni gene.

Tabella 1.VI: Calcolo dei valori osservati della statistica test SAM; esempio per due gruppi bilanciati

	Casi			*Controlli*			d_g
gene 1	$x'_{1,1}$	$\cdots$	$x'_{1,ne}$	$x''_{1,1}$	$\cdots$	$x''_{1,ne}$	d_1
$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$
gene k	$x'_{k,1}$	$\cdots$	$x'_{k,ne}$	$x''_{k,1}$	$\cdots$	$x''_{k,ne}$	d_k
$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\vdots$
gene ng	$x'_{ng,1}$	$\cdots$	$x'_{ng,ne}$	$x''_{ng,1}$	$\cdots$	$x''_{ng,ne}$	d_{ng}

Dal momento che questi valori esprimono una distanza tra i livelli medi dei gruppi esaminati, ci si aspetta che i geni attivi abbiano valori della statistica test con valore assoluto maggiore rispetto a quelli inattivi.

Per questo motivo è utile ordinare i geni in esame in maniera decrescente rispetto al valore assoluto di d_g, da quello che può essere considerato più verosimilmente come DE a quello che l'evidenza empirica indica come il meno plausibile. Questa lista sarà utilizzata in due maniere differenti nei due passi della comparazione tra normalizzazioni.

Il *test SAM* utilizza poi un approccio permutazionale per il calcolo della significatività, approccio che nel presente elaborato non viene preso in considerazione. La bontà delle normalizzazioni verrà invece valutata (come descritto nel successivo paragrafo) per liste crescenti di geni DE selezionate solo sulla base dell'ordinamento della statistica test.

1.3.2. Curve ROC per dati simulati

Nella prima fase dello studio comparativo la conoscenza dei geni effettivamente DE (veri positivi) permette di calcolare, per ogni possibile classificazione in espressi e non espressi, i due indici di bontà che si è soliti usare per valutare questo tipo di tecniche:

- la *sensibilità*: capacità di individuare i positivi (attivi), ottenuta come percentuale di veri positivi tra i geni classificati come positivi

- la *specificità*: capacità di classificare correttamente i negativi (inattivi), ottenuta come percentuale di veri negativi tra i negativi

Tabella 1.VII: Calcolo di sensibilità e specificità per un criterio di classificazione

		Test SAM		
		EE	**DE**	
Reale	**EE**	a	b	→ spec=a/a+b
	DE	c	d	→ sens=d/c+d

1.3.2.Curve ROC per dati simulati

A partire dalla lista dei geni ordinata in maniera decrescente secondo il valore assoluto della *statistica test SAM*, è possibile dividere i geni in due gruppi, a seconda che il valore d_g sia al di sopra o al di sotto di una soglia scelta. Facendo variare questa soglia su tutta l'ampiezza dell'intervallo tra il minimo e il massimo dei valori osservati, si ottengono tante coppie di sensibilità e specificità che corrispondono a diversi criteri per classificare i geni sulla base dello stesso test.

Con questo procedimento si costruiscono le cosiddette *curve ROC (Receiver Operating Characteristic)* che riportano il livello di sensibilità all'aumentare dell'errore di secondo tipo, cioè il complementare della specificità.

Nel *Grafico 1.IV* è riportato un esempio di tre *curve ROC* che mostrano l'andamento che lega i due tipi errore: all'aumentare di uno diminuisce l'altro, anche se non in maniera lineare né necessariamente secondo una funzione determinata. Questo tipo di grafici è costruito in modo che punti collocati in alto a sinistra hanno alta sensibilità e alta specificità, mentre quelli in basso a destra hanno bassa sensibilità e bassa specificità.

La diagonale rappresenta l'andamento di un criterio di classificazione completamente casuale il quale, al variare della percentuale di geni da classificare come DE, li seleziona assegnando a ciascuno la stessa probabilità.

Nell'esempio presentato, il criterio B ha sensibilità uniformemente maggiore di C e di A, mentre A è migliore di C per specificità maggiori di 0,75 e peggiore per quelle minori.

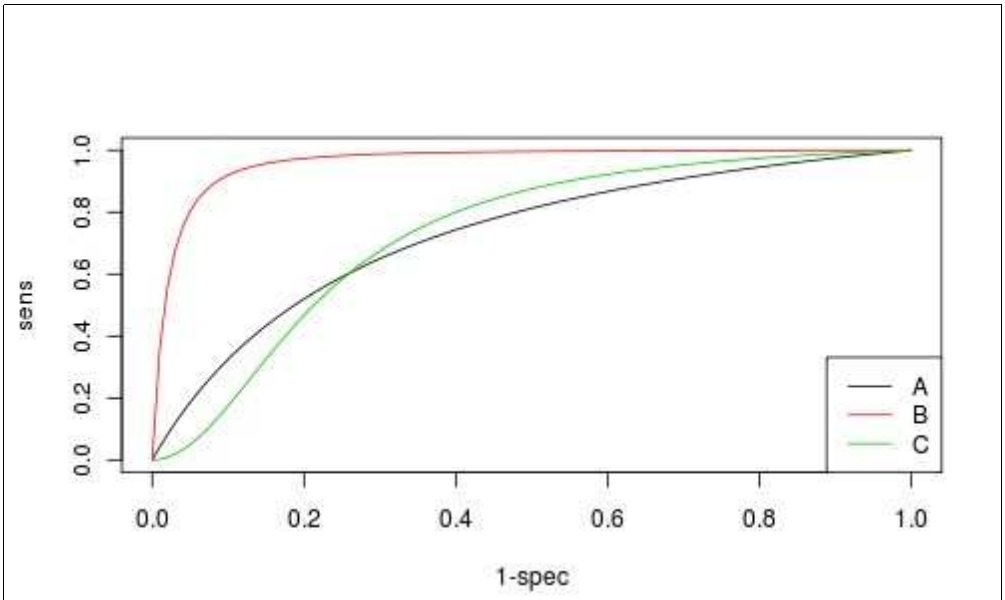


Grafico 1.IV: Esempio di curve ROC

In questo caso si potrebbe concludere che il criterio B è universalmente il migliore, mentre se fosse necessario scegliere tra A e C sarebbero necessarie considerazioni circa la quantità di geni DE che ci si attende: se, come nelle applicazioni classiche, questa è molto piccola allora sarebbe plausibile pensare che si dovrà lavorare con basse sensibilità e alte specificità, preferendo quindi il criterio A al criterio C.

1.3.3. Criterio della titolazione per dati reali

Nel momento in cui si andrà a valutare la bontà dei metodi di normalizzazione per mezzo di dati reali, non sarà disponibile l'informazione riguardante i geni effettivamente DE.

Shippy et al. hanno presentato nel 2006 il metodo di titolazione di campioni di RNA per la valutazione della riproducibilità di piattaforme di microarray e di tecniche di normalizzazione. Il metodo proposto consiste nell'utilizzare due campioni di RNA indipendenti (*A* e *B*), in cui la maggior parte dei geni è DE, e due titolazioni di *A* e *B* in rapporti di 3:1 *(C)* e 1:3 *(D)*.

I geni DE avranno livelli di espressione molto diversi, con *A>B* o *B>A*. Sfruttando questa informazione e i rapporti ci concentrazione di *C* e *D* è possibile ricavare le relazioni che seguono.

Se i livelli di espressione del gene *g* sono tali per cui i livelli nella classe *A* sono maggiori di quelli nella classe B *(A>B)*, allora

$$A > B \;\rightarrow\; \frac{1}{4}A > \frac{1}{4}B \;\rightarrow\; A > \frac{3}{4}A + \frac{1}{4}B \;=\; C \;\rightarrow\; A > C \tag{1.19}$$

$$A > B \;\rightarrow\; \frac{1}{2}A > \frac{1}{2}B \;\rightarrow\; \frac{3}{4}A + \frac{1}{4}B > \frac{1}{4}A + \frac{3}{4}B \;\rightarrow\; C > D \tag{1.20}$$

$$A > B \;\rightarrow\; \frac{1}{4}A > \frac{1}{4}B \;\rightarrow\; D \;=\; \frac{1}{4}A + \frac{3}{4}B > B \;\rightarrow\; D > B \tag{1.21}$$

Unendo le informazioni ricavate in (1.19), (1.20) e (1.21) si ottiene che i livelli di espressione del gene g nelle quattro classi dovranno seguire l'andamento:

$$A > C > D > B \tag{1.22}$$

Se invece per un determinato g i livelli di espressione sono tali per cui *A<B,* allora, con analoghi passaggi, si ricava che i livelli nelle quattro classi dovranno seguire l'andamento:

$$B > D > C > A \tag{1.23}$$

In virtù di queste relazioni, per ogni gene che si suppone essere DE è ragionevole attendersi di trovare un ordinamento stretto tra le medie di espressione nelle quattro classi. Infatti, al di là di errori casuali e deviazioni sistematiche, il livello medio di intensità dei quattro campioni dovrà rispettare l'ordinamento dettato dalla natura del processo di diluizione.

Per i geni EE, invece, i livelli di intensità avranno un ordinamento che varia nelle diverse ripetizioni, portando a livelli medi non necessariamente ordinati secondo i due trend indicati in (1.22) e (1.23).

Per valutare la bontà delle normalizzazioni si contano i geni con trend del tipo (1.22) o (1.23) tra i primi k della lista ordinata secondo il valore assoluto del *test SAM*, dopo ciascuna normalizzazione. Variando il parametro k si ottengono diverse valutazioni che nel loro complesso forniscono una informazione circa la plausibilità di ciascun metodo.

Da notare infine che un'alta percentuale di trend corretti non garantisce in realtà la correttezza della normalizzazione, dal momento che va a verificare una caratteristica dei dati normalizzati che è necessaria ma non sufficiente per la corretta individuazione dei geni DE.

$$\begin{aligned} DE &\Rightarrow (A>C>D>B)\cup(B>D>C>A) \\ DE &\not\Leftarrow (A>C>D>B)\cup(B>D>C>A) \end{aligned} \tag{1.24}$$

Pur con tale consapevolezza è però ragionevole ipotizzare che percentuali molto alte di trend corretti siano una indicazione di un funzionamento sostanzialmente buono del metodo di classificazione e quindi della normalizzazione precedentemente applicata.

2. Risultati

Dopo dopo aver illustrato in dettaglio gli strumenti usati, in questo capitolo si presentano i risultati dei diversi passi dello studio: la simulazione, la normalizzazione e il confronto.

Lo studio ha preso in considerazione tre diversi tipi di dati:

1. dati simulati con caratteristiche simili alle assunzioni classiche: bassa percentuale di geni DE e bilanciamento tra geni sovra- e sottoespressi;

2. dati simulati con alta incidenza di espressione differenziale (*boutique-array*) e, in un caso, anche con sbilanciamento;

3. dati reali, con molti geni DE.

Per quanto riguarda la simulazione e la normalizzazione, si mostreranno solo alcuni grafici di esempio, con il solo scopo di dare conto del tipo di dati generati e dei cambiamenti che le trasformazioni introducono.

I risultati dei confronti saranno invece presentati nel dettaglio, essendo questo lo scopo primario di tutto il lavoro.

2.1. Dati simulati secondo le assunzioni classiche

A scopo esemplificativo si mostrano i risultati della simulazione di una matrice di dati di espressione con 10,000 geni e 30 esperimenti bilanciati. Il *Grafico 2.I* mostra l'*MA-plot* della simulazione, in cui è evidente una leggera distorsione sistematica per bassi livelli medi di espressione.

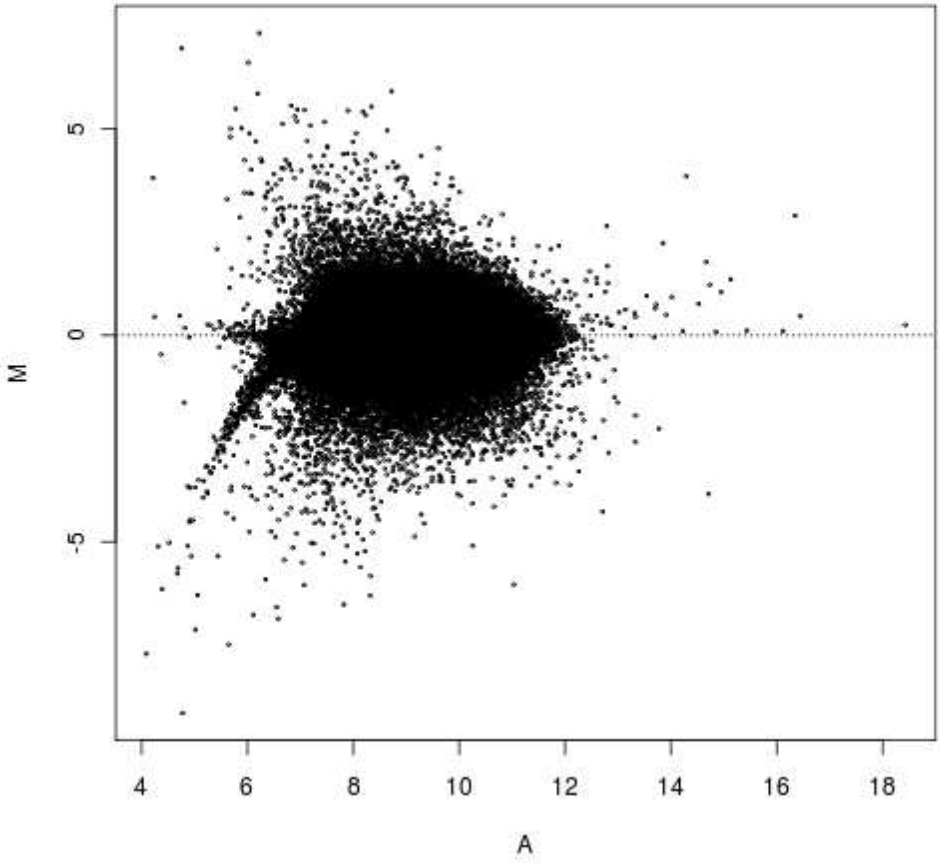


Grafico 2.I: Grafico MA dei dati simulati mediante il modello di Balagurunathan riadattato per dati a singolo canale

Il *Grafico 2.II* mostra il cambiamento della nuvola *MA-plot* dovuto alle normalizzazioni considerate (in rosso).

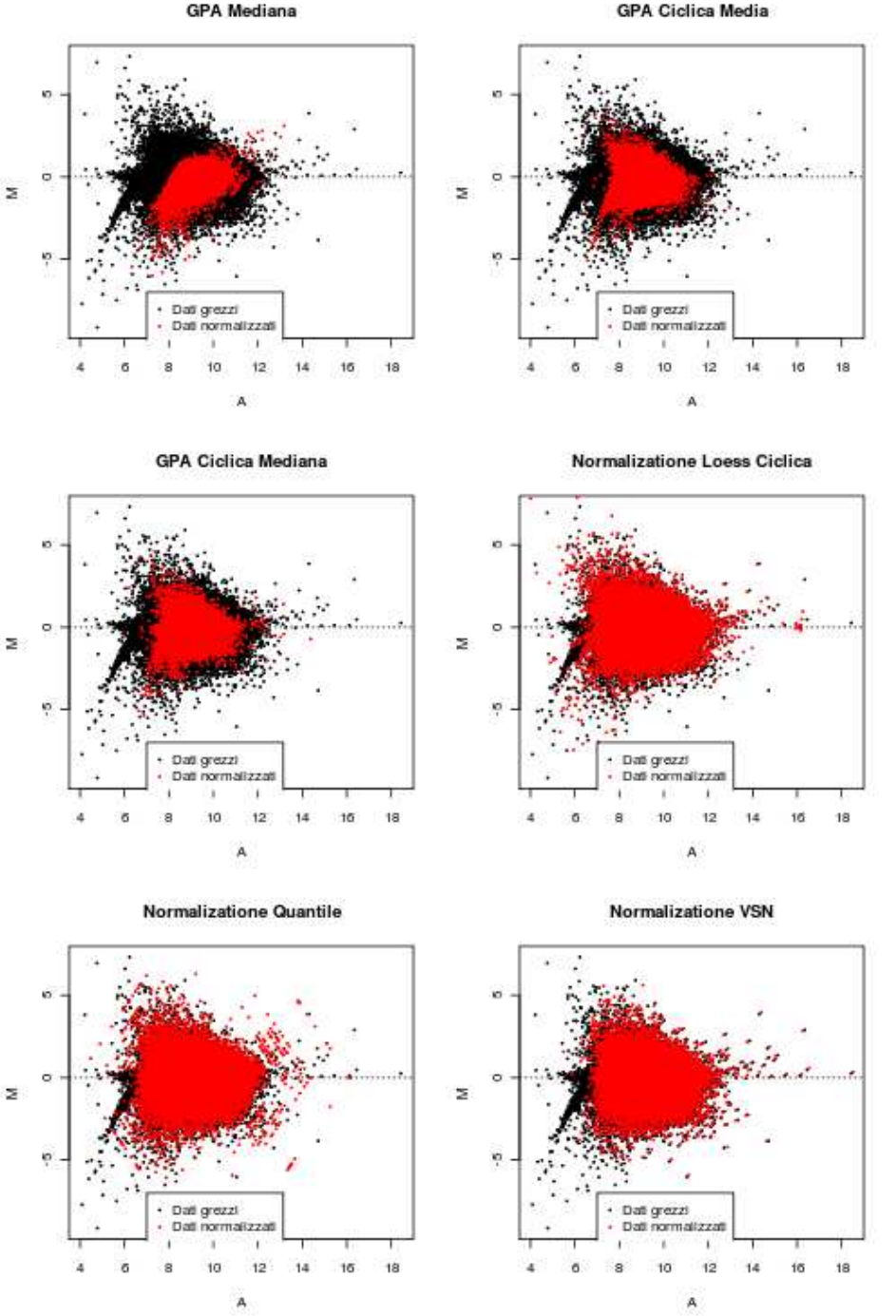


Grafico 2.II: Normalizzazioni del grafico MA della prima simulazione

La prima cosa da notare è come tutte le normalizzazioni riescono ad eliminare la distorsione sistematica presente per valori bassi di intensità media. Inoltre, le tre normalizzazioni classiche (*Loess ciclica*, *Quantile* e *Vsn*) introducono modificazioni minime rispetto ai dati grezzi, mentre le tre *GPA* sembrano modificarli in modo più marcato, pur mantenendo la struttura e la forma nei loro tratti generali.

2.1.1. Dati da un singolo laboratorio

Dopo aver trasformato tutte le matrici con le normalizzazioni selezionate, si sono calcolati i valori osservati della *statistica test SAM* per ogni gene e si sono costruite le sette *curve ROC* rispettivamente per i dati grezzi e per le sei normalizzazioni, nel modo descritto nel capitolo precedente.

Nel *Grafico 2.III* (relativo ai risultati derivati da una sola matrice simulata), come in tutti quelli analoghi che seguiranno in questo capitolo, è stata aggiunta una retta che rappresenta il livello di *sensibilità+specificità* massimo tra tutte le *curve ROC*.

Formalmente si tratta della curva di livello della funzione $g(1-spec,\ sens)$ più alta tra quelle che contengono almeno un punto di almeno una delle *curve ROC*, con

$$g(x,\ y)=y-x \tag{2.1}$$

Il colore della retta è scelto in analogia con quello della curva che massimizza $g(\bullet)$. Inoltre è indicato anche il punto di tangenza con essa e i relativi valori di *sensibilità* e *specificità*.

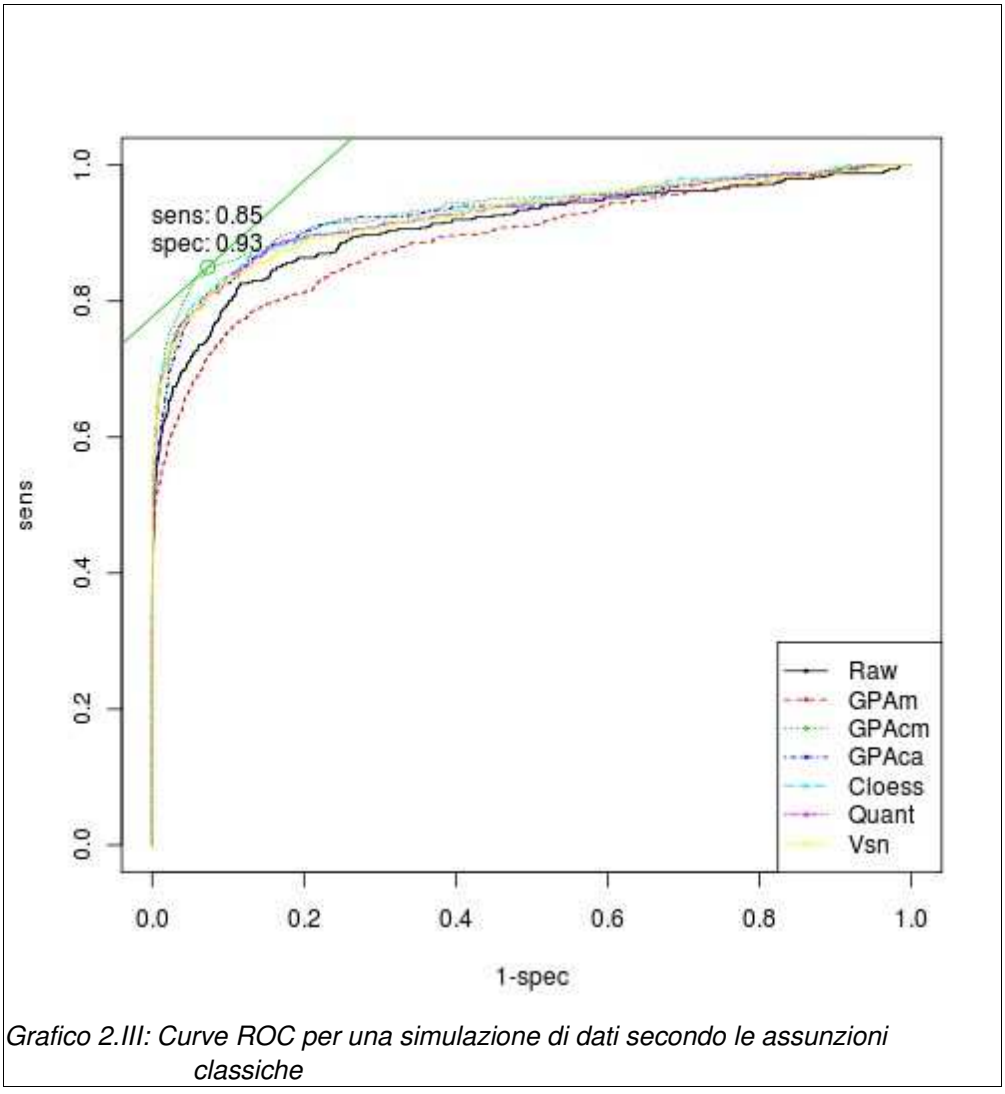


Grafico 2.III: Curve ROC per una simulazione di dati secondo le assunzioni classiche

Ripetendo la simulazione dei dati per dieci volte, effettuando i test sulle dieci matrici di dimensioni 10,000 × 30 e calcolando infine gli errori di primo e secondo tipo sull'insieme di tutte e dieci, si ottengono risultati analoghi, ma meno soggetti a fluttuazioni casuali (*Grafico 2.IV*).

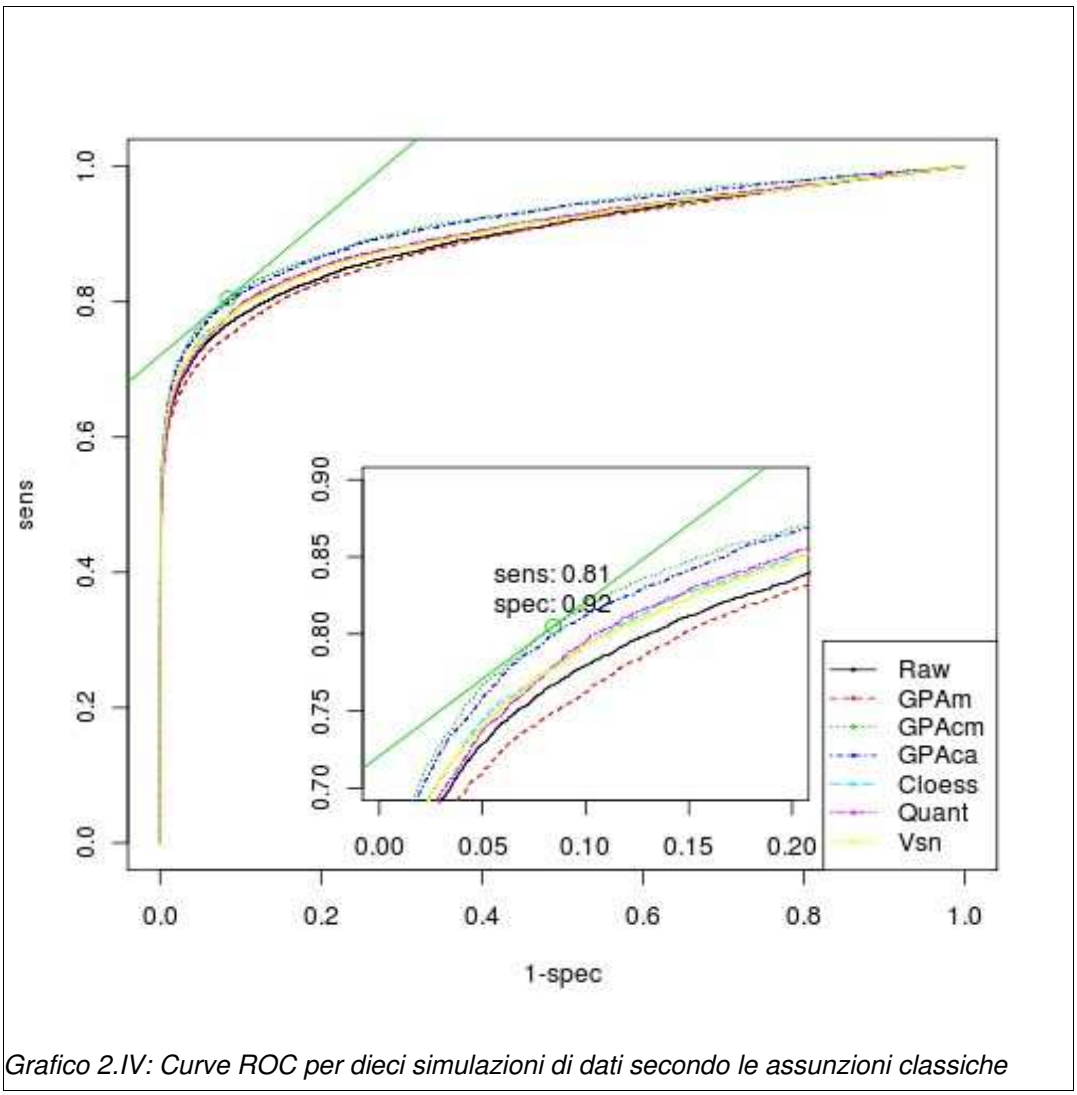


Grafico 2.IV: Curve ROC per dieci simulazioni di dati secondo le assunzioni classiche

2.1.2. Dati da due laboratori

L'analisi precedente può essere ripetuta in una situazione meno adatta ai metodi di normalizzazione, ma più interessante per le applicazioni pratiche: due gruppi di dati con caratteristiche piuttosto diverse possono costituire un test del rendimento delle normalizzazioni quando si ha a che fare con dati provenienti da due studi o laboratori diversi. Un esempio della diversità dell'MA-plot dovuto a simulazioni con parametri differenti è riportato nel *Grafico 2.V*.

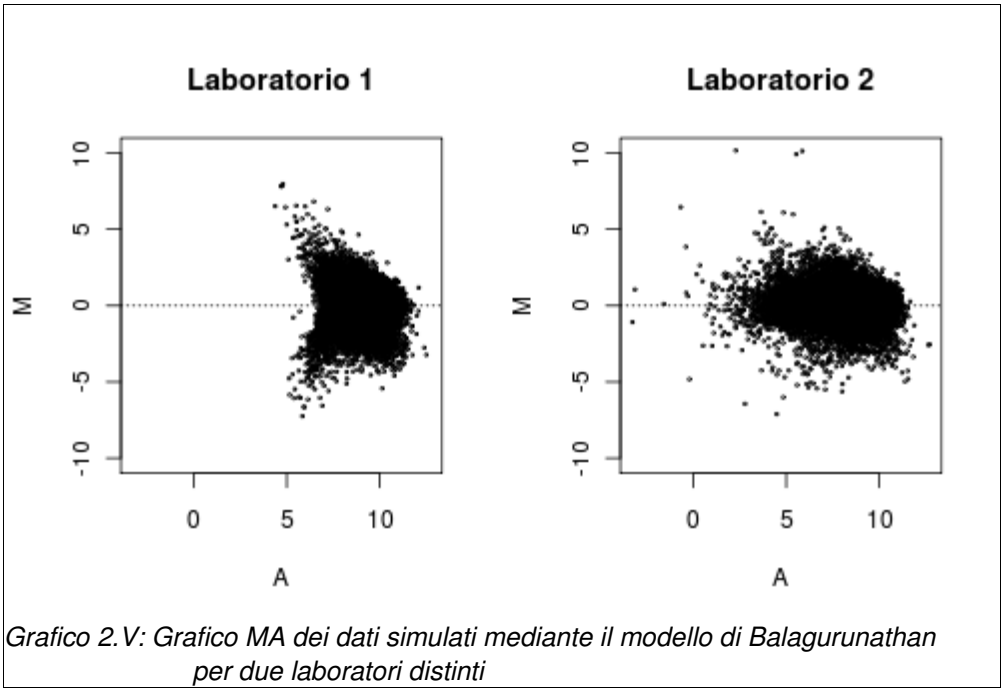


Grafico 2.V: Grafico MA dei dati simulati mediante il modello di Balagurunathan per due laboratori distinti

In questa situazione si ripete l'analisi con il *test SAM,* precedentemente illustrata e nei grafici seguenti sono riportate le *curve ROC* per una (*Grafico 2.VI*) e per dieci simulazioni (*Grafico 2.VII)*.

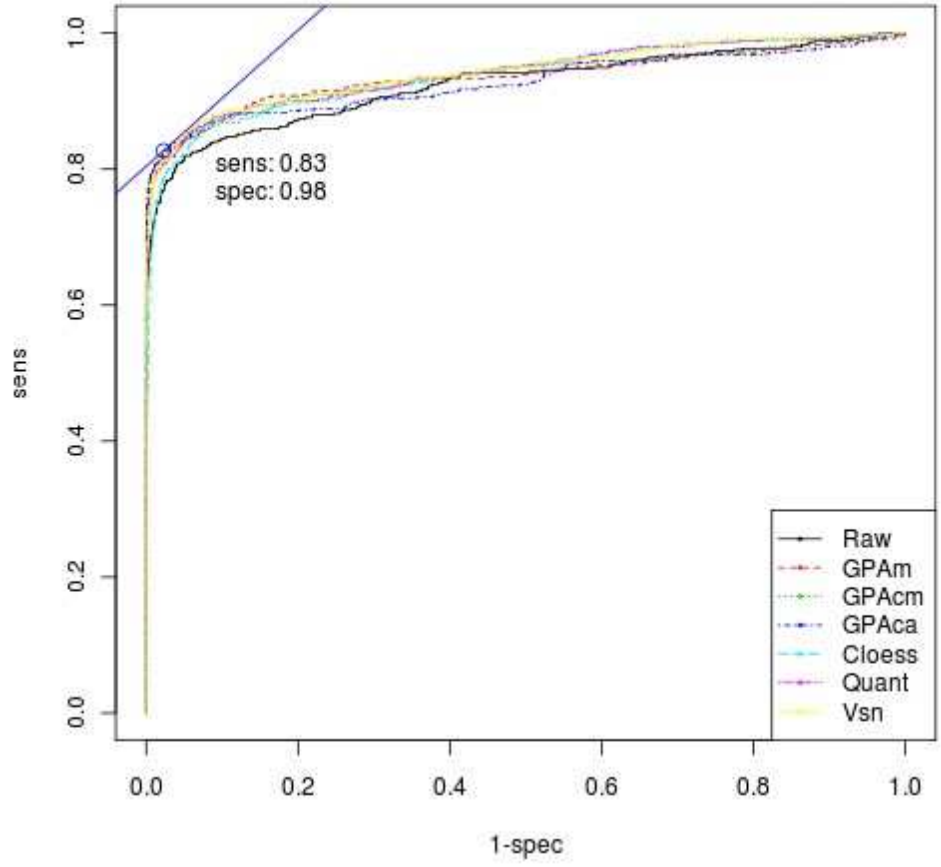


Grafico 2.VI: Curve ROC per una simulazione di dati da due laboratori, secondo le assunzioni classiche

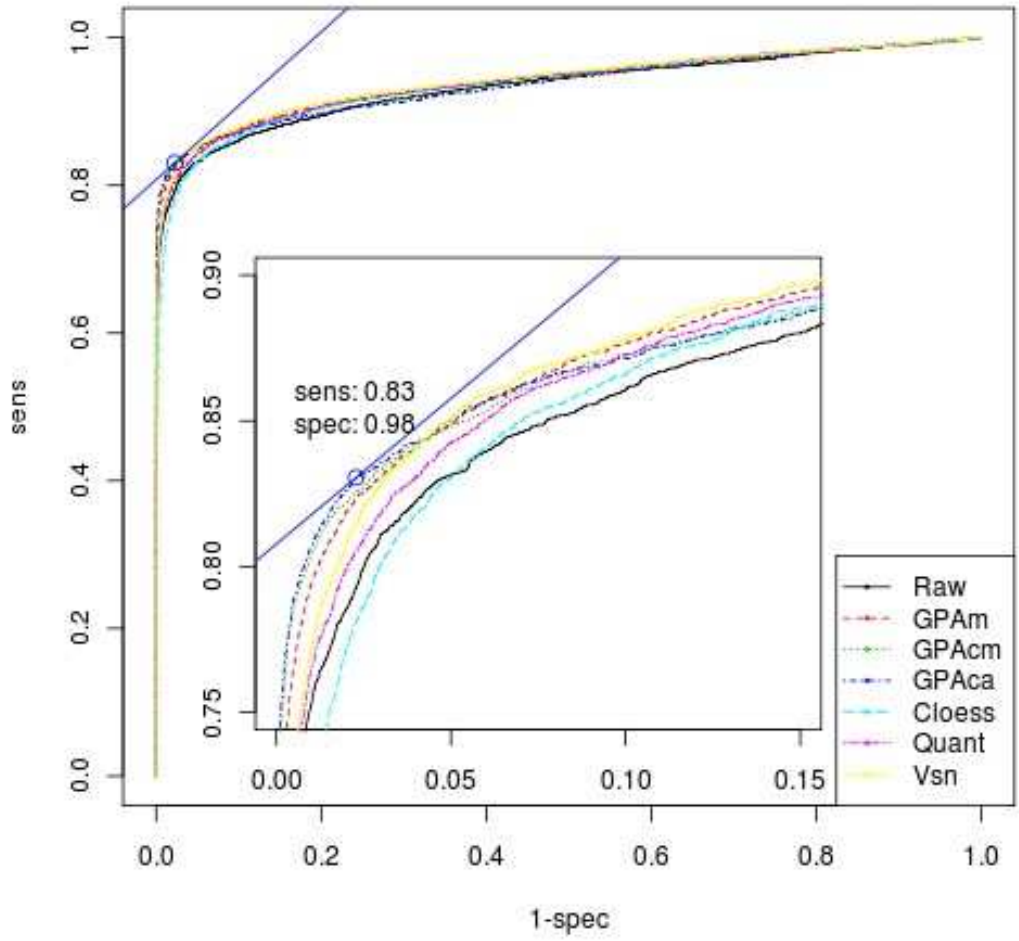


Grafico 2.VII: Curve ROC per dieci simulazioni di dati da due laboratori, secondo le assunzioni classiche

Come nel caso precedente, cambiando il numero di simulazioni si ottengono risultati simili, ma con fluttuazioni casuali di ampiezza visibilmente diversa.

2.2. Dati simulati per boutique-array

2.2.1. Dati da un singolo laboratorio

Cambiando le caratteristiche di espressione differenziale, tipiche dei cosiddetti boutique-array, le curve ROC cominciano a separarsi identificando chiaramente delle differenze significative tra normalizzazioni (*Grafico 2.VIII*).

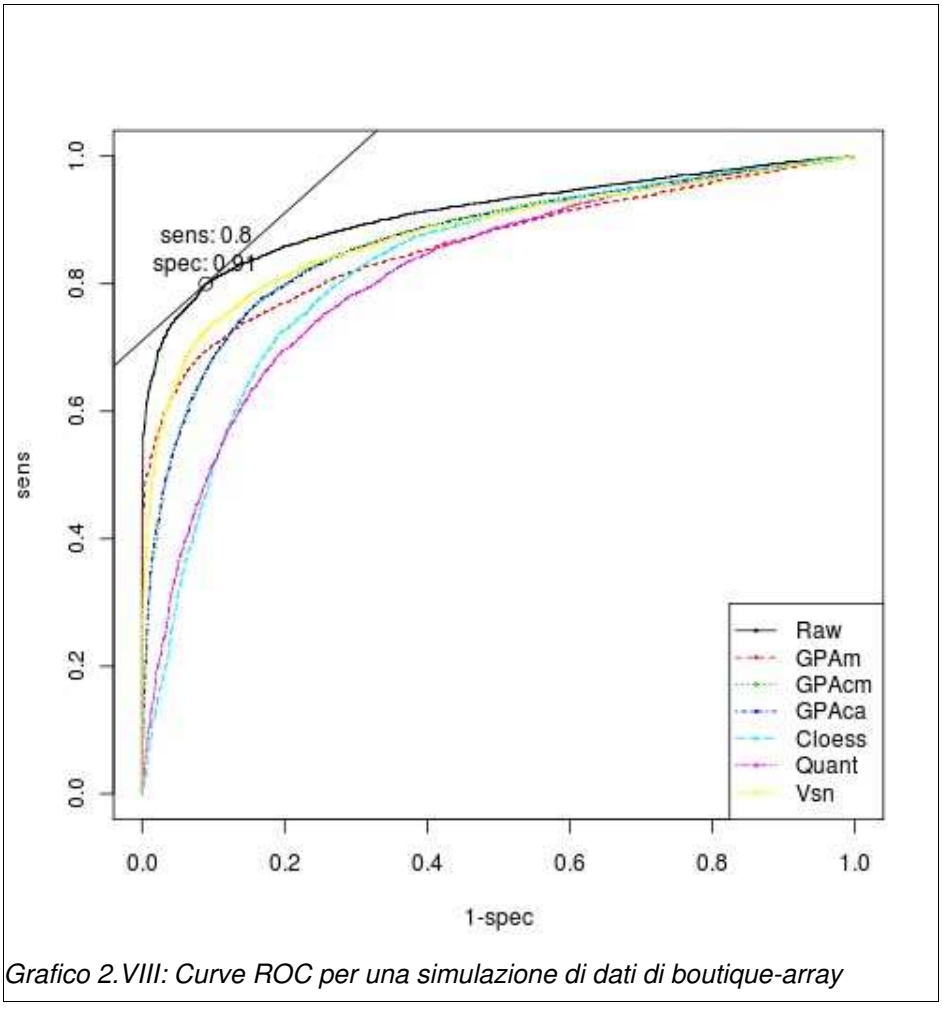


Grafico 2.VIII: Curve ROC per una simulazione di dati di boutique-array

Ripetendo la simulazione per dieci volte e reiterando il *test SAM* si ottengono le *curve ROC* del *Grafico 2.IX*.

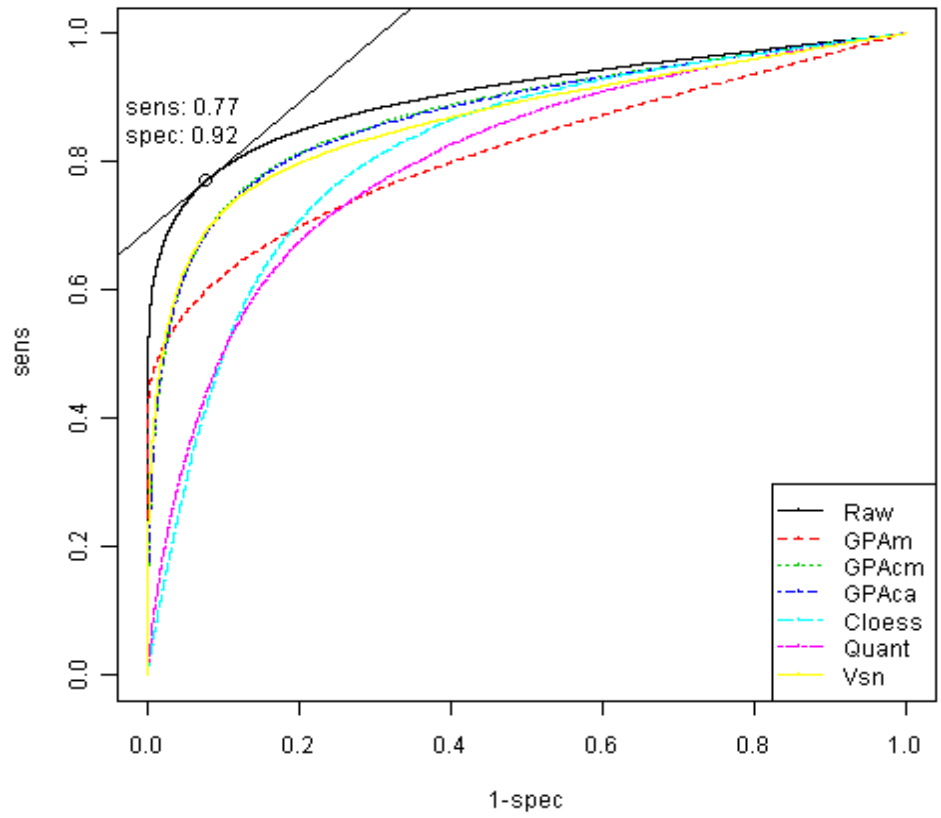


Grafico 2.IX: Curve ROC per dieci simulazioni di dati di boutique-array

2.2.2. Dati da due laboratori

L'analisi effettuata su dati simulati secondo le caratteristiche di due laboratori diversi, ma con caratteristiche di espressione differenziale tipiche dei boutique-array, dà origine alle *curve ROC* dei grafici *2.X* e *2.XI*.

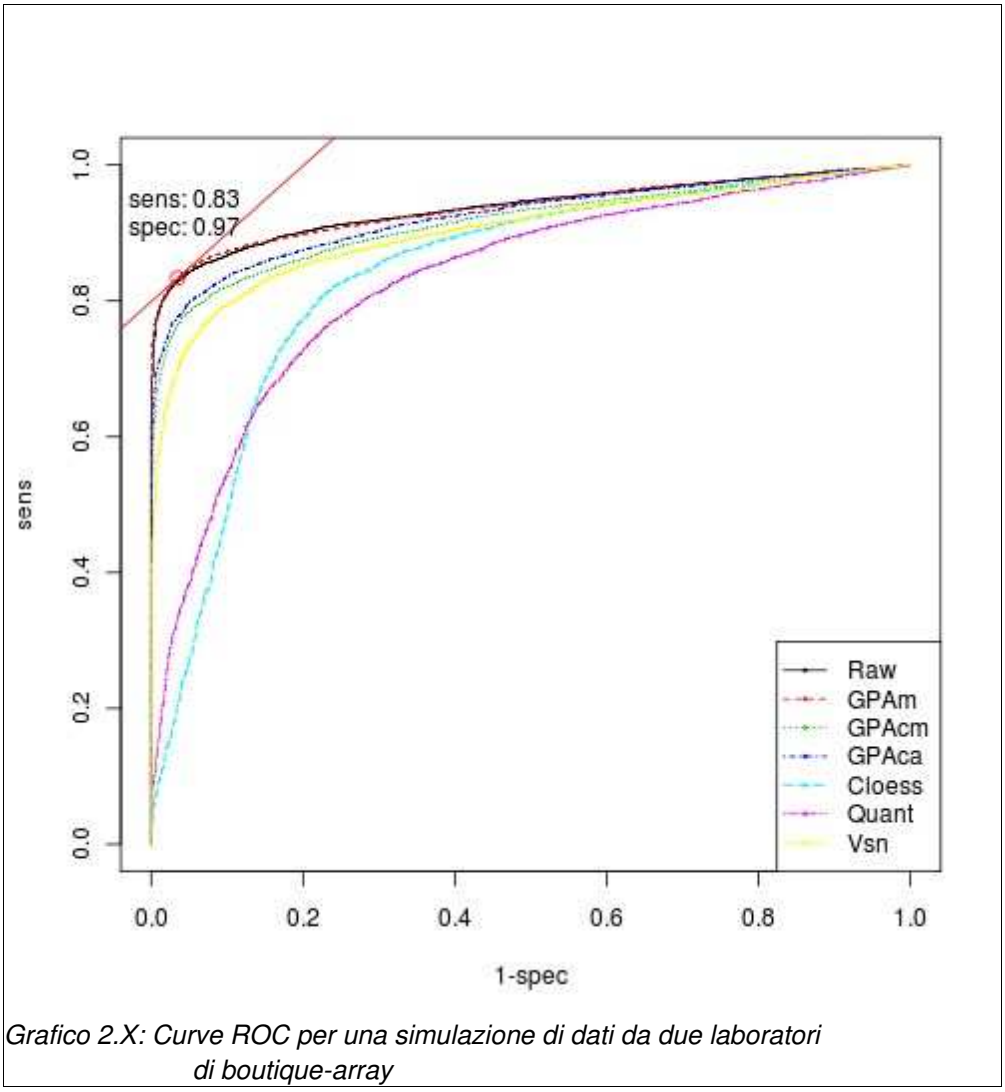


Grafico 2.X: Curve ROC per una simulazione di dati da due laboratori di boutique-array

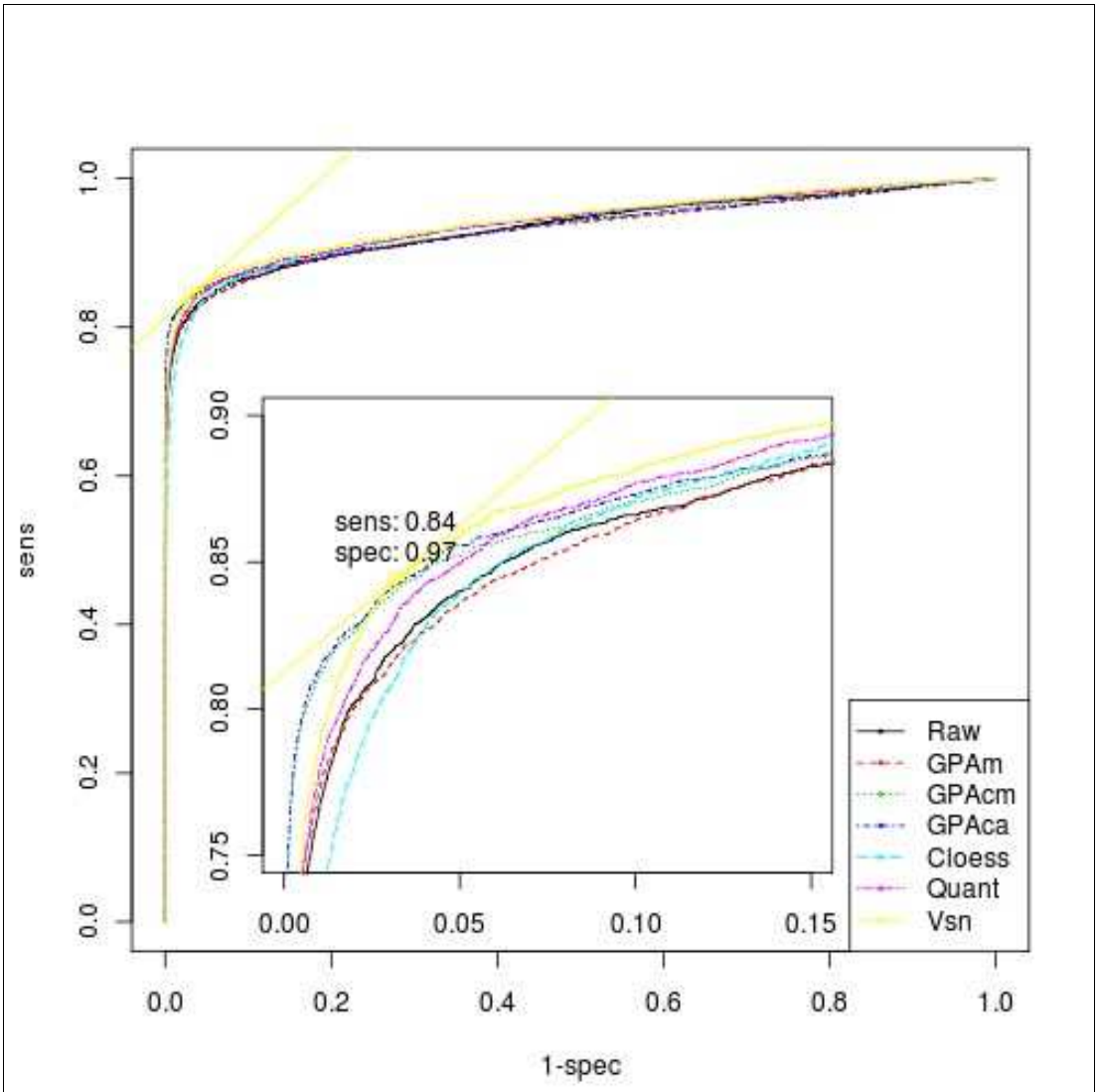


Grafico 2.XI: Curve ROC per dieci simulazioni di dati da due laboratori di boutique-array

2.2.3. Dati da due laboratori con sbilanciamento

Lo sbilanciamento tra sovra- e sottoespressi è una situazione abbastanza comune soprattutto in array piccoli che hanno molti geni geni DE. Il gran numero di DE insieme alla loro asimmetrica distribuzione tra sovra- e sottoespressi dovrebbe penalizzare tutte quelle normalizzazioni che invece si basano sulle assunzioni di poca e bilanciata espressione differenziale. In questo paragrafo si valuta la bontà delle normalizzazioni nel caso in cui il 49% di tutti i geni sono sovra, e il 21% sottoespressi.

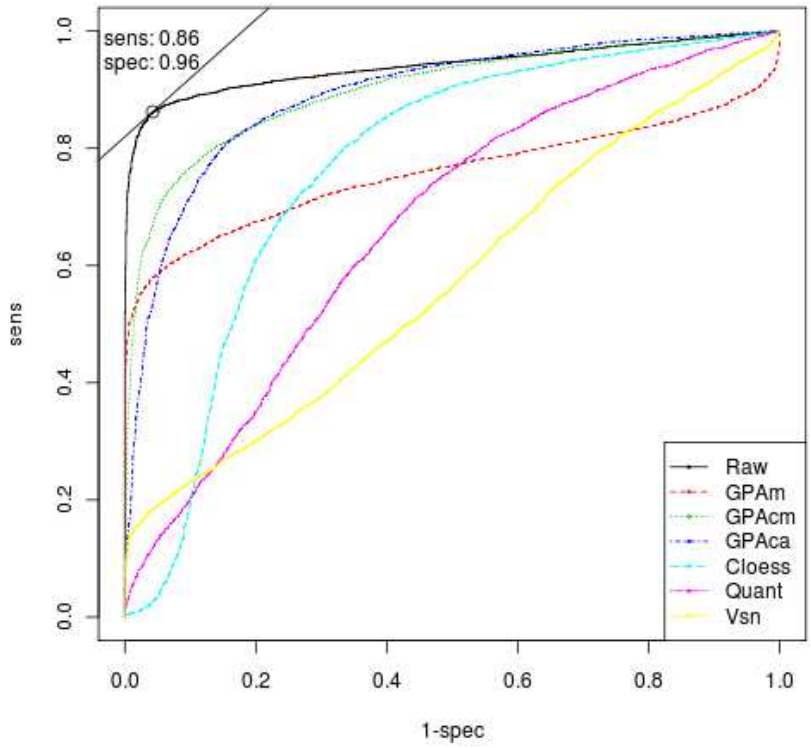


Grafico 2.XII: Curve ROC per una simulazione di dati da due laboratori di boutique-array con sbilanciamento

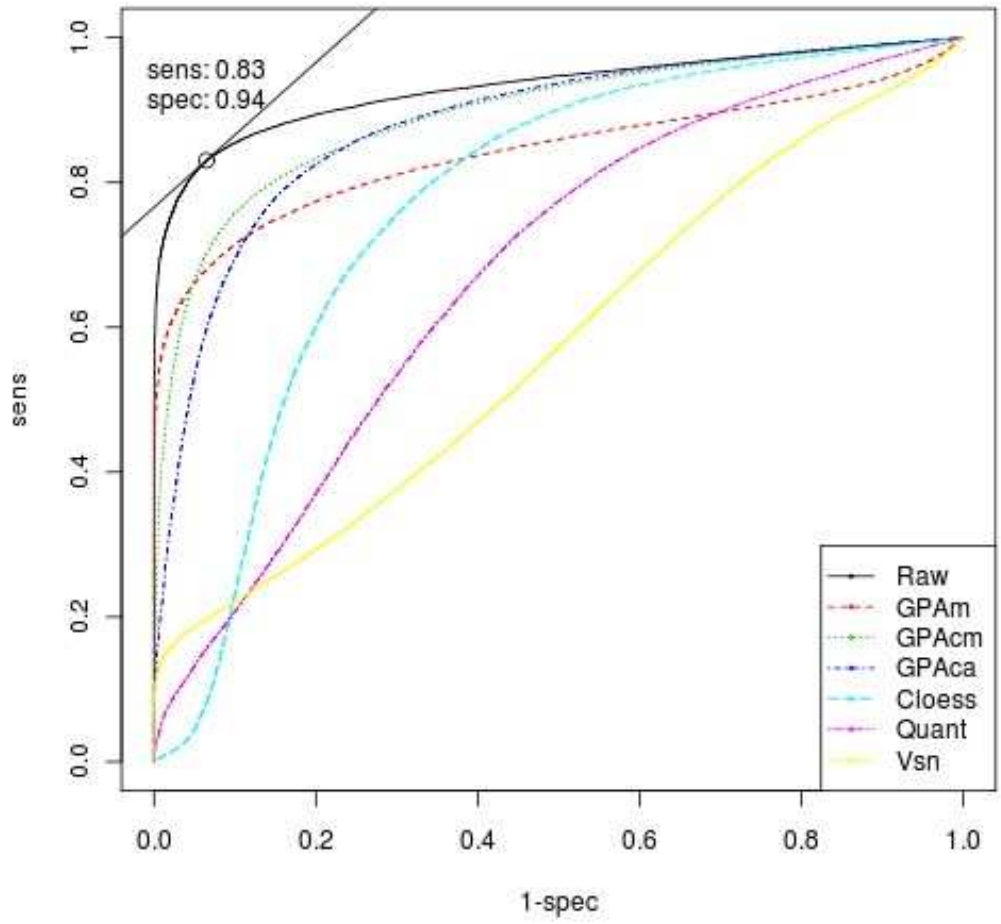


Grafico 2.XIII: Curve ROC per dieci simulazioni di dati da due laboratori di boutique-array, con sbilanciamento

Come atteso si nota come le differenze tra le normalizzazioni sono molto più marcate con conseguente peggioramento di tutte quelle che hanno come assunzione una equidistribuzione dell'espressione differenziale.

2.3. Dati reali

Come descritto nel *Paragrafo 1.3.3*, nell'ultima fase della valutazione della bontà delle normalizzazioni si fa uso di dati reali dove non è possibile calcolare gli errori di prima e seconda specie.

Per valutare il grado di plausibilità delle conclusioni che si traggono dal *test SAM* a seguito di ciascuna normalizzazione, si confrontano quindi la percentuale di geni individuati come DE che hanno un trend del tipo (1.22) o (1.23), coerente con la presunta espressione differenziale. Questa percentuale varia però al variare della soglia di confronto del valore osservato d_g, rispetto alla quale i geni vengono classificati come DE o EE.

Quello che si può fare è un confronto, tra i diversi metodi, della curva che descrive l'andamento del *True Discovery Rate* al variare del numero dei geni selezionati attraverso la statistica ordinata d_g. Il *True Discovery Rate* è definito come

$$TDR = \frac{\#(geni\ con\ trend\ corretto)}{\#(geni\ con\ d_g\ più\ alto)} \tag{2.2}$$

2.3.1. Dati da un singolo laboratorio

Per i dati provenienti da un solo laboratorio, costituiti da 17,416 geni e cinque ripetizioni per ciascuno dei quattro campioni, il grafico del *TDR* risultante è riportato di seguito (*Grafico 2.IX*)

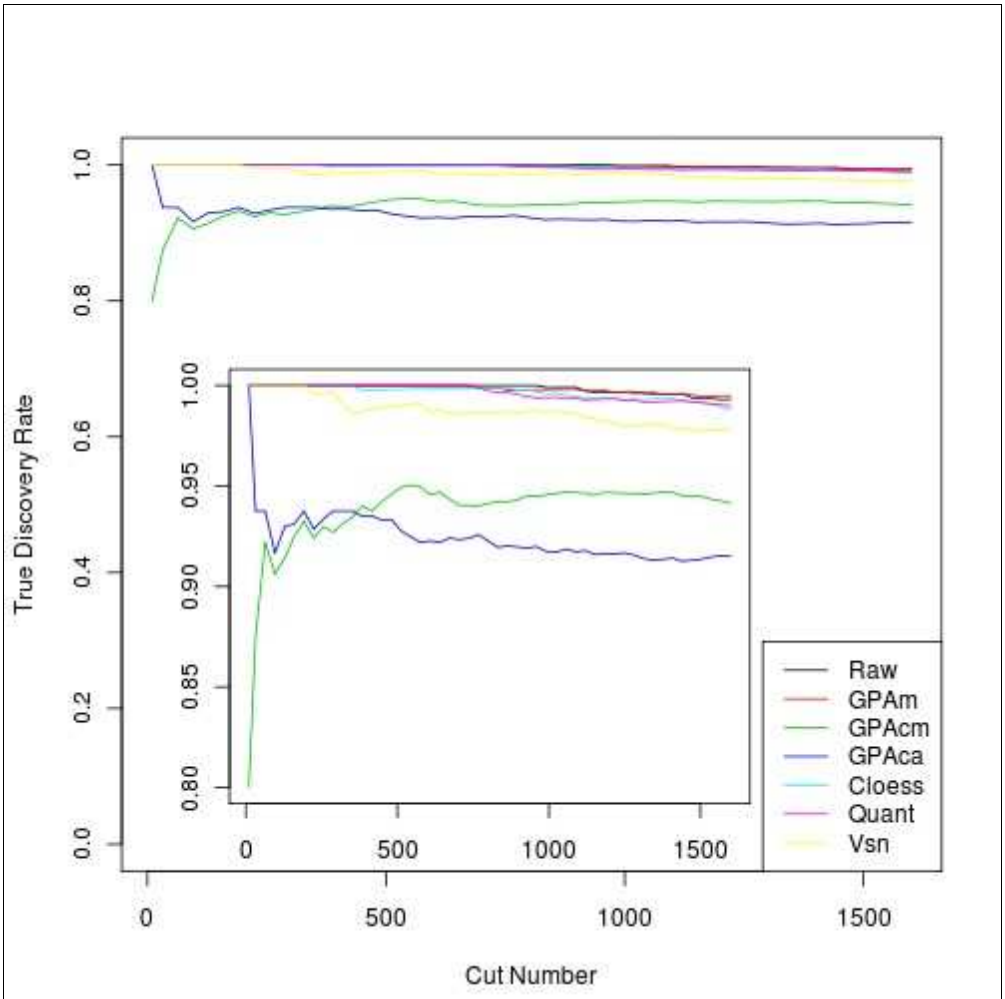


Grafico 2.XIV: True Discovery Rate delle sei normalizzazioni, al variare del numero di geni con più alto d_g, da classificare come DE; laboratorio singolo

2.3.2. Dati da due laboratori

I dati precedenti, se normalizzati congiuntamente ad un altro dataset di un altro laboratorio, danno i risultati del *Grafico 2.XV*.

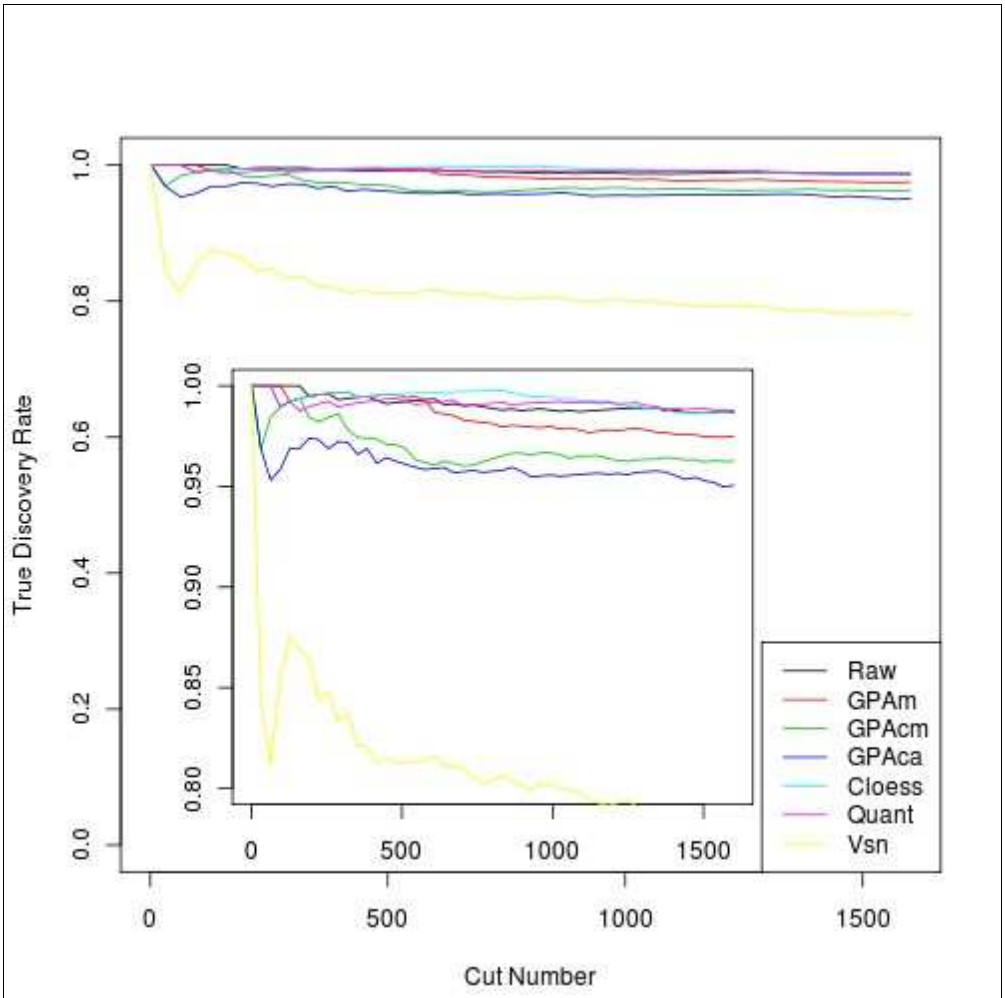


Grafico 2.XV: True Discovery Rate delle sei normalizzazioni, al variare del numero di geni con più alto d_g, da classificare come DE; due laboratori

3. Discussione e conclusioni

3.1. Discussione dei risultati

I risultati ottenuti con le analisi precedenti mostrano un funzionamento sostanzialmente buono dei metodi proposti, con alcune distinzioni a seconda delle situazioni.

3.1.1. Dati simulati secondo le assunzioni classiche

I dati simulati per un laboratorio con caratteristiche conformi alle classiche assunzioni dei metodi di normalizzazione mostrano livelli medi di intensità (e^{A}) che cadono tra $e^{3} \approx 20$ ed $e^{12} \approx 1{,}6 \cdot 10^{5}$. Le coppie di intensità presentano rapporti ($e^{\pm M}$) che in gran parte non superano proporzioni di 1:55 ($e^{4} \approx 55$)

Per questi dati le due *GPA cicliche*, *media* e *mediana*, generano *curve ROC* (grafici *2.III* e *2.IV*) uniformemente al di sopra di quelle di ogni altro metodo. La situazione ottima è raggiunta dalla versione *mediana* con un valore di sensibilità pari a 81% e di specificità pari a 92%.

Al contrario, la versione *mediana semplice* restituisce risultati peggiori rispetto ad ogni altro metodo e persino rispetto ai dati non normalizzati.

Queste differenze tra i diversi metodi, tuttavia, sono di ampiezza piuttosto ridotta, con differenze massime intorno al 10% di sensibilità a parità di specificità.

3.1.1.Dati simulati secondo le assunzioni classiche

I dati simulati da due laboratori diversi (*Grafico 2.V*) si distribuiscono su livelli di intensità medi piuttosto diversi: $(e^5 : e^{12}) \approx (150 : 10^5)$ il primo e $(e^3 : e^{12}) \approx (20 : 10^5)$ il secondo. Anche i rapporti tra intensità nelle coppie si attestano su livelli non del tutto simili: i primi raggiungono facilmente anche valori intorno ad $e^5 \approx 150$, i secondi difficilmente superano $e^4 \approx 55$.

In questa seconda situazione si vede (grafici 2.VI e *2.VII*) che la *GPA mediana* semplice raggiunge risultati apprezzabilmente migliori rispetto al caso precedente, risultando una delle normalizzazioni con *curva ROC* più alta.

Le *GPA cicliche* sembrano essere i metodi preferibili in presenza di pochi geni DE, quindi con sensibilità più contenute e specificità più alte, ossia nella parte sinistra del grafico, mentre nella parte destra la *Vsn* e la *GPA mediana* sembrano lavorare meglio .

Bisogna rilevare però che le differenze appena illustrate sono di entità minima e che complessivamente le curve sono notevolmente ravvicinate.

3.1.2. Dati simulati per boutique-array

I dati simulati con procedimento analogo al precedente, ma con una percentuale di geni DE non più del 5% ma del 70%, portano a risultati piuttosto diversi rispetto a quanto visto nel paragrafo precedente.

Nei grafici *2.VIII* e *2.IX* la curva dei dati grezzi è quella uniformemente più alta, mentre tutte le normalizzazioni, ad ogni livello di specificità alterano molto la struttura dei dati nascondendo o riducendo la differenziale espressione. Questo, ovviamente, non deve preoccupare: la distorsione dipendente dall'intensità con cui sono stati simulati i dati è di gran lunga inferiore alla differenziale espressione introdotta tra i due gruppi; in questo caso, infatti, i dati grezzi riescono già da soli ad identificare quasi tutti i DE.

Nel *Grafico 3.I* è mostrato un esempio di dati simulati per boutique-array in cui è evidente come la distinzione tra geni EE, in nero, e DE, in rosso, sia considerevole già nei dati grezzi sui quali perciò è possibile una analisi *SAM* con buoni livelli di successo. Le normalizzazioni, da parte loro, mostrano una correzione globale della distorsione della nuvola dei punti, senza però riuscire a separare maggiormente i due gruppi rispetto ai dati originali. Nelle applicazioni reali, però, le distorsioni dipendenti dall'intensità sono comunemente molto più forti rispetto alla simulazione e dunque questo problema non si presenta se non in dimensione ridotta.

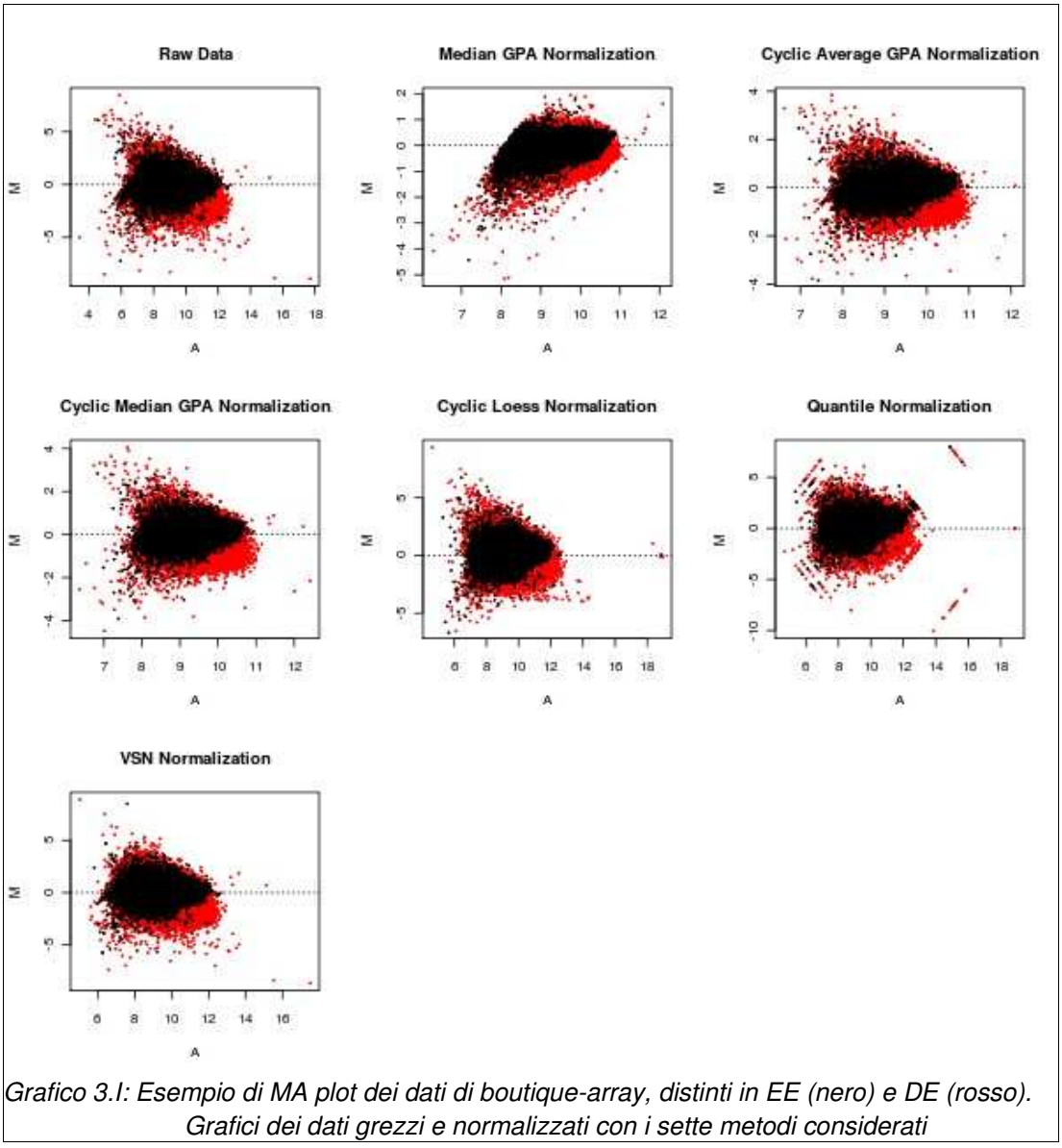


Grafico 3.I: Esempio di MA plot dei dati di boutique-array, distinti in EE (nero) e DE (rosso). Grafici dei dati grezzi e normalizzati con i sette metodi considerati

Le *curve ROC* mostrano in particolare risultati pessimi per la *Quantile*: a livello di specificità intorno al 10%, per esempio, questa ha una sensibilità nell'ordine del 20% contro il 77% dei grezzi che raggiungono in questo punto l'ottimo globale.

La *Loess Ciclica*, pur avendo un andamento generalmente non buono, raggiunge sensibilità piuttosto buone nella parte destra del grafico, quella di maggiore interesse, dato che ci si trova in condizioni di alta percentuale di geni DE.

La *GPA mediana* non ha un buon comportamento, in particolare è la peggiore nella zona destra, di maggiore interesse.

Le *GPA cicliche* al contrario sono, insieme alla *Vsn*, quelle che globalmente portano a risultati migliori e che, a differenza di questa, mantengono buoni livelli anche a destra.

Nel caso di una simulazione con dati provenienti da due laboratori (grafici *2.X* e *2.XI*) le migliorie introdotte dalle normalizzazioni sono impercettibili, rendendole sostanzialmente tutte equivalenti tra loro e simili ai dati non normalizzati.

Nel primo dei due grafici (*Grafico 2.X*), relativo ad una sola simulazione, si ritrovano *curve ROC* con andamenti simili a quelli individuati in precedenza, ad eccezione della *GPA mediana* che sembra operare meglio rispetto a prima. Questo miglioramento può essere riscontrato anche nel secondo grafico (*Grafico 2.XI*), relativo a dieci simulazioni dello stesso tipo di dati.

L'ultima simulazione è stata effettuata con boutique-array sbilanciati, cioè con prevalenza di sovraespressi tra i DE. Le *curve ROC* in questo contesto (grafici *2.XII* e *2.XIII*) sono in parte analoghe alle precedenti: le due *GPA cicliche* sono ancora le migliori, sia nel complesso, sia nella zona di interesse, quella destra.

La normalizzazione *Loess Cilcica* è, anche in questo caso, caratterizzata globalmente da un bassa performance, ma è tra le migliori a destra. La *Quantile* e la *GPA mediana* si comportano coerentemente con la situazione precedente, mentre l'unico metodo che risente visibilmente della nuova violazione delle assunzioni è la *Vsn*, che pare soffrire lo sbilanciamento molto di più dell'alta concentrazione di geni DE, pur essendo entrambe assunzioni di base di questo metodo.

Quindi, come atteso, nel caso in cui le classiche assunzioni tipiche dei dati di espressione vengano violate, le normalizzazioni che presentano sensibilità e specificità migliori sono le due versioni cicliche della GPA, proprio per la loro assenza di assunzioni.

3.1.3. Dati reali

Il *Paragrafo 2.3* mostra i risultati dell'analisi effettuata su dati reali per mezzo del metodo della titolazione di campioni biologici. Il *Grafico 2.XIV* riporta il *True Discovery Rate* per i dati di un singolo laboratorio normalizzati con i diversi metodi: la percentuale di successo dei metodi classici e della *GPA mediana* è indistinguibilmente vicina al 100% per ogni soglia di taglio. Le *GPA cicliche* invece presentano percentuali di errore più elevate, in particolare la versione *media*, ma pur sempre di piccola entità, tra il 5% e il 10%.

Il grado di attendibilità delle *GPA cicliche* sembra migliorare nel momento in cui ai dati precedenti si affianca un ulteriore insieme di dati, provenienti da un secondo laboratorio. Le percentuali di errore rilevabili nel *Grafico 2.XV* scendono su livelli che arrivano a sfiorare appena il 5%, valore minimo nel caso del laboratorio singolo. Al contrario, tutti gli altri metodi di normalizzazione presentano percentuali di successo che, pur sempre eccellenti, mostrano un calo rispetto a prima.

La *Vsn* è la normalizzazione che più di ogni altra presenta un crollo nella percentuale di geni correttamente classificati. Con livelli di errore che oscillano intorno al 20%, la *Vsn* suggerisce, per analogia con i risultati sui dati simulati, che i dati reali analizzati potrebbero presentare uno sbilanciamento tra sovra- e sottoespressi.

I dati reali utilizzati in questo elaborato fanno parte di un più ampio studio sperimentale presentato dalla *Food and Drug Administration* per la valutazione della riproducibilità dei microarray. Il disegno sperimentale e i protocolli utilizzati erano quindi molto standardizzati e molto riproducibili. Non c'è da stupirsi quindi che anche in questo i caso i dati non normalizzati presentino un elevato numero di andamenti corretti identificati. Le distorsioni presenti nei dati erano molto ridotte. I tipi di tessuti confrontati poi, *cervello* vs *pool di tessuti*, prevede che ci siano tra A e B molti geni DE ma bilanciati tra sovra- e sottoespressi. Quindi i risultati sono concordi con quelli trovati nei paragrafi *2.2.1* e *2.2.2*.

3.2. Conclusioni

I metodi di normalizzazione basati sull'*Analisi Procrastica Generalizzata* implementata in maniera ciclica forniscono risultati complessivamente comparabili con quelli dei metodi attualmente più diffusi, in diverse situazioni.

Inoltre, grazie all'assenza di assunzioni di tipo statistico e probabilistico circa la natura dei dati, le *GPA cicliche* risultano sempre più convenienti via via che ci si allontana da quelle che sono le assunzioni su cui, al contrario, sono basate le altre normalizzazioni. In generale le *GPA cicliche* possono essere utilizzate senza tenere conto della forma e della struttura dei dati.

Dal presente studio non emergono differenze apprezzabili tra le due versioni cicliche della *GPA*, quella *media* e quella *mediana*. Rimangono le considerazioni preliminari di cui si è discusso nel *Paragrafo 1.2.3.iii* che possono far preferire una all'altra a seconda del contesto sperimentale.

La *GPA mediana* semplice, al contrario, sembra dare risultati molto spesso deludenti o al più comparabili con quelli di altri metodi nelle situazioni meno classiche, con il solo vantaggio, seppur non trascurabile, della maggiore agilità computazionale.

Bibliografia

Balagurunathan Y, Dougherty ER, Chen Y, Bittner ML, Trent JM (2002) Simulation of cDNA microarrays via a parameterized random signal model . *Journal of Biomedical Optics* **7(3)**: 507–523.

Berge JMFT (1977) Orthogonal Procrustes Rotation for Two or More Matrices. *Psychometrika,* **42(2)**:267-276.

Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19(2)**: 185-193.

Chu, G., Narasimhan, B., Tibshirani, R. and Tusher, V. (2001) *Significant analysis of microarrays. Users Guide and Technical Document*, Department of Biological Science, University of Tulsa, USA.

Durbin BP, Hardin JS, Hawkins DM, Rocke DM (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18**(Suppl 1): S105–S110.

Goodal C (1991) Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society B*, **53(2)**: 285-339.

Gower JC (1975) Generalized procrustes analysis. *Psychometrika*, **40**: 33-55.

Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**(Suppl 1): S96-S104.

Jiang H, Deng Y, Chen HS, Tao L, Sha Q, Chen J, Tsai CJ, Zhang S (2004) Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes , *BMC Bioinformatics*, **5**: 81.

Park T, Yi SG, Kang SH, Lee S, Lee YS, Simon R (2003) Evaluation of normalization methods for microarray data. *BMC Bioinformatics,* **4**: 33.

Ramasamy A, Mondry A, Holmes CC, Altman DG (2008) Key Issues in Conducting a Meta-Analysis of Gene Expression Microarray Datasets, *PLoS Med* **5(9)**: e184

Risso D (2008) *Analisi dei dati di espressione genica: studio comparativo per la valutazione dell'impatto della normalizzazione sull'inferenza statistica*, Facoltà di Scienze Statistiche, Università di Padova.

Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, Luo Y, Sun YA, Willey JC, Setterquist RA, Fischer GM, Tong W, Dragan YP, Dix DJ, Frueh FW, Goodsaid FM, Herman D, Jensen RV, Johnson CD, Lobenhofer EK, Puri RK, Scherf U, Thierry-Mieg J, Wang C, Wilson M, Wolber PK, Zhang L, Amur S, Bao W, Barbacioru CC, Lucas AB, Bertholet V, Boysen C, Bromley B, Brown D, Brunner A, Canales R, Cao XM, Cebula TA, Chen JJ, Cheng J, Chu TM, Chudin E, Corson J, Corton JC, Croner LJ, Davies C, Davison TS, Delenstarr G, Deng X, Dorris D, Eklund AC, Fan XH, Fang H, Fulmer-Smentek S, Fuscoe JC, Gallagher K, Ge W, Guo L, Guo X, Hager J, Haje PK, Han J, Han T, Harbottle HC, Harris SC, Hatchwell E, Hauser CA, Hester S, Hong H, Hurban P, Jackson SA, Ji H, Knight CR, Kuo WP, Leclerc JE, Levy S, Li QZ, Liu C, Liu Y, Lombardi MJ, Ma Y, Magnuson SR, Maqsodi B, McDaniel T, Mei N, Myklebost O, Ning B, Novoradovskaya N, Orr MS, Osborn TW, Papallo A, Patterson TA, Perkins RG, Peters EH, Peterson R, Philips KL, Pine PS, Pusztai L, Qian F, Ren H, Rosen M, Rosenzweig BA, Samaha RR, Schena M, Schroth GP, Shchegrova S, Smith DD, Staedtler F, Su Z, Sun H, Szallasi Z, Tezak Z, Thierry-Mieg D, Thompson KL, Tikhonova I, Turpaz Y, Vallanat B, Van C, Walker SJ, Wang SJ, Wang Y, Wolfinger R, Wong A, Wu J, Xiao C, Xie Q, Xu J, Yang W, Zhang L, Zhong S, Zong Y, Slikker W., Jr (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, **24**:1151–1161.

Shippy R, Fulmer-Smentek S, Jensen RV, Jones WD, Wolber PK, Johnson CD, Pine PS, Boysen C, Guo X, Chudin E, Sun YA, Willey JC, Thierry-Mieg J, Thierry-Mieg D, Setterquist RA, Wilson M, Lucas AB, Novoradovskaya N, Papallo A, Turpaz Y, Baker SC, Warrington JA, Shi L, Herman D (2006) Using RNA sample titrations to assess microarray platform performance and normalization techniques. *Nature Biotechnology*, **24:** 1123-1131

Smyth GK, Speed T (2003) Normalization of cDNA microarray data. *Methods,* **31(4)**: 265-273.

Wilson DL, Buckley MJ, Helliwell CA, Wilson IW (2003) New normalization methods for cDNA microarray data. *Bioinformatics,* **19(11)**: 1325-1332.

Wit E, McClure J (2004) *Statistics for microarrays: design, analysis and inference*. Wiley.

Xiong H, Zhang D, Martyniuk CJ, Trudeau VL, Xia X (2008) Using Generalized Procrustes Analysis (GPA) for normalization of cDNA microarray data , *BMC Bioinformatics*, **9**: 25.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, **30(4)**: e15.

Zien A, Aigner T, Zimmer R, Lengauer T (2001) Centralization: a new method for the normalization of gene expression data. *Bioinformatics*, **17**(Suppl 1): S323-S31.